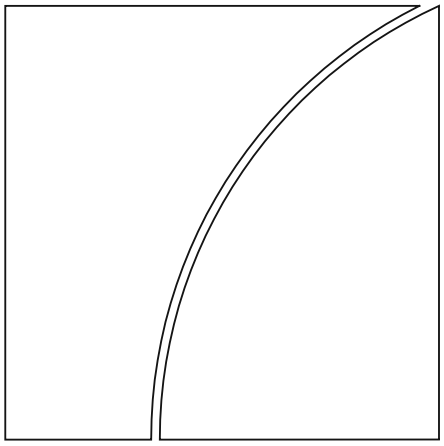
Irving Fisher Committee on
Central Bank Statistics

BIS



IFC Bulletin

No 67

Data science in central banking: exploring generative AI

March 2026

Contributions in this volume were prepared for the proceedings of the workshop on "Data science in central banking", co-organised by the Bank of Italy and the Irving Fisher Committee (IFC) in Rome, Italy, 18-20 February 2025.

The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

This publication is available on the BIS website (www.bis.org).

© *Bank for International Settlements 2026. All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1991-7511 (online)

ISBN 978-92-9259-936-2 (online)

Data science in central banking: exploring generative AI

IFC Bulletin No 67

March 2026

Proceedings of the workshop on "Data science in central banking", co-organised by the Bank of Italy and the Irving Fisher Committee (IFC) in Rome, Italy, 18-20 February 2025.

Overview

Generative artificial intelligence in central banking

Douglas Araujo, Research Economist, Central Bank of Brazil

Giuseppe Bruno, Director, Economics and Statistics Directorate, Bank of Italy

Adam Cap, Senior Macroeconomic Analyst, MED, BIS

Juri Marcucci, Economist, Economics and Statistics Directorate, Bank of Italy

Rafael Schmidt, Head of IT, MED, BIS

Olivier Sirello, Senior Statistical Analyst, MED, BIS

Bruno Tissot, Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat

Opening remarks

Alberto Naudon, IFC Chair, Central Bank of Chile

Keynote speech

Tailoring small language models for enterprise use cases

Julien Simon, Arcee.ai

Special feature

Data science: the role of statisticians

Elisabetta Carfagna, International Statistical Institute

1. Adoption of generative AI in central bank operations

IFC survey on central banks' use of AI and ML

Alberto Naudon, IFC Chair, Central Bank of Chile

Innovation project: large language model adaptation at Latvijas Banka

Mārtiņš Liberts and Rihards Ostrovskis, Bank of Latvia

A natural language processing toolbox for the National Bank of Romania

Claudia Voicilă and Daniel Șerbu, National Bank of Romania

2. Natural language processing tools for enhanced text analysis

Transforming survey analysis: tools for central banks

Nicholas Gray and Dominic Jones, Reserve Bank of Australia

From text to insights: AI-based topic modeling for central banking applications

Marlena Pietsch, Diana Sola, Felix Marx, Michaela Vollmer and Babak Zandi, Deutsche Bundesbank

Efficacy of AI RAG tools for complex information extraction and data annotation tasks: a case study using banks' public disclosures

Nicholas Botti, Flora Haberkorn, Charlotte Hoopes and Shaun Khan, Board of Governors of the Federal Reserve System

3. Generative AI for summarisation and information extraction

BankGPT: the use of large language models in official communications

Claudia Biancotti, Carolina Camassa, Marco Fruzzetti, Luigi Palumbo and Myriam Portaluri, Bank of Italy

Extracting economic issues from news data: with the help of generative AI

Younghwan Lee, Bank of Korea

What skills are employers looking for? Applying generative AI to Danish job postings

Per Andersen and Erik Grenestam, Danmarks Nationalbank

4. Text analysis for market monitoring and monetary policy purposes

The monetary policy statement database

Cory Baird, University of Tokyo, Jonathan Benchimol, Bank of Israel, Vira Vyshnevska and Wook Sohn, KDI School, and Iegor Vyshnevskyi, Sogang University

Read between the headlines: can news data predict inflation?

Alan Chester Arcin, Ma. Ellysah Joy Guliman, Genna Paola Centeno, Jacqueline Margaux Herbo, Sanjeev Parmanand and Cherrie Mapa, Bangko Sentral ng Pilipinas

The impact of monetary surprises on exchange rates: results from textual and high-frequency analysis

Jean-Charles Bricongne, Bank of France, and Louis Marolleau, Lille University

5. Using text analysis for novel economic insights

Turbulent times: capturing social unrest in the news

Sydney Eck, Flora Haberkorn, April Patel, Zach Proom and Abhinuv Uppal, Federal Reserve Board of Governors

Indexing and visualization of climate narratives using BERT and causal extraction

Noriyasu Kaneda, Bank of Japan, and Hiroki Sakaji, Hokkaido University

6. AI in supervisory technology and financial regulation

Gen AI in action: transforming data use in suptech

Patrick Hoffmann, Maha El Dimachki and Sivasubramanian Ramanathan, BIS Innovation Hub, Kenneth Gay and Harry Lee, Monetary Authority of Singapore

AI-driven legal summarization for financial regulation monitoring

Diana López Avilés and Isidora Díaz Jorquera, Central Bank of Chile

A novel multi-step prompt approach for LLM-based Q&As on banking supervisory regulations

Daniele Licari, Canio Benedetto, Daniele Bovi, Alessandro De Gregorio and Marco De Leonardis, Bank of Italy, Praveen Bushipaka and Tommaso Cucinotta, Sant'Anna School of Advanced Studies

7. Advanced forecasting and data analytics techniques

Project Neo: gaining economic insights with AI and novel data

Rachel Arulraj-Cordonier, Marko Nanut Petric and Flurina Strasser, BIS Innovation Hub

Enhancing external debt nowcasting: a data-driven approach using advanced analytics

Candra Puspita Dewi, Aditya Wisnugraha Sugiyarto and Sony Panji Wicaksono, Bank Indonesia

Generalised weighted framework for synthetic data evaluation

Chiung Ching Ho, Central Bank of Malaysia

From CPU to GPU: comparison of parallel computing techniques for numerical solutions

Maxim Ralchenko, Oleksandr Shcherbakov and Kim P Huynh, Bank of Canada

Istanbul electricity demand forecast with artificial neural networks

Hayriye Yasak Ozkal, Central Bank of the Republic of Türkiye, Buse Kaylan, Meltem Sipahi and Muammer Altan Cakir, Istanbul Technical University

Forecasting influenza-like illness in Italy using Wikipedia: a principal components regression approach

Gianluca Mura, Bank of Italy

A foundational model for macroeconomic times series forecasting and nowcasting

Batuhan Koyuncu, Byeungchun Kwon, Marco Lombardi, Fernando Perez-Cruz and Hyun Song Shin, BIS

8. Data privacy and anonymisation

Data anonymization principles at Banco de Portugal

Ana F Carvalho, Francisco Fonseca, Mário Lourenço and Ricardo Marques, Banco de Portugal

Gen AI in payments: secure synthetic data and more

Ajit Desai, Colin Jones and Masoud Nasari, Bank of Canada

Regression on encrypted datasets using fully homomorphic encryption

Adriano Baldeschi and Giuseppe Bruno, Bank of Italy, Mohammad Raeini, Paul Master and Vineet Chadha, Cornami

9. Metadata and data integration with SDMX

Metadata integration as a key enabling technique for AI and ML

Thomas Gottron and Daniel Suranyi, European Central Bank

SDMX chatbot: a trustworthy assistant to navigate the SDMX documentation

Samuel Melm and Brian Buffett, BIS

sdmx.io: where tools and learning resources come together to address official statistics use cases

Douglas Araujo, Central Bank of Brazil, and Brian Buffett, BIS

SDMX enabled users' processes at the Bank of Italy

Attilio Mattiocco and Valentino Pinna, Bank of Italy

10. AI assisted data search and retrieval

Supporting users in seeking data on the ECB data portal: a use case for retrieval augmented generation

Luca Petracca, Simone De Benedictis, Thomas Gottron and Zlatina Hofmeister, European Central Bank

Using retrieval-augmented generation (RAG) and LangChain to drive chatbots for interactive economic data analysis and insight

Seah Boon Keong, Eilyn Chong and Max Chew, Central Bank of Malaysia

Leveraging generative AI for granular credit data utilization: a multi-agent approach

Nontawit Cheewaruangroj, Kawinwish Laobundit, Peranut Nimitsurachat, Supachai Saengthong, Skunpoj Thanarojsophon and Anak Yodpinyanee, Bank of Thailand

Closing remarks

Paolo Angelini, Bank of Italy

Irving Fisher Committee on Central Bank Statistics

BIS

Generative artificial intelligence in central banking

Douglas Araujo, Giuseppe Bruno, Adam Cap, Juri Marcucci, Rafael Schmidt, Olivier Sirello, Bruno Tissot[1]

Executive summary

Central banks are increasingly exploring and using *generative artificial intelligence* (generative AI)[2] to support their activities, such as economic and monetary analysis, statistical production, financial supervision and payment oversight. *Large language models* (LLMs) in particular – a class of generative AI that can analyse and produce text by mimicking human intelligence – have unlocked unprecedented opportunities to deal with textual data, of which central banks are both heavy users and producers.

Continuous and rapid innovations are further expanding the capabilities and quality of generative AI's outputs. First, methods such as *retrieval-augmented generation (RAG)* can improve the accuracy and reliability of LLM-produced content, for example by drawing on verified sources and specialised knowledge. Second, a growing area of interest is the use of *small language models* (SLMs), which are more limited and cheaper versions of LLMs tailored to specific tasks. Third, and more generally, *agentic AI* systems can enhance task automation and assist decision-making in business processes, for example by orchestrating complex workflows.

With the proliferation of new data solutions and competing work priorities, **one question for central banks is how to make the most of these promising yet often embryonic AI advances in an efficient, effective, ethical and safe way**. Experience so far has highlighted the importance of adequately managing innovation, not least to balance its benefits with the associated risks. Three distinct areas of focus have emerged:

- The first relates to the design and implementation of **proper governance frameworks** for managing the associated risks and harnessing AI effectively and responsibly. Central banks have already taken significant steps in this endeavour,

[1] Respectively, Research Economist, Central Bank of Brazil (douglas.araujo@bcb.gov.br); Director, Economics and Statistics Directorate, Bank of Italy (giuseppe.bruno@bancaditalia.it); Senior Macroeconomic Analyst, Monetary and Economic Department (MED), Bank for International Settlements (BIS) (adam.cap@bis.org); Economist, Economics and Statistics Directorate, Bank of Italy (juri.marcucci@bancaditalia.it); Head of IT, MED, BIS (rafael.schmidt@bis.org); Senior Statistical Analyst, MED, BIS (olivier.sirello@bis.org); Head of BIS Statistics, MED, BIS, and Head of the IFC Secretariat (bruno.tissot@bis.org).

The views expressed here are those of the authors and do not necessarily reflect those of the BIS, the IFC, the Bank of Italy or any of those institutions mentioned.

The authors thank Michele Bullock, Michael Machuene Manamela, Gloria Peña, Luís Teles Dias and colleagues from the Reserve Bank of Australia, Banco de Portugal and South African Reserve Bank for helpful comments and suggestions. They are also grateful to Alison Arnot for editorial review and Márcia Cavalinhos for publication support.

[2] Expressions formatted in *italics* are further defined in the Glossary in the annex.

leveraging their long-standing expertise in data issues. Yet AI governance initiatives are still evolving, often reflecting the need for new skills and organisational strategies to support organisation-wide innovation.

- The second area of focus is on technical aspects, reflecting the need to **secure adequate IT resources as well as to adopt interoperable data processes and systems** to further advance the use of generative AI both within and across organisations. Open software tools and well-established information standards such as the Statistical Data and Metadata eXchange (SDMX) initiative can play a decisive role in this regard. They can ensure that data are used properly by AI and, more broadly, preserve the quality of the reference information produced by authoritative sources, such as central banks.
- Finally, promising yet often fast-paced technological advancements serve as a good reminder of the **importance of cooperation and knowledge exchange across central banks as well as with the other stakeholders involved in the data ecosystem**. Such collaborative initiatives can yield significant benefits, including sharing best practices and raising awareness of the challenges posed by AI. They also help to optimise the use of resources in the central banking community, by facilitating opportunities for co-investment in areas such as data-sharing techniques or open source software, to make data AI ready.

1. Introduction

Central banks are increasingly using generative AI to support their various activities (BIS (2025), IFC (2025a)). AI is commonly understood as computer systems capable of tasks that normally require human intelligence, while its recent generative form (also known as generative AI) produces new information content based on patterns learned from existing data.[3]

The versatility of most *generative AI* applications makes them suitable for a broad range of central banking activities that rely on extensive data usage. These typically span not only economic and monetary analysis and forecasting, supervision and payment oversight but also financial education and consumer protection. In particular, central banks have been active adopters of LLMs – a prominent class of generative AI models able to generate human-like text (IFC (2022, 2023), Araujo et al (2023)). Generative AI can also assist central banks in their role as data producers, by collecting, processing and disseminating reference statistics.

Yet, in practice, **the use of AI is not without challenges, especially related to data quality and IT infrastructure**. Regarding the first aspect, the main concern lies in preserving the high quality, credibility and integrity of central banks' data, while ensuring their accuracy, explainability and interpretability as well as addressing broader privacy, security and ethical considerations (Dilip et al (2026)). Turning to IT aspects, the infrastructure needed typically demands secure, scalable and high-performing solutions, often entailing additional costs and complex trade-offs – for example related to the use of cloud-based services, information protection and power

[3] See the Glossary for a list of more exhaustive definitions.

capabilities. Fortunately, a number of solutions are being continually developed and refined to address those challenges, ranging from models requiring less power to privacy-preserving methods. But keeping pace with these various evolutions can be cumbersome, adding uncertainty and volatility that may detract from the long-term foresight often characterising central bank activities.

To balance such opportunities and challenges, central banks have been advancing a number of strategies. First, the development of agile, resilient and clear governance frameworks has proved instrumental for mitigating the operational and ethical risks associated with generative AI (IFC (2021, 2025a, 2025b)). Second, central banks have been further upskilling their workforce towards AI-savvy profiles, while also capitalising on the broad expertise of their staff based on interdisciplinary background, scientific rigour and attention paid to data quality. Third, the increasing yet often still experimental use of generative AI has underscored the benefits of international cooperation, notably for exchanging best practices, raising awareness of the cross-cutting issues posed by AI and developing shared solutions.

Against this backdrop, the IFC, along with the Bank of Italy, has been organising recurrent data science workshops to review developments in the big data ecosystem and the ongoing adoption of innovative data techniques. The fourth edition in 2025 emphasised the **exploration of generative AI and its potential applications in central banking**. The overview of this *IFC Bulletin* provides a bird's eye view of the latest developments related to generative AI (section 2) and discusses related applications of interest to central banking (section 3). The various projects already being advanced also suggest a number of ways to make the most of the opportunities offered by this technology (section 4), especially regarding governance aspects, IT resources, information standards and data quality, as well as international cooperation.

2. Generative AI: a bird's eye view of recent developments

Recent and continuous technological advances in generative AI have opened new opportunities that are particularly relevant for central banks. First, RAG systems have emerged as a possible solution to meet the growing demand for tailoring LLMs' outputs to organisation-specific knowledge. Second, the development of *agentic AI* allows for flexible IT programming, supporting task automation solutions based on natural language capabilities instead of "traditional", hard-coded instructions. Third, SLMs effectively reduce the computational burden that is typical of large AI models, while safeguarding the relevance of their outputs.

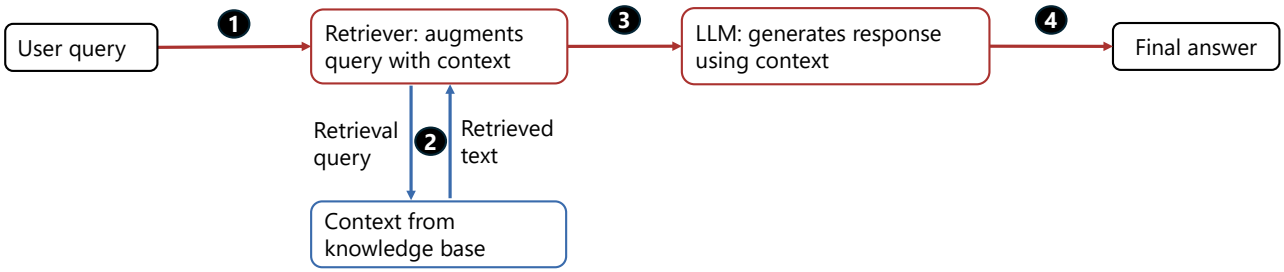
Retrieval-augmented generation (RAG) systems

RAG systems can enhance the quality and reliability of LLMs' outputs by integrating selected specialised documents forming the input to these models. This is in contrast with general purpose models, which typically answer user questions after first being trained on publicly available knowledge. In practice, LLMs are able to base their answers on specific and/or internal documents selected during the initial

setup phase, forming the "*knowledge base*" of the RAG (Graph 1).[4] As a result, when a user poses a question, the system will provide information derived exclusively from the customised knowledge base.

RAG simplified architecture model

Graph 1



Source: Authors' elaboration.

This approach has at least four advantages. First, it can significantly improve the quality, reliability and relevance of AI-generated answers by restricting the knowledge base, for instance to only verified and authentic sources. This makes RAGs well suited for use cases that require dealing with specialised and possibly restricted information – for example, financial supervisors may be interested in accessing data based on non-public bank assessments.[5] Second, RAGs can help control the quality of input information and thereby reduce the risk of false but seemingly true answers ("hallucinations"). This is a significant limitation of general purpose LLMs given the risk of producing faulty output because of incorrect or poor-quality input (the "garbage in garbage out" phenomenon). Third, they allow for more traceability, as input sources can be explicitly referenced and authenticated. Finally, RAG-based approaches can be simple and effective to deploy, although they require adequate IT infrastructure to operate. This advantage is especially notable in comparison to other techniques that similarly try to control the quality of models' outputs, for example by fine-tuning their input or customising their specifications – often entailing complex requirements and a large amount of computing resources.

Agentic AI

AI agents are language model (LM)-based autonomous and specialised software systems designed to automate tasks involving complex reasoning and decision-making. They can dynamically interact with data, users and external systems to perform various tasks such as workload scheduling, answering user queries and filtering emails.[6] While a key advantage of RAGs is allowing for improved accuracy

[4] See Lewis et al (2020) for a more exhaustive description of RAG systems.

[5] Specialised information may also encompass vast collections of archival records, which central banks often curate. RAG systems may be useful to effectively leverage such resources, as recently shown by the Reserve Bank of Australia (Bullock (2025)).

[6] While AI agents typically have a moderate and task-specific scope, "agentic AI" is characterised by greater autonomy in handling complex use cases. Its development has been further supported by the so-called Model Context Protocol (MCP), an open source universal interface to help AI agents better use information resources and software tools (see Box A).

and factual reliability in models' outputs by supplying relevant context at query time, they react to specific user queries and are limited to generating informed responses. AI agents, by comparison, can orchestrate processes and tools, execute various tasks and, in certain cases, feature learning capabilities that are essential for supporting decision-making. These extend beyond – at least to date – the simpler operations performed by most generative AI models that focus on transforming input information.

AI agents can be particularly useful in the context of central banking. For example, in the area of statistical compilation, one agent in charge of quality checks can help automate the detection of potentially erroneous values based on past patterns and quality checks. Meanwhile, another AI agent can estimate missing values, for instance based on historical trends and previous observations. Finally, a third AI agent may be used to check the accuracy of these estimates, supporting human validation tasks as needed.

Despite its advantages, experience suggests that **agentic AI requires properly designed workflows as well as human oversight** to be fully effective. On the one hand, it appears to work well with simple and modular workflows, for instance to optimise the use of various LMs' capabilities (Anthropic (2024)). On the other hand, rigorous human oversight, systematic testing, adequate validation and training as well as monitoring of agentic systems often prove critical to obtain high-quality outputs. Well-governed data and AI management processes therefore appear particularly important, as they play a decisive role in ensuring that security, privacy, accountability and broader ethical concerns are met.

Small language models (SLMs)

SLMs are language models that are significantly smaller than LLMs, trading off size for specialisation. Their limited number of parameters makes them faster, more efficient and less computationally intensive than larger models. As a result, SLMs are more conveniently fine-tuned to specific tasks. Their specific focus can help provide more relevant responses, while still having the ability to reply in natural language similar to larger LLMs.[7]

SLMs can offer notable advantages in the context of central banking, for at least two reasons. First, because of their relatively limited size, they can be fine-tuned locally, for example using on-premises IT infrastructure. This is particularly important for mitigating risks related to third-party dependency or confidentiality breaches. Second, SLMs excel in various specialised or detailed tasks that are essential components of many traditional workflows in central banks, such as in the areas of statistical compilation or micro-level supervision. Moreover, SLMs can be combined with LLMs in multi-agent frameworks, where SLMs focus on domain-specific tasks, while LLMs handle broader, general purpose requests.

[7] SLMs are trained through various techniques, such as "pruning" – ie where the architecture and size of the original LLM are simplified and reduced – and "knowledge distillation", where knowledge from a larger, more powerful "teacher" LLM is transferred into the smaller model. This latter technique has been used, for instance, to train the model DeepSeek-R1 (DeepSeek-AI et al (2025)), achieving comparable performance to OpenAI models but at a fraction of their size and cost.

3. Generative AI applications in central banking: main use cases

Generative AI has emerged as a versatile tool for supporting central banks across various tasks. Reported experience suggests that most use cases relate to four main areas: (i) economic and monetary analysis; (ii) forecasting exercises; (iii) statistical production; and (iv) supervision and oversight activities.

Economic and monetary analysis

One of the most promising areas where generative AI can support central banks is in **economic and monetary analysis, which is a fundamental area supporting the conduct of their monetary and financial stability policies**. Concrete applications include text analysis, assessing economic news and sentiment and performing classification and summarisation tasks.

First, **regarding text analysis**, **LLMs are unlocking new opportunities for monitoring economic and financial developments** by leveraging a wider range of economic indicators (Aprigliano et al (2023), Marcucci (2024)). For example, they facilitate a more structured analysis of press sources to provide timely insights on financial market developments, as shown by the Bank of Korea. These models can also support analysis of the real economy, with arguably greater detail and agility than traditional methods (Dahlhaus et al (2025)). The Danmarks Nationalbank, for instance, uses LLMs to analyse job postings and track movements in employer demand for specific skills over time, offering a more nuanced view of the labour market. Similarly, the Bank of Japan and the Hokkaido University have deployed LLMs to monitor climate-related news, providing useful insights in terms of international developments, regulatory aspects and business practices. News and other similar text-based sources can also inform risk assessments that can be out of the scope of traditional monitoring exercises. A notable example relates to episodes of social unrest, as shown by the Federal Reserve Board, which leverages voluminous data from the Global Data of Events, Language and Tone (GDELT) project.[8]

Second, **generative AI can help users gain a better grasp of "soft" indicators, such as economic sentiment and expectations**. This can be especially relevant for central banks when analysing economic agents' behaviour, for instance, to assess the influence of their policy communication and monitor feedback dynamics. For example, the National Bank of Romania developed a language sentiment index for measuring the impact of financial news, using OpenAI's ChatGPT complemented with a model specialised in financial and monetary policy statements (Araci (2019), Evdokimova et al (2023)). Similarly, a joint study by the Bank of France and the University of Lille investigates financial news related to surprises and expectations close to monetary policy meetings, relating them to the volatility of foreign exchange rates. Expectations by economic agents are also – and very often – captured by surveys, which often produce large volumes of unstructured text that need to be adequately integrated into policy frameworks. Case in point, the Reserve Bank of

[8] The GDELT aims to ensure comprehensive coverage of global news media to monitor events in near real time through sophisticated algorithms; see Leetaru and Schrodt (2013) and gdeltproject.org.

Australia shows that generative AI can help process these data with greater efficiency and ease.

Third, **another important aspect when dealing with vast amounts of textual documents is to be able to extract meaningful information for policy purposes.** This calls for adequate classification and summarisation processes, something that generative AI excels at (Zhang et al (2025)). Regarding classification, the Deutsche Bundesbank has developed an internal tool based on LLMs to help users categorise documents easily. The approach, using a bootstrapping resampling technique, allows for greater agility and lower computational and analytical costs. Turning to the capacity of generative AI to summarise documents, work by the Bank of Italy illustrates the usefulness of LLMs. Yet this research also finds the value of AI-generated summaries relative to the ones written by humans may vary, depending on users' literacy. For instance, individuals with higher educational backgrounds and greater familiarity with the Economic Bulletin published by the central bank tend to prefer human-written summaries.

Forecasting exercises

A key task of central banks that can typically be supported by generative AI is **forecasting short- and medium-term economic developments, including through nowcasting techniques** to make predictions about the current dynamics of key economic indicators.[9] For example, research by the BIS shows that the *transformer* architecture that underpins LLMs can be effectively adapted to both nowcast and forecast variables such as GDP, inflation and unemployment. Relatedly, the BIS Innovation Hub's Project Neo uses AI to nowcast the Swiss economy, leveraging a wide palette of data sources, including on retail sales, railway transport and health insurance, that can be accessed more rapidly than official statistics. Alternative data can also help to predict more specific developments outside the economic domain, such as the incidence of influenza, as demonstrated by the Bank of Italy using web-based information.

In addition to allowing for more timely forecasts, LLMs can also improve their accuracy, by integrating "traditional" macroeconomic time series with unstructured information like text (IFC (2025b)). For instance, Bangko Sentral ng Pilipinas shows that indicators built from domestic news can improve the performance of inflation forecasts. This reflects in particular the value of the information content of central bank communication as conveyed by the press, as argued by Araujo et al (2024). Joint work by the Bank of Israel, University of Tokyo, Korea's KDI School and Sogang University finds that the analysis of monetary policy statements – collected in a dedicated database encompassing 51 jurisdictions from 1990 to 2024 – can help to better predict the financial cycle.

Yet, in practice, **the effectiveness of generative AI in forecasting will often depend on the specific use case and data available**. Traditional statistical approaches may sometimes be more suitable, especially when considering the costs and/or risks associated with the use of innovative techniques. Reflecting this point, the approach followed by Bank Indonesia – applying standard clustering methods

[9] Nowcasting can also be helpful in predicting recent dynamics of economic indicators, especially when these are published with significant lags.

using models such as random forests to group borrowers – has proved well suited to capture external debt dynamics, especially in comparison with new methodologies.

Statistical production

Central banks are leveraging generative AI to support the production of statistics, a task that has become increasingly important over recent decades (Dilip et al (2026)).[10] A first key use is to simplify classification tasks, for instance by identifying products in price data,[11] mapping codes between different nomenclatures[12] and updating statistical classifications (UNECE (2023a)). A second strand of applications relates to data curation and quality assurance, including the optimisation of anomaly detection workflows. Metadata editing in particular appears to benefit considerably from LLMs' capability to generate and check text correctness, highlighting generative AI's potential to enhance existing editing processes (Sirello et al (2025)). A last and growing application of generative AI relates to the structured extraction and processing of unstructured data, such as text and images. For instance, generative AI allows for the compilation of experimental statistics on environmental, social and governance (ESG) indicators, derived from corporate reports and satellite imagery (IFC (2025c)).

Another important part of central banks' statistical work relates to dissemination, which can also be effectively supported by generative AI. One noteworthy application developed by the Central Bank of Malaysia is deploying RAG systems to improve users' access to internal and external statistical databases, including those from international organisations and the Malaysian statistical office. This tool also identifies and plots the relevant data, simplifying search and exploration. Another promising area is the development of LLM-based chatbots that facilitate data queries, such as the IMF's StatSChat and StatGPT projects (Farias (2025), Tebrake et al (2026)).

Statistical data retrieval can be further automated through the use of AI agents, as shown by the Bank of Thailand and the European Central Bank. Specifically, a chain of specialised LM–based software systems addresses the user's query all the way to the final response. For instance, a sample workflow related to the question "What was the inflation rate in Germany and Thailand last year?" first transforms the query into "Inflation rate in Germany and Thailand in 2024", then decomposes it into sub-queries (eg one for each country). A third agent would then construct specific requests[13] to retrieve the data from the original databases. The workflow could further include an agent to judge the quality of the response and another one to generate the answer, which is also subject to validation (Graph 2).

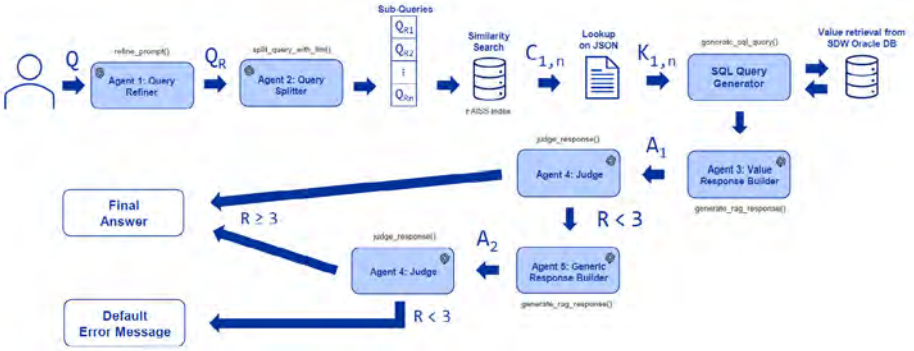
[10] For additional details on how generative AI can support statistical agencies, including central banks, across the various phases of the generic statistical business process model (GSBPM) and on the associated challenges faced in the related production workflows, see Piela (2024) and UNECE (2025a).

[11] For example, see Nunes and Palumbo (2025).

[12] For example mapping ISIC 4, ISIC 5 and NACE; see Pérez et al (2025). Another use case is mapping NACE 2008 to 2025; see Faria (2025).

[13] SQL, or Structured Query Language, is a widely used computer language for working with databases.

Using AI agents to facilitate access to the ECB Data Portal Graph 2



Source: L Petracca, S De Benedictis, T Grotton and Z Hofmeister, "Supporting users in seeking data on the ECB Data Portal: a use case for retrieval augmented generation", *IFC Bulletin*, no 67, March 2026.

Supervisory and oversight activities

Central banks are often tasked, depending on national circumstances, with supervising, monitoring and overseeing payment and financial systems, including banking and insurance activities. **These tasks often require access to vast amounts of specialised internal knowledge and/or restricted documents, where generative AI can provide valuable support**. In particular, RAGs can be very useful in dealing with sensitive information that is by its nature excluded from publicly available LLMs trained on non-confidential data.

A key area relates to the **microprudential supervision of banks**, as LLMs can be particularly helpful for analysing the activities of and risks posed by financial institutions. For example, the Federal Reserve Board follows a RAG approach to efficiently extract information from the vast amount of bank documents available (eg annual reports, financial statements). Generative AI can also be useful to keep track of the often complex and frequent changes characterising financial regulation (Brookes et al (2022)). For example, the Bank of Italy and the Sant'Anna School of Advanced Studies have built a RAG using public information related to European banking regulation on liquidity risk. The system enables easier identification of relevant legal references and provides tailored contextual guidance – including through pertinent examples – to eventually assist users in gaining more accurate and meaningful insights into complex regulatory topics. Relatedly, this approach can also be effective for the **oversight of payment systems.** As a compelling use case, the Central Bank of Chile, has developed a fine-tuned LLM – specifically trained in Spanish – to track, interpret and summarise the legal documents associated with payments.

Overall, a key lesson of these various projects is that **human involvement is often required to verify the accuracy of AI-generated responses**, reflecting the importance of specialised knowledge and accountability in supervisory processes.[14]

[14] This verification is also commonly referred to as "human-in-the-loop" (HITL).

4. Unlocking generative AI's full potential in central banking: the way forward

Central banks' experience with generative AI thus far highlights that making the most of the opportunities offered by this new technology calls for: (i) developing adequate governance and securing sufficient resources; (ii) strengthening information standards and (meta)data curation; and (iii) sharing best practices especially through international collaboration.

Governance and resources

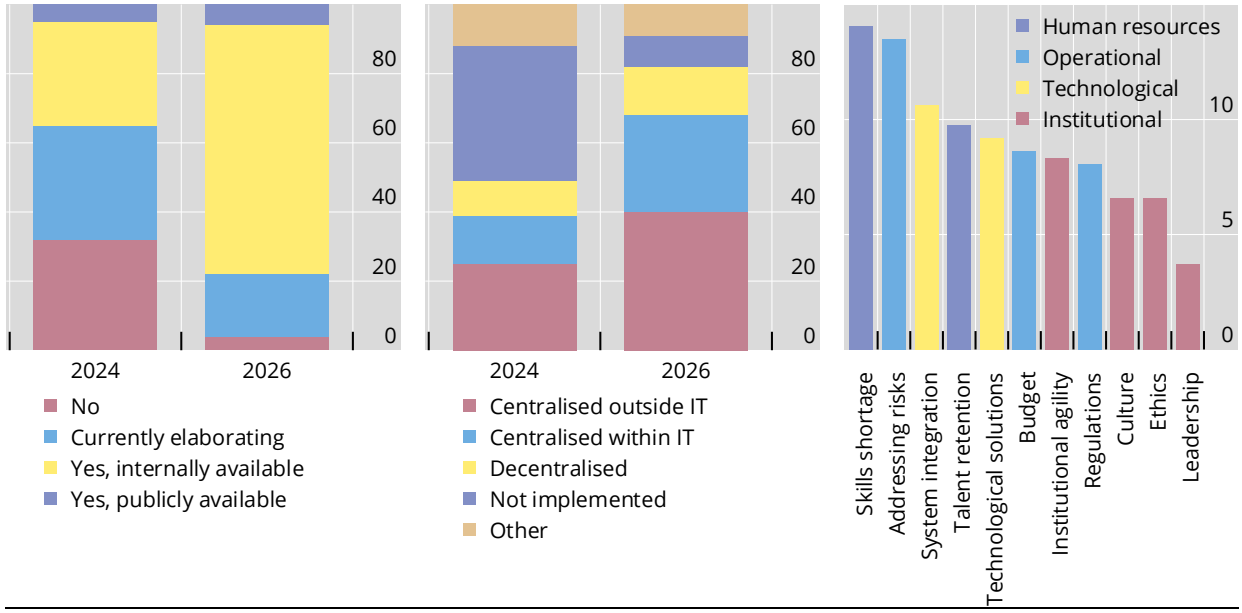
The integration of generative AI into central banks' processes requires an adequate governance framework, notably to balance innovation with AI-associated risks, as well as to ensure the availability of adequate IT, human and financial resources.

Fortunately, and reflecting the strategic importance of this new technology, **most central banks have been anticipating additional resources for AI applications as well as establishing frameworks** to adequately use and govern their use within the organisation (Graph 3; Box A; IFC (2025a)). Yet significant barriers to adoption remain,

Central banks have made notable advancements in AI governance since 2024[1]

In per cent of respondents

Graph 3



[1] Results are based on the IFC survey on AI in central banks conducted in 2024 and early 2026. Figures for 2026 are preliminary at the time of publication.

Sources: IFC; authors' calculations.

Box A

AI governance in central banks: recent developments and possible ways forward

AI governance is a broad and complex system of decision rights and accountabilities encompassing the principles, responsibilities, structures and frameworks intended to enable effective and responsible use of AI technology.

Central banks have been implementing AI governance through four key components.① The first is their established expertise in data governance and management, to ensure the quality, security and integrity of the data used by AI systems (Križman and Tissot (2022)). Second, they have been developing guidelines and policies governing the use of AI, such as terms of use as well as explainability requirements. Third, central banks have been adapting or setting up dedicated structures to enable organisation-wide development of AI. Typical arrangements include centralised units (eg steering committees), federated or hub-and-spoke models (eg AI hubs) and decentralised approaches at the level of business areas. Last but not least, risk management frameworks are another important component, including through renewed information classification rules, restrictions on AI tools and audit trails.②

Looking ahead, insights from a recent survey conducted among IFC central banks suggest three areas of focus to further develop governance, with the objective of making the most of AI opportunities while minimising the associated risks.

- *Quality of data and reference information.* Given AI's reliance on data, one question is whether currently disseminated statistical information is sufficiently machine-readable to be accurately processed, interpreted and returned by AI systems. In practice, delivering guidance on what "AI-ready data" means may call for reviewing existing (statistical) data quality frameworks, at least along three lines. First, while generative AI relies on structured and unstructured data, such as text,③ existing frameworks might not properly address the quality requirements of these novel data types – including those related to ethical considerations.④ Second, AI-readiness strategies for statistical data may also require developing frameworks to responsibly use and deploy AI, for example addressing issues related to ethical purpose, accountability, environmental sustainability and robustness.⑤ Third, gaining a better understanding of AI systems, including their limitations and potential inaccuracies stemming from training data, may require promoting statistical thinking and evaluation.⑥
- *Interoperability of data processes and systems.* Statistical standards, such as SDMX, can act as AI enablers,⑦ for instance to enhance data discoverability and interpretability – putting a premium on having curated centralised metadata that are easily accessible by machines. Additionally, these standards can improve the quality of AI systems, especially when paired with a Model Context Protocol (MCP), to ensure that the outputs generated by AI are accurate and traceable (see Box B). Yet further work might be required to enhance interoperability among data systems and ensure AI systems can reliably source high-quality data through established standards.⑧
- *Securing the adequate skills and resources for AI deployment.* Ensuring resilient AI applications in central banks demands building adequate skills at the organisational level, to raise awareness among staff about AI's opportunities and risks, eg related to information protection. It also calls for vigilance to avoid excessive and uncritical reliance on AI solutions that might eventually erode human expertise – a potential phenomenon described as "skill atrophy".⑨ Turning to IT resources, the deployment of AI typically requires advanced hardware and puts a premium on having adequate infrastructures (including the use of cloud services) to allow for more secure and scalable deployment. But such investments need to consider the associated potential risks, eg in terms of costs, third-party dependency and (data) sovereignty issues.
- *Evaluating generative AI outputs against existing methods.* Although benchmarks for evaluating natural language systems are becoming increasingly available,⑩ frameworks for assessing systematically the quality of AI-generated outputs – particularly for statistical purposes – remain limited to date. Furthermore, the opaque nature of these models (often referred to as "black boxes") may undermine the ability to verify data accuracy and integrity. This lack of transparency can result in unchecked inaccuracies, with generative AI systems producing erroneous outputs that are challenging to validate through established statistical methods.⑪

① IFC (2025a). ② CGRM (2025). ③ UNECE (2025a). ④ Floridi and Taddeo (2016) and UNECE (2026). ⑤ Haddou et al (2025) and UNECE (2025b). ⑥ RSS (2026). ⑦ IFC (2025d) and Anvar (2025). ⑧ Araujo et al (2025). ⑨ OECD (2026). ⑩ Examples include the General Language Understanding Evaluation (GLUE) benchmark and the Stanford Question Answering Dataset (SQuAD). ⑪ UNECE (2025a).

most notably in securing skilled personnel and suitable IT infrastructure. Moreover, ethical and technological risks – especially related to security, privacy and dependency on external providers – are reported to stay elevated (Graph 3.C). This puts a premium on developing suitable risk management frameworks, both at the level of the organisation – eg to avoid overreliance on a few external technological companies, including by adopting open source solutions – and at the level of the financial system – eg to prevent phenomena such as content homogenisation, potentially leading to correlated decisions and amplifying risks of procyclicality, contagion and systemic financial stress (FSB (2024), Zhang (2026)).

Turning to more operational aspects, implementing generative AI in the organisation calls for **adequate computing resources**. Specifically, high-performance hardware such as graphics processing units (GPUs) are in demand for running generative AI models, as seen in the Bank of Canada's use of AI to solve complex economic models. Yet, these IT resources – particularly when deployed on premises – may entail significant investments, relating to both their purchase and maintenance. One option, increasingly being explored by central banks, is cloud-based services which may feature lower adoption costs but also present a number of important trade-offs, especially in terms of data protection and sovereignty (IFC (2025a)).

Lastly, making the most of generative AI opportunities also demands both recruiting specialised staff that are proficient in data science techniques and **advancing general AI literacy across the organisation**. One solution has been to foster interdisciplinary approaches, including by attracting adequate staff profiles combining IT, statistical and data expertise (Araujo et al (2023), OECD (2026)). Another, perhaps more practical way is to favour more staff diversity within units, allowing for the combination of dedicated AI skills with other, more subject-oriented competencies. One interesting experiment from this perspective is the Innovation Laboratory (iLab) set up by the Bank of Latvia, with **cross-functional AI working groups** mixing various profiles across the organisation.

The need for adequate documentation and data integrity

As data are the bedrock of generative AI, their rigorous documentation (ie data about the data, or "metadata") and their integrity (ie the assurance that the information is kept accurate, authentic, consistent, reliable and traceable along its entire life cycle, including through robust information standards) are **key to ensure the adequate interpretation and reliability of the outputs generated by the new techniques**.

A way to improve the accuracy and sustainability of generative AI applications is to **focus on the quality of their input data. This will help to address the so-called "garbage in, garbage out" phenomenon**, where AI outputs are polluted by low-quality inputs (Bogani et al (2022)). Hence, an important action point is to ensure the provision of rich and standardised documentation about the information provided, such as details about the sources, compilation methodology and other quality assessments (Sirello et al (2025)).[15] From this perspective, the experience of the

[15] Other noteworthy quality aspects are replicability of statistical processes and the reproducibility of outputs, for example through so-called reproducible analytical pipelines (RAPs). These are even more

European Central Bank has underlined the relevance of building a common metadata architecture in the organisation as a way to integrate and standardise documentation across various data sets. Another option is the *Model Context Protocol* (MCP), which is emerging as an open solution for connecting AI models to external systems, including specialised tools or databases. A relevant use case, increasingly explored by statistical authorities, is to develop such MCP-based solutions to retrieve data from well-documented dissemination portals, with important benefits in terms of data traceability, accuracy and interpretability (see Box B).

Box B

Improving generative AI's accuracy with MCP: benefits and challenges

The Model Context Protocol (MCP) is a technical solution enabling LLMs to interact with external tools, such as software and data sets.① This approach brings various advantages. First, it can improve LLMs' accuracy by allowing models to interact with customised resources, such as specific databases or in-house software tools. This can be a key benefit for central banks, whose specialised workflows often rely on proprietary data sets and tools. It also helps increase the accuracy and reliability of AI models by reducing the risks of misuse and inaccurate inputs. Second, MCP works as a universal and standardised protocol (like a "USB-C port") for AI resources, eliminating the need for ad hoc solutions, which can be difficult to set up and maintain. Third, MCP can access resources both over the internet and locally. This is particularly important for central banks seeking to secure the confidentiality of their internal information and limit access rights.

Yet MCP solutions also feature some important challenges. First, their adoption can be complex, making them more suited for sophisticated business cases. Additionally, there is a risk of introducing specific cyber vulnerabilities, such as unintentionally revealing restricted data. Third, this approach is still relatively new, suggesting that major changes and compatibility issues could arise. Fourth, while MCP has experienced fast-paced adoption,② its long-term viability as a "universal" standard remains to be seen and will ultimately depend on whether it will be supported by a large share of AI providers.

Despite these challenges, experience suggests that MCP can benefit central banks in performing various tasks. For one, MCP can tap into central banks' internal information without compromising its overall security or leading to (risky) third-party dependencies. Perhaps more fundamentally, MCP can be effective in ensuring that AI applications source authoritative and trustworthy data, for instance through the use of the SDMX standard.③ This can contribute to having more accurate AI-generated outputs and also ensure that the data produced by central banks are accurately sourced by AI, reducing the risk of misuse and misinformation.

① MCP was introduced by the company Anthropic in November 2024; see Emilia (2024). ② For example, a number of AI players such as OpenAI, Microsoft and Google have been adopting MCP; see openai.com. Available MCP servers can be found in so-called directories; one popular directory is pulsemcp.com/servers. ③ Case in point, MCP solutions have become available for a number of national statistical offices, such as the National Statistics Office under the Ministry of Statistics and Programme Implementation (MoSPI) in India, to enable AI assistance to query their data through application programming interfaces (APIs); see pib.gov.in.

Another, complementary objective is to advance, promote and implement **interoperable information standards**. One practical way is to leverage well-established initiatives **to make data "*AI-ready*"**, that is, to ensure their accurate consumption by LLMs (Dossé (2025)).[16] A number of specialised standards, semantic models and common vocabularies, including the Data Documentation Initiative (DDI),

compelling given the probabilistic nature of AI/ML modelling techniques and the related need to quantify uncertainty, particularly to address issues such as model overconfidence (Dumpert (2025)).

[16] See the recommendations on making data "findable, accessible, interoperable, reusable and AI-ready" (or "FAIR^2"; Hill (2025)).

Dublin Core Metadata Initiative (DCMI) and the Data Catalog Vocabulary (DCAT),[17] already support both machine readability and actionability.[18] Another key standard actively used in the central bank community is the Statistical Data and Metadata eXchange (SDMX), which provides a rich information model that standardises both data and metadata through harmonised and unambiguous statistical concepts. This can be particularly effective to facilitate LLMs accessing accurate documentation, as shown by the BIS. One important feature from this perspective is that the SDMX standard can deal with various types of data, ranging from "traditional" macro time series to micro-level records and geospatial information (IFC (2025d)). In addition, beyond its usefulness as a tool for data and metadata management, SDMX can effectively support both data validation (for example the experience of the Bank of Italy) and data search (for instance through the sdmx.io initiative led by the BIS).[19]

A final focus point is to **implement active and continuous curation processes for creating, organising and improving metadata**. This is essential to enhance the authenticity, reproducibility and traceability of information, in turn ensuring adequate and transparent use by generative AI. First, the use of unique identifiers – including persistent ones such as the Digital Object Identifier (DOI) – can play an important role in identifying, locating and naming data both over time and across data sets. A related focus area is the development of attribution requirements and their related conventions, such as standardised (open) licences, disclaimers and terms of use (UNSC (2025, 2026)). Furthermore, innovative techniques such as blockchain technology can be leveraged to ensure data integrity and authenticity, particularly when handling information from official sources (Lambe et al (2025)). Finally, generative engine optimisation (GEO) tools have become of increasing interest for ensuring that official data sources are clearly and adequately referenced in AI-generated outputs provided by third-party providers.

Sharing and collaboration

Central banks' experience shows that **running collaborative projects can be a very cost-effective way to deploy or fine-tune generative AI techniques**. Yet this also requires adequate cooperation mechanisms, to allow for the needed collection and exchange of data while also securing their protection (IFC (2025d)).

One strand of options relates to leveraging existing *privacy-enhancing and privacy-preserving technologies* (OECD (2025a), Kim et al (2025)). These can be particularly useful in ensuring that AI models – which often need to be trained on vast information sets – make use of all the available data sources, while appropriately complying with privacy, copyright and data protection laws (OECD (2025b)). For

[17] DDI is an international standard for describing surveys, questionnaires, statistical data files and social sciences study-level information. It encompasses standards that support the entire life cycle of data (DDI-Lifecycle) and cross-domain integration (DDI-CDI). The DCMI supports innovation in metadata design, best practices and standards for resource description. Finally, DCAT – developed by Eurostat – enables semantic interoperability, for example by supporting standardised data set descriptions.

[18] For example, the World Bank has publicly released a metadata editor supporting the documentation of a wide range of structured data types, mapping them to established metadata standards and schemas (World Bank (2025)).

[19] The sdmx.io website contains an ecosystem of resources and tools – including those geared towards data scientists.

instance, pseudonymisation methods can be combined with the provision of fine-grained access rights depending on the user profile and purpose, as shown by Banco de Portugal. Another possibility is using data encryption, a technique which can be resource-intensive but offers greater security. A specific method is homomorphic encryption combined with multiparty secure private computing, which allows computations to be performed on restricted data in a decentralised way and without directly exchanging them (Ricciato (2024)). For instance, a project developed at the Bank of Italy allows for implementing secure multiparty linear regressions on data owned by different institutions, each with different sets of variables for the same individuals. Finally, LLMs can assist in improving statistical disclosure control by analysing sensitive textual content and automatically determining whether the content should be flagged as confidential (Rigaud (2025)).

Another important avenue increasingly adopted by central banks is the generation of synthetic data (IFC (2025e)). This technique replaces confidential information with "seemingly true" data that preserve the statistical properties of the original set. In this way, it allows users to perform analyses without accessing sensitive raw data (UNECE (2023b)). For instance, the Bank of Canada has used a synthetic version of its payment system data to train generative AI agents and analyse intraday liquidity patterns. However, not all synthetic data can adequately reproduce the characteristics of the original data set, as shown by the experience of the Central Bank of Türkiye and Istanbul Technical University. Consistent with this observation, the Central Bank of Malaysia has developed an evaluation framework to assess the quality of synthetic data. This framework measures statistical distance, structural similarity, performance in machine learning tasks and the risk of disclosing sensitive information to produce a composite weighted score for evaluating synthetic data.

Finally, collaboration between central banks and other stakeholders involved in the data ecosystem can play a decisive role in advancing generative AI applications. For example, the BIS Innovation Hub Ellipse Data and Knowledge Platform Community (EDKP) provides a platform for central banks and financial regulators to share source code, pool expertise and jointly develop AI-based supervisory technology solutions supporting financial supervision. Collaboration can also involve actors beyond the central banking/supervisory community, such as statistical offices, international organisations, academia and the private sector, for example by sharing knowledge, software and co-investment opportunities (IFC (2025a)).

Glossary

Artificial intelligence (AI): computer systems capable of tasks that normally require human intelligence (FSB (2017)), for instance by inferring how to generate outputs from the inputs they receive (OECD (2024)). It includes applications such as machine learning (ML) and generative AI.

AI agents: systems designed to act autonomously, making decisions and taking actions based on their programming and objectives. These systems are capable of performing tasks without constant human intervention.

AI-ready data: data that are suitable for consumption by AI systems and models, such as LLMs. Key dimensions for readiness include quality, interoperability, accessibility and compliance (both ethical and legal).

Generative AI: a subset of AI that produces new content, such as text, images, audio or video, based on patterns learned from existing data (or "training data"), usually in response to a user's prompt.

Knowledge base: a repository of information that provides domain-specific or general data to AI models. Knowledge bases can include various data types, including textual documents, which can be used to provide organisation-specific context through techniques such as RAG.

Large language models (LLMs): AI models trained on vast amounts of text data with the goal of generating human-like language. These models are designed to perform a wide range of natural language processing tasks, including summarising, translating, answering questions and interacting with users.

Model Context Protocol (MCP): an open framework and standard that enables the integration of AI systems with selected external software and data sources.

Privacy-enhancing and -preserving technologies: technologies to minimise risks of privacy infringement, such as anonymisation, encryption, differential privacy, homomorphic encryption and secure multiparty computation.

Retrieval-augmented generation system (RAG): a technique aimed at improving the quality of generative AI responses by integrating relevant context from pre-defined knowledge sources to each user prompt, offering reliable and well-defined sources for the generation of responses.

Small language models (SLMs): AI models similar to LLMs, but with significantly fewer parameters that are trained on less data than LLMs. While their smaller size reduces their capabilities compared with LLMs, SLMs are less computationally intensive and can often serve more specialised roles in specific domains.

References

Anthropic (2024): "Building effective agents", *Engineering at Anthropic,* 19 December.

Anvar, E (2025): "Why SDMX matters? A community journey towards SDMX as an AI-enabler and a data mesh enabler", *IFC Bulletin*, no 64, May.

Aprigliano, V, S Emiliozzi, C Guaitoli, A Luciani, J Marcucci and L Monteforte (2023): "The power of text-based indicators in forecasting Italian economic activity", *International Journal of Forecasting*, vol 39, no 2.

Araci, D (2019): "FinBERT: financial sentiment analysis with pre-trained language models", *arXiv*, no 1908.10063.

Araujo, D, N Bokan, F A Comazzi and M Lenza (2024): "Word2Prices: embedding central bank communication for inflation prediction", *CEPR Discussion Papers*, no 19784, December.

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2023): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October.

Araujo, D, A Cap, I Mattei, R Schmidt, O Sirello and B Tissot (2025): "Data science in central banking: unlocking the potential of data", *IFC Bulletin*, no 64, May.

Bank for International Settlements (BIS) (2025): *The use of artificial intelligence for policy purposes*, report submitted to the G20 Finance Ministers and Central Bank Governors, October.

Bogani, R, A Theodorou, L Arnaboldi and R Wortham (2022): "Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements", *AI and Ethics*, vol 3, October.

Brookes, J, M Everitt and Q-A Vo (2022): "The Basel regulatory framework: evolution of its textual complexity", *Bank Underground*, 24 August.

Bullock, M (2025): "Technology and the future of central banking at the RBA", speech at the 60th Shann Memorial Lecture, Perth, 3 September.

Consultative Group on Risk Management (CGRM) of the BIS Representative Office for the Americas (2025): *Governance of AI adoption in central banks*, January.

Dahlhaus, T, R Ellwanger, G Galassi and P-Y Yanni (2025): "From online job postings to economic insights: a machine learning approach to structuring naturally occurring data", *American Economic Association Papers and Proceedings*, vol 115, May.

DeepSeek-AI et al (2025): "DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning", *arXiv*, no 2501.12948.

Dilip, A, Z Mdingi, O Sirello, B Tissot (2026): "The evolving role of central bank statistics", *IFC Bulletin*, no 66, February.

Dossé, J (2025): "Is SDMX data AI-ready? Is AI SDMX-ready?", presentation at the SDMX Global Conference, Rome, September.

Dumpert, F (ed) (2025): *Foundations and advances of machine learning in official statistics*, Springer.

Emilia, D (2024): "Anthropic releases Model Context Protocol to standardize AI-data integration", *VentureBeat*, 26 November.

Evdokimova, T, P Nagy Mohácsi, O Ponomarenko and E Ribakova (2023): "Central banks and policy communication: how emerging markets have outperformed the Fed and ECB", *Peterson Institute for International Economics Working Papers*, no 23-10, October.

Faria, T (2025): "NACE Rév. 2.1 update: retraining an ML model in production using LLM", UNECE generative AI and official statistics workshop, Geneva, 12–14 May.

Farias, A (2025): "Talk2Manuals: your genAI-powered guide to statistical standards", UNECE generative AI and official statistics workshop, Geneva, 12–14 May.

Financial Stability Board (FSB) (2017): *Artificial intelligence and machine learning in financial services – market developments and financial stability implications*, November.

——— (2024): *The financial stability implications of artificial intelligence*, November.

Floridi, L and M Taddeo (2016): "What is data ethics?", *Philosophical transactions of the Royal Society*, vol 374, no 2083, December.

Haddou, M, T Seimandi and A Del Monaco (2025): "Fairness in machine learning: companion paper to the framework on responsible AI for official statistics", *High-Level Group for the Modernisation for Official Statistics report*, October.

Hill, S (2025): "FAIR^2 for official statistics: preparing data for responsible and interoperable AI use", UNECE generative AI and official statistics workshop, Geneva, 12–14 May.

Irving Fisher Committee on Central Bank Statistics (IFC) (2021): "Issues in data governance", *IFC Bulletin*, no 54, July.

——— (2022): "Machine learning in central banking", *IFC Bulletin*, no 57, November.

——— (2023): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October.

——— (2025a): "Governance and implementation of artificial intelligence in central banks", *IFC Report*, no 18, April.

——— (2025b): "Data integration: opportunities and challenges in central banks", *IFC Guidance Note*, no 6, July.

——— (2025c): "New insights from financial statements", *IFC Bulletin*, no 65, August.

——— (2025d): "SDMX adoption and use of open source tools", *IFC Report*, no 17, February.

——— (2025e): "Data science in central banking: enhancing the access to and sharing of data", *IFC Bulletin,* no 64, May.

Kim, S, R Jansen, M Jug, D Buckley, R de Fondeville, A C Fadel, A Saxena, J Stahl and W Hsiao (2025): "A Delphi study on the role of privacy enhancing technologies (PETs) in data sharing ecosystems", *Statistical Journal of the IAOS*, vol 41, no 3, July.

Križman, I and B Tissot (2022): "Data governance frameworks for official statistics and the integration of alternative sources", *Statistical Journal of the IAOS*, vol 38, no 3, August.

Lambe, E, G Tice, R Schmidt, C Schmieder and M Rusev (2025): "Blockchain and SDMX: immutable data sharing for trust", presentation at the SDMX Global Conference, September.

Leetaru, K and P A Schrodt (2013): "Gdelt: global data on events, location and tone, 1979-2012", *ISA annual convention*, vol 2, no 4.

Lewis, P, E Perez, A Piktus, F Petroni, V Karpukhin, N Goyal, H Küttler, M Lewis, W Yih, T Rocktäschel, S Riedel and D Kiela (2020): "Retrieval-augmented generation for knowledge-intensive NLP tasks", *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol 33.

Marcucci, J (2024): "Macroeconomic forecasting with text-based data" in M Clements and A B Galvão (eds): *Handbook of research methods and applications in macroeconomic forecasting*, chapter 16, Edward Elgar Publishing.

Nunes, I and L Palumbo (2025): "Classification at scale", UNECE generative AI and official statistics workshop, Geneva,12–14 May.

Organisation for Economic Co-operation and Development (OECD) (2024): "Explanatory memorandum on the updated OECD definition of an AI system", *OECD Artificial Intelligence Papers*, no 8, March.

——— (2025a): "Sharing trustworthy AI models with privacy-enhancing technologies", *OECD Artificial Intelligence Papers*, no 38, June.

——— (2025b): "Mapping relevant data collection mechanisms for AI training", *OECD Artificial Intelligence Papers*, no 48, October.

——— (2026): *OECD digital education outlook 2026: exploring effective uses of generative AI in education*, January.

Pérez, A, C Sáez, S Gallego and A Jurado (2025): "Synthetic LLM-generated text to train small models for automatic coding", UNECE generative AI and official statistics workshop, Geneva, 12–14 May.

Piela, R (2024): "Incorporating AI into statistical standards: enhancing GSBPM with (generative) AI", UNECE ModernStats World Workshop, Geneva, 21–22 October.

Ricciato, F (2024): "Steps toward a shared infrastructure for multi-party secure private computing in official statistics", *Journal of Official Statistics*, vol 40, no 1, March.

Rigaud, T (2025): "Enhancing statistical disclosure control using large language models", presentation at the UNECE Expert Meeting on Statistical Data Confidentiality, Barcelona, 15–17 October.

Royal Statistical Society (RSS) (2026): *AI is statistics: why statistical thinking is vital for the effective, ethical and safe use of AI*, March.

Sirello, O, B Bogdanova and E Erdem (2025): "Metadata in the age of AI: the role of official statistics in securing a virtuous cycle", *Statistical Journal of the IAOS*, vol 41, no 3, July.

Tebrake, J, E B Boukherouaa, J Danforth and M N Harikrishnan (2026): "StatGPT: AI for official statistics", *IMF Departmental Papers*, vol 2026, no 004, March.

United Nations Economic Commission for Europe (UNECE) (2023a): *Large language models for official statistics*, High-Level Group for the Modernisation for Official Statistics white paper, December.

——— (2023b): *Synthetic data for official statistics – a starter guide*, January.

——— (2025a): *Generative artificial intelligence for official statistics*, High-Level Group for the Modernisation for Official Statistics report, September.

——— (2025b): *Framework on responsible AI for official statistics*, High-Level Group for the Modernisation for Official Statistics report, October.

——— (2026): *Ethics in official statistics: foundations and practices to foster public trust*, January.

United Nations Statistical Commission (UNSC) (2025): *Fostering AI-readiness and responsible redistribution of official statistics*, room document by the World Bank, Fifty-sixth session, New York, 4–7 March.

——— (2026): *Joint report of the Committee of the Chief Statisticians of the United Nations System and the Committee for the Coordination of Statistical Activities*, items for information: coordination of statistical programmes, Fifty-seventh session, New York, 3–6 March.

World Bank (2025): *Metadata editor: user and practice guide*, Office of the Chief Statistician, May.

Zhang, T (2026): "The financial stability implications of artificial intelligence and digital finance", speech at International Financial Week – Asian Financial Forum, Hong Kong, 26 January.

Zhang, H, P S Yu and J Zhang (2025): "A systematic survey of text summarisation: from statistical methods to large language models", *ACM computing surveys*, vol 57, no 11, June.

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Opening remarks[1]

Alberto Naudon,
IFC Chair and Central Bank of Chile

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Opening Remarks for the 4th IFC Workshop on Data Science in Central Banking

Alberto Naudon, Chair of the BIS Irving Fisher Committee on Central Bank Statistics (IFC)

18 February 2025, Rome, Italy

Good morning, distinguished guests, colleagues, and friends,

It is my great pleasure to welcome you all to the **4th Workshop on Data Science in Central Banking** organized by the BIS Irving Fisher Committee on Central Bank Statistics (IFC) and hosted by the Bank of Italy.

As we gather today, we are reminded of the **rapid advancements in data science and its profound impact on central banking**. Indeed, the sheer volume and complexity of financial data now available call for more sophisticated techniques for data management and analysis. This trend is reinforced by the new opportunities opened up by artificial intelligence and machine learning. This workshop is a testimony to our collective commitment to harnessing innovation to enhance central bank' operations, policy-making, and overall effectiveness.

As emphasized in the last 2024 IFCs Annual Report just endorsed by the BIS All Governors a few weeks ago, **the current focus on data science and AI supports the broader objective of improving statistical methods and fostering innovation in central banks**. This IFC report underscores that leveraging new technologies can be instrumental to enhance data quality, improve analytical capabilities, and support evidence-based policymaking. The Report also calls for reviewing the related ongoing initiatives pursued by central banks and for providing a platform for sharing knowledge and best practices.

Let me recall that the **three previous IFC data science workshops** have been dealing with, respectively, (1) machine learning applications; (2) applications and tools in data science; and (3) data access and sharing. This time we will over the next three days delve into the various aspects related to the **use of generative AI in**

central bank activities. We will hear from esteemed experts and practitioners who will share their insights and experiences, providing us with valuable knowledge and practical tools to navigate the evolving landscape of data science.

I would like first to extend a special welcome to **our keynote speaker, Julien Simon, Chief Evangelist at Arcee.ai**, who will be discussing the tailoring of small language models for enterprise use cases. His expertise and vision will undoubtedly set the tone for our discussions.

Then the sessions of the **workshop will cover various critical areas**, such as natural language processing tools, AI for summarization and information extraction, supervisory technology, text analysis for market monitoring and monetary policy purposes, and data privacy and anonymization.

* *

*

Let me share with you a few thoughts on these issues:

First, the new techniques we will discuss are not only very timely, but they are also essential to leverage data science to address the complex challenges we face in modern central banking. In particular, **the integration of generative AI and advanced data analytics into central banks' operations can significantly enhance their ability** to make informed decisions, assess economic trends, and work to promote monetary and financial stability. More generally, IT innovation provides brand new perspectives. For instance, open-source software offer numerous benefits supporting official statistics and data analysis, including cost savings, flexibility, and the ability to customize solutions to meet specific needs. Another example is that modern data management approaches such as data lakes and data meshes architectures allow for new ways to store, organize, and access data [*Note that I said data lakes not "data ponds", and data mesh not "data mess"!!*]. This calls for careful planning and for not blindly following the crowd and fashionable buzz words. The main goal is to concretely help central banks to more effectively leverage their information assets, improve the integration and quality of their data, and support more sophisticated analytical techniques.

Second, your presence here today, coming from various jurisdictions all over the world and representing central banks, other public authorities, international organizations, academia and the private sector, **underlines the importance of the goal of this workshop, which is to showcase concrete projects, share**

experiences, develop in-house knowledge and also reduce reliance on external service providers.

Third, central banks, as producers of official data, have a key role to play to promote the access and dissemination of credible information to various external stakeholders, including other domestic authorities, international institutions, academia, and the general public. But **better data is also key for supporting real-time, evidence-based policymaking in central banks**, which increasingly rely on trustworthy data and sophisticated analytical and forecasting capacities to support their decisions.

Fourth, the relevance of artificial intelligence for central banks cannot be overstated, as it offers immense opportunities to enhance productivity, improve decision-making, and foster innovation. In particular, **Generative AI has the potential to revolutionize data analysis and interpretation, offering deeper insights and more accurate predictions.** For instance, the use of large language models can significantly enhance our ability to process and understand vast amounts of unstructured data, ranging from economic reports to news articles, thereby enabling us to make more informed policy decisions especially in the areas of monetary policy, financial stability, and regulatory oversight.

However, and this is my fifth point, **GenAI also presents significant challenges and risks**. Central banks must navigate issues such as data privacy, security, and ethical considerations. The potential for systemic risks, such as homogenization of information and procyclicality, requires careful management. As central banks increasingly rely on data-driven approaches, it is essential to ensure that sensitive information is protected, and that data is used ethically and responsibly.

And my last point is that addressing these challenges calls for **developing robust governance frameworks.** This is key so that we can harness the power of AI while mitigating its risks, ensuring that our financial systems remain stable and resilient. At the same time investing in advanced IT infrastructure and fostering collaboration and coordination as we do today can help to stay abreast of emerging threats and implement best practices.

* *

*

To conclude, this workshop aims to gather a diverse audience of practitioners, specialists, and interested stakeholders from central banks, international

organizations, national statistical offices, and beyond. **Our primary objective is to highlight ongoing projects and exchange experiences that can help foster in-house expertise and lessen reliance on external service providers.** For instance, a number of projects that will be presented in the next few days have replicable codes developed with open-source software and can be usefully shared among all interested stakeholders. Moreover, the presentations will enhance our understanding of the opportunities and risks associated with new Generative AI technologies. This is key for central banks willing to navigate the evolving financial landscape and ensure that they are well-positioned to meet future challenges.

I therefore encourage you all to actively participate in the sessions, engage with the speakers, and share your own experiences and perspectives. **It is through this collaborative spirit that we can truly advance our understanding and application of data science in our field**. Before closing, I would like to thank the organizers, speakers, and all participants for your dedication and contributions to this workshop. I am confident that our time together will be both enlightening and inspiring, and I look forward to the fruitful discussions and innovative ideas that will emerge.

Thank you, and welcome once again to the 4th Workshop on Data Science in Central Banking.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Tailoring small language models for enterprise use cases[1]

Julien Simon,
Arcee.ai

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.



Tailoring small language models for enterprise use cases

Julien Simon, Chief Evangelist
julien@arcee.ai
linkedin.com/in/juliensimon
youtube.com/juliensimonfr

Hard truths, part 1

If GenAI looks deceivingly simple, then I’m afraid you are deceived

- How deep is the model's knowledge on your business domain?
- Not just your industry, but also your company, your department, your customers, etc.
- How fresh is that knowledge?
- How compliant are the answers? Structure, length, tone of voice, safety
- Does the model make it easy to trust its answers? Step by step "reasoning", quotes

- Prompt tuning is mostly a bad joke 🤷
- RAG is mandatory for fresh data, but doesn't deliver domain adaptation
- And what about privacy, security, compliance, cost, scaling, etc.



Hard truths, part 2

The « best » closed LLM is not all you need

- Closed LLMs have been built to be a mile wide and an inch deep
- They're convenient for vanilla tasks but quickly show their limits
- They don't know what they don't know and they can't be efficiently fine-tuned
  « *How well do commercial fine-tuning APIs infuse knowledge into LLMs?* » https://arxiv.org/abs/2409.05059 (11/2024)
- Users try to fix this with huge-context RAG, which doesn't work well, if at all
  « *Lost in the Middle: How Language Models Use Long Contexts* » https://arxiv.org/abs/2307.03172 (06/2023)
- You have very few options to tweak cost/business performance/technical performance
- Many business processes require not only a model, but possibly several models, and often external IT tools (documents, APIs, SQL, etc.)



Hard truths, part 3

The « best » open-source SLM is not all you need

- Yes, open-source SLMs solve most, maybe all, the closed LLM problems.
- There are still challenges to overcome (we'll talk about them in a minute)
- Last night, while you were sleeping, somebody released a new « best » SLM 😳
- Many business processes require not only a model, but possibly several models, and often external IT tools (documents, APIs, SQL, etc.)



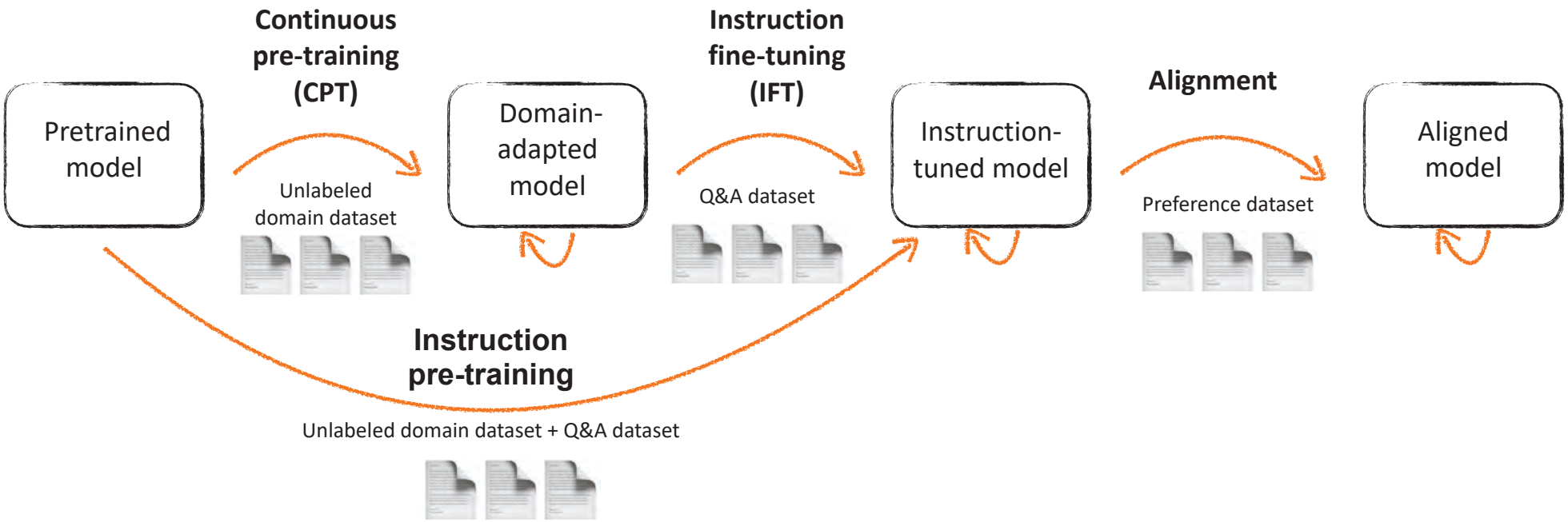
Our mission at Arcee AI

- Build SLMs that consistently outperform not only other SLMs, but also LLMs
- Build them quickly and cost-efficiently based on the « best » current architectures
- If need be, make it possible for you to tailor our SLMs
- Make it easy for you to consume models (ours and others)
  - Directly with simple APIs a la OpenAPI - Arcee Model Engine
  - In our agentic workflow platform - Arcee Orchestra
  - In our next-gen router-based inference platform - Arcee Conductor
- Give you several deployment options: SaaS, Cloud (VPC), on-prem
- Deliver industry-leading GenAI solutions, both on quality and ROI

How do we do it?



A typical SLM adaptation workflow



« Language Models are Few-Shot Learners » https://arxiv.org/abs/2005.14165 (05/2020)
« Finetuned Language Models Are Zero-Shot Learners » https://arxiv.org/abs/2109.01652 (09/2021)
« Efficient Continual Pre-training for Building Domain Specific Large Language Models » https://arxiv.org/abs/2311.08545 (11/2023)
« Instruction Pre-Training: Language Models are Supervised Multitask Learners » https://arxiv.org/abs/2406.14491v1 (06/2024)
« How Do Large Language Models Acquire Factual Knowledge During Pretraining? » https://arxiv.org/abs/2406.11813v1 (06/2024)



Challenges of SLM adaptation

- Building datasets is hard work
  - Continuous Pre-Training requires a large corpus (at least 1 billion of tokens)
  - Instruction Fine-Tuning and Alignment requires high-quality, diverse Q&A pairs

- Training / fine-tuning models: accuracy or cost, pick one
  - Working with the full model in original precision (say, BF16)
    - Compute-heavy and expensive… assuming you can even get the amount of compute you need
  - Parameter Efficient Fine Tuning (PEFT), e.g. LoRA or QLoRA
    - Much more memory-efficient, enabling smaller GPUs and shorter training times
    - Effective for Instruction Fine-Tuning and Alignment
    - Significant accuracy degradation for CPT
      - https://blog.arcee.ai/why-methods-like-qlora-fall-short-in-domain-knowledge-injection-2/
      - « *LoRA vs. Full Fine-Tuning: An Illusion of Equivalence* » https://arxiv.org/abs/2410.21228 (10/2024)

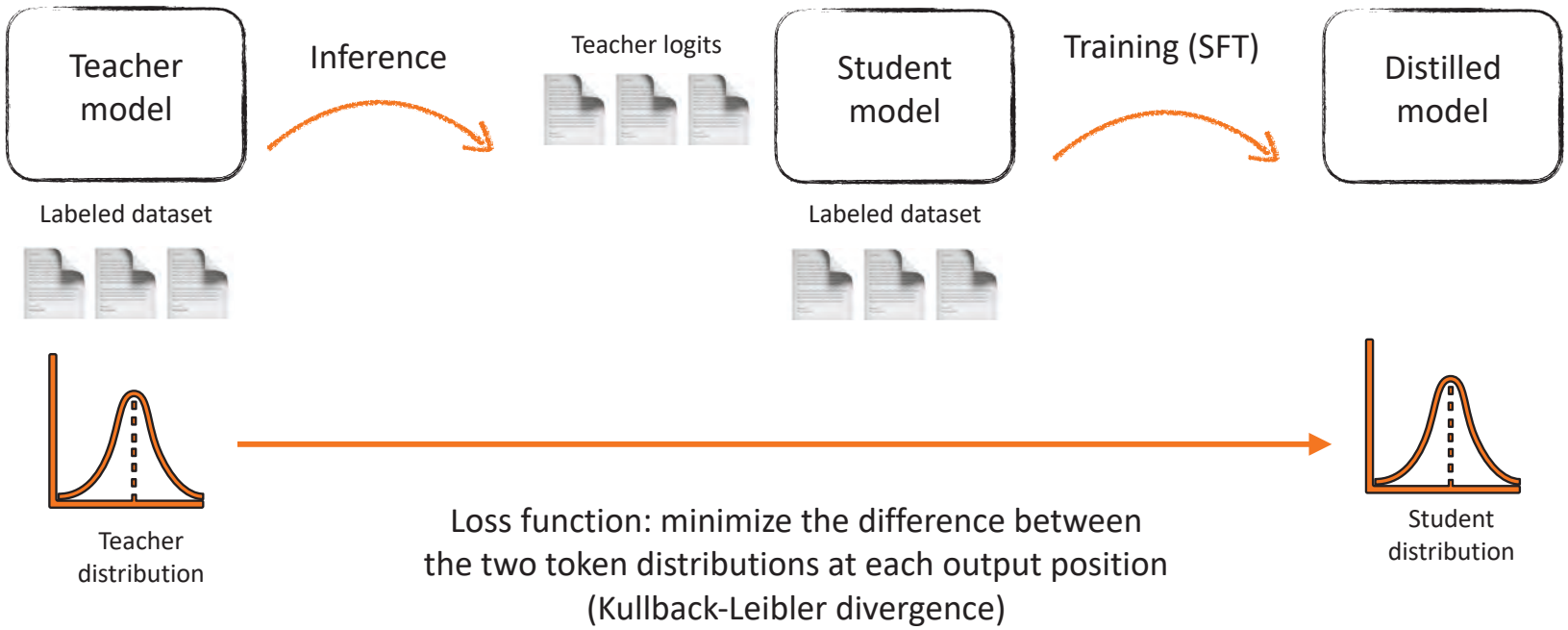
- Can we get FFT-like quality at a fraction of the cost? Yes



Model distillation: DistillKit

https://blog.arcee.ai/distillkit-v0-1-by-arcee-ai/ (08/2024) + https://github.com/arcee-ai/DistillKit

Intuition: if we have excellent large open-source models available,
can we teach smaller models to mimic their output?



Parameter-efficient training: Spectrum

https://arxiv.org/abs/2406.06623 (06/2024)+ https://github.com/cognitivecomputations/spectrum

- Intuition: not all layers contribute equally to the output.
- Some layers have a higher signal-to-noise ratio (SNR) than others.
- Spectrum identifies the high SNR layers.
- Run full fine-tuning on high SNR layers, and leaves the other layers untouched.

$$SNR = \frac{\sum_{k|\sigma_k > \varepsilon} \sigma_k}{\sum_{n|\sigma_n < \varepsilon} \sigma_n}$$

```
python spectrum.py --model-name <insert local or HF repo here>
                   --top-percent <top % of SNR ratios to target>
```

```
    "model.layers.10.self_attn.o_proj": {
        "snr": 0.25031203031539917,
        "type": "self_attn.o_proj"
    },
    "model.layers.11.self_attn.o_proj": {
        "snr": 0.2547757625579834,
        "type": "self_attn.o_proj"
    },
    "model.layers.12.self_attn.o_proj": {
        "snr": 0.2616233825683594,
        "type": "self_attn.o_proj"
    },
    "model.layers.13.self_attn.o_proj": {
        "snr": 0.2736438810825348,
        "type": "self_attn.o_proj"
    },
    . . .
```

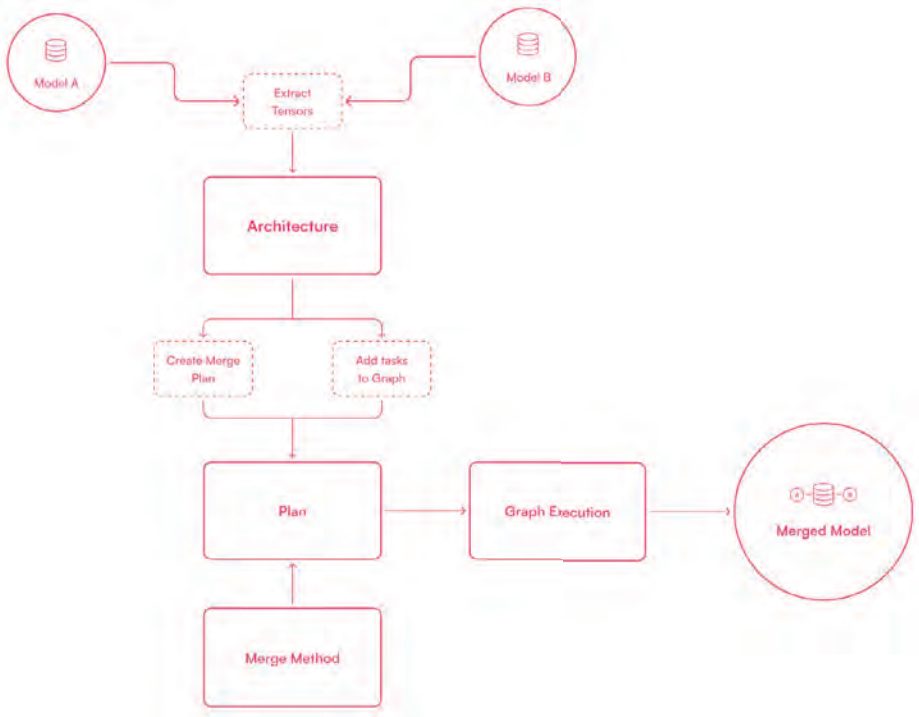
```
unfrozen_parameters:
- ^lm_head.weight$
- ^model.embed_tokens.weight$
# input_layernorm layers
- model.layers.0.input_layernorm
- model.layers.1.input_layernorm
- model.layers.2.input_layernorm
- model.layers.3.input_layernorm
- model.layers.4.input_layernorm
- model.layers.5.input_layernorm
# lm_head layers
# mlp.down_proj layers
- model.layers.21.mlp.down_proj
- model.layers.20.mlp.down_proj
- model.layers.22.mlp.down_proj
- model.layers.19.mlp.down_proj
- model.layers.23.mlp.down_proj
- model.layers.24.mlp.down_proj
. . .
```



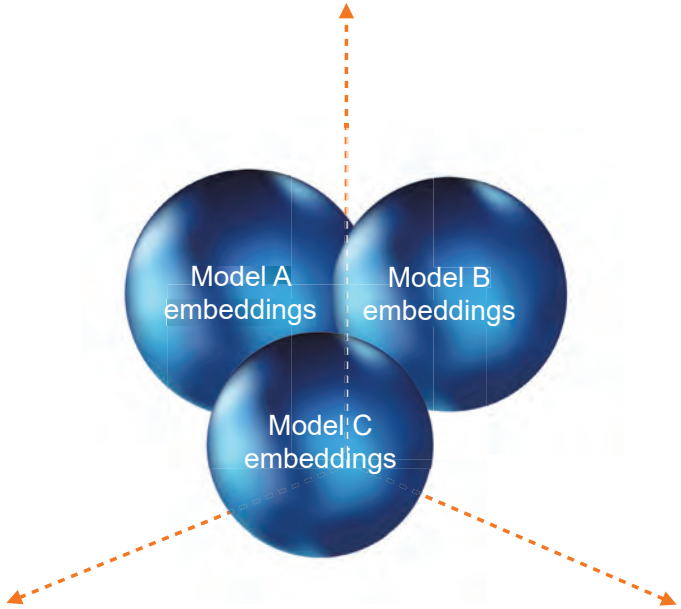
Model merging: MergeKit

https://arxiv.org/abs/2403.13257 (03/2024) + https://github.com/arcee-ai/mergekit

Intuition: if we have plenty of open-source models, can we build a model by merging several models that already have the properties we need?



- Combine several task-specific models (or different checkpoints) into a single model with zero additional training
- Not an ensembling technique: only one model
- Merging only requires lightweight compute
- Fast process, no extra inference latency



Demo: merging models with `mergekit`

```
$ git clone https://github.com/arcee-ai/mergekit.git
$ cd mergekit
$ pip install -e .
```

```yaml
models:
  - model: defog/llama-3-sqlcoder-8b
    parameters:
      density: 1.0
      weight: 0.2
  - model: MathGenie/MathCoder2-Llama-3-8B
    parameters:
      density: 1.0
      weight: 0.6
  - model: ajibawa-2023/Code-Llama-3-8B
    parameters:
      density: 1.0
      weight: 0.2
merge_method: ties
base_model: meta-llama/Llama-3.1-8B
dtype: float16
```

72.48%

65.35%

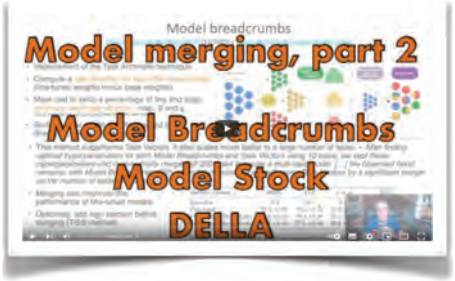
64.37%

```
(env-mergekit) → examples git:(main) ✗ mergekit-yaml ties-demo.yml ties-demo
Fetching 11 files: 100%| | 11/11 [00:00<00:00, 19886.79it/s]
Fetching 11 files: 100%| | 11/11 [00:00<00:00, 14377.48it/s]
Fetching 9 files: 100%| | 9/9 [00:00<00:00, 10960.72it/s]
Fetching 10 files: 100%| | 10/10 [00:00<00:00, 25070.56it/s]
Warmup loader cache: 100%| | 4/4 [00:00<00:00,  5.05it/s]
Executing graph: 100%| | 2039/2039 [01:05<00:00, 30.95it/s]
```

Merging three 8B models with TIES
1 minute on my Mac :)

73.62%

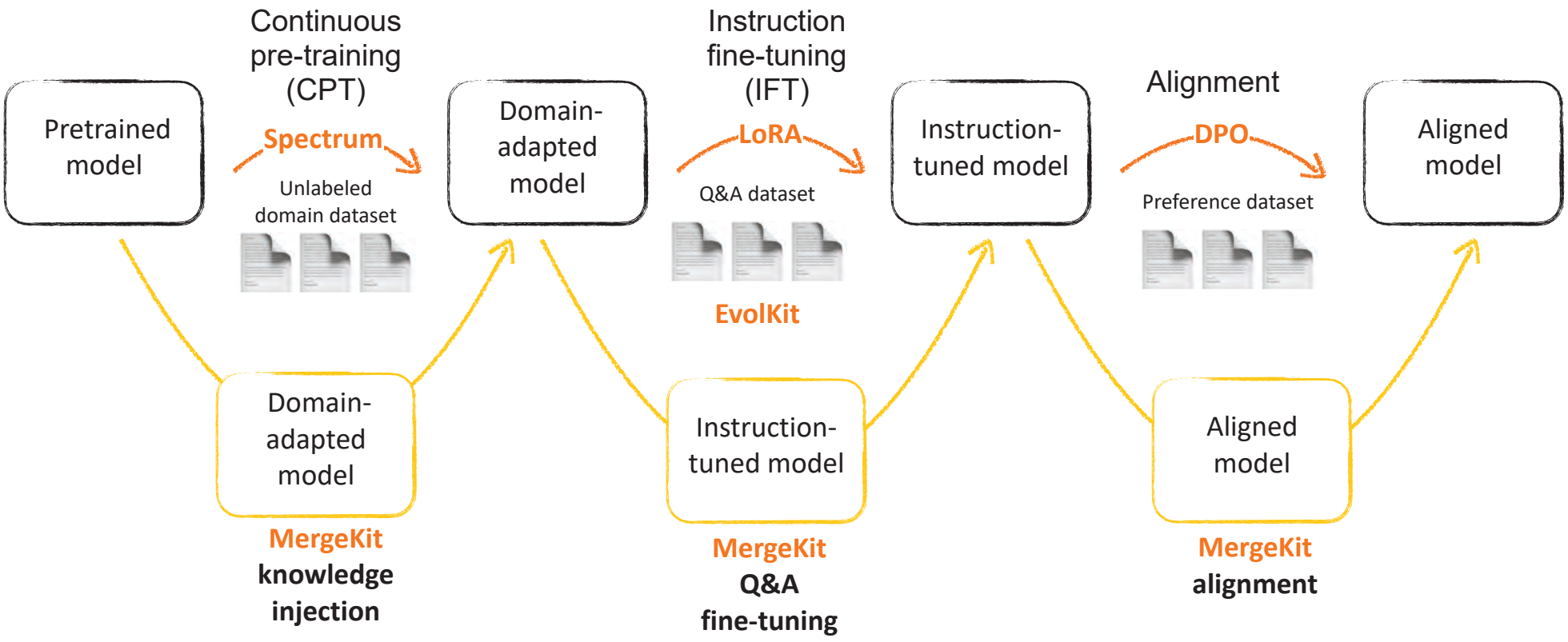




GSM8K scores - lm_eval 12/16/2024



A modern SLM adaptation workflow



« Arcee's MergeKit: A Toolkit for Merging Large Language Models » https://arxiv.org/abs/2403.13257 (03/2024)
« Spectrum: Targeted Training on Signal to Noise Ratio » https://arxiv.org/abs/2406.06623 (06/2024)
« Merging in a Bottle: Differentiable Adaptive Merging (DAM) and the Path from Averaging to Automation » https://arxiv.org/abs/2410.08371 (10/2024)

Arcee AI - The Open SLM leader

State-of-the-art tech stack based on open-source libraries

Spectrum (continuous pre-training), MergeKit (merging), DistilKit (distillation), EvolKit (dataset improvement)

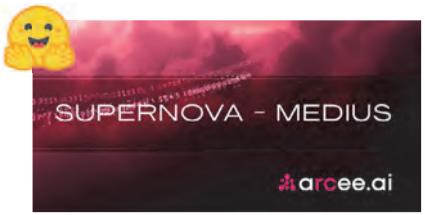
Best-in-class models based on open-source architectures



Qwen2 1.5B



Llama 3.1 8B



Qwen2.5 14B



Qwen2 72B



Llama 3.1 70B



Best 1.5B model



Best 8B model



Best 14B model



Best Arabic model

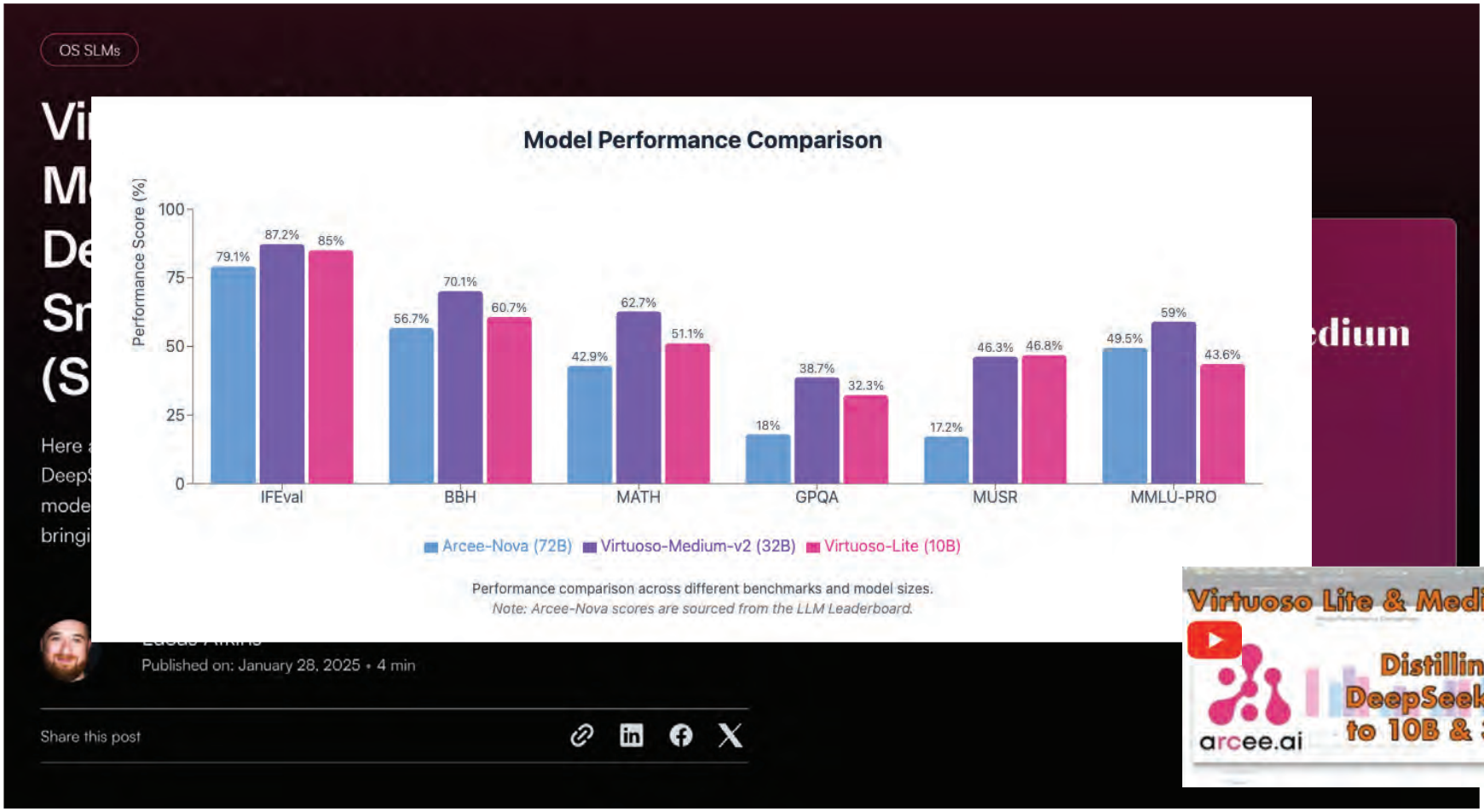


Best 70B model

Hugging Face OpenLLM Leaderboard benchmarks

Latest releases: 10B and 32B

https://www.arcee.ai/blog/virtuoso-lite-virtuoso-medium-v2-distilling-deepseek-v3-into-10b-32b-small-language-models-slms





Arcee SuperNova training pipeline

https://blog.arcee.ai/arcee-supernova-training-pipeline-and-model-composition/

Llama-3.1 405B Instruct

Labeled dataset 500M tokens

Distillation (Distillkit)

32 H100, 5 days

Llama-3.1 70B Instruct #1

Llama-3.1 70B Base

Synthetic data (EvolKit) 76K samples

Training (Spectrum) + checkpoint merging

8 H100

Llama-3.1 70B Instruct #2

Merging (Mergekit)

Llama-3.1 70B Base

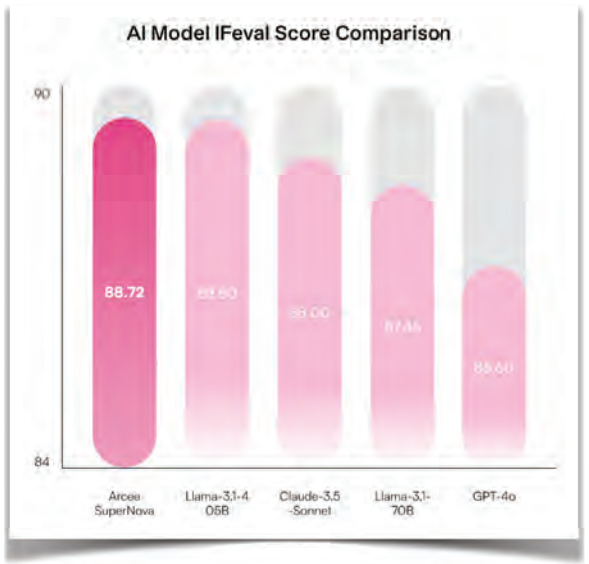
User preference datasets

Alignment (DPO)

8 H100

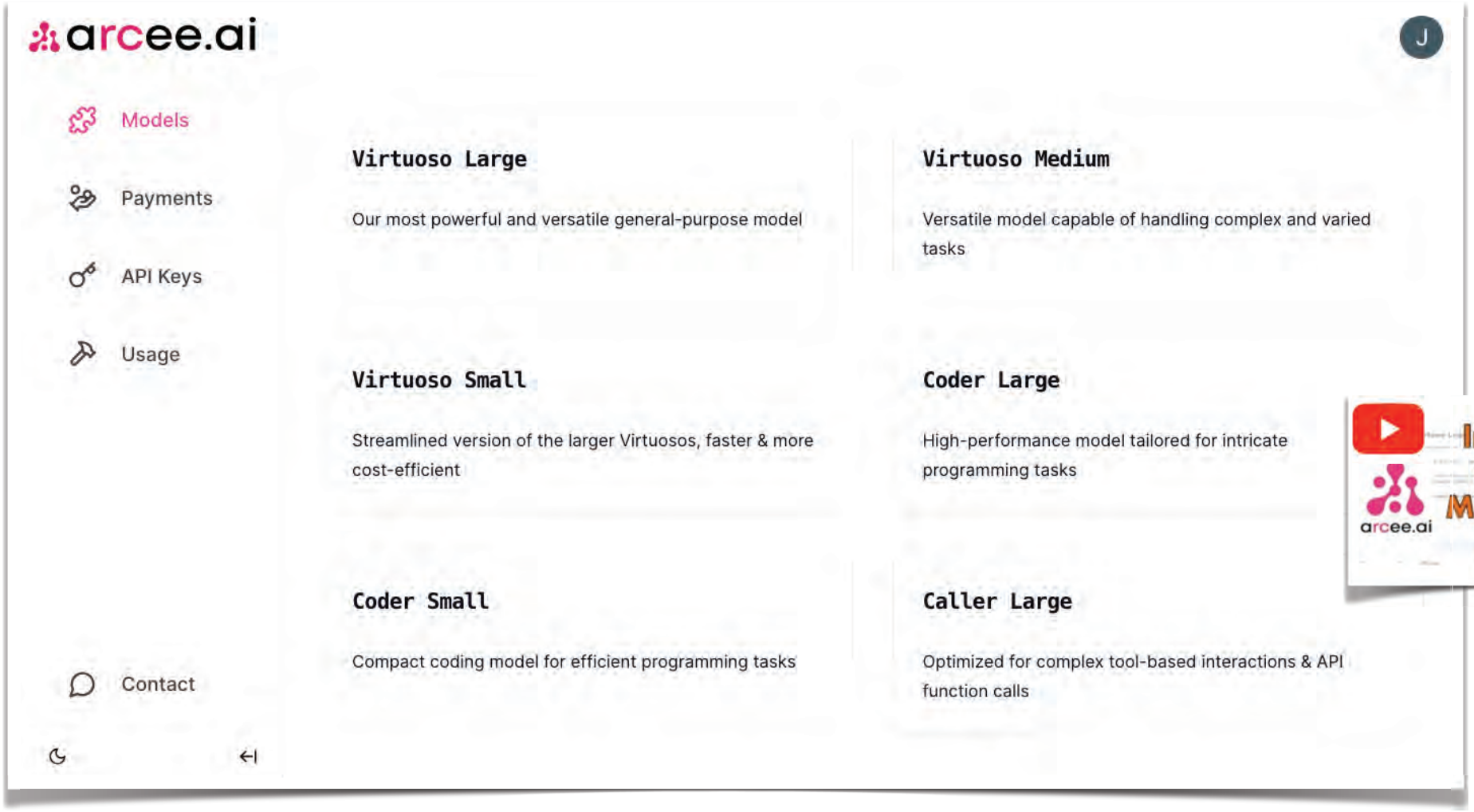
Llama-3.1 70B Instruct #2





Arcee Model Engine - SLM inference platform

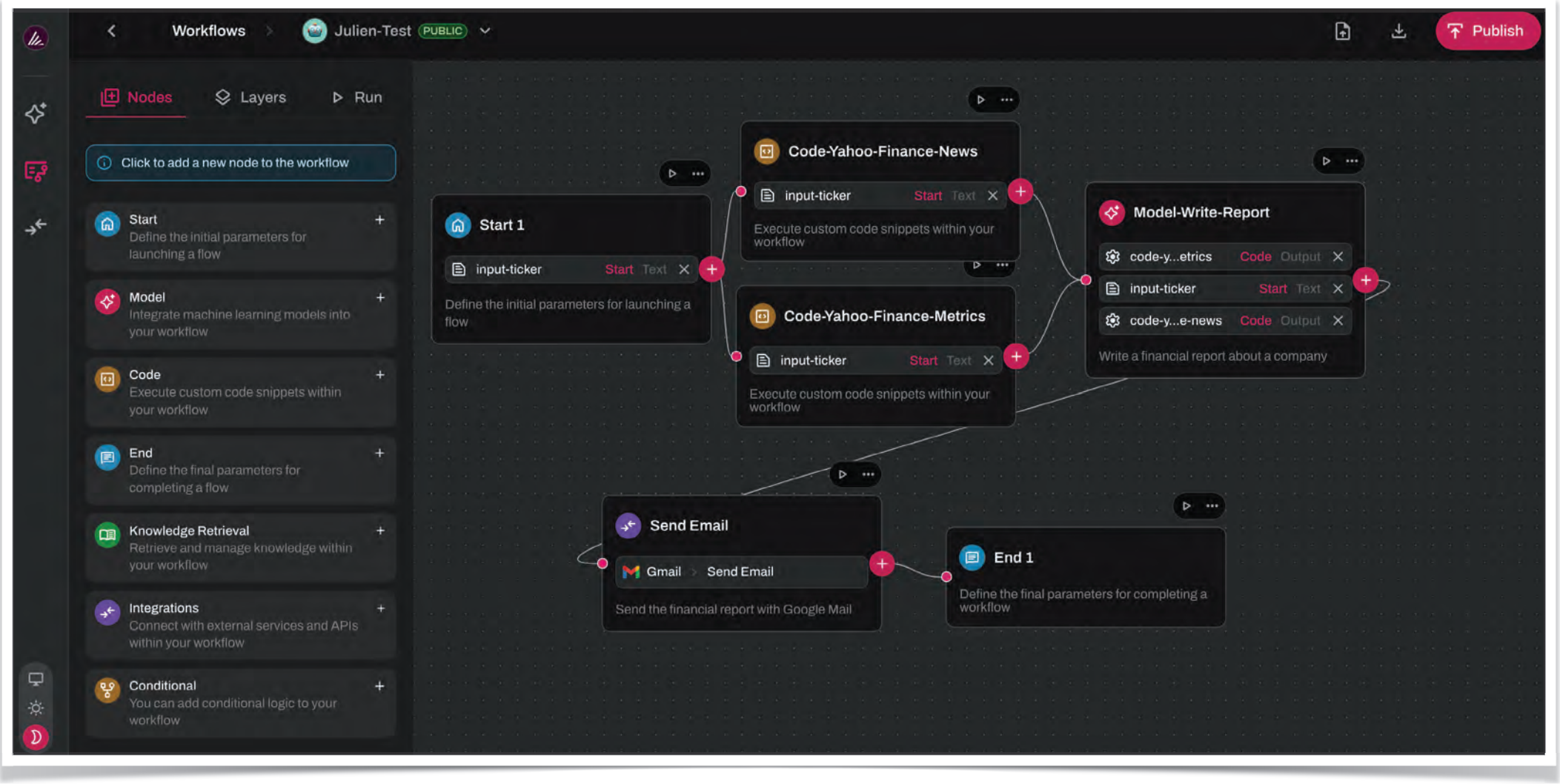
https://www.arcee.ai/product/model-engine





Arcee Orchestra - SLM agentic workflows

https://www.arcee.ai/product/orchestra



Arcee AI

Leader in Small Language Models (SLMs)

End-to-end Training and Agentic Platforms

SaaS and Private Deployments

Julien Simon, Chief Evangelist
julien@arcee.ai
linkedin.com/in/juliensimon
youtube.com/juliensimonfr

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Data science: the role of statisticians[1]

Elisabetta Carfagna,
International Statistical Institute

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

4th IFC and Bank of Italy Workshop on “Data Science in Central Banking”

Data science: the role of statisticians



Elisabetta Carfagna, University of Bologna
Chair of the Special Interest Group on Data Science
International Statistical Institute (ISI)

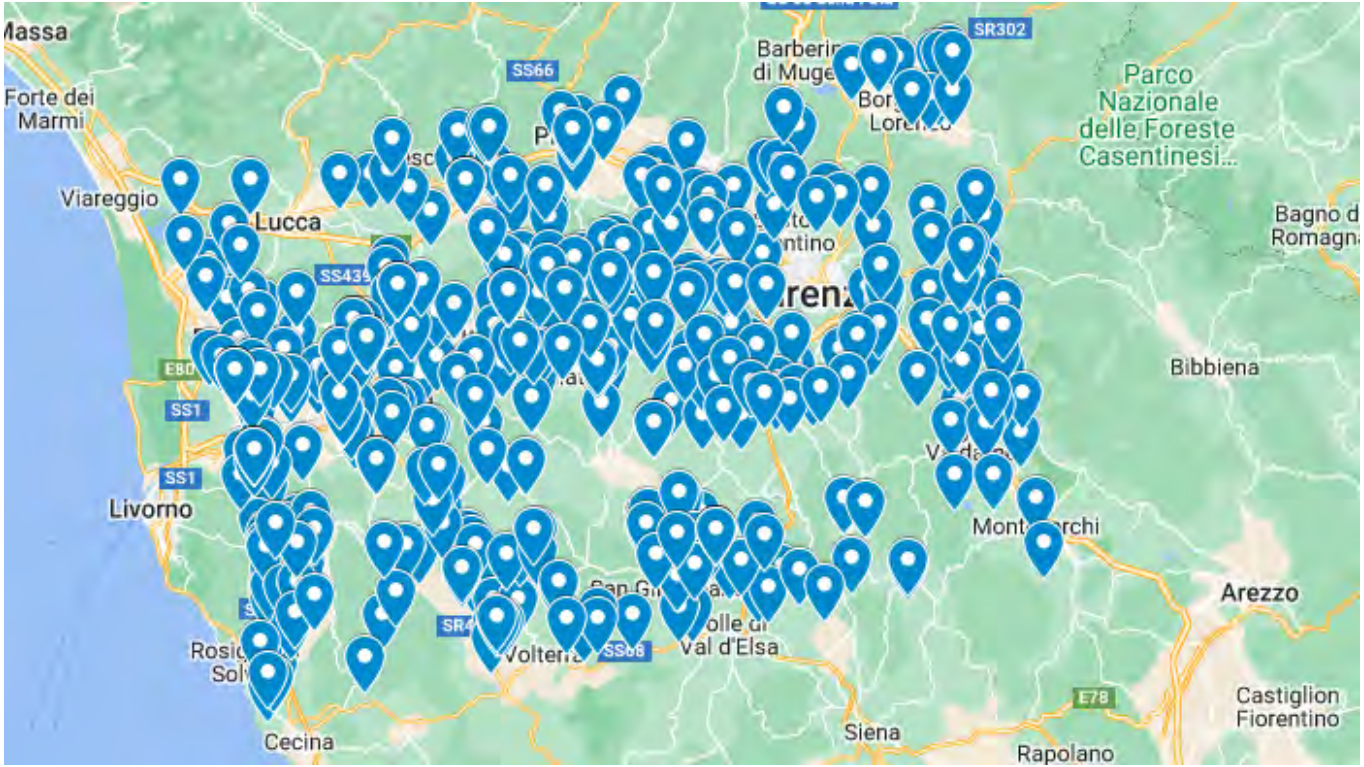
Alternative data sources and statistics

- **Non-probability data:**
  - Large datasets, new data sources, administrative registers, satellites and aircrafts, webcams, data voluntarily provided by the internet users, data harvested from the web
- **Questions**
  - How reliable are statistics produced through data science applied to big data?
  - Do machine learning classifiers outperform parametric models?
  - Which is the role of statisticians?

Data elaborated for estimating acreage of crops

- Satellite data:
  - 6 images from Sentinel 1 and Sentinel 2
  - 6 vegetation indexes for each of the Sentinel 2 images:
  - Normalized Difference Vegetation Index (NDVI)
  - Green Normalized Vegetation Index (GNDVI)
  - Two-band Enhanced Vegetation Index (EVI2)
  - Normalized Difference Water Index (NDWI)
  - Chlorophyll Red-Edge (CIRed-edge)
  - Soil-Adjusted Vegetation Index (SAVI)
- Digital elevation model

Area sampling frame - Un-clustered point sampling



Training and test data: ground data collected by Italian Ministry of Agricultural Food and Forestry Policies (MiPAAF) in 2016 (AGRIT project) 574 geo-referenced points in the north of Tuscany Region

Information collected on the ground on the sample of 574 geo-referenced points

- Land use
  - Agricultural land use
  - Cropping patterns
- Farm management
  - Soil cover
  - Tillage practices
  - Ground cover technique
  - Presence and kind of irrigation
  - Presence of fences







Esempio di 781 - SIEPI E FILARI inseriti nella matrice agricola





Esempio di 781 - SIEPI E FILARI pertinenti alla rete stradale principale: da non rilevare

Comparison of the performance of classifiers

- Regularized multinomial regression - penalized logistic classifier estimated using the maximization of the likelihood function combined with a lasso penalty term to deal with many explicative variables

$$Pr(G = k|X = x) = \frac{\exp\left(\beta_{k0} + b_k^T x\right)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + b_l^T x)}, k = 1, \dots, K-1$$

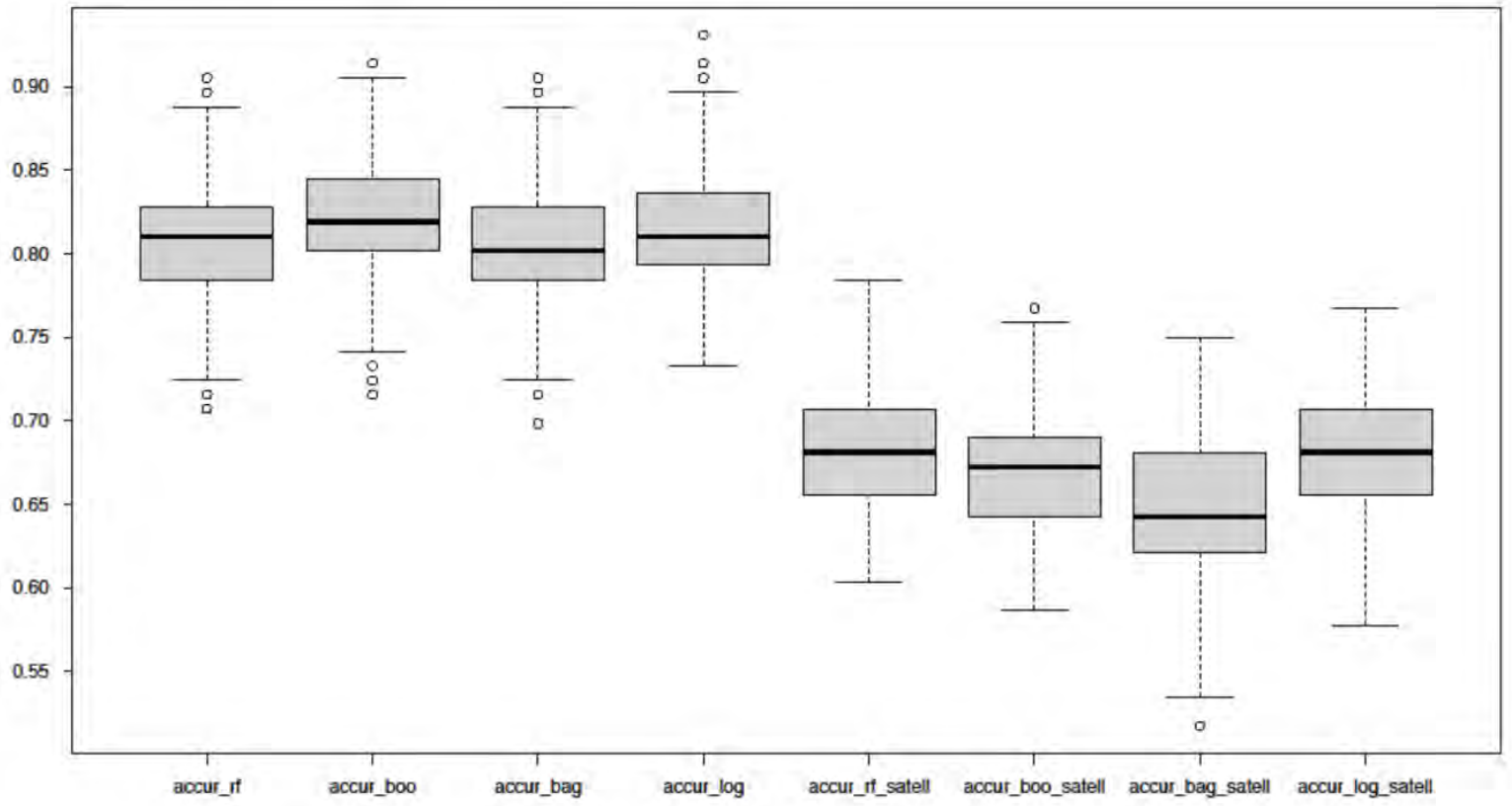
- 3 machine learning classifiers:
  - random forest
  - boosting
  - bagging

Each classifier has been repeated for different training and test sets, maintaining the same proportions of training and test sets (1000 simulations):

- 80% of the sample used for training the classifiers (459 points)
- 20% of the sample used for testing (115 points)

Distribution of accuracy of Random Forest, Boosting, Bagging and Regularized multinomial regression

- with all explicative variables
- with only explicative variables derived from satellite data

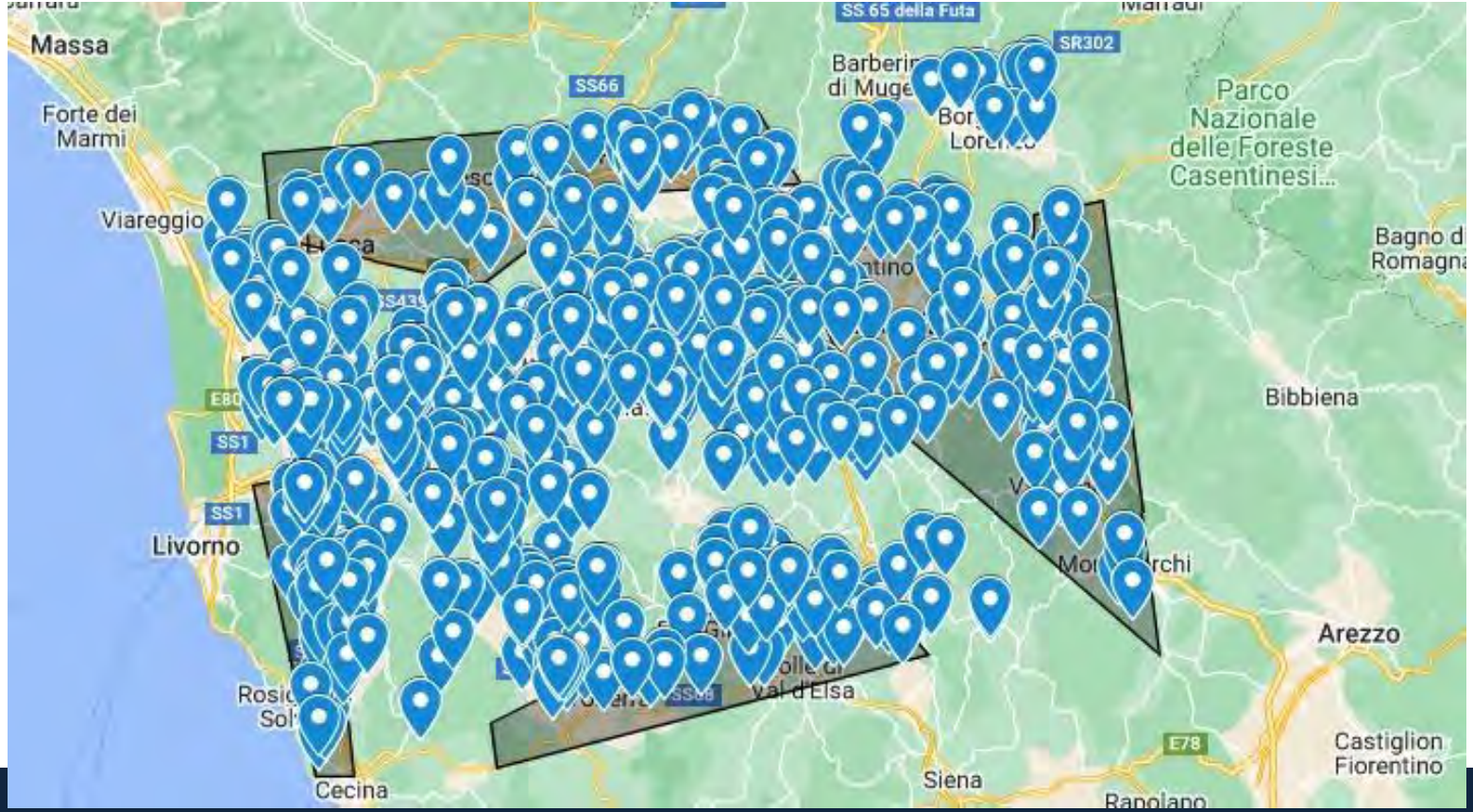


Median accuracy and Kappa value for the various classifiers

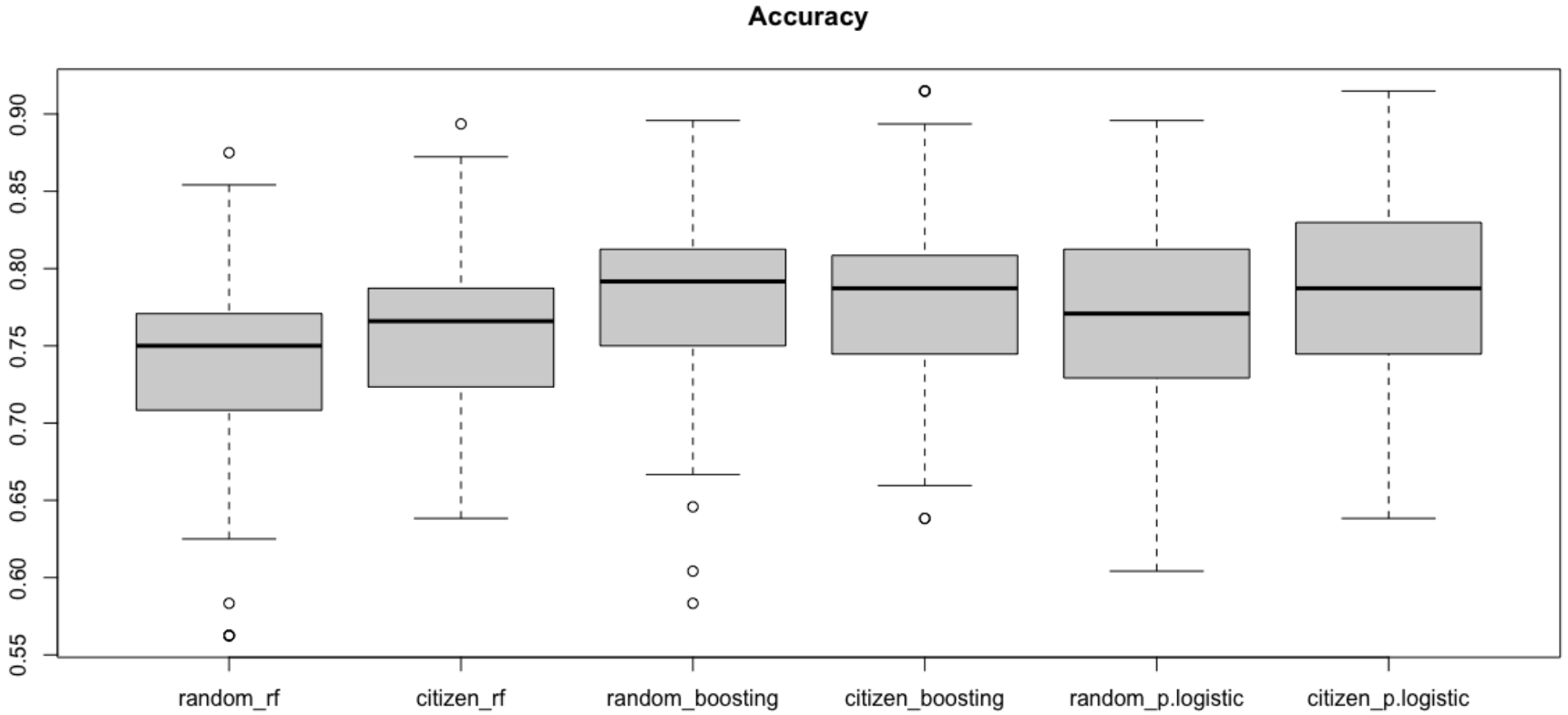
Vineyards; olive groves; sunflower; winter cereals, other crops

All explanatory variables			Satellite explanatory variables only		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.845	0.777	Random forest	0.691	0.526
Regularized multinomial regression	0.836	0.767	Regularized multinomial regression	0.689	0.521
Random forest	0.819	0.731	Boosting	0.677	0.528
Bagging	0.812	0.713	Bagging	0.652	0.495

Subset of ground data - 177 out of the 574 points close to cities and coastal areas selected to simulate a probable spatial distribution of data collected by citizens (citizen science)



Distribution of accuracy of Random Forest, Boosting, and Regularized multinomial regression with 177 sample points out of the 574: stratified random subsample and citizens subsample - 142 points (80%) for training – 35 points



Median accuracy and Kappa value for the various classifiers with 177 sample points: stratified random subsample and citizen subsample

Vineyards; olive groves; sunflower; winter cereals, other crops

Stratified random subsample			Citizen subsample		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.77	0.66	Boosting	0.77	0.63
Regularized multinomial regression	0.76	0.64	Regularized multinomial regression	0.79	0.65
Random forest	0.72	0.59	Random forest	0.74	0.55

Comparisons of area estimates based on: 1) citizen subsample 2) entire AGRIT sample

Total study area 741,450 ha

	Area estimate with citizen subsample (ha)	Area estimate with AGRIT sample (ha)	Area estimate citizen-AGRIT	Relative difference %
Other	148,654	152,364	-3,709	-2
Olive groves	52,896	47,043	5,853	12
Vineyard	50,577	37,368	13,209	35
Winter cereals	47,488	42,877	4,611	11
Sunflowers	5,285	9,070	-3,785	-42

Expansion factor based on photo-interpreted systematic sample (29,658 points) has been adopted also for estimates based on citizen subsample

Role of statisticians in data science

- **Population and variables of interest - under and over coverage**
- **Data collection methodology - training and test data**
- **Data quality - sampling and non sampling errors**
- **Quantify uncertainty - Classification accuracy - small classes**
- **Classifiers - assumptions - interpretation**
- **Models for combining proxies with real data - bias-variance tradeoffs**
- **Results analysis**

ISI welcomes IFC sessions at ISI regional conference, Malta, June 3-6, 2026

Thank you elisabetta.carfagna@unibo.it

Main references

Carfagna, E., Gallego, F.J. (2005) Using remote sensing for agricultural statistics. International Statistical Review, 73: 389-404.

Defourny, P. (2017) Land cover mapping and monitoring. In: J. Delincé (ed.), Handbook on Remote Sensing for Agricultural Statistics (Chapter 2). Handbook of the Global Strategy to improve Agricultural and Rural Statistics (GSARS): Rome.

Friedman J., Hastie T., Tibshirani R. (2001) The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA.

Gómez, C., White, J.C. & Wulder, M.A. (2016) Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. ISPRS Journal of Photo-grammetry and Remote Sensing, 116: 55–72.

Hamza M. and Larocque D. An empirical comparison of ensemble methods based on classification trees. Journal of Statistical Computation and Simulation, 75(8):629–643, 2005

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, Brian Alan Johnson, (2019) Deep learning in remote sensing applications: A meta-analysis and review, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 152, 2019, Pag-es 166-177, ISSN 0924-2716

Pratesi M. (2023) Letter from the President, The Survey Statistician, 2023, Vol. 88, 4.

Tillé Y., Debusschere, M., Luomaranta, H., Axelson, M., Elvers, E., Holmberg, A. & Valliant, R. (2022) Some Thoughts on Official Statistics and its Future (with discussion), Journal of Official Statistics,38(2) 557-598. https://doi.org/10.2478/jos-2022-0026

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

IFC survey on central banks' use of AI and ML[1]

Alberto Naudon,
IFC Chair and Central Bank of Chile

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Irving Fisher Committee on Central Bank Statistics

BIS



IFC survey on central banks' use of AI and ML

Alberto Naudon, *IFC Chair and Central Bank of Chile*

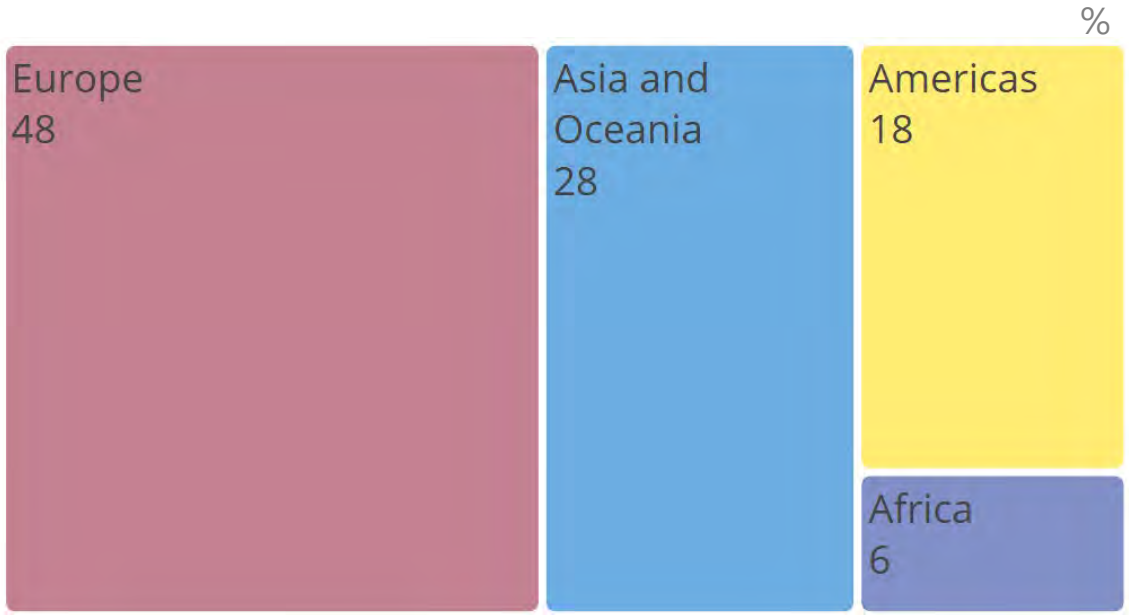
4th IFC and Bank of Italy workshop on "Data Science in Central Banking"

18 February 2025, Rome (Italy)

IFC survey on central banks' use of AI and ML

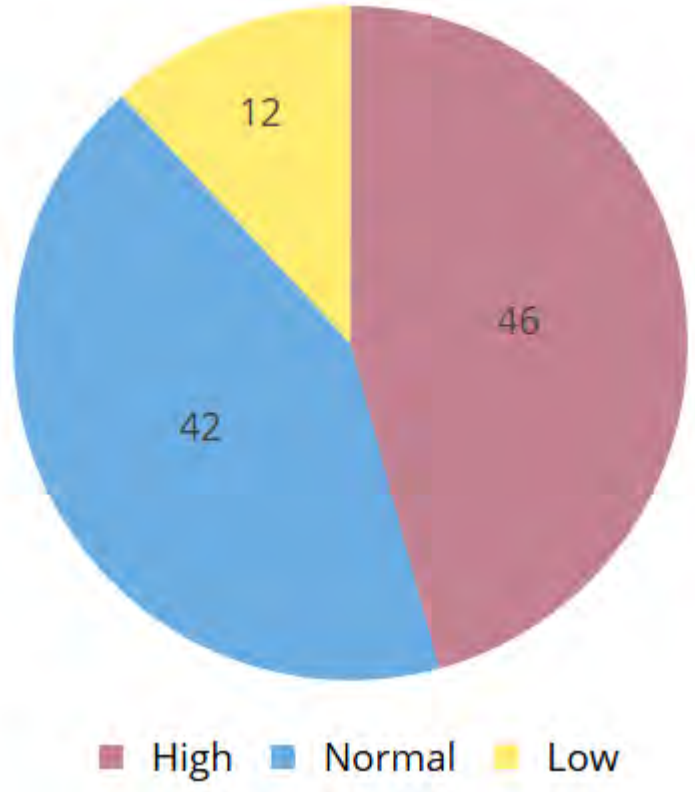
- **Period**: September – October 2024
- **Coverage**: 60 jurisdictions
- **Six chapters**:
  - Scope and interest
  - Expectations
  - Applications
  - Organisational policies, governance and risks
  - IT stack
  - Collaborative strategies

Share of respondents by geographical region



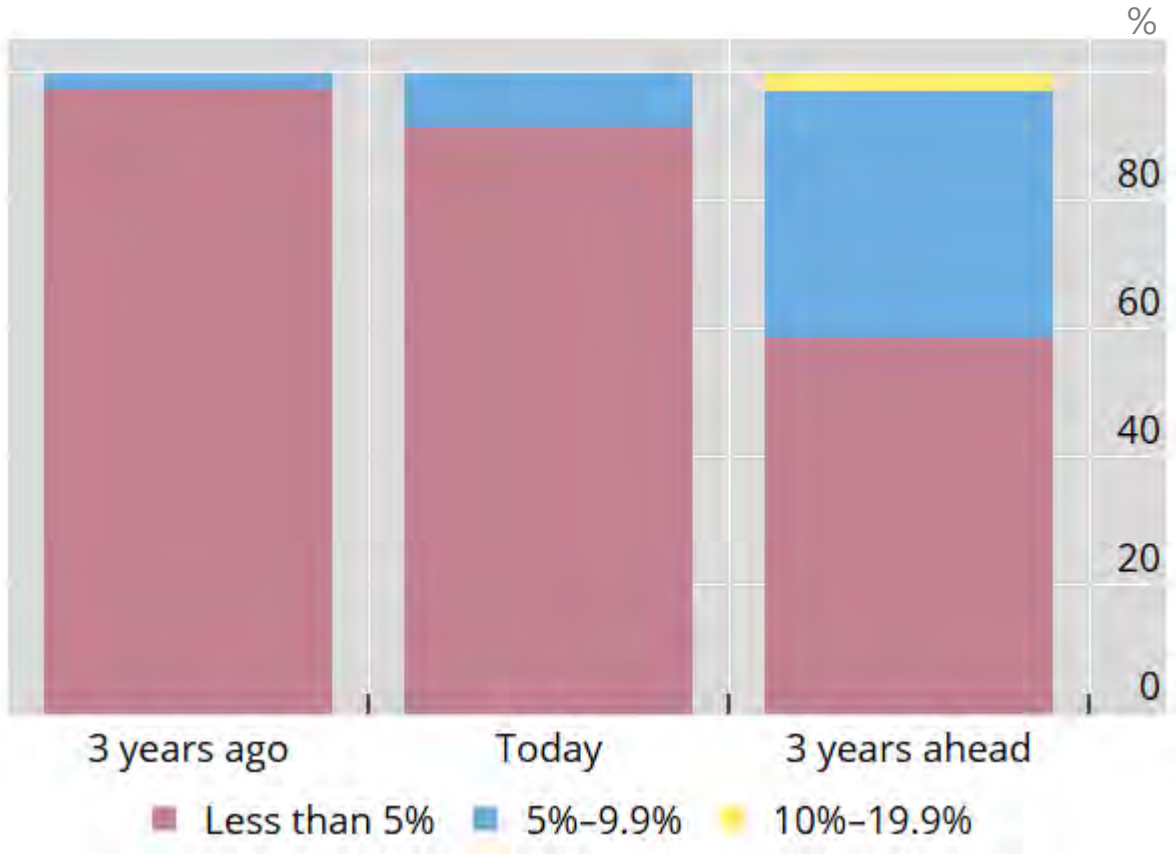
1 Artificial intelligence is a **priority** for central banks

AI/ML importance in strategic planning



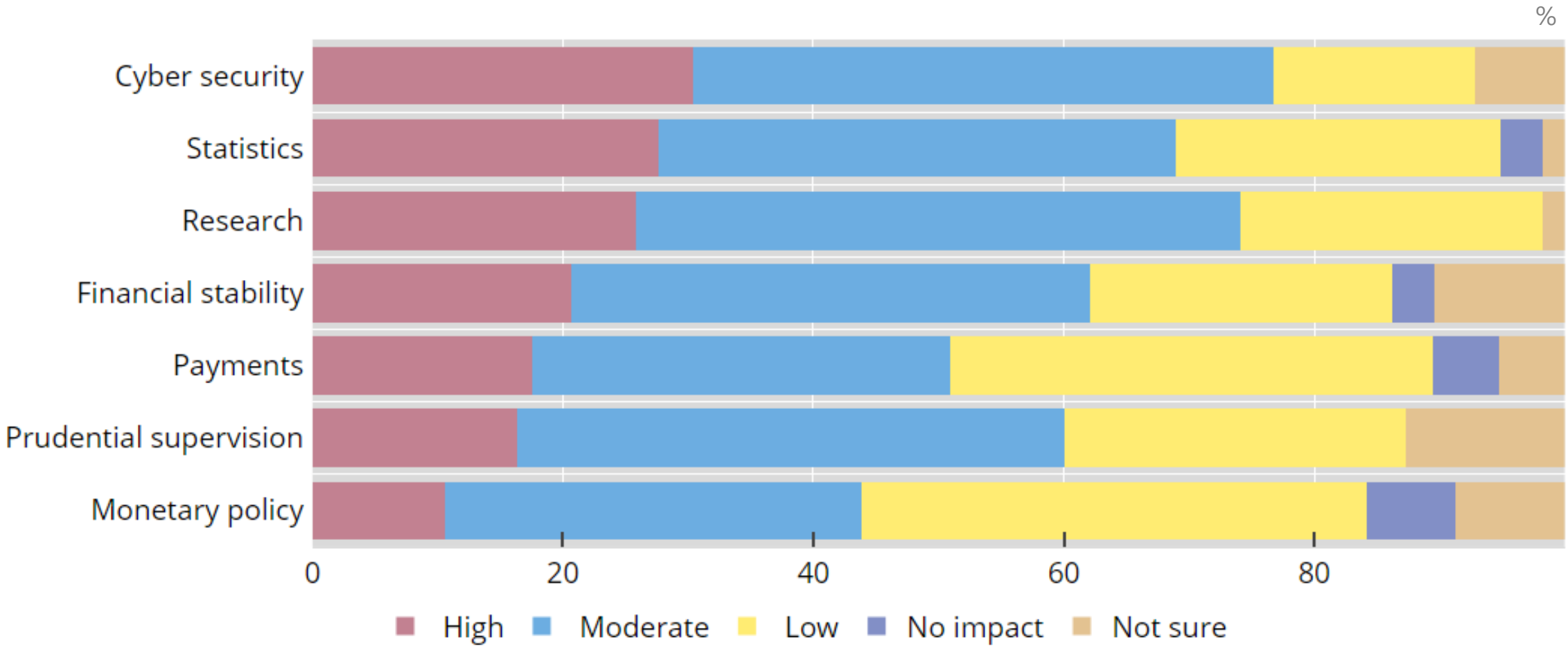
Question: On a scale from 1 (low priority) to 3 (high priority), how important is AI/ML to your institution's strategic goals over the next two years?

Budget allocated to AI/ML projects



Question: What percentage of the total budget of your institution is intended to finance AI/ML projects?

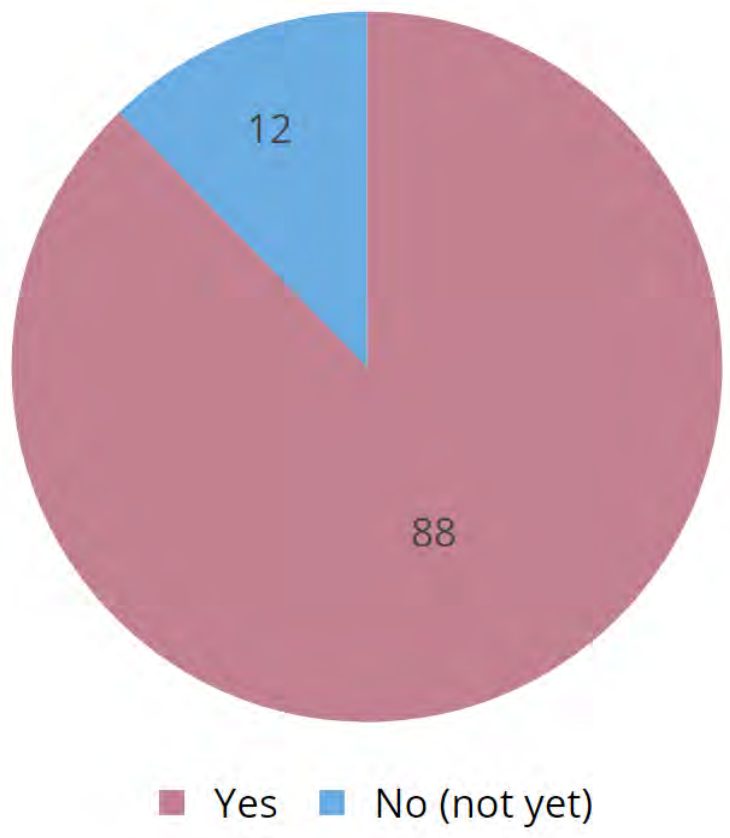
1 AI is expected to have a **significant impact** on **cyber security**, **statistics** and **research**



Question: Please rate the impact you expect from AI/ML in the following functional domains in the next two years.

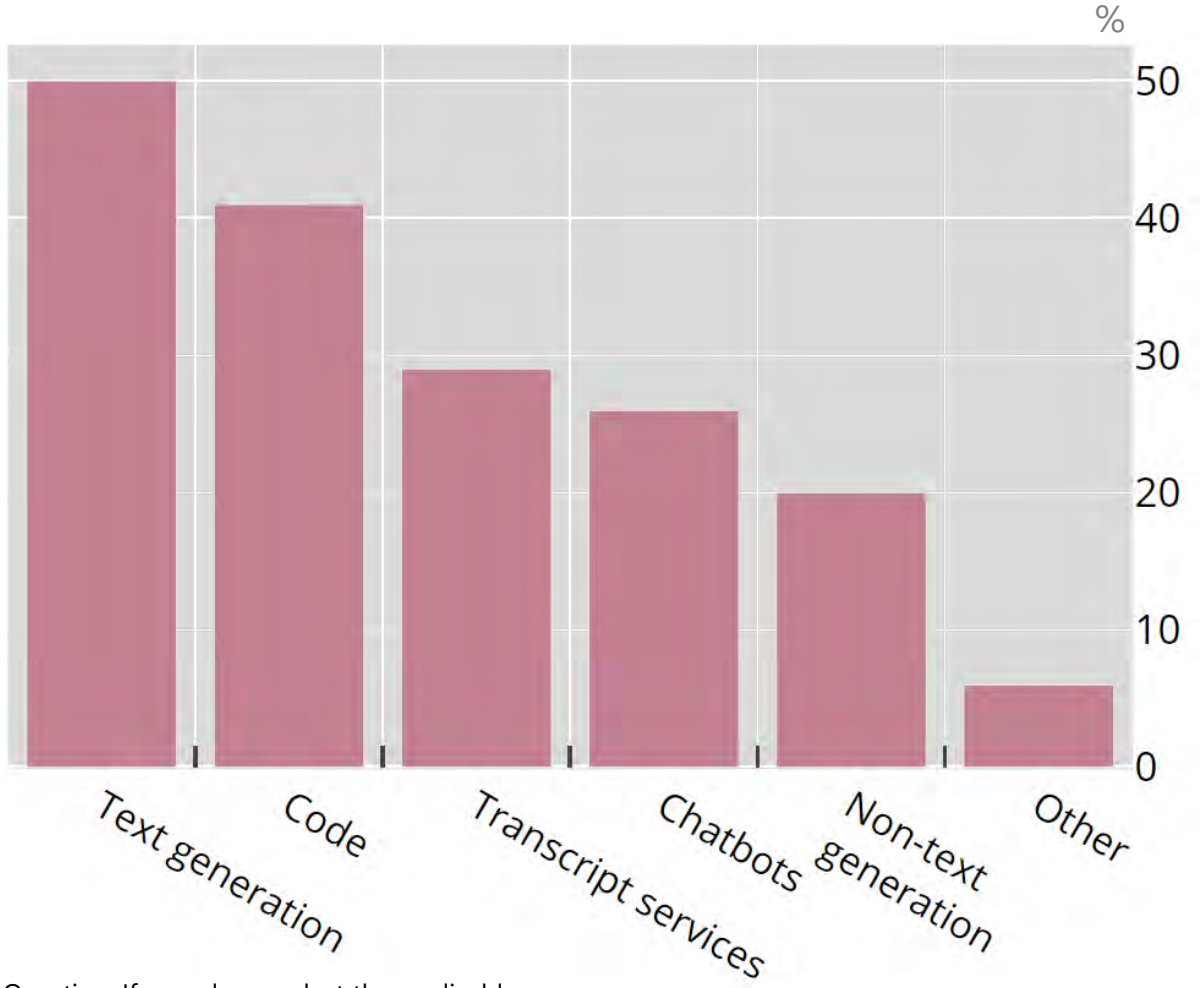
1 Central banks widely use **generative AI**

Share of central banks using generative AI



Question: Are you currently using generative AI in your institution?

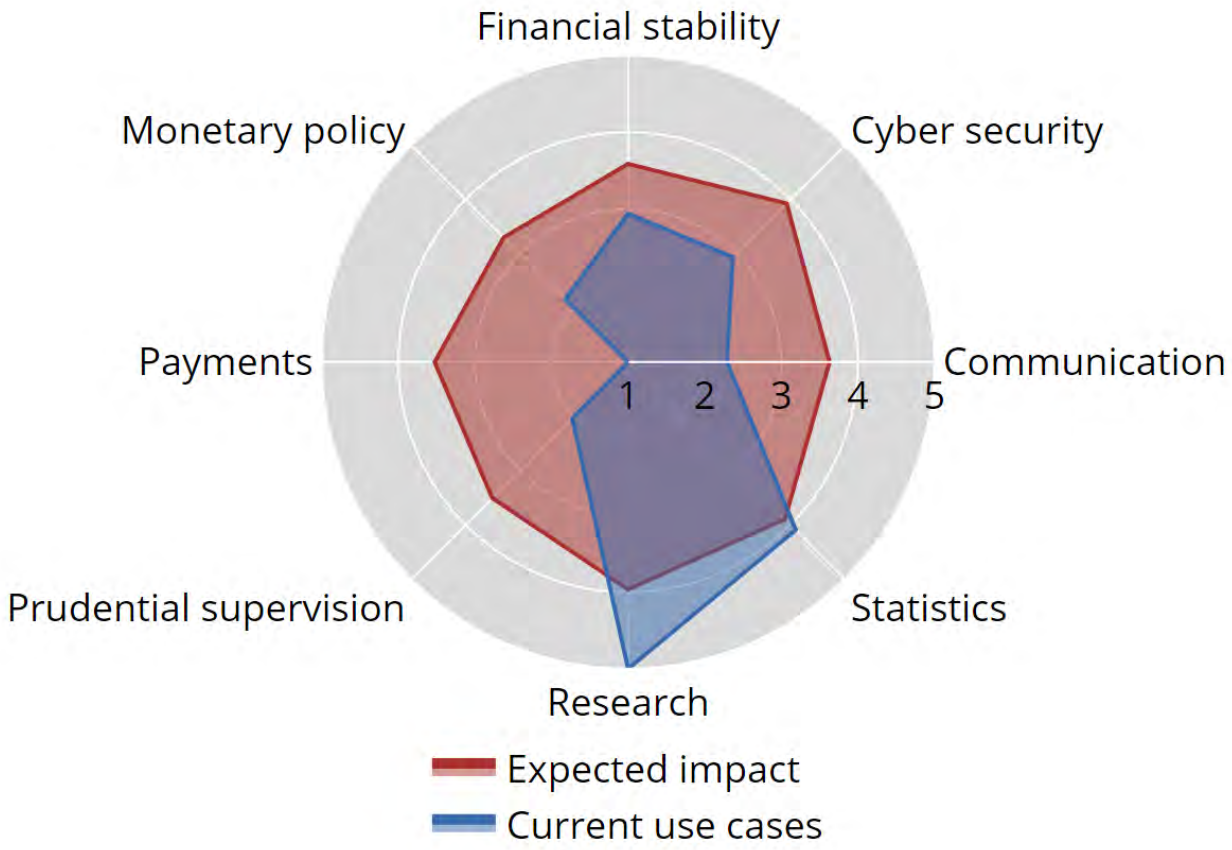
Use cases mostly involve text and code generation



Question: If yes, please select the applicable use cases.

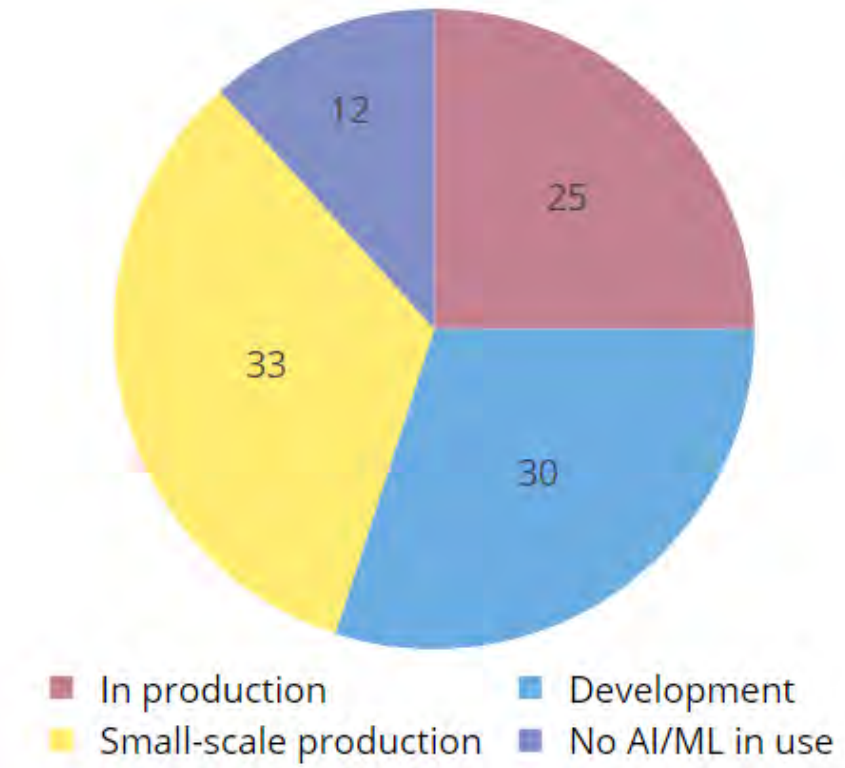
1 Yet **expectations** do not necessarily match **applications** which remain limited

Expected impact and current applications of AI/ML[1]



[1] The expected impact is calculated as the average of the response rates on a scale from 1 to 5 (1: not sure; 2: not impactful at all; 3: slightly impactful; 4: moderately impactful; 5: highly impactful). Current use cases are derived as the total count normalised on a scale from 1 (minimum) to 5 (maximum).

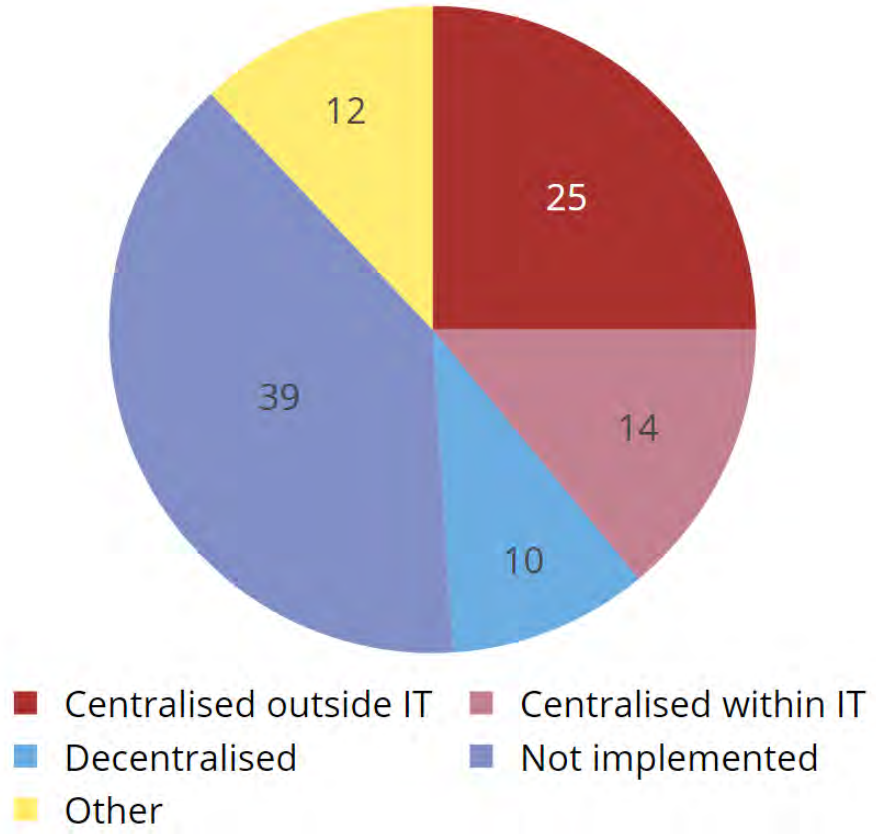
AI/ML use cases



Question: What is the state of AI/ML adoption in your institution?

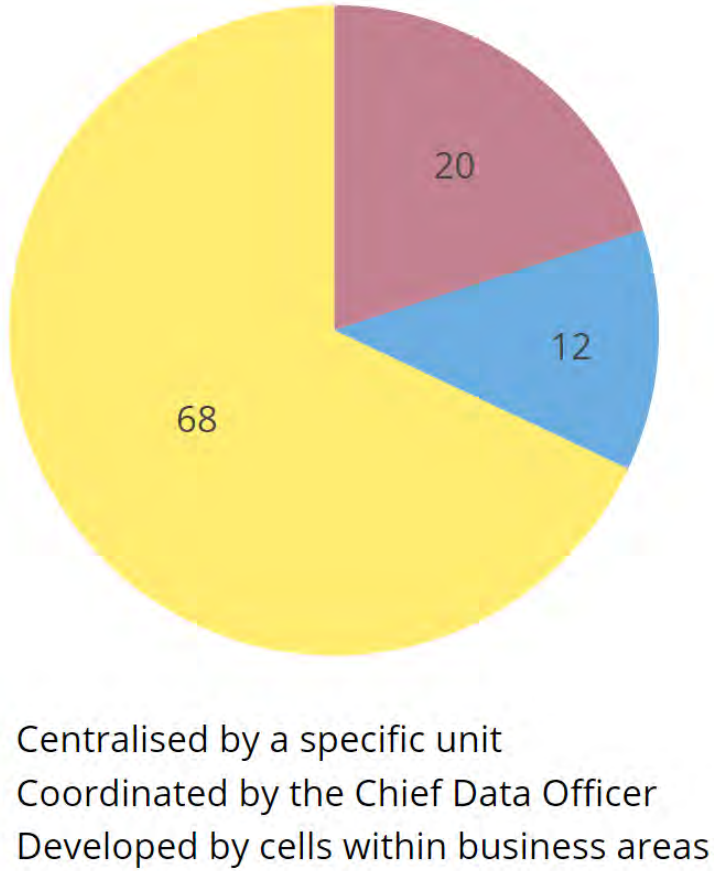
2 **Governance is still in its infancy** and involves multiple stakeholders …

While AI governance is mostly not implemented or centralised…



Question: How is the AI governance role implemented in your institution?

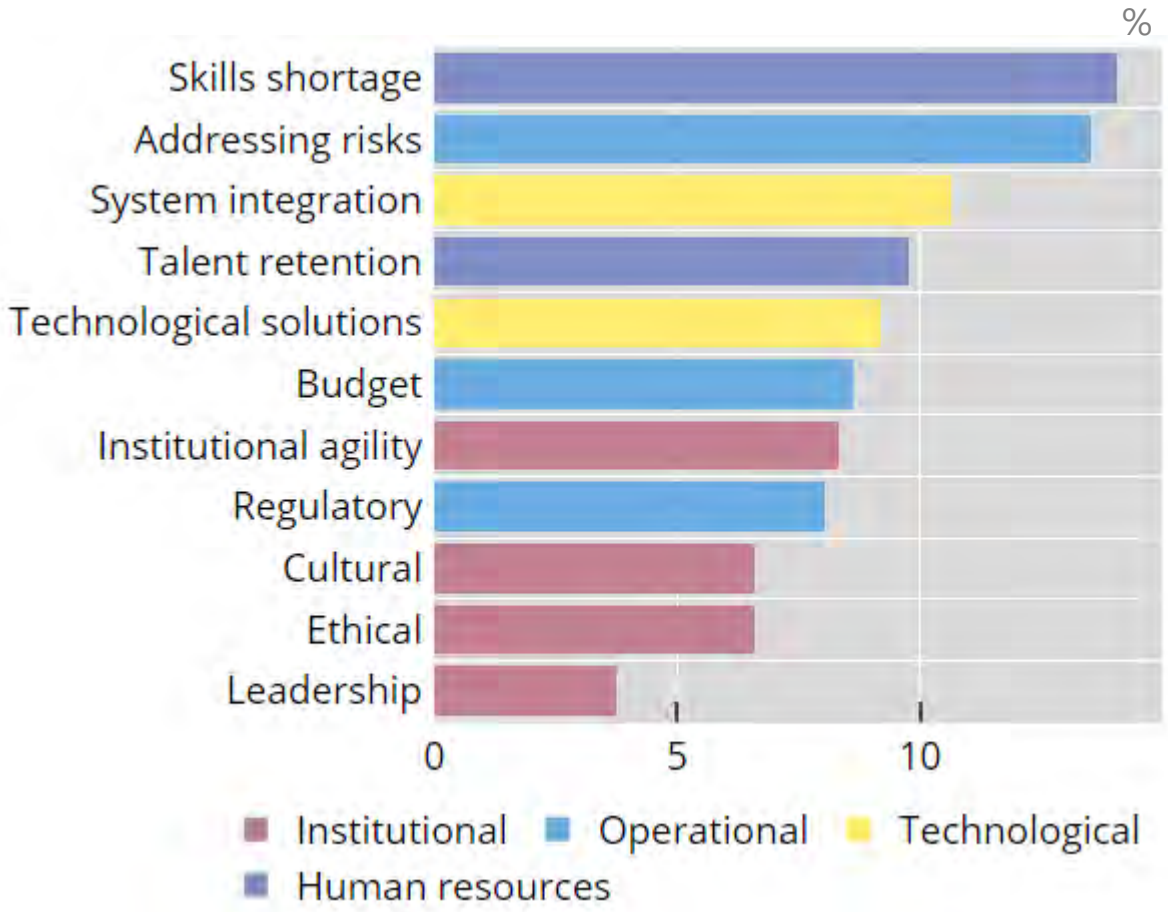
… AI/ML projects are mostly developed across business areas



Question: How is the work on AI/ML projects organized in your institution?

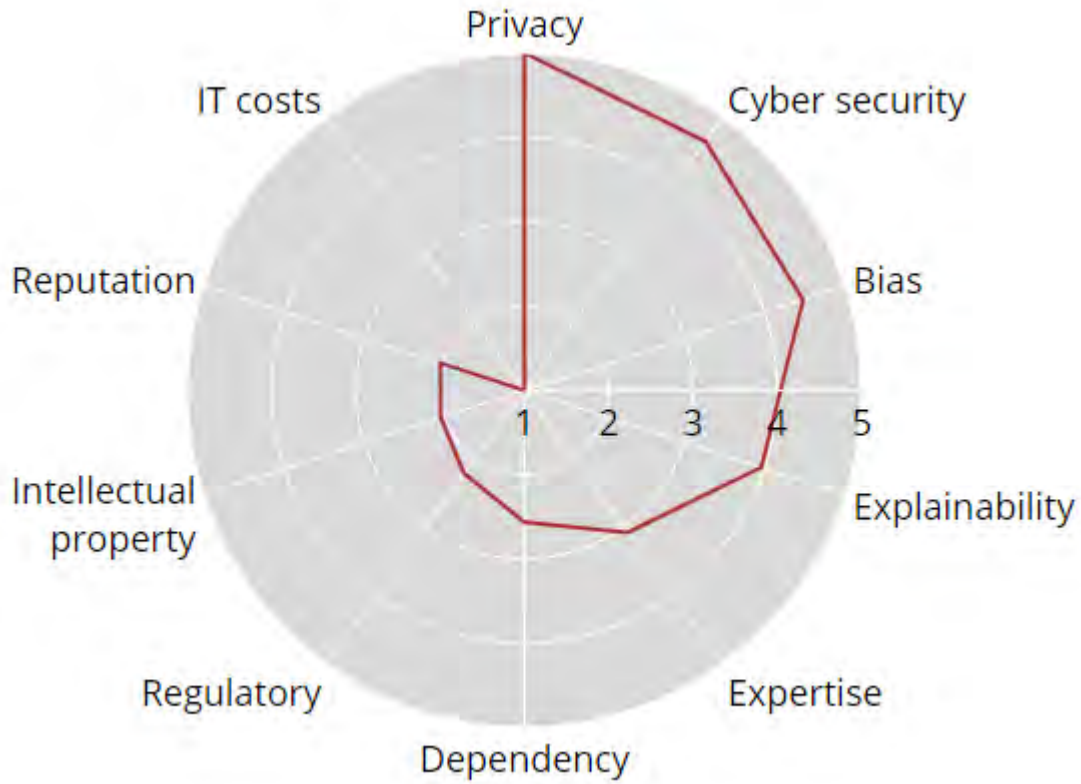
2 AI is also posing some key **challenges** and **risks**

Key barriers to AI/ML adoption



Question: What are the biggest challenges to business adoption of AI/ML?

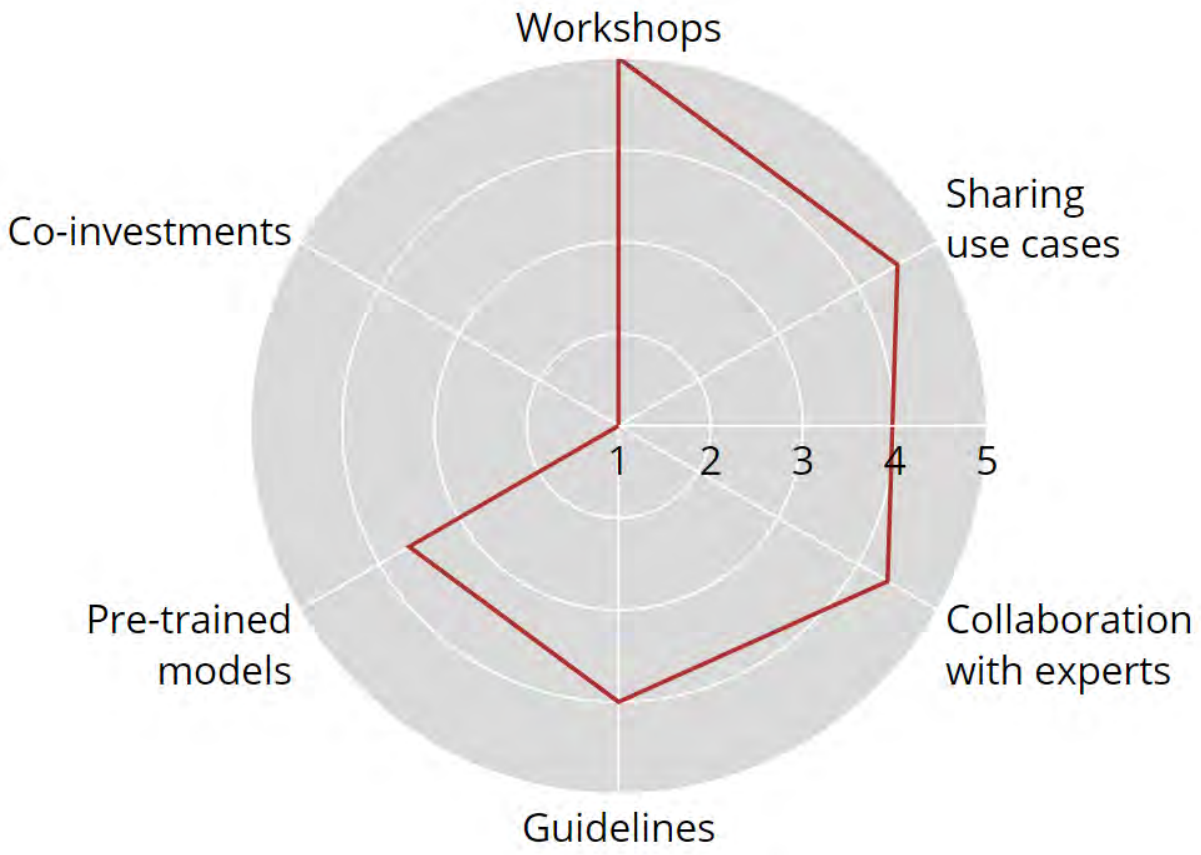
Main risks



Question: Which AI/ML risk does your institution consider relevant?

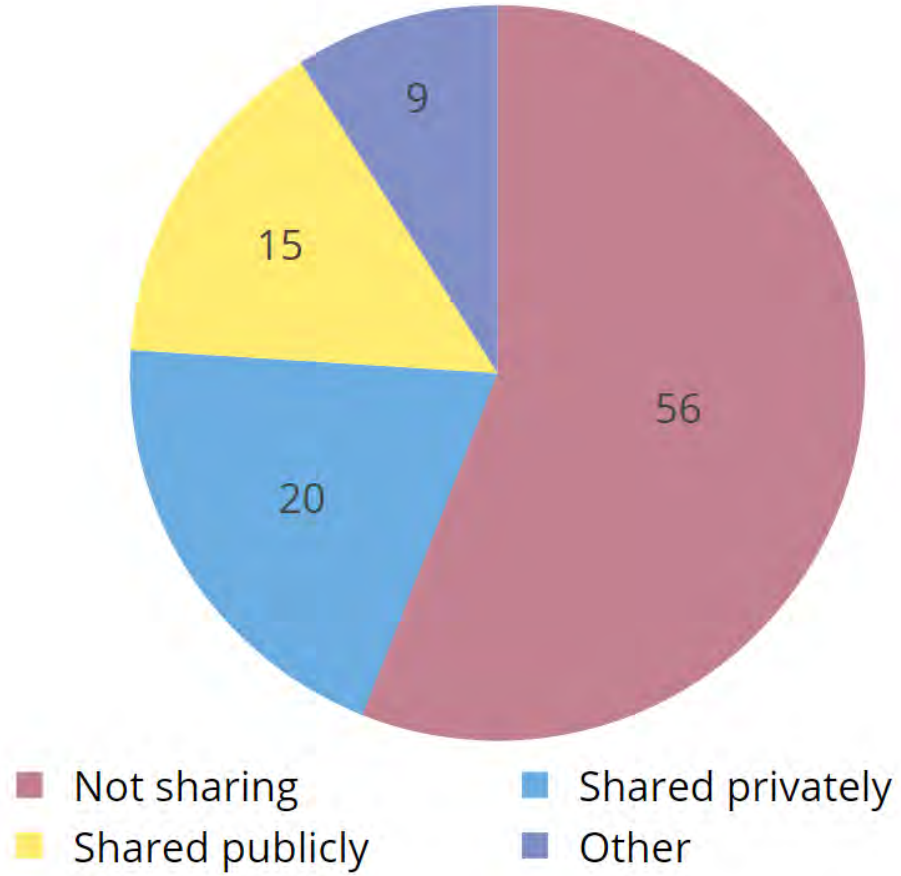
3 Looking forward, three priorities: **sharing**

Sharing knowledge, code and use cases



Question: What are the main avenues for central bank collaboration to address the challenges of AI going forward?

Sharing AI/ML code and models has yet to mature

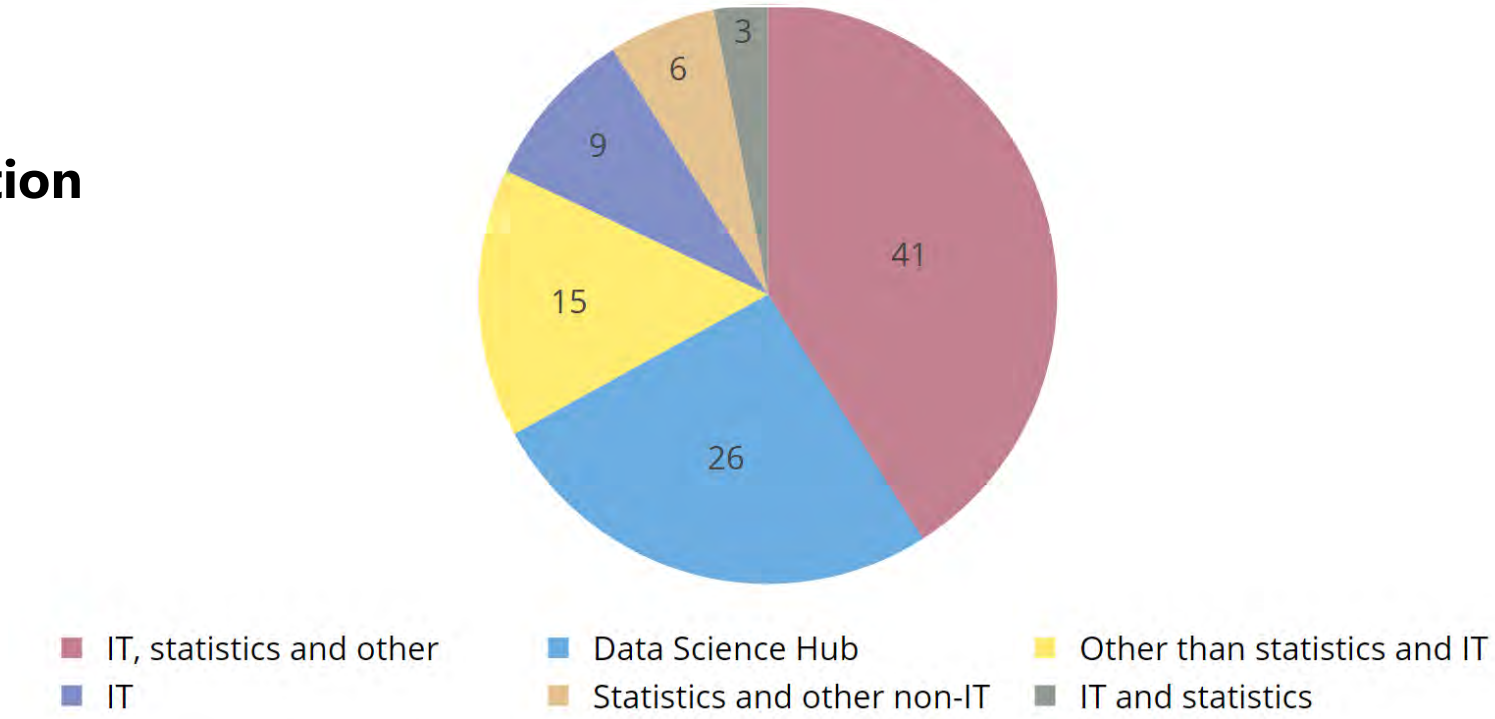


Question: Does your institution share AI/ML code or models, or plans to?

3 Looking forward, three priorities: **collaboration**

- **Collaboration** *within*, *across* and *outside* central banks
- Key role of **international cooperation**
- Coordination to **co-invest** but also to **monitor** and **manage risks**

AI/ML involve strong collaboration within the central bank



Question: Departments involved in current AI/ML-related projects by central banks.

3 Looking forward, three priorities: genAI, **data quality** and data platforms

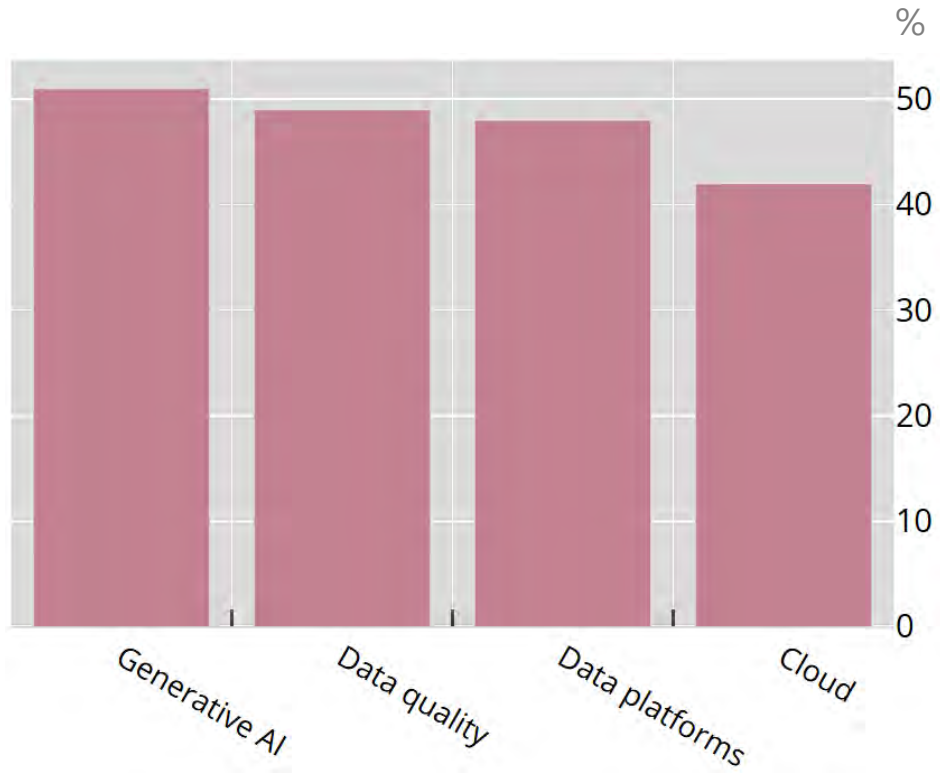
Beyond expanding the use of generative AI, **prioritizing data quality** as a top priority :

- Good **AI** is good **data**
- AI to get the **data right**
- AI to get the **right data**

Central banks as providers of **high-quality data for AI**

Maximize data value through **data platforms**, including **leveraging cloud** services and infrastructures

Priorities in the near future



Question: In your opinion, what will be the emerging technologies topics that will concentrate interest and resources in the upcoming future?

Irving Fisher Committee on
Central Bank Statistics

BIS

Thank you for your attention !

IFC publications →

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Innovation project: large language model adaptation at Latvijas Banka[1]

Mārtiņš Liberts and Rihards Ostrovskis,
Bank of Latvia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Innovation Project: Large Language Model Adaptation at Latvijas Banka

Mārtiņš Liberts and Rihards Ostrovskis[1]

Abstract

Latvijas Banka has conducted an innovation project "Large Language Model Adaptation at Latvijas Banka (LLM-Lab)" according to the *iLab* framework. Innovation laboratory (*iLab*) framework is a method for targeted and rapid evaluation of new ideas, methods, or technologies. The project had two main objectives: identifying bank functions where large language models (LLMs) could be utilized, and developing a prototype for at least one of these identified functions. The project spanned four months and was executed by a team of nine employees from various departments, all working part-time for the project.

Keywords: Innovation framework, large language models, AI infrastructure

JEL classification: O31, O32, O33

[1] Chief Data Scientist (Martins.Liberts@bank.lv) and Data Scientist (Rihards.Ostrovskis@bank.lv), Data and Statistics Department at Latvijas Banka

1. iLab Framework

The Need for Innovation

Innovation is crucial for improving efficiency, optimising processes, and addressing organisational challenges. By embracing new ideas, institutions can streamline operations, freeing up valuable time for more strategic tasks. This approach not only enhances productivity but also helps reduce work overload, creating a more effective and sustainable work environment.

What is iLab?

iLab[2], or the "Innovation Laboratory," is not a physical space or organisational unit but a structured collaboration platform designed to turn ideas into real solutions. It provides a dynamic environment where professionals from different departments can work together, integrating technological advancements with business processes. iLab is a key tool for executing organisational strategies and fostering innovation within Latvijas Banka.

At iLab, employees develop ideas that align with the bank's strategic goals. The platform allows teams to test new concepts before they become official projects, encouraging cross-departmental collaboration and knowledge sharing. By bringing together expertise from various fields, iLab ensures that innovative ideas are explored and refined effectively.

How iLab Works

Each year, approximately three iLab projects are implemented, with participation open to employees from different departments. Participation is entirely voluntary, supported by direct managerial backing. To encourage a healthy competitive spirit, multiple teams may work on the same idea.

iLab follows a structured framework to ensure efficient idea development. The process begins with employees submitting their ideas for iLab projects and voting for ideas. The top voted ideas are then reviewed and prioritised by the Innovation Management Committee. Once selected, teams are formed with clearly defined roles, fostering a collaborative working environment.

During the research and development phase, teams explore potential technological solutions, define business processes, and create prototypes. Projects follow a three-month sprint cycle, using agile methods to experiment and refine their solutions. Teams provide monthly progress updates, culminating in a final review and presentation to the Innovation Management Committee. Finally a seminar is organised about lessons learned and results achieved during the iLab project. This is

[2] Artis Šmits (9 May 2023) "iLab – veids kā inovēt"[presentation slides in Latvian]. Latvijas Banka (internal publication).

the last deliverable of the iLab project. The seminar is open to all employees of Latvijas Banka.

Lessons Learned and Participant Insights

iLab projects produce several valuable outcomes, including the identification of business processes that can benefit from new technologies, a market analysis of available solutions, and an evaluation of the costs and benefits of implementing these innovations. Additionally, teams develop prototypes to demonstrate the feasibility of their ideas.

Participants have found iLab to be an inspiring and rewarding experience. It provides a structured way to explore new ideas, experiment with emerging technologies, and implement creative solutions. Employees appreciate the hands-on learning approach, which allows them to gain practical experience in an innovative environment.

Several key lessons have emerged from iLab's projects. Managerial support is crucial for allowing team members to engage with their projects. Flexibility is also essential, as enabling employees to join or leave teams easily helps maintain enthusiasm and productivity. Additionally, the initiative encourages a mindset where failure is seen as a learning opportunity rather than a setback. Even unsuccessful projects provide valuable insights and contribute to future innovation efforts. Finally, ensuring that iLab projects align with the organisation's strategic goals is critical for long-term success.

Participating in iLab is an exciting opportunity to contribute to organisational development while gaining new skills. It fosters a culture of innovation, encourages collaboration, and allows employees to challenge conventional ways of working. Moreover, iLab participation enhances professional growth and serves as a valuable addition to an employee's CV.

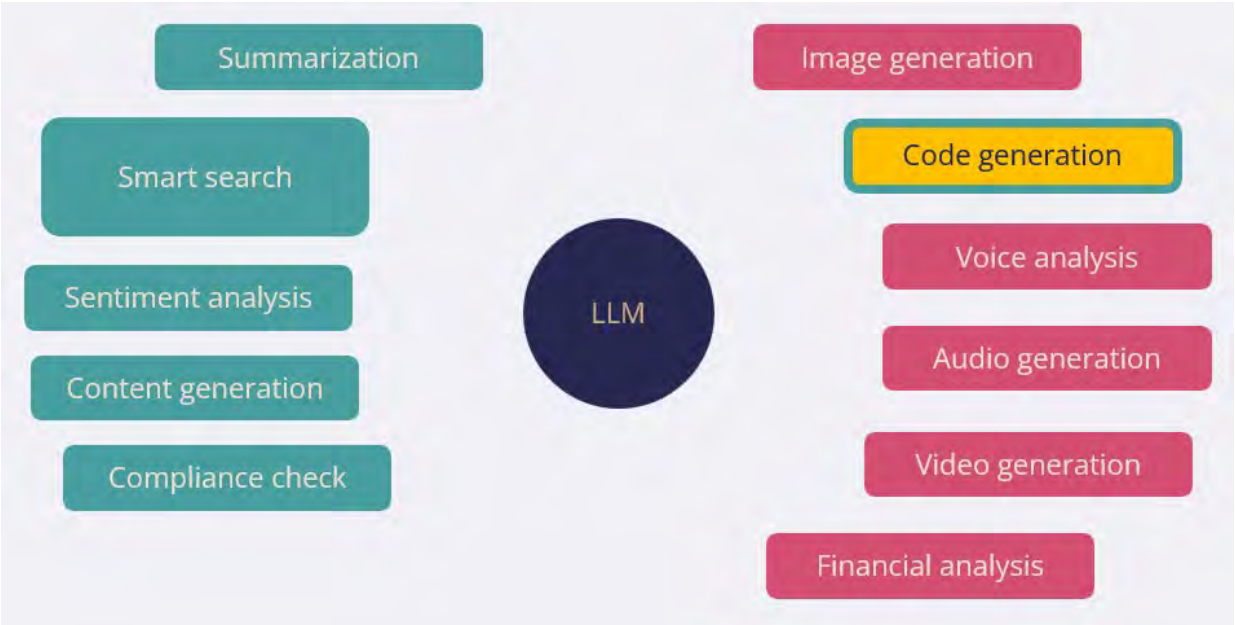
By fostering a culture of continuous learning and innovation, iLab has become an important driver of technological advancement and strategic growth within Latvijas Banka. It empowers employees to experiment, collaborate, and develop creative solutions.

2. LLM-Lab

One of the iLab projects in 2024 at Latvijas Banka was LLM-Lab. Name of this iLab is combination from the name of the iLab framework and technology researched in it – LLM. The goal of this project was to build at least one prototype that is useful for addressing real-world daily challenges. As LLM is general-use technology it was required to narrow down the scope, so it was decided that we will focus on projects that involve only text and not any multi-modal models. As shown in Graph 1, summarization, smart search, sentiment analysis, content generation and compliance appear to be the most promising categories at Latvijas Banka. The existing market products for code generation are effective, while tools for image generation, voice analysis, audio generation, and video generation do not align with the organization's primary objectives. For financial analysis, specialized tools are already in use. Therefore, the ones on the left were chosen as the primary focus.

Scope of LLM-Lab: Key Use-Case Categories

Graph 1



Source: Latvijas Banka LLM-Lab report to Innovation Management Committee

Course of the project

The project team was highly cross-functional including nine colleagues with technical (data scientists, IT architect, SupTech developers and experts) and business (experts from cash operations, law and strategy) backgrounds from several different departments. As it is typical for the iLab projects, also for LLM-Lab there were no members whom this was sole job responsibility.

We remained quite flexible throughout the project by having regular weekly meetings to monitor progress. Overall length of the project was set to around four months.

The overseeing body of this project is Innovation Management Committee involving board members, heads of departments and seasoned experts, ensuring that progress remains aligned with the organization's overall objectives.

Objectives

In order to achieve the goal of the project, first objective was to identify functions at Latvijas Banka where this technology could be effectively applied and prioritize them. Although this type of task differs from typical projects, which usually arise from a clearly identified problem, the technology and experience were still at an early stage of maturity, making it necessary to step back and identify problems to be solved by the LLM technology. After surveys and communication with every department at the organization 39 use cases were identified, two of which had already been initiated by

a few members of LLM-lab before. It is worth noting that all of the use cases widely differ by their complexity, scale and gains.

To delve a bit further into the list of use cases it is worth adding that the list of use cases is not final and remains open to revision; modifications and additions are only welcome. The larger the list, the higher the probability to find the opportunities with the highest value. We defined the top use cases by:

- categorization by categories mentioned in Graph 1;
- information classification levels at Latvijas Banka[3];
- alternatives in other institutions;
- complexity of implementation;
- initiative from stakeholders;
- impact of finished product.

Our first pilot prototypes were chosen based on initiative from stakeholders, impact and information confidentiality level appropriate for Microsoft Azure for easier experimentation and implementation so we can focus more on the solution engineering itself, less on supporting technology. As developing entirely new solutions extra effort from stakeholders is mandatory, and for most of the use cases it was by far the most important aspect for prioritization.

Second objective was to develop at least one generative AI solution within Latvijas Banka. As the participation rate and effort from the team was high, it was decided to work on more than one potential solution and continue working and building prototypes for four initiatives:

- Emission Prospectus compliance checks against Commission Delegated Regulation (EU) 2019/980[4] for licensing and sanctions purposes;
- Smart search in large documents related with cash instructions and procedures;
- Data extraction from letters received from State Police regarding cash issues;
- Assistant for drafting specification for environmentally friendly public procurements.

Additionally, another very relevant aspect – as we are a relatively small central bank, making the correct decisions and spending our resources wisely is very important for us.

A key supporting objective was to gain knowledge about the technology, the methodology for applying it, and the typical real-world scenarios in which it is used. Notably, the project's objectives were not pursued strictly in sequence but were approached iteratively, particularly regarding the learning process. Some theoretical

[3] More in detail in *Information Classification* chapter.

[4] Some of the Annexes of Delegated regulation - 2019/980 - EN - EUR-Lex are being used to find information inside the Prospectus document. If all necessary information is included in a Prospectus document then it complies to the rules and is correct. Available: https://eur-lex.europa.eu/eli/reg_del/2019/980/oj/eng

groundwork was completed in advance, while additional insights emerged during hands-on activities.

LLM-Lab LLM System Prototypes

LLM System Prototypes

Table 1

	Prospectus	Cash Instruction Documents	Data Extraction from State Police Letters	Environmentally Friendly Public Procurement Assistant
Key Use-Case Categories	Smart search + compliance check	Smart search + summarization	Smart search	Smart search + content generation
Complexity (relative)	Medium	Medium	Easiest	Hardest
Knowledge base	Prompt	RAG	Prompt	RAG
Knowledge size	80-100 A4 p.	500 A4 p.	1-2 A4 p.	1000 A4 p.
Language	English	Latvian	Latvian	Latvian
Base model	gpt-4/gpt-4o	gpt-4/gpt-4o	gpt-4/gpt-4o	gpt-4/gpt-4o
Information classification level	Restricted access	Restricted access + specially protected (excluded from prototype)	Restricted access	Public + restricted access
Result	50 documents per year, multiple hours saved per document	Harder to measure in time	200 documents per year, 20 minutes saved per document	Harder to measure in time
Interface	Background service integrated in internal platform	Chat	Background service	Chat
Types of prompts	Pre-defined	Flexible	Pre-defined	Flexible

Table 1 showcases four prototype systems that demonstrate the multifaceted nature of smart search when integrated with large language models. Although each prototype presents a slightly different combination of functionalities – from compliance checking and summarization to content generation – they all rely, at their core, on some form of information extraction. In an LLM context, smart search is understood as the capability to interpret user queries and retrieve, extract, or synthesize relevant content from large text corpus. Complexity is relative among the chosen use-cases – it helped us to decide which prototypes to focus more on at the beginning.

The size of the knowledge base varies for each of our business cases. We began with Prospectus and data extraction from State Police letters because they rely primarily on what we consider the simpler aspect of an LLM system – smart search – and involve a much smaller knowledge base. In particular, data extraction from State Police letters has very short inputs, so it was clear from the outset that a vector database was unnecessary. This simplified the overall LLM system and avoided any risk of inaccuracies when extracting chunks of text.

Language is not an obstacle anymore if the most recent and powerful LLMs are used. According to OpenAI already gpt-4 was performing well for Latvian[5] – far above average. Smaller models for smaller languages could bring some obstacles.

We later concluded that Prospectus smart search also did not require Retrieval Augmented Generation (RAG). It produced better results without it, especially after the release of the GPT-4o model, which features an expanded input context window of 128,000 tokens[6] – compared to the 8,000 or 32,000 tokens[7], of the GPT-4 model. Since 100 pages amount to roughly 80,000 tokens, they comfortably fit within a context window, and we did not observe "lost in the middle"[8] issues.

However, for our other two prototypes, the knowledge base is too large to fit into a context window, making a vector database and RAG necessary to handle the data effectively. The information classification level for all prototypes permits the use of cloud-based LLM APIs, except for some cash-instruction documents, which were excluded due to their relatively small quantity. Nevertheless, we designed the systems to be as modular as possible so that we can switch to locally hosted models at any time, for any reason.

For solutions integrated into established processes – such as Prospectus or data extraction from State Police letters – measuring results is relatively straightforward, given the clear input, output, and time savings compared to previous methods. However, for chatbots, which are used for a broader range of tasks by more colleagues, quantifying benefits is more challenging. This also applies to the user interface: while background services can be seamlessly integrated into existing workflows, adopting a chat interface requires changes in both user behaviour and organizational culture.

The types of prompts significantly affect the complexity of evaluating the system. When using a fixed set of predefined prompts – especially with closed-ended questions, such as extracting data from state police letters – it becomes easier to build a corresponding evaluation dataset. However, developing chat systems with flexible, open-ended questions requires a larger and more diverse dataset, ensuring that the distribution of questions aligns with the underlying knowledge base.

Current status, as of writing this paper (March 2025), is that Prospectus and data extraction from State Police letters have been in production for several months and is being used on regular basis. The rest of the solutions are at the last cycles of evaluation and soon to be implemented in production.

[5] OpenAI "GPT-4". Available: https://openai.com/index/gpt-4-research/

[6] Azure OpenAI Documentation. Available: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models?tabs=global-standard%2Cstandard-chat-completions

[7] Except gpt-4 turbo had 128k input context window length, but gpt-4o is a superior model

[8] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *arXiv preprint arXiv:2307.03172*, 2023. [Online]. Available: https://arxiv.org/pdf/2307.03172

Conclusion and results of LLM-Lab

To begin with, not only was the initial goal achieved, but four prototypes, instead of one, were developed, and future projects and priorities were clearly outlined. Beyond the direct impact, the initiative laid a flexible yet rigorous foundation by involving participants from across the organization, thus fostering a much broader understanding of applied GenAI among both technical and business colleagues. As of March 2025, this groundwork has already contributed to numerous projects and will continue to support the development of future initiatives.

Overall, LLM-Lab provided a highly impactful and well-structured approach to exploring this new and immensely broad topic.

3. Way forward

Democratising LLM Technology

To fully unlock the potential of large language models (LLMs), we aim to democratise their use across the organisation, ensuring that every employee – not just technical specialists – can create and customise their own AI models. This includes providing a no-code interface, allowing users to build and fine-tune models effortlessly, regardless of their technical expertise. By making AI model development accessible to all, we foster innovation, improve efficiency, and enable employees to tailor AI solutions to their specific needs.

Achieving this vision requires a robust infrastructure that is functional, intuitive, and secure. Employees need a seamless and reliable environment where they can experiment with LLMs, refine their models, and integrate them into their workflows. Security is paramount, ensuring that sensitive data remains protected while maintaining ease of use.

To ensure broad usability, each custom model must be accessible through two key interfaces. First, a web-based chat interface will allow employees to interact with their models in a user-friendly manner, making AI-driven assistance readily available for various tasks. Second, an API integration will enable these models to be embedded into existing information systems, streamlining operations and enhancing automation.

By providing these tools and infrastructure, we create an ecosystem where AI is not just a specialised technology but a widely available resource that empowers every employee to innovate and improve their daily work.

Information Classification

The current information classification framework used at Latvijas Banka distinguishes between three levels of information:

- publicly available information;
- restricted access information;

- specially protected restricted access information.

Publicly available information can be stored and processed across all environments without restrictions. However, restricted access information may only be stored and processed on on-premises infrastructure or Microsoft Azure Cloud, while general cloud platforms are not permitted for this type of data storage. The most sensitive category, specially protected restricted access information, must be stored exclusively on on-premises infrastructure to ensure the highest level of control.

This is a crucial requirement we should be taking into account when building our LLM infrastructure. This is a reason we have chosen to build two environments, namely:

- Environment with cloud models hosted in Microsoft Azure – usable for restricted access information;
- Environment with local models hosted on premises – usable for restricted access information and also specially protected restricted access information.

This dual-environment approach ensures that AI capabilities can be leveraged securely while adhering to strict data protection standards.

Technology Stack

We have chosen Open WebUI[9] as a core technology for building the LLM infrastructure. Open WebUI is an open-source project delivering extensible, feature-rich, and user-friendly self-hosted AI platform. Open WebUI serves as the primary interface for user interactions, whether through direct usage or system integrations.

In the cloud environment, LLM models are hosted on Azure AI Services[10], while in the on-premises environment, they are managed by Ollama[11]. The backend processes, including request handling and data storage, are managed by supporting components such as LiteLLM[12] (an open-source gateway for cloud hosted models) and PostgreSQL[13] (database engine), but these remain transparent to users. Graph 2 explains several usage scenarios. These usage scenarios illustrate how different users interact with the AI system according to their specific roles and security requirements.

[9] Timothy Jaeryang Baek et. al. "Open WebUI" [Online]. Available: https://docs.openwebui.com/

[10] Microsoft "Azure AI services documentation" [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/

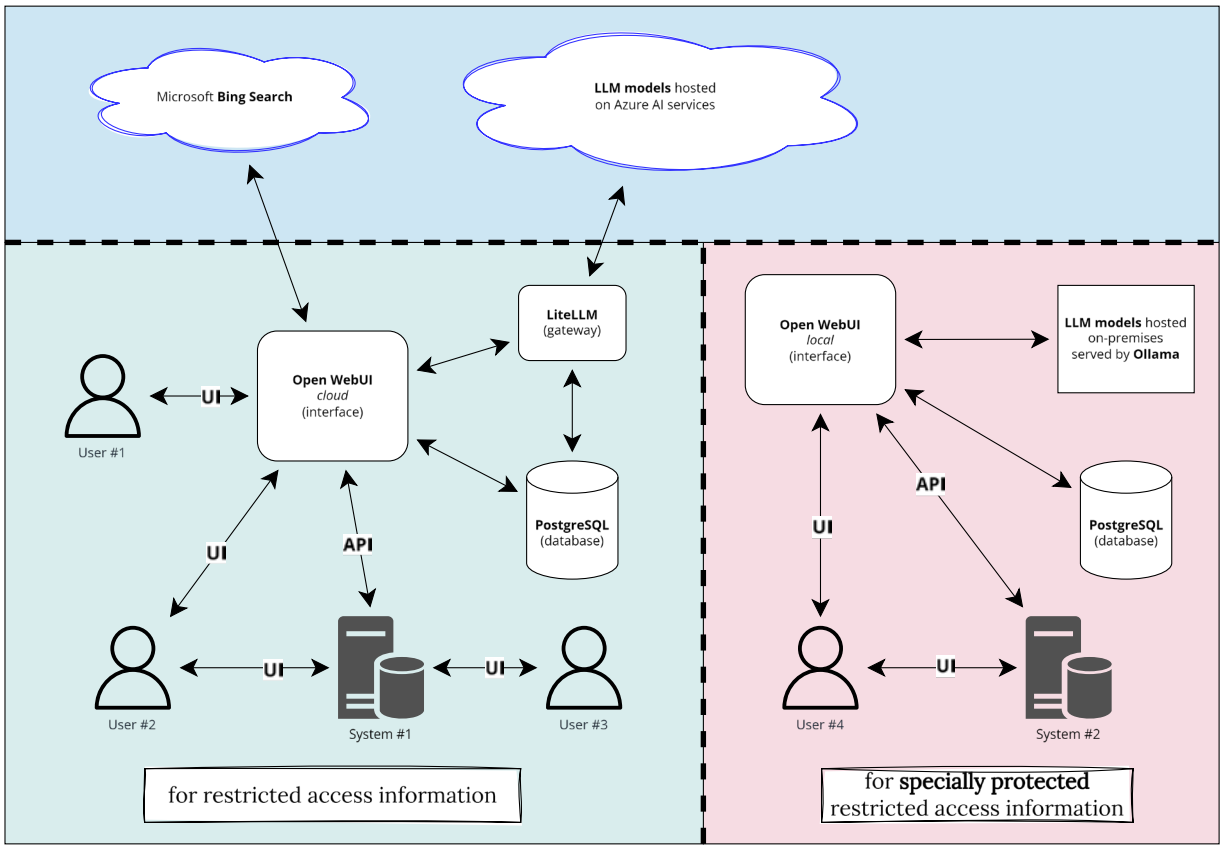
[11] Jeffrey Morgan et. al. "Ollama Documentation" [Online]. Available: https://github.com/ollama/ollama/blob/main/docs/README.md

[12] Ishaan Jaffer and Krrish Dholakia "Lite LLM" [Online]. Available: https://docs.litellm.ai/

[13] The PostgreSQL Global Development Group "PostgreSQL 17.4 Documentation" [Online]. Available: https://www.postgresql.org/docs/current/

Architecture sketch

The sketch explains two environments for cloud based and on-premises hosted models Graph 2



[1] A sketch for an Open WebUI architecture for restricted and specially protected restricted information

Sources: Authors own drawing

User #1 represents researchers using cloud-based AI. A researcher requires AI-powered assistance for analysing publicly available or restricted access information. They interact with Open WebUI (*cloud*), which provides access to LLM models hosted on Azure AI Services. If needed, the system can retrieve additional data from Microsoft Bing Search.

User #2 represents analysts integrating AI with internal systems. They work with restricted access information and requires AI support within an internal system. They interact with Open WebUI (*cloud*) and integrate AI functionalities via an API, allowing automation and enhanced workflow efficiency. The infrastructure ensures secure handling of restricted access information while providing easy access to AI capabilities.

User #3 represents users utilising LLM technology exclusively through system integration rather than direct interaction with Open WebUI. Their internal system connects via an API, enabling seamless access to LLM models hosted in Azure AI Services. This integration allows AI-powered automation while maintaining security and compliance with information classification policies.

User #4 represents analysts working with specially protected restricted access information. Their interaction mirrors that of User #2, except they use Open WebUI

(*local*) to access LLM models hosted on-premises, managed by Ollama. This setup ensures that all data remains within a highly secure infrastructure while enabling AI-powered analysis and decision-making.

The integration of these technologies offers several benefits. A unified interface provided by Open WebUI simplifies interactions with various LLMs through a single, cohesive platform. The framework supports custom model development through techniques such as prompting and retrieval-augmented generation (RAG), enabling employees to customise AI models according to their specific needs. Finally the framework allows usage and sharing of created custom models through the web based chat interface or API integration in existing information systems.

Conclusions

The iLab framework proved to be an effective platform for initiating a structured and goal-oriented exploration of LLM technology at Latvijas Banka. Through this process, we identified multiple potential use cases for LLM integration and successfully developed four functional prototypes. The positive outcomes of the LLM-Lab project have reinforced our commitment to further implementing LLM technology across various banking processes.

A key priority moving forward is the democratisation of LLM technology, ensuring that AI-powered tools are accessible to all employees. We recognise this as a crucial step toward modernising and optimising our institution's operations. To achieve this, we are actively developing a functional, intuitive, and secure environment that enables seamless interaction with both generic and custom LLMs, fostering innovation and efficiency across all departments.

Innovation project: large language model adaptation at Latvijas Banka

Mārtiņš Liberts and Rihards Ostrovskis
Latvijas Banka

February 18, 2025

Table of contents

- *iLab* framework
- *LLM-Lab* project
- Way forward
- Summary

iLab framework

iLab framework

- *iLab* = **Innovation Laboratory**
- A method for targeted and rapid testing of new ideas, solutions, and technologies
- Benefits of *iLab* – **structured innovation**:
  - Work smarter in the long term, not harder
  - Achieve the bank's strategic goals
  - Enhance professional capabilities
  - Foster constructive collaboration with colleagues from other departments
  - Purposefully go the *extra mile*
  - Diversify work life with interesting and exciting projects

Key principles of *iLab*

- **Permission to Fail**
  - Mistakes are allowed, but inaction is not tolerated
  - “Fail fast” is considered a positive outcome
- **Structured and Purposeful Work**
  - Focused, structured work without unnecessary bureaucracy
- **Deliverables**
  - Ends with a deliverable: feasibility of the idea, prototype, potential costs, benefits, or a well-evaluated decision to abandon the idea

The process of *iLab*

- Submission of ideas
- Voting for submitted ideas
- Approval of *iLab* projects by the Innovation Management Committee
- Project work for three months
- Monthly interim reports
- Final deliverable

LLM-Lab project

LLM-Lab objectives

- **Identify Bank Functions**
  - Determine the bank functions where implementing an LLM-based solution would be most beneficial
- **Develop AI Solution**
  - Create a generative AI solution within Latvijas Banka

LLM-Lab innovation steps

- Vision
- Technical capacity
- Business problem
- User identification
- Experimenting and prototyping

Vision for solutions

- Summary creation
- Smart search
- Sentiment analysis
- Text generation
- Compliance assessment

Business scenarios

- 39 possible business scenarios were identified, some examples:
  - Processing email correspondence with market participants
  - Evaluating EU regulatory projects for compliance with Latvian interests
  - Processing EU questionnaires and preparing responses
  - Generating flowcharts from text to evaluate procedure quality
  - Obtaining information on large monetary transactions that could affect the balance of payments

Prototypes

- Emission Prospectus compliance checks against the Commission Delegated Regulation (EU) 2019/980 for licensing and sanctions purposes
- Smart search in large documents related with cash instructions and procedures
- Information extraction from letters received from State Police regarding cash issues (**in production**)
- Assistant for drafting specification for green public procurements

Way forward

Information classification

	On-premises	Microsoft Azure Cloud	Cloud
Publicly available information	✓	✓	✓
Restricted access information	✓	✓	✗
Specially protected restricted access information	✓	✗	✗

Technology stack

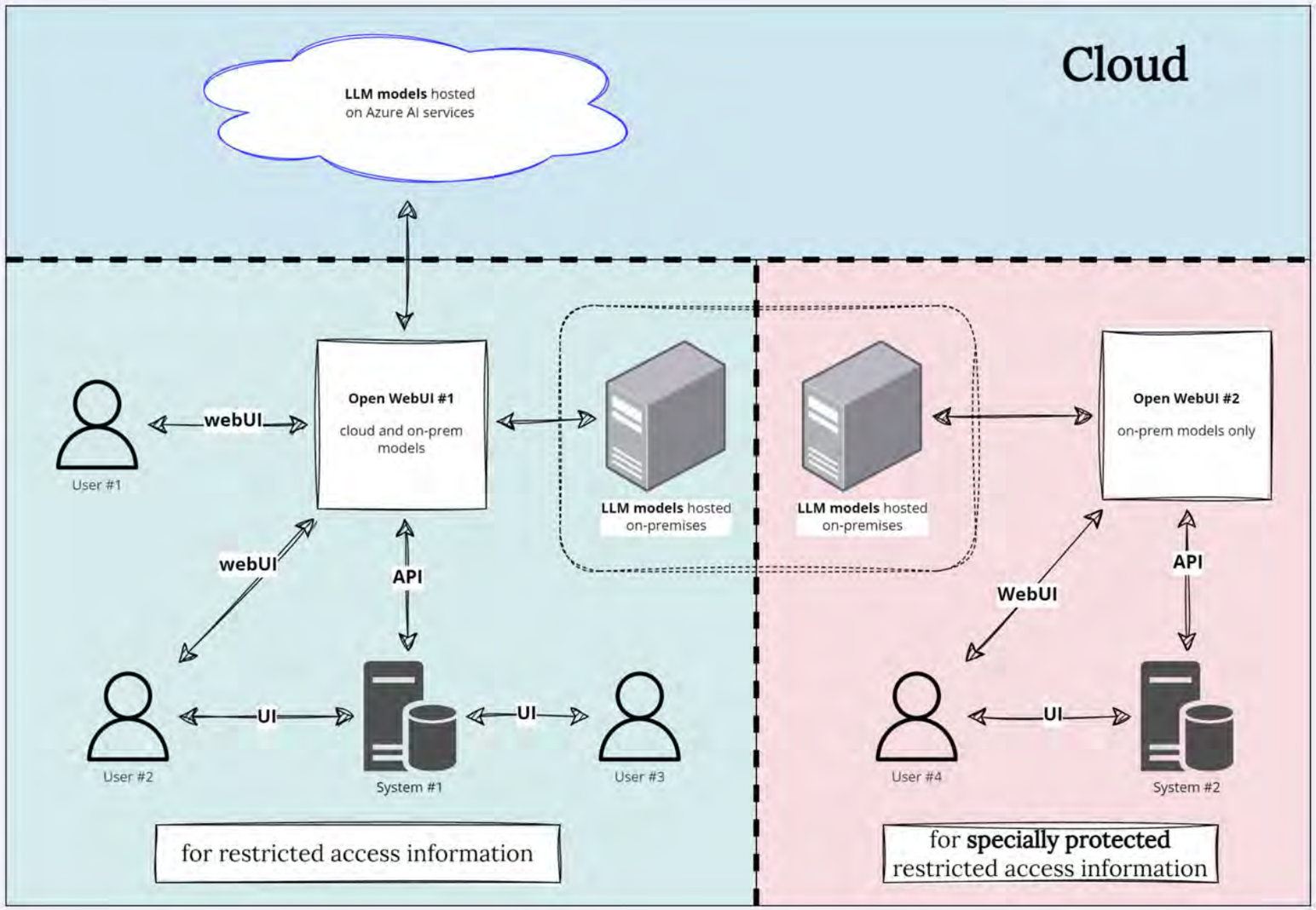
Azure AI Services for hosting OpenAI GPT models in cloud

Ollama for hosting open-source models on-premises

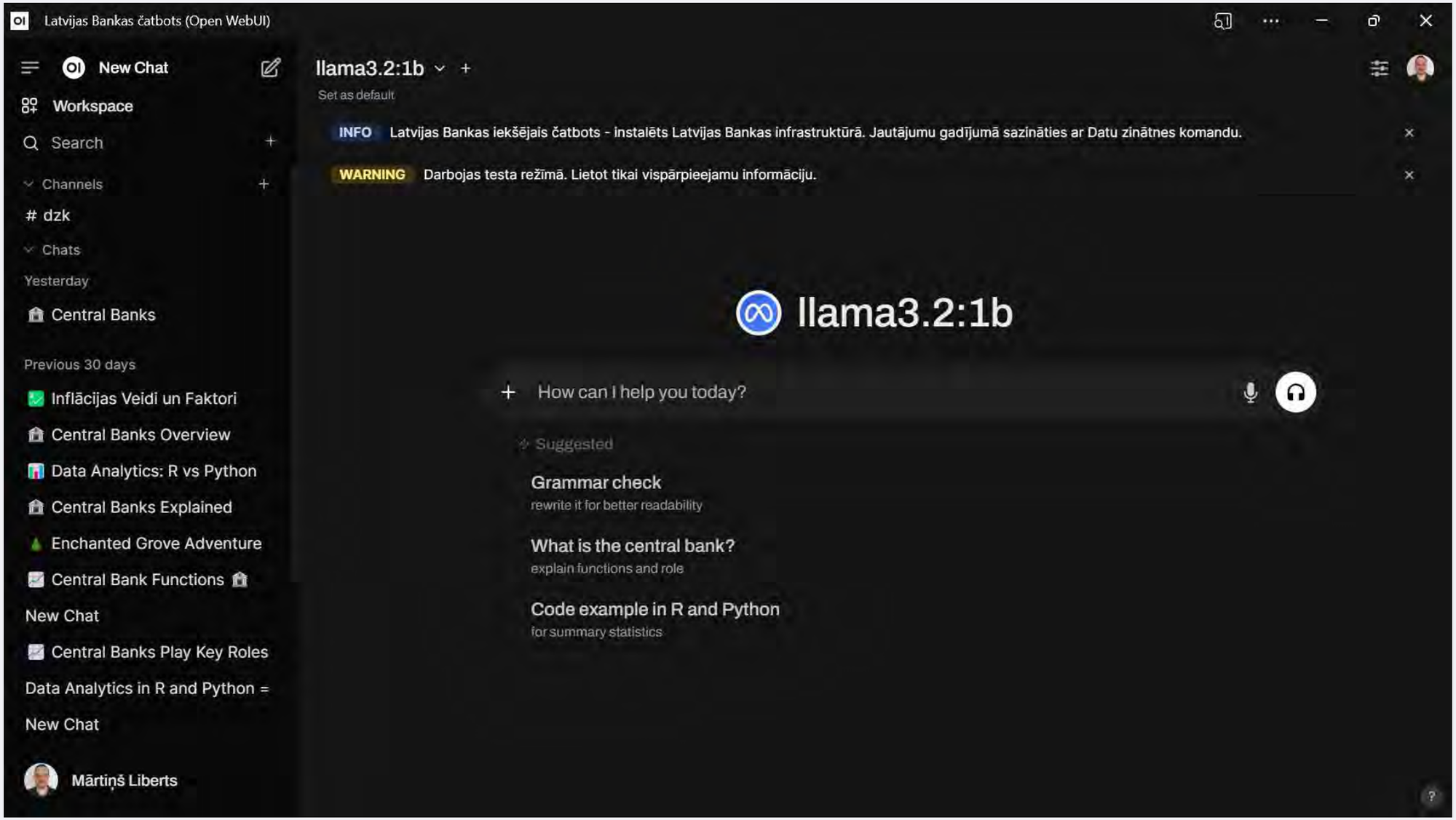
Open WebUI as user and application interface for LLM models

- Unified Interface: Simplify interactions with different LLMs through a single, integrated platform.
- Ease of Implementation: Quick start integration with comprehensive documentation and community support.
- Custom models built by prompting and retrieval-augmented generation (RAG)

Architecture sketch



UI example


Latvijas Bankas čatbots (Open WebUI)
New Chat
Workspace
Search
Channels
dzk
Chats
Yesterday
Central Banks
Previous 30 days
Inflācijas Veidi un Faktori
Central Banks Overview
Data Analytics: R vs Python
Central Banks Explained
Enchanted Grove Adventure
Central Bank Functions
New Chat
Central Banks Play Key Roles
Data Analytics in R and Python =
New Chat
Mārtiņš Liberts
llama3.2:1b
Set as default
INFO Latvijas Bankas iekšējais čatbots - instalēts Latvijas Bankas infrastruktūrā. Jautājumu gadījumā sazināties ar Datu zinātnes komandu.
WARNING Darbojas testa režīmā. Lietot tikai vispārpieejamu informāciju.
llama3.2:1b
How can I help you today?
Suggested
Grammar check
rewrite it for better readability
What is the central bank?
explain functions and role
Code example in R and Python
for summary statistics

Summary

Summary

- *iLab* framework – a method for targeted and rapid testing of new ideas, solutions, and technologies
- *LLM-Lab* project – the first step for implementing an LLM-based solutions
- Hybrid architecture with LLM models hosted on-premises and in cloud accessible through an unified interface provided by the Open WebUI

Thank you!

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

A natural language processing toolbox for the National Bank of Romania[1]

Claudia Voicilă and Daniel Șerbu,
National Bank of Romania

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

A Natural Language Processing toolbox for the National Bank of Romania

Claudia Voicilă, Daniel Șerbu[1]

Abstract

This paper advocates extending beyond traditional numerical data by enhancing central bank analyses with textual information. We propose integrating Natural Language Processing (NLP) tools into the National Bank of Romania's (NBR) analytical toolkit to align with the EU AI Strategy, national and international digitalization priorities, and Romania's objectives to strengthen financial system supervision in line with OECD standards. To this end, we introduce an NLP toolbox specifically tailored for the NBR, aimed at increasing the efficiency of economists in processing textual sources such as financial news and press releases. The toolbox's primary objective is to align the NBR with the best practices in central banking innovation.

The first part of the analysis examines the use of NLP in processing monetary policy statements, starting with the NBR's adoption of inflation targeting. The dataset was expanded to include statements from the European Central Bank, the U.S. Federal Reserve, and 25 other central banks from emerging economies. A supervised sentiment analysis using built labels was conducted by calibrating FinBERT, a Transformer-based model designed to capture contextual relationships in financial text. This model enables more accurate sentiment classification by identifying domain-specific terms that frequently convey either negative or positive sentiment.

The second part of the paper explores the NBR's daily Romanian financial news archive. Here, a specialized Romanian financial dictionary was developed to build a sentiment index. Sentiment intensity scores were assigned using a Generative AI agent. Additionally, topic modelling via Latent Dirichlet Allocation (LDA) was employed to identify and track the evolution of major news topics, highlighting key themes in the financial news over time.

Keywords Natural Language Processing, Financial Stability, Central Bank Communication

JEL classification: E58, C55

1. Introduction

The domain of Natural Language Processing in Central Banking is still a novelty, different applications emerging constantly, as the desire to keep up with the innovation in commercial banks is high. Despite the level of complexity being lower

[1] The opinions are those of the authors and do not necessarily reflect the views of the National Bank of Romania

than research in computer science areas, economists from central banks around the world gather often to discuss ideas about digital transformation in banking. This paper aims to provide an overview of common NLP analyses on text data from the National Bank of Romania, offering a fresh perspective on a task that has not been explored in depth within the institution. Therefore, the main goal of the present work is to align NBR to the good practice of central banks and get inspired by the rapid technological changes of the last years in terms of text analysis. Building a toolbox for Romanian language is not an easy task as it is not a widely spoken language and since Romanian has many different particularities than English, the language of most NLP tools.

Communication of central banks has an important role in ensuring the effectiveness of monetary policy and supporting the credibility of the institution (on a more general note). In a way, communication can be considered a monetary policy tool (due to the impact it has on inflation expectations and economic behaviour) but can also reveal financial stability concerns. A central bank can communicate in many ways: monetary policy decisions, governor and other board members' speeches, reports (Inflation Report, Financial Stability Report) and presentations of spokespersons etc.

In monetary policy decision statements words are chosen with extreme care, as even the slightest nuance can influence financial markets and shape economic expectations. Usually, a central bank's communication is likely to be efficient only if it is perceived as a credible institution by the public. The relationship between the bank's credibility and effective communication has long been recognized by central bankers, as credibility is an important factor in shaping economic agents' expectations.

This paper proposes a series of custom-made NLP tools for analysing financial news in Romanian and Monetary Policy Statements in both Romanian and English, offering insights into the interaction of central bank communication and financial market perception. Sentiment analysis plays a key role in assessing how the press reacts to monetary policy decisions and that's an essential factor for monetary policy efficiency. At the same time, topic modelling allows us to uncover the most relevant themes in financial news over time, providing a clearer list of subjects of public interest. By leveraging these tools, we aim to enhance the understanding of how monetary policy decisions are communicated, perceived, and ultimately, how they shape financial expectations.

2. English Monetary Policy Statements Database

Monetary policy is the primary responsibility of a central bank, with price stability usually as its foremost objective. Generally, words are chosen with extreme care in monetary policy decision statements, as even the slightest nuance can influence financial markets and shape economic expectations. A central bank's communication is effective only when it is perceived as a credible institution, as credibility is essential for shaping economic agents' expectations and has long been recognized as a key factor by central bankers.

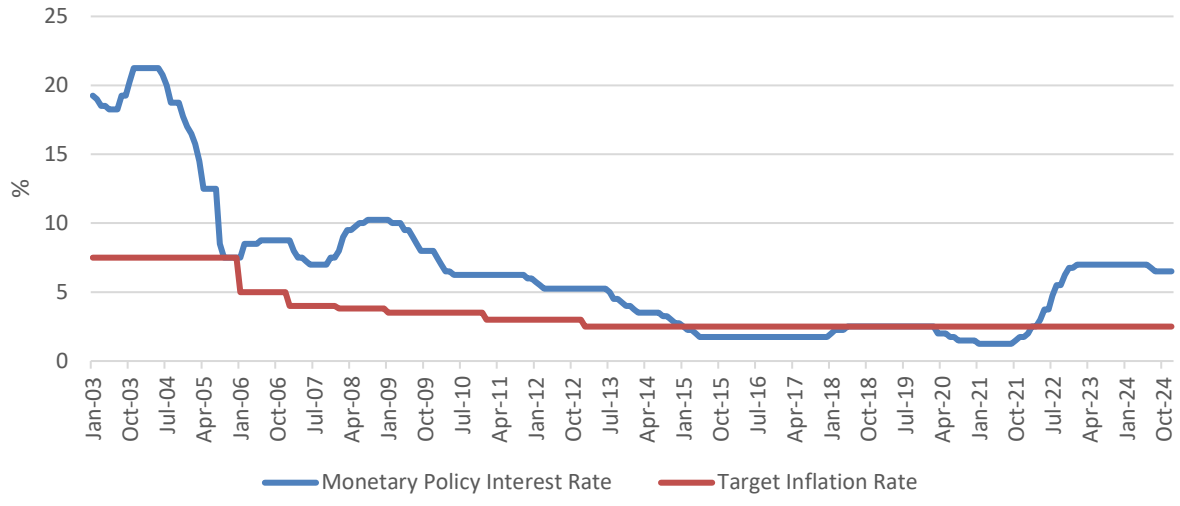
It is possible to get a sense of how efficiently central banks communicate by analysing the clarity, coherence, and influence of their messages on financial markets, economic expectations, and public sentiment. Using NLP techniques, we can also examine how well central banks convey important policy decisions by identifying key events—such as financial crises, policy shifts, and macroeconomic shocks—in textual data. This helps us understand how central banks respond to these events and the broader economic impact of their communication strategies.

The monetary policy statement released by the central bank provides insights into current economic conditions and future outlooks, while also highlighting a range of market risks and vulnerabilities. We employed methods for automated content analysis to extract sentiment from each monetary policy statement since the NBR adopted inflation targeting (2005). By developing a sentiment indicator that measures the frequency of positive or negative words in each board decision published on the website, we constructed credibility indexes for monetary policy based on these statements.

We have web-scraped using a tool called Beautiful soup by Richardson (2007) the monetary policy decisions from the press releases of the Board of the National Bank of Romania that are also published in English (apart from the original Romanian text. The National Bank of Romania released a total of 164 monetary policy statements during 2003 and 2024, out of which 50 were linked to loosening monetary policy (meaning cutting down the reference interest rate), 26 were related to tightening monetary policy (meaning raising the reference interest rate), whereas 88 monetary policy decisions were neutral, meaning they maintained the same level as before for the monetary policy rate. The main factor is the inflation rate, as the purpose of the Romanian central bank is to ensure and maintain price stability, with an inflation target of 2.5% ±1 p.p. since 2013 (Figure 1).

Evolution of the monetary policy rate and the target inflation rate

Figure 1



Source: NBR

We enlarged our monetary policy database with monetary policy statements in English from 25 emerging market central banks (Asia: Armenia, Georgia, Israel, India, Kazakhstan, Malaysia, Mongolia, Pakistan, Philippines, South Korea and Thailand; Europe-Asia: Turkey and Russia; Europe: Czech Republic, Hungary, Poland and Ukraine, South America: Brazil, Chile, Colombia and Peru; Africa: Egypt, Nigeria and South Africa; North America: Mexico) along with the Federal Reserves in the United States and the European Central Bank's press-conference introductory statements that rules over the eurozone countries (Austria, Belgium, Croatia, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Portugal, Slovakia, Slovenia and Spain). Using a larger data set we can underline common global trends and better understand the monetary decisions of the National Bank of Romania since they are strongly influenced by the decisions made by European Central Bank (ECB) and the Federal Reserve's Board (FED) in the USA. The extended database contains 4112 monetary policy statements from 1998 to 2023.

The extended dataset was gathered by Evdokimova et al (2023) and offered by Araujo (2023) that introduced "gingado", an open-source Python library designed to enhance the use of machine learning in the fields of economics and finance. This tool has relevant datasets and integrates seamlessly with major machine learning libraries and offers functionalities that simplify the augmentation of user datasets with authoritative data via the SDMX protocol. Gingado includes a benchmarking feature that optimizes random forest models for robust performance across various scenarios without extensive user input. Furthermore, the library aids in the rigorous documentation of machine learning models, incorporating ethical considerations into the documentation process. The platform is under continuous development, with new features being added to address the specific needs of economic and financial data analysis.

We assigned labels to our monetary policy statements database based on the reference monetary policy interest rate obtained from the Refinitiv database. Based on the monetary policy rate at each decision date, we created three categories: positive sentiment if the interest rate was lowered (loosening monetary policy), negative sentiment if the interest rate was raised (tightening monetary policy), and neutral if the interest rate was unchanged. We then regrouped the neutral category as positive if the previous decision was positive (indicating that the positive economic sentiment continues) and negative if the previous decision was negative (indicating that the negative economic sentiment continues), but only for interest rate levels around 3%, which is considered a low value.

On the extended English Monetary Policy Statements database with labels, we leveraged a pre-trained FinBERT model by ProsusAI, which is specifically designed for financial sentiment analysis. This model, originally trained on financial texts such as SEC filings, earnings calls, and analyst reports, is better suited for monetary policy sentiment classification than general-purpose models like our first option, DistilBERT by Sanh (2019). While DistilBERT is optimized for speed and trained on general sentiment datasets (such as SST-2), it lacks domain-specific knowledge, making it less effective in capturing the nuances of central bank communication.

To enhance sentiment analysis accuracy in monetary policy statements, we fine-tuned FinBERT on a custom dataset of central bank communications, classifying text statements as either positive or negative (removed the neutral sentiment).

Additionally, we extended its application to score individual words based on their sentiment polarity, allowing a deeper exploration of sentiment-laden terms within monetary policy.

The performance of the fine-tuned FinBERT model on the sentiment classification task demonstrates strong accuracy of 90% on the test set (Table 1) Training over 10 epochs resulted in a notable reduction in training loss, particularly in early epochs, signifying effective learning. In terms of classification metrics, the model achieved a precision of 86.5% for negative sentiment and 92.5% for positive sentiment, with recall values that indicate balanced performance across both classes. The F1-scores of 87.8% (negative) and 91.6% (positive) further confirm FinBERT's ability to classify sentiment with high accuracy, ensuring reliable sentiment extraction from monetary policy statements.

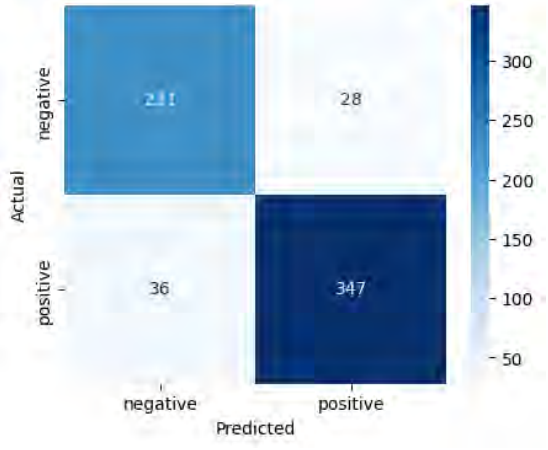
Classification report of fine-tuned FinBERT on test sample (10 epochs) Table 1

	Precision	Recall	F1-score	Support
Negative	86.52%	89.19%	87.82%	259
Positive	92.53%	90.60%	91.56%	383
Accuracy			90.03%	642
Macro accuracy	89.53%	89.89%	89.69%	642
Weighted average	90.11%	90.03%	90.05%	642

Source: own computations

After training, we saved the fine-tuned FinBERT model and tokenizer for future use. Given its specialization in financial text analysis, we leveraged this pre-trained model to extract and validate a list of positive and negative words in English, which served as the foundation for our sentiment dictionary. Rather than using the model for document-level sentiment scoring, we focused on identifying sentiment-laden vocabulary relevant to the financial domain. This approach highlights FinBERT's utility not only in contextual sentiment classification but also in supporting the development of fine-grained lexical resources tailored to central bank communication analysis.

Confusion Matrix of FinBERT on test sample from the extended English database Figure 2



Source: own computations

These results (Figure 2) indicate that FinBERT is probably the most effective choice for financial text analysis, capturing nuanced policy sentiment and providing a robust tool for assessing central bank communication. We used the fine-tuned on monetary policy statements FinBERT model as a starting point for the list of words in the specialized proposed dictionary, however, a tool like this can be very useful in a central bank, as it can analyse sentiment trends over time, assess the impact of policy decisions on financial markets, extract key phrases and topics for explainability, identify comparable past policy statements, classify monetary policies based on sentiment, and even predict market reactions to central bank communications.

3. Romanian Financial News Articles Database

Crafting NLP Tools for Romanian Financial News

The National Bank of Romania has an archive of news starting from 1998, manually collected and selected by employees in the Communication Department, making the text relevant and eliminating the need to filter out unrelated news. The older news, scanned from physical newspapers, requires optical character recognition (OCR). In total, there are 27 years of news, amounting to more than 4,000 PDFs, each containing approximately 30-40 daily news articles. Each day reflects the news from the previous working day. Due to time constraints and limitations in the length of this paper, we will focus on the last 7 years: 2018-2024.

In the preprocessing of Romanian text mainly with NLTK by Loper (2002), several systematic steps were undertaken to ensure the data was appropriately cleaned and prepared for further analysis. The process began with the removal of any watermarks from the text and was followed by the normalization of specific Romanian characters to their correct forms: for example, characters like 'ş' with cedilla were replaced with 'ș' with comma, and 'ţ' with cedilla with 'ț' with comma, ensuring consistency in the representation of Romanian diacritical marks[2]. Next, whitespace normalization was applied to the text, reducing any irregular spacing to single spaces, thereby creating a uniform text structure. Punctuation was then systematically removed using Python's string translation method, which helped in simplifying the text by eliminating non-alphabetic characters that could interfere with tokenization and subsequent analysis.

Following the removal of punctuation, any numerical digits present in the text were filtered out. This step was crucial to focus the analysis on the linguistic content of the text, excluding numeric data which was not relevant for the sentiment and topic modelling tasks at hand. The cleaned text was then processed through Stanza's Romanian language pipeline by Qi et al. (2020), which included tokenization, part-of-speech tagging, and lemmatization. During this stage, the text was converted to lowercase to maintain consistency and reduce redundancy. Lemmatization was employed to reduce words to their base dictionary form, therefore normalizing

[2] There is no uniform approach in the news archive, even though in 2003, a scientific council of the Institute of Linguistics of the Romanian Academy determined that the correct sign is the comma, not the cedilla.

different inflected forms of words to a common base. We used spaCy's NER model to identify named entities labelled as *PERSON* from the Romanian text corpus.

Simultaneously, stop words—common words in Romanian that do not carry significant meaning—were removed from the text. This was done to ensure that the analysis focused on the more meaningful words that contribute to the overall sentiment and thematic content of the text. Additionally, only alphabetic tokens were retained, with non-alphabetic tokens being discarded to further refine the dataset. The resulting tokens, which were now cleaned, lemmatized, and stripped of irrelevant content, were then stored for subsequent analysis. This comprehensive preprocessing ensured that the Romanian text data was in an optimal state for accurate and meaningful linguistic analysis, sentiment scoring, and topic modelling.

One of the most useful NLP tools is topic modelling which can be applied to news such as in our case, but also central bank reports, economic reviews and monetary policy statements to identify key topics and main themes discussed by the policymakers. This tool can help identify emerging trends, priorities and risks in monetary policy. Using this topic modelling analysis, we can track main trends in the news and this type of list can be useful for decision makers at the central bank level.

We constructed a dictionary from the pre-processed Romanian text data using the Gensim library, which systematically mapped each unique term in the corpus to a specific integer ID. By doing so, we established a comprehensive lexicon that enabled us to translate the textual content into a numerical format suitable for computational analysis. To refine this dictionary, we implemented a filtering process that excluded words appearing in fewer than 20 documents and those present in more than 50% of the documents. This step ensured that our analysis focused on terms that were neither too rare nor overly common, thereby raising the model's ability to discover meaningful topics within the text.

Following the dictionary construction and filtering, we generated a Bag-of-Words (BoW) representation for each document. This representation provided a numerical summary of the documents, capturing the frequency of each term in the text as per the dictionary's mappings. Using this BoW format, we applied Latent Dirichlet Allocation (LDA) according to Blei et al.(2003) to the corpus to identify underlying topics. The LDA model was trained with four topics, and we specified parameters to control the training process, including the number of passes over the data and the update frequency. Upon completion of the training, we extracted and examined the topics, each represented by the most significant words associated with it. Additionally, we employed pyLDAvis by Mabey (2018) to visualize these topics, enabling an interactive exploration of the thematic structures present within the documents. This visualization provided a deeper understanding of the distribution and significance of each topic across the corpus.

We identified the five most prominent topics along with their associated keywords, providing a structured overview of the key thematic areas within the corpus for each year (Table 2). In addition to analysing the relative importance of terms within each topic, we also examined the temporal evolution of these themes to assess how their significance has shifted over time (Figure 3). This allowed us to capture emerging trends and changing priorities reflected in the Romanian text data, offering a dynamic perspective on the underlying discourse.

LDA topic analysis – top 10 words per topic for each year

Table 2

	2018	2019	2020	2021	2022	2023	2024
Topic 1- Global Economy, Gold Reserves & Geopolitics	Bancpost, cryptocurrency, UniCredit, Piraeus, Bitcoin, BERD, clause, card, OTP, insolvency	Parliament, Politician, pillar, deputy, greed, IRCC, political-party, treasury, hearing, adviser	Growth, QE, advisor, monetary policy, anticipation, capitalism inflation, monetary	CSALB, ton, insolvency, Treasury, Russian, laundering, grain, governance, corporate, circumstantial	Ruble, grain, ton, oil, Senate, gas, tanker, metal, Gazprom, federation	Minister, political-party, Fidelis, BT, exemption, SME, polish, Israel, OUG	Tezaur, Trump, corporation, ton, Moscow, election, trillion, CEC, Russian, candidate
Topic 2 - Interest Rates, Inflation & Housing Market	Greed, OUG, card, Turkey, quarterly, gov order, revision, normative, pillar, pound	MF, prim-minister, revision, execution, liability, RONEUR, pillar, president, correction	Governor, CSALB, ASF, pillar, file, president, Bitcoin, S&P, polish, reconcile	Dividends, UniCredit, anticipation, adviser, aggregated, absorption, residential, capping, growth	Schengen, admission, Austrian, ASF, pillar, Erste, corporation, SME, FTX, administrator	Suisse, Swiss, UBS, Silicon-Valley, SVB, First Bank, Switzerland, depositor, cryptocurrency	IRCC, governance, capping, instant, ton, poverty, CEC, margin, expensive, mortgage
Topic 3 – Crypto currency, Financial Markets & Digital Assets	Convergence, Italian, poverty, inhabitant, intermediation, circulation, Brussels, insolvency, union, Deutsche	Prim-minister, Chinese, EximBank, RONEUR, ton, polish, Libra, crypto, inhabitant, poverty	UniCredit, user, term, grant, Revolut, FNGCIMM, unemployed, Bitcoin, telephone, authentication	City-Insurance, ASF, court, political-party, RONEUR, blockchain, constitutional, insolvency, deputy	Lombard, HICP, interest rate, Schengen, associate, anticipated, inflation, decrease	ANPC, OTP, Revolut, Tezaur, Intesa, inclusion, ARB, Moscow, SanPaolo, CFA	Trump, Bitcoin, Revolut, flat, cryptocurrency, Fitch, elections, gov decision, amnesty, Alpha
Topic 4 - Social & Economic Trends, Policy & Regulations	Pillar, Parliament, capping, president, political-party, Senate, prim-minister, Bancpost, deputy, saving	Alpha, OTP, adviser, Lagarde, convergence, Christine, deputy-governor, revision, profitable, wealth	Vaccine, pandemic, trillion, S&P, lockdown, associate, provision, rescheduling, Biden, mobility	Ton, OTP, Erdogan, pound, inequality, wheat, wealth, privatization, Turkey, MF	CSALB, revision, postponing, non-reimbursable, execution, MF, court, taxation, duties, suspension	Bitcoin, interbank, lombard, insolvency, crypto currency, decrease, trillion, non-performing, ton, associate	Spontaneous, tradition, correction, abstract, freedom, conduct, decrease, knowledge, western, OTP
Topic 5 - Legal & Financial Disputes, Banking & Consumer Protection	Food, OTP, ASF, repo, vegetables, non-food, expensive, abusive, facility, fuel	CSALB, execution, debt, clause, resolution, unpredictable, abusive, litigation, S&P, forced	Duties, RONEUR, grant, Swiss, fuel, taxpayer, Erste, CFA, EximBank, ton	Banknote, coal, revision, taxpayer, compensation, ANAF, HICP, manufacturing, capping, OPEC	Revolut, user, instant, respondent, crypto, corporation, Swiss, Suisse, Tezaur, pound	CSALB, SIF, Alpha, pillar, privatization, ASF, court, resolution, litigation, solution	CSALB, Moldova, court, resolution, Republican, litigation, Intesa, Alpha, farmer, SanPaolo

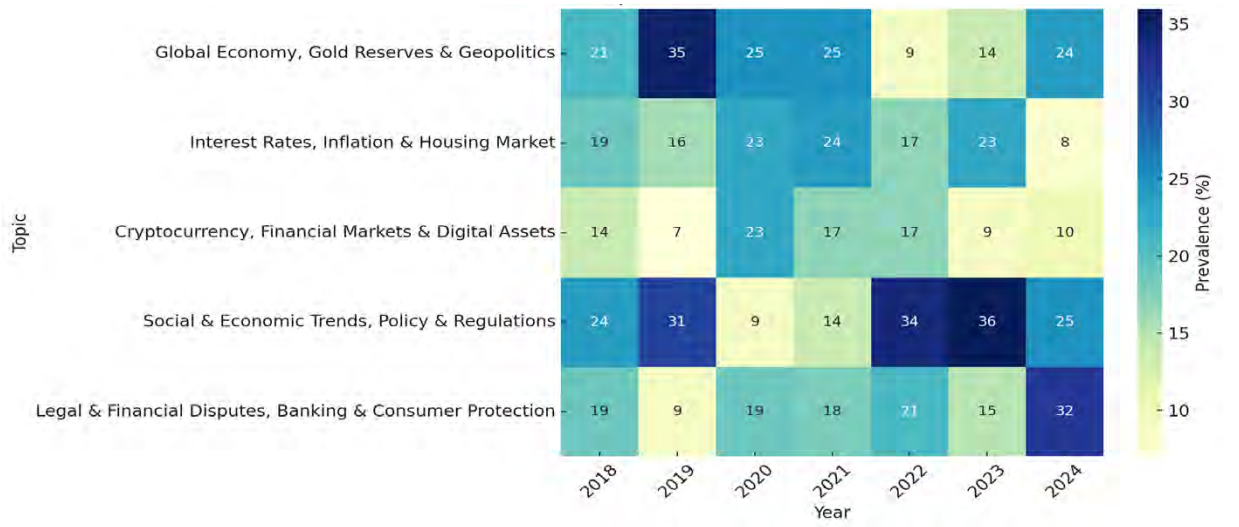
[1] Mentions of individuals were automatically identified using Named Entity Recognition (NER) with the spaCy language model. Entities labelled as PERSON were later replaced manually with institutional roles to protect personal data.

[2] The words were translated from Romanian to English for reader's understanding

Source: created by the authors

Heatmap of the Importance of the Main Five Topics Over Timef

Figure 3



Source: own computations

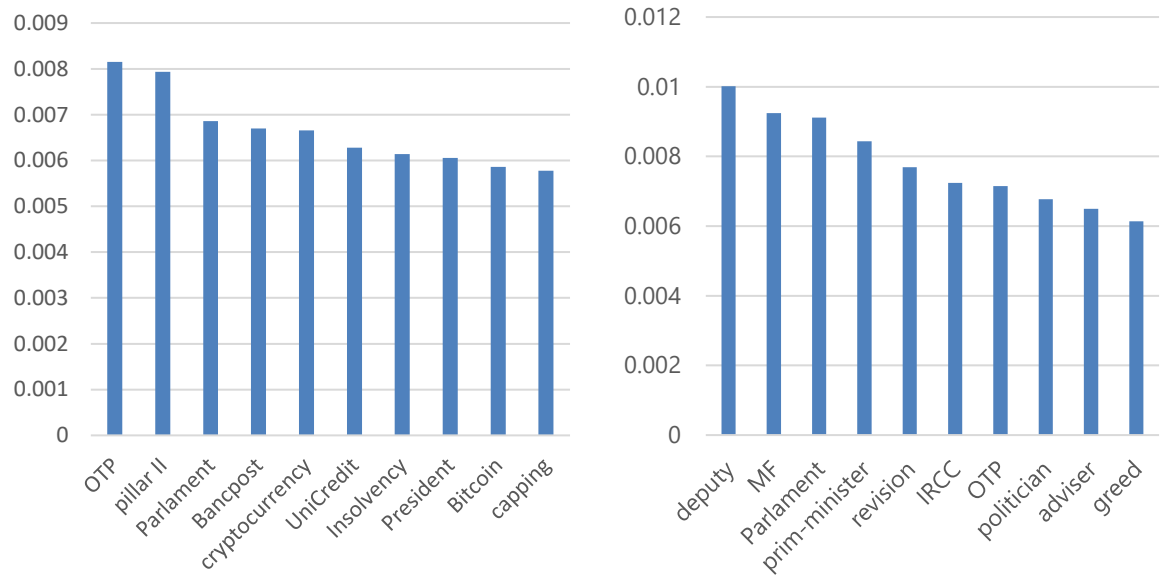
After constructing the LDA model, we aggregate word frequencies across all topics to create a comprehensive view of the most significant terms within the corpus. In LDA, each word receives a weight within each topic, reflecting its probability of occurrence given that topic. To assess the most thematically significant terms for the corpus, we computed a cumulative score for each word by summing its weights across all topics. Formally, for each word *w*, its overall importance is calculated as:

$$LDA\ Weighted_Score(w) = \sum_{t=1}^{T} \phi_t(w)$$

Where $\phi_t(w)$ denotes the probability of word *w* in topic *t*, and *T* represents the total number of topics (in our case, T=5). This aggregated score captures the extent to which a word is representative across the entire thematic structure of the corpus, rather than within a single topic alone. We then ranked all terms according to their LDA-weighted scores and extracted the top 10 words per year. These words are visualized and reported as key indicators of the dominant semantic patterns found in the annual financial news review.

An analysis of the top 10 words aggregated across all five LDA topics for the year 2018 points out that the year (Figure 4) was marked by the wide debate on the "tax on greed," which aimed to impose a significant tax on bank profits linked to the level of interest rates, sparking concerns about financial sector stability and economic impact. Additionally, in 2018, Romania experienced some discussions surrounding the country's second-pillar pension system reform, which is a mandatory private pension fund scheme, and Bancpost, one of Romania's major banks, was acquired by Banca Transilvania, marking a significant consolidation in the Romanian banking sector.

Top 10 words based on LDA topic-weighted scores (Year 2018/ 2019) Figure 4/Figure 5



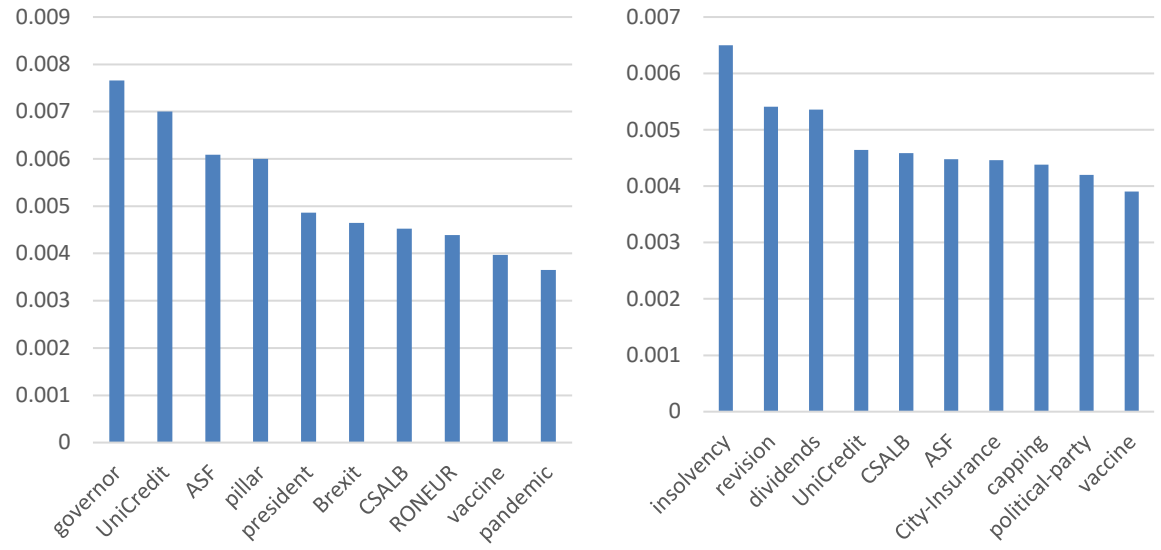
[1] Note: the words were translated from Romanian to English

Source: created by the authors

In 2019 (Figure 5), political discourse was dominated by themes of political instability and leadership challenges, largely shaped by electoral dynamics. Legal controversies involving prominent political figures influenced governance and public trust, while opposition efforts focused on contesting fiscal policies, particularly in the context of budgetary decisions that raised concerns over public finances. At the same time, legislative initiatives aimed at consumer protection in financial services gained traction, reflecting broader discussions on financial stability and regulatory oversight. Overall, the year was marked by intense political debates and policy disputes, highlighting the intersection of governance, economic decision-making and public confidence.

In 2020 (Figure 6), Romania tackled the COVID-19 pandemic with various public health measures and economic interventions, while the National Bank of Romania implemented monetary policies to stabilize the economy amidst the global crisis. Simultaneously, the country navigated the impacts of Brexit, while Revolut expanded its services in Romania, adapting to the changing financial landscape. In 2021 (Figure 7), Romania faced rising inflation, prompting companies to reassess dividend distributions, while the CSALB (Center for Alternative Dispute Resolution in Banking) continued to mediate disputes between consumers and banks. Additionally, there was growing interest in blockchain technology that can be attributed to the significant increase in valuation of the cryptocurrency sector, with increased discussions around its potential impact on the financial sector.

Top 10 words based on LDA topic-weighted scores (Year 2020/ 2021) Figure 6/Figure 7



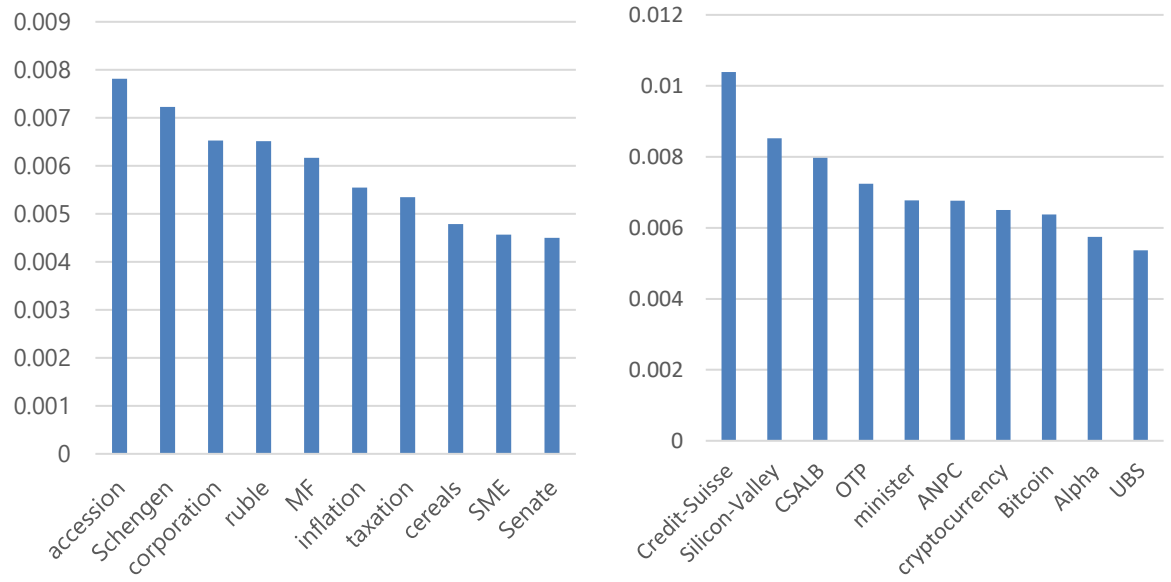
Source: created by the authors

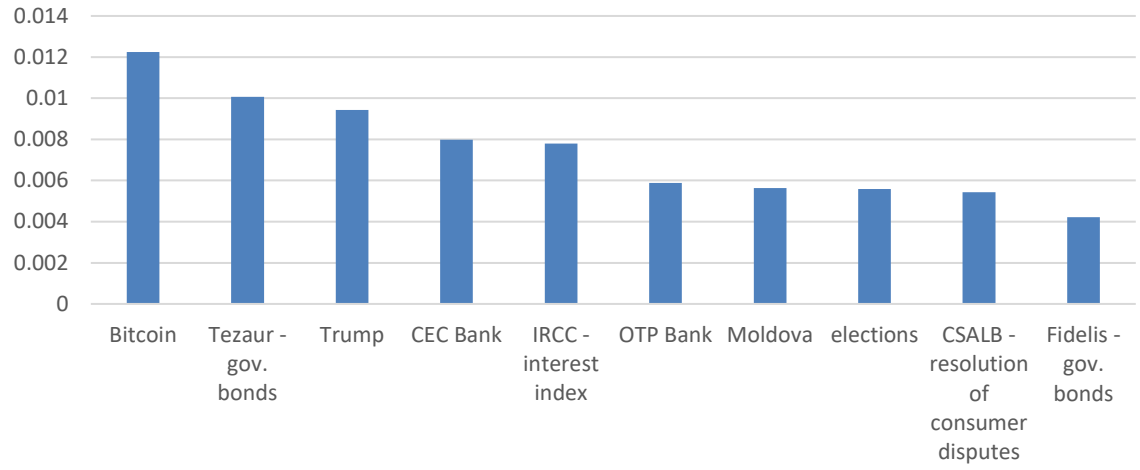
In 2022 (Figure 8), Romania's efforts to join the Schengen Area faced ongoing delays, while the depreciation of the Russian ruble amid global economic instability influenced regional trade dynamics. Additionally, Romania grappled with rising inflation, which impacted purchasing power and economic stability. In 2023 (Figure 9), the turmoil surrounding Credit Suisse's acquisition by UBS and the collapse of Silicon Valley Bank had ripple effects in Romania, contributing to increased market volatility and heightened concerns among investors and financial institutions. The events underscored the interconnectedness of global financial systems, leading to cautious responses from Romanian banks and regulators to ensure stability in the face of external shocks.

In 2024 (Figure 10), the key themes emerging from unsupervised analysis of public discourse in Romania highlight the prominence of electoral events, financial sector developments, and digital asset trends. The U.S. presidential election and Romania’s local elections dominated public interest, reflecting the growing impact of political dynamics on economic expectations. At the same time, the rise of cryptocurrencies, particularly Bitcoin, remained a focal point, driven by increased institutional adoption and regulatory discussions. In the financial sector, Revolut’s decision to establish a local branch in Romania marked a significant milestone in the country's fintech evolution, enhancing competition in digital banking. Additionally, UniCredit's acquisition of Alpha Bank signalled further consolidation in the banking industry, potentially reshaping market dynamics and credit accessibility. These trends underscore Romania’s integration into global financial and technological shifts while navigating the challenges of an election year.

Top 10 words based on LDA topic-weighted scores (Year 2022/2023)

Figure 8/Figure 9/ Figure 1010





[1] Additional information was added for reader's better understanding

Source: created by the authors

Building the Romanian Financial Stability Dictionary

There are two primary approaches for sentiment analysis that do not use LLMs: supervised learning, which requires many labelled documents, and rule-based, which relies on a previously built dictionary of positive and negative words that can be adapted to the subject of the documents. In our case, we do not have readily available labels and creating them would require an immense effort to go through the daily text due to its sheer size and length, which would also be subjective to the reader to some extent. The best approach, given the limitations, was to build a financial stability

dictionary for the Romanian language starting English words identified by FinBERT and fine-tuned with expert-level knowledge in Romanian.

General tools included the lexicons by Hu and Liu (2004), Baccianella et al. (2010) with SentiWordNet 3.0, and VADER by Hutto and Gilbert (2014), which are tailored to consumer reviews and online content but proved inadequate for technical financial texts. Financial-specific dictionaries such as Loughran and McDonald (2011), designed for corporate filings, and the IMF Financial Stability Dictionary by Correa et al. (2017), better captured sentiment but still showed limitations. While the IMF dictionary yielded the most relevant results, its effectiveness was constrained by language, as NBR communications are originally in Romanian. Consequently, the study highlights the need to develop a domain-specific sentiment dictionary in Romanian to improve analysis accuracy in the context of central banking and financial news.

Our positive and negative list of words was draft from FinBERT results in English and was also inspired by the IMF dictionary, the best English option for monetary policy statements. However, the Romanian language has many particularities, so simply machine translating from English would result in poor-quality outcomes. For that reason, we manually translated the terms and ended up with a list of Romanian words that did not perfectly match the English ones therefore we expert added specific Romanian terms. We also investigated RoWordNet by Dumitrescu et al. (2018) to find all possible synonyms and antonyms for our Romanian words on the existing list.

The resulted Romanian Dictionary focuses on Financial Stability issues and can be used to measure the sentiment from the Romanian news database. It contains 150 positive words in lemma form[3] and 182 negative terms in lemma form. We had the list assessed by Gen-AI model (ChatGPT) who was prompt as a Romanian financial stability expert and had to score positive and negative terms on an intensity scale: -0.25 (ambiguous), -0.5 (quite negative), -0.75 (clear negative), -1 (strong negative), and 0.25, 0.5, 0.75, to 1 for positive.

The formula to evaluate the sentiment index is simple and relies on the dictionary with polarities (positive/negative) and their intensity score levels (weights) for each lemma.

For any day and document (t) the simple sentiment index is computed (Figure 11):

$$Sentiment\ Index = \frac{P(t)-N(T)}{P(t)+N(t)} \quad (1)$$

Where:

- P(t) is the list of positive lemmas in the text t that matches the positive words in the dictionary
- N(t) is the list of negative lemmas in the text t that matches the negative words in the dictionary

[3] A lemma is the base or dictionary form of a word, used in linguistic processing to group inflected forms (e.g., "running," "ran," and "runs" all share the lemma "run").

For any day and document (t) the sentiment index using scores (Figure 15):

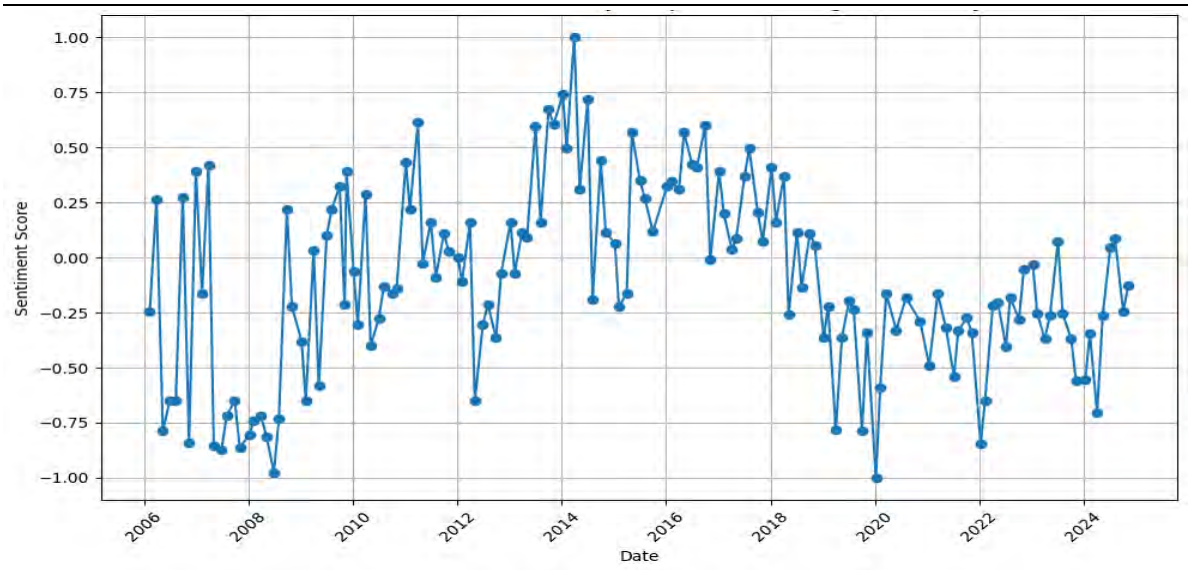
$$Sentiment\ Score\ Index = \frac{\sum_{i=1}^{P(t)} S_p(i) + \sum_{j=1}^{N(t)} S_n(j)}{\sum_{i=1}^{P(t)} S_p(i) + \sum_{j=1}^{N(t)} |S_n(j)|} \quad (2)$$

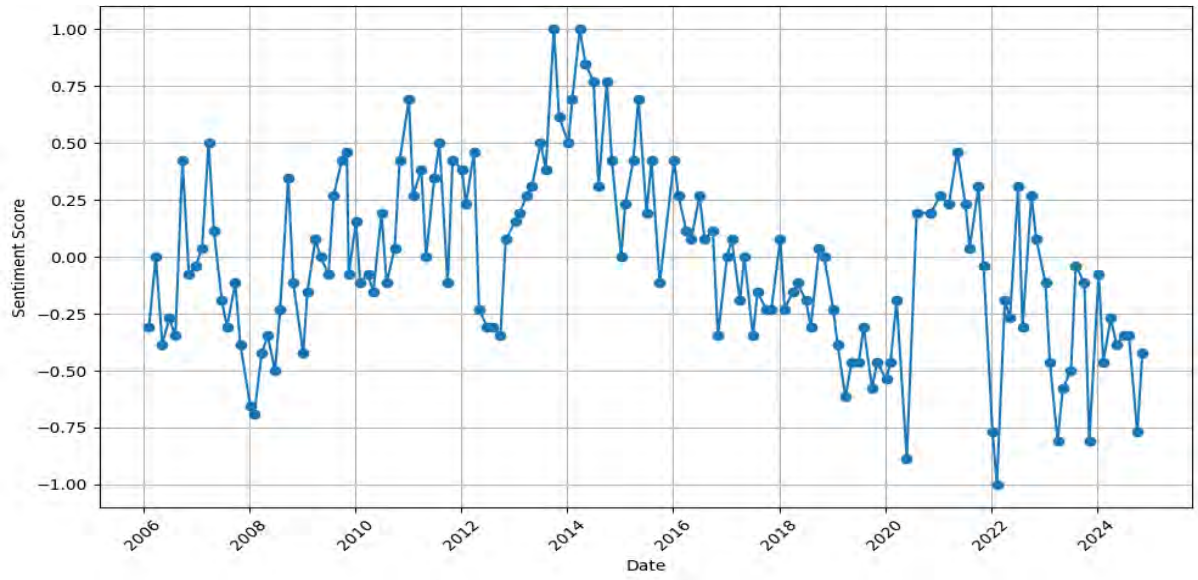
Where:

- P(t) is the list of positive lemmas in the text t that matches the positive words in the dictionary
- N(t) is the list of negative lemmas in the text t that matches the negative words in the dictionary
- $S_p(i)$ the score (intensity) of the i-th positive lemma in P(t) according to the positive words dictionary
- $S_n(i)$ the score (intensity) of the j-th positive lemma in N(t) according to the negative words dictionary

Sentiment Index of Monetary Policy using Romanian Financial Stability Dictionary (without and with scores)

Figure 11/ Figure 12





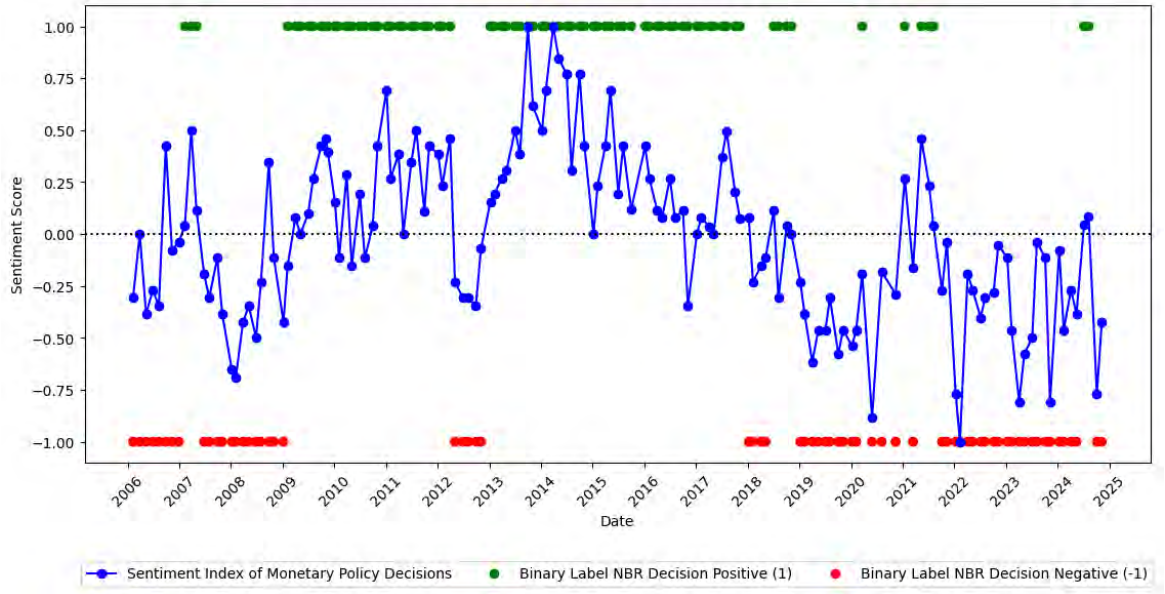
Source: authors' own computations

By leveraging the Romanian Financial Stability Dictionary, we managed to correctly identify trends in monetary policy that the unspecialized method failed to identify, such as the global financial crisis of 2007-2008, the 2012 NBR intervention of raising the monetary policy rate due to rising inflationary pressures related to the Eurozone crisis, the 2014 peak of positive sentiment driven by strong economic growth that boosted consumer confidence, the 2019 drop due to the proposed 'greed tax' for on banks and energy companies, as well as the polarized presidential elections and the 2022 exponential rise in inflation that led to 8 negative interventions by the NBR, etc.

We tested our results against binary labels that we've built starting from the movement of the monetary policy interest rate (series from Refinitiv): positive sentiment if the monetary policy rate was lowered in that meeting, and negative if the monetary policy rate was raised. All neutral decisions, meaning meetings where the monetary policy rate was maintained, were labelled with the same sentiment as the last non-null decision movement; therefore, we considered that the sentiment was maintained until further notice.

The version of the sentiment index with scores outperforms the simple sentiment index by 2% (78% versus 76%), and the final index that we obtained by mixing the two methods matches 92.5% of the labels. This performance is considered excellent, given that neutral decisions were made using the last decision rule and not manually evaluated by a human (Figure 13).

Sentiment Index of Monetary Policy versus Binary NBR Decision Labels Figure 13



Source: authors' own computations

To assess the performance of our NLP analysis, we examined the correlation between the sentiment index of monetary policy decisions and the most relevant economic variable: the inflation rate, since controlling inflation is the primary goal of monetary policy. We found an expected negative correlation between our computed sentiment and the all-items HICP (Harmonized Index of Consumer Prices) of over 40%, while the benchmark correlation between the monetary policy interest rate and HICP was -60%.

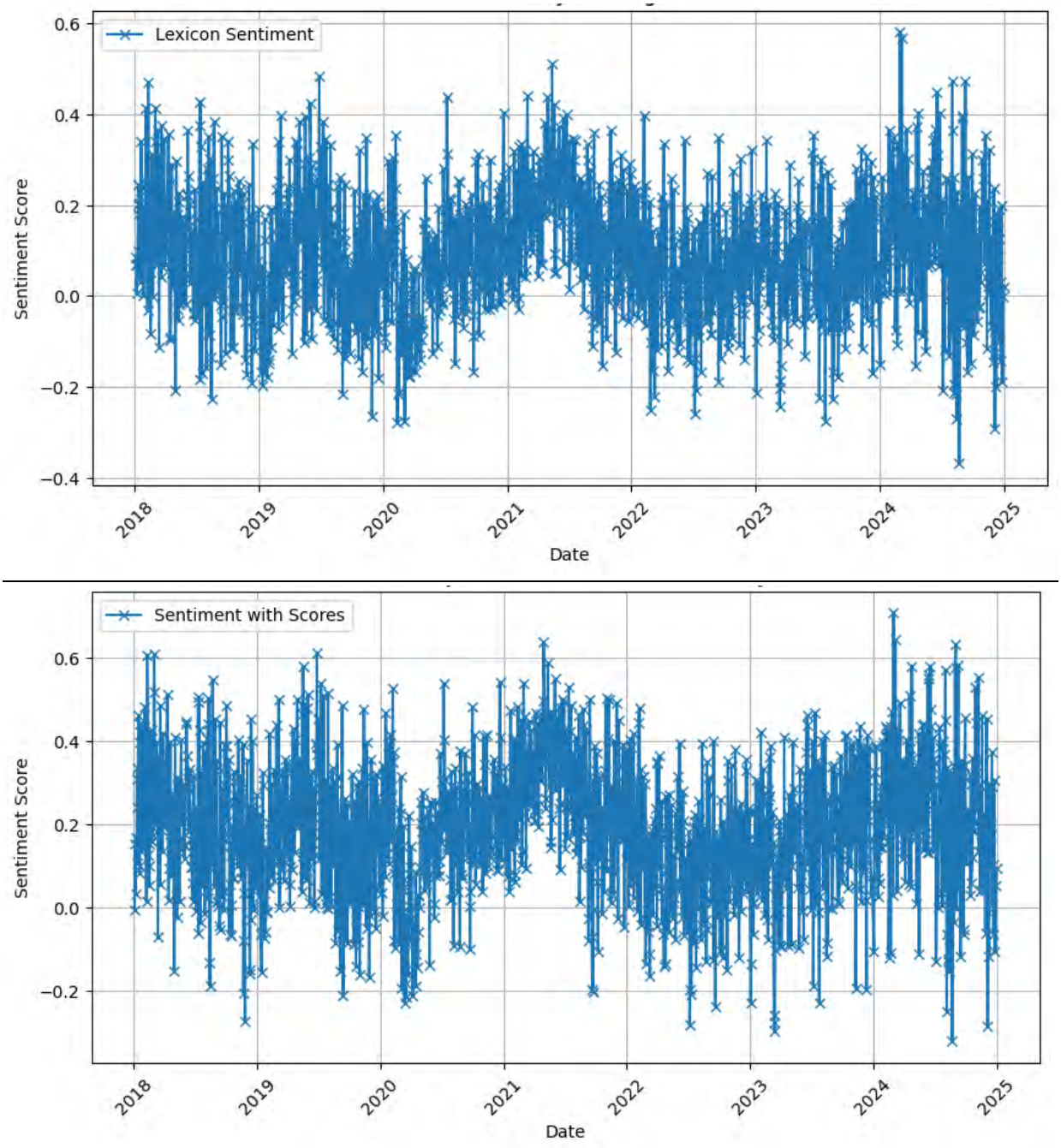
Extracting Sentiment from Financial News Articles

We used the tailor-made dictionary to compute the sentiment in daily financial news using the simple method of computing positive/negative lemmas (Figure 14) and the method that also uses scores (Figure 15).

Daily Sentiment of News using Romanian Financial Stability Dictionary

(Without and with scores)

Figure 14/Figure 15



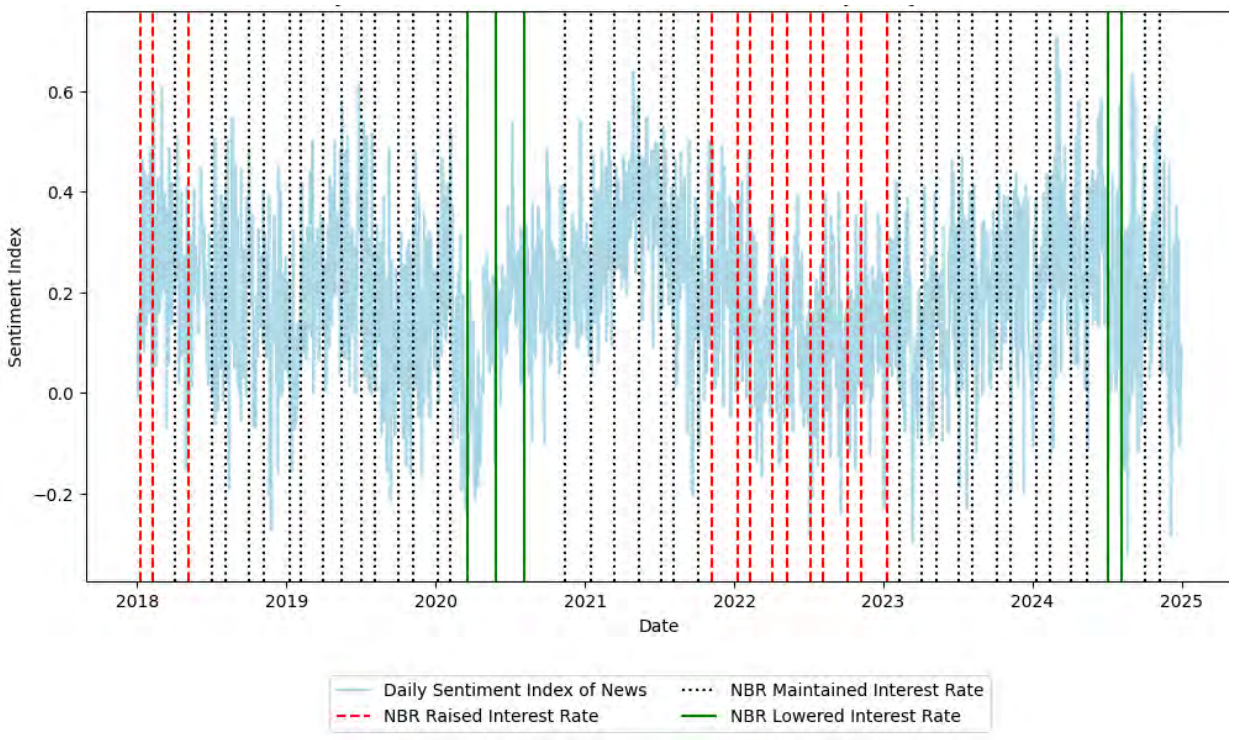
Source: authors' own computations

Sentiment analysis of economic and political discourse in Romania throughout the most recent year, 2024, reveals a predominantly positive outlook in the first half of the year, with sentiment scores remaining above the neutral threshold. However, the second half of the year saw a shift towards negative sentiment, driven by key adverse events. Notably, the annulment of the elections on December 9, 2024,

introduced significant political uncertainty, while Fitch's decision to assign a negative outlook to Romania at year-end further weighed on investor confidence. Additionally, the decline in industrial production in August signalled potential economic headwinds, contributing to a deteriorating sentiment. These developments collectively shaped an overall negative sentiment trajectory for the year, highlighting the sensitivity of economic perceptions to both political instability and macroeconomic indicators.

We compared the daily sentiment of financial news with the moments of monetary policy decisions (Figure 16) and showed how the positive sentiment of a monetary policy decision, such as a lowering of the reference interest rate, positively influences the economy. There was a rise in sentiment in mid-2020 despite the COVID-19 pandemic. The opposite pattern can be identified in early 2018 and especially in 2022, where there were 8 negative monetary policy decisions, meaning the NBR raised the interest rate to combat inflation pressures. In just one year, the reference monetary policy rate rose by almost 5 percentage points due to the need to counteract the international inflationary environment. We concluded that there is strong evidence that the press responds to monetary policy decisions and that the communication of the central bank is efficient in influencing the markets.

Daily Sentiment of Financial News and Monetary Policy Decisions — Figure 16



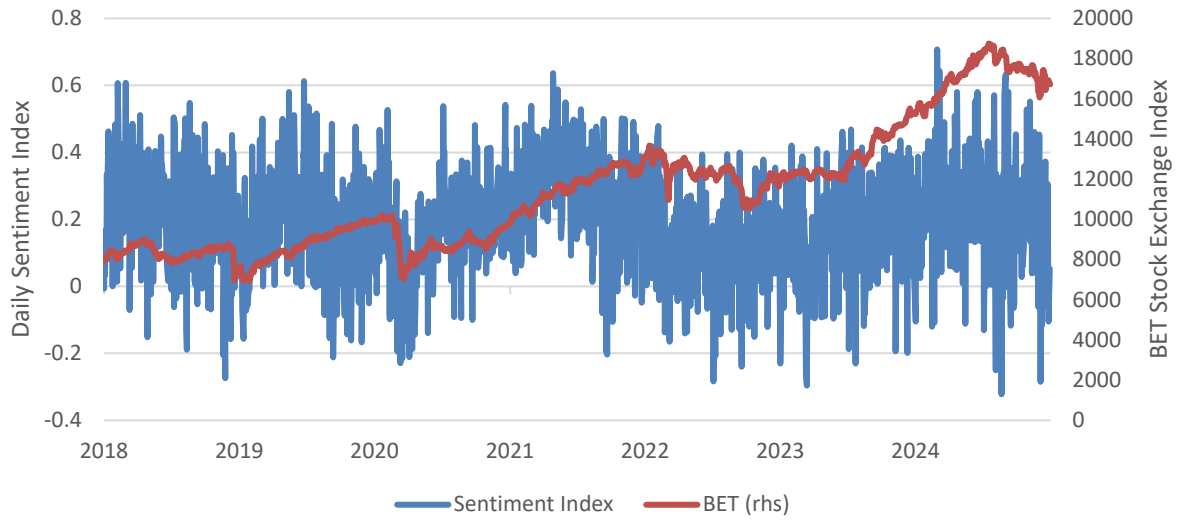
Source: authors' own computations

In conclusion, there is great value in building our own Romanian Dictionary for Financial Stability issues, and we achieved that by combining terms translated from FinBERT in English with manually identified Romanian terms. We computed scores for special terms related to financial stability using Gen-AI. To assess the performance of our daily sentiment indicator, we developed an unsupervised K-Means++ analysis

according to Kanungo (2002) on embeddings using the BERT multilingual model by Devlin (2018) after reducing their size using PCA. The results show a 93% overlap between the clusters and the binary sentiments from the simple index computed by counting lemmas, while a slightly lower 96% overlap was observed between the clusters and the binary sentiments from the index computed also using scores.

Daily Sentiment of News and Stock Market Index

Figure 17



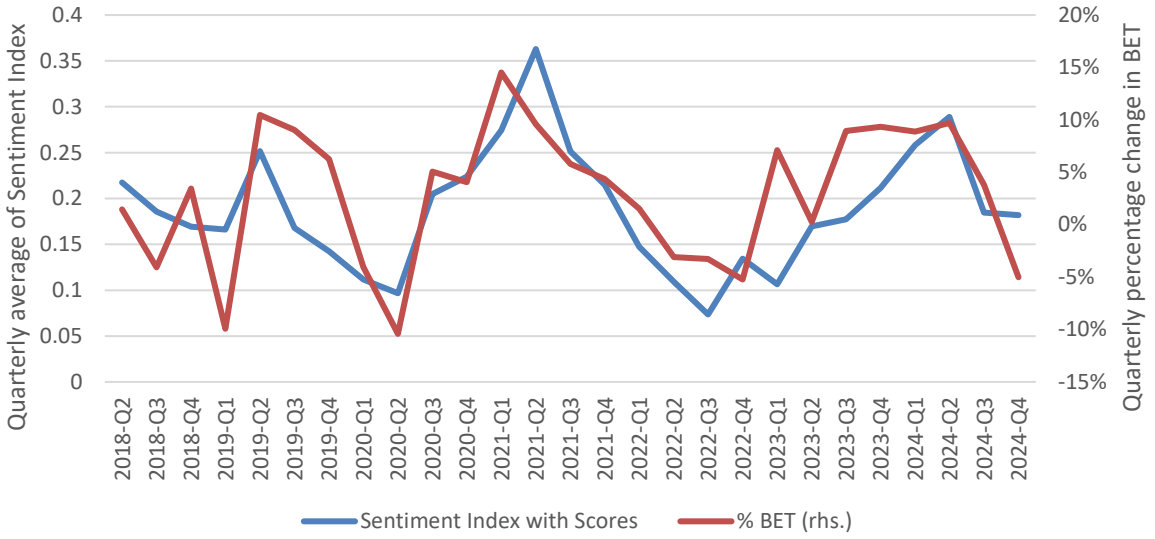
Source: authors' own computations

There is a positive correlation (Figure 17) between our daily sentiment index of financial news and Romanian Stock Exchange Index (BET), but at the quarterly level, the interconnectivity between the two series is most evident, with a correlation degree of 62% (Figure 18).

A negative correlation can be identified (Figure 19) between our sentiment index of financial news and the short-term yields of Romanian government bonds (1 year maturity - indicate the cost of state financing and also reflect inflation pressure), as well as the overnight and 3-month interest rate benchmarks known as the Romanian Interbank Offered Rate (ROBOR3M - part of the variable interest rate households pay for their loans).

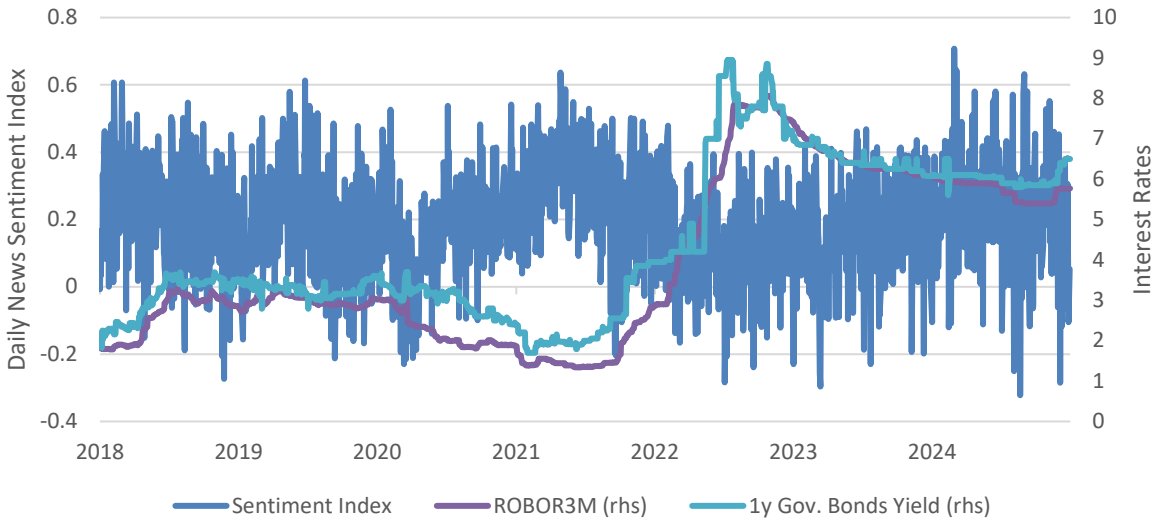
Quarterly Sentiment of News and Percentage Change in Stock Market Index

Figure 18



Daily Sentiment of News, Interbank Interest Rates and Yield Rates

Figure 19



Source: authors' own computations

4. Conclusions

In conclusion, this paper proposes stepping aside from classical numerical, tabular data and investing more time and effort in the use of nontraditional data, not just time series, therefore implementing Natural Language Processing methods at a large scale for some types of text that exist in a central bank. The primary goal is to provide the National Bank of Romania with a guideline for pre-processing Romanian text, building NLP tools, and benchmarking sentiment analysis. Creating a toolbox for

Romanian, when most tools are in English, was challenging due to the need to consider all language particularities.

We believe that text is a rich resource for central banks, and therefore, in this paper, we proposed integrating Natural Language Processing into the National Bank of Romania's set of analyses through a thorough analysis on lexicon-based approaches to sentiment analysis using news and monetary policy decisions. We can only hope that by following these guidelines for text in Romanian, the list of potential use cases for Natural Language Processing will continue to be extended at the National Bank of Romania and that economists will commit to a long-term adoption of text data in their research, as there are many other types of text sources available to them.

The core contribution of this study is the construction of a Romanian Dictionary for Financial Stability, developed by identifying positive and negative terms extracted with FinBERT, fine-tuned on English monetary policy statements. We enriched this dictionary by assigning sentiment intensity scores using a generative AI agent. Our results demonstrate that the tailor-made dictionary outperforms existing solutions in capturing sentiment, with the scoring-enhanced approach yielding superior performance. The monetary policy sentiment index correctly identified sentiment in 92.5% of binary-labelled monetary policy statements. For the unsupervised analysis of daily financial news articles, we applied K-Means++ clustering on multilingual BERT embeddings, achieving overlapping rates of 93% and 96% with our results. Additionally, the daily news sentiment index showed consistent correlations with key macroeconomic variables. Integrating both components of the analysis, we find strong evidence that the press responds to monetary policy decisions and that central bank communication effectively influences financial markets.

5. References

1. Araci, D. (2019). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:1908.10063.
2. Araujo, Douglas KG. 2023. "Gingado: A Machine Learning Library Focused on Economics and Finance." Working Paper 1122. BIS Working Paper. Bank for International Settlements.
3. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
5. Correa, R., Garud, K., Londono-Yarce, J. M., & Mislang, N. (2017). Constructing a dictionary for financial stability.
6. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
7. Edison, H., & Carcel, H. (2021). Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. Applied Economics Letters, 28(1), 38-42.

8. Evdokimova, T., Nagy-Mohacsi, P., Ponomarenko, O., & Ribakova, E. (2023). Central banks and policy communication: How emerging markets have outperformed the Fed and ECB. Peterson Institute for International Economics Working Paper, (23-10).

9. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).

10. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

11. Jolliffe, I. T. (1986). Principal component analysis. Springer Series in Statistics. Springer-Verlag.

12. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.

13. Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. arXiv preprint cs/0205028.

14. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of finance, 66(1), 35-65.

15. Mabey, B. (2018). pyLDAvis documentation.

16. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

17. Richardson, L. (2007). Beautiful soup documentation.

18. Sanh, V. (2019). DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108.

19. Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019, October). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In 2019 sixth international conference on social networks analysis, management and security (SNAMS) (pp. 338-343). IEEE.

20. Vasiliev, Y. (2020). Natural language processing with Python and spaCy: A practical introduction. No Starch Press.

NATIONAL BANK OF ROMANIA

Rome, February 18th, 2025

A NATURAL LANGUAGE PROCESSING TOOLBOX FOR THE NATIONAL BANK OF ROMANIA

Claudia Voicilă, Senior Economist

Note 1: The opinions are those of the author and do not necessarily reflect the views of the National Bank of Romania

Note 2: Part of my dissertation thesis for the Artificial Intelligence Master's program at the Faculty of Mathematics and Informatics, University of Bucharest.

Agenda

Introduction

Chapter 1. English Monetary Policy Statements Database
Chapter 2. Romanian Financial News Articles Database

Conclusions

Abstract

- This paper introduces a specially designed toolbox aimed at leveraging Natural Language Processing (NLP) to unlock insights for the National Bank of Romania (NBR), enhancing economists' capacity to process text data such as financial news and press releases, exploring areas of Financial Stability and Central Bank Communication.
- We propose implementing scalable Natural Language Processing methods within the National Bank of Romania's analysis toolkit, focusing on lexicon-based sentiment analysis of daily news and monetary policy decisions. This study seeks to align the institution with the best practices of other central banks and highlights the untapped potential of textual data as a valuable resource in central banking.

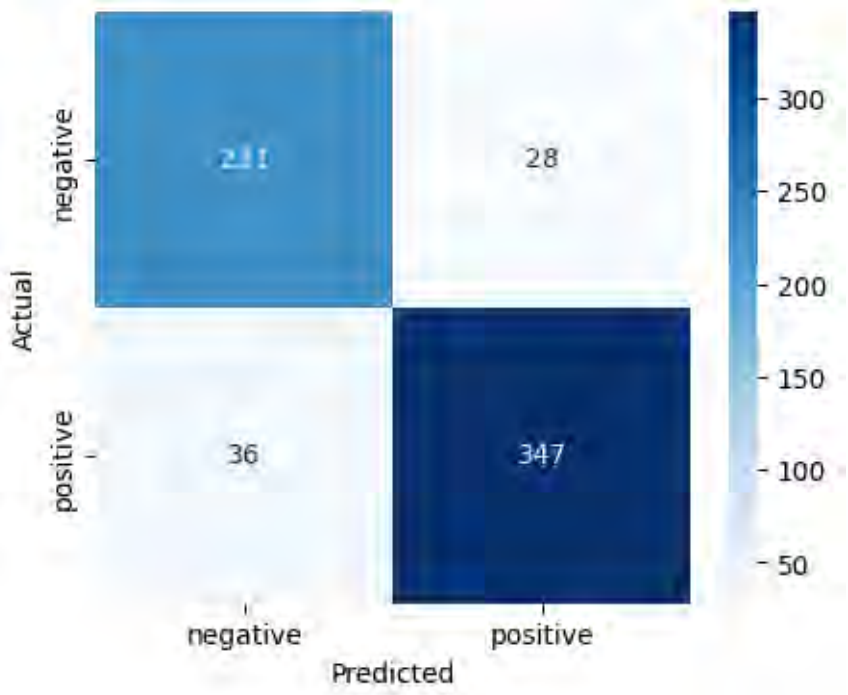
Introduction

- The paper is structured in two chapters: the first one addressed monetary policy statements in English from the National Bank of Romania, the European Central Bank, the United States's Federal Reserves and 25 other emerging market countries, with the scope of building a list of positive and negative words specific for central bank text, the starting point of the Romanian Financial Stability Dictionary.
- The second chapter focus on the Romanian news archive database, and we finalize the Financial Stability Dictionary that contains specialized positive and negative words. We augment the method with GenAI scores to measure text-based sentiments in financial texts.

Chapter 1. The English Monetary Policy Statements Database

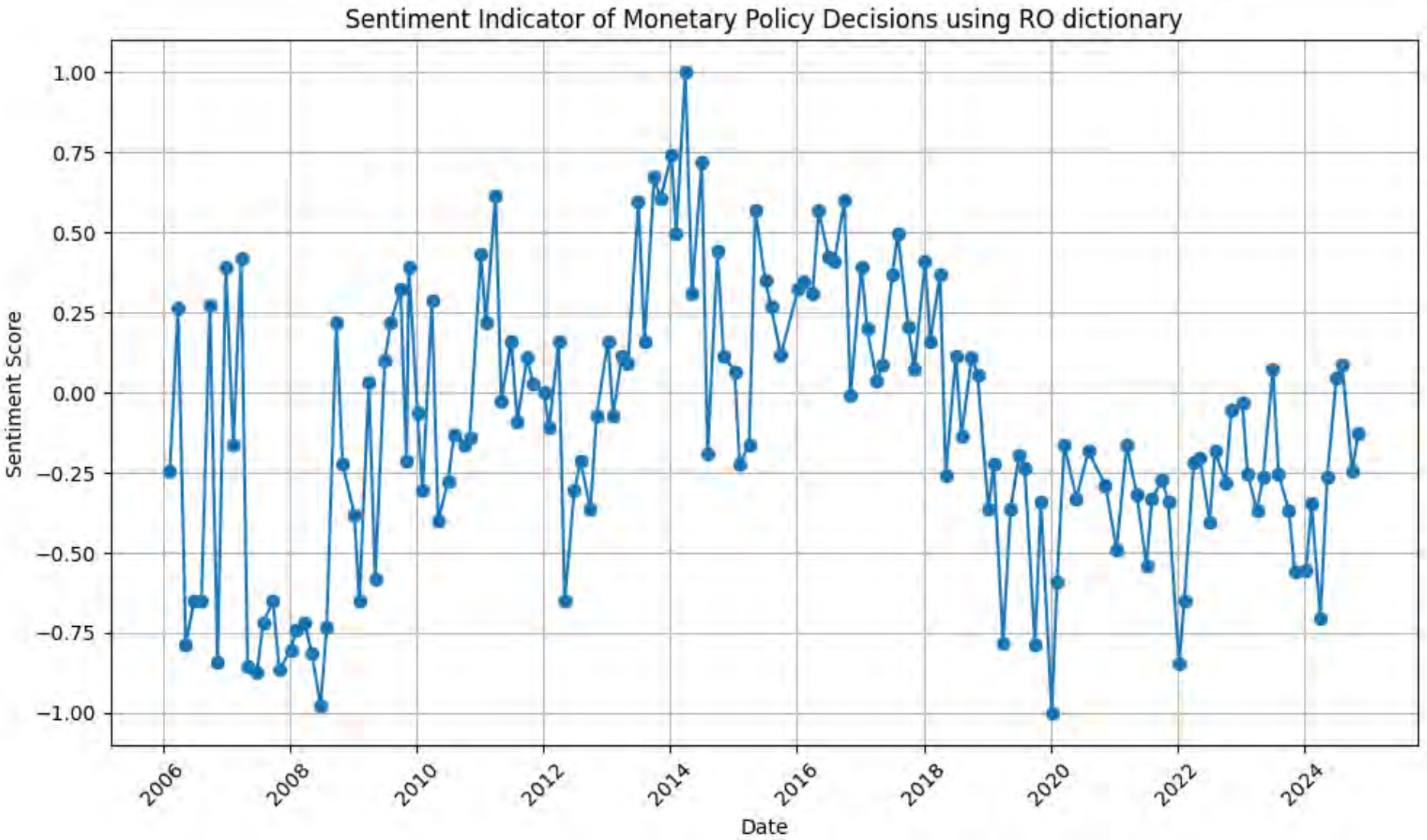
- The first part focuses on English monetary policy statements starting from the moment NBR adopted inflation targeting. We extended the dataset by including ECB, FED, and 25 other emerging market countries.
- We fine-tuned a Transformer model, FinBERT, on the binary sentiment labels we created for the monetary policy statement based on movements in the monetary policy interest rate of each country (positive sentiment if the interest rate is lowered, negative sentiment if the interest rate is raised).
- After trying unspecialized lexicons, we demonstrated that using specialized Financial Stability Dictionary is crucial for constructing a sentiment index for monetary policy, and we measure the polarity of words using GenAI.

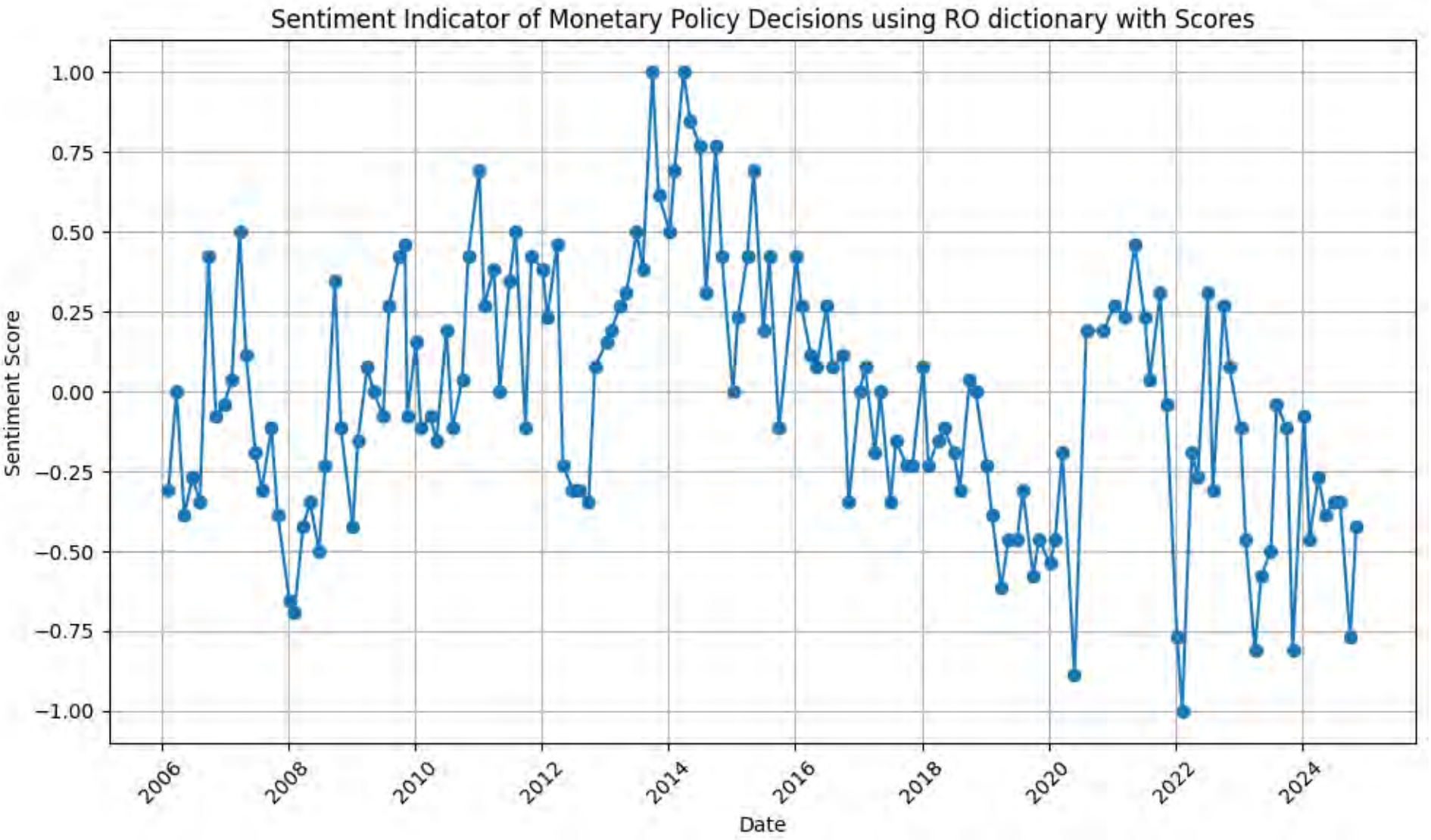
Confusion Matrix of FinBERT on test sample from the extended English database



Classification report on test sample (10 epochs)

	precision	recall	F1-score	support
positive	86.52%	89.19%	87.82%	259
negative	92.53%	90.60%	91.56%	383
accuracy			90.03%	642
macro accuracy	89.53%	89.89%	89.69%	642
Weighted average	90.11%	90.03%	90.05%	642

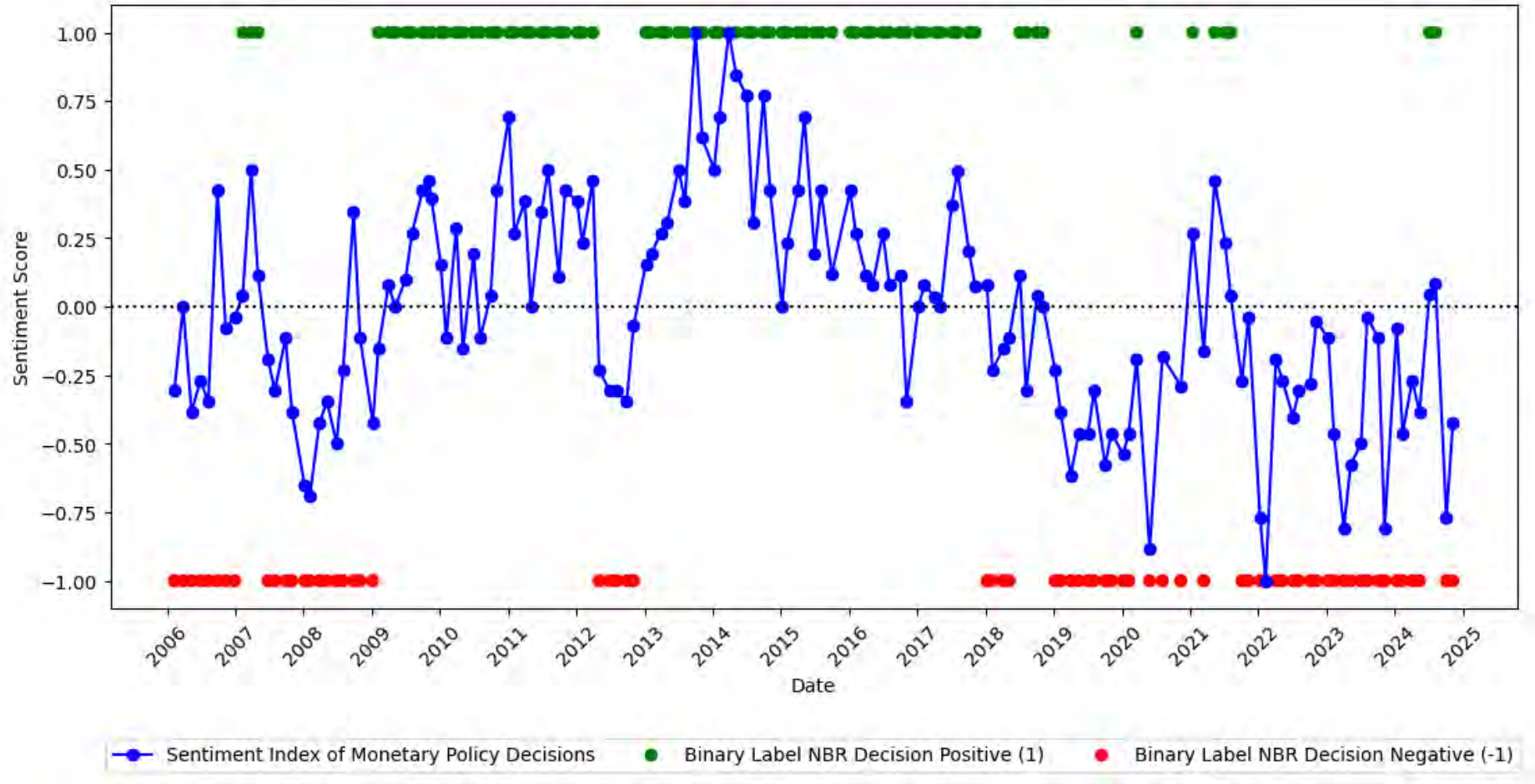

Sentiment Indicator of Monetary Policy Decisions using RO dictionary
Sentiment Score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Date


Sentiment Indicator of Monetary Policy Decisions using RO dictionary with Scores
Sentiment Score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
Date

- The version of the sentiment index with scores outperforms the simple sentiment index by 2% (78% versus 76%), and the final index that we obtained by mixing the two methods matches 92.5% of the labels. This performance is considered excellent, given that the neutral decisions were built using the last decision rule (negative sentiment persists if last decision was to raise the interest rate, positive sentiment persists if the last decision was to lower interest rate) and not manually evaluated by a human.

- To assess the performance of our NLP analysis, we examined the correlation between the sentiment index of monetary policy decisions and the most relevant economic variable: the inflation rate, since controlling inflation is the primary goal of monetary policy. We found an expected negative correlation between our computed sentiment and the all-items HICP (Harmonized Index of Consumer Prices) of over -40%, while the correlation between the monetary policy interest rate and HICP was around -60%.

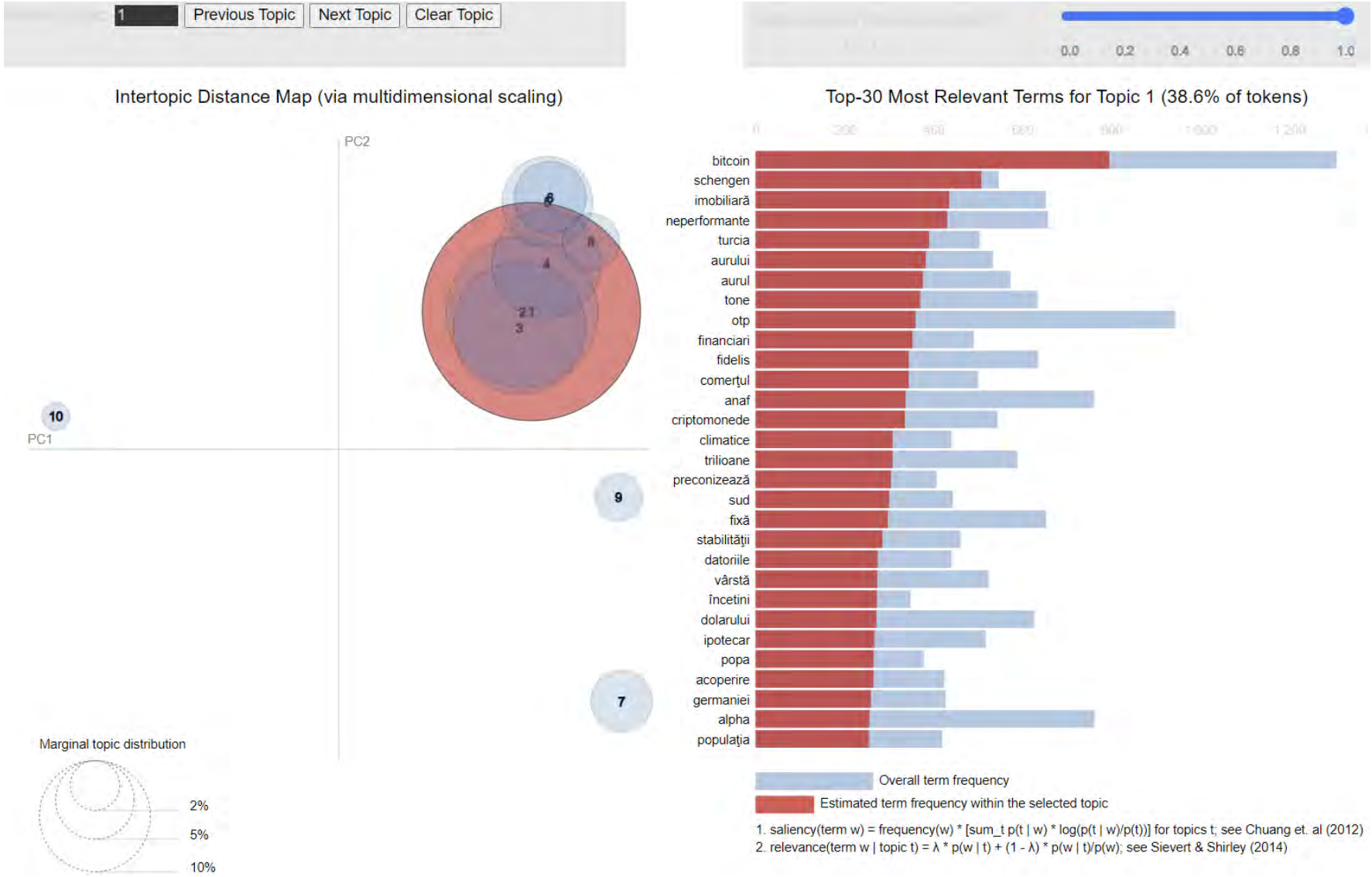
Sentiment Index of Monetary Policy compared to Binary Labels



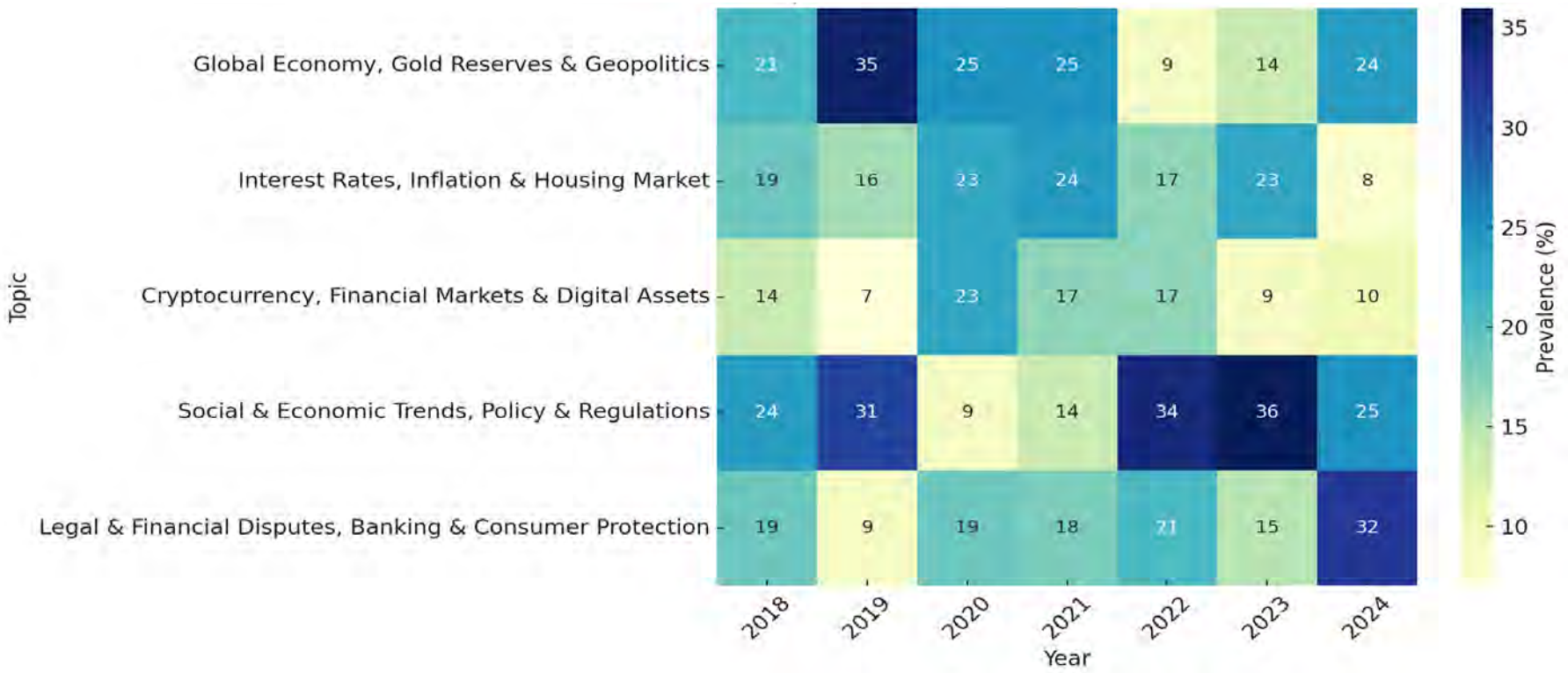
Chapter 2. The Romanian Financial News Articles Database

- The second part addresses the NBR's archive of daily financial news, for which we conduct topic modeling using LDA to track key trends in the news, and propose a tailor-made, Romanian Financial Stability Dictionary to measure sentiments.
- We computed polarity intensity scores by prompting as a financial stability expert a Gen AI agent.
- In absence of labels, we assess the performance by comparing our method with K-Means++ clustering applied to embeddings obtained with multilingual BERT Transformer and reduced with PCA, and by demonstrating correlations with macroeconomic and market variables as well as interactions with monetary policy decisions.

Interactive LDA analysis, Example for year 2023



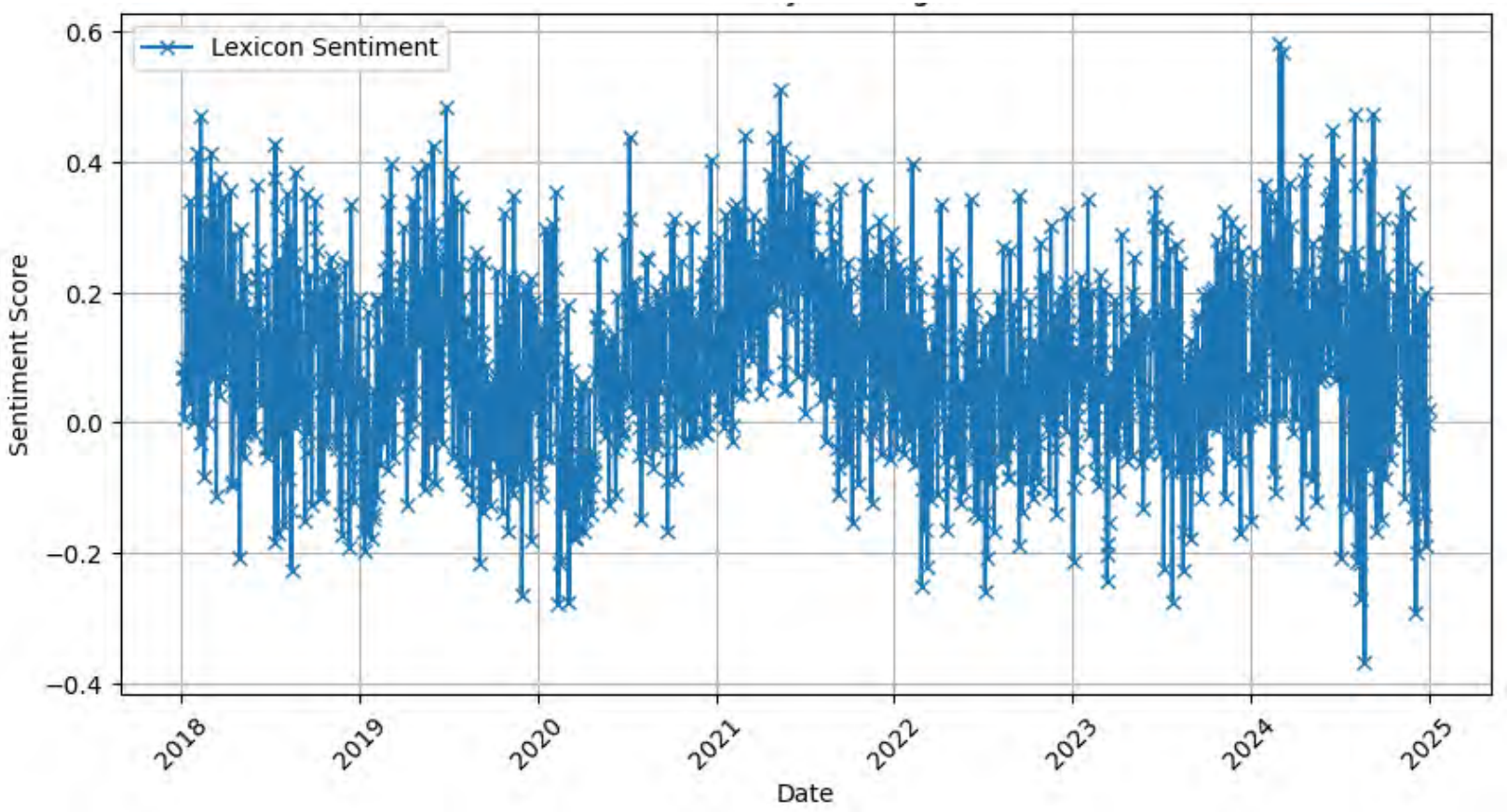
Heatmap of the Importance of the Main Five Topics Over Time



Word Cloud analysis, Example for year 2024



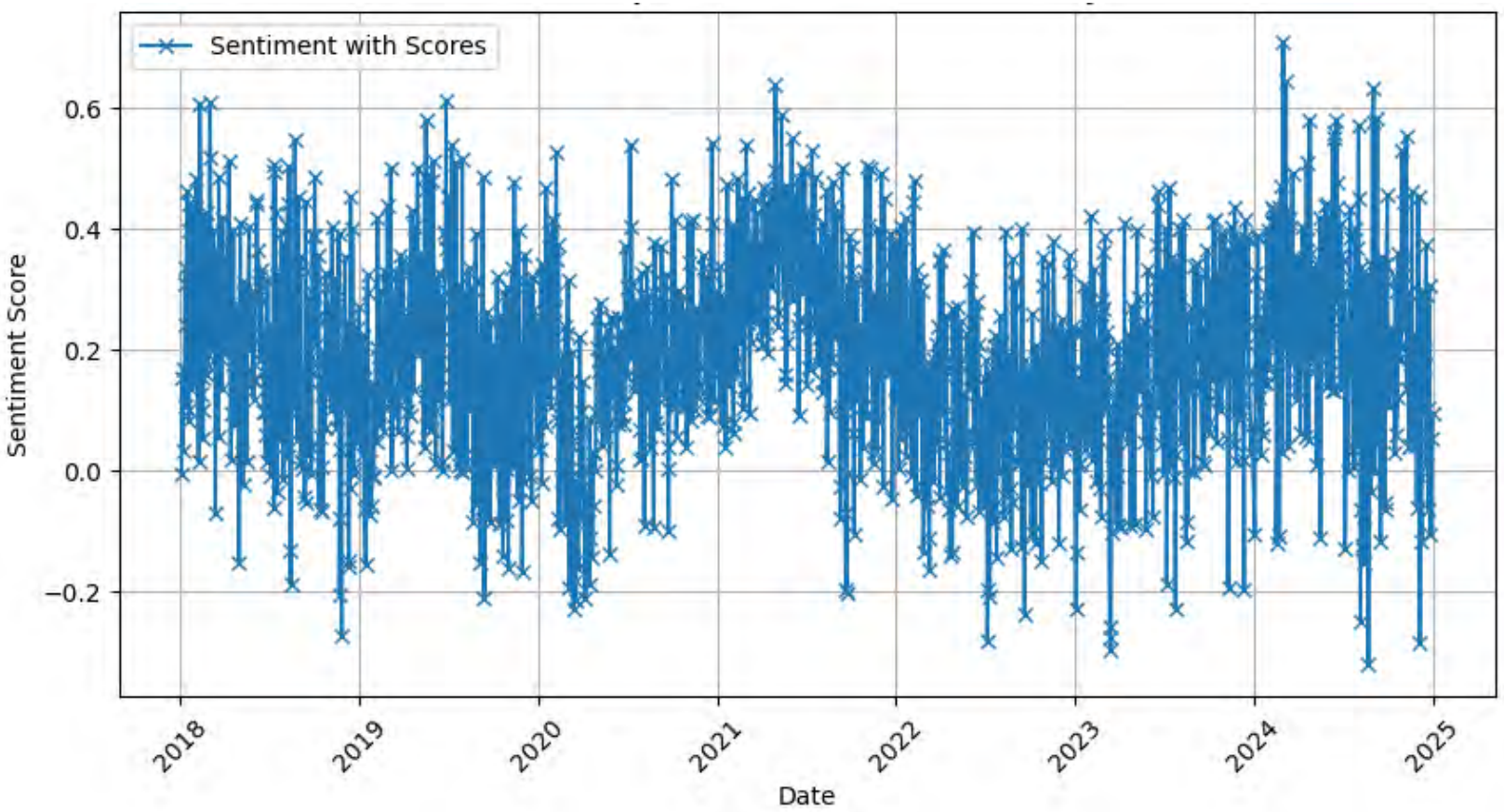
Daily Sentiment of News using Romanian Financial Stability Dictionary



$$Sentiment\ Index = \frac{P(t) - N(T)}{P(t) + N(t)}$$

where P number of positive lemmas and N number of negative lemmas

Daily Sentiment of News using Romanian Financial Stability Dictionary augmented with scores



$$Sentiment\ Score\ Index = \frac{\sum_{i=1}^{P(t)} S_p(i) + \sum_{j=1}^{N(t)} S_n(j)}{\sum_{i=1}^{P(t)} S_p(i) + \sum_{j=1}^{N(t)} |S_n(j)|}$$

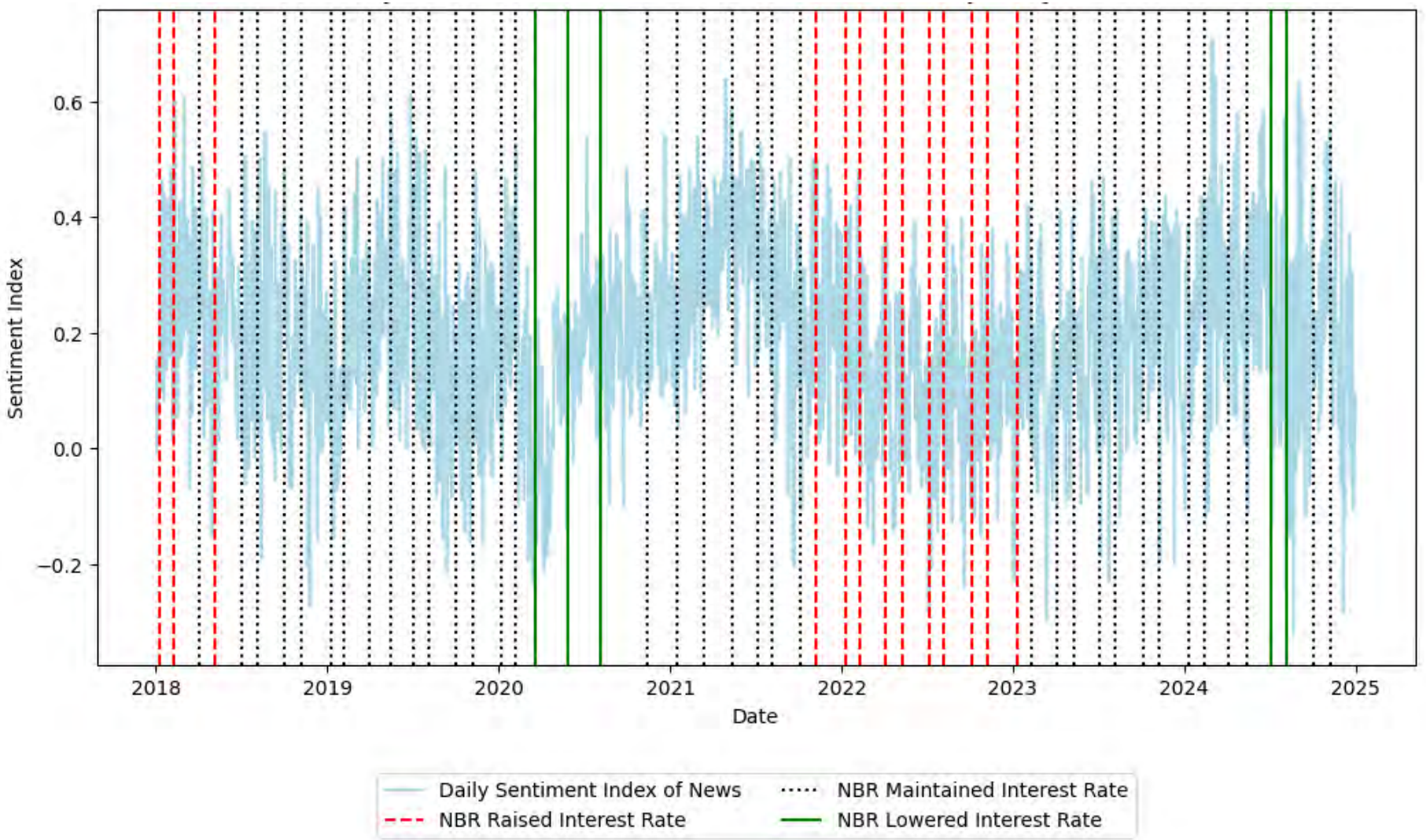
where Sp scores of positive lemmas and Sn scores of negative lemmas

- To assess the performance of our daily sentiment indicator, we developed an unsupervised K-Means++ analysis according to Kanungo (2002) on embeddings using the BERT multilingual model by Devlin (2018), after reducing their size using PCA. The results show a 93% overlap between the clusters and the binary labels from the simple index, while a slightly higher 96% overlap for the index using scores.

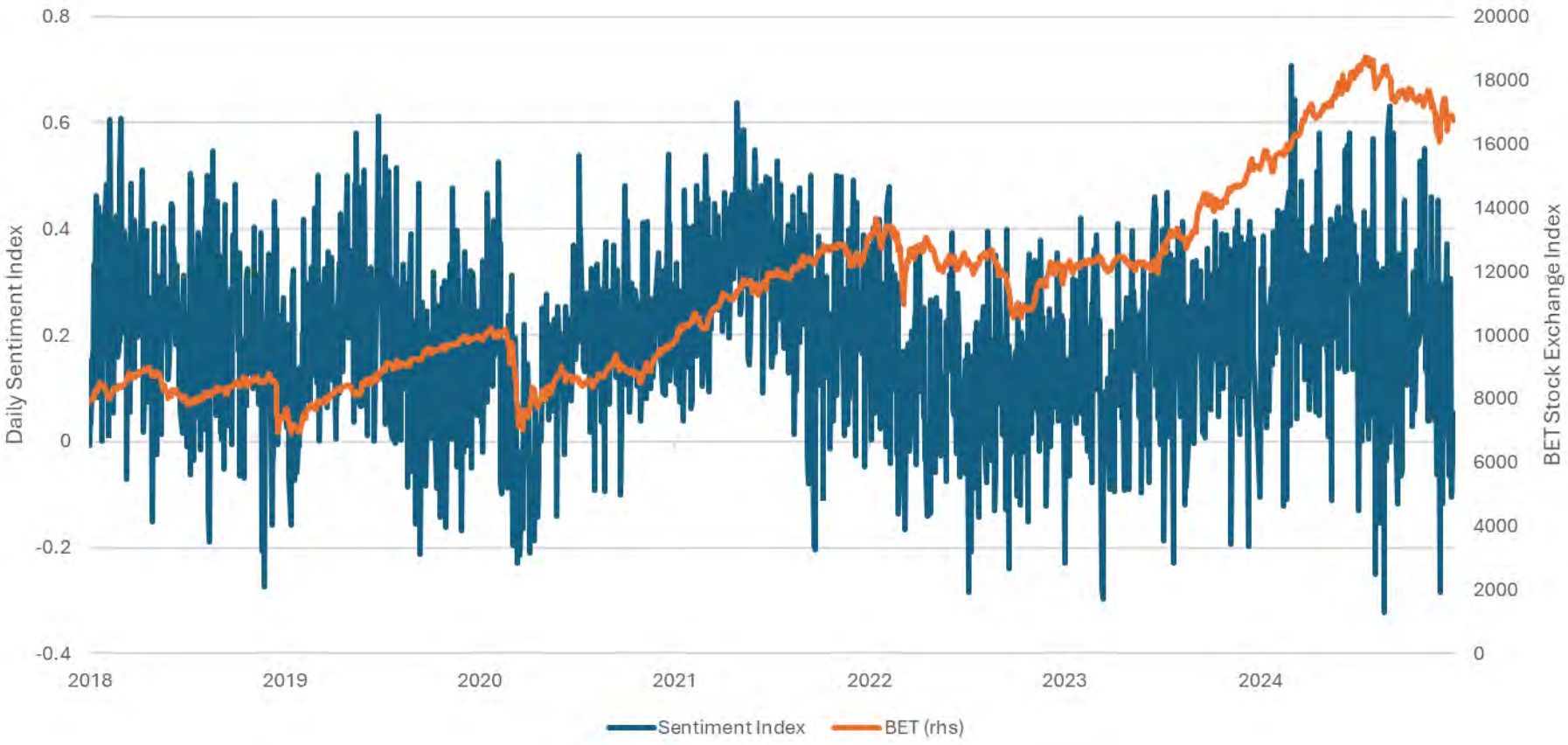
- We compared the daily sentiment of financial news with the moments of monetary policy decisions and showed that there is strong evidence that the press respond to monetary policy decisions and that the communication of the central bank is efficient in influencing the markets.

- We also found a clear correlation with macroeconomic variables: BET, the stock market index, yields of short-term bonds, as well as interbank interest rates ROBOR3M.

Daily Sentiment of Financial News and Monetary Policy Decisions



Daily Sentiment of News and Stock Market Index



Quarterly Sentiment of News and Percentage Change in Stock Market Index



Daily Sentiment of News, Interbank Interest rates and Sovereign Yield Rates



Conclusions

- This paper proposes stepping aside from classical numerical, tabular data and investing more time and effort in the use of nontraditional data, not just time series, thus implementing natural language processing methods at a large scale for some types of text that exist in a central bank. The primary goal is to provide the NBR with a guideline for pre-processing Romanian text, building NLP tools, and benchmarking sentiment analysis.
- We can only hope that by following this NLP Toolbox for the National Bank of Romania, the list of potential use cases for NLP will continue to be extended at the NBR and that economists will commit to a long-term adoption of text data in their research, as there are many other text sources available.

Thank you!

claudia.voicila@bnro.ro



References

Abodayeh, A., Hejazi, R., Najjar, W., Shihadeh, L., & Latif, R. (2023, March). Web Scraping for Data Analytics: A BeautifulSoup Implementation. In *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)* (pp. 65-69). IEEE.

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., & Monteforte, L. (2023). The power of text-based indicators in forecasting Italian economic activity. *International Journal of Forecasting*, *39*(2), 791-808.

Araci, D. (2019). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:1908.10063.

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., & Tissot, B. (2023). Data science in central banking: applications and tools. *IFC Bulletin*, (59).

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., & Tissot, B. (2023). Machine learning applications in central banking. *Journal of AI, Robotics & Workplace Automation*, *2*(3), 271-293.

Araujo, Douglas KG. 2023. "Gingado: A Machine Learning Library Focused on Economics and Finance." Working Paper 1122. BIS Working Paper. Bank for International Settlements.

Astuti, V., Bruno, G., Marchetti, S., & Marcucci, J. (2023). *News and banks' equities: do words have predictive power?* (Vol. 59). Bank for International Settlements.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, No. 2010, pp. 2200-2204).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chong, E., & Ho, S. J. (2023). *Measuring text-based sentiments from monetary policy statements–a Malaysian case study using natural language processing* (Vol. 59). Bank for International Settlements.

.

Chong, E., Ho, C. C., Ong, Z. F., & Ong, H. H. (2022). *Using news sentiment for economic forecasting: a Malaysian case study* (Vol. 57). Bank for International Settlements

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74-77).

Correa, R., Garud, K., Londono-Yarce, J. M., & Mislang, N. (2017). Constructing a dictionary for financial stability.

De Bandt, O., Jean-Charles, B., Julien, D., Alexandre, D., De Gaye, A., & Pierre-Antoine, R. (2023). *Utiliser la presse pour construire un nouvel indicateur de perception d'inflation en France*. Banque de France.

Denes, J., Lestrade, A., & Richardet, L. (2022). *Using twitter data to measure inflation perception* (Vol. 57). Bank for International Settlements.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doerr, S., Gambacorta, L., & Garralda, J. M. S. (2021). Big data and machine learning in central banking. *BIS Working Papers*, (930).

Dumitrescu, S. D., Avram, A. M., Morogan, L., & Toma, S. A. (2018, June). Rowordnet–a python api for the romanian wordnet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1-6). IEEE.

Edison, H., & Carcel, H. (2021). Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts. *Applied Economics Letters*, *28*(1), 38-42.

Evdokimova, T., Nagy-Mohacsi, P., Ponomarenko, O., & Ribakova, E. (2023). Central banks and policy communication: How emerging markets have outperformed the Fed and ECB. *Peterson Institute for International Economics Working Paper*, (23-10).

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Istrefi, K., Odendahl, F., & Sestieri, G. (2022). ECB communication and its impact on financial markets.

Istrefi, K., Odendahl, F., & Sestieri, G. (2023). Fed communication on financial stability concerns and monetary policy decisions: Revelations from speeches. *Journal of Banking & Finance*, *151*, 106820.

Jabbar, M. A., Wibisono, O., Widjanarti, A., & Zulen, A. A. (2023). Machine learning for measuring central bank policy credibility and communication from news. *IFC Bulletins chapters*, *59*.

Jolliffe, I. T. (1986). Principal component analysis. *Springer Series in Statistics*. Springer-Verlag.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.

Lee, Y., & Seo, B. (2023). Extracting economic sentiment from news articles: the case of Korea. *IFC Bulletins chapters*, *59*.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65.

Mabey, B. (2018). pyLDAvis documentation.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Richardson, L. (2007). Beautiful soup documentation.

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019, October). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 sixth international conference on social networks analysis, management and security (SNAMS)* (pp. 338-343). IEEE.

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Transforming survey analysis: tools for central banks[1]

Nicholas Gray and Dominic Jones,
Reserve Bank of Australia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Transforming survey analysis

Tools for central banks

Nicholas Gray and Dominic Jones[1]

Abstract

We provide recommendations based on learnings from survey exploratory text analysis at the Reserve Bank of Australia. We provide recommendations around a broad framework of *gradual supervision*, *exploration and iterative refinement*, and *modelling co-occurrences*.

The views expressed in this paper are those of the authors and should not be attributed to the Reserve Bank of Australia. Any errors are the sole responsibility of the authors.

Keywords: text analysis, survey analysis, natural language processing, weakly supervised, text classification.

JEL classification: C80, C83, C55, C45, C19

[1] Reserve Bank of Australia

Motivation

Surveys are commonly used at central banks. For example, surveys on consumer payments, trust, banknote design, inflation expectations, or business conditions.

Surveys often contain questions with free-text response. Sometimes they are used to complete incomplete lists (e.g. "if other, please describe…"), but frequently they are open-ended. For example, a question that asks a respondent about their views on holding cash. These questions are useful for eliciting what's "top of mind" [2]. The downside is that free-text responses are unstructured – they cannot easily be graphed, modelled, or otherwise analysed. Some sort of transformation needs to occur first. In our experience, this often means "coding": extracting a set of key themes and tagging each response with the relevant ones. In the language of text analysis, this is a classification task.

If the survey sample is small enough, manually reading the responses is the best option. In many cases, the sample is large – on the order of thousands of responses, sometimes more. Often, the burden of coding so many responses is too high. However, modern NLP techniques can help.

Framework

In our experience, analysing surveys (or other text corpora) follows a few key themes. We have written them down as a framework to share our learnings and contribute to discussion in the central banking community. The framework has three parts. The parts are not mutually exclusive, but we think they are usefully considered separately.

The first part of the framework is *gradual supervision*. Subject-matter experts often learn throughout the analysis process, and it's important to keep the analysis up to date with the evolving understanding. We make heavy use of 'weakly-supervised' classification techniques here and close collaboration with subject-matter experts.

The second part is *exploration and iterative refinement*. Here, we provide tools and recommendations that help subject-matter experts to explore and comprehend the analysis at any stage.

The third and final part is *modelling co-occurrences*. We are often asked about how key themes co-occur. It has been useful for us to formally model this in many analyses. We provide detail and recommendations about how to do it.

Gradual supervision

Usually, subject-matter experts start knowing perhaps a little – a prior, or maybe some analysis of a subsample. As analysis proceeds, we learn more about the subject. Ideally, the analysis itself should be responsive to this growing knowledge.

Traditionally, NLP techniques are grouped under two headings: unsupervised and fully supervised. Unsupervised techniques take nothing or very little as input, and therefore are easy to use, but can be difficult to interpret. Fully supervised techniques take expert-annotated 'gold' data as input, and allow for state-of-the-art classifications, but are expensive. Both are, of course, useful. Recently, 'weakly supervised' techniques have become more powerful and common, bridging the gap between un- and fully-supervised methods. We now use these methods to keep up with the growing understanding – and thus supervisory power – resulting from analysis. In this section, we will briefly discuss the methods we use, why, and conclude with some recommendations.

Unsupervised

Descriptive statistics like term or phrase frequencies are easy to run and tend to be interpretable. In jargon-rich fields, term frequencies tend to be more useful. We often pair word-level frequencies with phrase-level frequencies. Phrases – several words better understood as a single unit of meaning – can be extracted in several ways. The simplest is to compute n-gram frequencies, but more sophisticated phrasal segmentation algorithms exist.

Unsupervised topic models – we tend to use BERTopic [1] – are workhorses in this category. Rather than term frequencies, topic models extract 'topics' from a set of documents. Often, these topics are a set of terms, and a document can be represented by a mixture of topics. This maps neatly on to questions usually asked at early stages of an analysis: "what sort of things are people saying?"

Plotting "embedding spaces" and iteratively refining them is a common technique in this category, discussed below.

Weakly supervised

Weakly supervised approaches fill in the gap between unsupervised and fully supervised methods. When we don't have labelled data, these methods allow us to better use whatever guidance we *do* have – be it a list of queries, keywords, examples, or a taxonomy.

Weakly supervised methods have much similarity with information retrieval methods. There is some degree of supervision given to the algorithm – which might otherwise be called a *query*. We then ask what documents are relevant to the query. Therefore, traditional information retrieval methods can be useful here whenever the question can be cast appropriately. For example: "which responses are relevant to unemployment?" We frequently use keyword-based relevance schemes (e.g. BM25 [6]) and transformer-based retrievers (e.g. sentence transformers [5]) for this task.

Topic models can also be used in a weakly supervised fashion by providing a list of "seed" words to build topics around. CatE [4] or seeded BERTopic (or many others) can be used for this purpose. These techniques are especially useful to enrich current understandings.

Finally, often survey responses are "coded" into several categories defined by some codebook. For example, a response might be about unemployment and future

expectations but not about investment intentions. While the construction of the codebook itself is mostly appropriate for subject-matter experts, the actual coding is a *classification* task.

Since constructing a fully supervised classifier is often expensive, weakly supervised classifiers are a useful alternative. These may take the form of a dataset seeded by information retrieval techniques, a prompt-based LLM approach, or specialised algorithms like X-Class [8]. While the performance of a weakly supervised classifier is lower than that of a fully supervised one, it is often not too much lower while being much cheaper to put together. We have found this to be important for speeding up feedback loops between data scientists and subject-matter experts.

Fully supervised

That is not to say fully supervised classifiers are obsolete. When circumstance demands the best possible performance, or when the survey is repeated and temporal consistency is useful, traditional fully supervised classifiers are effective. We often reach for a fine-tuned RoBERTa [10] (though we have been experimenting with ModernBERT [9]).

Recommendations for gradual supervision

Work closely with the subject-matter experts. Communicate the nature of what you are doing and be honest about limitations. When you are eliciting guidance – for example, lists of relevant keywords or even a taxonomy – be clear about what format you want it in. Even consider making a form for it.

Reporting matters: spend some time making your reports digestible.

Finally, here are a few questions we tend to ask experts when we're thinking about moving up the chain towards more supervised techniques:

- Is this worth more investigation?
- Should we cut this out of the analysis?
- Are there any indicators of this sort of response – textual or otherwise?
- Are there any sub-categories you might be interested in?
- Would you like to track this over time, if the survey is repeated?

Exploration and iterative refinement

Document clustering is a useful exploratory technique. The most basic form of exploration is simply looking through responses manually. Using spreadsheets can be unwieldly, especially if there are many similar responses. Instead, we often use "embedding spaces". Documents – survey responses – are first converted into numerical "embeddings", which are then plotted in an "embedding space", with potentially some sort of clustering algorithm applied. It's often common to colour points in visualisations of these embedding spaces according to however they've

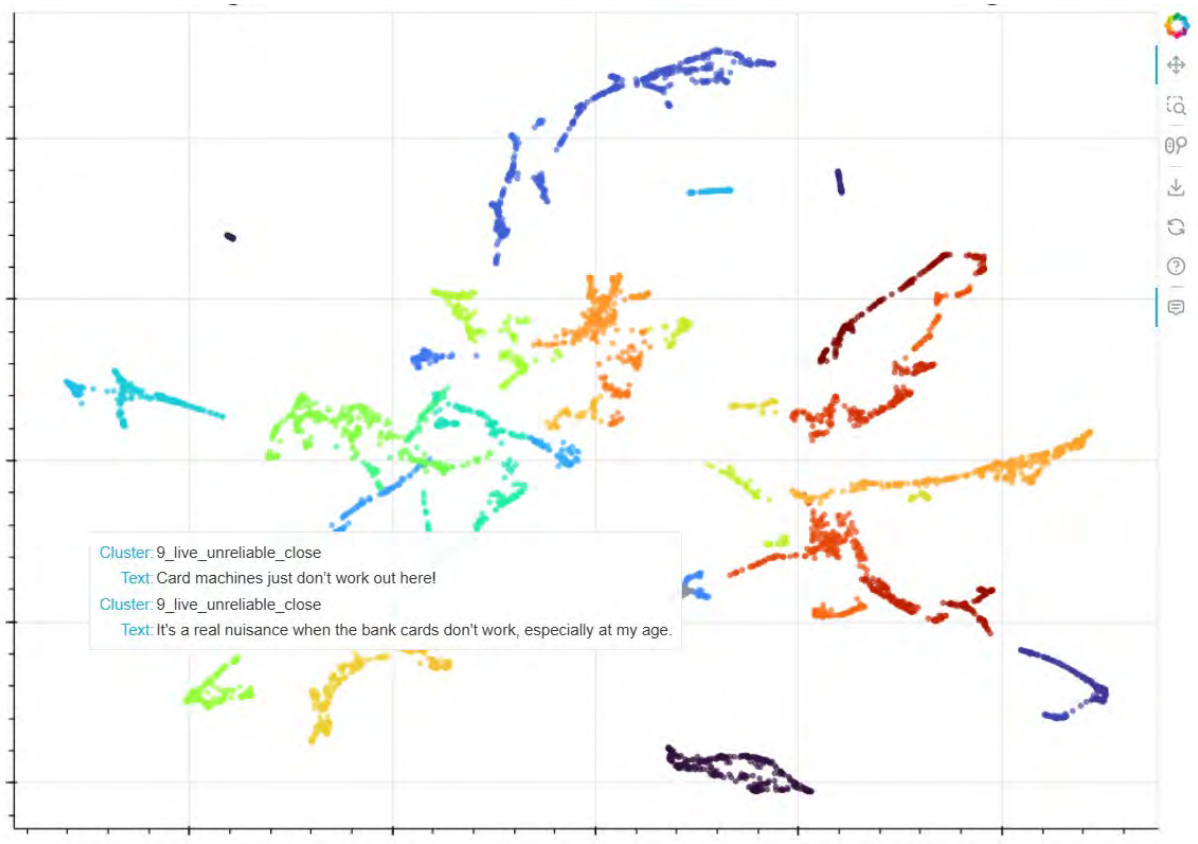
been categorised (e.g. via a gradual supervision pipeline). It's also common to use the embedding space *itself* to recover categories. When working with these embedding spaces we often think of two interrelated steps – *exploration* and *refinement*.

Exploration

Exploring is crucial for building a better understanding of themes in the data. It also makes it easier to explain themes to or loop in subject-matter experts, helping the refinement step. Crucial aids for exploration are *interactivity* and *summarisation*.

2D Visualisation of Sentence Transformer Embedding Space

Dimension reduction with PaCMAP; Clustering with HDBSCAN; Artificial data — Graph 1



Data is artificially generated using ChatGPT to simulate expected responses in a hypothetical "Cash Preference" survey.[2] We used the default settings for PaCMAP.

The user is currently hovering over responses assigned to "9_live_unreliable_close" cluster.

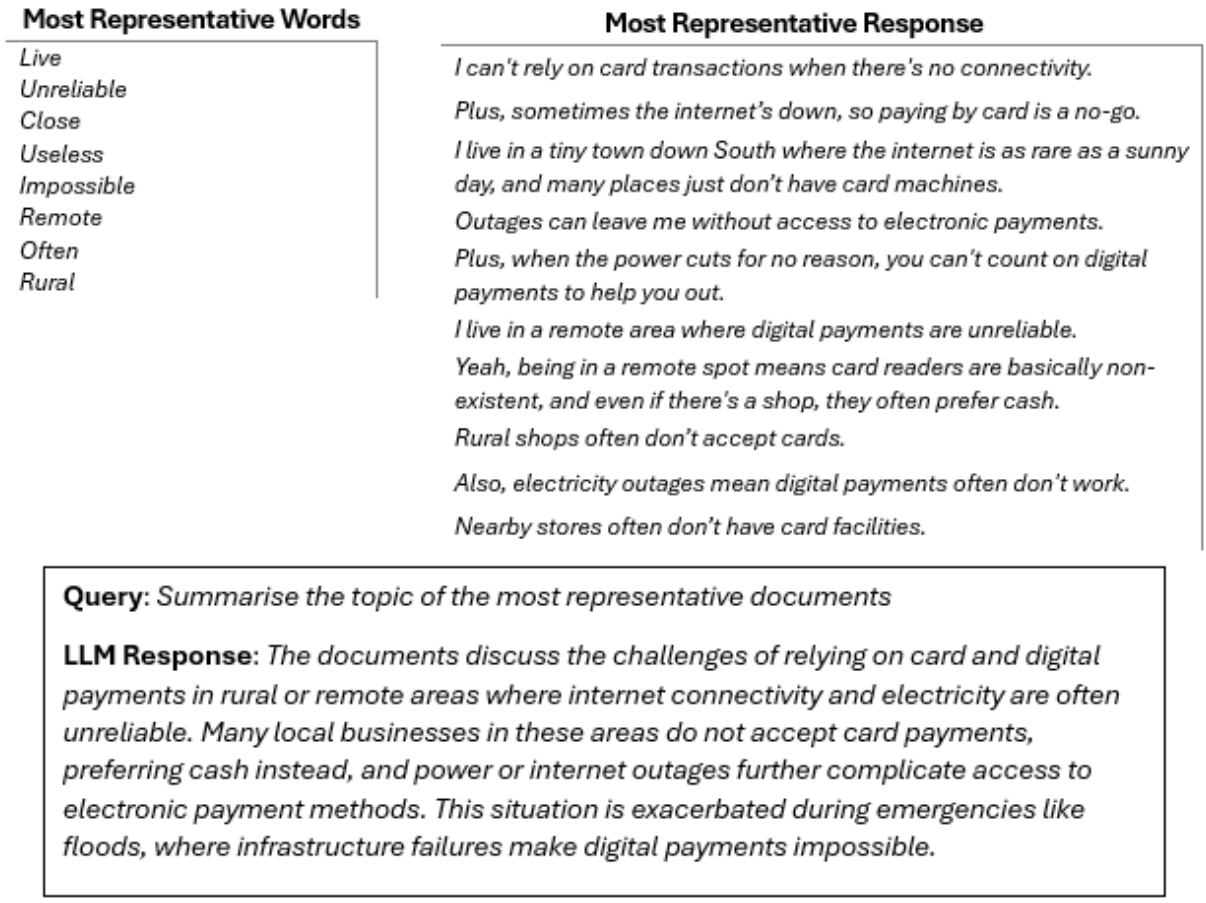
Sources: RBA

[2] All responses are created by the model based on prompts for generalised themes expected in such a survey. No real survey data is provided to the model.

Interactivity. The simplest technique is to apply a dimensionality reduction algorithm such as UMAP [3] or PaCMAP [7] to reduce the embedding space to 2 dimensions. This allows us to draw a scatter plot, where each point represents a document (survey response). The distance between points is proportional to the similarity (however defined) between documents. We can also colour the points based on some categorisation, such as cluster assignments. Interactivity allows users to zoom in on clusters and hover over points to read the text that each point represents. Visually exploring responses to surveys can reveal interesting characteristics of responses and clusters, which can help refine analysis.

Summarisation of Response Cluster

Characteristics for cluster "9_live_unreliable_close"; Artificial data — Graph 2



Same artificially generated data as Graph 1.

LLM summary using locally hosted model[3]

Sources: RBA

[3] Microsoft's Phi-4 model.

Summarisation. Hovering and reading clustered responses one-by-one can only take one so far. To speed up feedback loops, we often *summarise* clusters of responses. For this, we use cluster-indicative keywords, representative responses, and LLM-based summaries.

For cluster-indicative keywords, we often use "class-based" TF-IDF. It extracts words or phrases that are more informative to one class (cluster) than others by weighting term frequency and how common that term is across different clusters. High scoring terms are more representative of responses in the cluster. Informative keywords identified here can also be fed into a weakly supervised classifier or other method to refine response clusters.

For representative responses, we use the geometry of the clusters themselves. Since they live in some embedding space, we can retrieve responses close to the centroid of the cluster. This gives us specific responses that are more representative of the cluster. Reading these responses can provide a quicker and clearer overview of the main themes. Conversely, we can look at the most unusual responses in the cluster by extracting the responses furthest from the cluster centroid. Often these responses are not typical to the central theme of the cluster. These responses can be informative for subject-matter experts or may signal that the cluster needs refinement. By looking at responses from across the cluster geometry, we aim to avoid biases that arise from only considering measures of central tendency.

Finally, we sometimes use (small, local) LLMs to generate summaries of clusters. These are also effective at providing quick and clear summaries of clusters. Depending on how sophisticated the LLM or prompt is, we can also ask it to extract information: for example, keywords, entities. We may even ask it specific questions about responses. If there are token limits, we often use responses near the centroid as above, or even just a random sample.

(Iterative) refinement

Improvements are always identified in whatever preliminary clustering we look at. This is often the step where techniques from gradual refinement are useful. In any case, we may want to combine, split, throw away, or even completely refresh clusters. By gaining insight from exploration techniques and then updating the analysis using appropriate techniques closes the exploration-refinement loop.

The interactive plot, despite being a highly summarised version of the embedding space, can quickly reveal clusters in need of refinement (such as splitting or merging). Looking at the clusters summarised by keywords or representative examples can also indicate whether clusters need further analysis. Finally, using LLMs to summarise clusters speeds up review, especially if there are many.

Recommendations

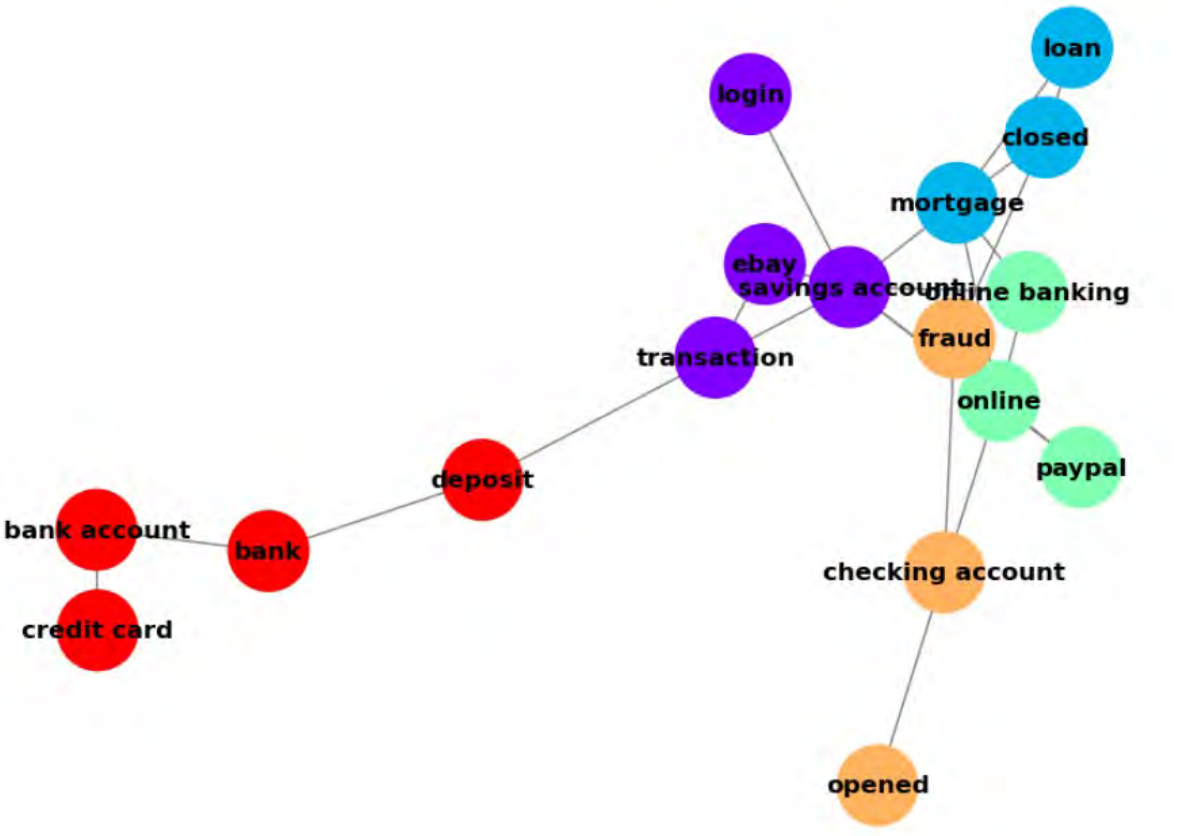
Sit down with your subject-matter experts and explain your charts. Let them see the data underlying any LLM-generated summary. And, where possible, automate the process (e.g. generating summaries or keywords) to tighten feedback loops. Feed refined clusters back into the gradual supervision pipeline – this is often most useful between the un- and weakly-supervised phases.

Modelling co-occurrences

Once we've classified, tagged, coded, or otherwise categorised a set of responses, we're often asked about how categories co-occur. For example, what is the probability of a response being about monetary policy, *conditional* on it being about inflation? Or are there groups of categories that often appear together? Answers to these sorts of questions are often useful for subject-matter experts. There are two ways that we often think about answering these questions. We'll refer to the categories as *concepts* for generality.

Example Concept Co-Occurrence Network

Graph 3



Edges between topics indicate these two topics often co-occur. This was generated from the publicly available CFPB Consumer Complaint Database.

Sources: CFPB, RBA

Bag-of-concepts

A natural way to think about concept co-occurrence is by analogy to topic models. In a simple scenario, a document is a "bag" of words (i.e. the order doesn't matter), and topics are words that often co-occur. One way to estimate these topics is by representing the document bags-of-words as a document-word matrix and then factorising this matrix. Principal Component Analysis or Non-negative Matrix Factorisation are appropriate here. For concept co-occurrence, we have a document-concept matrix (i.e. the output of your classifiers), with each row (column) being a document and each column (row) recording which categories it belongs to. We then factorise in the same way as a simple topic model. The resultant output will highlight groups of concepts that co-occur.

Concept co-occurrence networks

An alternative (but not unrelated) method of modelling co-occurrences is to use tools from network science. For each pair of concepts, we can define conditional probabilities or Jaccard similarity (intersection over union). This is naturally represented as an adjacency matrix, which is interpreted as a graph. From there, we might use filtering heuristics (e.g. we are only interested in links with this much weight) or "backboning" algorithms to filter noise from the network. To find groups of concepts that co-occur, we then apply community detection algorithms to the filtered network. We find this effective as a visual device – network plots can be intuitive once explained. (See graph 3.)

Conclusions

Based on our recent experiences in exploratory analysis of textual survey responses, we present a small framework based around *gradual supervision, exploration and iterative refinement,* and *modelling co-occurrences*. Whatever part of the framework we're thinking about – our general recommendations apply:

- Work closely with your subject-matter experts.
- Use the whole gamut of techniques from un- to fully-supervised.
- Think hard about how to speed up exploration of the data (e.g. via interactivity and summarisation).
- Build tight interactive loops with subject-matter experts to refine your analysis (e.g. cluster groups).
- Use LLMs to reduce cognitive load on experts, but ensure they have access to the underlying data.
- Think about formally modelling co-occurrences if there are a lot of potential overlaps.

References

[1] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794

[2] Ingar K. Haaland, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2024. Understanding Economic Behavior Using Open-ended Survey Data. https://doi.org/10.3386/w32421

[3] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (September 2018), 861. https://doi.org/10.21105/joss.00861

[4] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative Topic Mining via Category-Name Guided Text Embedding. In *Proceedings of The Web Conference 2020*, April 20, 2020. ACM, Taipei Taiwan, 2121–2132. https://doi.org/10.1145/3366423.3380278

[5] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

[6] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *FNT in Information Retrieval* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[7] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research* 22, 201 (2021), 1–73. Retrieved April 30, 2025 from http://jmlr.org/papers/v22/20-1061.html

[8] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021. Association for Computational Linguistics, Online, 3043–3053. https://doi.org/10.18653/v1/2021.naacl-main.242

[9] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. https://doi.org/10.48550/arXiv.2412.13663

[10] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, August 2021. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. Retrieved April 30, 2025 from https://aclanthology.org/2021.ccl-1.108/

RESERVE BANK OF AUSTRALIA

Transforming Survey Analysis

Tools for Central Banks

Nicholas Gray and Dominic Jones
2025 BIS IFC Workshop on Data Science in Central Banking
February 2025
The views expressed in this presentation are those of the authors and should not be attributed to the Reserve Bank of Australia. Any errors are the sole responsibility of the authors.

Surveys

- Surveys often include free-text open-ended questions.
  - They’re great for eliciting what’s top of mind – or getting at unstructured thoughts or things the survey had missed.

- If the survey sample is small enough, reading responses is the best option.

- But usually, the sample is huge! NLP techniques can help build a qualitative understanding using quantitative techniques.

Surveys at central banks

- Consumer payments surveys
- Trust and credibility
- Banknote design
- Inflation expectations
- Business surveys
- Lending surveys

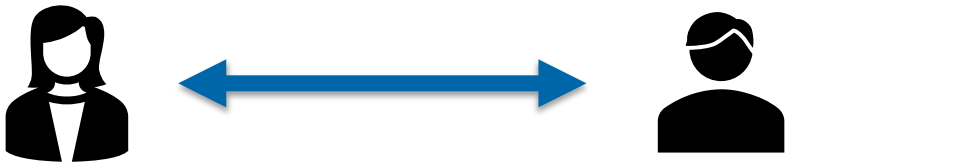
Framework

- Gradual supervision
  - Making use of expert knowledge where we have it…
  - …providing guides where we don't.
- Cluster exploration
  - Providing richer descriptions of topics to help with comprehension
- Iterative refinement
  - Subject-matter experts in the loop.
- Concept matrices
  - Are there any complex interactions worth thinking about?

Gradual supervision

- Usually, SMEs start knowing not much (but not zero) and gradually learn more about survey responses.
- Our old model – just run a topic model – didn't reflect this growing understanding of the data.

- Luckily, techniques now exist that cover the whole gamut of unsupervised to fully supervised – and allow us to incorporate growing expert knowledge!



Unsupervised

Weakly supervised

Fully supervised



- Input: not much
- Easy
- Hard to interpret sometimes

- Input: anything you can think of
- Cheap
- Very good quality

- Input: labelled examples
- Costly + high expert effort often required
- High quality

Unsupervised techniques

- Descriptive statistics: wordclouds, TF-IDF scores, phrasal segmentation algorithms.

- (Unsupervised) Topic modelling
  - BERTopic

- Information retrieval
  - BM25, pretrained sentence transformers.
  - Great for incorporating priors!

Weak supervision

- When SMEs start building a picture it helps to iterate fast.
  - Full supervision is expensive and slows down the loop.
  - Trying candidate models should be cheap.

- Guided topic models
  - BERTopic (again)
  - CatE
- Prompt-guided embeddings
  - Using natural language to shape the embedding space.
- Weakly-supervised classifiers
  - Fast to build!

Full supervision

- When SMEs have a strong understanding of the data – or want to be consistent over multiple surveys – it pays to train a fully supervised classifier.

- We often use **taxonomy classification** and train cross-encoders or text matching networks.

- It’s useful to think about how to communicate these models and to track performance over time.

Recommendations

- Work closely with your SMEs.
  - Communicate the nature of what you're doing and be honest about limitations.
  - When eliciting guidance be clear about what format you want it in. Consider making a form for it!

- Reporting matters.
  - Do spend some time making reports easily digestible.

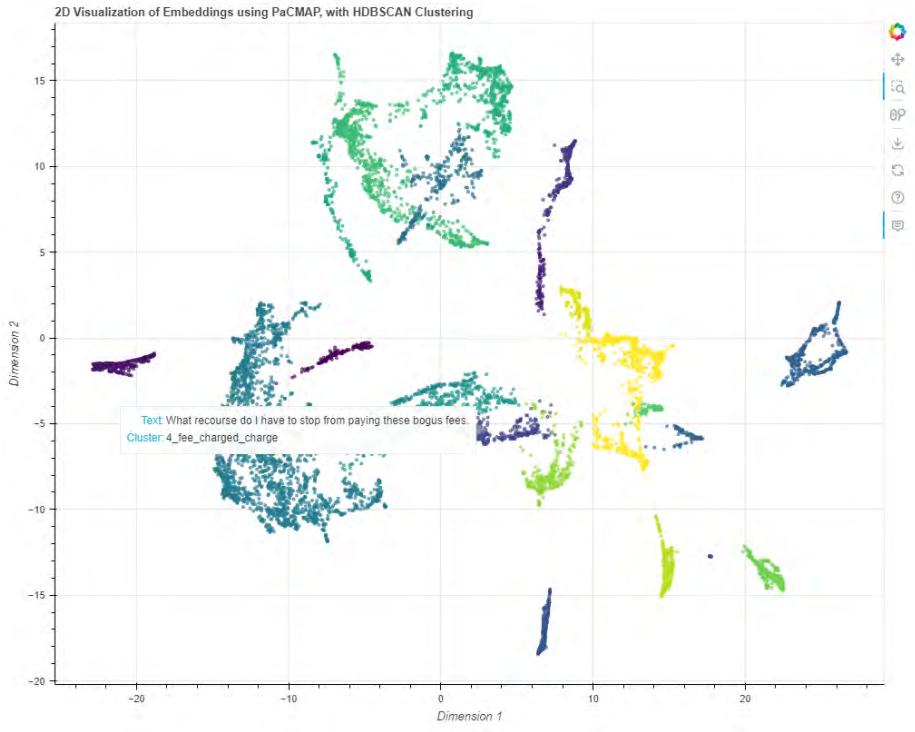
- When should we move up the chain? Questions to ask SMEs:
  - "Is this worth more investigation?"
  - "Should we cut this out of the analysis?"
  - "Are there any indicators of this sort of response?"
  - "Are there any sub-categories you're interested in?"
  - "Do you want to track this over time for the next survey?"

Cluster exploration

- Clusters, groups, classes don’t need to be set in stone.
- Especially when understandings change!
- Agnostic of method, geometry of embedding space can be utilised.
- Providing interactivity and multiple perspectives important for reducing potential bias.

Cluster exploration

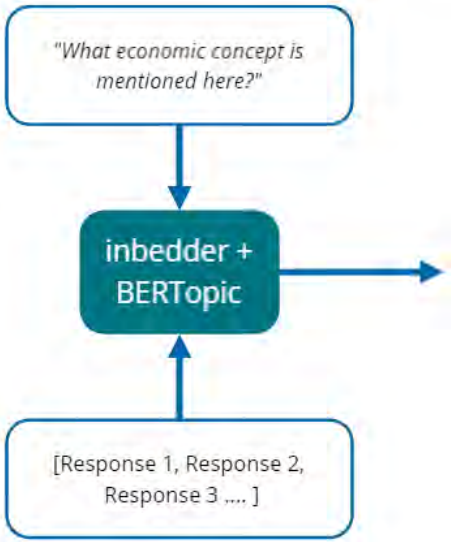
Interactivity:

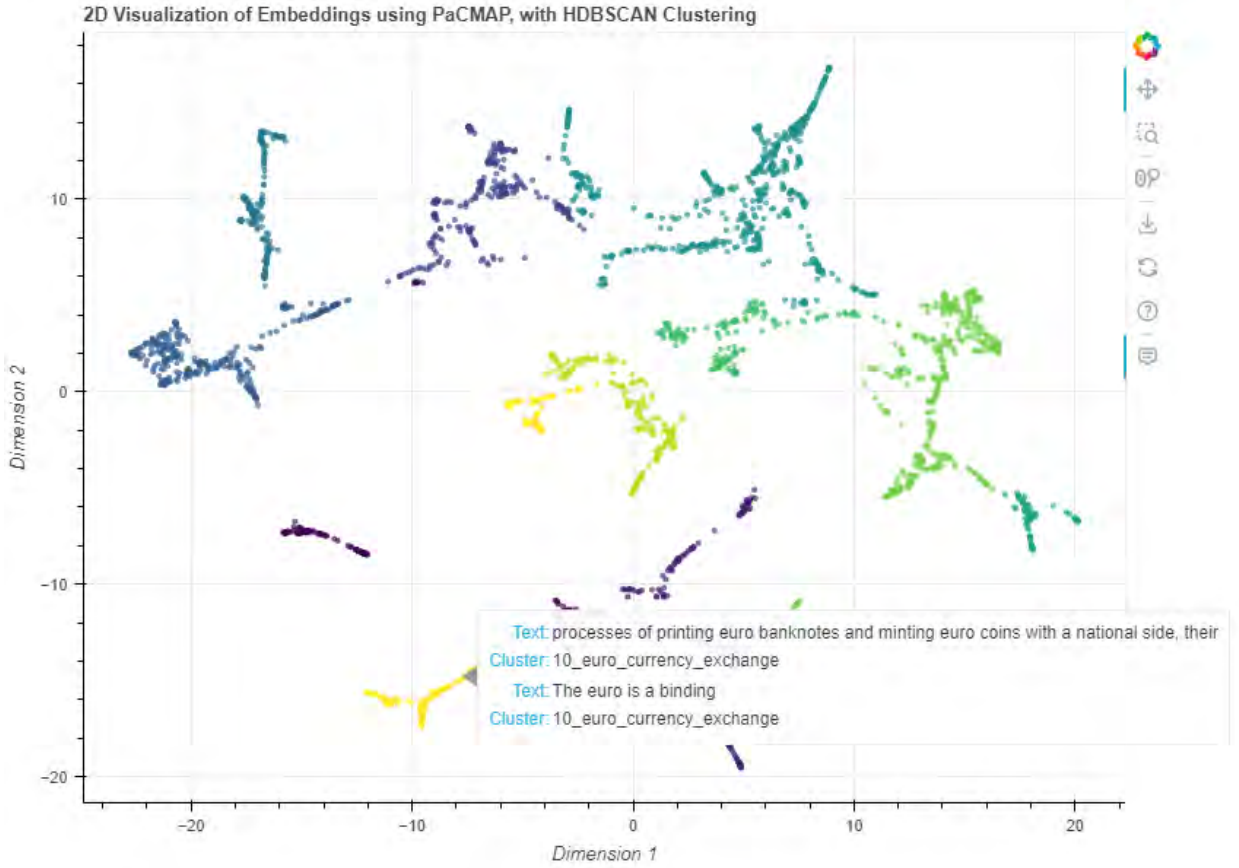


Representative responses:

Closest to centroid of *7_credit_score_report* cluster:	
Distance from centroid	**Reponses**
0.856	I didn't know this was being reported to the credit company until after it was done, I am trying to buy a house and this is causing problems.
0.827	They reported my account to the credit bureaus as 30 days late.
0.813	I contacted my bank in regards to this issue as they are reporting every month to the credit reporting agency 's with remarks which is affecting my credit score in a negative way.
0.805	I told them that was not fair because I cleared up my credit and it is in good standing.
0.804	This company has pulled my credit multiple times for no reason as well.

Prompts, too!


"What economic concept is mentioned here?"
inbedder + BERTopic
[Response 1, Response 2, Response 3 ....]


2D Visualization of Embeddings using PaCMAP, with HDBSCAN Clustering
Dimension 2
Dimension 1
10
0
−10
−20
−20
−10
0
10
20
Text: processes of printing euro banknotes and minting euro coins with a national side, their
Cluster: 10_euro_currency_exchange
Text: The euro is a binding
Cluster: 10_euro_currency_exchange

LLM summaries

- All running on local LLMs for privacy reasons.

Q: Provide summaries of the most unusual document in cluster*

Response:

Scam Involving Timeshare Conversion:

- Sale of a timeshare condo as a campground membership.
- Down payment of $10,000 without delivery of promised timeshare product.
- Denial of refund despite a contract addendum.
- Accusation of credit card fraud protection guarantee non-honoring by Chase.
- Loss of significant funds without receiving promised item or refund.
- Seeking governmental and legal assistance due to fraud.

Unauthorized Charges and Identity Theft:

- Unauthorized changes to bank account details and subsequent unauthorized transactions.
- Denial of dispute by bank despite evidence of fraudulent activity.
- Credit bureaus listing unauthorized transactions as late payments and charge-offs.
- Loss of funds due to identity theft and fraudulent charges.
- Seeking assistance to recover stolen money and address unauthorized account changes.

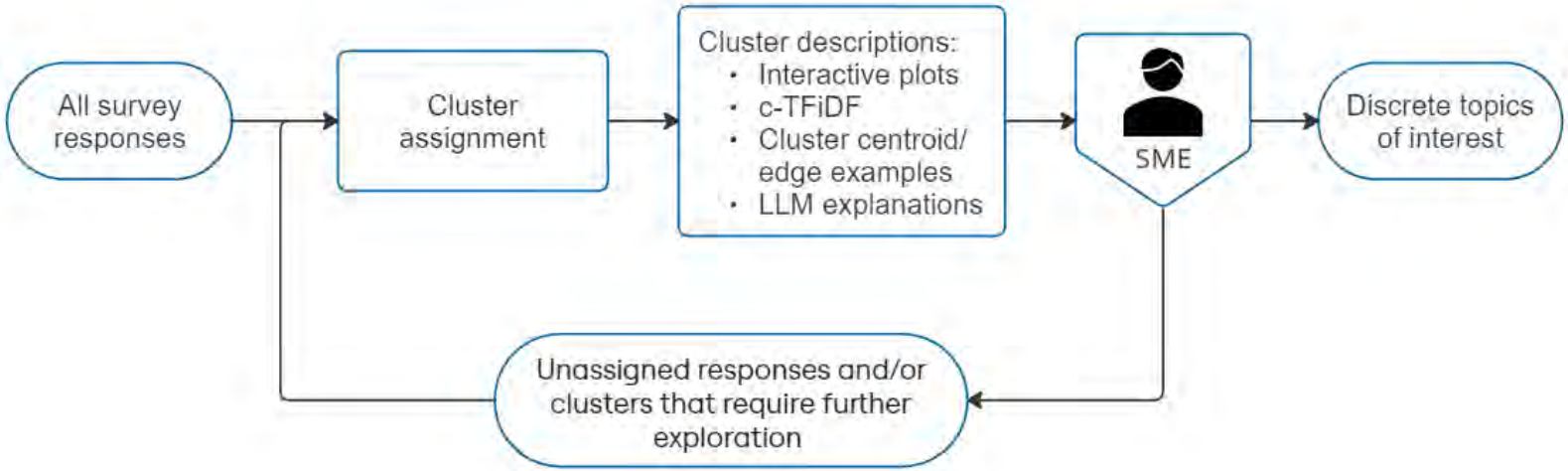
Fraudulent Transactions and Disputes:

- Merchant fraudulently disputing charges and misrepresenting to credit bureaus.
- Credit bureaus listing unauthorized transactions as legitimate.
- Credit card companies failing to resolve disputes and continuing to charge customers.
- Scenario involving a guest with a Capitol One card charged twice for a non-existent hotel stay.
- White Dial Stainless Steel watch purchase fraudulently disputed due to account access by an unauthorized third party.

* More prompt engineering behind the scenes

Iterative refinement

- Cluster assignments – no matter what algorithm you use – can be a mixed bag.
- More problematic: sometimes they're hard to explain!
- Having SMEs "in the loop" helps here.



Iterative refinement

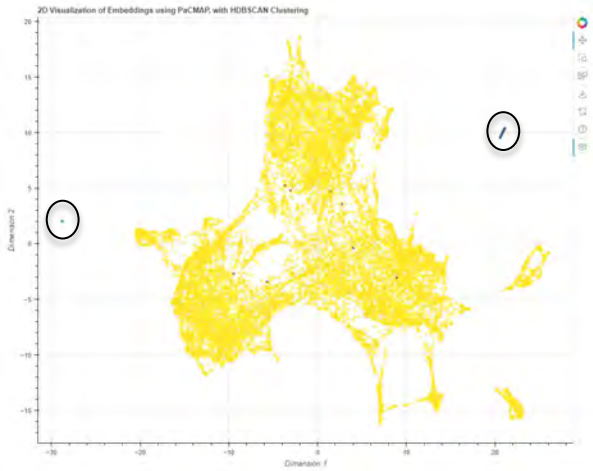
- Cache discrete clusters (circled)
- Rerun clustering on large central cluster

→

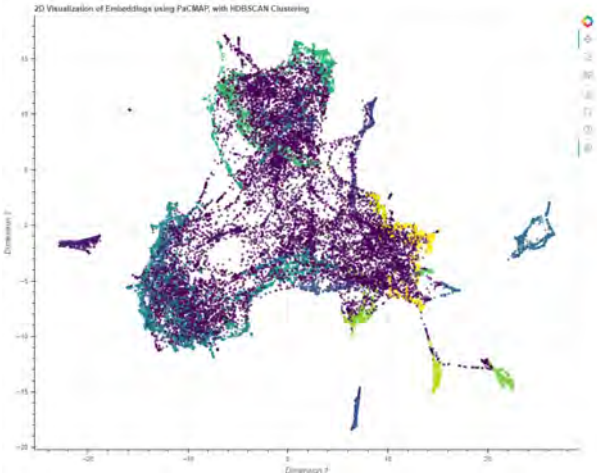
- Remove noise cluster (for downstream re-clustering)
- Focus in on clustered response

→

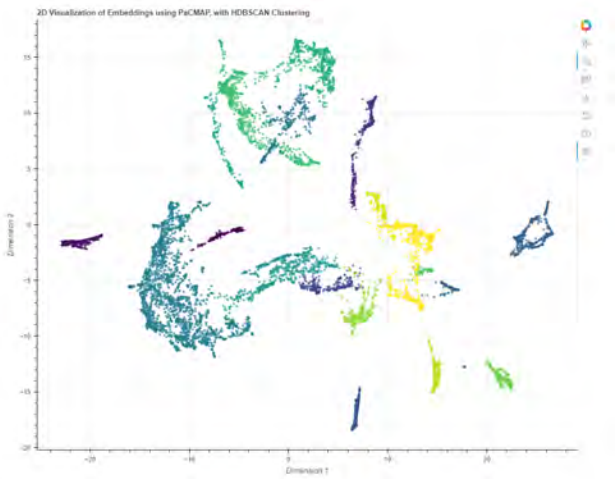
Run analysis on 20 clusters (plus 2 cached clusters from first iteration)



sentences = 23049
number of clusters = 3
noise cluster count = 12



sentences = 22882
number of clusters = 20
noise cluster count = 13818



sentences = 9064
number of clusters = 20
noise cluster count = 0

Recommendations

- Sit down with your SMEs and explain all your charts.
- Let them see the data underlying any LLM summaries – the chain of trust is important!
- Automate the process where possible – tight feedback loops are good.

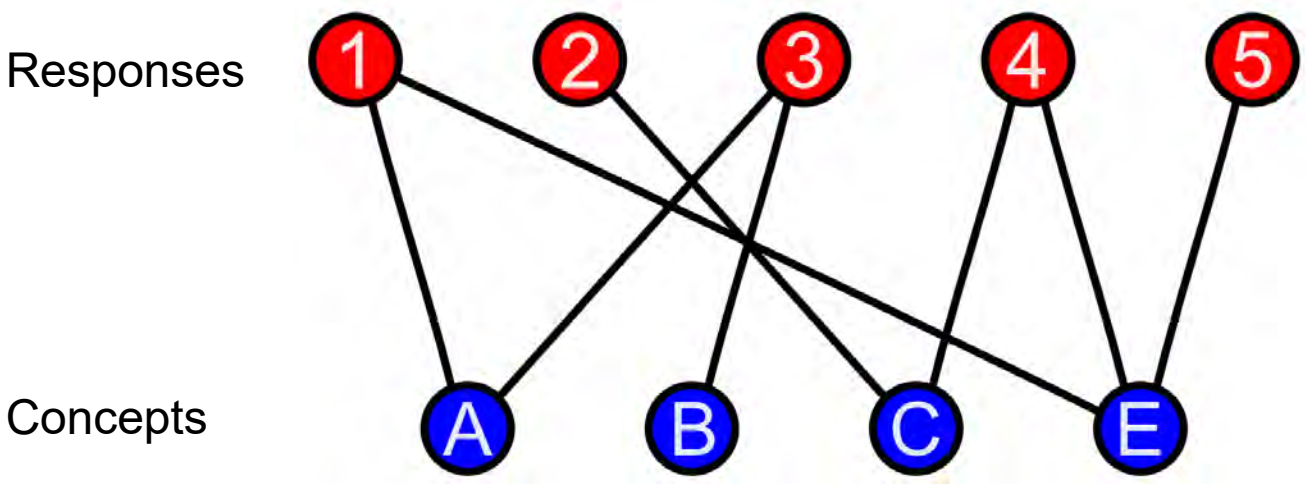
- Feed the refined clusters back into the gradual supervision pipeline.
  - Usually, iterative refinement takes place somewhere between the un- and weakly-supervised steps.

Concepts

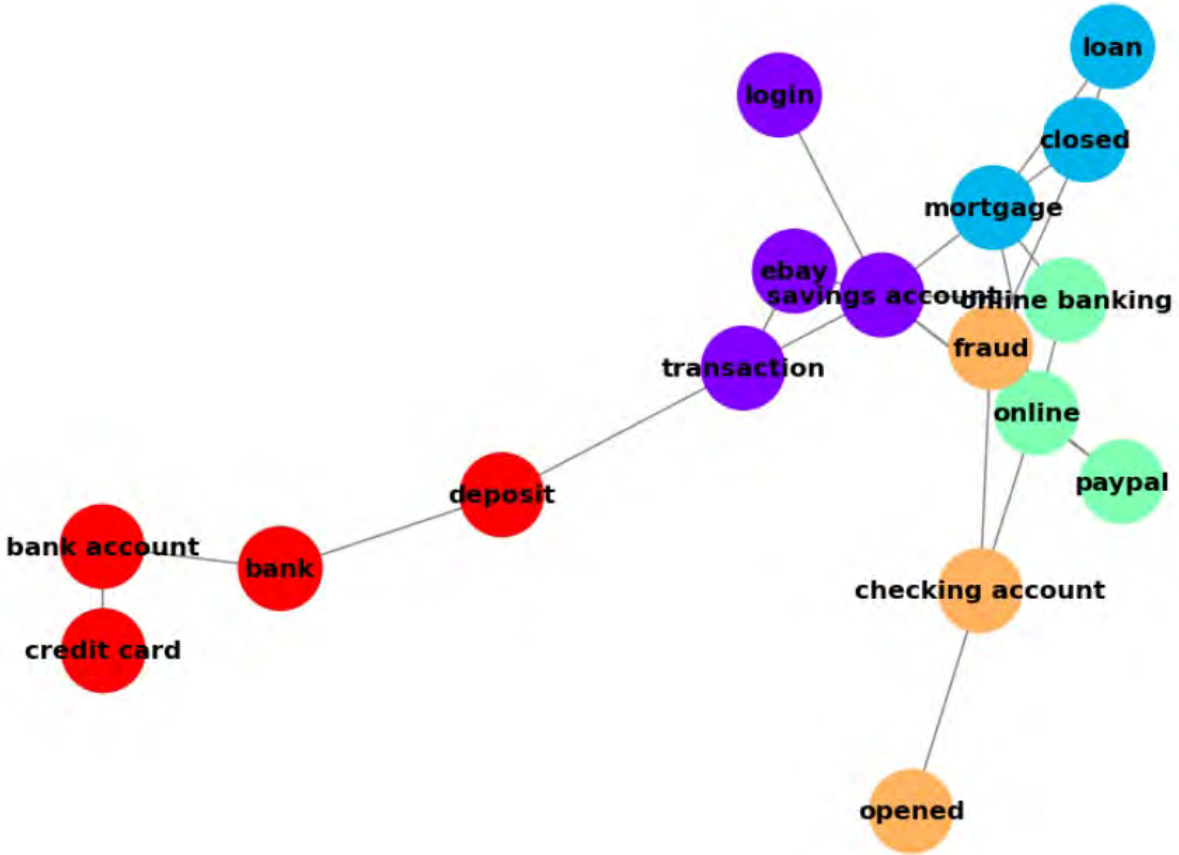
- When the output of gradual supervision is a set of classifiers, we call the classifier output *concepts*.
- Reporting on concepts is straightforward: bar charts, line graphs, etc.
- But what about more complex interactions?

Concept co-occurrence

- Each text response is associated with a set of concepts. This forms a “bipartite graph”. We can “project” these into concept-concept matrices.



Concept co-occurrence networks



Interpreting concept co-occurrence networks

- An edge means that two concepts are related (i.e. many responses are tagged with both concepts).
- Clusters on this graph can reveal more complex relationships that bar charts or group-bys wouldn't!
- If we see tight clusters (e.g. cliques) it's a sign that three separate concepts might be usefully thought of as a group.
  - In some cases!

Conclusions

- Guided supervision
  - Work closely with your SMEs
  - Use the whole gamut of techniques from un- to weakly- to fully-supervised.
- Cluster exploration
  - Important for reducing potential bias from basic topic summarization.
- Refinement
  - Build tight interactive loops with your SMEs to allow them to refine group assignments.
  - Use LLMs to reduce cognitive load on SMEs, but ensure they have access to the underlying data.
- Interactions
  - It’s worth plotting a co-occurrence network if you’ve got a lot of concepts that might be interrelated.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

From text to insights: AI-based topic modeling for central banking applications[1]

Marlena Pietsch, Diana Sola, Felix Marx, Michaela Vollmer and Babak Zandi,
Deutsche Bundesbank

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

DEUTSCHE BUNDESBANK EUROSYSTEM

SCAI Service & Community Center for Artificial Intelligence

From Text to Insights
AI-Based Topic Modeling for Central Banking Applications

Marlena Pietsch, Diana Sola, Felix Marx, Michaela Vollmer, Babak Zandi

4th IFC Workshop on Data Science in Central Banking

Rome, 18-20 February 2025

Motivation
AI Services

Which topics can frequently be found in these documents?



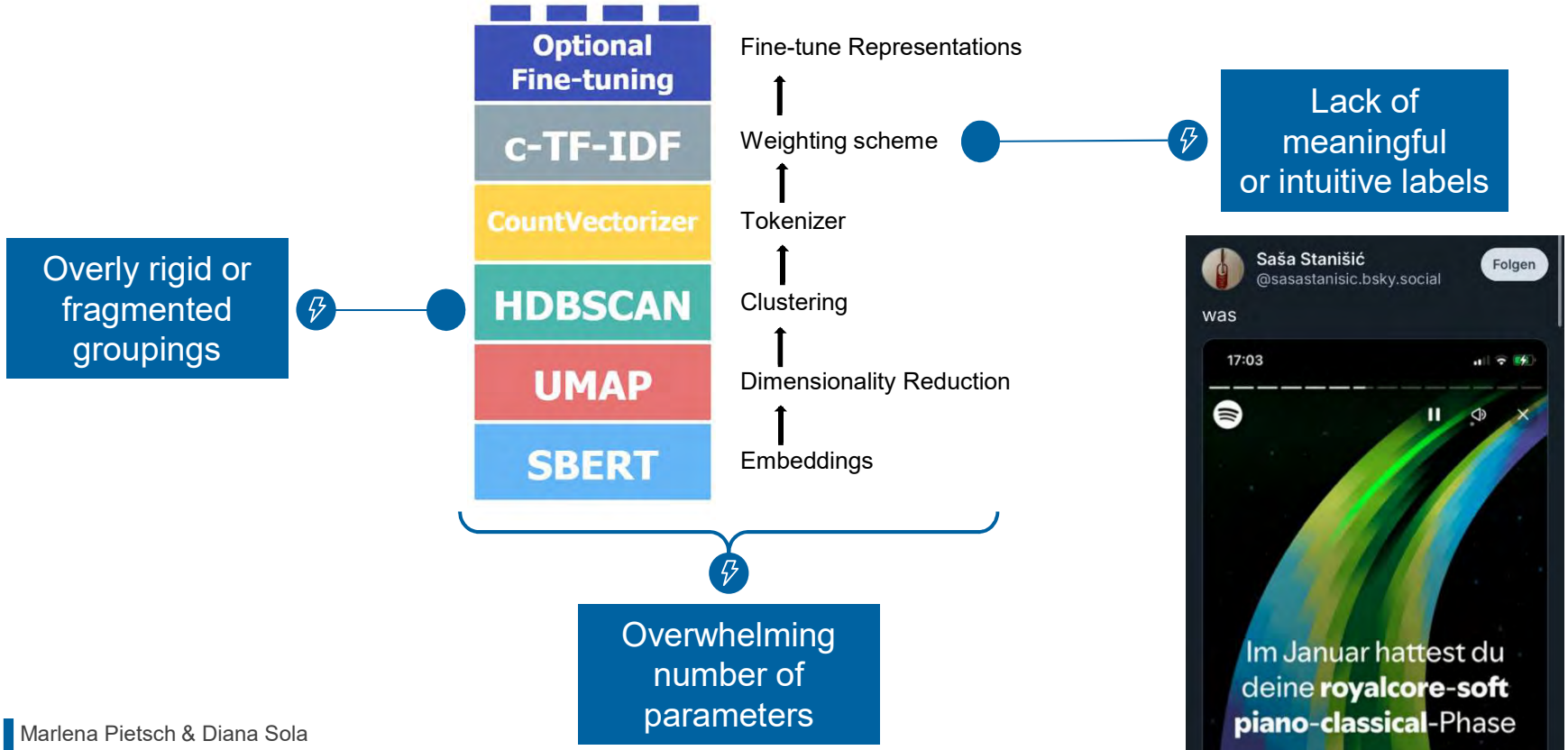
How can we extract structured information from tables in documents?



How can we quickly summarize key points from dozens of pages?



Early steps
Topic modeling with BERTopic



A step beyond traditional methods
Topic Modeling with LLMs

Key benefits of using LLMs for topic modeling:



Contextual understanding of the models



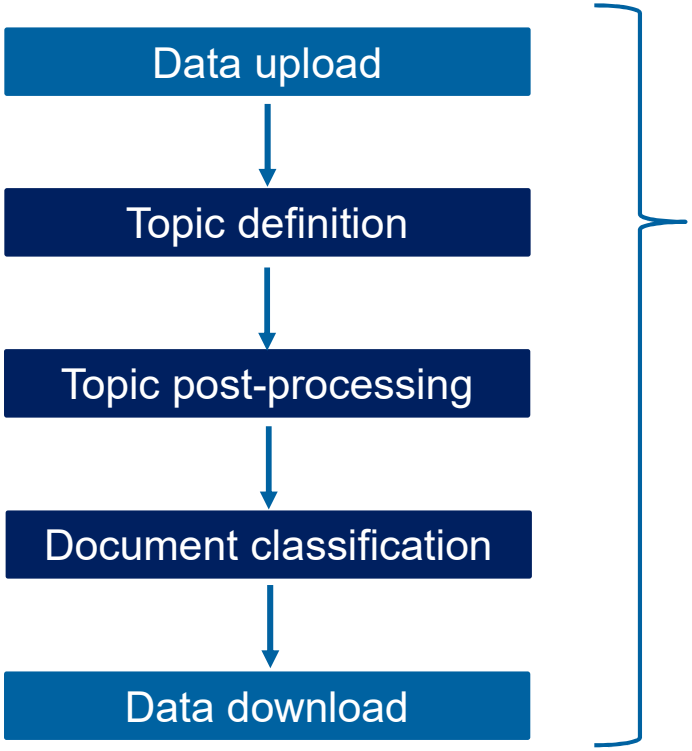
Aligned and human-like topics



Adaptability without hyperparameter tuning

Our Topic Model App bridges LLM capabilities with a user-friendly interface

Topic Model App
From Data to Insights in Action



Emphasis on user-centric design for non-technical audiences

- Step-by-step guidance through the topic modelling process
- Option of adapting the AI's work in every step manually

✓ transparency ✓ debuggability ✓ controllability

The Topic Modeling Process
From Data to Insights in Action

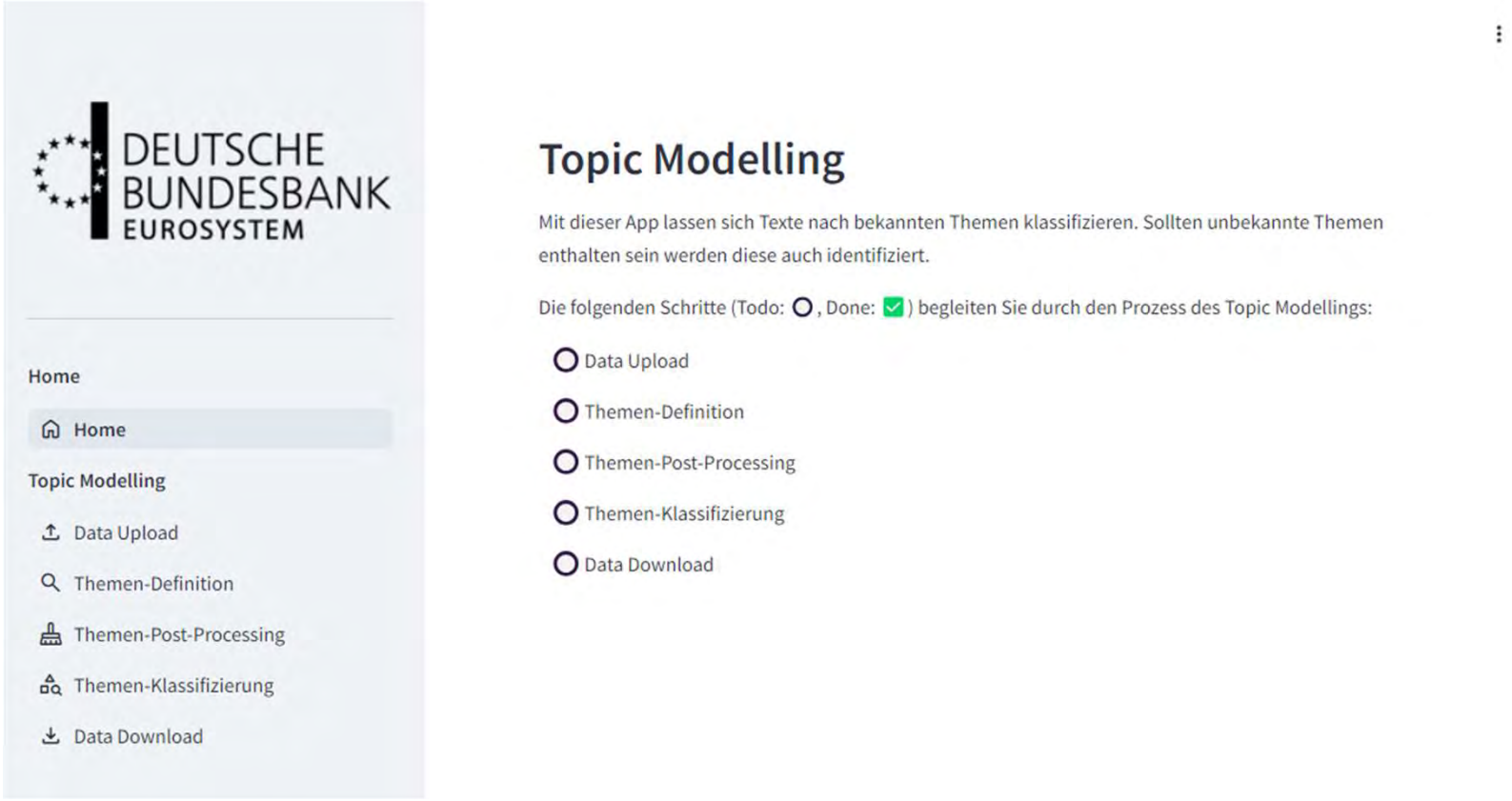
Topic definition → **Topic post-processing** → **Document classification**

- Manually, or
- Uploading pre-defined topics, or
- Few-shot prompting an LLM to identify topics from data

- Adding, deleting and altering topics manually, or
- Few-shot prompting an LLM to cluster subsume similar topics and identify a hypernym

- Few-shot prompting an LLM to assign one or several topics to the given document

Topic Model App Demo



Lessons learned
Challenges & Experiences

Challenges

- Difficulties in evaluating the quality of generated topics
- Scaling to large data inputs
- Aligning model architecture and outputs with domain-specific needs
  - How specific or general do topics need to be to be meaningful?

Experiences

- Importance of interdisciplinary collaboration in app development and understanding of business needs
- Test users preferred pre-written prompts and limited tuning options to the high flexibility in hyperparameter tuning

Outlook
Future Directions for the App

Planned Enhancements

- Exploration of automatic and meaningful evaluation methods
- Enabling hierarchical topic modelling for more complex topic structures
- Integrating the app in existing workflows
- Adding statistics and visualization of model outputs

Thank you!

Contact:

Marlena Pietsch
marlena.pietsch@bundesbank.de

Diana Sola
diana.sola@bundesbank.de

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Efficacy of AI RAG tools for complex information extraction and data annotation tasks: a case study using banks' public disclosures[1]

Nicholas Botti, Flora Haberkorn, Charlotte Hoopes and Shaun Khan,
Board of Governors of the Federal Reserve System

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Efficacy of AI RAG tools for complex information extraction and data annotation tasks: a case study using banks' public disclosures[1]

Nicholas Botti[2], Flora Haberkorn, Charlotte Hoopes, Shaun Khan[3]

Abstract

We utilize a within-subjects design with randomized task assignments to understand the effectiveness of using an AI retrieval augmented generation (RAG) tool to assist analysts with an information extraction and data annotation task. We replicate an existing, real-world annotation task on a set of thousands of pages of public disclosure documents from global systemically important banks (GSIBs). We test two treatment conditions. First, a "naive" AI use condition in which annotators use only the tool and must accept the first answer they are given. And second, an "interactive" AI treatment condition where annotators use the tool interactively, and use their judgement to follow-up with additional information if necessary. Compared to the human-only baseline, the use of the AI tool accelerated task execution by up to a factor of 10 and enhanced retrieval accuracy—particularly in the interactive condition. We find that when extrapolated to the full task, these methods could save up to 268 hours. Additionally, our findings suggest that annotator skill, not just with the subject matter domain, but also with AI tools, is a factor in both the accuracy and speed of task performance.

Keywords: Artificial Intelligence (AI), Retrieval Augmented Generation (RAG), Machine Learning, Data Extraction, Bank Data Analysis, Financial Disclosure, AI-Human Interaction, Human-Computer Interaction (HCI)

JEL Codes: C45, G21, O33, M41, D83, C55, G28

---

[1] The views expressed in this paper are those of the authors, and do not represent the views of the Board of Governors of the Federal Reserve System or the US Department of the Treasury.
We thank Daniel Beltran for his mentorship and useful comments, and participants at the 4th BIS-IFC Workshop on Data Science in Central Banking for feedback and comments. All errors are our own. We thank the authors of the original study referenced for generously allowing us to use their dataset and for their assistance with various aspects of this analysis: Cord Barnes, Daniel Beltran, Kellen Lynch, Monica Sanz, and Pinar Uysal.

[2] Corresponding Author Email: nicholas.botti@frb.gov

[3] Nicholas Botti, Flora Haberkorn, and Charlotte Hoopes are affiliated with the Board of Governors of the Federal Reserve System. Shaun Khan is affiliated with the US Department of Treasury but his work was performed while affiliated with the Board of Governors of the Federal Reserve System.

1. Introduction

Background

Recent advancements in artificial intelligence (AI) tools for document analysis and information extraction, particularly the development of large language models (LLMs) and retrieval augmented generation systems (RAG) have generated intense interest in social science research and among policy analysis. The IMF's 2024 Annual Report cited how AI will transform the work of many industries (IMF 2024). Therefore, it is imperative for us to understand both when and how to use these tools effectively and safely, especially in the realm of economics, finance, and policy analysis.

One particular use case that has drawn considerable interest in policy analysis, is the ability to turn unstructured comprehensive and widely available text into structured datasets suitable for use for further assessment. Using unstructured text data to create a clean structured dataset suitable for financial and economic analysis has historically been a tedious process that starts with manual data collection. While some data vendors already use certain AI tools to extract data from unstructured text documents and make these datasets commercially available, they are often produced with a lag. And this kind of vendor data is not necessarily available for many niche topics. Our study helps bridge this gap by exploring the effectiveness of AI tools for extracting useful information quickly from unstructured data sources, such as public reports and investor disclosure statements by banks.

Policymakers are acutely aware of the need to be able to access information on economic and financial conditions quickly. The example of Silicon Valley Bank (SVB) is instructive. It has been called the first "digital bank run," having taken place over a rapid period, and requiring investors and policymakers to act quickly based on incomplete information (Tett 2023).

Rather than studying a complex data pipeline, we seek to understand how financial analysts can effectively extract this type of information on an ad-hoc basis in a real-world task setting. This is consistent with available information that suggests that thus far the most common way of using these AI tools is as a personal productivity assistant (Maslej, et al. 2025).

Therefore, we set out to study how effective human annotators and analysts are at this type of task, utilizing a within-subjects design to model the use of AI tools as a treatment effect. We investigate the accuracy of this approach, the magnitude of time-savings, as well as determining the effect of other factors on these outcomes.

2. Literature Review

AI as an Assistant and Productivity

Numerous studies have been published focusing on AI and productivity, such as the study done on how introducing a generative AI tool benefited junior software engineers compared to their seniors (Jaffe, et al. 2024) as well as a review done by the BIS that found using LLM's can boost productivity of engineers (Gambacorta, et al. 2024). Investigation into knowledge workers productivity has shown only half the time savings gained as software engineers (Bick, et al. 2025), but doesn't review a variety of factors such as how the worker's background and AI knowledge is a factor in how effective they are in using the tool.

Measuring the Performance of AI Models on Tasks

The literature in artificial intelligence and machine learning is dedicated to studying the performance of LLMs on benchmarks and has continued to struggle to capture the speed of development and domain-specific applicability. ARC (Chollet 2019) and DROP (Dua, et al. 2019) are two of the premier assessment tools used to evaluate the ability for an LLM to gather important information from documents in a manner that emulates research assistance. RAG specific benchmarks such as TAT-QA (Zhu, et al. 2021), cover the effectiveness of finance information retrieval while benchmarking other domains. However, these benchmarks and others often fail to capture the complexities of real-world tasks, particularly in the economics and financial policy domains. Even with the introduction of agents, few benchmarks have emerged looking at making economic-based choices such as EconEvals (Fish, et al. 2025). However, such benchmarks often also focus on automated analysis, rather than the effects of pairing a human analyst with these tools, leaving them limited from a qualitative scope.

In our study, we investigate a real-world information extraction and data annotation task previously completed by human analysts. By replicating an existing, completed task, we can create a robust evaluation of the impact of introducing these AI tools on this specific workflow. Additionally, we structure our evaluation as a within-subjects design, allowing us to understand the effect of introducing use of the AI tool as a treatment effect compared with a control human-only approach, all while controlling effectively for per-annotator skill.

3. Methodology

3.1 Case study using banks' public disclosures.

We utilized the annotated dataset compiled by Beltran et al. (2023), which was created through a data collection process conducted by human annotators in 2023 to extract diverse information from bank documents. The goal of the collection was to extract information from the public and investor disclosure documents of global systemically important banks (GSIBs), and create an annotated dataset of GSIB policies, commitments, and progress on certain key metrics. This unified dataset across banks would allow researchers to make cross-bank comparisons and better understand industry trends over time. This dataset was compiled using a human only approach, so it offers a natural benchmark to compare to our RAG approach.

Human annotators examined thousands of pages of documentation per bank, across annual reports, non-financial disclosures, and specialized bank reports for each financial institution. The targeted information ranged from banks' broad investment goals to banks exact dollar exposure to certain sectors. The complete dataset included 150 questions for each of the 29 GSIBs, a mix of true/false questions, categorical assessments of policies and activities, and exact figure questions.

This task makes for an excellent case-study in assessing the impact of exposing analysts to AI tools to help complete their work. It is a complex, real-world task in the financial analysis domain, not a multiple choice question benchmark. The questions often required combining information from multiple different sources within the documents to produce a final answer. Additionally, because of heterogeneity in how banks disclose information, we can assess the efficacy of the AI-assisted approach under incomplete information. Because the documents do not always contain complete information to answer all the

questions for every single bank, we were able to assess whether AI reacted appropriately in scenarios when information wasn't available by correctly refusing to answer.

3.2 Study Design

This study employed a mixed experimental design to evaluate the impact of an Artificial Intelligence (AI) Retrieval-Augmented Generation (RAG) tool on the performance of human annotators performing a complex information extraction task using public disclosure documents from Global Systemically Important Banks (GSIBs). The core experiment utilized a within-subjects design, where specific questions within each bank annotation task were randomly assigned to either a 'Control' condition or an 'AI' (naive use) condition for each participant. This allowed for comparing the performance under each condition while controlling for individual annotator differences.

A subsequent repeated-measures phase was conducted later, where participants revisited the questions previously assigned to the 'AI' condition, this time operating under an 'AI+' (interactive use) condition. This allowed for a direct comparison of naive versus interactive AI tool usage on the same set of questions while controlling for recall effects.

3.3 Participants

Three professional financial analysts participated in the study. Participants were recruited via a convenience sample of volunteers within the researchers' organization and completed the experimental tasks during their regular work hours. Prior to the experiment, participants completed a multiple-choice quiz designed to assess their baseline familiarity with both the subject matter (GSIB disclosures and related finance topics) and AI tools. These baseline skill assessments are used descriptively and for subsequent analysis exploring potential interactions between skill profiles and condition effects.

3.3 Materials

Bank Documents: The materials consisted of comprehensive sets of public disclosure documents retrieved from bank websites and formatted as pdf files (e.g., annual reports, non-financial reports, regulatory filings) for a sample of nine European GSIBs. These document sets were identical to those compiled and then used in the original analysis by Beltran et al. (2023) and were provided by the original research team, ensuring consistency and relevance.

Annotation Questions: A standardized set of 30 questions was used for all banks and participants. These questions were selected in consultation with the original Beltran et al. (2023) research team to be representative of the full 150-question set used in their study. The selection aimed for diversity across question complexity (high: 9, medium: 11, low: 6), required answer format (categorical: 9, Yes/No: 10, exact figure/value: 7), and thematic category (risk management: 11, financed activity: 10, oversight rules: 3, external project ratings: 2). Each question included detailed instructions and definitions identical to those used by Beltran et al. (2023).

3.4 Experimental Conditions

Participants completed annotation questions under three distinct conditions:

1. **Control Condition:** Participants answered assigned questions using only the provided PDF documents and a standard PDF reader. They were permitted to use the reader's built-in text search ("find") functionality but were explicitly prohibited from using any other external tools, web searches, or AI assistance. The focus was on manual information retrieval and synthesis based solely on the source documents.
2. **AI (Naive) Condition:** Participants answered assigned questions using a provided AI RAG chat tool pre-loaded with the relevant bank's document collection. Participants were instructed to formulate a single prompt based on the annotation question and submit it to the tool. They were explicitly instructed to record their final answer based *only* on the AI tool's initial response, without consulting the original PDF documents. This condition aimed to simulate a naive or high-trust approach to using the AI assistant.
3. **AI+ (Interactive) Condition:** This condition was implemented in the follow-up phase for questions previously tackled under the 'AI' condition. Participants initiated the process identically to the 'AI' condition (formulate single prompt, review AI response). However, they were then given discretion: based on their judgment of the initial AI response, they could choose to (a) ask additional follow-up questions within the AI chat tool, (b) consult the original PDF documents, or (c) do neither or both, before recording their final answer. This condition aimed to simulate a more interactive, discerning, and potentially iterative approach to using the AI assistant.

AI RAG Tool Specifications: The same AI RAG tool configuration was used for both the 'AI' and 'AI+' conditions. It was implemented using OpenWebUI (version 0.3.2) as the interface. The backend utilized Claude 3.5 Sonnet (model version 20240622) via Amazon Bedrock as the Large Language Model (LLM) with default sampling parameters. Vector embeddings were generated using Amazon Titan Text Embedding v2 via Amazon Bedrock. For the RAG process, documents were chunked using a size of 2000 characters with a 200-character overlap. The retrieval mechanism used a "top_k" setting of 20, meaning the 20 most relevant chunks were retrieved to inform the LLM's response. This setup was chosen to represent a commonly accessible, though not state-of-the-art, RAG implementation suitable for typical researcher workflows.

3.5 Procedure

Participants received initial training covering the task background, the experimental procedures for each condition, instructions on using the AI RAG tool interface, and guidelines for self-reporting time. The detailed instructions embedded within each annotation question served as the primary guide for content interpretation.

For each assigned bank, the 30 questions were presented in the same fixed order. Prior to the task, each question was randomly assigned (using a computer script with a random number generator, resulting in an approximately 50/50 split) to either the 'Control' or 'AI' condition for that specific annotator-bank instance. Randomization was not stratified by question characteristics.

Participants performed the tasks at their own discretion regarding location (e.g., office, remote). For each question, participants self-reported their start and end time (wall clock time) to the nearest minute. Durations under one minute were recorded as 0.5 minutes, based on observational validation for a subset of tasks.

The initial experiment (Control vs. AI) was completed first. The follow-up experiment (AI+) was conducted after a minimum delay of two weeks for all participants (up to three months for two participants) to mitigate potential recall effects from the initial phase.

Bank assignments were as follows: Annotator A completed tasks for six banks, Annotator B for three banks, and Annotator C for four banks. Annotator C's assignments included two banks also completed by Annotator A and two banks also completed by Annotator B, creating deliberate overlap to enable the assessment of inter-annotator agreement.

3.6 Outcome Measures

Time Efficiency: Measured as the self-reported time taken per question (in minutes) for each condition.

Accuracy/Agreement: The primary measure of answer quality. Annotator responses for all conditions were compared against the final, vetted answers established by the Beltran et al. (2023) research team. Although the realistic nature of this task implies that there are no known true values, given the thoroughness of the original research team's work, we consider comparison against these values as an acceptable proxy for task accuracy. A member of the original team performed this agreement assessment, consulting with other original team members for complex cases. For each answer, agreement was categorized as "exact match," "substantially similar," "unclear," or "incorrect." For the main analysis, "exact match" and "substantially similar" were coded as Agreement = True (representing acceptable answers based on the original team's criteria), while "incorrect" was coded as Agreement = False. Answers rated "unclear" (20 answers, roughly 4% of total answers) were excluded from the accuracy analysis. Robustness checks that did not count "substantially similar" as agreement yielded broadly similar results.

AI+ Interaction Metrics: For the AI+ condition only, participants recorded (a) the number of distinct follow-up messages sent to the AI tool after the initial prompt, and (b) whether they chose to consult the original PDF documents (Yes/No) before finalizing their answer.

3.7 Ethical Considerations

Informed consent was obtained orally from all participants prior to their involvement in the study. Participants were informed about the study's purpose, procedures, data handling, and the voluntary nature of participation. Data collected, including annotator responses, timing, and quiz scores, were de-identified before the main statistical analysis to protect participant confidentiality.

4. Results

4.1 Participant Characteristics

Prior to the study, participants were evaluated on a short assessment. The assessment consisted of two parts. The first portion was devoted to their skill and understanding of AI - assessed by questions about prompt techniques, characteristics of different models, understanding of the fundamentals of RAG, knowledge about hallucinations, etc. The second portion measured their understanding of the subject matter domain of the documents - regulatory frameworks, bank policies, balance sheet considerations, etc.

Participant Skill by Annotator

Table 1

Participant	Subject Matter Skill (%)	AI Skill (%)
A	68.8	52.2
B	68.8	82.6
C	87.5	21.7

As shown in Table 1 all three participants had generally high scores for their understanding of the subject matter domain, consistent with their roles, but their scores on the AI portion of the assessment varied much more, consistent with anecdotal questions we asked participants about their experience with AI tools. Participant B and C provide the largest contrast between their skill sets, while participant A is more in the middle.

4.2 Question Characteristics

The questions used in the task are generally well balanced, and are a representative subset of the full 150 question dataset produced by Beltran et al. 2023. Tables 2, 3, and 4 below outline the makeup of questions tested by annotators over the course of different treatments (a total of 494 question-answer pairs).

Question Formats

Table 2

Format	Count	Percentage
Yes/No	192	38.9
Categorical (cat)	178	36.0
Numerical FIgure	124	25.1

Question Complexity

Table 3

Complexity	Count	Percentage
Low	114	23.1
Medium	209	42.3
High	171	34.6

Question Categories

Table 4

Category	Count	Percentage
Risk Management	204	40.9
Financing Activities	199	40.3
Oversight	57	11.5
External Organizations	26	7.3

4.3 Time Efficiency

Time to Complete One Question, in Minutes

Table 5

Condition	Obs	Mean	SD	Median	Min	Max
Control	153	4.0	2.7	4.0	0.5	15
AI	168	0.3	0.3	0.5	0.5	1
AI+	173	1.2	1.2	1.0	0.5	6

AI assistance reduced task completion time by 90% in the naive AI condition and 70% in the interactive AI+ condition, with minimal variance. It should be noted that this time to answer the question represents a first-pass or "best effort" attempt to answer the question. While we do not have exact figures, the original research team reports that they spent on average significantly longer than this per question before arriving at their final results - such as by incorporating multiple levels of review and discussing some questions as a group.

If these numbers generalize to the full task (150 questions per bank, 29 banks), we would estimate the following time savings:

Time Saving for Complete Task (Estimated)

Table 6

Condition	Mean time per Question (mins)	Total Time (Estimated Hours)	Total Savings vs Control (Estimated Hours)
Control	4.0	290	-
AI	0.3	21.75	268.25
AI+	1.2	87	203

Time Efficiency by Annotator

Table 7

Participant	Obs	Mean	SD	Median	Min	Max
A	227	2.2	2.8	0.5	0.5	15
B	118	2.0	2.0	1.0	0.5	12
C	149	1.2	1.0	1.0	0.5	5

[1] Time reported in Minutes.

Average Time Taken per Task by Participant by Condition

Table 8

Participant	Obs	Control	AI	AI+
A	227	5.5	0.5	1.0
B	118	3.9	0.6	1.8
C	149	1.9	0.7	1.2

[1] Time reported in Minutes. Average time calculated with simple mean.

Participant C was the fastest overall, with the lowest mean and median times, possibly due to stronger subject matter knowledge (87.5%) helping streamline the annotation process. Note the high variation in task time, ranging from 0.5 to almost 15 minutes depending on the question. Also note the unusually high amount of time taken by Participant B during the interactive AI condition, perhaps indicating their high level of experience with AI tools allowed them to more effectively take advantage of the capability.

4.4 Accuracy (Agreement)

Agreement Rates by Condition

Table 9

Condition	Total Responses	Agreements	Agreement Rate (%)
Control	153	81	52.9
AI	168	99	58.9
AI+	173	114	65.9

Descriptively, accuracy increased somewhat with AI use. The highest agreement rate occurred in the AI+ condition, suggesting that an interactive approach allows annotators to better assess and refine AI outputs.

Agreement by Question Type and Condition

Table 10

Format	Control(%)	AI (%)	AI+(%)	AI+ vs Control (pp)
Yes/No	62.5	76.6	84.4	+21.9
Categorical	29.2	35.9	40.9	+11.7
Numerical	65.9	67.5	76.7	+10.8

Unsurprisingly, annotators performed best on the True/False questions, and worst on the categorical questions, which often required tricky subjective judgement in how to assess whether the information met complex criteria. We see the greatest increase in performance with the AI+ method on these simpler yes/no questions.

Agreement by Complexity and Condition

Table 11

Complexity	Control(%)	AI(%)	AI+ (%)	AI+ vs Control (pp)
Low	66.7	76.3	90.0	+23.3
Medium	61.3	69.9	73.0	+11.7
High	34.5	33.3	40.7	+6.2

Consistent with the findings from the question type, using the AI tool had the strongest positive effect on lower complexity questions. For higher complexity questions, it still improved accuracy somewhat, but much less so. Note the large variation in performance, consistent across conditions, between the low/high complexity questions, validating the effectiveness of the complexity assessments.

Agreement by Participant

Table 12

Condition	Total Responses	Agreements	Agreement Rate (%)
A	227	123	54.2
B	118	72	61.0
C	149	99	66.4

Participants' overall accuracy scores were generally consistent with their evaluations of subject matter expertise. But this masks strikingly high variation when breaking down by experimental condition.

Agreement by Participant and Condition

Table 13

Participant	Control(%)	AI(%)	AI+ (%)	AI+ vs Control (pp)
A	48.6	53.2	60.3	+11.7
B	33.3	68.3	75.0	+41.7
C	72.9	60.0	66.7	-6.2

Annotator B, who had by far the highest experience working with AI tools, was able to more than double their chance of correctly answering the question when using the AI tool, despite performing the worst in the control condition. Meanwhile Annotator C, who had the highest subject matter expertise but very little experience with AI tools, actually performed somewhat worse under the treatment conditions.

While the sample size here is clearly insufficient (three annotators) to generalize these findings, this is a very interesting result nonetheless that we believe should motivate future research.

5. Statistical Analysis

5.1 Accuracy

We perform a logistic regression on the binary variable representing whether an answer (observation) correctly agrees with the findings of the original research team, per the agreement methodology discussed previously. We control for both the individual baseline of each annotator and the question complexity. We use the "Control" condition, annotator A, and medium complexity as the baseline. Dummy variables for the AI and AI+ conditions, annotators B and C, and complexity of high or low are included in the regression.

Our hypothesis for this experiment was that there would be a considerable increase in speed under the AI conditions, particularly the naive AI condition, but that there would be some loss of accuracy. We hypothesized that the key question would be how large that loss of accuracy would be, and under what conditions that loss of accuracy might be acceptable for certain tasks

We find that under the naive AI condition, accuracy increases, but is not statistically significant whereas under the interactive AI condition, accuracy increases significantly (p=0.014). High question complexity, and Annotator C both significantly decreased and increased accuracy (respectively).

$$log(P(Agreement_{\mathrm{n}} = 1) / (1 - P(Agreement_{\mathrm{n}} = 1))) = \beta_0 + \beta_1 \cdot Condition_AI_{\mathrm{i}}$$
$$+ \beta_2 \cdot Condition_AIplus_{\mathrm{i}} + \beta_3 \cdot ParticipantID_B_{\mathrm{i}} + \beta_4 \cdot ParticipantID_C_{\mathrm{i}}$$
$$+ \beta_5 \cdot Complexity_Low_{\mathrm{i}} + \beta_6 \cdot Complexity_High_{\mathrm{i}}$$

Logistic Regression on Agreement_Boolean

Table 14

Variable	Human vs AI vs AI+ n = 494
Intercept	0.2554 (0.231)
Condtion_AI	0.2557 (0.243)
Condition_AI+	0.6010* (0.246)
Participant_B	0.2484 (0.249)
Participant_C	0.5655* (0.236)
Complexity_Low	0.5257 (0.274)
Complexity_High	-1.3483*** (0.221)

[1] Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05

Additionally, we also investigate whether checking the original documents or changing the number of messages to send to the AI affects the accuracy during the interactive (AI+) phase. Filtering to the AI+ (interactive) condition only, we include a dummy variable for whether the original documents were checked or not, as well as the number of messages sent to the AI tool before recording a final answer to the question.

Logistic Regression on Agreement_Boolean, Within AI+ Condition Only

Table 15

Variable	AI+ (n = 173)
Intercept	1.1349**
	(0.05400)
Doc_Check	0.0466
	(0.396)
ParticipantID_B	0.9713
	(0.547)
ParticipantID_C	0.2191
	(0.411)
Question_Complexity_low	1.2258*
	(0.597)
Question_Complexity_high	-1.2821***
	(0.385)
Num_Messages	-0.3787
	(0.269)

[1] Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05

We find that neither the number of messages sent and whether the analyst checks the original documents are statistically significant, which lends to multiple interpretations. This could indicate that these additional measures weren't helpful, while on the other hand this could be a selection effect. Thus, indicating that the analysts are correctly using their judgement to apply these measures only to harder questions instead of leveraging the AI for tasks they can complete confidently. We consider the latter the more straightforward and likely explanation, but further work could explore setting this as an explicit experimental condition. Analysts checked the original document 24% of the time for low complexity questions, but 38% and 39% of the time for medium and high complexity questions respectively, giving at least some indication that their judgement of when these additional steps were necessary was correct.

5.2 Time per Task

We then run a generalized linear mixed model (GLMM) to evaluate the effects of AI on time spent to answer a question. We include a gamma distribution and a log link to address the positive and right-skewed characteristics of our time data.

$$log(E[Time_{it}]) = \beta_0 + \beta_1 Condition_i + \beta_2 Question_Complexity_i + \beta_3 ParticipantID_B_i$$
$$+\beta_4 ParticipantIDC_i + \beta_5 Complexity_Med_i + \beta_6 Complexity_High_i + u_i$$

Our results suggest that AI assistance significantly reduces the time required for annotation. Complexity also has a marginal effect, implying that more complex questions may demand more time. Additionally, we notice individual differences among annotators. The results show that annotator B spent significantly more time answering the questions. This could be attributed to the annotator's lower institutional knowledge on the subject or their higher skill with AI tools leading them to use slower but more advanced techniques.

GLMM Regression on Time_Minutes

Table 16

Variable	Human vs AI vs AI+ (n = 494)
Intercept	1.0629***
	(0.05400)
Condtion_AI	-1.65987***
	(0.05824)
Condition_AI+	-0.96922***
	(0.05809)
Participant_B	0.21772***
	(0.05907)
Participant_C	-0.03431
	(0.05513)
Complexity_Low	-0.12708*
	(0.06003)
Complexity_High	0.07441
	(0.05307)

[1] Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05

6. Conclusion

AI RAG tools show significant promise in helping financial analysts with extracting and categorizing information in lengthy sets of financial documents. Even when used in a naive fashion using relatively simple RAG architecture, we see no statistically significant decrease in task accuracy, while when used carefully and with judgement, we see a statistically significant increase in task accuracy. In both cases the time savings are significant, up to an order of magnitude. We find some evidence of the importance of training analysts how to use these tools most effectively, and that they may struggle with the most complex questions. Due to low sample size though, further research is needed into these areas.

Despite improvements in accuracy vs the control however, it is important to note that we only measure here first-pass accuracy. Both in the control and AI treatment methods, accuracy still fails to breach 80% of what could be done with a more thorough team process involving multiple levels of review and verification. This underscores the importance of verifying critical information before making important decisions, with or without AI.

References

Bank for International Settlements (BIS) (2024): "Generative AI and labour productivity: a field experiment on coding", BIS Working Papers, no 1208, September.

Beltran, D O, H Bensen, A Kvien, E McDevitt, M V Sanz, and P Uysal (2023): "What are large global banks doing about climate change?", International Finance Discussion Papers, no 1368, Board of Governors of the Federal Reserve System, January.

Bick, A, A Blandin, and D Deming (2025): "The rapid adoption of generative AI: productivity effects and time savings across occupations", NBER Working Paper Series, no 32966, February

Chollet, F (2019): "On the measure of intelligence", arXiv preprint, arXiv:1911.01547.arxiv

Dua, D, Y Wang, P Dasigi, G Stanovsky, S Singh, and M Gardner (2019): "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs", arXiv preprint, arXiv:1903.00161.

Fish, S, J Shephard, M Li, R Shorrer, and Y A Gonczarowski (2025): "EconEvals: benchmarks and litmus tests for LLM agents in unknown environments", arXiv preprint, arXiv:2503.18825.

Gambacorta, L, H Qiu, S Shan and D M Rees (2024): "Generative AI and labour productivity: a field experiment on coding", BIS Working Papers, no 1208, September.

International Monetary Fund (IMF) (2024): "Gen-AI: Artificial intelligence and the future of work", Staff Discussion Notes, no 2024/001, January. IMF eLibrary

Jaffe, S, N P Shah, J Butler, A Farach, A Cambon, B Hecht, M Schwarz, and J Teevan (eds) (2024): "Generative AI in real-world workplaces: The second Microsoft report on AI and productivity research", Microsoft Research, July.

Maslej, N, L Fattorini, R Perrault, Y Gil, V Parli, J Manyika, J Etchemendy, K Ligett, T Lyons, and Y Shoham (2025): "Artificial Intelligence Index Report 2025", Stanford Institute for Human-Centered Artificial Intelligence, April.

Tett, Gilian (2023): "Wake up to the dangers of digital bank runs" Financial Times 4/20/23

Zhu, F, W Lei, Y Huang, C Wang, S Zhang, J Lv, F Feng, and T-S Chua (2021): "TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance", arXiv preprint, arXiv:2105.07624.

Efficacy of AI RAG Tools for Complex Information Extraction and Data Annotation Tasks on Bank Documents

Nicholas Botti (Presenter), Flora Haberkorn, Shaun Khan

Board of Governors of the Federal Reserve System

The views expressed in this presentation are solely those of the authors and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System or the Federal Reserve System.

Introduction

- There is tremendous interest in utilizing Large Language Models (LLMs) and related techniques such as retrieval augmented generation (RAG) to extract information from financial documents and other types of unstructured data
- Some general benchmarks exist for this type of task, such as RAGAS (Es et. al. 2023) or needle in a haystack (Kamradt 2023)
- We apply these techniques to a **challenging, real-world information extraction task** on bank documents that took multiple human annotators months to complete (Beltran et. al. 2023)

Structured Data Annotation Task

- Using dataset created by (Beltran et. al. 2023)
- Large global banks report information about their actions, policies, commitments, etc. relating to various topics in their annual reports and other public disclosure documents
- However, because there is no structured dataset of this information, it is difficult to look at industry-wide trends, etc.
- Task objective: Have annotators review these unstructured text data sources, and qualitatively code variables of interest to create a high-quality dataset for researchers

Note: Information obtained during this project, or during the original research paper, is not used for supervisory purposes by the Federal Reserve Board

A Unique Challenge for AI

- ~1000-2000 pages worth of documentation to review per bank
- Complex multi-part questions require combining/cross-referencing multiple sections of text
- Heavy usage of finance domain-specific terminology
- ~150 questions to answer per bank.
- Not all questions have answers in the provided documents.
- Some questions require subjective interpretation.
- **Research goal: understand how effectively AI tools can assist human annotators on this type of complex task**

Experiment Methodology

- Select a representative subset of 30 questions from the original task, and select a subset of nine European banks
- Randomly assign half of the questions (15) to the control group, and half (15) to the treatment group for each bank
- Use a pre-selected group of documents for each bank
- Annotators given general background on the task, but no extensive training. Annotators recorded time to answer each question
- Control group: annotators answer the questions using only the provided documents
- Treatment group: annotators given access to an AI tool, pre-populated with the set of documents for their bank. They answer based ONLY on the information provided by the tool, and are instructed not to look at the documents themselves

AI Tool Usage Procedure

- Annotators were given access to an AI document chat interface
- Wrote their own prompts based on the question
- Used their judgement to qualitatively code and record answer according to provided criteria
- Encouraged to send only one prompt per question
  - Could ask the tool up to one additional follow-up question if necessary

AI Tool Architecture

- Used a basic retrieval augmented generation (RAG) approach
  - Embedding Model: Amazon Titan Embed Text v2 (1024 dimension). 2000-character passages with 100-character overlap.
  - Top 20 most relevant passages sent as context along with prompt to LLM
  - LLM: Claude 3.5 Sonnet (20240620 version) via AWS Bedrock
- Note: This is not a state-of-the-art RAG approach
  - Objective of this research was to establish the feasibility of AI tools on this type of task using tools commonly available to researchers
  - Models and architecture consistent for all questions

Determining Results

- We compare the annotations from our experiment with the final values used in the original research teams report (“original values”) and measure “agreement” with those values
  - 100% agreement unlikely – many questions require subjective judgement
  - Can’t measure “accuracy” directly – no known true values to compare
- Original values based on multi-step process: initial annotation by multiple different annotators, multiple review phases, etc.
  - Therefore, the original values provide a reasonable proxy for accuracy
- NOTE: For about 20% of the questions, we were able to find an answer for question original team marked as “unknown”
  - Further work will consider how to best evaluate these, excluded from current analysis
  - Could indicate even higher accuracy, or could indicate hallucinations

Research Questions

- Does utilizing the AI tool save time compared with manual annotation?
- Are annotators less accurate when using an AI tool, and if so by how much?
- Are there any patterns to accuracy/time savings in the types of questions/tasks that can help inform other researchers considering using AI tools to assist with data annotation and information extraction tasks?

Results

Mean Time per Question

Overall

Condition	Time per Question	Sample Size (n)
Control (Human Only)	4.90 minutes	91
Treatment (AI Tool)	**0.54 minutes*** **	106
Time Savings	**89%**	

By Annotator

	Annotator A	Annotator B
Control	5.28 minutes	4.00 minutes
Treatment	**0.50 minutes*** **	**0.625 minutes*** **
Time Savings	**91%**	**84%**
Sample Size (n)	134	63

Key Findings:

- As expected, using the AI tool produced strongly significant time savings
- A nearly 10x reduction in the amount of time taken to complete each question
- Some difference between the two annotators
- Subjectively, annotators reported completing the task with the tool assistance as being much more "pleasant"

* p<0.10, ** p<0.05, *** p<0.01. Mann-Whitney U test, control vs treatment. Time recorded to the nearest 30s.

Time Taken for Annotation (Per Question)

By Question Complexity

Condition	Low	Medium	High
Control	4.31 minutes	5.00 minutes	5.06 minutes
Treatment	**0.50 minutes*** **	**0.53 minutes*** **	**0.58 minutes*** **
Time Savings	**89%**	**89%**	**89%**
Sample Size (n)	40	85	72

Key Findings:

- More complex questions took somewhat longer using both the control and treatment methods, but the relative time savings remained consistent.

* p<0.10, ** p<0.05, *** p<0.01. Mann-Whitney U test, control vs treatment. Time recorded to the nearest 30s.

Accuracy (Agreement with "Original Values")

Overall

Condition	Accuracy	Sample Size (n)
Baseline (Random Guess)	27.5%	197
Control (Human Only)	43.9%	91
Treatment (AI Tool)	**61.3%****	106

By Annotator

Condition	Annotator A	Annotator B
Control	50.0%	29.6%
Treatment	**57.1%**	**69.4%*** **
Sample Size (n)	134	63

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test, control vs treatment

Key Findings:

- Using the tool significantly **improves** accuracy, an unexpected result
- Annotator A had some subject matter expertise with the task/domain, but little experience using AI tools. Annotator B was the opposite, having less experience with the task/domain, but considerable experience using AI tools
- Annotator skill level had a much larger impact than expected. Annotator B, despite performing much worse at the control task, performed considerably better when using the tool.

Accuracy (Agreement with "Original Values")

By Question Complexity

Condition	Low	Medium	High
Control	56.2%	50.0%	30.3%
Treatment	**83.3%***	**72.1%****	**35.9%**
Sample Size (n)	40	85	72

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test

Key Findings:

- Using the tool significantly improved accuracy for low/medium complexity question, but not for the most complex questions.
- Accuracy was low for both the control and treatment groups on high complexity questions.
- These questions tended to require the most subjective judgement, and could be explained by more limited contact with the original research team or lower annotator skill
- However, these questions also often required referring to many different portions of the documents and synthesizing complex conclusions. This may indicate a limitation of the fairly basic RAG approach used

Accuracy (Agreement with "Original Values")

Low/Med Complexity Questions Only

Condition	Annotator A	Annotator B
Control	54.5%	42.9%
Treatment	**76.2%****	**76.0%***
Sample Size (n)	86	39

High Complexity Questions Only

Condition	Annotator A	Annotator B
Control	40.0%	15.4%
Treatment	**28.6%**	**54.5%***
Sample Size (n)	48	24

Key Findings:

- Given the limited sample size, difficult to draw strong conclusions about the high complexity questions
- Annotator A performed worse with access to the tool on high complexity questions (though not statistically significant), while Annotator B performed significantly better.
- This could indicate that skill using AI tools is especially important on more complex data annotations (vs not as important on less complex ones)

* p<0.10, ** p<0.05, *** p<0.01. Fisher's exact test

Conclusions

- In highly complex, challenging data annotation tasks, even a basic RAG approach was able to substantially **improve** the performance of untrained human annotators, an unexpected result
- Allowing human annotators to use AI tools reduced task time by a factor of 10x
  - This could allow researchers to assemble data annotations for much larger datasets than would otherwise be feasible
- We find some evidence that annotator skill/experience with AI tools is a key factor in their effectiveness
  - Researchers should consider incorporating training for annotators if they intend to utilize these tools
- We find some evidence tool use may be less effective on extremely high complexity tasks, but further research is needed in this area

Limitations/Further Work

- Accuracy comparison is limited due to no known true values
  - Further work should include a detailed qualitative review of a sample of non-agreement questions to attempt to discern true values
- Limited sample size
  - Human annotation is cost prohibitive, the motivation for this research
- RAG approach used is simplistic, and only 1-2 prompts used per question
  - Further work should evaluate newer and more robust approaches
- Line between AI and human judgement is blurred
  - Further work should include a 100% AI approach as comparison
  - Further study is needed for the effect of annotator skill with AI tools

References

- Beltran, Daniel O., Hannah Bensen, Amy Kvien, Erin McDevitt, Monica V. Sanz, and Pinar Uysal (2023). “What are Large Global Banks Doing About Climate Change?,” International Finance Discussion Papers 1368. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2023.1368.
- Es, Shahul, Jithin James, Luis Espinosa-Anke, Steven Schockaert (2023). “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv. https://doi.org/10.48550/arXiv.2309.15217
- Kamradt, Greg (2023). “Needle in a Haystack Analysis,” X, https://x.com/GregKamradt/status/1722386725635580292

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

BankGPT: the use of large language models in official communications[1]

Claudia Biancotti, Carolina Camassa, Marco Fruzzetti,
Luigi Palumbo and Myriam Portaluri,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

BankGPT: the use of large language models in official communications

Claudia Biancotti, Carolina Camassa, Marco Fruzzetti, Luigi Palumbo, Myriam Portaluri

Abstract

This study investigates the summarization performance of two prominent large language models (LLMs) – *GPT-3.5 turbo* and *LLaMA 2* – by evaluating their ability to generate summaries of the Bank of Italy's *Economic Bulletin*, a standardized institutional publication. The analysis compares model-generated summaries with the official human-produced summary, using evaluations from domain experts within the Bank's Directorate General for Economics, Statistics and Research. Employing the Bradley-Terry model for pairwise comparisons, the study assesses summaries across three criteria: fluency, relevance, and consistency. Results indicate no overall preference between human and machine-generated summaries; however, experts with greater familiarity with the *Bulletin* and higher educational attainment showed a marked preference for the official summaries. In addition, preferences correlated with demographic factors, with women and individuals over 35 more likely to prefer human-authored summaries. These findings highlight both the strengths and limitations of LLMs for tasks requiring specialized domain knowledge and audience alignment.

Keywords: Large Language Model (LLM), central banks, Bradley–Terry model, logistic regression.

JEL classification: O39, Z19, A19, C88, H89

1. Introduction

Large language models (LLMs) are artificial intelligence systems trained to understand the syntax and semantics of human language, making them appear well-suited for summarization tasks. These models can process large volumes of text and generate grammatically correct summaries at high speed.

This study assesses the summarization capabilities of two widely used LLMs - *GPT-3.5 turbo* and *LLama 2* - by analyzing their ability to summarize central bank publications, with a particular focus on the Bank of Italy's *Economic Bulletin*, an institutional report with a format and editorial structure that are repeated in every edition. The *Economic Bulletin* is meticulously drafted by senior economists and its English language version undergoes final revisions by professional English editors and translators to ensure clarity and accessibility for the general public.

To evaluate the summaries, we conducted a survey with domain experts from the Directorate General for Economics, Statistics and Research, comparing three types of summaries: the official summary from the *Economic Bulletin* and those generated by *GPT-3.5 turbo* and *Llama 2*. The evaluation was based on pairwise comparisons using three key criteria: fluency, consistency, and relevance. Following the definitions in Fabbri et al. (2021), fluency measures both grammatical correctness and stylistic coherence with the original text. Relevance assesses the summary's efficacy in distilling essential information from the source text. Finally, consistency examines factual accuracy, ensuring the absence of imprecisions or fabricated content.

We used the Bradley-Terry model for paired comparisons (Bradley and Terry 1952) to analyze the results, which – although limited by the small, non-random sample – are consistent with established patterns (T. Zhang et al. 2024). Overall, evaluators showed no clear preference for human and machine-generated summaries. However, evaluators with a deeper knowledge of the *Economic Bulletin* and higher education levels tended to favor human-generated summaries, indicating that LLMs may not yet fully capture specialized knowledge.

Our findings also suggest a preference for human summaries among women and individuals over 35. This pattern may reflect how preference training and fine-tuning processes shape LLM outputs in a way that aligns with different demographic groups' content preferences (Kirk et al. 2024; Kotek, Dockum, and Sun 2023; Pervez and Titus 2024).

2. Literature Review

Large language models (LLMs) are taking the world by storm. In 2018 OpenAI developed the initial Generative Pre-trained Transformers (GPT) model, a state-of-the-art language model, capable of understanding and replicating patterns in human language with good accuracy. Since September 2022, when OpenAI released "Chat Generative Pre-Trained Transformer", commonly known as "ChatGPT", the significant potential of these AI tools was clear to the public. Subsequently, more proprietary and open LLMs have been released over time with ever-increasing capabilities (OpenAI et al. 2023; Gemini Team et al. 2023; Touvron et al. 2023).

Large language models (LLMs) represent state-of-the-art artificial intelligence systems trained on extensive textual corpora to generate contextually relevant, human-like text. These models demonstrate remarkable capabilities in producing coherent and grammatically sophisticated content with high computational efficiency. Their applications span diverse tasks including text correction, writing assistance, and content summarization. In October 2023, an OECD survey of various institutions (Mitton 2023) revealed that while 23% of respondents reported using Generative AI (GenAI) technologies in their work, 76% lacked clear policies and guidelines for its use in communication tasks.

LLMs have found applications in different domains, including finance. Fatouros et al. (2023) investigated the capabilities of large language models in financial sentiment analysis, in the context of the foreign exchange market, and demonstrated their significant effectiveness in understanding market behavior. Further studies showed the potential of LLMs in generating investment signals given a set of information (Fatouros et al. 2024).

In the realm of central banking, Hansen and Kazinnik (2023) assessed whether ChatGPT can decipher FOMC announcements. Smales (2023) tested whether ChatGPT can classify hawkish/dovish monetary policy decisions made by the Reserve Bank of Australia (RBA). Gambacorta et al. (2024) introduced central bank language models (CB-LMs) designed for two primary purposes: (i) predicting masked words within central bank-specific terminology and (ii) classifying the monetary policy stance conveyed in Federal Open Market Committee (FOMC) statements. The authors advocated advancing Natural Language Process analysis in monetary economics and central banking, emphasizing the importance of knowledge sharing among central banks to enhance the application of language models in the processes underlying monetary policy decision-making. However, they also highlighted critical concerns regarding confidentiality, transparency, replicability, and cost-efficiency associated with integrating LLMs into central banking workflows.

The use of LLMs in text summarization is a widely studied topic, we refer to Zhang, Yu, and Zhang (2024) for a comprehensive review. Recent studies found that performance can be comparable to expert humans on generic news, although with consistent stylistic differences (T. Zhang et al. 2024). However, the evaluation of LLM-Generated summaries is a task that remains methodologically challenging, and current approaches remain incomplete (Liu et al. 2023; Schaik and Pugh 2024).

The evaluation of AI-generated content quality and the detection of AI-authored content are active areas of research. Recent studies on poetry found that non-expert readers were not able to identify AI-generated poems, and those were rated more favorably than human-authored ones (Porter and Machery 2024). Another study on artworks found that participants were unable to consistently distinguish between human and AI-created images, and they generally preferred AI-generated artworks over human-made ones (Grassini and Koivisto 2024). Additionally, the same study showed that participants displayed a negative bias against artworks that were perceived as AI-generated, independently on their true source.

3. Experiment design

The objective of this study is to assess whether a professional audience demonstrates a clear preference for human-generated summaries over those produced by LLMs in the context of central bank communications. As a case study, we analyze the quarterly *Economic Bulletin* of the Bank of Italy, a highly standardized institutional publication. This *Bulletin* offers a comprehensive overview of economic developments in Italy, contextualized within the broader international and euro-area context. It focuses on key aspects of the economy, including the real economy, inflation, banking, and the financial markets. Each section has an introductory summary, edited by the editorial committee and professional reviewers at the Library Division of Bank of Italy.

The *Economic Bulletin* is originally written in Italian and undergoes a multi-step translation process before publication. First, an institutional automated tool generates an initial translation. This version is then reviewed and refined by a professional translator before undergoing a final validation by editors to ensure accuracy and adherence to the publication's standards.

For our analysis, we randomly selected five out of the 13 sections from the April 2023 issue of the *Economic Bulletin*[1]. The selected sections included: (1.2) *The World Economy – The Euro Area*, (2.2) *The Italian Economy – Firms*, (2.6) *The Italian Economy – Price Developments in Italy*, (2.7) *The Italian Economy – Banks*, and (2.8) *The Italian Economy – Financial Markets*. We extracted the original text from each section and generated summaries using GPT-3.5 Turbo and LLaMA 2 70B. To ensure comparability with the original text, we designed the prompts to produce summaries with a word count approximately matching that of the human-authored introductory summaries at the beginning of each section.

In order to evaluate the preferences of a professional audience, we conducted a survey (full text available in Appendix B) among all employees in the Directorate General for Economics, Statistics, and Research at the Bank of Italy, comprising approximately 600 professionals. A number of divisions in the directorate are involved, in different capacities, in writing and editing of the *Economic Bulletin*. The survey was designed and distributed via an internal platform, which also recorded completion time for each survey response. Participants received a series of emails containing a link to the survey, facilitating access and encouraging participation.

The survey gathered demographic information about respondents, including gender, age, level of English proficiency and educational credentials including degree level and field. Additionally, it collected details on professional background, such as familiarity with LLMs, years of work experience, whether respondents had ever contributed to the authorship of the Bank of Italy's *Economic Bulletin*, and how often they read the publication.

For each selected section, we conducted a randomized pairwise comparison of two summaries chosen from three possible sources: the official human-generated introductory summary, GPT-3.5 Turbo, or LLaMA 2. To minimize potential judgment biases, the survey characterized all summaries as generated by LLMs. Participants were asked to evaluate their preferences for each pair of summaries based on three

[1] https://www.bancaditalia.it/pubblicazioni/bollettino-economico/2023-2/index.html

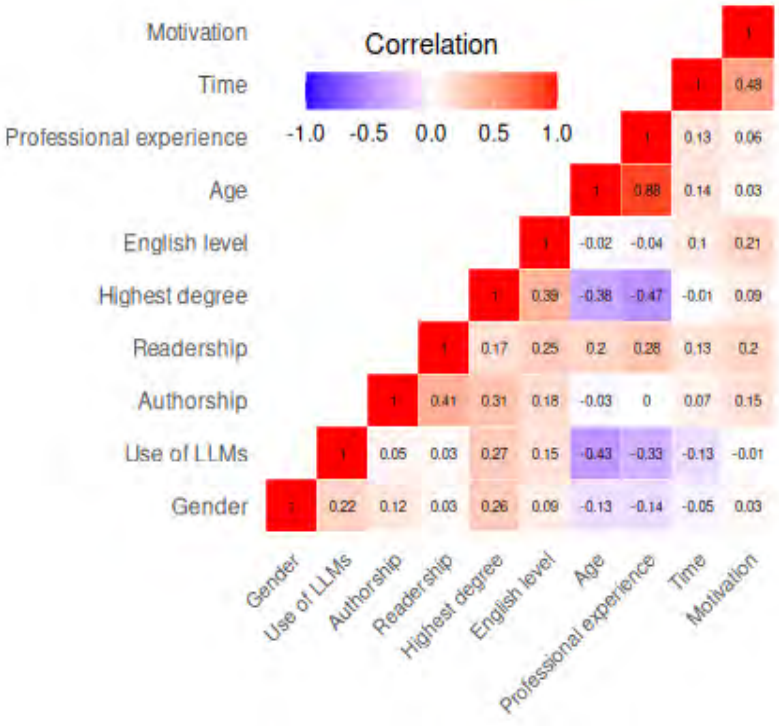
metrics - Fluency, Consistency, and Relevance - as defined by Fabbri et al. (2021). Preferences were recorded using a five-point Likert scale: Strong Preference for Text A, Preference for Text A, Neutral, Preference for Text B, and Strong Preference for Text B. Respondents were also given the opportunity to provide written justifications for their evaluations.

4. Data description

The survey was conducted in February 2024, yielding 175 responses, representing approximately 30% of the target population. Of these respondents, 133 completed evaluations for all randomly proposed text pairs, while 42 participants provided assessments for at least one text pair. Nine responses contained incomplete demographic information, which was subsequently imputed using the Random Forest algorithm provided by the mice package in R (Buuren and Groothuis-Oudshoorn 2011). Analyses performed both with and without imputed data yielded consistent results. Table 4 in Appendix A reports the summary figures for the covariates.

We conducted a preliminary analysis of correlations across covariates, with variables transformed into ordinal levels. The results are visualized in Figure 1. As expected, age and professional experience showed strong positive correlation. We observed a negative correlation between LLM usage and age, and a positive correlation with educational attainment. Finally, there is a positive correlation between the time spent for answering the survey and having left at least a written motivation for the preferences indicated between different texts.

Covariates correlation plot — Figure 1



We computed the mean preference scores for each metric and section using a weighted scoring system. This system allocates +1 point to the preferred text and -1 point to the non-preferred text, with strong preferences weighted at±2 points and neutral responses at zero points. As shown in Figure 2, preferences varied across writing samples: human-generated text was significantly preferred in sections 1.2 (*The World Economy – The Euro Area*) and 2.7 (*The Italian Economy – Banks*), whereas LLM-generated text performed better in sections 2.6 (*The Italian Economy – Price Developments in Italy*) and 2.8 (*The Italian Economy – Financial Markets*). Section 2.2 of the *Bulletin* (*The Italian Economy – Firms*) yielded heterogeneous results across different metrics, with no clear preference pattern emerging.

Mean preference scores

Figure 2

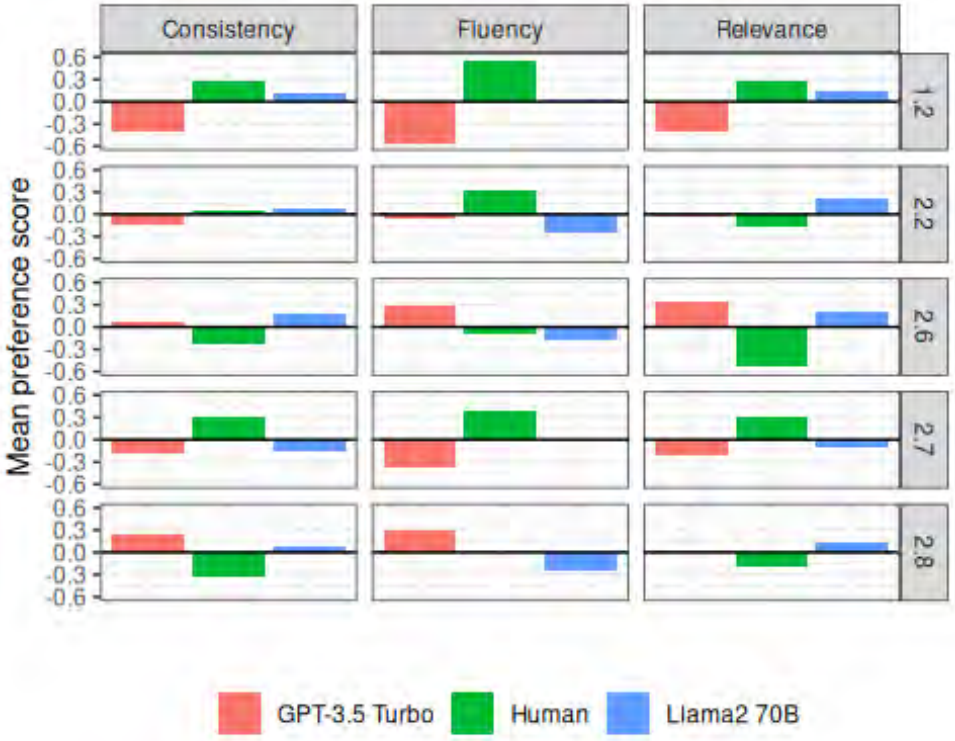
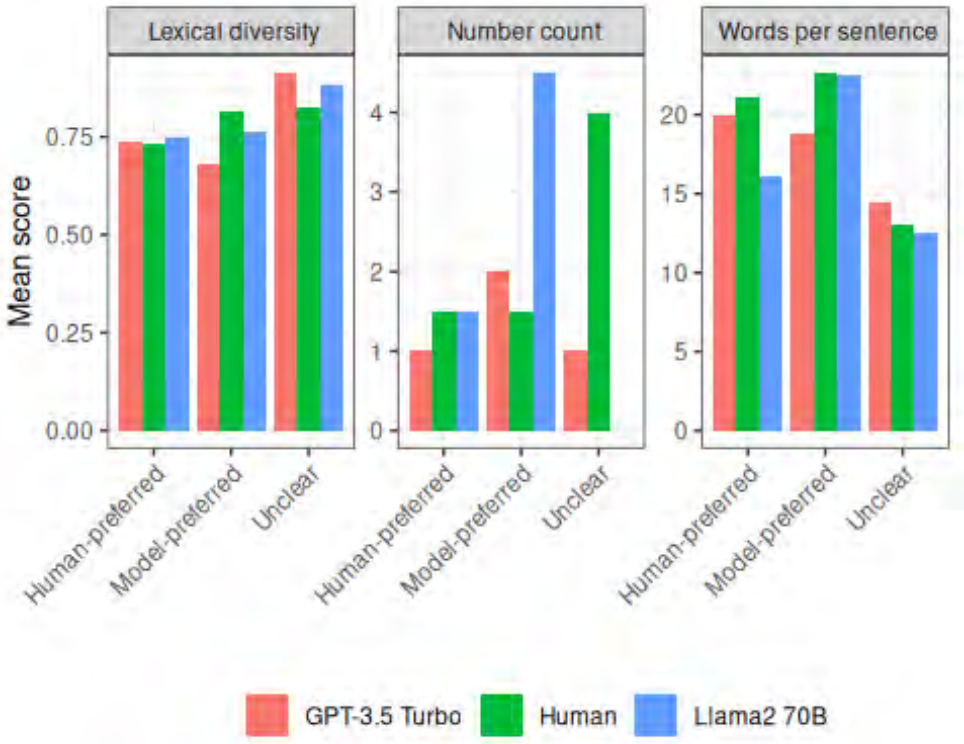


Figure 3 presents a set of descriptive metrics for each source and summary, including the average number of words per sentence, the frequency of numerical values in each summary, and lexical complexity, which is defined as to the ratio of different unique word stems to the total number of words. Based on the preferences outlined above, we classify sections 1.2 and 2.7 as "Human preferred," section 2.2 as "Unclear," and the remaining sections as "Model preferred." The emerging pattern shows that the favored summaries in human-preferred sections have longer sentences, and lower number counts, while model-preferred sections show higher number counts and shorter sentences. This pattern aligns with the different nature of these sections - model-preferred sections cover data-heavy topics like prices and financial markets, while human-preferred sections handle more complex, context-rich topics like macroeconomic analysis and banking sector developments.

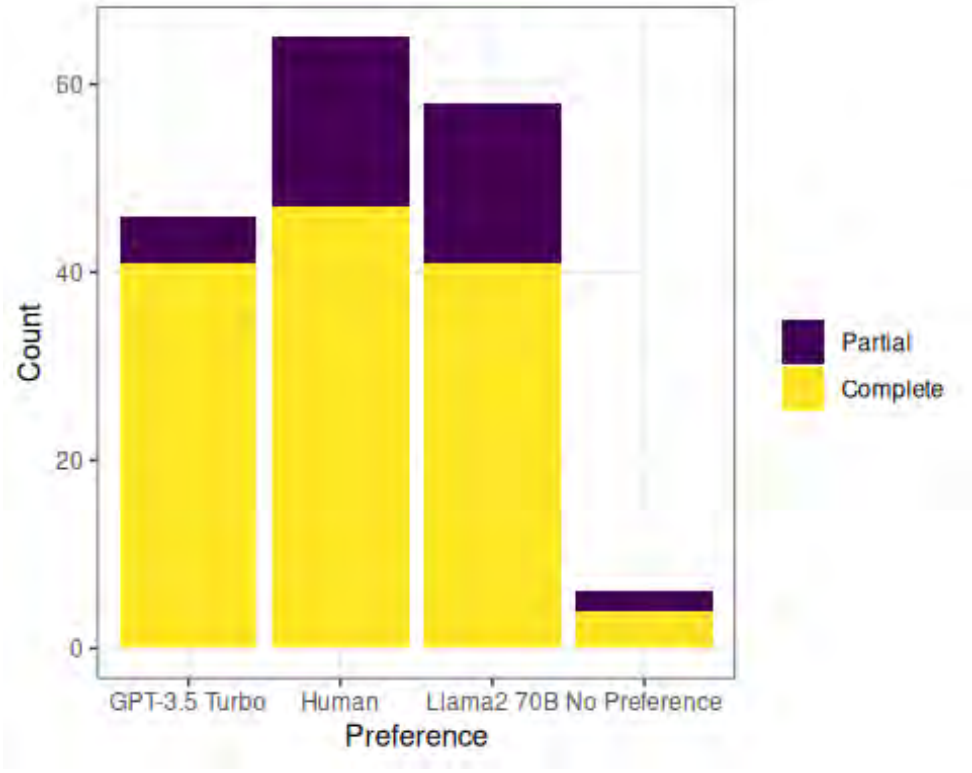
Text metrics

Figure 3



We chose to use The Bradley-Terry model (Bradley and Terry 1952) to convert the pairwise comparisons of the summaries into a comprehensive ranking of preferences for each respondent (see Appendix C for the formal specification and estimation procedure). This model is a popular choice to process the pairwise preference data, and it is commonly used to derive LLM rankings from conversational data (Chiang et al 2024). The results, shown in Figure 4, indicate that human-generated summaries are the most preferred overall, particularly among respondents who completed the survey. GPT-3.5 Turbo and LLaMA2 70B are competitive, with similar aggregate scores, while "No Preference" ranks lowest, indicating a strong tendency for respondents to choose a clear favorite.

Individual preferences for summary sources, derived from Bradley-Terry scores Figure 4



5. Main results

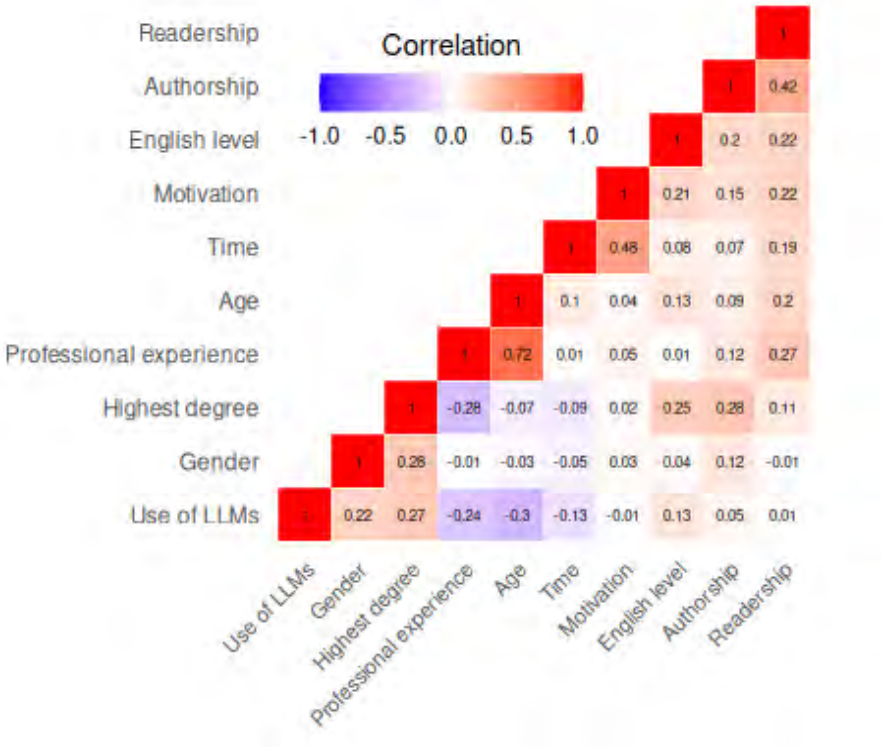
We employed a logistic regression model to explore whether specific respondent characteristics influence their preference for human-generated summaries over those created by LLMs. In this analysis the dependent variable is the Bradley-Terry calculated preference for human-generated text.

In order to streamline the model and reduce dimensionality, we recoded the covariates into binary variables. Age was categorized into two groups: "Under35" and "Over35." Professional experience was dichotomized into "Junior" (10 or fewer years of professional experience) and "Senior." English proficiency levels of "Elementary" and "Intermediate" were combined into a single category, as were all educational levels below the doctoral level. Additionally, readership intensity of the *Economic Bulletin* was simplified into two groups: "Low" for respondents who indicated reading the publication "Sometimes" or less frequently, and "High" for those reporting the two highest levels of readership intensity. "Motivation" indicates whether the respondent provided at least a written motivation for the preferences assigned.

This approach ensures a parsimonious model while retaining the ability to identify key patterns and associations in the data. Figure 5 presents the correlation matrix for the dichotomized variables.

Dichotomized covariates correlation plot

Figure 5



The logistic regression analysis using the full set of covariates, with preference for human-generated text as the dependent variable, is reported as (1) in Table 1. Since most variables were not statistically significant, we implemented a stepwise backward selection procedure using Akaike Information Criterion (AIC) (Akaike 1969, 1971) to identify the most parsimonious model. We present this reduced model as (2) in Table 1.

Logistic regression on Bradley-Terry preferences for human-generated text Table 1

	(1)	(2)
Intercept	-2.064 **	-2.175 ***
	(0.638)	(0.528)
Age: Over35	1.255 *	1.117 **
	(0.627)	(0.429)
Gender: Male	-1.008 *	-0.968 *
	(0.395)	(0.389)
Authorship: Yes	0.793 *	0.797 *
	(0.397)	(0.360)
Motivation: Yes	0.952 *	0.858 *
	(0.414)	(0.352)
Highest degree: PhD	1.119 *	1.052 **
	(0.447)	(0.396)
English level: Advanced	-0.660	
	(0.479)	
Professional Experience: Senior	0.016	
	(0.557)	
Use of LLMs: Yes	0.247	
	(0.428)	
Readership: High	0.164	
	(0.406)	
Time	-0.000	
	(0.000)	
N. obs.	175	175
AIC	218.931	211.088

*** p < 0.001; ** p < 0.01; * p < 0.05.

While the model demonstrates modest predictive performance, with an F1 score of 0.42, it achieves statistically significant improvement over the no-information rate baseline.

The first takeaway is that age and gender have the largest coefficients in absolute terms: economists older than 35 years tend to prefer human text while men tend on average to prefer LLMs. The second takeaway is that having already used LLMs does not substantially influence the choice. Our findings also showed that holding a doctoral degree is positively correlated with preference for human text. Lastly, not surprisingly, authors of past *Economic Bulletin* editions favor the human text on average.

Wong and Kim (2023) demonstrated that individuals are more likely to perceive ChatGPT as male rather than female. Additionally, research on gender differences in reading preferences (Summers 2013) found that male respondents were slightly more inclined to favor books authored by men. For our study, it is relevant to mention that the translation team for the *Economic Bulletin* is mostly composed by women, and 4 out of the 5 human-generated summaries under analysis were translated by a woman.

These findings suggest that the writing style of LLMs, shaped by their training and fine-tuning processes, may influence their appeal across different demographic groups.

Furthermore, having previously authored the *Economic Bulletin* and the provision of a written justification for preferences may reflect a heightened level of involvement and an emotional attachment to the publication and its distinctive writing style. Consequently, it is not surprising that these factors are associated with a stronger preference for human-generated text.

6. Robustness analysis

In this section we show the results of several robustness exercises that we performed in order to check the strength of our findings. The first exercise is the estimation of the reduced model presented as (2) in Table 1 on the three separate evaluation metrics: consistency, fluency, and relevance. Figure 6 presents the Bradley-Terry preferences for each individual metric.

Bradley-Terry preference scores for individual metrics Figure 6



Human-generated summaries seem to have an edge over LLMs in terms of consistency and fluency, while they score less favorably for relevance. Table 2 presents the results for logistic regression of preferences for human-generated text on each individual metric.

Logistic regression on individual metrics Bradley-Terry preferences for human-generated text

Table 2

	Consistency	Fluency	Relevance
Intercept	-1.772 ***	-1.647 ***	-2.707 ***
	(0.496)	(0.489)	(0.616)
Age: Over35	0.568	0.960 *	1.424 **
	(0.407)	(0.406)	(0.523)
Gender: Male	-0.604	-1.003 **	-0.665
	(0.373)	(0.375)	(0.397)
Authorship: Yes	0.588	0.588	0.215
	(0.355)	(0.356)	(0.384)
Motivation: Yes	0.579	1.032 **	0.854 *
	(0.344)	(0.342)	(0.376)
Highest degree: PhD	0.711	0.682	0.664
	(0.383)	(0.377)	(0.410)
N. obs.	175	175	175
AIC	216.710	220.133	194.082

*** p < 0.001; ** p < 0.01; * p < 0.05.

We note that the coefficients for age and having written a motivation for the preference retain significance for two metrics, and gender retains significance for fluency.

The second robustness check we perform is variable selection using LASSO (Tibshirani 1996). We select the optimal shrinkage coefficient with cross-validation using the R package glmnet (Friedman, Tibshirani, and Hastie 2010). The results are shown in Table 3.

Lasso logistic regression on Bradley-Terry preferences for human-generated text

Table 3

	(1)
Intercept	-1.688
Age: Over35	0.766
Gender: Male	-0.554
Authorship: Yes	0.603
Motivation: Yes	0.583
Highest degree: PhD	0.667
Professional experience: Senior	0.000
Use of LLMs: Yes	0.000
Readership: High	0.053
English level: Advanced	-0.034
Time	0.000

Variable selection via LASSO leads to results very similar to the stepwise regression, retaining the same sign and relative magnitude. Two additional predictor - readership and English level - are retained by LASSO, but only with marginal coefficients.

The third robustness check involves excluding past authors of the *Economic Bulletin* from both the descriptive statistics and the logistic regression analysis. Figure 7 presents the distribution of preferences among respondents who have not contributed to previous editions of the *Economic Bulletin*. As expected, summaries generated by humans are the least preferred among this group.

Bradley-Terry preference scores excluding past authors of the *Economic Bulletin* Figure 7

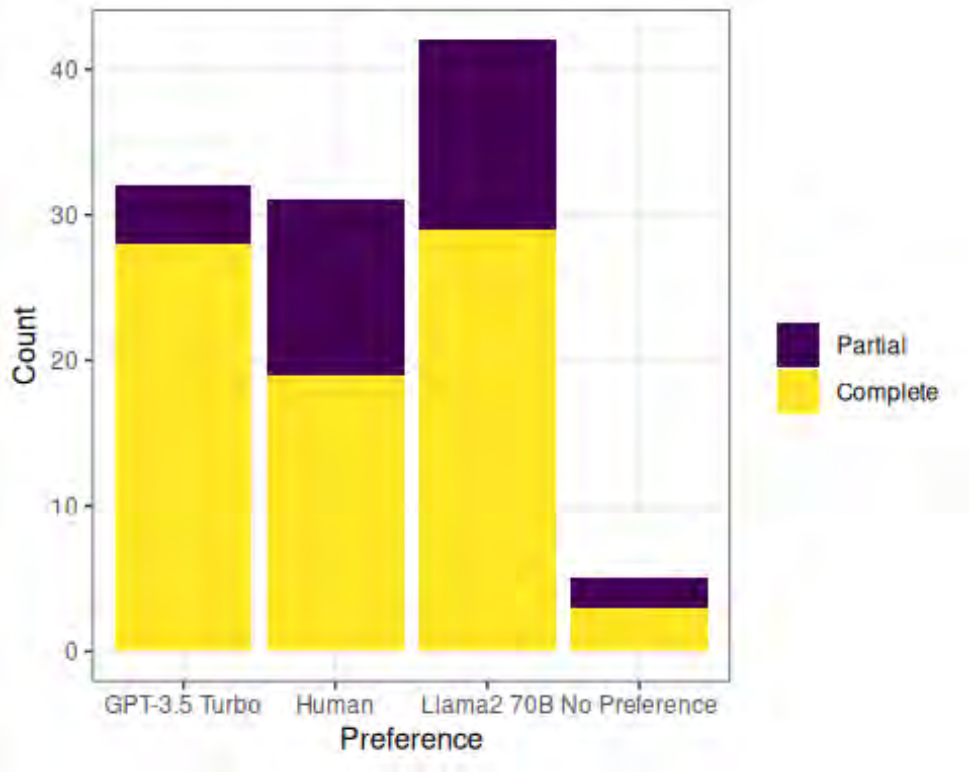


Table 4 reports the results of the logistic regression and stepwise selection procedure, excluding past authors of the *Economic Bulletin*, following the same methodology as in Table 1. The findings remain consistent with the analysis of the full sample, with the sole exception that the "Motivation" parameter is not statistically significant in the reduced sample.

Logistic regression on Bradley-Terry preferences for human-generated text excluding past authors of the *Economic Bulletin* Table 4

	(1a)	(2a)
Intercept	-2.858 **	-2.965 ***
	(0.910)	(0.750)
Age: Over35	2.580 **	2.207 **
	(0.912)	(0.696)
Gender: Male	-1.050 *	-0.967
	(0.533)	(0.513)
Motivation: Yes	0.759	
	(0.605)	
Highest degree: PhD	1.787 **	1.721 **
	(0.639)	(0.525)
English level: Advanced	-0.913	
	(0.630)	
Professional Experience: Senior	-0.404	
	(0.759)	
Use of LLMs: Yes	0.021	
	(0.578)	
Readership: High	0.733	
	(0.582)	
Time	-0.000	
	(0.000)	
N. obs.	110	110
AIC	121.835	114.818

*** p < 0.001; ** p < 0.01; * p < 0.05.

7. Conclusions

Our study confirms that large language models (LLMs) are capable of producing text of comparable quality to that produced by expert humans. However, the findings also highlight the presence of audience segmentation in preferences, influenced by demographic factors such as age, gender, and education level. Specifically, older respondents and those with doctoral degrees are more likely to prefer human-generated text, suggesting that experience and advanced education may shape expectations or familiarity with certain writing styles. Conversely, male respondents show a tendency to prefer LLM-generated content, which could be influenced by gendered perceptions of writing styles, as supported by previous research.

Additionally, the study highlights the role of emotional engagement and connection with the source material, as evidenced by the positive association between past authorship of the *Economic Bulletin* or providing a written justification for preferences and the inclination towards human-generated text. These findings

suggest that factors such as personal involvement and familiarity with a publication's distinctive style play a crucial role in shaping audience preferences.

While the models used to analyze these patterns show moderate predictive accuracy, they offer statistically significant insights over a no-information baseline. The consistency of the results across robustness checks supports the validity of the observed relationships between demographic characteristics and preferences.

These findings have important implications for the use of LLM-generated content in professional communications, suggesting that organizations should carefully consider their audience composition when determining optimal content generation strategies. Future research might productively explore whether these preference patterns persist across different document types and institutional contexts, and investigate the specific textual features that drive these demographically-based preferences.

References

Akaike, Hirotugu. 1969. "Fitting autoregressive models for prediction." *Annals of the Institute of Statistical Mathematics* 21 (1): 243–47.

Akaike, Hirotugu. 1971. "Autoregressive Model Fitting for Control." *Annals of the Institute of Statistical Mathematics* 23 (1): 163–80.

Bradley, Ralph Allan, and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39 (3/4): 324. https://doi.org/10.2307/2334029.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r" 45: 1–67. https://doi.org/10.18637/jss.v045.i03.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2025. "Chatbot arena: an open platform for evaluating LLMs by human preference." In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vol. 235. JMLR.org, Article 331, 8359–8388. https://dl.acm.org/doi/10.5555/3692070.3692401.

Fabbri, Alexander R., Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. "SummEval: Re-Evaluating Summarization Evaluation." *Transactions of the Association for Computational Linguistics* 9: 391–409. https://doi.org/10.1162/tacl_a_00373.

Fatouros, Georgios, Kostas Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. "Can Large Language Models Beat Wall Street? Evaluating GPT-4's Impact on Financial Decision-Making with MarketSenseAI." *Neural Computing and Applications*, December. https://doi.org/10.1007/s00521-024-10613-4.

Fatouros, Georgios, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. 2023. "Transforming Sentiment Analysis in the Financial Domain with ChatGPT." *Machine Learning with Applications* 14 (December): 100508. https://doi.org/10.1016/j.mlwa.2023.100508.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent" 33. https://doi.org/10.18637/jss.v033.i01.

Gambacorta, Leonardo, Byeungchun Kwon, Taejin Park, Pietro Patelli, and Sonya Zhu. 2024. "CB-LMs: language models for central banking." BIS Working Papers 1215. Bank for International Settlements. https://ideas.repec.org/p/bis/biswps/1215.html.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2023. "Gemini: A Family of Highly Capable Multimodal Models." https://doi.org/10.48550/ARXIV.2312.11805.

Grassini, Simone, and Mika Koivisto. 2024. "Understanding How Personality Traits, Experiences, and Attitudes Shape Negative Bias Toward AI-Generated Artworks." *Scientific Reports* 14 (1). https://doi.org/10.1038/s41598-024-54294-4.

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can ChatGPT Decipher Fedspeak?" *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4399406.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. 2024. *The PRISM Alignment Dataset*. https://doi.org/10.57967/hf/2113.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. "Gender Bias and Stereotypes in Large Language Models." *Proceedings of The ACM Collective Intelligence Conference*, November, 12–24. https://doi.org/10.1145/3582269.3615599.

Liu, Yixin, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, et al. 2023. "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2023.acl-long.228.

Mitton, Terri. 2023. "Results from the Survey on the Use of Generative AI for Communication." In *UNECE Expert Meeting on Dissemination and Communication of Statistics 2023*. Geneva. https://unece.org/sites/default/files/2023-10/DissComm2023_S3_OECD_Mitton_PPT.pdf.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. "GPT-4 Technical Report." https://doi.org/10.48550/ARXIV.2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback". *Advances in neural information processing systems*, *35*, 27730-27744. https://dblp.org/rec/conf/nips/Ouyang0JAWMZASR22.html.

Pervez, Naseela, and Alexander J. Titus. 2024. "Inclusivity in Large Language Models: Personality Traits and Gender Bias in Scientific Abstracts." https://doi.org/10.48550/ARXIV.2406.19497.

Porter, Brian, and Edouard Machery. 2024. "AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably." *Scientific Reports* 14 (1). https://doi.org/10.1038/s41598-024-76900-1.

Schaik, Tempest A. van, and Brittany Pugh. 2024. "A Field Guide to Automatic Evaluation of LLM-Generated Summaries." *Proceedings of the 47th International ACM*

SIGIR Conference on Research and Development in Information Retrieval, July, 2832–36. https://doi.org/10.1145/3626772.3661346.

Smales, Lee A. 2023. "Classification of RBA Monetary Policy Announcements Using ChatGPT." *Finance Research Letters* 58 (December): 104514. https://doi.org/10.1016/j.frl.2023.104514.

Summers, Kate. 2013. "Adult Reading Habits and Preferences in Relation to Gender Differences." *Reference & User Services Quarterly* 52 (3): 243–49. https://doi.org/10.5860/rusq.52.3.3319.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." https://doi.org/10.48550/ARXIV.2302.13971.

Turner, Heather, and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package". *Journal of Statistical Software* 48 (9):1-21. https://doi.org/10.18637/jss.v048.i09.

Wong, Jared, and Jin Kim. 2023. "ChatGPT Is More Likely to Be Perceived as Male Than Female." https://doi.org/10.48550/ARXIV.2305.12564.

Zhang, Haopeng, Philip S. Yu, and Jiawei Zhang. 2024. "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models." https://doi.org/10.48550/ARXIV.2406.11289.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. "Benchmarking Large Language Models for News Summarization." *Transactions of the Association for Computational Linguistics* 12: 39–57. https://doi.org/10.1162/tacl_a_00632.

Appendix A

Summary of demographic characteristics for survey's participants.

Distribution of demographic details — Table 4

Gender	*Female*	*Male*					
	61	114					
Age	*26-35*	*36-45*	*45-55*	*>55*			
	48	63	43	21			
Professional experience	*≤5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*>30*
	31	37	27	31	24	7	18
Highest degree	*High School*	*Bachelor*	*Master*	*PhD*			
	5	21	50	99			
Highest degree field	*Economic and Finance*	*Engineering*	*Law and Humanities*	*Mathematics and Statistics*	*Other*		
	118	10	12	23	12		
English level	*Elementary*	*Intermediate*	*Advanced*				
	5	32	138				
Authorship of the Economic Bulletin	*No*	*Yes*					
	110	63					
Readership of the Economic Bulletin	*Never*	*Rarely*	*Sometimes*	*Often*	*Regularly*		
	8	34	49	41	43		
Use of LLMs	*No*	*Yes*					
	51	124					
Directorate	*Economic Outlook and Monetary Policy*	*Financial Stability*	*International Relations and Economics*	*Statistical Analysis*	*Statistical Data Collection and Processing*	*Structural Economic Analysis*	*Other*
	45	21	20	23	23	35	8

Appendix B

Survey administered to the participants.

Large Language Models

This survey aims to **evaluate the usage of Large Language Models** (LLMs) by personnel in the Economics and Statistics Department and **assess the quality of LLMs output** in the context of official Central Bank communications.

Your answers will help support the development of Bank of Italy's strategy and guidelines for LLMs usage. Please read each question carefully and answer using your best judgement.

The expected time to complete the questionnaire is **20 minutes**.

Selected chapters from the Bank of Italy Economic Bulletin (April 2023) were processed to obtain a concise summary using various LLMs. Your task is to express your preference between five pairs of text based on three metrics:

- **Fluency**: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.
- **Relevance**:A summary is fully relevant if it extracts the most important points from the source.
- **Consistency**: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Your responses will be processed anonymously and only used for this research.

There are 33 questions in this survey.

Demografics

Have you ever used ChatGPT or other Large Language models (LLMs)? *

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Yes
- ◯ No

How often do you use ChatGPT or other LLM? *

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '1 [LLMUSEYN]' (Have you ever used ChatGPT or other Large Language models (LLMs)?)

ⓘ Choose one of the following answers
Please choose **only one** of the following:

- ◯ I have only used it once or twice
- ◯ Once a month
- ◯ Once per week
- ◯ Several times a week
- ◯ Every day

How do you use ChatGPT or other LLMs?

*

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '1 [LLMUSEYN]' (Have you ever used ChatGPT or other Large Language models (LLMs)?)

ⓘ Check any that apply
Please choose **all** that apply:

- ☐ Institutional purposes
- ☐ Drafting text
- ☐ Summarizing text
- ☐ Writing software code
- ☐ Translating documents
- ☐ Personal purposes (research, etc)
- ☐ Other: _____

What applications of ChatGPT or other LLMs may better support your work?

Check any that apply
Please choose **all** that apply:

- ☐ Drafting text
- ☐ Summarizing text
- ☐ Writing software code
- ☐ Translating documents
- ☐ Other:

Assume you have access to fully private LLM services suitable for confidential data.

Gender:

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Female
- ◯ Male
- ◯ Prefer not to say

Age

Choose one of the following answers
Please choose **only one** of the following:

- ◯ =< 25
- ◯ 26 - 35
- ◯ 36 - 45
- ◯ 46 - 55
- ◯ > 55

Familiarity with English:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Beginner
- ◯ Elementary
- ◯ Intermediate
- ◯ Advanced

Highest degree obtained:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ High School
- ◯ Bachelor
- ◯ Master
- ◯ PhD

Degree topic:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Economics and Finance
- ◯ Mathematics and Statistics
- ◯ Engineering
- ◯ Law and Humanities
- ◯ Other

Years of professional experience:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ =< 5
- ◯ 6 - 10
- ◯ 11 - 15
- ◯ 16 - 20
- ◯ 21 - 25
- ◯ 26 - 30
- ◯ > 30

Did you author or edit any edition of Bank of Italy's Economic Bulletin in the last 5 years? *

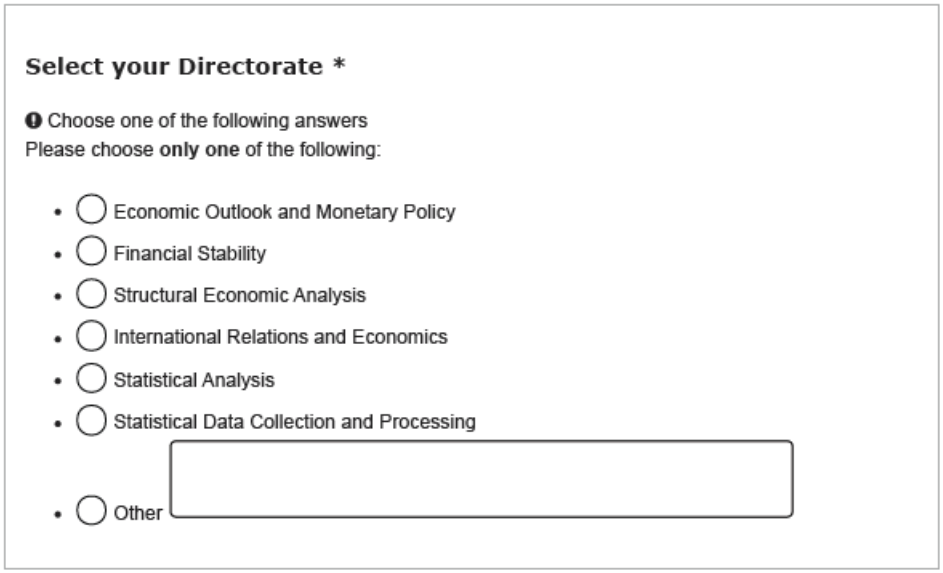
ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Yes
- ◯ No

How often do you read Bank of Italy's Economic Bulletin? *

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Never
- ◯ Rarely
- ◯ Sometimes
- ◯ Often
- ◯ Regularly



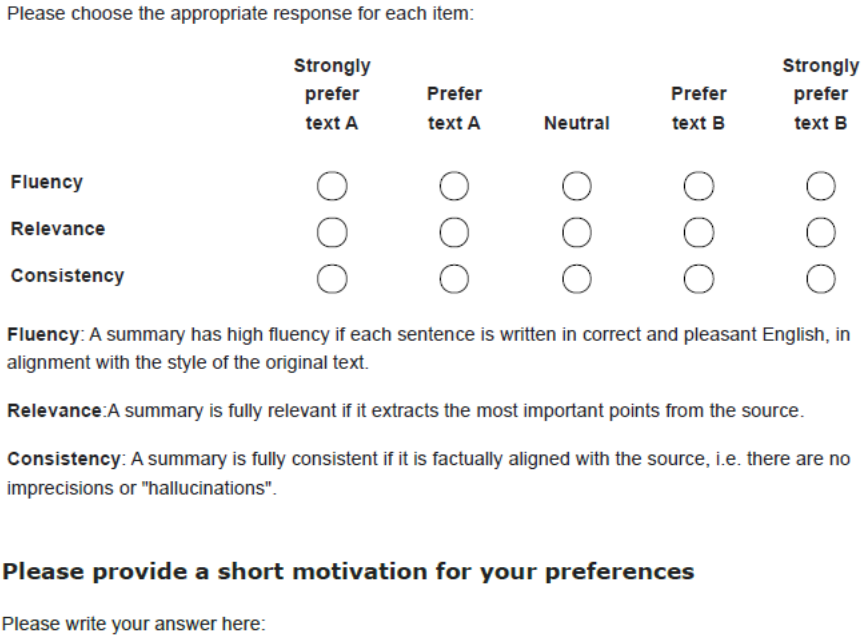
Text comparison 1

Text A and text B are two alternative summaries of chapter 1.2 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (Human)	Euro-area GDP stagnated in the last quarter of 2022, following a prolonged period of growth. The available indicators point to a slight expansion in economic activity in the early months of 2023. The fall in consumer price inflation continued, driven by the energy component; core inflation, which already stood at the highest levels, rose again. The Governing Council of the European Central Bank further raised the key interest rates, indicating that the high level of uncertainty is making it all the more important to adopt future decisions meeting by meeting and based on the data that become available. The European Union's plan to reduce energy dependence on Russia and to step up the green transition (REPowerEU) was approved.
Option 2 (GPT-3.5 Turbo)	The Euro Area witnessed stagnant GDP in Q4 2022 but hints at a mild upturn in early 2023. Despite falling consumer price inflation due to energy prices, core inflation rose. The European Central Bank responded by further raising key interest rates, emphasizing the need for adaptive decisions based on evolving data amid pervasive uncertainty. The Euro Area's plan, REPowerEU, aimed at reducing energy reliance on Russia and advancing the green transition, gained approval. GDP growth remains cautious, inflation forecasts revised downward, and although wage growth strengthens, risks of an inflationary spiral appear subdued. The ECB's decisive monetary tightening continues alongside reduced asset purchases and moderate credit to firms amidst increased loan costs. The Council of the EU endorsed the REPowerEU initiative to diminish energy reliance on Russia and promote sustainability.

Option 3 (LlaMa2 70B)	The eurozone economy experienced nil GDP grow that the end of 2022, but it slightly improved in the first quarter. Household consumption and gross fixed investment declined while net foreign demand contributed to GDP growth. The inflation rate fell for the fifth consecutive month, driven by energy product prices, while core inflation continued to rise. Wage growth strengthened, but risks of an upward wage-price spiral remain low. The ECB tightened monetary conditions by increasing key interest rates and announced that it will continue monitoring ongoing financial market tensions. Credit to firms decreased, while lending to households continued to grow at a slower pace. The cost of loans rose following the increase in key interest rates. The European Commission approved the REPowerEU plan to reduce energy dependence on Russia and step up the green transition.

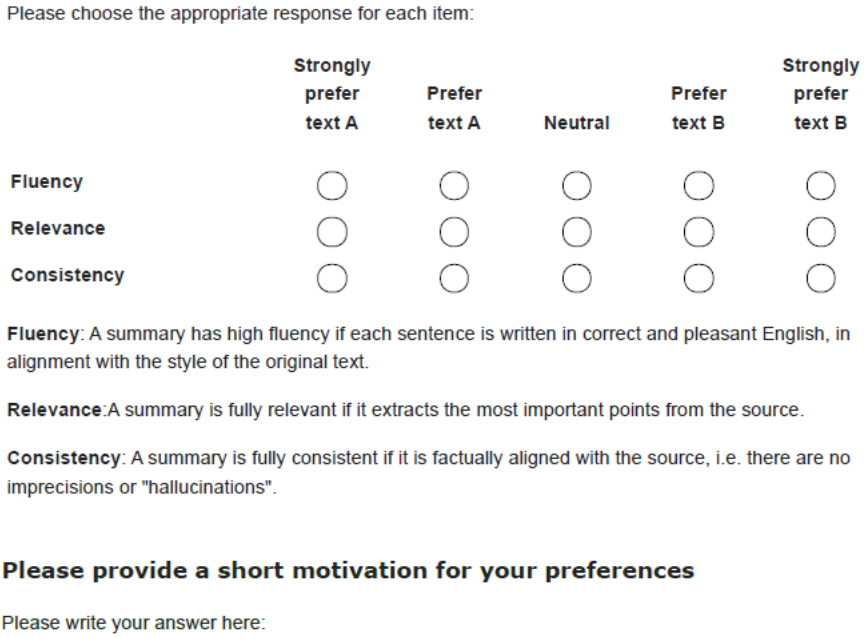


Text comparison 2

Text A and text B are two alternative summaries of chapter 2.2 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (GPT-3.5 Turbo)	Manufacturing activity has resumed growth in Q1 2023, while service sector indicators suggest a recovery. Business confidence has improved across all sectors, and investment has strengthened. However, the real estate market remains weak, with declining house sales and prices rising less than consumer price inflation. The debt-to-GDP ratio of Italian non-financial corporations has declined further.
Option 2 (LlaMa2 70B)	Industrial production appears to return to growth in Q1, despite a drop in capital goods and intermediate goods. Business confidence improved across all sectors, with positive signs from PMIs. Investment accelerated in Q4, while the real estate market weakened. Firms' debt-to-GDP ratio declined further, with liquidity remaining high.
Option 3 (Human)	After having declined fortwo consecutive quarters, manufacturing activity appears to have resumed its expansion in the first three months of 2023. The qualitative indicators for firms in the service sector point to a recovery inactivity. According to firms, capital accumulation will continue this year. Real estate market conditions remain weak.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

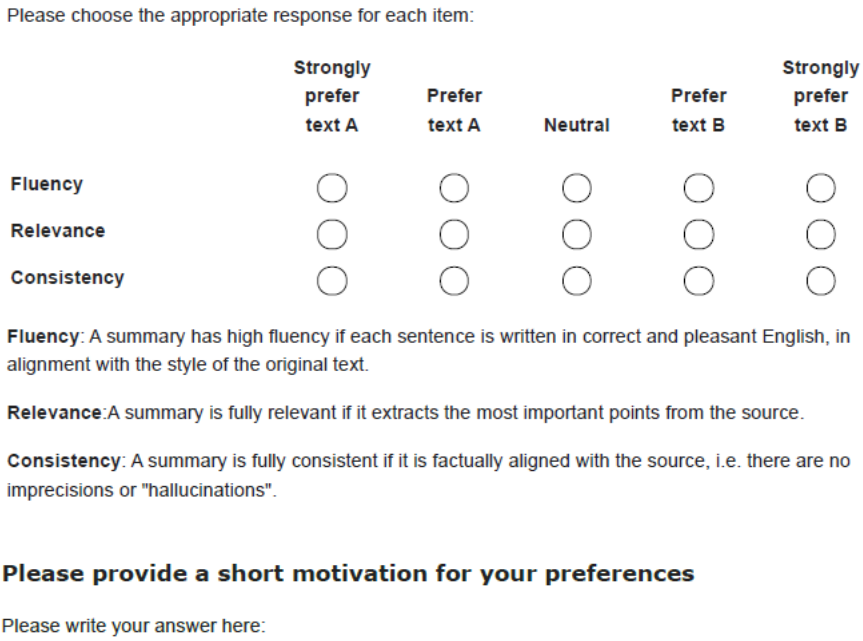
Please write your answer here:

Text comparison 3

Text A and text B are two alternative summaries of chapter 2.6 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (LlaMa2 70B)	Harmonized consumer price inflation in Italy decreased to 8.2% in March 2023 from its peak of 12.6% last autumn, primarily due to a drop in energy prices. Food prices continued to rise sharply, driven by the pass-through of past energy price increases. Producer price pressures eased, and wage pressures on prices remained subdued. Inflation expectations among firms and households continue to decline.
Option 2 (Human)	After peaking at the end of last year, consumer price inflation started to decline, reflecting the sharp drop in the energy component, while core inflation increased. Households' and firms' expectations point to continued signs of inflationary pressures easing.
Option 3 (GPT-3.5 Turbo)	Consumer price inflation in Italy decreased to 8.2% in March due to a reduction in energy prices, while core inflation increased. Food prices continued to rise sharply, driven by the pass-through of energy price increases. Producer price pressures continue to ease, and wage pressures on prices remain muted. Inflation expectations of firms and households continue to decline.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

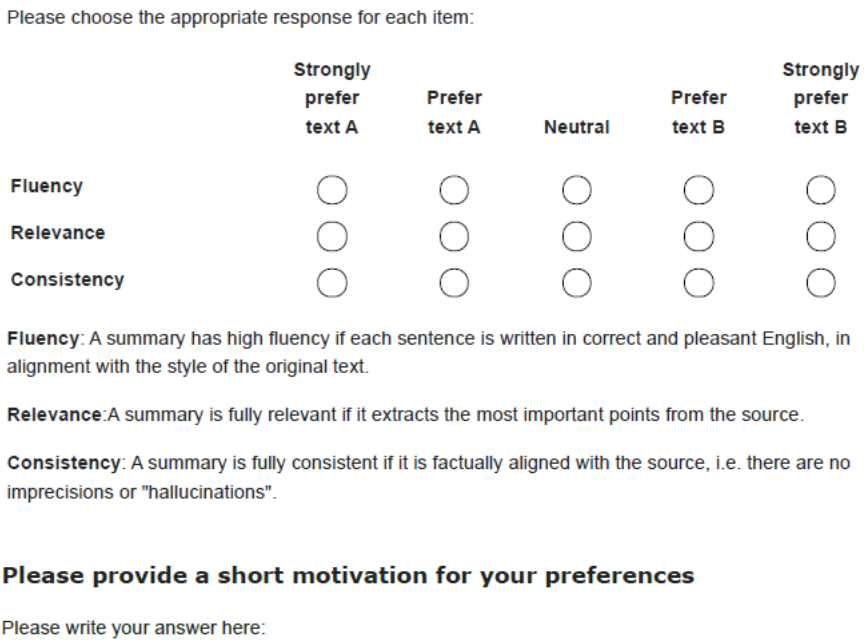
Please write your answer here:

Text comparison 4

Text A and text B are two alternative summaries of chapter 2.7 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (LlaMa2 70B)	Lending to non-financial corporations declines due to tighter credit standards and higher funding costs. Household lending is also slowing down, mainly due toa slowdown in home mortgages. The average interest rate on new bank loans to firms and households increases, while bank funding decreases. Despite tensions across the international banking system, Italian banks are in a better position compared to past crises. The non-performing loan rate remains low, and the profitability of significant banking groups increases. Capital ratios rise driven by reduction in weighted assets and Tier-1 capital growth.
Option 2 (GPT-3.5 Turbo)	The Italian banking sector saw a decline in lending to businesses due to tightened credit standards and higher funding costs, notably impacting firms' borrowing and repayment behaviors. The interest rate hikes continue to influence borrowing costs for both firms and households. Bank funding contracted due to repayments and reduced liabilities to the Eurosystem, as deposits decreased while funding costs rose. Despite global banking tensions post US and Swiss bank failures, Italian banks stand in a relatively stronger position. Non-performing loan rates remain low, and significant Italian banking groups exhibit improved profitability and increased capital ratios, attributed to reduced loan loss provisions and capital growth.
Option 3 (Human)	Between November and February, the growth rate of bank loans decreased sharply, particularly for loans to firms, reflecting higher funding costs and tighter credit standards. Key interest rate hikes continue to pass through to the cost of lending. Despite recent tensions following the failures of some banks in the United States and Switzerland, the Italian banking system is in a better position compared with crises in the past. The non-performing loan rate is still low. The profitability and capital ratios of Italian banking groups classified as significant improved in2022.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

Please write your answer here:

Text comparison 5

Text A and text B are two alternative summaries of chapter 2.8 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (Human)	In Italy as in other major financial markets, the tensions sparked by the collapses of some banks in the US and in Switzerland starting from 9 March have influenced the trends in asset prices. After rising in January and February, government bond yields and share prices fell sharply amid the first signs of instability in the international banking sector. Volatility increased as well, while yield spreads between Italian ten-year government bonds and the corresponding German Bunds held stable.
Option 2 (LlaMa2 70B)	Yields on Italian government bonds edged up 9 basis points to 4.1% in March, driven by increased concerns about policy rate hikes and risks to financial stability. Bond funding became more expensive, with yields on non-financial corporate and bank bonds rising 9 and 22 basis points respectively. Equity prices gained 4.8% for the Italian stock market index and5.2% for the banking index in January and February, but have since been impacted by increased volatility and downward pressure following the collapse of US banks and Credit Suisse.
Option 3 (GPT-3.5 Turbo)	Italian government bond yields edged up by 9 basis points to 4.1% at the end of March, while yields on bonds issued by Italian non-financial corporations and banks also rose, albeit to a lesser extent than in the euro area. Equity prices have gained since mid-January, but have been impacted by the collapses of some US banks and Credit Suisse, resulting in a spike in volatility and strong downward pressure on share prices, especially in the financial sector.

Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

Please write your answer here:

Appendix C

C.1 The Bradley-Terry Model: formal specification and estimation

The Bradley-Terry model (Bradley and Terry 1952) is a probabilistic model for analysing paired comparison data, which we use in this study to convert respondents' pairwise preferences into a comprehensive ranking of summary sources. Let $i, j \in \{1, \ldots, m\}$ index the alternative summaries under comparison. The Bradley–Terry model assigns each summary a latent score $\beta_k \in R$ such that the probability that i is preferred to j is:

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j},$$

with $Pr(j \succ i)$ obtained by swapping the indices. This can be equivalently expressed in logistic form:

$$Pr(i \succ j) = \frac{1}{1 + e^{-(\beta_i - \beta_j)}},$$

where $\beta_i = \log \pi_i$. β_i and β_j are positive-valued parameters which can be thought of as representing 'ability', and in the context of the study the ability to produce valid summaries of the *Economic Bulletin*.

The key advantage of the Bradley-Terry model is that it allows us to derive a complete ranking of multiple sources (in our case, human-generated, GPT-3.5 Turbo, and LLaMA2 70B summaries) from pairwise comparison data. While each individual comparison only involves two sources, the model can estimate ability parameters for all three sources by combining information across all pairwise contests. For identifiability, one source must be set as the reference category with its parameter constrained to zero. The estimated ability parameters for the other sources represent log-odds relative to this reference. These parameters can then be transformed back to the original scale, providing the basis for an absolute ranking of all sources.

To estimate the parameters of the Bradley-Terry model from our data, we use maximum likelihood estimation. We perform the estimation using the *BradleyTerry2* package in R (Turner and Firth 2012). To account for the 5-point Likert scale used in our survey, we weighted the comparisons as follows: strong preferences were given a weight of 2, regular preferences a weight of 1, and neutral responses were treated as 0.5 preference for each option. The model was first applied to each individual across all the answers given, in order to derive a general preference indication, and then separately to each quality dimension (fluency, consistency, and relevance), yielding dimension-specific preference parameters for each source.

C.2 Pairwise comparisons and the Bradley Terry Model in LLM evaluations

Pairwise comparison serves as an effective methodology to collect human preferences on outputs generated by Large Language Models. On one hand, pairwise preference data have been utilized to obtain data for the "human alignment" stage

of LLM training (Ouyang et al. 2022). On the other hand, they have been used to compare performance across different models, with statistical frameworks such as Bradley-Terry and Elo enabling the extraction of global rankings from these comparisons. Chiang et al. (2024) employed this latter approach in the Chatbot Arena framework, framework, which has emerged as a standard for human evaluation of LLM outputs.

The pairwise comparison approach reduces the cognitive load on evaluators, who need only compare two texts directly rather than assign absolute scores. It also eliminates scale biases that may arise when different evaluators interpret rating scales differently.

BankGPT: the use of Large Language Models in official communications

CLAUDIA BIANCOTTI, CAROLINA CAMASSA, MARCO FRUZZETTI, **LUIGI PALUMBO**, MYRIAM PORTALURI

The views expressed in this presentation are those of the authors only and do not necessarily represent the views of the Bank of Italy or the Eurosystem.

LLMs and official communications



The **use of LLMs is increasingly pervasive** in business and official settings.

Potential efficiency gains and lagging guidelines for official use.

Our research questions:

- Is there a **quality gap** between text generated by LLMs and human experts?
- Are **preferences** related to specific **characteristics of the audience**?

Some references

OECD institutions' staff

23% use GenAI
76% do not have guidelines
(Mitton, 2023)

LLMs for economic analysis

Financial sentiment analysis
(Fatouros et al., 2023)
Investment signals
(Fatouros et al., 2024)

LLMs and CB communication

Deciphering Fedspeak
(Hansen and Kazinnik , 2023)
RBA monetary policy statements
(Smales, 2024)

CB Language Models

Special-purpose text models
LLMs for CB workflows
(Gambacorta et al., 2024)

LLMs and summarization

LLMs on par with humans on news… (Zhang et al., 2024)
…but falling short on specific tasks (Lui et al., 2023)

GenAI vs. Human

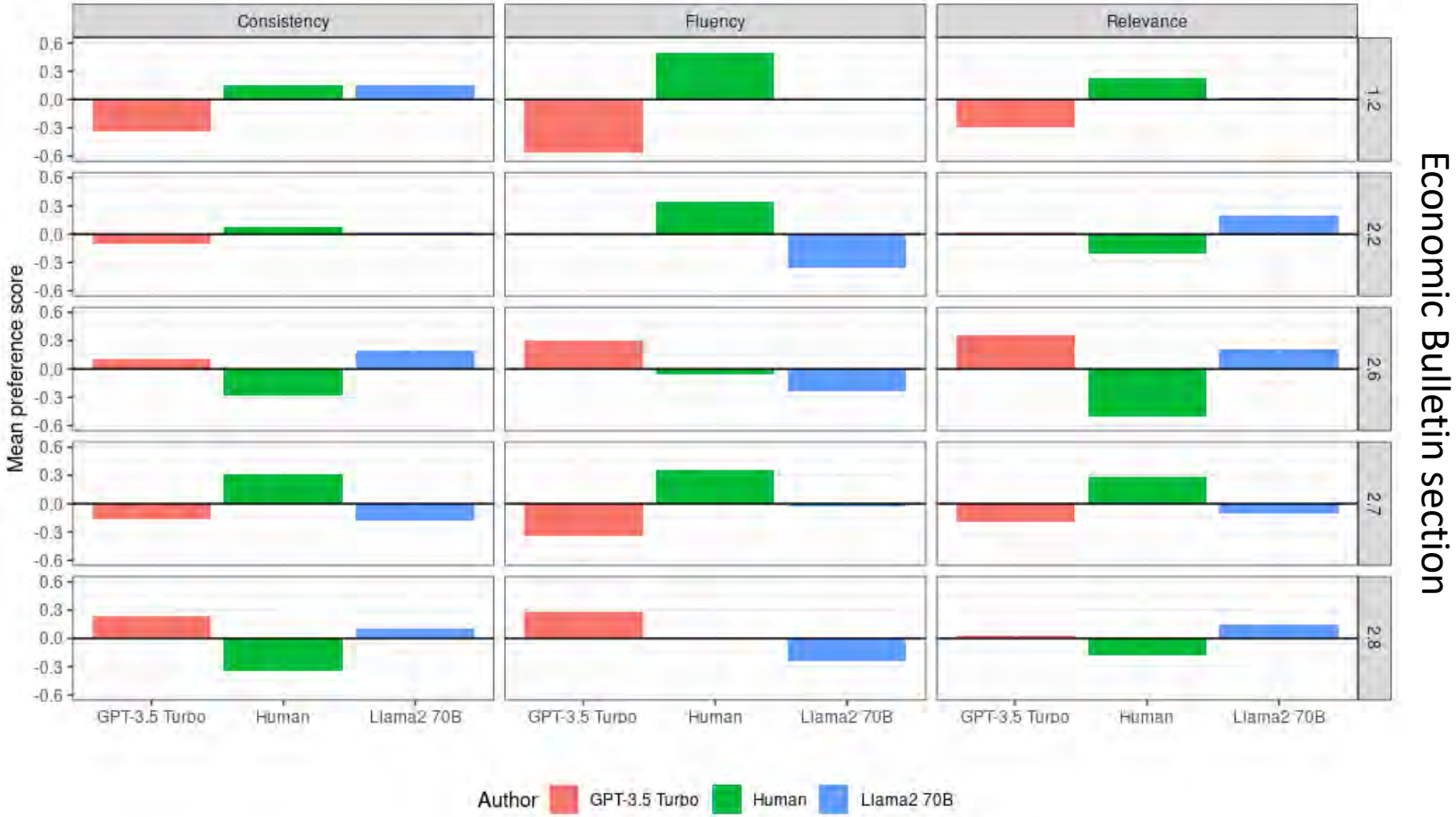
GenAI text better than human
(Porter and Machery, 2024)
Cannot distinguish, and bias
(Grassini and Koivisto, 2024)

Our experiment



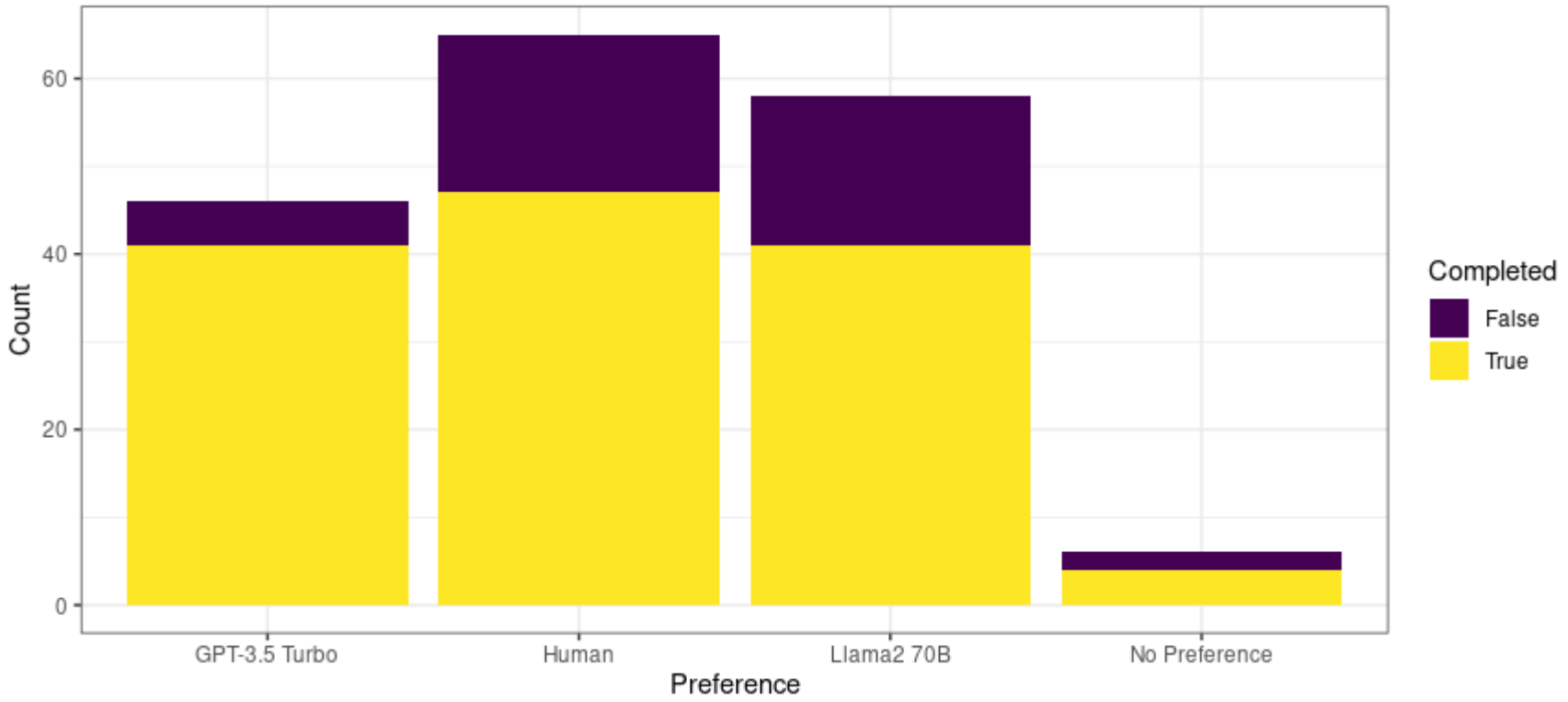
- Target: Bank of Italy's **Economics and Statistics Directorate General**
- Good participation, with **175** answers out of about 600 colleagues
- Pairwise comparison **of *Economic Bulletin* summaries**, all presented as LLM-generated
- Three evaluation metrics: **fluency, consistency and relevance** (Fabbri et al., 2021)
- Bradley-Terry model to convert comparisons into **comprehensive ranking of models** (Bradley and Terry, 1952)

Preferences are varied across sections and metrics



Economic Bulletin section

Human authors are (slightly) preferred



Age, gender, and education

	(1)		(2)	
Intercept	-2.064 **	*(0.638)*	-2.175 ***	*(0.528)*
Age: Over35	1.255 *	*(0.627)*	1.117 **	*(0.429)*
Gender: Male	-1.008 *	*(0.395)*	-0.968 *	*(0.389)*
Authorship: Yes	0.793 *	*(0.397)*	0.797 *	*(0.360)*
Motivation: Yes	0.952 *	*(0.414)*	0.858 *	*(0.352)*
Highest degree: PhD	1.119 *	*(0.447)*	1.052 **	*(0.396)*
English level: Advanced	-0.660	*(0.479)*		
Professional Experience: Senior	0.016	*(0.557)*		
Use of LLMs: Yes	0.247	*(0.428)*		
Readership: High	0.164	*(0.406)*		
Time	-0.000	*(0.000)*		
N. obs.	175		175	
AIC	218.931		211.088	

*** p < 0.001; ** p < 0.01; * p < 0.05.

Conclusions

- LLMs are capable of producing text of **comparable quality** to that produced by expert humans…
- …but stronger **domain expertise** steers readers' preferences towards human-generated text.
- Salience of **demographic factors** such as age and gender for preferences' prediction raise questions about **potential bias** in LLMs' training material and process.
- Implications for institutional communications: organizations should carefully **consider their audience composition** when determining optimal content generation strategies.

Q&A

THANKS!

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Extracting economic issues from news data: with the help of generative AI[1]

Younghwan Lee,
Bank of Korea

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Extracting economic issues from news data: with the help of generative AI

Younghwan Lee[1]

Abstract

This study proposes a method for extracting relevant economic issues in a timely manner and presenting them succinctly. Specifically, I introduce a method for directly assessing economic issues using news data. By closely tracking trending keywords in economic news articles, I identify key terms warranting attention during different periods. Based on these keywords and their contexts, several important issues driving the economic dynamics were inferred using the capabilities of generative AI. The primary advantage of this method is twofold. First, it provides policymakers with a powerful tool to directly focus on issues driving economic dynamics, unlike traditional approaches that indirectly identify the shocks behind economic variables. This is particularly important because policymakers deal with real issues and not mere economic variables. Second, the method employs algorithms that process vast volumes of data that are updated daily, making them timely, readily available, and potentially less prone to individual biases. The identified issues resulted primarily from the algorithm, with only minor refinements, allowing for replication with minimal human intervention. The usefulness of the proposed method is demonstrated through an example implementation using Korean news data from 2005.

Keywords: Economic Issue Extraction, Economic Trends Detection, Generative AI, Natural language processing

Table of Contents

Extracting economic issues from news data: with the help of generative AI 1

1. Introduction 2

2. Implementation 4

   2.1 Sentiment Classification 5

   2.2 Trending Keyword Extraction 7

   2.3 Summarization 9

3. Concluding Remarks 10

References 11

[1] Gangneung Branch, Bank of Korea

E-mail: yhlee@bok.or.kr

Disclaimer: The views expressed in this article do not necessarily reflect those of Bank of Korea.

1. Introduction

One of the most important mandates of a national statistical institute is to keep policymakers informed of current economic issues, so that they can take timely and effective policy actions. Traditionally, economic indicators such as GDP, inflation, and sentiment indices have been regularly published and closely monitored to fulfill this mandate. This raises the fundamental question: Are traditional indicators sufficient to capture the evolving economic landscape? Should policymakers rely on well-defined statistics to understand the economy?

This is analogous to how doctors diagnose patients. Symptoms, such as body temperature and blood pressure, provide valuable information about a patient's condition. However, the ultimate goal of doctors is not just to relieve symptoms but to cure the underlying disease. Similarly, economic indicators reflect outcomes and not necessarily the root causes of economic issues that policymakers should truly be aware of. If we focus only on economic statistics, we risk confusing the indicator with the true underlying issue. To develop effective policies, we must move beyond superficial symptoms and directly analyze the economic forces shaping these indicators.

Despite the apparent need to assess trending economic issues directly, their linguistic nature poses a major challenge to econometricians and statisticians. Traditional statistics are inherently numeric, allowing for quantitative analysis such as direct comparison and aggregation. On the contrary, economic issues are often expressed in text rather than numbers. Consequently, conventional statistical tools cannot be applied directly.

For a long time, this was our good excuse—economic text was too complex to process effectively. However, recent breakthroughs in Natural Language Processing (NLP) and generative AI have eliminated this issue. These advancements have enabled us to analyze economic issues directly from text data in a systematic and scalable manner.

To assess key economic issues directly from a vast volume of text data, I present a method augmented by a Large Language Model (LLM), which is also known as generative AI. This method consists of three key steps: sentiment classification, trending keyword extraction, and summarization. First, each sentence in the collected articles is classified according to its revealed sentiment as positive, negative, or neutral. Second, trending keywords are extracted from sets of positive and negative sentences. That is, keywords are identified from negative sentences to monitor emerging negative issues and vice versa for positive sentences. Third, sentences containing each trending keyword are embedded into a semantic vector space and clustered based on their similarity. Each cluster is then summarized using a pretrained Generative AI model, producing concise descriptions of the key issues associated with each keyword.

One may wonder why summaries are not directly generated from large text corpora using large language models. However, this approach has significant practical limitations. General-purpose models are computationally intensive and impractical for daily, large-scale processing. More importantly, their output often lacks specificity and verifiability. For instance, when asked to explain a surge in negative sentiment in news articles, the model may return a generic statement such as "Macroeconomic conditions have worsened due to various factors," without identifying concrete

drivers or providing links to underlying evidence. These models are also susceptible to hallucinations and generate confident but unfounded explanations. Such shortcomings are critical for policy applications that demand transparency and accountability.

The proposed method, which is simple and effective, has three key advantages: (i) verifiability, (ii) flexibility, and (iii) ease of implementation. First, verifiability arises from the modular structure of the method. Each step—sentiment classification, trending keyword extraction, and clustering with summarization—produces intermediate outputs that can be inspected and traced back to the original text, ensuring transparency and accountability. Thus, policymakers and researchers can verify the final results down to the level of the original data — the raw sentences actually published in newspapers — ensuring full transparency and reliability.

Second, the method can be customized for specific tasks flexibly without significant modifications to the existing framework. A simple example is the use of a predefined set of keywords to monitor specific domains of economic activity, such as "exchange rate." This small adjustment allows policymakers to remain informed of up-to-date issues related to exchange rate dynamics. A more intriguing application would be to use this method as a tool to monitor the transmission and interpretation of central bank communications in the media[2]. By restricting the target sentences to those that mention or quote central bank statements, the framework can be adapted to track how policy messages are conveyed, distorted, or received by the public and market participants. This opens up the possibility of assessing not only the content but also the effectiveness of monetary policy communication.

Most importantly, the proposed method is easy to implement. Instead of reinventing the wheel, I assembled existing building blocks, such as the K-means clustering algorithm from the open-source FAISS library and the Word2Vec algorithm from Mikolov et al. (2013), with minimal modifications. In addition, the summarization step was powered by an open-source pretrained LLM, Llama 3.2. This approach allows practitioners to adopt and scale a system easily without the need for complex in-house model development, making it practical and accessible for real-time policy applications. Moreover, because the method relies on well-established algorithms with known computational costs, it can be efficiently executed on high-performance laptops, eliminating the need for specialized hardware.

The remainder of this paper is organized as follows. Section 2 details the implementation of the proposed method, outlining the conceptual framework and practical considerations. Each of the three main components—sentiment classification, keyword extraction, and clustering with summarization—is described, with attention paid to design choices and the rationale behind them. Section 3 concludes the paper with a discussion of remaining challenges and directions for future research.

[2] I am grateful to Bruno Tissot for bringing attention to the importance of central bank communication during the workshop discussion, which inspired the idea of applying this method to monitor how central banks' messages are conveyed in the media.

2. Implementation

The core idea is to go beyond conventional numerical indicators and directly assess the underlying economic forces that drive fluctuations in a systematic and scalable manner. The proposed method builds a system that processes a large volume of economic news articles and returns concise summaries. These summaries capture the most relevant economic issues within a predefined target period, categorized by sentiment and thematic content. A three-step approach was adopted to ensure that the process remains verifiable, flexible, and easy to implement. Figure 1 briefly illustrates the overall procedure.

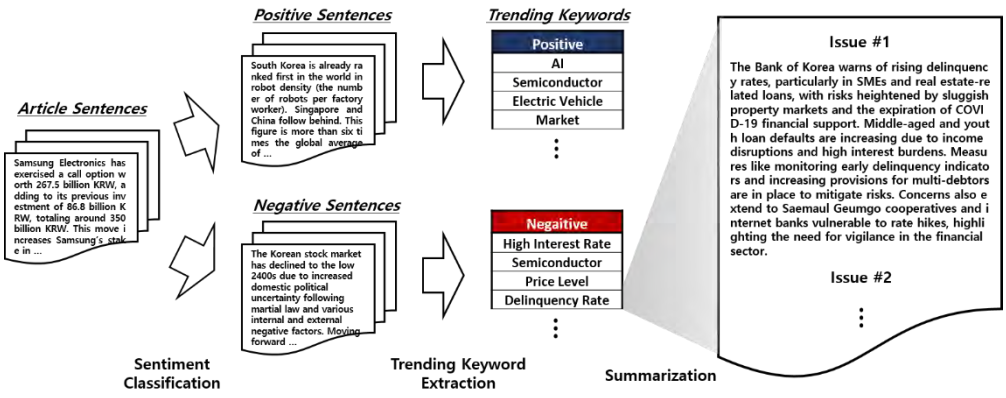
First, each sentence in the article corpus is classified by sentiment (positive, negative, or neutral) using a sentiment classifier trained on a large volume of human-labelled data. To properly assess the significance and economic impact of a factual event, the sentiment in which it is reported must be considered. For example, the statement "exports increased by 5%" may appear benign or even mildly favorable at first glance. However, if it frequently appears in a negative context, such as when driven by a sharp depreciation of the domestic currency, it should be interpreted with caution, as it may reflect broader economic concerns.

Second, trending keywords for the target period are extracted separately from the pools of positive and negative sentences. These keywords represent recurring themes or economic concepts gaining salience in the media, such as "inflation," "COVID-19," or "AI." A trending score is defined and calculated for each keyword to measure its informativeness with respect to the target period. Keywords with high trending scores are selected as the trending keywords.

The final step groups sentences containing each keyword into semantically coherent clusters. Each cluster is then summarized using a pretrained generative model, yielding concise and readable descriptions of the underlying economic issues. This three-stage pipeline transforms raw text into structured outputs that are verifiable, policy-relevant, and easy to interpret. The following sections describe each step in greater detail.

Snapshot

Figure 1.



Source: Author's illustration based on implementation results.

2.1 Sentiment Classification

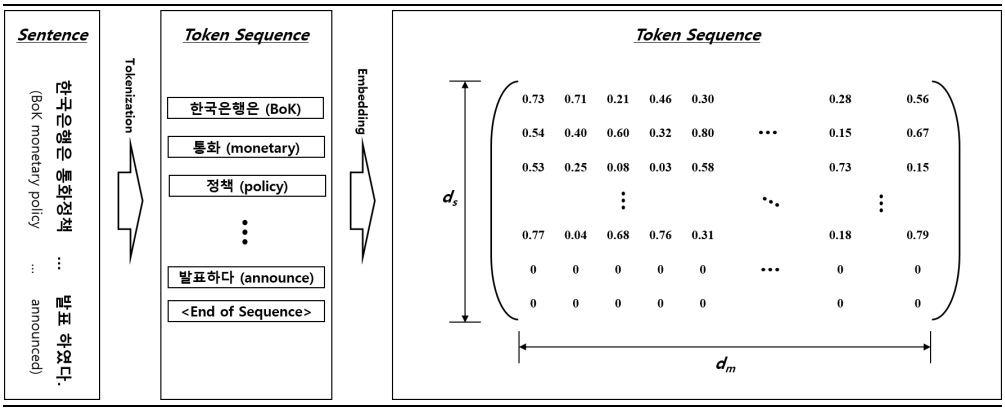
The sentiment classifier can be viewed as a function $\mathrm{C}: \mathrm{D} \rightarrow \mathrm{L}$ that maps a sentence $\mathrm{d} \in D$ to a sentiment label $\mathrm{l} \in L$, where D denotes the set of article sentences and $\mathrm{L} = \{$"positive", "negative", "neutral"$\}$. Rather than estimating C directly, I decompose it into two components: a preprocessing function $\mathrm{h}: \mathrm{D} \rightarrow E$ that transforms text into a numerical representation called embedding where E represent the embedding space, and a classification function $\mathrm{g}: \mathrm{E} \rightarrow \mathrm{L}$ that assigns sentiment labels based on these representations. This modular design enhances flexibility in implementation and improvement. Using the preprocessing procedure, classification function g was calibrated using a labelled dataset consisting of approximately 400,000 sentences annotated by human operators.

The goal of preprocessing is to convert unstructured sentences into structured numerical representations that are both computationally tractable and semantically meaningful. This involves two key steps: tokenization, which breaks a sentence into smaller units (tokens), and embedding, which maps each token into a continuous vector space. The result is a matrix in $\Re^{\mathrm{d_s} \times \mathrm{d_m}}$, where $\mathrm{d_s}$ is the maximum sequence length and d_m is the embedding dimension. Figure 2 illustrates this process.

Specifically, tokenization divides a sentence into smaller linguistic units, called tokens, using a pretrained tokenizer from the spaCy library[3]. These tokens typically represent words, punctuation marks, or other meaningful elements. For example, the sentence "The Bank of Korea announced changes to its monetary policy." is tokenized as ["The", "Bank", "of", "Korea", "announced", "changes", "to", "its", "monetary", "policy", "."]. This step structures the input for further processing while preserving syntactic and semantic content. Consequently, each sentence is transformed into a sequence of tokens that can be systematically analyzed and embedded into a numerical representation.

Preprocessing

Figure 2.



Source: Author's illustration

[3] spaCy is a general-purpose natural language processing library with multilingual support. See Honnibal and Montani (2017) and https://spacy.io and

Following tokenization, each token is mapped to a vector in a continuous, real-valued space using the Word2Vec algorithm[4]. This neural embedding method captures semantic and syntactic relationships by training the model to predict the surrounding words, given a target word (skip-gram) or vice versa (CBOW). In the resulting semantic space, similar words are located close to one another—for example, "Bank of Korea" and "Fed" appear nearby due to their shared institutional role. Semantic proximity is typically measured using the cosine similarity, which quantifies the contextual resemblance between word vectors.

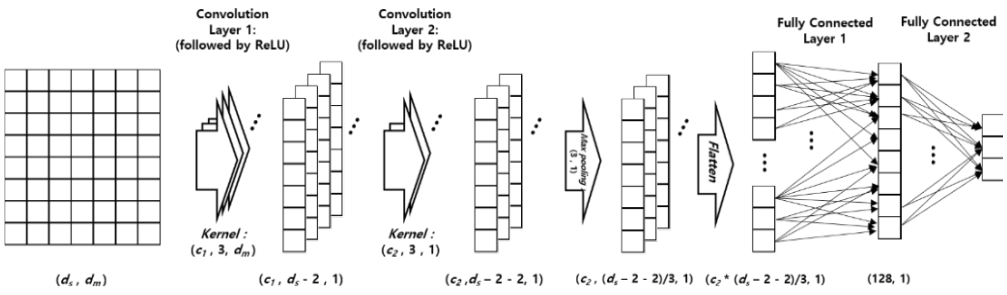
Given a maximum token sequence length d_s and embedding dimension d_m, this step yields a $d_s \times d_m$ real-valued matrix, where each row corresponds to the embedding vector of a token. If the sequence contains fewer than d_s tokens, zero-padding is applied to ensure consistent dimensionality across the inputs. For instance, the sequence ["Bank," "of," "Korea"] with $d_s = 6$ is padded to ["Bank," "of," "Korea," 0, 0, 0], with padding tokens mapped to zero vectors that contribute no semantic information.

Once the matrix representation of a sentence has been prepared, the next step is to construct a classification model g: $\mathrm{E} \backslash \mathrm{rightarrow} \mathrm{L}$ that maps it to a sentiment label. A Convolutional Neural Network (CNN) is employed for this purpose. A CNN interprets the matrix as a grayscale image, where d_s is the sequence length and d_m is the embedding dimension[5]. Similar to how CNNs identify features in images, the network applies filters to the sentence matrix to detect semantic structures that are useful for classifying sentiments.

The CNN used in this study comprises two convolutional layers that extract hierarchical features from the input. The first layer applies c_1 filters to capture localized semantic cues across token sequences, whereas the second layer builds higher-level abstractions from these features with c_2 number of channels. The max-pooling layer follows, selecting the most salient features and reducing the sequence length. The resulting feature maps are flattened and passed through two fully connected layers. The final output is a three-dimensional probability vector over the sentiment classes computed using the softmax function. Figure 3 shows the model architecture.

CNN Classifier

Figure 3.



Source: Author's illustration

Note: d_s, $\mathrm{d_m}$, $\mathrm{c_1}$, and c_2 are set to 100, 128, 16, and 8, respectively.

[4] For the details of the algorithm, see Mikolov et al. (2013).

[5] For a comprehensive review of the CNN, see Li et al. (2021).

2.2 Trending Keyword Extraction

Trending keywords are defined as sets of terms that capture emerging or salient economic issues during the target period. In other words, they were the keywords that were most likely to come to mind when a specific period was recalled. To identify such keywords, I approach the problem from a dual perspective rather than solving the primal problem directly. Specifically, I ask which period is most likely to come to mind when a given keyword is recalled. Based on this perspective, a trending score is computed for each keyword, rooted in the Neyman–Pearson lemma, and used for keyword extraction.

For example, consider the following hypothetical scenario. Suppose you are presented with a set of keywords randomly selected from news articles originating at an unknown point in time. Based on these observed keywords, you are asked to determine whether the articles belong to a specific target period or base (normal) period. Under the formal hypothesis-testing framework, this task is equivalent to testing the null hypothesis that the keywords originate from the base period against the alternative that they originate from the target period.

The Neyman–Pearson lemma states that under certain conditions, the likelihood ratio test is uniformly most powerful for distinguishing between two simple hypotheses. Suppose the true probabilities under the null and alternative hypotheses are given as $\Pr[\,\mathrm{i}\,|\mathrm{base\ period}]$ and $\Pr[\,i\,|target\ period]$, respectively. In the aforementioned keyword-based hypothesis test context, the informativeness of observing keyword i to discern the null and alternative hypotheses is measured by the following score S_i:

$$\mathrm{S_i} = \Pr[\,i\,|target\ period] \times ln\left(\frac{Pr[\,i\mid target\ period]}{Pr[\,i\mid base\ period]}\right).$$

To assess the validity of the trending keyword extraction method, I compare the yearly trending keywords with the evolution of economic sentiment implied in news articles based on the sentiment classification results from the previous section. Specifically, a sentiment classifier is applied using 10,000 randomly sampled article sentences per day. Based on the number of positive and negative sentences obtained from the sentiment classification described above, the following index is constructed:

$$SI_t = \frac{\#\ of\ positive\ sentences\ -\ \#\ of\ negative\ sentences}{\#\ of\ positive\ sentences\ +\ \#\ of\ negative\ sentences}\ .$$

The sentiment index SI_t is normalized and standardized such that its mean and standard deviation are equal to 100 and 10, respectively.

Figure 4 illustrates the evolution of the sentiment index between 2020 and 2023, including the period of the COVID-19 outbreak and its aftermath (solid red line), alongside the corresponding yearly trending keywords extracted from positive (blue headings) and negative sentences (red headings). The figure conveys, in a single comprehensive view, both the trajectory of economic sentiment over time and the key issues underlying these shifts, offering a simultaneous understanding of how and why sentiment has evolved.

Trending Keywords

Figure 4.



Source: Author's illustration based on implementation results.

In 2020, the coronavirus disease (COVID-19) outbreak emerged as the most dominant economic and social issue in Korea. After the first confirmed case on January 20, sentiment began to decline sharply. This downward trend accelerated following the World Health Organization's declaration of the pandemic, reflecting heightened public anxiety and growing economic uncertainty.

The following year, sentiments recovered to some extent, supported by strong government intervention and growing optimism surrounding vaccine development. Although concerns such as the emergence of new variants and public debates over vaccination persisted, the tone of public discourse began to shift. New opportunities, particularly those related to ESG (Environmental, Social, and Governance) themes, have gained traction and were increasingly discussed in a positive light.

At the beginning of 2022, Russia's invasion of Ukraine had a profound and far-reaching impact on the global economy, including that of Korea. The resulting geopolitical instability, coupled with excess liquidity in financial markets, has led to a surge in inflation. In response, central banks worldwide, including the Bank of Korea, tightened their monetary policies through a series of interest rate hikes. These developments remained central throughout 2022 and continued to influence economic sentiment well into 2023.

Interest in ESG themes began to materialize more concretely, particularly through rising expectations for the electric vehicle and battery industries. Additionally, the launch of ChatGPT in early 2023 brought significant public attention to AI and semiconductor technologies, contributing to a rebound in positive market sentiment. Figure 4 aligns closely with the narrative described above, illustrating that the proposed method can serve as an effective tool for monitoring economic conditions in a structured and data-driven manner.

2.3 Summarization

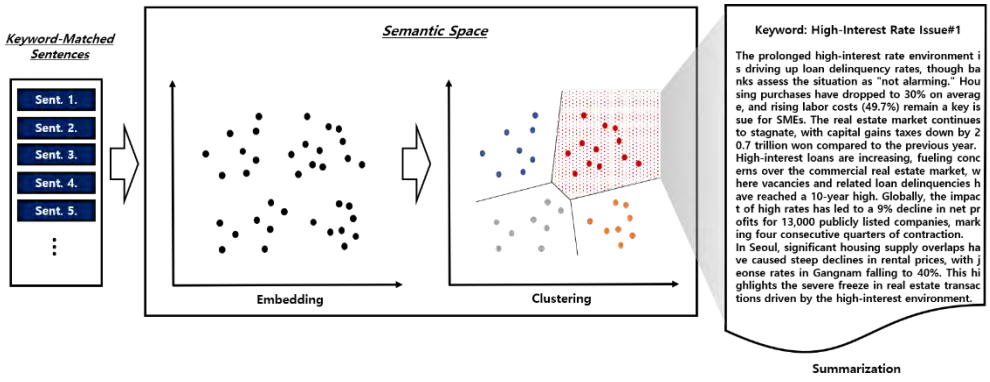
Although trending keywords help reveal the economic issues driving macroeconomic conditions, they often fail to provide a comprehensive perspective. Some keywords can oversimplify or even mislead users because they fail to capture the full range of implications. For instance, Figure 4 shows that the keyword "vaccine" was prominent in 2021 in both positive and negative contexts. However, the keyword alone does not clarify whether the discussion focuses on breakthroughs in vaccine development or concerns over side effects and public resistance.

The goal of the summarization step is to produce outputs that are succinct yet intellectually accessible while preserving the multidimensional perspectives and sentiments surrounding the trending keywords. It consists of three key components: embedding, clustering, and summarizing. First, each sentence containing a given keyword is embedded into a semantic space. In the system implemented in this study, the Word2Vec algorithm is employed for this task. Second, the embedded sentences are clustered using the K-means algorithm[6]. Because the semantic space preserves semantic similarity, clustering based on vector distance allows each cluster to be interpreted as distinct. Third, each cluster representing a specific issue is summarized using a Large Language Model (LLM). In this implementation, a version of LLaMA 3.2[7] is used to generate these summaries.

The output of this process is presented in Figure 5, which illustrates how the system operates and the results produces. A set of positive sentences containing the keyword "high-interest rate" was processed using the aforementioned procedure. The system organizes these sentences by subject matter into issue clusters, and each cluster is then summarized into a concise description of the core issue. In experiments with various keywords across different periods, the method consistently succeeded in summarizing the underlying issues within several minutes on a personal laptop.

Trending Keywords

Figure 5.



Source: Author's illustration based on implementation results.

[6] For the implementation, FAISS is used. For the details, see Johnson and Jégou (2017).

[7] Specifically, llama-3.2-Korean-Bllossom-3B is used. See Choi et. al (2024).

3. Concluding Remarks

Motivated by the desire to monitor economically significant issues in a timely manner, this study presents a method for extracting and summarizing such information from a large volume of news articles. Owing to recent advancements and the rapid commoditization of natural language processing (NLP) technology, the proposed method is easy to implement on a personal laptop without requiring significant investment in specialized hardware or expensive software. The ease of implementation and cost-efficiency of the proposed method lowers barriers for many central bankers and economists, helping alleviate their reluctance to expand their monitoring scope beyond traditional structured data. Despite its simplicity, this approach is versatile and can be adapted to various policy needs. For example, in environments where specific economic issues, such as bank stability, become prominent, this method can be used to monitor related issues in greater depth and in a timely manner. Additionally, with slight modifications, it can be employed for comprehensive monitoring of central bank communications, allowing policymakers to adjust the content or tone of their messaging by closely tracking how the media perceives and frames their communications.

Despite its contributions, this study had several limitations. Fine-tuning and data scalability are notable examples; however, other important challenges remain. Fine-tuning is required to tailor the system to specific use cases. The experimental results were generally robust, but the quality of the summarization was influenced by both the number of clusters and sentences included. Therefore, careful calibration of cluster counts relative to sentence volume is necessary. A more fundamental issue is whether the data scope can be expanded beyond news articles to include social media data. Social media can serve as a valuable complementary source for identifying public attention to economic issues owing to its immediacy and scale. However, unlike news articles written in formal or edited language, social media content tends to be noisy and linguistically inconsistent. To effectively leverage such diverse data sources, future research must develop methods capable of handling this linguistic variability and noise.

References

Choi, C., Jeong, Y., Park, S., Won, I., Lim, H., Kim, S., Kang, Y., et al. (2024). Optimizing language augmentation for multilingual large language models: A case study on Korean. *arXiv Preprint, arXiv:2403.10882. https://doi.org/10.48550/arXiv.2403.10882*

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *https://spacy.io*

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, *33*(12), 6999-7019.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Extracting Economic Issues from News Data: With the Help of Generative AI

Younghwan Lee

Bank of Korea

Disclaimer: The views expressed in this talk do not reflect necessarily those of Bank of Korea

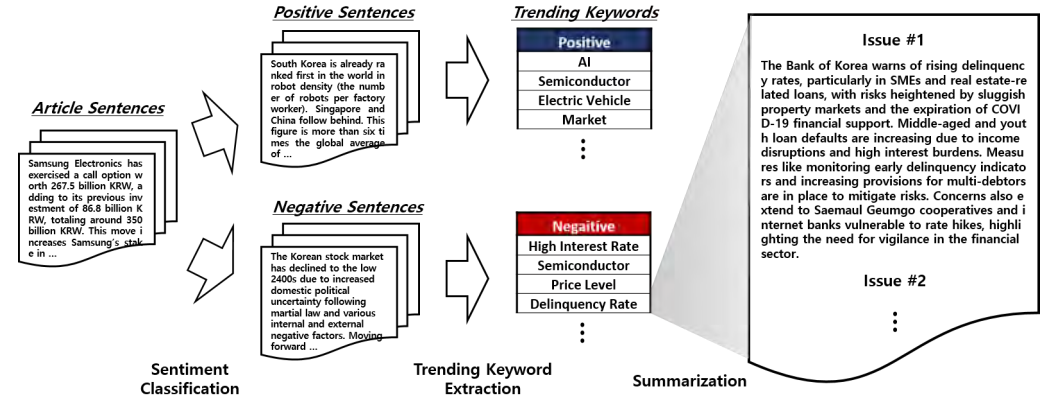
Motivation

- Traditional economic monitoring frameworks that rely on economic variables such as GDP and CSI might only be a second-best method.
  - Doctors treat diseases, not symptoms.
  - Similarly, policymakers must address economic issues, not just economic variables.

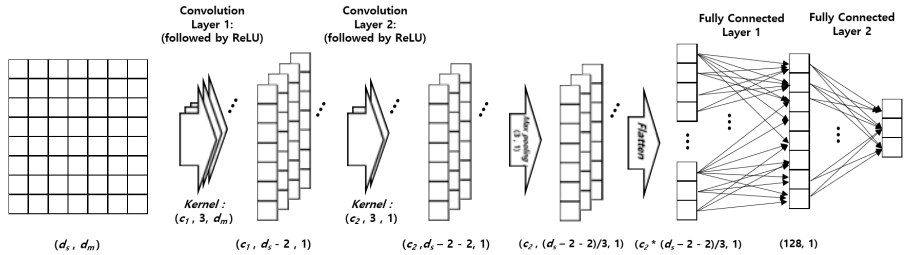
- However, directly monitoring economic issues is challenging in practice, as they are often linguistic rather than numeric.

- In this study, I present a methodology to incorporate economic issues into the monitoring framework that is:
  i) Verifiable,
  ii) Flexible,
  iii) **Easy to implement**.

Snapshot

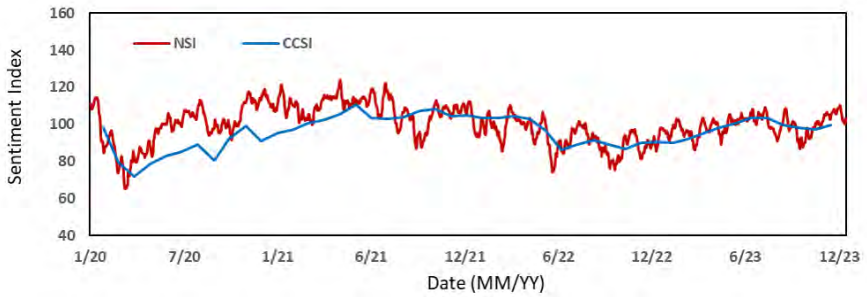

Article Sentences
Samsung Electronics has exercised a call option worth 267.5 billion KRW, adding to its previous investment of 86.8 billion KRW, totaling around 350 billion KRW. This move increases Samsung's stake in ...
Positive Sentences
South Korea is already ranked first in the world in robot density (the number of robots per factory worker). Singapore and China follow behind. This figure is more than six times the global average of ...
Negative Sentences
The Korean stock market has declined to the low 2400s due to increased domestic political uncertainty following martial law and various internal and external negative factors. Moving forward ...
Sentiment Classification
Trending Keywords
Positive
AI
Semiconductor
Electric Vehicle
Market
Negaitive
High Interest Rate
Semiconductor
Price Level
Delinquency Rate
Trending Keyword Extraction
Issue #1
The Bank of Korea warns of rising delinquency rates, particularly in SMEs and real estate-related loans, with risks heightened by sluggish property markets and the expiration of COVID-19 financial support. Middle-aged and youth loan defaults are increasing due to income disruptions and high interest burdens. Measures like monitoring early delinquency indicators and increasing provisions for multi-debtors are in place to mitigate risks. Concerns also extend to Saemaul Geumgo cooperatives and internet banks vulnerable to rate hikes, highlighting the need for vigilance in the financial sector.
Issue #2
Summarization

Classification: Methodology



- CNN (Convolutional Neural Network) is applied to classify sentiment into positive, neutral, and negative categories.
  - i) Tokens are generated using **spaCy** and embedded into 128-dimensional vectors via **Word2Vec**.
  - ii) Token vectors are concatenated into a 100×128 matrix, with zero-padding added if needed for consistent dimensions.

Classification: Result



- Sample 10,000 sentences daily to derive an index. The results show it is highly correlated with existing sentiment indices.
  - $$X_t = \frac{\#\,\text{Positive Sentences} - \#\,\text{Negative Sentences}}{\#\,\text{Positive Sentences} + \#\,\text{Negative Sentences}}$$
    - Scale and translate X_t to resemble a standardized index (Mean = 100, Std. = 10).

Trending Keyword: Idea

- It is inefficient to directly work with a large set of sentences. We need a method to focus on a few important issues that characterize current economic conditions.
- ***Trending Keywords*** are defined as keywords extracted from news headlines during a specific time period that best describe the key issues of that period.
- Then, how can we capture the ***Trending Keywords***?
- Suppose that you are provided with a collection of news headlines, and you are asked to guess when those headlines were published based on the keywords in the headlines.
- I define ***Trending Keywords*** as a set of keywords that are most useful for us to address the question above.

Trending Keywords: Methodology

- The Neyman-Pearson lemma states that the log-likelihood test is the most powerful test. By utilizing the empirical likelihood ratio of keyword frequency, we can address the question at hand.

- For each keyword $i \in I$ define trending score S_i as follow:

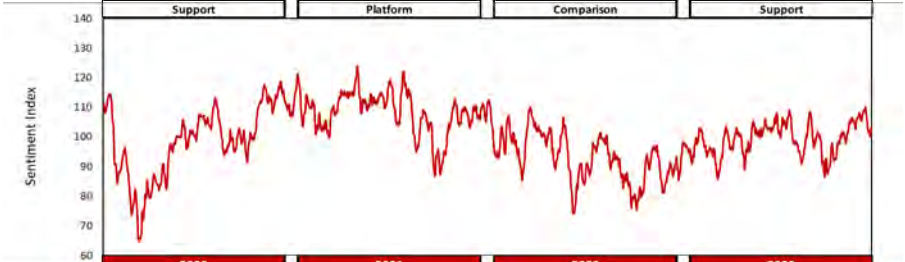
$$S_i = \Pr[i|\text{target period}] \ln \frac{\Pr[i|\text{target period}]}{\Pr[i|\text{base period}]}.$$

- S_i increase when the keyword i
  - i) is encountered more frequently in the headlines during the target period,
  - ii) is relatively uncommon in the base period.
- We can extract ***Trending Keywords*** for the target period by ranking keywords according to their trending score.

Trending Keywords: Result

- ***Trending Keywords*** from 2020 summarize the captures important issues after COVID-19 outbreak.

2020	2021	2022	2023
COVID19	COVID19	ESG	AI
Spread	Vaccine	Electric Vehicle	Semiconductor
Crisis	ESG	LG Energy Solution	Electric Vehicle
Non-face-to-face	Eco-friendly	Price	Market
Support	Platform	Comparison	Support



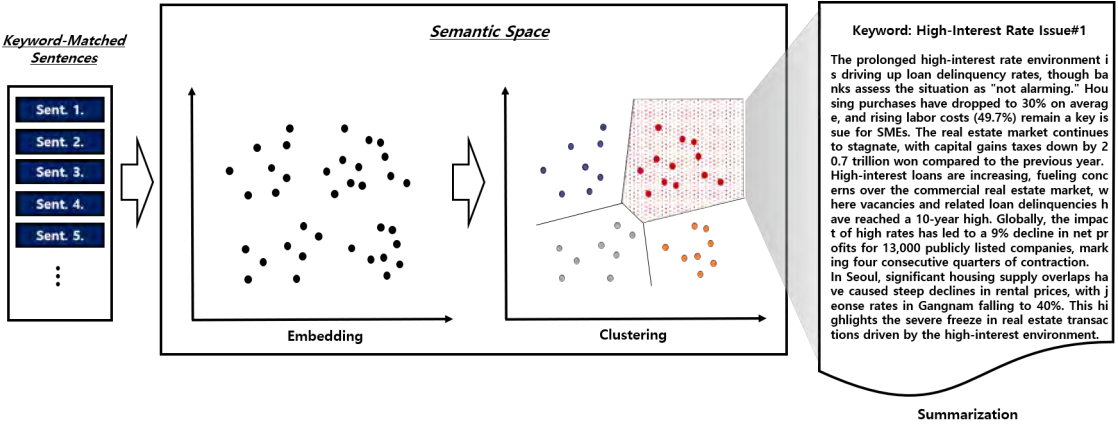
2020	2021	2022	2023
COVID19	COVID19	Ukraine	High Interest Rates
Spread	Vaccine	Inflation	Semiconductor
Crisis	Variant	Russia	Inflation
Cases	Inflation	Interest Rates	Delinquency Rate
Resurgence	Bitcoin	Increase	Real Estate

Summarization: Methodology

- The goal of summarization is to extract information comprehensively and deliver it succinctly.
  - Working with keywords is efficient but often lacks clarity and sufficient context.
  - Reviewing all keyword-matched sentences is inefficient as many convey similar meanings.

- Strategy:
  i) **Embed**: Transform keyword-matched sentences into semantic vectors (**Word2Vec**).
  ii) **Cluster**: Group sentences by semantic similarity using K-means (via **FAISS**).
  iii) **Summarize**: Generate concise summaries for each cluster with a pretrained LLM (**Llama 3.2**).

Summarization: Result



Concluding Remarks

- This study proposes a verifiable, flexible, and easy-to-implement methodology to extract economic issues from a large volume of news articles.
  - With the help of generative AI and pretrained LLMs, the unstructured nature of economic issues is effectively addressed.
- Future Research:
  - Exploring how to predict next quarter's trending keywords and summarized issues.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

What skills are employers looking for? Applying generative AI to Danish job postings[1]

Per Andersen and Erik Grenestam,
Danmarks Nationalbank

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

DANMARKS
NATIONALBANK

What skills are employers looking for?
Applying generative AI to Danish job postings

4th IFC workshop on data science in central banking, February 18th 2025

Per Andersen & Erik Grenestam, Danmarks Nationalbank

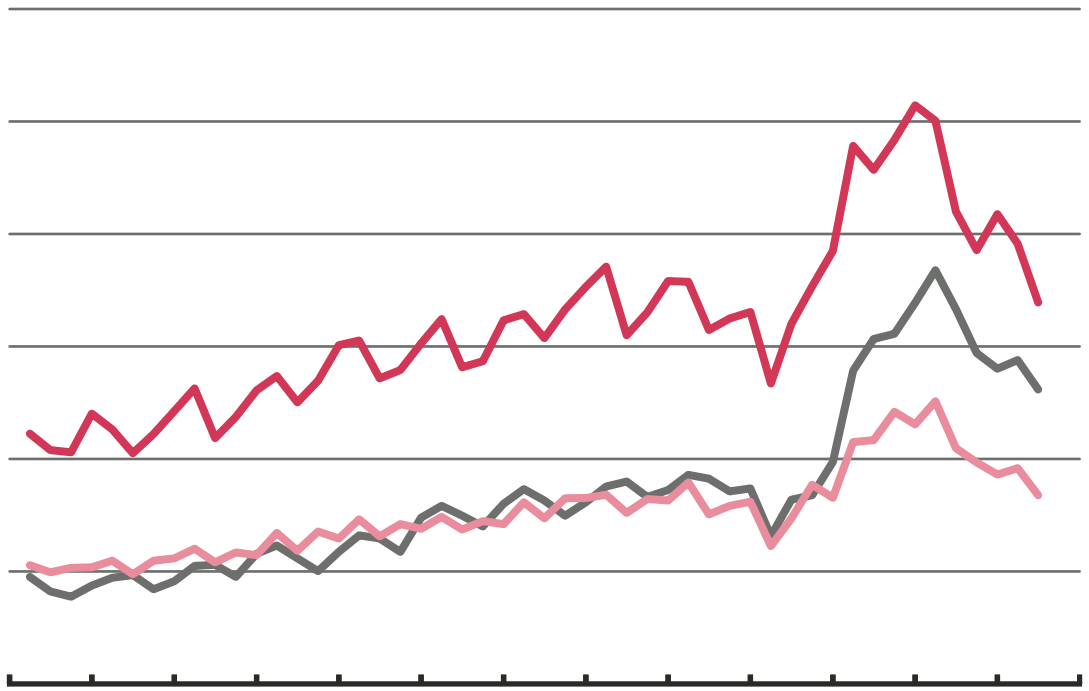
Our goal is to build a useful dataset describing skill demand in Denmark

~ 3.2 million job postings from *jobindex.dk* (2011-2023)

Generative AI

Panel of granular **skill mentions** using an established **skill taxonomy**

Our job postings data cover most of the open positions in Denmark



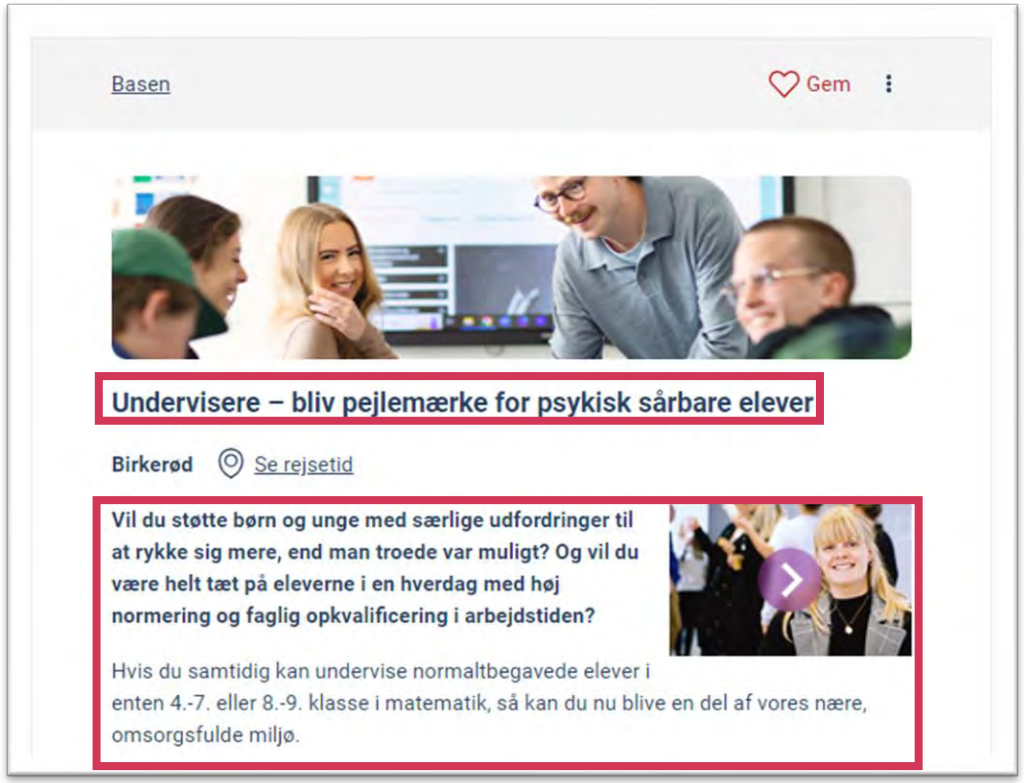
DANMARKS NATIONALBANK

Why should we care about skills?

- Labor economics has seen a paradigm shift from a static view of labor to a task- and skill-based approach (Autor, 2013).
- Skill-biased technological change shifts the demand for specific skills (Acemoglu & Restrepo, 2018; Autor, 2019).
- Evidence of a rising premium for non-cognitive (”soft”) skills (Edin et al., 2017)

→ **Data on skill demand could be critical for understanding labor market trends**

We extract a list of required skills, competencies, and knowledge using OpenAI's GPT 4o-mini



You are given a job posting written in Danish. Your tasks are:

1) *Extract all the required skills, competencies, and knowledge.*
2) *Identify the job title as accurately as possible.*

You must perform both tasks with equal care and thoroughness. Focus first on identifying and listing all relevant skills and competencies in detail. Then, identify the job title. Translate your findings into English. Do not include any information about the employer.

{job posting text}

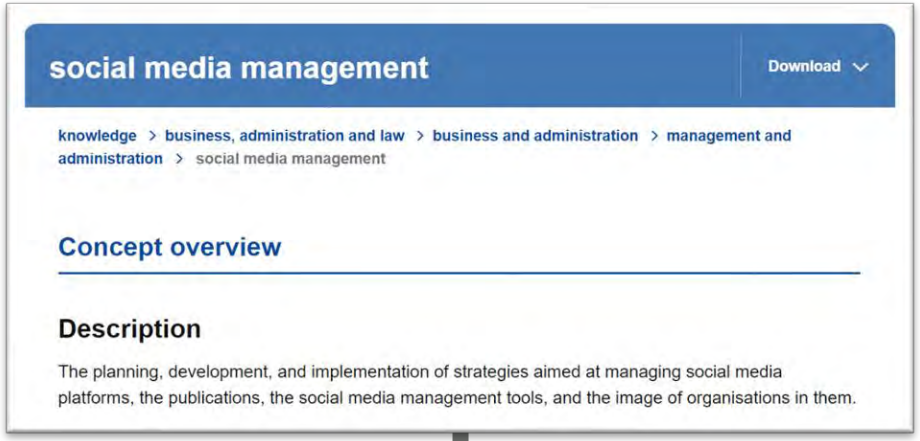
Matching the GPT response to ESCO enables us to create a structured dataset of skills

Example response from GPT 4o-mini

Skill	Job title
planning and coordinating content	Social Media Specialist
performance reporting and analysis	
influencer management	
trend spotting	
interest in social media	
aesthetic sense	
experience with trends	
fluency in English (written and spoken)	
strategic thinking	
creative thinking	
content creation for social media	

For structure, we match each skill to it's closest ESCO skill

ESCO (*European Skills, Competences, Qualifications and Occupations*) is a classification system maintained by the European Commission. ESCO lists about 14.000 skills that are "relevant to the European labor market".



Pre-trained embedding model

DANMARKS NATIONALBANK

We use an LLM to choose the best ESCO matches for each extracted skill-title pair

- The embedding model reduces the scope of the classification problem from ~14.000 down to 5-10 labels.
- However, the closest match is not always the best…

Skill	Job title	ESCO title	Rank
perform statistical analysis using SAS	biostatistician	mathematics and statistics	1
		apply statistical analysis techniques	2
		SAS language	3
		statistical analysis system software	4
		statistical modeling techniques	5
		statistics	6
		data analytics	7

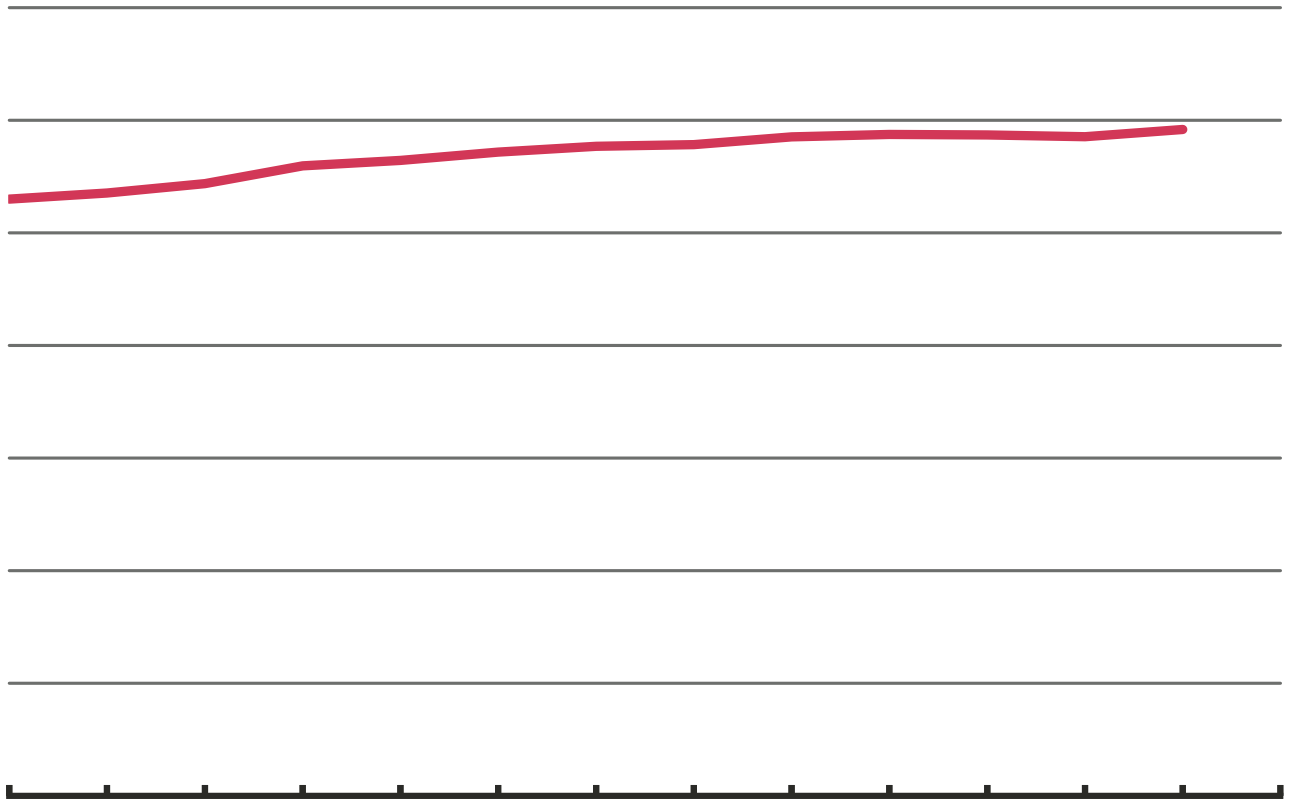
Prompt for reranking labels

The skill *'perform statistical analysis using SAS'* is required for the job *'biostatistician'*. Which of the following options is the best match for this skill?

1) mathematics and statistics
2) apply statistical analysis techniques
3) SAS language
4) statistical analysis system software
5) statistical modeling techniques
6) statistics
7) data analytics
8) none of the above

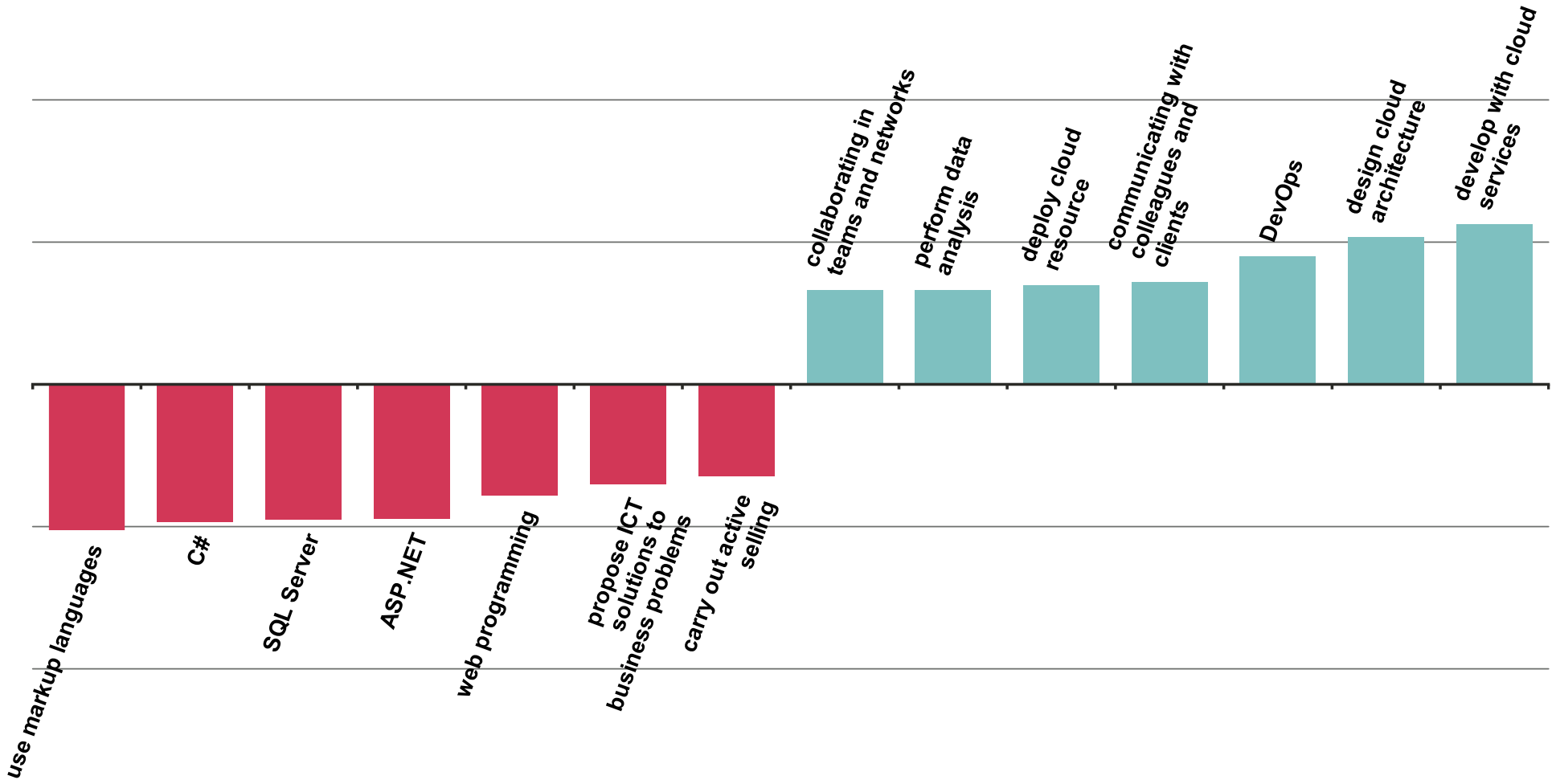
24 million prompts later…

- We’ve extracted 36.3 million skills mentioned across 3.2 million job ads.
- In 2023, we identify 11.8 skills per ad, up from 10.5 in 2011.
- Total cost of tokens: **~$800**



DANMARKS NATIONALBANK

Case: Skill trends in the information sector reflect programming languages falling out of favor and the rise of the cloud



DANMARKS NATIONALBANK

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

The monetary policy statement database[1]

Cory Baird,
University of Tokyo

Jonathan Benchimol,
Bank of Israel

Vira Vyshnevska and Wook Sohn,
KDI School

Iegor Vyshnevskyi,
Sogang University

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

The Monetary Policy Statement Database: An LLM Application to Global Financial Conditions*

Cory Baird† Jonathan Benchimol‡ Wook Sohn§ Vira Vyshnevska§
Iegor Vyshnevskyi¶

March 11, 2026

Abstract

This study introduces the Monetary Policy Statement Database (MPSD), comprising 6,693 statements from 51 central banks worldwide (1990–2024). We develop a reproducible pipeline combining standard natural language preprocessing with large language model (LLM) tools for cross-country analysis. Four key findings emerge. First, statements lengthened substantially after the Global Financial Crisis while readability improved modestly. Second, inflation references comove across countries during global inflation episodes. Third, LLM-based question answering and aspect-based sentiment reveal that central banks attribute global financial conditions primarily to broad U.S. macroeconomic developments rather than to Federal Reserve policy actions specifically. Fourth, using a benchmark dictionary-based sentiment index and LLM-derived aspect-based sentiment indicators, Granger causality tests suggest that statement sentiment predicts the Global Financial Cycle rather than merely responding to it. The MPSD and accompanying codebase support reproducible research on monetary policy communication and international transmission.

Keywords: Central bank communication, Large language models, Text analysis, Generative database, Machine learning.

JEL Classification: C55, C63, E52, E58, G15.

*This paper does not necessarily reflect the views of the Bank of Israel or any other institution.

†University of Tokyo, Tokyo, Japan. Corresponding author. Email: cbaird@g.ecc.u-tokyo.ac.jp
‡Bank of Israel, Jerusalem, Israel.
§KDI School, Sejong, South Korea.
¶Sogang University, Seoul, South Korea.

1 Introduction

The evolution of monetary policy over recent decades has elevated central bank communication to a primary policy instrument, fundamentally transforming how monetary authorities influence economic outcomes (Woodford, 2005; Blinder, 2018). As former Federal Reserve Chairman Ben Bernanke observed, monetary policy has become "98 percent talk and 2 percent action," reflecting the critical role of communication in shaping market expectations and enhancing policy effectiveness (Blinder et al., 2008). This transformation has become particularly salient as central banks increasingly rely on forward guidance and other communication-based tools to influence both financial markets and household expectations (Hansen et al., 2019; Coibion et al., 2022).

The empirical analysis of central bank communications faces several critical constraints that limit our understanding of how monetary policy messages are crafted, transmitted, and interpreted across global markets. First, despite significant advances in communication strategies (de Haan and Hoogduin, 2024), the absence of a centralized, standardized database of monetary policy communications has restricted analysis to individual institutions (Born et al., 2014; Benchimol et al., 2025) and small groups of advanced or emerging economies (Cieslak and Schrimpf, 2019; Vyshnevskyi et al., 2024). Current research relies primarily on the Bank for International Settlements' repository of speeches, leaving a significant gap in the systematic analysis of monetary policy statements (Picault and Renault, 2017). Second, the lack of robust, open-source preprocessing pipelines has hindered the reproducibility of text-based monetary policy research, a methodological concern highlighted by Gentzkow et al. (2019) and Hansen and McMahon (2016). Third, while natural language processing (NLP) techniques have advanced rapidly, their application to monetary policy analysis has been constrained by insufficient computational infrastructure and domain-specific training data.

Recent studies confirm the complex and crucial influence of central bank communication. Hansen and McMahon (2016) demonstrate how linguistic analysis effectively reveals policy intentions and economic forecasts. Kryvtsov and Petersen (2021) provide experimental evidence on the effectiveness of different communication strategies, while Hayo and Neuenkirch (2015) document communication's crucial role during financial crises. These studies, alongside research on information effects by Nakamura and Steinsson (2018) and Bauer and Swanson (2023), highlight both the potential and challenges of systematically analyzing monetary policy communications.

This paper addresses these challenges by introducing the Monetary Policy Statement Database (MPSD), a comprehensive infrastructure for analyzing monetary policy statements across 51 central banks. Our contribution is fourfold.

First, we develop and publicly release a standardized database of monetary policy statements enabling rigorous cross-country analysis of central bank

communications. This resource empowers researchers to investigate policy transmission and coordination across diverse economies, including those beyond the traditional focus on major central banks (Armelius et al., 2020).

Second, we implement a sophisticated technical architecture that combines advanced NLP capabilities with robust code and data management tools. The infrastructure leverages data pipeline management (DVC), version control (Git), and experiment tracking while integrating cloud computing solutions through Amazon S3 buckets for storage and Hugging Face for open-source language model integration. This framework supports automated text summarization and question answering (QA) systems specifically designed for central bank applications, enabling efficient analysis of large-scale monetary policy communications. Following best practices in computational reproducibility (Gentzkow et al., 2019), our implementation uses object-oriented programming principles, with modular Python classes providing standardized tools for text preprocessing and advanced NLP analytics.

Third, we offer a practical framework enabling users to leverage large language models (LLMs), including Claude, Gemini, and ChatGPT. Building on methodological advances in text analysis, we develop tools for extracting policy-relevant information from complex monetary policy documents (Gonzalez and Tadle, 2022).

Fourth, we provide novel empirical evidence on the communication channel of international policy transmission. Our analysis yields four key findings with implications for understanding global monetary policy coordination: (i) Central banks worldwide exhibit remarkable narrative convergence during major economic shocks, with communications showing strong thematic comovement that precedes—rather than follows—observable financial market stress. (ii) When attributing global financial conditions to external factors, central banks systematically cite broad U.S. economic developments (housing markets, consumer spending, growth prospects) rather than Federal Reserve policy actions specifically. This distinction has important implications for modeling policy spillovers. (iii) Dictionary-based communication sentiment and LLM-derived sentiment indicators Granger-cause movements in the Global Financial Cycle (GFC), suggesting that central bank narratives actively shape global financial conditions rather than merely describing them. (iv) Substantial heterogeneity exists in how advanced-economy versus emerging-market central banks discuss global conditions, with emerging markets showing more idiosyncratic patterns, particularly at times of financial stress. These findings suggest that communication may be an underappreciated channel through which monetary policy transmits across borders.

Table 1 illustrates our approach to ensuring end-to-end transparency in artificial intelligence (AI) applications for monetary policy analysis. The pipeline shows how we maintain scientific rigor from raw data collection through preprocessing, model inference, and analysis—addressing concerns about

reproducibility in computational economic research, particularly when applying advanced language models to policy analysis.

Table 1: End-to-End Transparency for Replicable AI

Data Collection	**Featurization**	**Replicable AI Use**
Identify statement links	Cleaning	Replicable LLM infrastructure
Scrape statements	Sentence segmentation	Track inputs (prompts, texts)
Automate retrieval	Part-of-speech tagging	Track outputs (responses)

Source: Authors.

We leverage our data and AI infrastructure to investigate several key questions regarding monetary policy, the Global Financial Cycle (Miranda-Agrippino and Rey, 2020; Rey, 2015), and the transmission of policy narratives across borders (Armelius et al., 2020). We employ LLMs for QA and aspect-based sentiment (ABS) analysis. In QA, we prompt the model to summarize the drivers of the global economy, particularly focusing on instances where central banks identify the Federal Reserve as a key influence. For ABS, we use LLMs to extract sentiment for specific themes related to the impact of global financial conditions on domestic economies, such as economic growth, financial conditions, capital flows, and monetary policy spillovers.

Our findings from this LLM analysis reveal that central banks often cite the U.S. economy as a driver of global economic conditions rather than specifically attributing influence to the Federal Reserve. This nuanced distinction—where broader economic factors in the United States are perceived as more impactful than central bank actions alone—has important implications for modeling policy spillovers. Furthermore, our sentiment analysis shows that central banks expressed greater concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period, likely reflecting the lasting impact of the Global Financial Crisis and increased global economic interconnectedness.

While recent discussions of AI in economics have focused primarily on closed-source foundation models, we emphasize the fundamental importance of robust, open-source data science infrastructure. The MPSD provides comprehensive tools for text analysis specifically designed for central bank communications, enabling researchers to extract and analyze policy narratives systematically across institutions and time. This approach builds on seminal work in computational text analysis (Gentzkow et al., 2019) while addressing specific challenges in monetary policy research (Hansen and McMahon, 2016).

Our work connects to the growing literature on international monetary policy spillovers. Obstfeld (2015) highlights the challenges that global financial integration poses for monetary policy independence, Bruno and Shin (2015) demonstrate significant effects of U.S. monetary policy on capital flows to

emerging markets, and Kalemli-Özcan (2019) show that these spillovers primarily operate through global banks and risk premia in international asset markets. We extend this literature by focusing specifically on the communication channel of policy spillovers, providing direct evidence that central bank narratives—beyond policy actions themselves—significantly influence global financial conditions. Our findings complement Brusa et al. (2020), who document that ECB communications affect U.S. asset prices through signaling about global economic conditions.

Our work contributes to multiple strands of literature. First, we extend research applying NLP to central bank communications (Gonzalez and Tadle, 2022), moving beyond specific aspects like sentiment (Loughran and McDonald, 2011; Correa et al., 2021) or clarity (Bulír et al., 2013; Benchimol et al., 2023) to provide a comprehensive analytical framework. Second, we advance the literature on monetary policy transmission by enabling systematic cross-country analysis of communication strategies and their effects (Hansen et al., 2019; Picault and Renault, 2017). Third, we contribute to the literature on global financial cycles and policy spillovers (Miranda-Agrippino and Rey, 2020; Armelius et al., 2020). Fourth, we extend the methodological literature on text analysis in economics (Gentzkow et al., 2019; Benchimol et al., 2022) by developing reproducible procedures specifically for central bank communications. This project also complements recent large-scale data collection efforts in central bank communication research. In particular, Campiglio et al. (2025) introduce the cbspeeches.com corpus comprising over 35,000 speeches from 131 central banks (1986–2023). While their corpus focuses on speeches and covers a broader institutional range, our database concentrates specifically on monetary policy statements—the official communications released in direct connection with policy rate decisions. The two resources are therefore complementary and together support a comprehensive ecosystem for reproducible research on central bank communication.

Our approach also connects to recent surveys on central bank communication. Masciandaro et al. (2024) provide a comprehensive review of how central banks have evolved from silence to active social media engagement, documenting techniques used to analyze communication effectiveness. Our infrastructure complements their focus on social media by providing systematic analysis of formal policy statements. Additionally, our work relates to the broader literature on monetary policy transmission mechanisms. De Grauwe and Ji (2023) demonstrate that using forward-looking data in monetary policy affects policymaking quality differently across policy regimes, while Kumar et al. (2023) show that market volatility significantly influences the term premium through channels interacting with monetary policy. These findings underscore the importance of understanding how central bank communications shape expectations and market responses—precisely the mechanisms our database enables researchers to investigate.

The remainder of the paper is organized as follows. Section 2 details the

technical architecture and construction of the MPSD, including data collection procedures, preprocessing pipelines, and the LLM inference framework. Section 3 presents an exploratory analysis of the database, highlighting key patterns in central bank communications across time and institutions. Section 4 examines our application of generative AI to central bank communications, focusing on extracting economic narratives related to the GFC. Section 5 discusses policy implications and potential extensions, and Section 6 concludes.

2 Database Construction and Architecture

The MPSD constitutes a comprehensive research infrastructure designed to facilitate systematic analysis of monetary policy communications across global central banks. The database encompasses monetary policy statements from 51 central banks, with coverage varying by institution. Statement frequency ranges from monthly to quarterly depending on each central bank's policy framework, yielding 6,693 unique policy statements.

We define a monetary policy statement observation as a text communication released in direct connection with a policy rate decision event (whether a change or no change in rates). Because central banks employ varying naming conventions—including "decisions," "statements," "announcements," and "press releases"—we systematically map each institution's official document series to the corresponding decision dates. This mapping is fully documented in the accompanying metadata file. For example, the ECB's communication evolved from the "Introductory Statement" (1999–mid-2021) to the "Monetary Policy Statement" (since mid-2021), while short "Monetary Policy Decisions" press releases—which historically contained only rate decisions without supporting rationale—are not treated as equivalent to full policy statements in our corpus. Similarly, for the Reserve Bank of Australia, we include the post-meeting "Monetary Policy Decision" statements rather than the separate quarterly "Statement on Monetary Policy" inflation report. The corpus excludes ancillary documents such as inflation reports, minutes, and speeches, which serve different communicative functions. All documents were preprocessed to remove non-textual elements including tables, charts, and lists of meeting attendees, ensuring that linguistic analysis captures substantive policy communication content.

Monetary policy communications take different formal labels across central banks, including "decisions," "statements," and "press releases," as documented in Table 4. While these document types can be identified, they are not treated as separate analytical categories in this study. Despite differences in naming conventions and format, these communications are released in direct connection with policy rate decision events (change or no change) and serve a functionally comparable role: conveying the policy outcome together with varying degrees of economic assessment and forward-looking guidance. In many cases, a single

document simultaneously fulfills multiple functions, making a clean and objective partition by document type difficult. To preserve transparency and cross-country comparability, we therefore treat these texts as institution-specific realizations of a common policy-decision communication instrument, while explicitly accounting for observable sources of possible heterogeneity such as language of publication.

The technical architecture implements three fundamental design principles that address specific challenges in monetary policy research. The infrastructure enables scalable data collection and standardization through advanced web scraping and preprocessing pipelines, essential for maintaining consistency across the diverse spectrum of central bank websites and communication formats (PDF, DOC/DOCX, HTML). We also maintain data integrity through version control, addressing reproducibility challenges identified in computational social science research (Gentzkow et al., 2019).

2.1 Technical Infrastructure Implementation

The MPSD ensures integrity of economic research data by strictly enforcing lineage. Using Data Version Control (DVC), the system tracks the full data lifecycle—from web scraping to LLM-based analysis—via parameterized pipelines. To support full computational reproducibility, the architecture couples Git for code versioning with uv for unified Python version and dependency management (ensuring that data manipulation tools such as pandas work consistently across different environments).

Table 2 illustrates the LLM infrastructure we developed specifically for central bank communication analysis. This diagram shows the integration of multiple components (modules), including text retrieval, preprocessing, and LLM integration. The structure enables seamless incorporation of different AI models (such as ChatGPT and Gemini, as well as open-source models hosted on Hugging Face) and supports various analytical tasks including classification, summarization, and QA. This modular design ensures reproducibility while maintaining flexibility for users to adapt the infrastructure to their specific use cases.

Data storage relies on Amazon S3 buckets managed directly through DVC. The system segregates assets into distinct partitions—raw data, processed outputs, and model artifacts—ensuring clear data lineage and secure access control.

2.2 Code and Folder Infrastructure

The MPSD employs a modular design with discrete functional components for text retrieval, text cleaning, and LLM inference. As illustrated in Figure 1, the codebase follows a standardized folder structure based on the "cookiecutter" template for data science projects. This approach organizes the project into logical sections to facilitate maintenance and collaboration:

Table 2: AI Infrastructure for Central Bank Communication Analysis

Modular LLM-based Data Processing Framework for Academic Research		
Data Acquisition Module (Abstract Class)	**LLM Processing Module (Abstract Class)**	**LLM Data Pipeline Workflow**
Methods: 1. Initialize (source parameters) 2. Scrape Statement Links 3. Download/Save Statements	**Methods:** 1. Initialize (LLM Model) 2. Process/Load Prompts 3. Generate Response	**Methods:** 1. Initialize (Input Corpus) 2. Filter via Keywords or POS 3. LLM Inference 4. Save Response
Output: Text Files CSV/Parquet Text Dataframes	**Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML)	**Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) Figure for Research (PNG/PDF)
Abstraction: Extend abstract functions to accommodate new data sources (e.g., central bank speeches, minutes) and communication modes.	**Abstraction:** Implement model-agnostic functions to leverage diverse LLM architectures (e.g., Gemini, Claude, HuggingFace models).	**Use-Case:** Utilize concise, reusable scripts for seamless integration with LLM APIs (e.g., Gemini, Claude) by researchers.

Source: Authors.

- `src`: Contains Python scripts organized by functionality
- `data`: Stores raw and transformed text files and processed outputs
- `references`: Houses GenAI prompts and documentation

Complementing this structure, the system implements automated database updates through DVC pipelines. This enables researchers to access data and replicate the entire analysis—including the graphs and charts utilized in this paper—with a single command. This structure ensures end-to-end reproducibility from data manipulation through NLP analysis, addressing a critical need in text analysis.

2.3 Data Collection and Processing Framework

Our database houses two data types: statements retrieved programmatically via Python, stored in the 'API' folder, and statements downloaded manually, termed 'static downloads' (Vyshnevskyi et al., 2024). As illustrated in Table 2, the process for programmatically sourced data uses an agnostic class with three abstract methods: (1) downloading statement links, (2) iterating through these links to download each statement, and (3) parsing each statement and saving the output as Parquet and CSV files. The pipeline's final stage applies format-specific processing for PDF, Word, and HTML documents.

Figure 1: Folder Structure Based on "Cookiecutter" Template for Data Science Projects

```
project_root/
├── src/
│   ├── data_processor.py
│   ├── figure_generator.py
│   ├── processor.py
│   ├── pipeline.py
│   ├── sentiment_lexical.py
│   ├── sentence_split.py
│   └── web_scraper.py
├── references/
│   ├── dictionaries/
│   │   ├── centralbank_info.py
│   │   ├── dict_regions.py
│   │   ├── dict_country_converter.py
│   │   └── sentiment_label.py
│   ├── lexical_sentiment_dictionary/
│   │   └── lexical_master_dict.xlsx
│   └── paper_econsurveys/
│       ├── all_figures_settings.yaml
│       ├── llm_sentiment_prompt.yaml
│       └── llm_gfc_driver_prompt.yaml
├── data/
│   └── database/
│       └── txt_files/
├── external/
├── sentences/
└── all_analytical_output/
    ├── cache/
    │   └── [Intermediate processed data (.parquet)]
    ├── tables/
    │   └── [Final manuscript tables (.xlsx, .tex)]
    ├── sentiment/
    │   └── llm_processsor/
    │       └── [LLM sentiment scoring outputs (.csv, .parquet)]
    └── figures/
        └── [15+ Final Analytical Figures (.pdf)]
```

Source: Authors.

The Python code employs a modular structure, maintaining an agnostic design while permitting targeted modifications. Base classes define abstract interfaces

(including the three methods previously mentioned), complemented by specific implementation classes tailored to individual central bank websites to handle their variations. We have decided not to release the scraping code due to concerns about potential misuse that could overload central bank web servers. This approach is consistent with established practices in empirical economics, where reproducibility is defined by the ability to replicate published results using shared data and code rather than by replicating the original data collection process itself. We hope our project highlights best practices for similar text database initiatives, demonstrating how abstract and agnostic designs can help optimally organize scraping code.

Continuous integration through GitHub Actions implements sophisticated workflows for database maintenance and updates. These workflows allow us to perform frequent checks for new policy statements, enabling us to cross-check any findings against the links stored in our system.

2.4 LLM Inference Framework

The MPSD incorporates a framework for leveraging language models in central bank applications. We created an abstract LLM class that seamlessly integrates Claude, ChatGPT, Gemini, and open-source models hosted on Hugging Face. This class provides methods to select models, process prompts from YAML files, and generate and save responses, enabling researchers to test various prompts and assess the robustness of their results across multiple models.

2.5 Future Development and Maintenance

The MPSD features a forward-looking architecture designed for scalability to incorporate other types of central bank communication, such as speeches, summaries, and minutes. The modular design also enables easy addition of new central banks. Furthermore, integration with GitHub and DVC enables automated workflows that ensure the database is regularly updated while preserving versioned, point-in-time snapshots of the data.

3 Database Content

3.1 Interest Rate Announcements

Over the last century, central bank communication has evolved from indirect and cryptic messaging to becoming an independent and significant monetary policy tool in both developed and developing economies (Blinder et al., 2024). This communication involves disseminating information about monetary policy objectives, economic outlook, and future policy trajectories (Blinder et al., 2008). Assenmacher et al. (2021) defines monetary policy communication as a central

bank's transmission of its objectives, plans, instruments, policy decisions, and economic assessments.

Monetary policy statements—official communications explaining policy decisions and their underlying rationales (Dincer et al., 2022)—represent a crucial component of this communication framework. Their primary objectives include managing inflation expectations, enhancing policy predictability, ensuring transparency, and building institutional credibility with markets, governments, and the public (Kahn, 2007; Blinder et al., 2008; Issing, 2019; Lewis et al., 2020).

Central banks employ various nomenclature for these statements: the Reserve Bank of Australia (RBA) issues a post-meeting "Monetary Policy Decision," the Bank of Japan (BoJ) releases a "Statement on Monetary Policy," the Federal Reserve's primary monetary policymaking body (Federal Open Market Committee) produces a "FOMC Statement," and the European Central Bank (ECB) publishes an "Introductory Statement" (until mid-2021) or "Monetary Policy Statement" (since mid-2021)—documents that provide the detailed rationale accompanying each policy decision. This practice is relatively recent, with the RBA releasing one of the first official policy statements on January 23, 1990, following New Zealand's pioneering efforts in transparent communication during the mid-1990s (Blinder et al., 2001).

While the Australian media reported on the RBA's January 1990 monetary policy change, a notable article in a business-focused newspaper, potentially the first centered on this type of monetary policy statement, did not explicitly describe the statement itself as a new communication format.[1] Searches in other publications yielded similar findings.

The statement's novelty as a communication tool may have been overlooked because senior RBA officials were already actively communicating with markets and the public through other channels, such as speeches, prior to the January 23, 1990, statement, as indicated by news searches from that period.[2] This practice of announcing policy changes began formally in January 1990, marking a shift from previous operational methods where market participants inferred policy changes from the RBA's market actions.

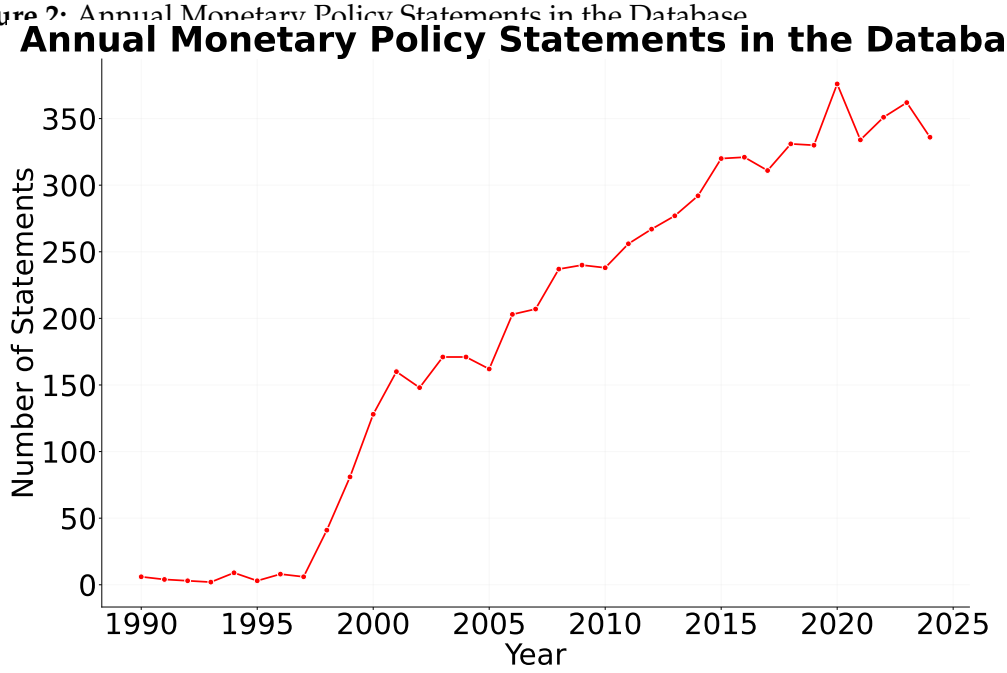
The U.S. Federal Reserve issued its first press release on February 4, 1994, initially publishing statements only when modifying the federal funds rate. By May 1999, the FOMC began releasing statements after every meeting, and by May 2002, included voting details and dissenting vote explanations (Acosta and Meade, 2015). Other central banks followed similar patterns: the Reserve Bank of New Zealand in 1996, the BoJ in 1998, the Swiss National Bank in 2000, and the ECB in 2001. As Figure 2 illustrates, statement issuance accelerated significantly following the 2008 financial crisis.

While formatting varies across institutions, monetary policy statements

[1] Tim Dodd. (January 24, 1990 Wednesday). Timing Surprises Markets. Australian Financial Review.

[2] Tony Kaye. (April 12, 1989 Wednesday). Timing 'Crucial' in Monetary Policy.

Figure 2: Annual Monetary Policy Statements in the Database



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

typically include inflation and growth forecasts, economic assessments, and justifications for policy adjustments (Blinder, 2009). Federal Reserve statements generally assess domestic economic conditions, reference the dual mandate, announce policy decisions, provide forward guidance, and report voting outcomes (Acosta and Meade, 2015). The ECB structures its statements around economic and monetary analyses, policy decisions, rationales, and forward guidance (Assenmacher et al., 2021). The Bank of Korea employs a six-paragraph framework covering policy decisions, global and domestic economic conditions, inflation trends, financial market developments, and future policy direction. These statements typically become more detailed during economic crises, such as the 2008 financial crisis or the COVID-19 pandemic (Macklem and Vardy, 2023).

Statement style has evolved toward greater accessibility and clarity (Blinder et al., 2024), increasingly incorporating visual elements to communicate complex concepts (Assenmacher et al., 2021). Most central banks have converged on approximately eight annual policy meetings, balancing the need for policy flexibility with regular updates. Notable exceptions include the Swiss National Bank (four annually) and the RBA (eleven annually, reduced to eight in 2024). As meeting frequency has declined at some institutions, statement length and detail have increased, particularly for the FOMC (Davis and Wynne, 2019), reflecting the rise of non-standard monetary policy tools such as forward guidance and asset purchases.

The monetary theory consensus now emphasizes open communication as essential for effective policy transmission (Eijffinger and Masciandaro, 2014). Research demonstrates communication's importance in improving monetary

transmission and maintaining trust (Ehrmann and Wabitsch, 2022). Global monetary policy has become increasingly transparent, with central banks—including those in emerging economies—publishing policy statements to enhance market stability and policy effectiveness (Dincer et al., 2022). Despite this expanded outreach, audience engagement remains uneven: firms and consumers demonstrate limited awareness, while financial market participants respond significantly to communications, including policy statements (Blinder et al., 2024). Semantic analyses of ECB, Federal Reserve, and other central bank statements reveal a trend toward increased precision and transparency following the Global Financial Crisis (Benchimol et al., 2020).

3.2 Database Overview

This section examines the evolution and characteristics of monetary policy statements (MPS). We define a *monetary policy statement observation* as a written statement released by a central bank on date t, associated with a policy-rate decision event (change or no change). To account for variations in naming conventions across central banks, we map each bank's official document series—released on or explicitly linked to a decision date—to the corresponding decision event t and document this mapping. We begin with an overview of our database, which encompasses approximately 6,693 unique monetary policy statements from 51 central banks globally, spanning the period from 1990 to 2024. As depicted in Figure 8, this standardized and diverse corpus provides an extensive dataset ideal for training and benchmarking AI systems.

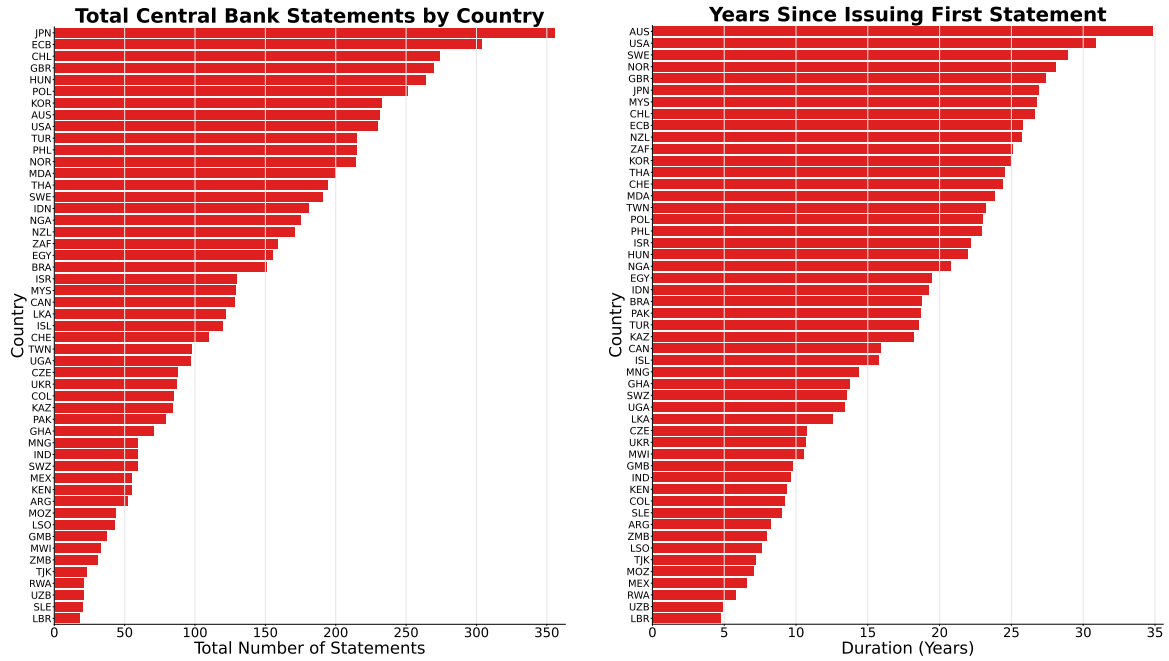
A descriptive analysis of our dataset, as summarized in Tables 3 and 4, reveals substantial heterogeneity in central bank communication patterns across institutions.[3] The Bank of Japan (BoJ) and European Central Bank (ECB) emerge as prolific publishers of statements, while central banks in middle-income countries exhibit more limited publication frequency and shorter historical coverage. These cross-institutional variations reflect differences in institutional frameworks and communication strategies, as illustrated in Figure 3. The statistics on coverage periods, statement frequency, linguistic characteristics, and readability metrics demonstrate systematic differences in communication style and complexity, highlighting the evolution of central bank transparency practices.

Figure 3 provides comparative evidence of central bank communication practices across countries, with Panel (a) documenting the total number of monetary policy statements per country and Panel (b) illustrating the duration of coverage in years. The data reveal considerable heterogeneity in both dimensions. For instance, while the BoJ leads in total statement volume, it does not exhibit the

[3]Appendix A.4 documents that unusually long statements in the corpus are concentrated in countries where the policy decision is communicated through a broader analytical document rather than a concise press release, reflecting institutional differences in communication formats and document bundling.

longest historical coverage. This institutional variation reflects differences in transparency regimes, governance structures, and the evolution of communication strategies among monetary authorities.

Figure 3: Statement Volume and Coverage by Country

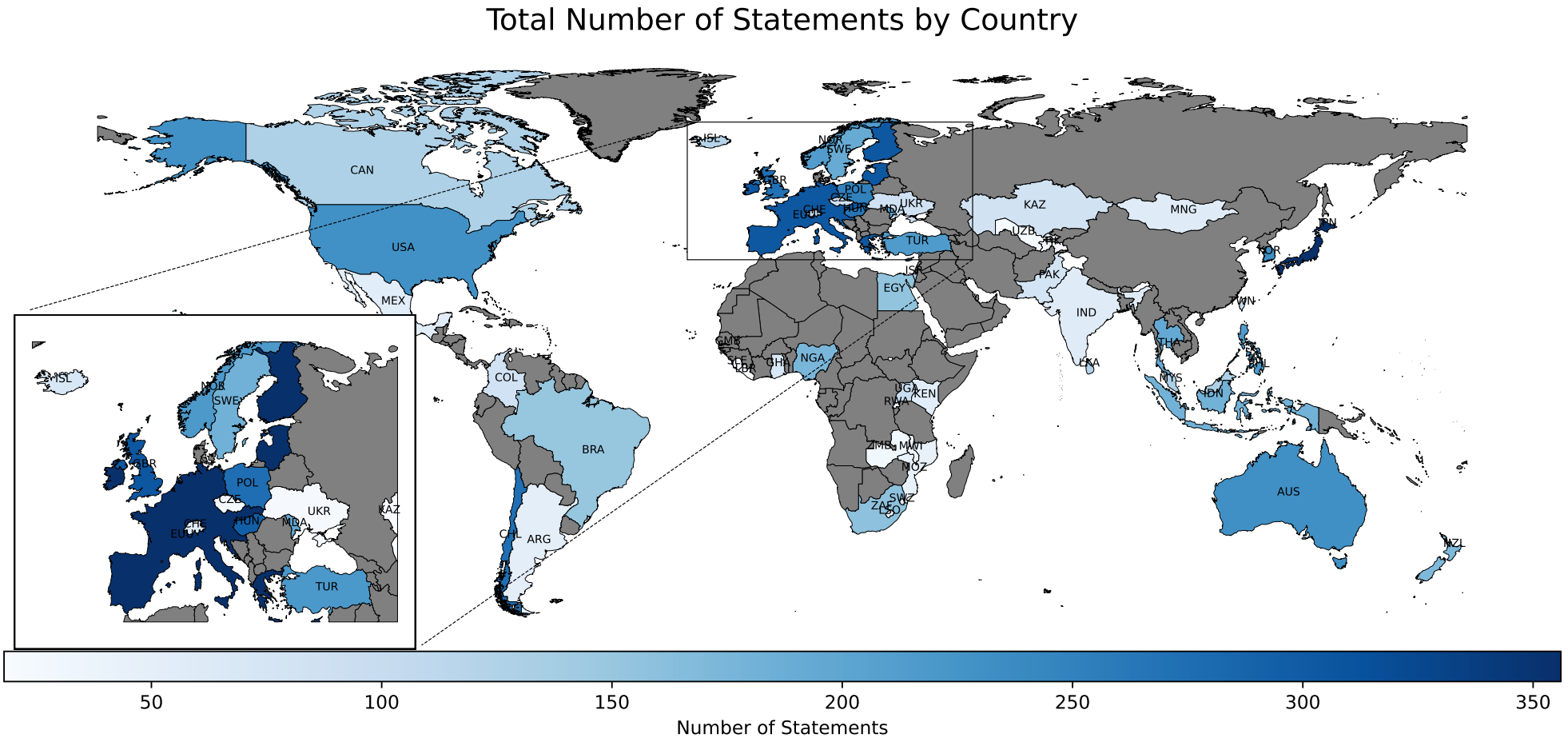


(a) Statements per Country

(b) Coverage Period (Years)

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

The map in Figure 4 illustrates the global reach of the MPSD, encompassing central banks from advanced economies, emerging markets, and developing countries across different continents. This geographical diversity is crucial for studying how monetary policy communication varies across different economic contexts and institutional settings, enabling more robust cross-country analyses than previously possible with existing datasets.



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Table 3: Central Bank Monetary Policy Statements: Coverage and Communication Frequency

Country	Region	Income Group	Number of Statements	Avg. Statements per Year	Modal Statements per Year	Statements in 2023	Database
Japan (JPN)	East Asia & Pacific	High income	356	13	8	8	Programmatic
Euro area (EUU)	Europe & Central Asia	High income	304	11	12	8	Programmatic
Chile (CHL)	Latin America & Caribbean	High income	274	10	12	8	Programmatic
United Kingdom (GBR)	Europe & Central Asia	High income	271	9	12	8	Programmatic
Hungary (HUN)	Europe & Central Asia	High income	264	11	12	12	Programmatic
Poland (POL)	Europe & Central Asia	High income	251	10	11	11	Programmatic
South Korea (KOR)	East Asia & Pacific	High income	233	9	12	8	Programmatic
Australia (AUS)	East Asia & Pacific	High income	231	6	11	11	Programmatic
United States (USA)	North America	High income	230	7	8	8	Programmatic
Turkey (TUR)	Europe & Central Asia	Upper middle income	215	11	12	12	Programmatic
Philippines (PHL)	East Asia & Pacific	Lower middle income	215	9	8	9	Programmatic
Norway (NOR)	Europe & Central Asia	High income	214	7	8	8	Programmatic
Moldova (MDA)	Europe & Central Asia	Upper middle income	200	8	12	9	Programmatic
Thailand (THA)	East Asia & Pacific	Upper middle income	194	7	8	6	Programmatic
Sweden (SWE)	Europe & Central Asia	High income	191	6	6	5	Programmatic
Indonesia (IDN)	East Asia & Pacific	Upper middle income	181	9	12	12	Programmatic
Nigeria (NGA)	Sub-Saharan Africa	Lower middle income	175	8	11	9	Programmatic
New Zealand (NZL)	East Asia & Pacific	High income	171	6	8	7	Programmatic
South Africa (ZAF)	Sub-Saharan Africa	Upper middle income	159	6	6	6	Programmatic
Egypt (EGY)	Middle East & North Africa	Lower middle income	155	7	8	8	Programmatic
Brazil (BRA)	Latin America & Caribbean	Upper middle income	151	7	8	8	Programmatic
Israel (ISR)	Middle East & North Africa	High income	130	5	8	8	Programmatic
Malaysia (MYS)	East Asia & Pacific	Upper middle income	129	4	6	6	Programmatic
Canada (CAN)	North America	High income	128	8	8	8	Programmatic
Sri Lanka (LKA)	South Asia	Lower middle income	122	9	8	8	Manual

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD). Region and Income Group data are from the World Bank country classifications by income level for 2022–2023.

Notes: This table presents statistics on monetary policy statements for the 25 central banks with more than 120 statements in our database. Coverage period spans from January 1990 through December 2024. "Programmatic" indicates that data are collected via a Python script through an automated interface that updates regularly, while "Manual" indicates manually collected data. Modal statements per year represent the most frequently occurring number of statements issued annually. Income group classifications follow World Bank definitions. Table 3 reports this information for the illustrative subsample discussed in the text, while the complete classification for all countries in the database is provided in the accompanying *metadata.csv* file in the public repository.

Table 4: Central Bank Monetary Policy Statements: Complete Database

Country	Code	Number of Statements	First Statement	Last Statement	Coverage (Years)	Database Type	Document Type	English Language	Central Bank	Region
Argentina	ARG	52	2016-08-09	2024-11-01	9	Programmatic	MPC	Translated	Central Bank of Argentina	Latin America & Caribbean
Australia	AUS	231	1990-01-23	2024-12-10	35	Programmatic	MPD / IRD	Official	Reserve Bank of Australia	East Asia & Pacific
Brazil	BRA	151	2006-03-08	2024-12-11	19	Programmatic	CS	Translated	Central Bank of Brazil	Latin America & Caribbean
Canada	CAN	128	2009-01-20	2024-12-11	16	Programmatic	MPD	Official	Bank of Canada	North America
Chile	CHL	274	1998-01-08	2024-09-03	27	Programmatic	PR of RPM	Translated	Central Bank of Chile	Latin America & Caribbean
Colombia	COL	85	2015-08-21	2024-10-31	10	Programmatic	PR	Translated	Central Bank of Colombia	Latin America & Caribbean
Czechia	CZE	88	2014-02-06	2024-11-07	11	Programmatic	BD	Translated	Czech National Bank	Europe & Central Asia
Egypt	EGY	155	2005-06-02	2024-11-21	20	Programmatic	MPCD	Translated	Central Bank of Egypt	Middle East & North Africa
Eswatini	SWZ	59	2011-05-13	2024-11-22	14	Manual	MPS	Official	Central Bank of Eswatini	Sub-Saharan Africa
Euro area	EUU	304	1999-03-04	2024-12-12	26	Programmatic	IS / MPS	Official	European Central Bank	Europe & Central Asia
Gambia	GMB	37	2015-02-24	2024-11-26	10	Manual	MPC PR	Official	Central Bank of The Gambia	Sub-Saharan Africa
Ghana	GHA	71	2011-02-18	2024-11-29	14	Manual	MPC PR	Official	Bank of Ghana	Sub-Saharan Africa
Hungary	HUN	264	2002-12-16	2024-12-17	23	Programmatic	PR	Translated	Magyar Nemzeti Bank	Europe & Central Asia
Iceland	ISL	120	2009-01-29	2024-11-20	16	Programmatic	Statement of MPC	Official	Central Bank of Iceland	Europe & Central Asia
India	IND	59	2015-04-07	2024-12-06	10	Programmatic	MPS	Official	Reserve Bank of India	South Asia
Indonesia	IDN	181	2005-08-16	2024-11-20	20	Programmatic	PR	Translated	Bank Indonesia	East Asia & Pacific
Israel	ISR	130	2002-09-23	2024-11-25	23	Programmatic	IRA	Translated	Bank of Israel	Middle East & North Africa
Japan	JPN	356	1998-01-16	2024-12-19	27	Programmatic	Statement on MP	Translated	Bank of Japan	East Asia & Pacific
Kazakhstan	KAZ	84	2006-09-05	2024-11-29	19	Programmatic	PR	Translated	National Bank of Kazakhstan	Europe & Central Asia
Kenya	KEN	55	2015-08-05	2024-12-05	10	Manual	PR	Official	Central Bank of Kenya	Sub-Saharan Africa
South Korea	KOR	233	1999-12-02	2024-11-28	26	Programmatic	MPD	Translated	Bank of Korea	East Asia & Pacific
Lesotho	LSO	43	2017-04-04	2024-11-26	8	Manual	Statement of MPC	Official	Central Bank of Lesotho	Sub-Saharan Africa
Liberia	LBR	18	2019-11-20	2024-08-27	6	Manual	Communique	Official	Central Bank of Liberia	Sub-Saharan Africa
Malawi	MWI	33	2014-04-30	2024-11-04	11	Manual	Statement of MPC	Official	Reserve Bank of Malawi	Sub-Saharan Africa
Malaysia	MYS	129	1998-02-06	2024-11-06	26	Programmatic	MPS	Translated	Central Bank of Malaysia	East Asia & Pacific
Mexico	MEX	55	2018-05-17	2024-12-19	7	Programmatic	MPS	Translated	Bank of Mexico	Latin America & Caribbean
Moldova	MDA	200	2001-01-25	2024-12-05	24	Programmatic	MPD	Translated	National Bank of Moldova	Europe & Central Asia
Mongolia	MNG	59	2010-07-23	2024-12-13	15	Manual	MPS	Translated	Bank of Mongolia	East Asia & Pacific
Mozambique	MOZ	44	2017-10-26	2024-11-27	8	Manual	Communique	Translated	Bank of Mozambique	Sub-Saharan Africa
New Zealand	NZL	171	1999-03-17	2024-11-27	26	Programmatic	MPD	Official	Reserve Bank of New Zealand	East Asia & Pacific
Nigeria	NGA	175	2004-02-18	2024-11-26	21	Programmatic	MP Communique	Official	Central Bank of Nigeria	Sub-Saharan Africa
Norway	NOR	214	1996-11-05	2024-12-19	29	Programmatic	Rate decision	Translated	Central Bank of Norway	Europe & Central Asia
Pakistan	PAK	79	2005-11-27	2024-07-29	20	Programmatic	MPS	Official	State Bank of Pakistan	South Asia
Philippines	PHL	215	2002-01-17	2024-12-19	23	Programmatic	MPD	Official	Central Bank of the Philippines	East Asia & Pacific
Poland	POL	251	2001-11-28	2024-12-04	24	Programmatic	MPCo PR	Translated	Narodowy Bank Polski	Europe & Central Asia
Rwanda	RWA	21	2019-02-07	2024-11-20	6	Manual	PR	Official	National Bank of Rwanda	Sub-Saharan Africa
Sierra Leone	SLE	20	2015-12-14	2024-12-23	10	Manual	MPS	Official	Bank of Sierra Leone	Sub-Saharan Africa
South Africa	ZAF	159	1999-10-13	2024-11-21	26	Programmatic	Statement of MPC	Official	South African Reserve Bank	Sub-Saharan Africa
Sri Lanka	LKA	122	2012-05-11	2024-11-27	13	Manual	MP Review	Translated	Central Bank of Sri Lanka	South Asia
Sweden	SWE	191	1996-01-09	2024-12-19	29	Programmatic	MPD	Translated	Sveriges Riksbank	Europe & Central Asia
Switzerland	CHE	110	2000-01-20	2024-06-20	25	Programmatic	MPA	Translated	Swiss National Bank	Europe & Central Asia
Taiwan	TWN	98	2001-06-28	2024-09-19	24	Programmatic	MPD	Translated	Central Bank of the Republic of China (Taiwan)	East Asia & Pacific
Tajikistan	TJK	23	2017-01-31	2024-04-26	8	Manual	PR	Translated	National Bank of the Republic of Tajikistan	Europe & Central Asia
Thailand	THA	194	2000-05-23	2024-12-18	25	Programmatic	MPCD	Translated	Bank of Thailand	East Asia & Pacific
Turkey	TUR	215	2006-01-23	2024-08-20	19	Programmatic	PR on Interest Rates	Translated	Central Bank of the Republic of Turkey	Europe & Central Asia
Uganda	UGA	97	2011-07-01	2024-12-05	14	Manual	MPS	Official	Bank of Uganda	Sub-Saharan Africa
Ukraine	UKR	87	2014-04-15	2024-12-13	11	Programmatic	MPD	Translated	National Bank of Ukraine	Europe & Central Asia
United Kingdom	GBR	271	1997-06-06	2024-11-07	28	Programmatic	MPSm	Official	Bank of England	Europe & Central Asia
United States	USA	230	1994-02-04	2024-12-18	31	Programmatic	FOMC statement	Official	Board of Governors of the Federal Reserve System	North America
Uzbekistan	UZB	21	2020-01-18	2024-12-12	5	Manual	PR	Translated	Central Bank of the Republic of Uzbekistan	Europe & Central Asia
Zambia	ZMB	31	2016-11-16	2024-11-13	9	Manual	MPC Statement	Official	Bank of Zambia	Sub-Saharan Africa

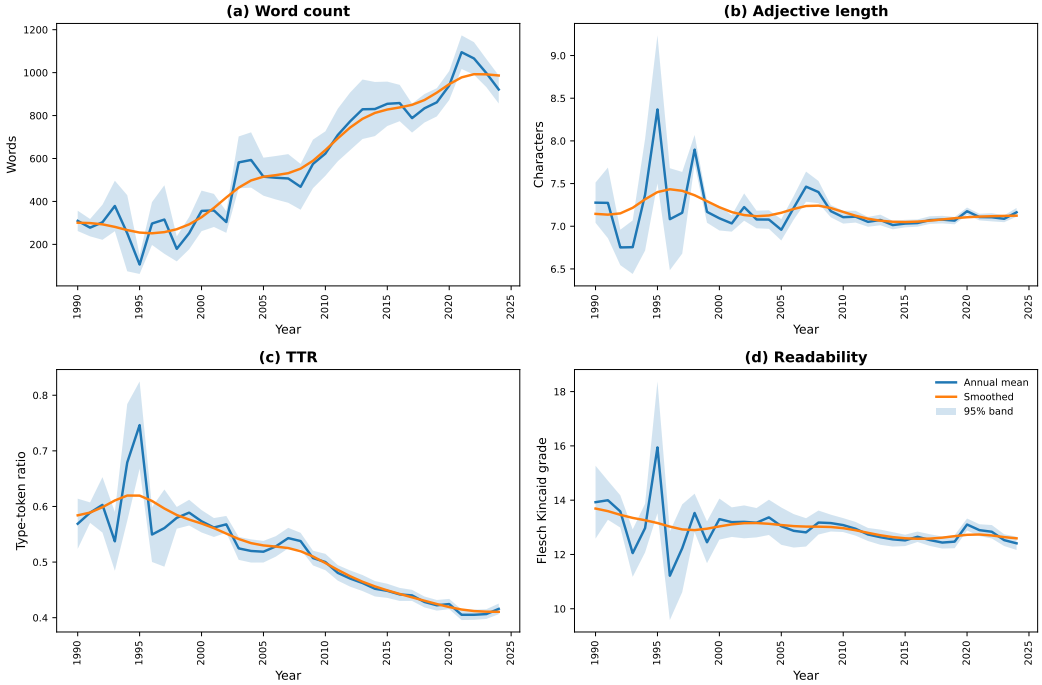
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD). Region data are from the World Bank country classifications.

Notes: This table presents the complete coverage of the MPSD across all 51 central banks. Document Type abbreviations: MPC = Monetary Policy Committee; MPD = Monetary Policy Decision; MPS = Monetary Policy Statement; IS = Introductory Statement; PR = Press Release; CS = Copom Statement; BD = Board Decision; IRD = Interest Rate Decision; MPCD = Monetary Policy Committee Decision; IRA = Interest Rate Announcement; MPA = Monetary Policy Assessment; RPM = Rate-Setting Policy Meeting; MPCo = Monetary Policy Council; MP = Monetary Policy. Coverage years are calculated as the time between the first and last statements in our database.

3.3 Linguistic Patterns

We characterize the evolution of monetary policy statements using four complementary linguistic measures that capture distinct dimensions of central bank communication: scale, morphological complexity, lexical diversity, and readability (Figure 5).

Figure 5: Evolution of Statement Length and Linguistic Characteristics



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: Four document-level linguistic indicators computed for each statement: (i) word count (total tokens); (ii) average adjective length (mean character count of adjectival tokens via part-of-speech tagging); (iii) type-token ratio (unique tokens divided by total tokens); and (iv) Flesch-Kincaid Grade Level. Statement-level measures are aggregated annually using cross-statement means. Shaded bands represent 95% confidence intervals based on the standard error of the annual mean. Smoothed trends obtained via Gaussian filtering.

Panel (a) documents a pronounced upward trend in average word count, with statements approximately tripling in length between 1990 and 2024. The expansion accelerates markedly following the Global Financial Crisis, consistent with the adoption of unconventional monetary policies and forward guidance, which require more extensive explanation. This finding aligns with Masciandaro et al. (2024)'s survey documenting the increasing complexity of central bank communication frameworks. Panel (b) reports average adjective length—a measure of morphological complexity in evaluative language. This indicator remains stable over the sample period, fluctuating modestly around 7.5 characters. The stability suggests that while statements have expanded substantially, central banks have not systematically adopted more complex descriptive vocabulary;

lengthening occurs primarily through additional content rather than denser prose. Panel (c) displays the type-token ratio (TTR), a standard measure of lexical diversity. The secular decline in TTR—from approximately 0.60 in the early 1990s to 0.40 by 2024—indicates increasing lexical repetition as statements lengthen. This pattern is consistent with the institutionalization of standardized policy language and templated communication structures documented in Hansen and McMahon (2016). Panel (d) presents Flesch-Kincaid Grade Level scores, where lower values indicate greater accessibility. Despite the substantial increase in length, readability improves modestly over time, suggesting that central banks have partially offset greater informational content with simpler sentence structures and more accessible word choices. This evolution reflects deliberate efforts to reach broader audiences beyond financial market specialists. Collectively, these patterns reveal a fundamental transformation in central bank communication: statements have become longer and more standardized while maintaining—and even improving—accessibility. This shift toward detailed yet readable policy communication represents a purposeful adaptation to the expanded role of transparency in modern monetary policy frameworks.[4]

A closer examination of monetary policy statements reveals a typical structure including economic forecasts, domestic and foreign economic analysis, policy announcements regarding interest rates and forward guidance, and comments on the policy mix. These statements exhibit recurring sentence patterns, often involving specific parts-of-speech sequences, as detailed in Table 5.

Table 5 illustrates the systematic relationship between economic themes and linguistic structures in monetary policy statements. By identifying these recurring syntactic patterns, we develop more effective NLP tools tailored to central bank communications. For instance, the sequence "policy + AUX + ADJ" is frequently associated with the Monetary Stance, as seen in examples like "policy is restrictive" or "stance remains loose," while the structure "NOUN + ADP + inflation" typically characterizes Inflation Dynamics. These patterns, identified using spaCy's Universal POS tags combined with lexical anchors, provide a robust framework for disentangling the stylistic conventions of central bank communication from simple word frequency.

Figure 6 highlights the most frequent part-of-speech patterns observed in monetary policy statements, providing further evidence of the structured nature of central bank communications.

As shown in Figure 6, common sentence patterns include `DET ADJ NOUN` (e.g., "The strong dollar," "The current rate"), `DET NOUN ADP` (e.g., "The impact of," "The level of"), and `ADP DET NOUN` (e.g., "In the market," "On the economy"). The prevalence of these patterns suggests a high degree of formalization in monetary policy discourse, with central banks employing consistent linguistic structures to

[4]Appendix A.3 presents these linguistic indicators separately for statements originally drafted in English and those translated into English. The similarity of trends across both groups suggests that the main results are not driven by translation-related measurement artifacts.

Table 5: Linguistic Patterns in Monetary Policy Statements by Economic Theme

Economic Theme	**Part-of-Speech Pattern**	**Representative Example**
Inflation Forecasts	inflation + AUX + VERB	inflation **will rise** inflation **might persist**
Growth Projections	growth/GDP + AUX + VERB	growth **should recover** GDP **may stall**
Inflation Dynamics	NOUN + ADP + inflation	risk **of** inflation outlook **for** inflation
Rate Levels	rate + ADP + NUM	rate **at** 5.25% rate **above** 2.0
Forward Guidance	policy/rate + AUX + VERB	policy **will remain** rates **might stay**
Monetary Stance	policy/stance + AUX + ADJ	policy **is restrictive** stance **remains loose**
Fiscal Policy	NOUN + ADP + spending/deficit	impact **of** deficit boost **from** spending
External Conditions	NOUN + ADP + trade/export	demand **for** exports uncertainty **about** trade

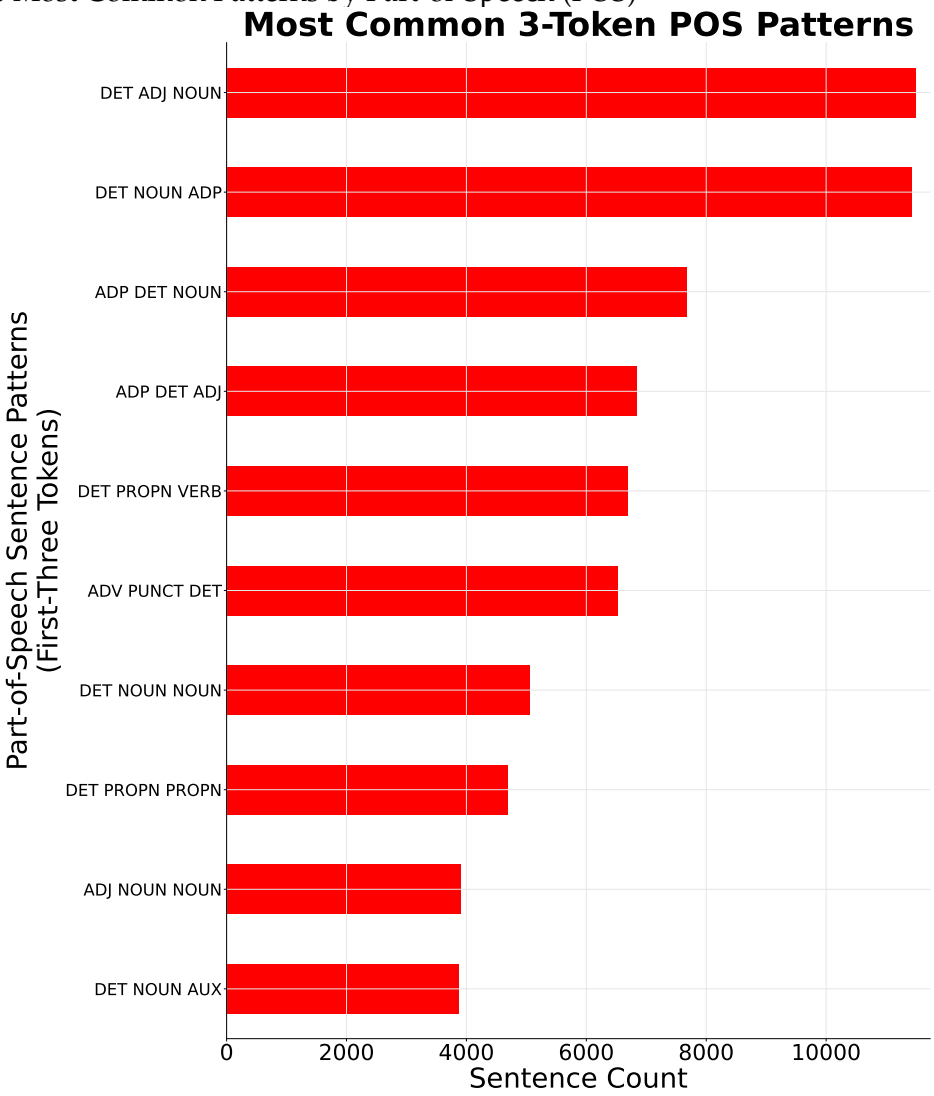
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: This table presents the simplified recurring linguistic patterns identified in our corpus. Pattern identification employs a structural approach combining Universal POS tags (AUX = Auxiliary, ADP = Adposition, ADJ = Adjective, NUM = Number) with single lemma-based lexical anchors. This structure allows for the capture of complex syntactic relationships (e.g., "risk of inflation") rather than simple word adjacency.

convey their assessments and decisions. Additionally, the `DET PROPN VERB` pattern is particularly associated with central bank commentary and actions, as illustrated in Appendix Figure 20.

Figure 7 presents a word cloud visualization of terms appearing in the `NOUN + ADP + inflation` pattern as shown in Table 5. This structure effectively filters for the context of inflation rather than mere mentions of the word. Dominant trigrams such as "risks to inflation," "outlook for inflation," and "decline of inflation" highlight that this syntactic pattern is primarily used to communicate uncertainty and future trajectory. The high frequency of terms like "rise," "pressures," and

Figure 6: Most Common Patterns by Part-of-Speech (POS)



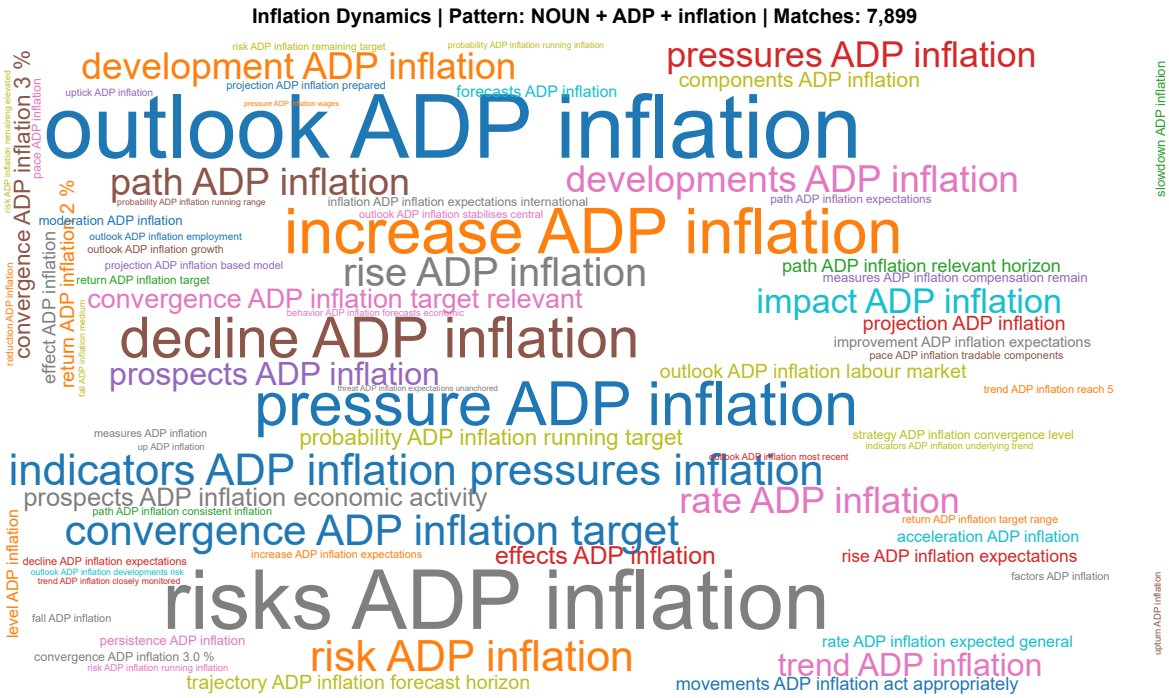
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

"convergence" confirms that this pattern acts as a linguistic marker for the central bank's assessment of price dynamics.

3.4 Corpus Characteristics and Word Frequency

Figure 8 presents a word cloud visualization of the most frequent terms in our corpus of central bank communications. The prominence of terms such as "inflation," "economic," "growth," "policy," and "financial" reflects the core concerns of monetary policymakers. This visualization not only illustrates the thematic focus of central bank communications but also demonstrates the richness and domain-specificity of our corpus, which is essential for training specialized language models that can accurately capture the nuances of monetary policy discourse.

Figure 7: Most Common Trigrams with the Part-of-Speech Pattern `NOUN + ADP + inflation`



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: The inflation dynamics (`NOUN + ADP + inflation`) pattern highlights contextual usage such as "risks to inflation" and "outlook for inflation." Matches: 7,899

3.5 Inflation Mentions Over Time

We analyze the evolution of key economic terms in monetary policy communications. Figure 9 tracks references to inflation across central banks over time, providing a measure of the prominence of price stability in policy discourse. To account for mechanical variation arising from changes in communication frequency and document length, we report two complementary normalized measures. Panel (a) of Figure 9 shows inflation mentions normalized by the number of statements issued per year, capturing the average intensity of inflation-related discussion per policy communication. Panel (b) normalizes inflation mentions by total word count, reporting mentions per 1,000 words and thus measuring the relative salience of inflation within the textual content of policy statements.

These normalizations ensure that observed changes in inflation-related communication are not mechanically driven by an increased number of policy meetings or longer statements, but instead reflect shifts in the emphasis placed on inflation by central banks. The resulting series reveal substantial comovement in inflation discourse across countries, with particularly pronounced increases during episodes of elevated global inflationary pressure. At the same time, persistent cross-country differences remain, reflecting heterogeneity in mandates,

Figure 8: Standardized and Diverse Corpus for Training/Testing A.I.



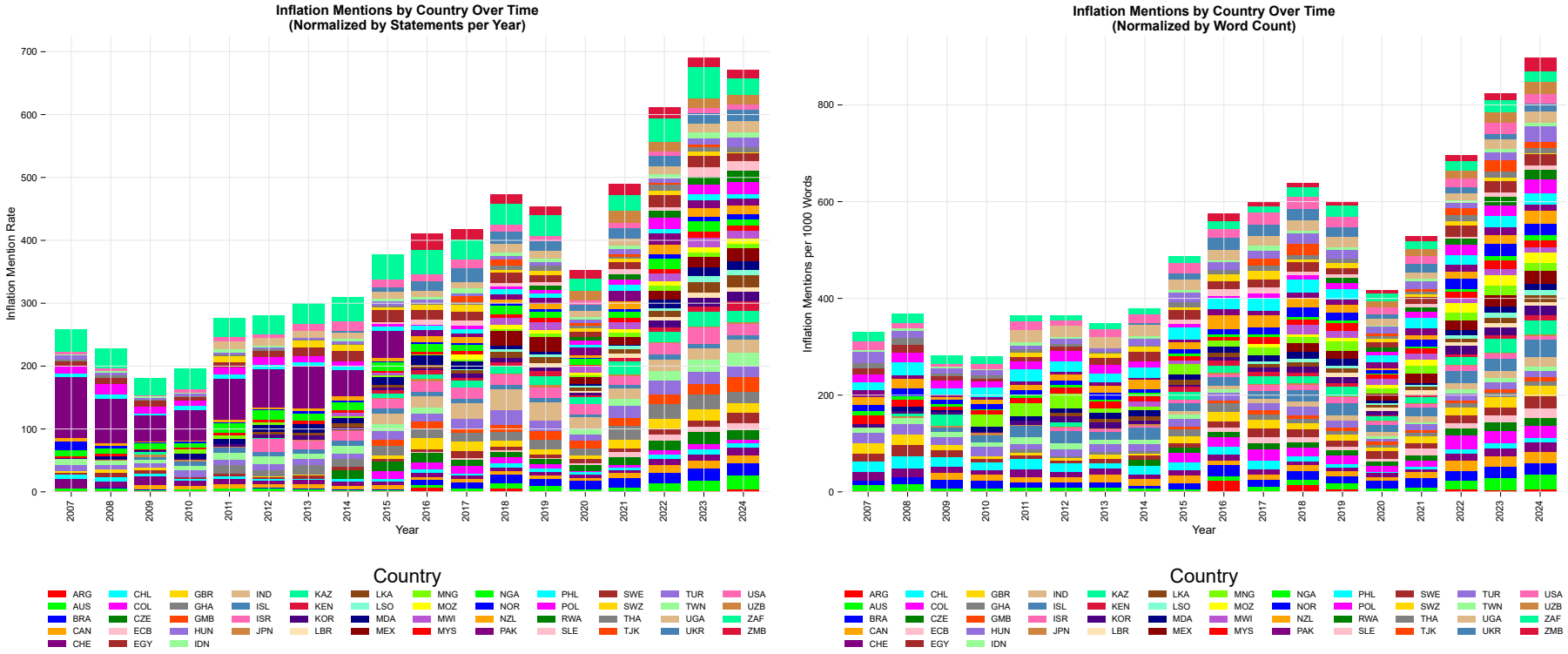
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: 51 central banks and 200,456 sentences.

inflation dynamics, and policy frameworks. Overall, the normalized measures provide a clearer interpretation of how central banks' communication priorities evolve in response to changing macroeconomic conditions, while facilitating meaningful comparison across institutions and over time.

This analysis of monetary policy statements contributes to a deeper understanding of how central banks communicate their policy stances and perspectives on economic developments. By dissecting the structure, content, and evolution of these statements, we gain valuable insights into the strategies and objectives of central banks.

Figure 9: Inflation Mentions by Country



(a) Normalized by statements per year

(b) Normalized by word count

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

4 A Generative AI Application to Central Bank Communications

This section illustrates the capabilities of generative AI, specifically LLMs, to analyze economic data and address pertinent challenges in economic research. Our application focuses on examining the GFC, a phenomenon characterized by the synchronized movement of capital, credit, and asset prices across countries (Rey, 2015; Miranda-Agrippino and Rey, 2020, 2022).

The applications presented here are illustrative and are intended to demonstrate how the database and infrastructure can be used, rather than to establish definitive causal claims.

4.1 Conceptual Framework

Central bank communications influence global financial conditions through multiple channels. First, statements from major central banks—particularly the Federal Reserve and ECB—serve as public signals about the state of the global economy, reducing information asymmetries among market participants worldwide (Morris and Shin, 2002). Second, policy communications coordinate expectations across borders, affecting asset prices and capital flows even in countries not directly targeted by the communication (Woodford, 2005). Third, the content and tone of central bank statements provide forward guidance that shapes expectations about future monetary policy paths globally (Blinder et al., 2008). These mechanisms suggest that systematically analyzing central bank communications can reveal how policymakers perceive and respond to global economic dynamics—precisely the objective our infrastructure enables.

4.2 Addressing Challenges in AI Applications to Economics

While LLMs offer immense analytical potential, they present significant methodological challenges for economic research. We identify and address two primary concerns:

1. **Reproducibility**: The lack of transparency surrounding the inputs and internal mechanisms of many LLMs hinders the replication of research findings, a cornerstone of the scientific method. To address this, our study prioritizes open-source code, providing access to LLM code inputs, prompts used for model querying, a clean code base, and user-friendly Jupyter notebooks. This approach ensures complete transparency and facilitates the reproduction of the research process, promoting rigorous scientific inquiry.

2. **Computational Efficiency**: High computational costs and energy consumption associated with LLMs present practical barriers to their

application in economic research. To mitigate this, we implement text-filtering pipelines that leverage keyword and part-of-speech analysis to refine the input data for the LLMs, effectively reducing the computational burden and enhancing efficiency.

While the internal mechanisms of LLMs are not fully transparent, reproducibility in empirical economics relies on the ability to replicate results using the same data, code, and documented procedures. We therefore ensure transparency at every stage under the researcher's control, including data collection, preprocessing, prompt design, model selection, and analytical workflows. This approach is consistent with current best practices in text-as-data research using modern NLP tools (Gentzkow et al., 2019). Importantly, scientific reproducibility does not require replicating the original data collection process; rather, it requires that published results can be exactly reproduced from the shared corpus and code. All figures, tables, and empirical findings in this paper satisfy this criterion.

Table 6 illustrates the output of our LLM-based analysis of global financial conditions over time. The visualization shows how central banks' discussion of global economic drivers has evolved through different economic cycles. This approach demonstrates how LLMs can systematically extract and analyze economic narratives from large corpora of policy communications, offering insights into how policymakers conceptualize and respond to global economic dynamics. The temporal patterns revealed in this analysis align with major economic events and shifts in the GFC, highlighting the value of applying advanced language models to economic research questions.

Table 6: LLM-Generated Analysis of Global Financial Conditions (2007–2013)

Year and Driver	Origin Country	Summary
2007 Global economic conditions, subprime mortgage crisis, global liquidity, U.S. economic slowdown/housing market, commodity prices.	U.S., China	Several central banks identify global economic conditions as a key driver, with concerns arising from the U.S. economic slowdown/housing market and the subprime mortgage crisis. Concerns regarding global liquidity and commodity prices (oil and food) are also noted. The U.S. Federal Reserve is implicitly referenced through discussions of the U.S. economy and monetary policy, but no central bank explicitly identifies the Federal Reserve as the main driver.
2008 U.S. subprime mortgage crisis, Global financial turmoil, Slowdown in the U.S. economy, Global economy slowdown.	United States	The main driver of global economic and financial conditions is the United States, specifically originating from the U.S. subprime mortgage crisis.
2009 Global Financial Crisis / Global Economic Slowdown / Global Recession.	United States (due to U.S. housing bust)	The main driver of global economic and financial conditions during this period was identified as the Global Financial Crisis, originating from the U.S. housing bust. Many central banks acknowledged the significant impact of the crisis on their domestic economies. No central banks directly identified the Federal Reserve as the driver, but the origin of the crisis is cited as the U.S. housing market.
2010 Global Economic Recovery.	Not explicitly one country (mentions of U.S., China, Asia)	The primary driver of global economic and financial conditions is identified as the global economic recovery, influenced by emerging market economies (especially China) and, to a lesser extent, advanced economies like the U.S. and Japan. Multiple texts mention fiscal stimulus measures as supporting the recovery.
2011 Sovereign debt crisis in Europe and global economic slowdown.	Euro Area	The main driver of global economic and financial conditions is the sovereign debt crisis in Europe, which is causing a global economic slowdown. Several central banks identify the European debt crisis as a significant risk. The Federal Reserve's Quantitative Easing (QE) is mentioned by Bank Indonesia as an event to monitor, but it is not portrayed as the primary driver of the global economy.
2012 Sovereign Debt Crisis in Europe / Weak Global Growth.	Euro Area / Multiple Countries	The main driver of global economic and financial conditions appears to be the sovereign debt crisis in Europe and the overall weak global growth. Many central banks across the globe cite the Eurozone crisis as a major source of risk and uncertainty. The central banks do not directly identify the Federal Reserve as the main driver of global economic and financial conditions, although some mention monetary policies of major central banks.
2013 Uncertainty regarding U.S. monetary policy (specifically tapering), fiscal policy, and debt ceiling issues.	United States (primarily), Euro Area and China	Several central banks identify the U.S. Federal Reserve's potential tapering of quantitative easing as a major driver of global financial market volatility and a downside risk to growth. The U.S. fiscal situation and debt ceiling debates also contribute to uncertainty. Concerns about the Euro Area's economic challenges and China's growth slowdown are also frequently mentioned.

Source: Structured summaries derived from the Monetary Policy Statement Database. We filtered 2,921 unique sentences containing the keywords "global" and "spillover" by year using spaCy (see Section 4.2) and generated summaries via Google's Gemini (gemini-2.0-flash) using the prompt in Appendix A.2. All input texts (sentences) are stored in JSON for replication and will be made available along with database metadata.

4.3 LLM Data Pipeline Infrastructure

Our methodological framework centers on a scalable LLM-based data pipeline architecture that integrates diverse language models—both proprietary (Claude, Gemini, ChatGPT) and open-source (hosted on Hugging Face)—within a unified computational environment. The pipeline's modular design permits seamless substitution of text corpora and models, enabling comparative analysis across central bank communications and addressing potential concerns regarding model-specific biases. This flexibility strengthens the external validity of our findings and ensures the analytical framework can seamlessly integrate emerging language models and additional data sources.

The architecture is designed for scalability along two dimensions:

- **LLM Agnosticism**: The system is compatible with any model (e.g., ChatGPT, Gemini, Claude) and task (e.g., sentiment analysis, QA)
- **Text/Prompt Flexibility**: The pipeline can be dynamically repurposed to evaluate alternative texts and prompts

4.4 Methodological Approach

To illustrate the application of LLMs to central bank communications, we explore the GFC through two key analytical approaches:

4.4.1 Question Answering

In the QA task, we prompt the LLM to provide summaries of the drivers of the global economy, with particular attention to the role of the Federal Reserve. The core directive of the prompt (the full prompt is available in Appendix A.2) is summarized as follows:

> "I want you to identify the main driver of global economic and financial conditions and the origin country driving these conditions. In the summary, please note if any central banks identify the Federal Reserve as the driver of global economic or financial conditions."

Table 7 presents a comparative analysis of how different LLMs interpret the *same* central bank text. This side-by-side comparison reveals significant variations in the summaries generated by different models, even when processing identical inputs and the same prompt. Notably, Claude identifies the U.S. Federal Reserve as the specific driver, whereas all other models—including Gemini (see Table 6)—attribute the influence to the U.S. economy more generally. These differences underscore the importance of transparency in LLM-based research and the need for careful model selection and evaluation when applying these tools to economic analysis. The figure illustrates why our pipeline's ability to integrate multiple

models is valuable—it allows researchers to compare outputs across different LLMs and assess the robustness of their findings.

Table 7: Comparative Analysis of LLM Interpretations of 2008 Global Economic Conditions

Model	Identified Mechanism	Origin Country	Summary Assessment
Claude	The Federal Reserve	United States	Based on central bank statements, the main driver is the Federal Reserve. Several central banks (e.g., Canada, New Zealand, Australia) identified the Fed's actions as a key factor impacting global growth. The global slowdown stemming from the U.S. housing bust was a recurring theme.
Xiaomi	Global financial market turmoil and economic slowdown	United States	The primary driver is global financial market turmoil stemming from the U.S. subprime mortgage crisis. The Federal Reserve is acknowledged by multiple central banks (e.g., PK, ZA, NG, GB, ECB, JP) as a key driver given its role in global stability.
Mistral	Slowdown in the U.S. economy / subprime mortgage crisis	United States	Multiple central banks cite the US economic slowdown—particularly the subprime crisis—as the primary driver. Notably, no central bank explicitly names the Federal Reserve as the driver in this interpretation.
Arcee	U.S. subprime mortgage crisis and broader U.S. economic weakness	United States	Multiple central banks explicitly cite the U.S. subprime crisis and economic slowdown as primary drivers. Several banks note that the Federal Reserve's policy actions are relevant to addressing these conditions.

Source: Structured summaries derived from the Monetary Policy Statement Database. We filtered 2,921 unique sentences containing the keywords "global" and "spillover" by year using spaCy (see Section 4.2) and generated summaries via Anthropic's Claude, mistralai/devstral-2512, arcee-ai/trinity-mini, xiaomi/mimo-v2-flash. All input texts (sentences) are stored in JSON for replication and will be made available along with database metadata.

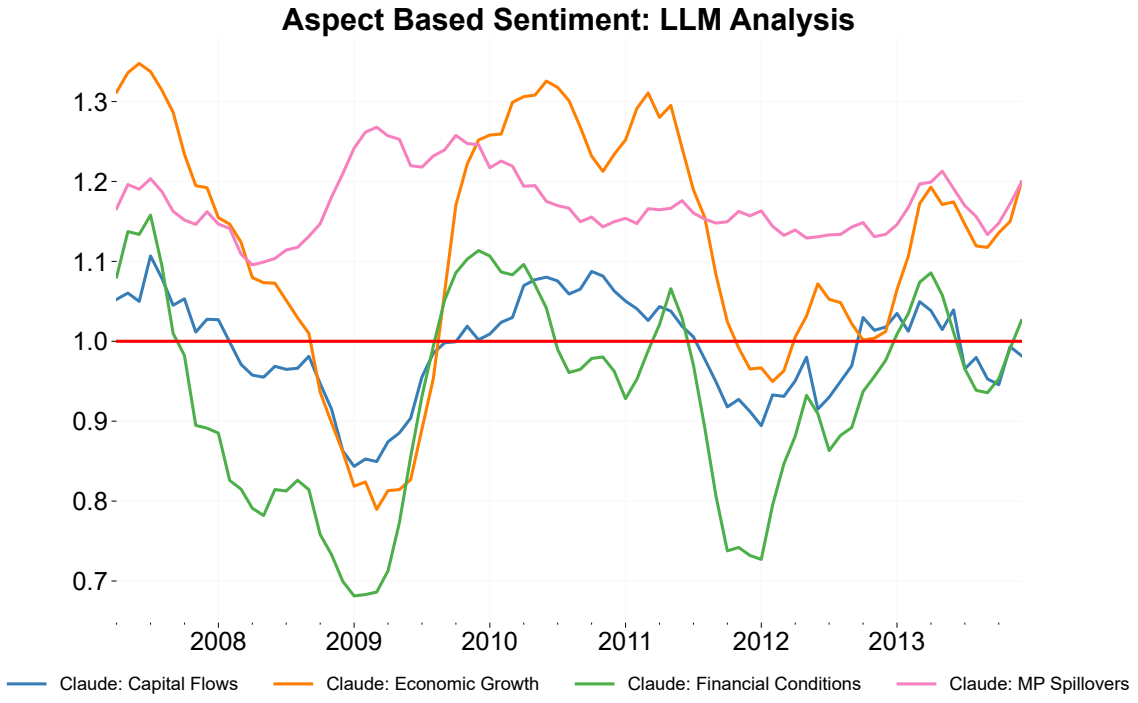
4.4.2 Aspect-Based Sentiment Analysis

For the ABS task, we instruct the LLM to extract sentiment toward specific themes from central bank communications. The core directive of the prompt (the full prompt is available in Appendix A.1) is summarized as follows:

> "Analyze the provided central bank text to identify discussions regarding the impact of global financial conditions on domestic economies, focusing on themes such as economic growth, financial conditions, capital flows, and monetary policy spillovers. For each explicitly discussed theme, assign a continuous sentiment score on a

scale from 0.0 (negative) to 2.0 (positive), where 1.0 represents a neutral view."

This allows us to track sentiment toward specific economic factors across time and central banks, providing insights into how financial conditions and policy spillovers are perceived globally.



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLM tools.

Notes: Sentiment scores are generated using a version of Anthropic's Claude LLM (claude-3-haiku-20240307). The series capture sentiment related to capital flows, economic growth, financial conditions, and monetary policy spillovers (MP Spillovers), averaged across all countries.

Figure 10 illustrates the evolution of sentiment across different economic aspects over time, as extracted by our LLM-based analysis. The figure demonstrates distinct patterns in how central banks express sentiment toward different themes, with notable divergences during periods of economic stress. For instance, the sentiment toward economic growth and financial conditions shows greater volatility compared to monetary policy spillovers and, to a lesser extent, capital flows, particularly during and after the Global Financial Crisis. These temporal patterns offer valuable insights into the shifting priorities and concerns of monetary policymakers across different economic cycles.

4.5 Validation of LLM-Based Sentiment

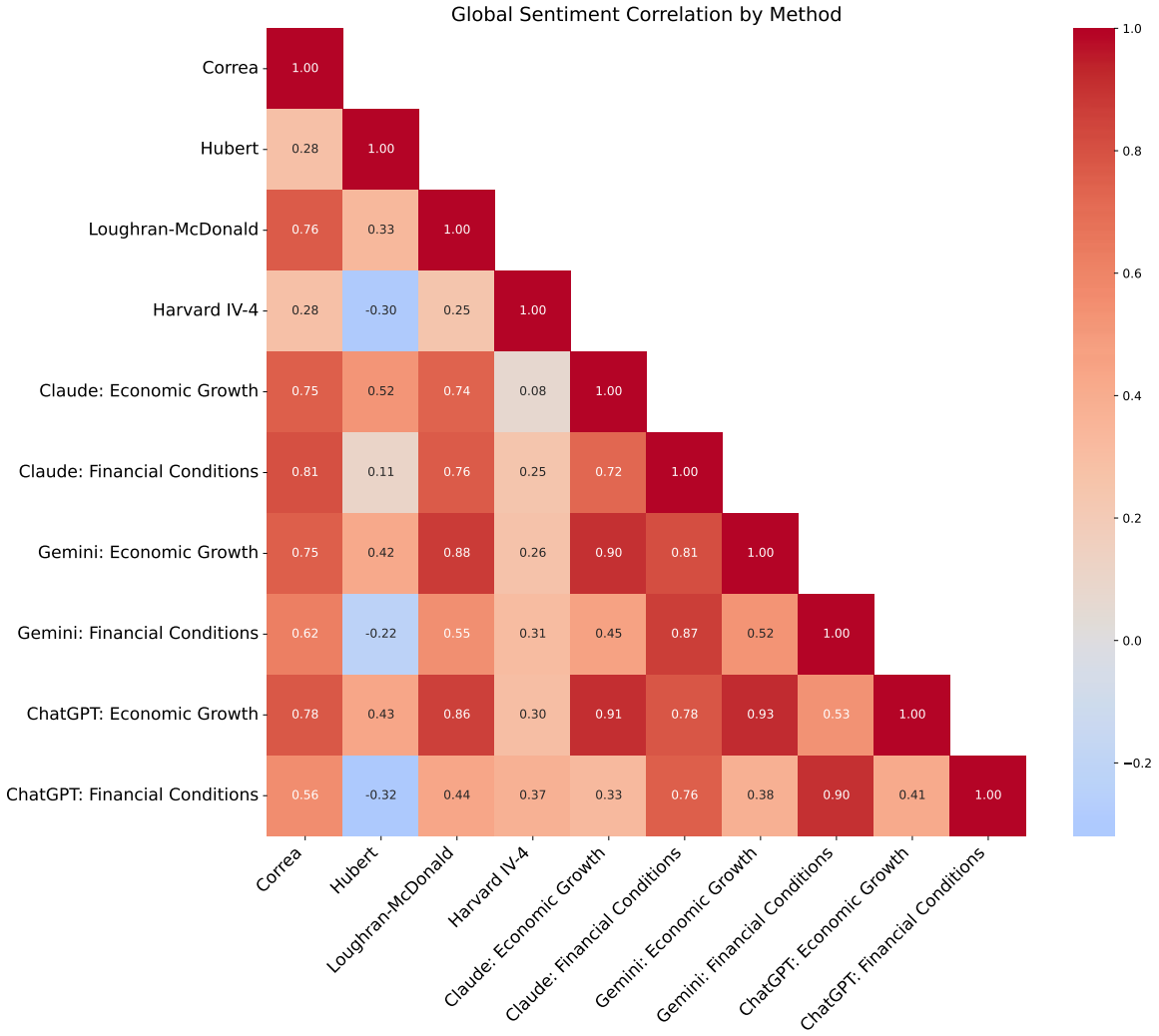
To assess the relationship between traditional lexical sentiment measures and LLM-derived sentiment indices, Figure 11 reports the correlation matrix across all sentiment metrics used in the analysis (see the extended matrix in Appendix Figure 18). Several clear patterns emerge. First, established finance-oriented lexical measures exhibit moderate to strong similarity among themselves. The Correa and Loughran-McDonald indices are strongly correlated (0.76), indicating that both capture a similar notion of financial tone in monetary policy statements. The Hubert index shows weaker alignment with these measures (correlations of 0.28 with Correa and 0.33 with Loughran-McDonald), while the general-purpose Harvard IV-4 dictionary displays substantially lower and less stable correlations, including a negative association with the Hubert index (-0.30). This pattern underscores the importance of domain-specific sentiment tools when analyzing central bank communication.

Second, LLM-derived sentiment measures show consistently strong correlations with finance-specific lexical indices, particularly for the Economic Growth dimension. Across models, Economic Growth sentiment correlates between 0.74 and 0.88 with the Loughran-McDonald index and between 0.75 and 0.78 with the Correa index. In contrast, correlations with the Harvard IV-4 dictionary remain markedly lower, typically below 0.26 (with even 0.08). This divergence suggests that LLMs align more closely with financial-domain sentiment constructs rather than with generic polarity-based lexicons.

Third, correlations differ systematically across sentiment dimensions. Financial Conditions sentiment exhibits weaker and more heterogeneous correlations with lexical measures than Economic Growth sentiment, with values ranging from 0.44 to 0.76 against Loughran-McDonald and from 0.56 to 0.81 against Correa. At the same time, Financial Conditions sentiment shows very strong agreement across LLMs themselves, with correlations exceeding 0.75 in most cases. This pattern indicates that while LLMs consistently identify a shared financial-conditions signal, that signal is only partially captured by traditional dictionary-based approaches.

Finally, the high cross-model correlations among LLMs—reaching 0.90–0.93 for Economic Growth sentiment—point to a robust consensus across different architectures and training paradigms. Rather than producing idiosyncratic sentiment measures, LLMs converge on similar interpretations of economic tone, reinforcing their reliability as measurement tools. Taken together, these results suggest that LLM-based sentiment measures are neither orthogonal to nor simple replications of traditional lexical indices. Instead, they complement established finance-specific dictionaries by capturing structured, dimension-specific sentiment—particularly for growth and financial conditions—while diverging from broad, non-domain-specific sentiment constructs. This finding aligns with the view that monetary policy communication conveys multiple informational

Figure 11: Correlation Matrix of Lexical and LLM-Based Sentiment Measures



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLM tools and Loughran-McDonald, Correa, Hubert, and Harvard IV-4 dictionaries.

Notes: Sentiment correlation by method. Sentiment scores are generated using Loughran-McDonald, Correa, Hubert, and Harvard IV-4 dictionaries, and a version of Anthropic's Claude LLM (claude-3-haiku-20240307), OpenAI's ChatGPT (gpt-4o-mini) and Google's Gemini (gemini-2.0-flash). Correlations are computed based on the cross-sectional mean of the sentiment scores, aggregated across all sample countries.

dimensions and that modern LLMs are well-suited to disentangling these dimensions in a way that remains consistent with established financial text-analysis benchmarks.

These results contribute to the literature on text-as-data benchmarks. While Gentzkow et al. (2019) caution that different measures often target different estimands, our findings demonstrate reassuring convergence when the domain is strictly defined. The results support the premise of Loughran and McDonald (2011) regarding the necessity of domain-specific sentiment lexicons: the LLMs

naturally emulate the financial-specific weighting of the Loughran–McDonald dictionary rather than the generic Harvard IV-4. Consequently, while Hansen and McMahon (2016) note that policy statements contain multiple informational dimensions, our analysis suggests that modern LLMs are highly effective at identifying the specific dimensions of financial risk and economic growth that traditional financial dictionaries were designed to capture.

We identify three mechanisms through which LLMs outperform dictionary methods in this setting. First, LLMs capture contextual negation and qualification that bag-of-words approaches miss. When a central bank states that "inflation pressures have not materialized as expected," dictionary methods count "inflation" and "pressures" as negative, while LLMs correctly interpret the overall sentiment as positive (declining concern). Second, LLMs handle domain-specific terminology more flexibly. Central bank communications frequently use technical phrases ("quantitative easing," "forward guidance," "macro-prudential") whose sentiment depends heavily on context; LLMs infer meaning from surrounding text while dictionaries either ignore such terms or assign fixed scores. Third, LLMs can extract ABS—distinguishing concern about inflation from concern about growth—whereas most dictionary approaches produce a single aggregate score. This granularity is essential for understanding how central banks balance competing objectives in their communications. Our correlation analysis confirms these advantages: LLM-derived sentiment for specific dimensions (growth, financial conditions) correlates more strongly with the GFC than aggregate dictionary-based measures, suggesting that aspect-specific extraction captures economically meaningful variation that aggregate approaches obscure.

4.5.1 Validation Framework

Given the absence of labeled ground truth for ABS in central bank communication, we validate LLM-derived measures using two complementary strategies. First, we benchmark LLM-based sentiment indices against established dictionary-based sentiment measures commonly used in the economics literature. We interpret agreement as evidence of plausibility and divergence as reflecting differences in the underlying estimands rather than measurement error per se. Consistent with this framework, we find that our LLM-derived sentiment tracks finance-specific lexical measures closely, validating the models' ability to capture relevant economic tone. Figure 11 summarizes cross-method correlations between lexical sentiment measures and LLM-derived aspect sentiments, while Figure 12 compares their time-series behavior around major global episodes. Second, we assess the robustness of LLM-based sentiment by comparing outputs across multiple LLMs. Specifically, in addition to our baseline results based on Claude, we replicate the ABS extraction using alternative LLMs (Gemini and ChatGPT). Figure 17 reports the resulting sentiment series for key economic dimensions, including economic growth and financial conditions. Despite differences in scale

and volatility across models, the series exhibit strong comovement and similar temporal dynamics, particularly around major macroeconomic events. This cross-model consistency provides an additional validation layer, indicating that the extracted sentiment patterns are not driven by idiosyncratic behavior of a single LLM architecture. Finally, to ground these statistical aggregates in the underlying text, we conducted a manual review of a random sample of statements using our validation tool. While this qualitative inspection does not constitute a formal scientific audit, we observed substantial agreement between LLM outputs and human interpretation of the underlying text. An illustrative example of this validation exercise is provided in the Appendix (Figure 21). Taken together, benchmarking against transparent lexical measures, comparing across independent LLMs, and spot-checking against human judgment enhances confidence in the stability and interpretability of the proposed sentiment indicators, while preserving their ability to capture contextual and aspect-specific information that is difficult to obtain using dictionary-based approaches alone.

We acknowledge three limitations inherent to LLM-based sentiment extraction. First, prompt sensitivity may affect results; we mitigate this concern by documenting all prompts in the repository and comparing outputs across different model architectures. Second, model updates may affect reproducibility over time; we therefore report exact model versions (e.g., claude-3-haiku-20240307, gemini-2.0-flash, gpt-5-nano) and archive all outputs. Third, LLMs may exhibit biases inherited from training data; our cross-model comparison suggests that such biases do not dominate, as models trained on different corpora produce highly correlated sentiment series. These limitations are common to all LLM-based text analysis and do not invalidate our findings; rather, they underscore the importance of the validation framework we employ.

4.6 Empirical Implementation

Our implementation employs a multi-stage filtering process to extract relevant narrative elements from the corpus:

1. Search using keywords (global, spillover, flows): yielding 6,135 sentences
2. Truncate for specific time periods (e.g., 2019 data): reducing to 1,228 sentences

This approach balances computational efficiency with analytical rigor, addressing the computational challenges of LLM analysis of large text corpora while ensuring systematic sampling for robust results.

4.7 Results

Our analysis using LLMs produces several important insights into the GFC:

4.7.1 Origins of Global Financial Conditions

Our QA analysis, shown in Table 6, indicates that central banks often cite the U.S. economy as a driver of global economic conditions, and in some instances, specifically mention the U.S. Federal Reserve as a factor influencing global economic and financial conditions, particularly during periods of high uncertainty such as 2007 and 2009. However, this is not a consistent trend, and in many years, broader economic factors in the U.S. and other regions, such as the Eurozone sovereign debt crisis and emerging market economies, are perceived as more impactful. This nuanced understanding of global economic influences, where the relative importance of different factors, including central bank actions, can vary over time, contrasts with some academic literature that tends to emphasize the direct and dominant influence of Fed policy on global financial conditions, suggesting a potential gap between academic models and policymakers' conceptualization of global economic linkages (Kalemli-Özcan, 2019).

4.7.2 Sentiment Analysis Results

The ABS analysis shows that central banks expressed greater concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period, likely reflecting the lasting impact of the Global Financial Crisis and increased global interconnectedness.

Our Granger causality analysis follows the formal specification:

$$\text{GFC}_t = \alpha + \sum_{k=1}^{p} \beta_k \text{GFC}_{t-k} + \sum_{k=1}^{p} \gamma_k \text{Sentiment}_{t-k} + \varepsilon_t \tag{1}$$

$$\text{Sentiment}_t = \alpha' + \sum_{k=1}^{p} \beta'_k \text{Sentiment}_{t-k} + \sum_{k=1}^{p} \gamma'_k \text{GFC}_{t-k} + \varepsilon'_t \tag{2}$$

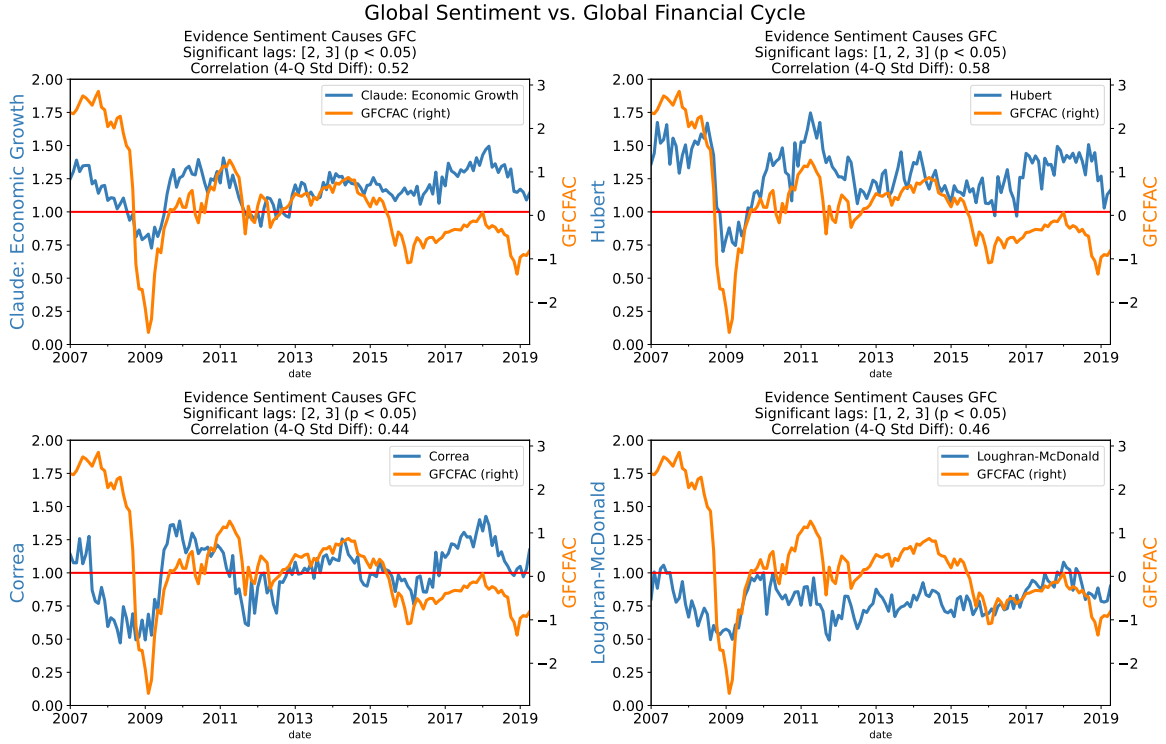
where GFC_t represents the GFC factor at time t, Sentiment_t represents our measure of central bank communication sentiment, and p represents the lag order. The null hypothesis that sentiment does not Granger cause GFC corresponds to $\gamma_k = 0$ for all $k \in \{1, 2, ..., p\}$. Similarly, the null hypothesis that GFC does not Granger cause sentiment corresponds to $\gamma'_k = 0$ for all $k \in \{1, 2, ..., p\}$.

We conduct these tests using lag orders $p \in \{1, 2, 3\}$ and report the resulting p-values in Figure 13. The results consistently reject the null hypothesis that sentiment does not Granger cause GFC, while failing to reject the null hypothesis that GFC does not Granger cause sentiment. This asymmetric causality pattern provides robust evidence that central bank communication sentiment influences global financial conditions rather than merely responding to them.

We emphasize that Granger causality is a statistical concept capturing predictive precedence rather than structural causation in the economic sense (Granger, 1969). The asymmetric pattern we document—where sentiment predicts GFC but not vice versa—is consistent with communication serving as a

transmission channel, though we cannot rule out omitted common factors or reverse causality operating at frequencies not captured by our quarterly specification. Nevertheless, the finding that communication sentiment leads the GFC aligns with theoretical models in which central bank signals coordinate expectations and thereby influence asset prices (Morris and Shin, 2002; Woodford, 2005).

Figure 12: Comparison of LLM-Generated Sentiment Indices with Lexical Approaches



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD), dictionaries and LLM-based (Anthropic's Claude LLM claude-3-haiku-20240307) sentiment measures, and GFC factor (GFCFAC).

Notes: Global Sentiment vs. GFC. GFCFAC denotes the GFC factor. Blue lines show standardized global sentiment indices; orange lines show GFCFAC (right axis). Reported correlations refer to four-quarter standardized differences.

Figure 12 compares global sentiment measures derived from LLM-based and traditional lexical approaches with the GFC factor (GFCFAC). Contrary to the hypothesis that ABS measures might decouple from the aggregate cycle, the figure highlights a strong structural alignment between the methods. The LLM-based economic growth sentiment displays a robust correlation (0.52) with the GFC, effectively outperforming traditional finance-specific dictionaries such as Correa (0.44) and Loughran-McDonald (0.46), and performing comparably to the Hubert index (0.58). All measures exhibit significant predictive power at lags 2 and 3, confirming that the LLM-derived signal captures the same fundamental turning points as established benchmarks.

Taken together with Figures 10 and 11, these patterns indicate that LLM-based sentiment is highly consistent with the signal captured by domain-specific lexical dictionaries. Rather than diverging, the LLM-derived indices appear to refine the signal found in traditional measures. The fact that the LLM-based growth sentiment tracks the GFC more closely than the Loughran–McDonald and Correa indices suggests that the model's contextual awareness allows it to filter out noise better than bag-of-words approaches. Consequently, the observed convergence between these methods serves as a robust cross-validation, demonstrating that LLMs can reliably recover key macroeconomic dynamics while offering the additional advantage of aspect-specific granularity.

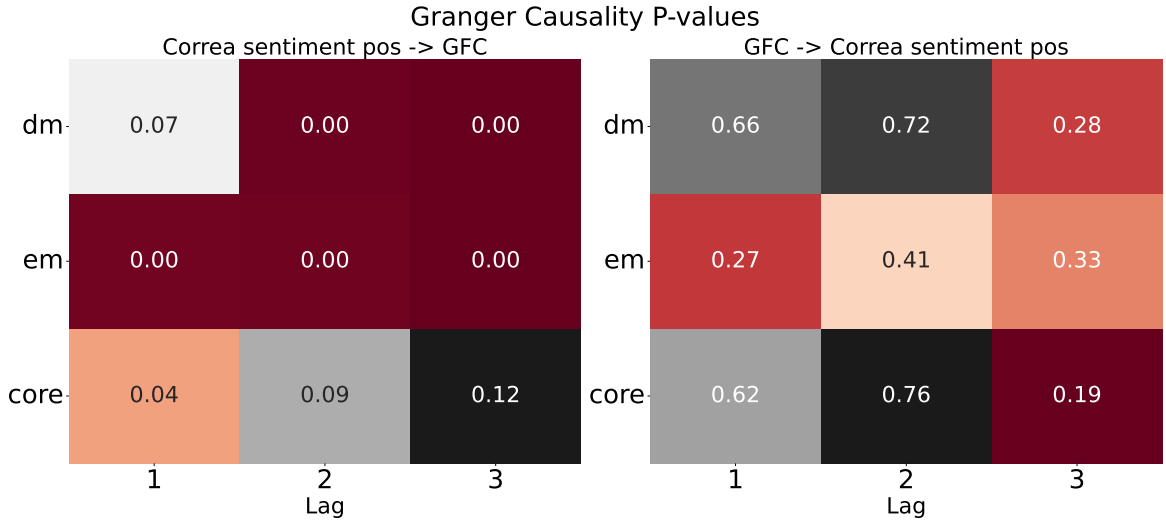
4.7.3 Empirical Evidence for Sentiment Driving the GFC

Beyond descriptive analysis, we find empirical evidence that dictionary-based sentiment expressed in central bank communications may drive components of the GFC. Figure 13 presents results from Granger causality tests using the benchmark Correa sentiment index, showing that aggregated sentiment from both developed and emerging markets Granger-causes movements in the GFC (measured as a 4-quarter difference). Importantly, we do not find evidence of reverse causality, suggesting that central bank communications may play a causal role in shaping global financial conditions. For this analysis, we classify countries following World Bank income categories: high-income economies (per capita GNI above $13,845 in 2022, following the WDI/Atlas-method cutoff) are designated as developed markets, while upper-middle and lower-middle income economies are designated as emerging markets. A complete mapping of each central bank to its classification category is provided in the accompanying metadata file.

Notably, emerging market (EM) sentiment appears to be a more significant driver of the GFC than sentiment from core central banks (ECB, Australia, Canada, Switzerland, UK, Japan, and US). This finding may reflect that EM sentiment captures global investors' risk-on/risk-off preferences, which themselves influence the GFC. Thus, the relationship between EM central bank sentiment and the GFC likely operates through the former's capacity to reflect and amplify global investor risk appetite, rather than through direct causation. These findings extend the literature on international policy transmission by identifying communication as a significant channel through which central banks influence global financial markets.

To assess whether this predictive relationship extends beyond dictionary-based measures, Appendix Figure 19 reports analogous Granger causality tests using LLM-derived, ABS indicators for economic growth and financial conditions. Consistent with the benchmark results based on the Correa index, we find strong evidence that LLM-based sentiment Granger-causes movements in the GFC across multiple model architectures and lag specifications. In particular, sentiment related to economic growth exhibits robust predictive power for the GFC at horizons of one to three quarters, while financial-conditions sentiment also shows

Figure 13: Dictionary-Based Sentiment Granger-Causes the GFC



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

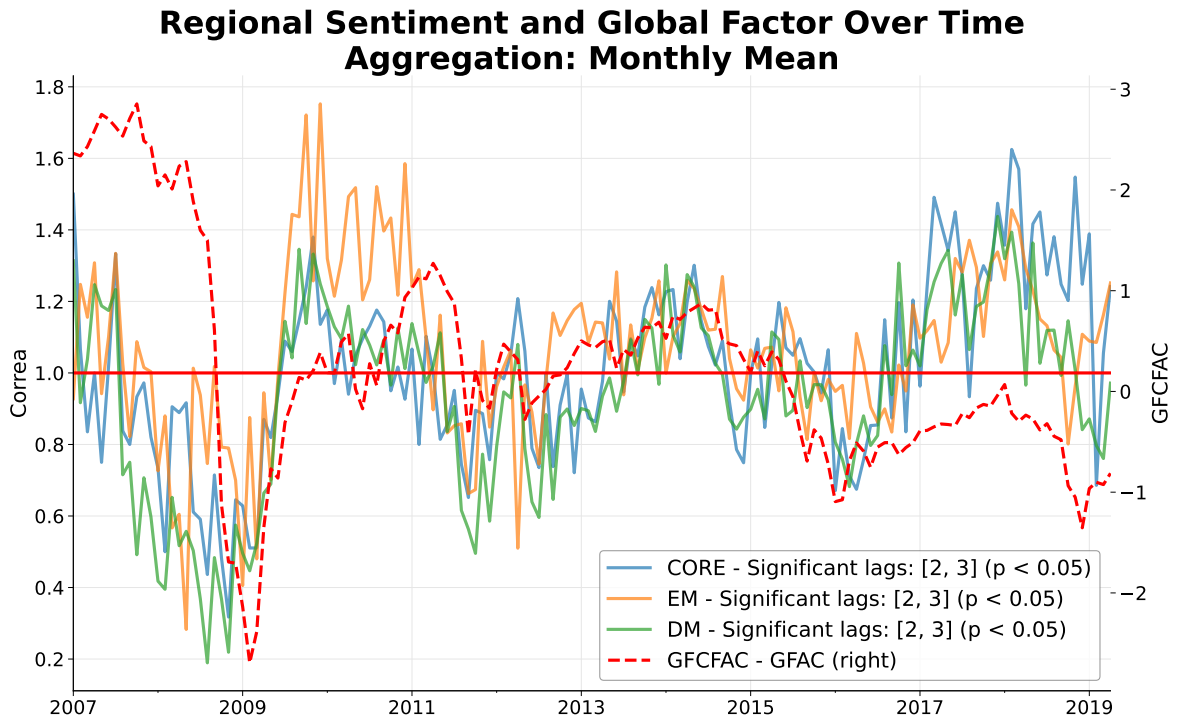
Notes: Sentiment is the benchmark dictionary-based index from Correa et al. (2021) computed from MPSD statements. The GFC is from Miranda-Agrippino and Rey (2020). The figure reports Granger-causality test p-values (rounded to two decimal places) for lag orders $p \in \{1,2,3\}$ (see text). "core" denotes a subset of advanced-economy central banks comprising the ECB and the central banks of Australia, Canada, Switzerland, the United Kingdom, Japan, and the United States. "dm" (developed markets) includes Australia, Canada, Switzerland, the Czech Republic, the United Kingdom, Hungary, Israel, Iceland, Japan, South Korea, Norway, New Zealand, Poland, Sweden, Taiwan, and the United States. "em" (emerging markets) includes Argentina, Brazil, Chile, Colombia, Egypt, Ghana, Gambia, Indonesia, India, Kenya, Kazakhstan, Sri Lanka, Liberia, Lesotho, Moldova, Mongolia, Malawi, Mexico, Malaysia, Mozambique, Nigeria, the Philippines, Pakistan, Rwanda, Sierra Leone, Eswatini, Thailand, Tajikistan, Turkey, Ukraine, Uganda, Uzbekistan, South Africa, and Zambia. Country groupings follow World Bank income classifications: high-income economies are classified as developed markets, while upper-middle and lower-middle income economies are classified as emerging markets.

significant predictive content, albeit with greater heterogeneity across models. Importantly, mirroring the main results, we again find little evidence of reverse causality from the GFC to LLM-based sentiment. This asymmetry suggests that LLM-derived sentiment captures forward-looking information embedded in central bank communications, rather than merely responding mechanically to contemporaneous global financial conditions. Taken together, these appendix results reinforce the interpretation that the predictive content of central bank communication for the GFC is not an artifact of a particular sentiment construction, but persists across both traditional lexical indices and modern LLM-based approaches.

4.7.4 Regional Variation in Sentiment

Our analysis also shows important regional variations in how central banks perceive and respond to global financial conditions. Figure 14 illustrates how

sentiment evolves differently across major economic regions, potentially reflecting heterogeneous exposure to global shocks and different policy responses.



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using Correa dictionary.

Notes: Regional sentiment and Global Factor over time. Aggregation: Monthly Mean"Core" denotes a subset of advanced-economy central banks comprising the ECB and the central banks of Australia, Canada, Switzerland, the United Kingdom, Japan, and the United States. "DM" (developed markets) includes Australia, Canada, Switzerland, the Czech Republic, the United Kingdom, Hungary, Israel, Iceland, Japan, South Korea, Norway, New Zealand, Poland, Sweden, Taiwan, and the United States. "EM" (emerging markets) includes Argentina, Brazil, Chile, Colombia, Egypt, Ghana, Gambia, Indonesia, India, Kenya, Kazakhstan, Sri Lanka, Liberia, Lesotho, Moldova, Mongolia, Malawi, Mexico, Malaysia, Mozambique, Nigeria, the Philippines, Pakistan, Rwanda, Sierra Leone, Eswatini, Thailand, Tajikistan, Turkey, Ukraine, Uganda, Uzbekistan, South Africa, and Zambia. Country groupings follow World Bank income classifications: high-income economies are classified as developed markets, while upper-middle and lower-middle income economies are classified as emerging markets.

Figure 14 displays distinct regional patterns, with periods of both convergence and divergence in sentiment. While advanced economies often show similar sentiment trends, emerging markets display more idiosyncratic patterns, particularly during periods of financial stress. These regional differences suggest that the GFC affects economies differently based on their level of financial integration, institutional characteristics, and economic fundamentals.

These findings contribute to the literature on the GFC by providing systematic evidence on how central banks perceive, discuss, and potentially influence global financial conditions through their communications.

5 Discussion

5.1 Policy Implications

Our findings have several implications for monetary policy implementation and coordination. First, the evidence that central bank sentiment Granger causes movements in the GFC (see Figures 13 and 19) suggests that communication strategies may have cross-border spillover effects beyond what is typically acknowledged. This raises questions regarding central banks' responsibility to consider the international implications of their communications, not just their policy actions, and implies that central banks should carefully calibrate their statements, recognizing that they may influence financial conditions well beyond their jurisdictions.

Second, the finding that central banks more frequently cite broad U.S. economic conditions rather than the Federal Reserve specifically as driving global conditions reveals a potential disconnect in how policy spillovers are conceptualized. While the academic literature often focuses on monetary policy transmission across borders (Miranda-Agrippino and Rey, 2020), central bankers themselves appear to emphasize broader economic linkages, suggesting that international policy coordination may need to encompass fiscal and regulatory dimensions as well. The distinction between attributing global influences to a country's economy versus its central bank has meaningful implications for how we model and address international policy spillovers.

Third, the observed regional variations in how sentiment responds to global conditions (as illustrated in Figure 14) highlight the potential for heterogeneous impacts of global shocks, which complicates the case for uniform policy approaches across different economies. This supports arguments for tailored policy frameworks that account for specific vulnerabilities and transmission channels in different economic contexts (Armelius et al., 2020). The substantial differences in how emerging market and advanced economy central banks discuss global conditions suggest that policy coordination efforts should account for these varying perspectives and constraints.

Our linguistic analysis of monetary policy statements (see Table 5 and Figure 6) also reveals how central banks have institutionalized certain communication patterns to convey policy intentions. These formalized linguistic structures may enhance the predictability of policy communications but could potentially constrain adaptability during rapidly evolving economic conditions. Policymakers might benefit from considering how these established communication patterns influence market interpretations and whether greater flexibility in communication styles could improve policy effectiveness in certain contexts.

Beyond these specific findings, our analysis suggests that if central bank narratives actively shape global financial conditions—as our Granger causality results suggest—then coordinating communication strategies may be as important

as coordinating policy rate decisions. The post-GFC trend toward longer, more detailed statements reflects a genuine tradeoff: while elaboration enhances transparency, it also increases the complexity of the signal that markets must interpret. Our readability analysis suggests that central banks have partially offset this complexity with simpler sentence structures, but the optimal balance between completeness and clarity remains an open question. The substantial cross-model agreement we document among LLMs suggests that these tools can reliably extract economically meaningful signals from text, but the systematic differences we observe (such as Claude's tendency to identify the Federal Reserve specifically while other models reference the U.S. economy more broadly) highlight the importance of methodological transparency when using these tools for policy analysis.

5.2 The MPSD as a Research Tool

The MPSD offers significant advantages for researchers in economics, finance, and related fields. By providing standardized access to monetary policy communications across multiple countries and time periods, it enables novel comparative analyses previously infeasible due to data limitations. For researchers interested in comparative economics, the database facilitates studies of how institutional differences across central banks manifest in their communication strategies. For those studying policy transmission, the coverage allows for analysis of how market reactions to policy communications vary across different economic environments and institutional frameworks. The significant variations in statement frequency, length, and readability documented in Tables 3 and 4 provide a rich set of institutional variations that can be exploited to identify the determinants and effects of different communication approaches.

The integrated technical infrastructure supporting the database (illustrated in Tables 1 and 2) is particularly valuable for researchers applying computational methods to economic questions. The modular design allows for customization to specific research needs while maintaining reproducibility, addressing a key challenge in computational social science (Gentzkow et al., 2019). Our standardized folder structure (Figure 1) and systematic approach to data processing ensure that analyses can be precisely replicated across different research contexts. The evolution of monetary policy statement characteristics over time, as documented in Figure 5, provides valuable insights into how central bank communication strategies have adapted to changing economic conditions and policy frameworks. Researchers can exploit this temporal variation to study how communication evolves in response to institutional learning, changes in monetary policy regimes, and shifts in the broader economic environment.

5.3 Future Extensions

Several promising directions exist for extending both the database and its applications. First, expanding the corpus to include speeches, minutes, and press conferences would provide a more complete view of central bank communication strategies and enable analysis of how different channels interact to shape market expectations.

Second, incorporating multimedia elements such as press conference transcripts and videos could enable multimodal analysis, including non-verbal cues that may convey policy signals. Advances in computer vision and audio processing make such analysis increasingly feasible.

Third, developing domain-specific pre-training approaches for language models—including continued pre-training on central bank corpora or specialized architectures that account for the hierarchical and temporal structure of policy communications—could improve performance on monetary policy tasks. Our analysis of part-of-speech patterns (Table 5 and Figure 6) provides a foundation for such specialized models.

Finally, integrating the database with high-frequency financial data could facilitate more granular analyses of market reactions to specific communication elements, advancing our understanding of the information effects of monetary policy communications (Nakamura and Steinsson, 2018; Bauer and Swanson, 2023). Building on our finding that sentiment Granger causes the GFC (Figures 13 and 19), such integration would enable more precise identification of which aspects of communication drive market responses across different asset classes.

6 Conclusion

This paper introduces the MPSD, a research infrastructure that addresses fundamental gaps in the empirical analysis of monetary policy communication. By compiling and standardizing 6,693 monetary policy statements from 51 central banks spanning 1990–2024, the database constitutes an unprecedented resource for systematic cross-country analysis, offering exceptional temporal and geographic coverage (Figures 2 and 4). Its technical architecture integrates version control, automated data collection, and standardized preprocessing pipelines (Table 1), ensuring precise replicability while leveraging advanced computational tools. The modular design facilitates extension to diverse research questions (Table 2), supporting the growing integration of NLP methods into economic analysis.

Our linguistic analysis uncovers systematic patterns in how central banks structure their communications (Table 5 and Figure 6), advancing the understanding of institutionalized discourse in monetary policy and informing the development of specialized NLP tools for this domain. Exploiting the breadth of the database, we document a fundamental transformation in central bank communication: statements have approximately tripled in length since 1990, with

the expansion accelerating markedly after the Global Financial Crisis, while readability has modestly improved as central banks offset greater informational content with simpler sentence structures (Figure 5). In addition, inflation references comove substantially across countries during episodes of elevated global inflationary pressure, with persistent cross-country differences reflecting heterogeneity in mandates, inflation dynamics, and policy frameworks (Figure 9).

Our empirical application investigates the communication channel of international policy transmission, focusing on the GFC. Using LLMs for question answering and aspect-based sentiment analysis, we document that central banks typically attribute global financial conditions to broad U.S. economic factors rather than specifically to Federal Reserve policy actions. Sentiment analysis reveals greater concern about growth and financial conditions than about monetary policy spillovers in the post-2008 period. Granger causality tests based on both a benchmark dictionary-based sentiment index and LLM-based indicators (Figures 13 and 19) suggest that statement sentiment may drive components of the GFC rather than merely respond to them.

Regional variations in central bank sentiment (Figure 14) reveal important heterogeneity in how central banks perceive and react to global financial conditions, with emerging-market central banks showing greater sensitivity to capital flow concerns and exchange rate pressures relative to their advanced-economy counterparts. These differences indicate that the transmission of global financial shocks depends significantly on institutional factors and the degree of financial integration, carrying important implications for international policy coordination. The comparative analysis of LLM-generated and lexical sentiment indices (Figure 12) illustrates both the potential and the limitations of advanced language models for economic research, underscoring the value of methodological pluralism in computational text analysis.

These findings contribute to the understanding of monetary policy transmission in an interconnected global economy and underscore the importance of the communication channel as a mechanism through which monetary policy transmits across borders. The systematic approach to text analysis developed here offers a replicable template for applying frontier NLP methods to economic research. By making the MPSD and all associated code publicly available, we aim to support future research on monetary policy communication, financial market reactions, and policy coordination. As central banks continue to refine their communication strategies, rigorous empirical analysis of these communications will remain essential for understanding how policy intentions translate into economic outcomes.

7 Acknowledgments

We are grateful to two anonymous referees and the Editor for their constructive comments and suggestions, which substantially improved this paper. We also thank participants at the 4th IFC and Bank of Italy Workshop on Data Science in Central Banking for helpful discussions. All remaining errors are our own.

8 Code and Data Availability

The full codebase used for data processing, analysis, and figure generation, together with the compiled Monetary Policy Statement Database, excluding the scraping code,[5] is publicly available at `https://github.com/CentralBankTexts/monetary-policy-statement-database`. The repository is released under the MIT License, allowing broad reuse while ensuring proper attribution. A permanent archived version of the repository, to ensure long-term accessibility and citability consistent with standard reproducibility practices, will be archived with a permanent DOI upon final publication.

A Appendix

A.1 LLM Sentiment Prompt

```
instruction: |
  You will be analyzing a text from a central bank. The text is provided
   ↪ below:

  <central_bank_text>
  {{CENTRAL_BANK_TEXT}}
  </central_bank_text>

  Your task is to identify discussions related to the impact of global
   ↪ financial conditions on domestic economies. Focus on the following
   ↪ themes:
  1. Economic growth
  2. Financial conditions
  3. Capital flows
  4. Monetary policy spillovers

  Carefully read through the text. For each theme explicitly discussed,
   ↪ assign a continuous sentiment score on a scale from 0.0 to 2.0.
```

[5] While institution-specific scrapers are not distributed, the compiled corpus and full analysis code enable reproduction of all reported results.

```
  Scoring Guidelines:
  * 0.0 to < 1.0 (Negative): The text suggests unfavorable impacts, risks,
   ↪  contraction, or tightening.
      * Closer to 0.0: Severe risks, crises, or strong negative shocks.
      * Closer to 1.0: Mild headwinds or slight downside risks.
  * 1.0 (Neutral): The text presents a balanced view, negligible impact,
   ↪  or uncertainty with no clear direction.
  * > 1.0 to 2.0 (Positive): The text suggests favorable impacts,
   ↪  expansion, loosening, or resilience.
      * Closer to 1.0: Mild tailwinds or slight improvements.
      * Closer to 2.0: Strong growth, robustness, or highly beneficial
       ↪  conditions.

  Output Rules:
  * Only include themes that are explicitly discussed.
  * Use floating-point numbers to capture nuance (e.g., 0.45, 1.2, 1.85).
  * Format the output as a Python list of lists, where each inner list
   ↪  contains the theme string and the score.

  Your final output should be a Python list containing only the
   ↪  theme-score pairs. Do not provide any additional commentary.

  Example Output:
  [["economic_growth", 1.8], ["financial_conditions", 0.45],
   ↪  ["capital_flows", 1.0]]

output_format: |
  A Python list of lists containing theme and float score pairs (e.g.,
   ↪  [["economic_growth", 1.25], ["financial_conditions", 0.8]]).
```

A.2 LLM Narrative Prompt

```
instruction: |
  The following texts come from central banks. The country is identified
   ↪  in the Country: and the Comment: is the text released by the central
   ↪  bank.
  I want you to identify the main driver of global economic and financial
   ↪  conditions and the origin country driving these conditions.
  In the summary please note if any central banks identify the Federal
   ↪  Reserve as the driver of global economic or financial conditions.

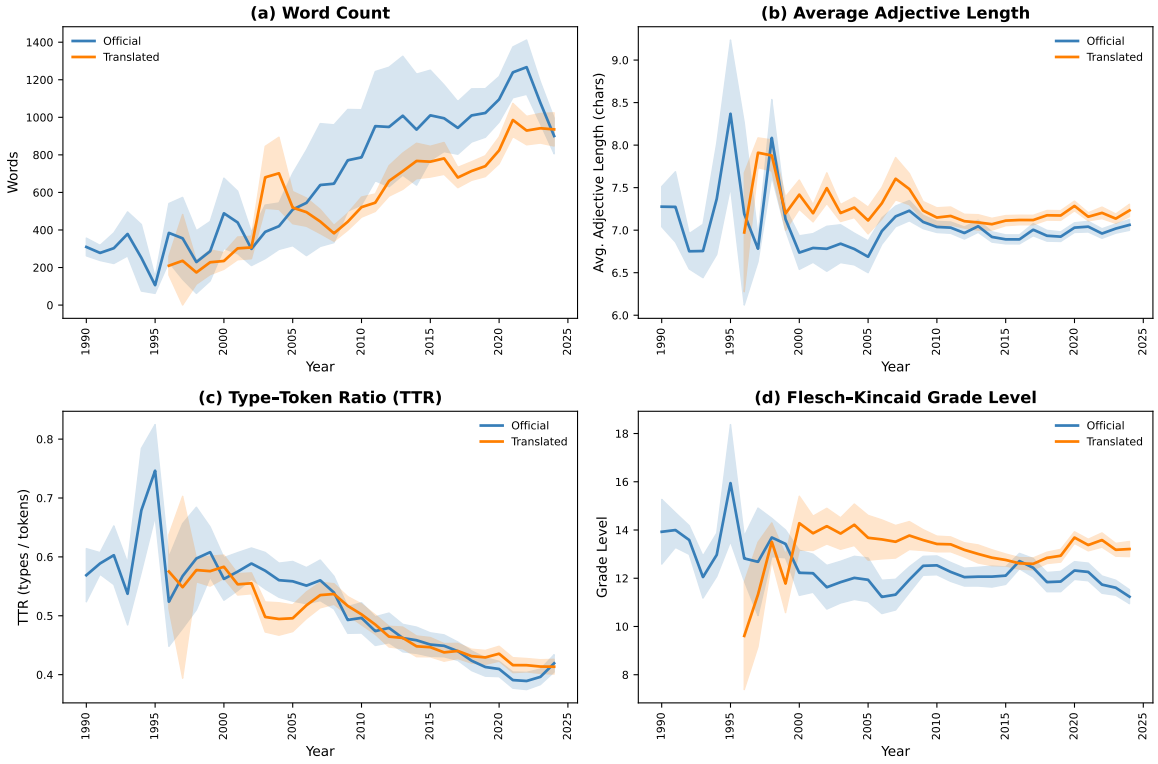
output_format: |
  Please respond in the following format:
  Driver: [Identified Driver]
  Origin Country: [Origin Country]
```

```
Summary: [brief explanation]
```

A.3 Language Groups Analysis

We distinguish between monetary policy statements originally drafted in English and those translated into English from other languages. Figure 15 reports the evolution of linguistic indicators separately for these two groups, allowing us to assess whether observed patterns are driven by language or translation effects. For statements originally drafted in English, the results closely mirror the full-sample trends. Average word counts increase markedly over time, while lexical diversity, as measured by the type-token ratio, declines. At the same time, readability exhibits only a modest downward trend, indicating that although statements have become substantially longer, they have not become proportionally more complex in terms of sentence structure and word choice. Statements translated into English display broadly similar dynamics. Despite differences in levels, the time-series trends closely track those observed for native-English statements: translated texts become longer over time, show declining lexical diversity, and maintain relatively stable readability. Average adjective length remains largely unchanged in both groups, suggesting that the expansion of communication occurs primarily through increased volume rather than more complex evaluative wording. The similarity of trends across native-English and translated-English statements suggests that the main results are not driven by translation-related measurement artifacts. Instead, they reflect systematic changes in central bank communication practices over time. Persistent level differences across groups are more plausibly attributed to institutional drafting conventions and communication styles rather than language effects per se.

Figure 15: Average Statement Length and Readability Measure by Language Groups



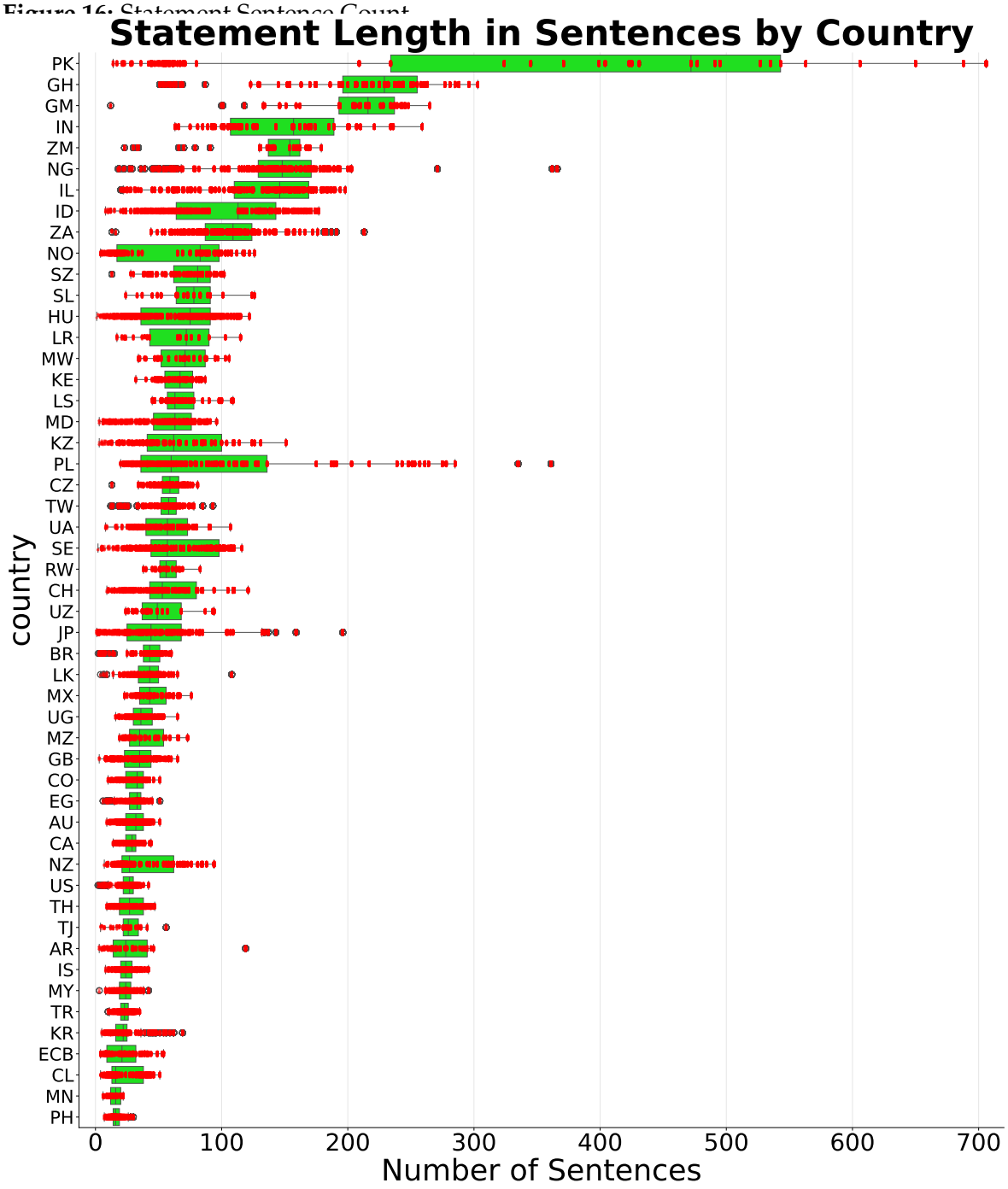
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: For each central bank statement, we compute four document-level linguistic indicators for countries with English as official language and English as translated language: (i) word count, defined as the total number of tokens; (ii) average adjective length, measured as the mean character length of adjectival tokens identified using part-of-speech tagging; (iii) the type–token ratio (TTR), defined as the number of unique tokens divided by total tokens; and (iv) readability, captured using standard readability indices (e.g., Flesch-Kincaid Grade Level). Each indicator is first computed at the statement level and then aggregated to the annual frequency by taking the cross-statement mean. To capture uncertainty and heterogeneity across institutions and over time, we report confidence bands constructed from the standard error of the annual mean.

A.4 Statement Sentence Count

We observe that unusually long "statements" in the corpus are concentrated in countries where the policy-rate decision is communicated through a broader analytical document rather than a short press release (see Figure 16). For example, in Pakistan (PK) the rate decision has often been published within, or together with, a Monetary Policy Statement/Report that includes detailed discussion of global and domestic economic developments and the outlook. A similar pattern appears in several other countries as well, such as Nigeria (NG) or Ghana (GH), for example, where MPC statements typically combine the decision with an extended review of macroeconomic conditions and risks. In India (IN), statement length is amplified by the release of a policy package around each meeting (e.g., MPC resolution together with accompanying policy statements), which can be captured as a single text. Large sentence-count variation primarily reflects institutional differences in communication formats and document bundling.

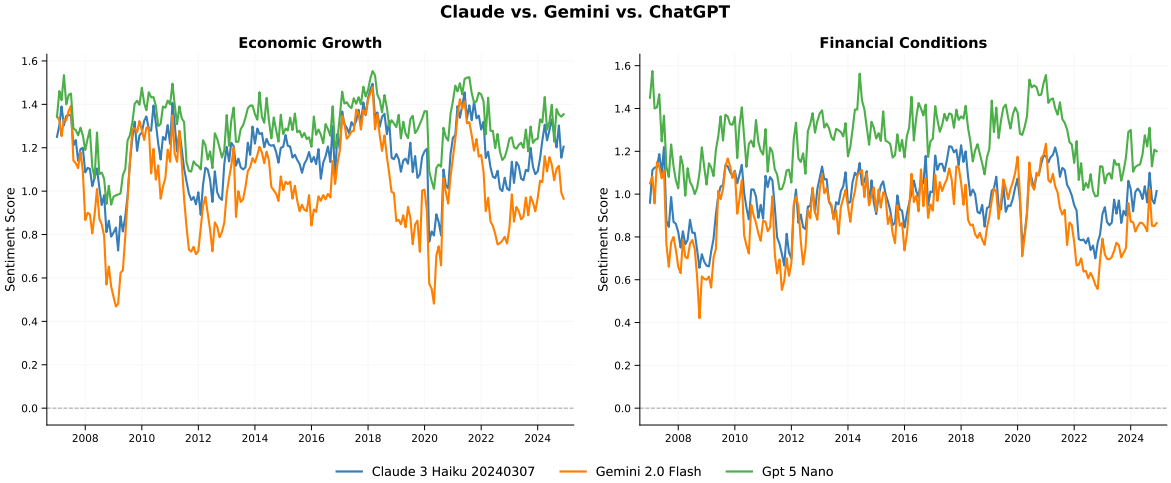
Figure 16: Statement Sentence Count



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.

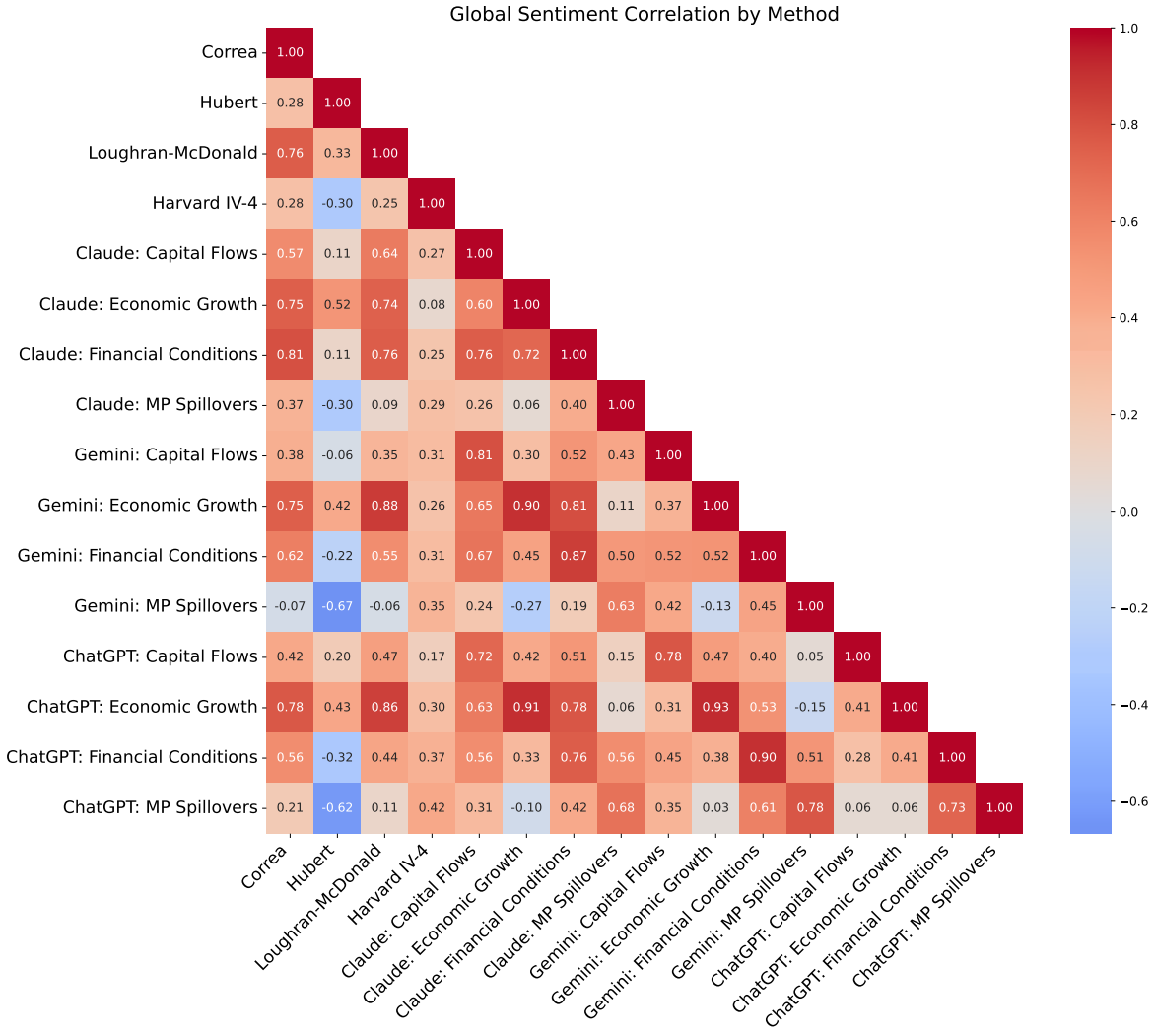
Notes: Statement length in sentences by country.

Figure 17: Cross-Model LLM Sentiment Comparison

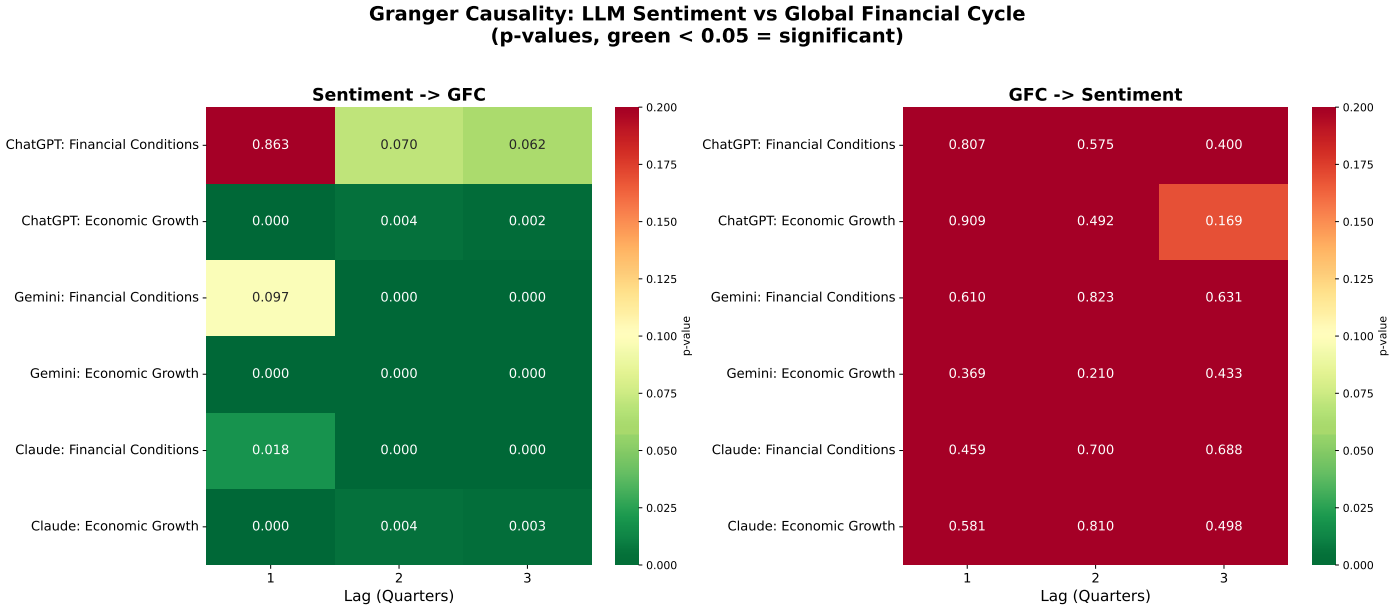


Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.

Figure 18: Extended Correlation Matrix of Sentiment Measures



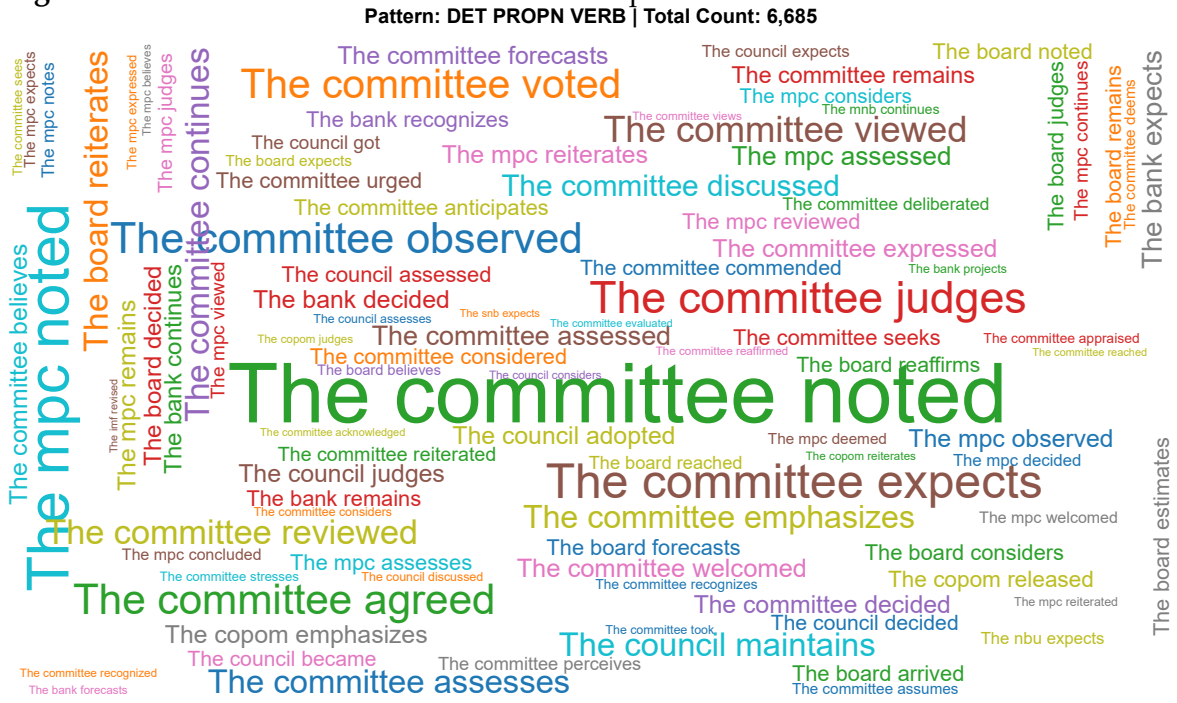
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: Sentiment measures are derived from MPSD statements using multiple LLMs (ChatGPT, Gemini, and Claude) and capture ABS related to economic growth and financial conditions. The GFC is from Miranda-Agrippino and Rey (2020). The figure reports p-values (rounded to three decimal places) from Granger causality tests at quarterly frequency. Green cells indicate statistical significance at the 5% level. Tests are conducted for lags of one to three quarters.

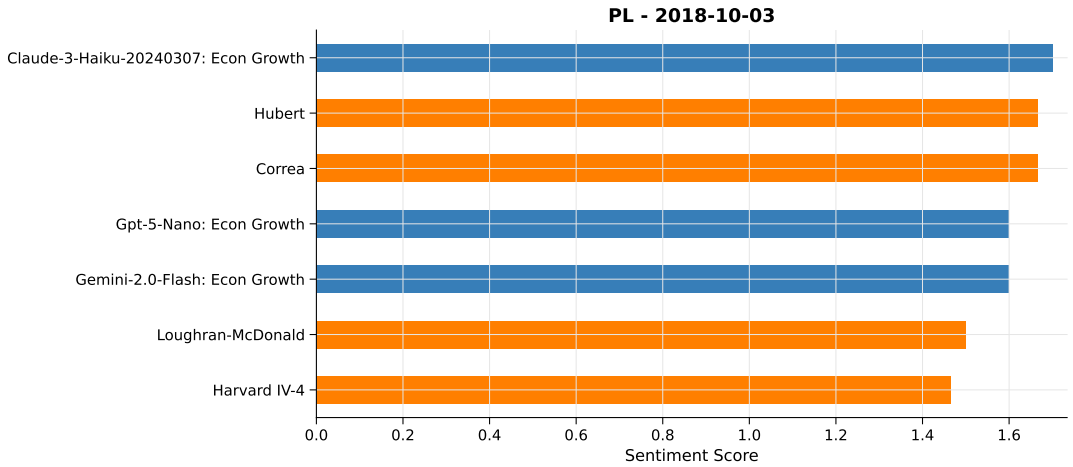
Figure 20: Most Common Three-Token Part-of-Speech Pattern `DET PROPN VERB`.



Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: The `DET PROPN VERB` pattern often forms the root of sentences associated with CB. Total count of three-token sentences: 6,685.

Figure 21: Sentiment Variation Across Example Statement, Poland 2018-10-03



Warsaw, 3 October 2018 Information from the meeting of the Monetary Policy Council held on 2-3 October 2018 The Council decided to keep the NBP interest rates unchanged: reference rate at 1.50%; lombard rate at 2.50%; rediscount rate at 0.50%; rediscount rate at 1.75%. Global economic conditions remain favourable. In the euro area, the economic situation continues to be favourable, despite slightly slower GDP growth than in 2017. In the United States, economic growth is higher than in the previous year, which confirms that economic conditions in this country are strong. In China, activity growth has continued at a stable pace for the past few quarters. Since the beginning of the year, global energy commodity prices, including those of oil, have risen substantially. This has contributed to higher inflation in many countries. At the same time, core inflation in the external environment of the Polish economy, including the euro area, remains moderate, despite continued strong econ

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: This figure illustrates the qualitative validation approach by displaying sentiment variation across methods for the same policy statement

References

Acosta, M. and E. E. Meade (2015). Hanging on every word: semantic analysis of the FOMC's postmeeting statement. FEDS Notes 2015-09-30, Board of Governors of the Federal Reserve System.

Armelius, H., C. Bertsch, I. Hull, and X. Zhang (2020). Spread the Word: International spillovers from central bank communication. *Journal of International Money and Finance 103*, 102116.

Assenmacher, K., G. Glöckler, S. Holton, P. Trautmann, D. Ioannou, S. Mee, K. Bakk-Simon, S. Bergbauer, M. Catenaro, E. Charalampakis, M. Ehrmann, G. Ferrero, D. Georgarakos, P. Gertler, A. Giovannini, R. Grandia, N. Hernborg, N. Herrala, D. Kedan, G. Kenny, T. Linzert, M. Manrique, R. Mestre, E. Mönch, S. Nardelli, N. Nomm, L. Pavlova, G. Schultefrankenfeld, M. Silgoner, I. Skotida, B. Winkler, M. T. Bitterlich, E. Argiri, C. Alonso, J. Byron, F. Arigoni, M. Deroose, M. Gardt, K. Istrefi, O. Goldfayn-Frank, J. Hellström, P. Huertgen, K. Kalnberzina, G. Kocharkov, V. Márquez, C. Ruhe, M. Šanta, R. Reedik, G. Sciot, A. Stylianou, and E. Taylor (2021). Clear, consistent and engaging: ECB monetary policy communication in a changing world. Technical Report 274, ECB Occasional Paper.

Bauer, M. D. and E. T. Swanson (2023). An alternative explanation for the "Fed Information Effect". *American Economic Review 113*(3), 664–700.

Benchimol, J., I. Caspi, and S. Kazinnik (2023). Measuring communication quality of interest rate announcements. *Economists' Voice 20*(1), 43–53.

Benchimol, J., S. Kazinnik, and Y. Saadon (2020). Communication and transparency through central bank texts. Paper presented at the 132nd Annual Meeting of the American Economic Association, January 3-5, 2020, San Diego, CA, United States.

Benchimol, J., S. Kazinnik, and Y. Saadon (2022). Text mining methodologies with R: an application to central bank texts. *Machine Learning with Applications 8*, 100286.

Benchimol, J., S. Kazinnik, and Y. Saadon (2025). Federal Reserve Communication and the COVID-19 Pandemic. *Manchester School 93*(5), 464–484.

Blinder, A. (2009). Talking about monetary policy: the virtues (and vice?) of central bank communication. BIS Working Papers 274, Bank for International Settlements.

Blinder, A., C. Goodhart, P. Hildebrand, D. Lipton, and C. Wyplosz (2001). *How Do Central Banks Talk?* Number 3 in Geneva Reports on the World Economy. London, UK: Centre for Economic Policy Research.

Blinder, A. S. (2018). Through a crystal ball darkly: The future of monetary policy communication. *AEA Papers and Proceedings 108*, 567–571.

Blinder, A. S., M. Ehrmann, J. de Haan, and D.-J. Jansen (2024). Central bank communication with the general public: promise or false hope? *Journal of Economic Literature 62*(2), 425–457.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. D. Haan, and D.-J. Jansen (2008). Central bank communication and monetary policy: a survey of theory and evidence. *Journal of Economic Literature 46*(4), 910–945.

Born, B., M. Ehrmann, and M. Fratzscher (2014). Central bank communication on financial stability. *Economic Journal 124*(577), 701–734.

Bruno, V. and H. S. Shin (2015). Capital flows and the risk-taking channel of monetary policy. *Journal of Monetary Economics 71*, 119–132.

Brusa, F., P. Savor, and M. Wilson (2020). One Central Bank to Rule Them All. *Review of Finance 24*(2), 263–304.

Bulír, A., M. Cihák, and D.-J. Jansen (2013). What drives clarity of central bank communication about inflation? *Open Economies Review 24*(1), 125–145.

Campiglio, E., J. Deyris, D. Romelli, and G. Scalisi (2025). Warning words in a warming world: Central bank communication and climate change. *European Economic Review 178*, 105101.

Cieslak, A. and A. Schrimpf (2019). Non-monetary news in central bank communication. *Journal of International Economics 118*, 293–315.

Coibion, O., Y. Gorodnichenko, and M. Weber (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy 130*(6), 1537–1584.

Correa, R., K. Garud, J. M. Londono, and N. Mislang (2021). Sentiment in central banks' financial stability reports. *Review of Finance 25*(1), 85–120.

Davis, J. S. and M. A. Wynne (2019). Central Bank Communications: A Case Study. In *Oxford Handbook of the Economics of Central Banking*. Oxford University Press.

De Grauwe, P. and Y. Ji (2023). On the use of current and forward-looking data in monetary policy: a behavioural macroeconomic approach. *Oxford Economic Papers 75*(2), 526–552.

de Haan, J. and L. Hoogduin (2024). ECB communication policies: An overview and comparison with the Federal Reserve. *Journal of International Money and Finance 142*, 103050.

Dincer, N., B. Eichengreen, and P. Geraats (2022). Trends in monetary policy transparency: further updates. *International Journal of Central Banking 18*(1), 331–348.

Ehrmann, M. and A. Wabitsch (2022). Central bank communication with non-experts - A road to nowhere? *Journal of Monetary Economics 127*, 69–85.

Eijffinger, S. and D. Masciandaro (2014). *Modern Monetary Policy and Central Bank Governance*. Cheltenham, UK: Edward Elgar Publishing.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gonzalez, M. and R. C. Tadle (2022). Monetary policy press releases: an international comparison. BIS Working Papers 1023, Bank for International Settlements.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Hansen, S. and M. McMahon (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics 99*(S1), 114–133.

Hansen, S., M. McMahon, and M. Tong (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics 108*(C), 185–202.

Hayo, B. and M. Neuenkirch (2015). Central bank communication in the financial crisis: Evidence from a survey of financial market participants. *Journal of International Money and Finance 59*(C), 166–181.

Issing, O. (2019). *The long journey of central bank communication*. Cambridge, MA: MIT Press.

Kahn, G. A. (2007). Communicating a policy path: the next frontier in central bank transparency? *Economic Review-Federal Reserve Bank of Kansas City 92*(1), 25.

Kalemli-Özcan, Ş. (2019). U.s. monetary policy and international risk spillovers. In *Proceedings - Jackson Hole Economic Symposium*, Jackson Hole, WY, pp. 95–191.

Kryvtsov, O. and L. Petersen (2021). Central bank communication that works: Lessons from lab experiments. *Journal of Monetary Economics 117*(C), 760–780.

Kumar, A., S. K. Mallick, M. Mohanty, and F. Zampolli (2023). Market volatility, monetary policy and the term premium. *Oxford Bulletin of Economics and Statistics 85*(1), 208–237.

Lewis, D., C. Makridis, and K. Mertens (2020). Do monetary policy announcements shift household expectations? CEPR Discussion Papers 14360, Centre for Economic Policy Research.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance 66*(1), 35–65.

Macklem, T. and J. Vardy (2023). 20 years of central bank communications, and lessons for the future. In *Central banking in the Americas: Lessons from two decades*, pp. 55–67. Bank for International Settlements.

Masciandaro, D., O. Peia, and D. Romelli (2024). Central bank communication and social media: from silence to Twitter. *Journal of Economic Surveys 38*(2), 365–388.

Miranda-Agrippino, S. and H. Rey (2020). U.S. monetary policy and the Global Financial Cycle. *Review of Economic Studies 87*(6), 2754–2776.

Miranda-Agrippino, S. and H. Rey (2022). The global financial cycle. In *Handbook of International Economics*, Volume 6, pp. 1–43. Elsevier.

Morris, S. and H. S. Shin (2002). Social value of public information. *American Economic Review 92*(5), 1521–1534.

Nakamura, E. and J. Steinsson (2018). High-frequency identification of monetary non-neutrality: the information effect. *Quarterly Journal of Economics 133*(3), 1283–1330.

Obstfeld, M. (2015). Trilemmas and tradeoffs: living with financial globalization. In C. Raddatz, D. Saravia, and J. Ventura (Eds.), *Global Liquidity, Spillovers to Emerging Markets and Policy Responses*, Volume 20, Chapter 2, pp. 013–078. Central Bank of Chile.

Picault, M. and T. Renault (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance 79*, 136–156.

Rey, H. (2015). Dilemma not trilemma: the global financial cycle and monetary policy independence. NBER Working Papers 21162, National Bureau of Economic Research, Inc.

Vyshnevskyi, I., W. Jombo, and W. Sohn (2024). The clarity of monetary policy communication and financial market volatility in developing economies. *Emerging Markets Review 59*, 101121.

Woodford, M. (2005). Central bank communication and policy effectiveness. *Proceedings - Economic Policy Symposium - Jackson Hole Aug*, 399–474.

The Monetary Policy Statement Database

An LLM Application to Global Financial Conditions*

Cory Baird[1] Jonathan Benchimol[2] Wook Sohn[3] Vira Vyshnevska[3] Iegor Vyshnevskyi[4]

[1]University of Tokyo, Tokyo, Japan
Corresponding author. Email: cbaird@g.ecc.u-tokyo.ac.jp

[2]Bank of Israel, Jerusalem, Israel

[3]KDI School, Sejong, South Korea

[4]Sogang University, Seoul, South Korea

February 18, 2025

This presentation does not necessarily reflect the views of the Bank of Israel or any other institutions.

*Code and data: `github.com/CentralBankTexts/monetary-policy-statement-database`

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Not Just a Database: Reproducible AI for Macroeconomics

- **Scientific Rigor**
  Infrastructure to maintain scientific standards in text analysis.

- **Unprecedented Scale**
  6,693 policy rate statements from 51 central banks worldwide.

- **Cutting-Edge Infrastructure**
  Transparent, automated, and scalable pipelines.

- **Applying AI to Macroeconomics**
  Leverages AI to identify economic narratives and sentiment over time and across countries.
  - Global Financial Cycle origins and drivers (Rey, 2015).

Current Literature

Key Limitations in CB Communication Research:

- **Data & Standardization:** Existing research focuses on a few countries or uses only BIS speeches (Born et al., 2014; Picault and Renault, 2017).
  - Contribution: standardized dataset of 51 CBs
    Larger than existing non-public databases, e.g., Gonzalez and Tadle (2022)

- **Reproducibility:** Lack of open-source pipelines impedes reproducible research (Gentzkow et al., 2019; Hansen and McMahon, 2016).
  - Contribution: transparent and accessible text cleaning, and Large Language Model (LLM) pipelines

Applying the Scientific Method to "AI" and NLP

- Failure to share data/algorithms hinders reproducibility and efficiency
  - **Can reproduce** lexical/dictionary methods at immense cost
  - **Cannot reproduce** black box LLMs (Claude, Gemini) because data, prompts, and models must be **exactly** the same

- **Irreproducible NLP methods** hinder scientific progress and data reliability in downstream modeling

Data Collection	**Featurization**	**Replicable AI Use**
Identify statement links	Cleaning	Replicable LLM infrastructure
Scrape statements	Sentence segmentation	Track inputs (prompts, texts)
Automate retrieval	Part-of-speech tagging	Track outputs (responses)

Table 1: End-to-End Transparency for Replicable AI

Source: Authors.

Infrastructure Overview

- **Transparency**
  - All code/scripts made public on GitHub[1]
  - Technologies: GitHub, Pyenv, Poetry
- **Automation**
  - Scripts executed regularly via a Data Version Control (DVC) pipeline
  - Technologies: DVC, GitHub Actions.
- **Scalability**
  - Modular scripts easily adaptable for other CB texts
  - Technologies: Python Object Oriented Programming, GitHub
- **Reproducibility**
  - Transparent text analysis
  - Reproducible pipelines for cleaning, featurization, and LLMs
  - Technologies: UV, Pyenv, Poetry, S3, DVC

[1] `github.com/CentralBankTexts/monetary-policy-statement-database`

MPSD Overview

- Monetary Policy Statement Database (MPSD)
  - **Scope**
    51 CBs (35 API and 16 static currently cleaned)
  - **Data**
    Approximately 6,693 unique MPSs
  - **Temporal Coverage**
    1990–2024

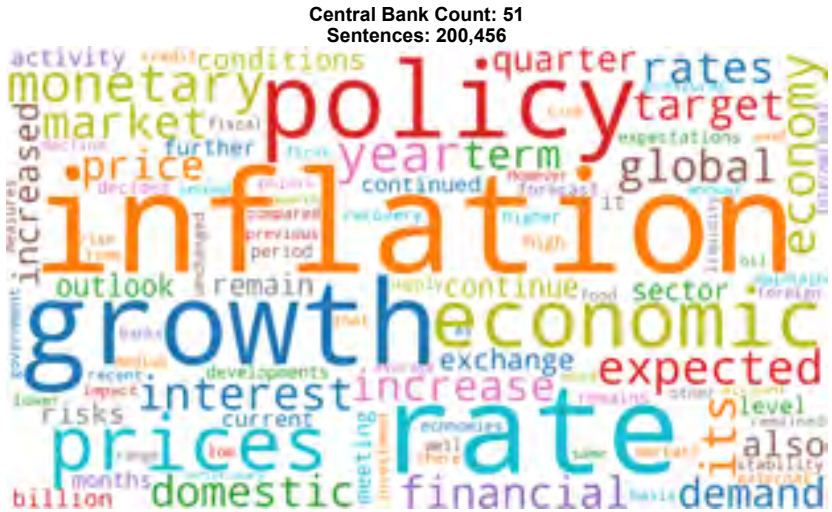


Figure 1: Standardized and diverse corpus for training/testing A.I.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Technical Infrastructure: Code & Data Management

- **Data integrity and reproducibility**
  - Data Version Control (DVC) tracks full lifecycle: web scraping $\rightarrow$ LLM analysis
  - Parameterized pipelines enforce data lineage at every stage
- **Code and environment management**
  - Code versioning: Git and GitHub
  - Python version and dependency management: uv
  - Cloud storage: AWS S3 (raw, processed, and artifact partitions)
- **Automated updates**
  - GitHub Actions: daily checks for new CB policy statements
  - Single-command replication of data, models, and outputs
- **LLM infrastructure**
  - Abstract LLM class integrating Claude, ChatGPT, Gemini, and Hugging Face
  - Methods: model selection, YAML prompt processing, response generation/saving

Code/Folder Infrastructure: Automated & Scalable

- **Modular Design:** Programs/scripts are modularized for:
  1. Text retrieval
  2. Text cleaning
  3. LLM inference
- **Standardized Folder Structure:**

  `src` : Python scripts
  `data` : Text files
  `references` : GenAI prompts
- **MPSD Updates:** Automated updates using DVC.
- **End-to-End Reproducibility:** A full pipeline ensures reproducible results from data acquisition to NLP analysis.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

The Rise of Central Bank Communication

- CB communication dramatically changed since the 1990s
  - First official MPS: Reserve Bank of Australia in 1990
  - Rapid growth in MPSs after the year 2000
- MPSs are a crucial tool for:
  - Managing inflation expectations
  - Enhancing policy predictability
  - Building credibility and trust

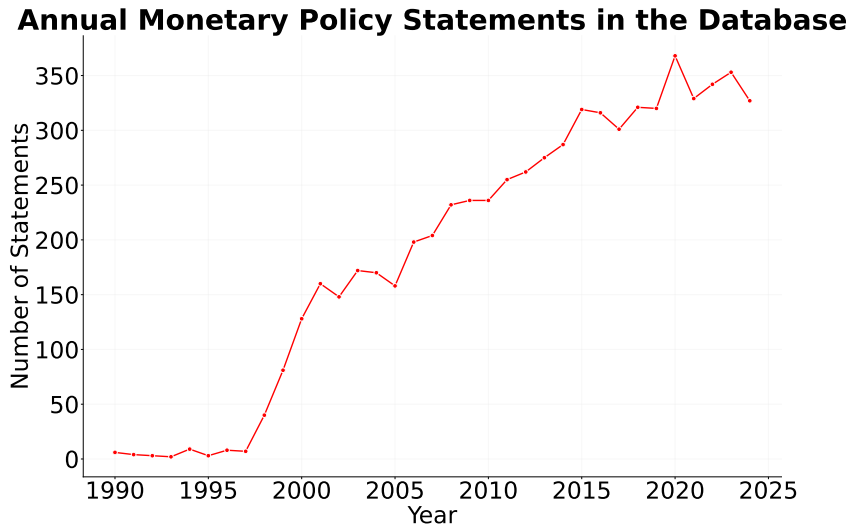


Figure 2: Annual MPSs in the MPSD

Dataset Coverage

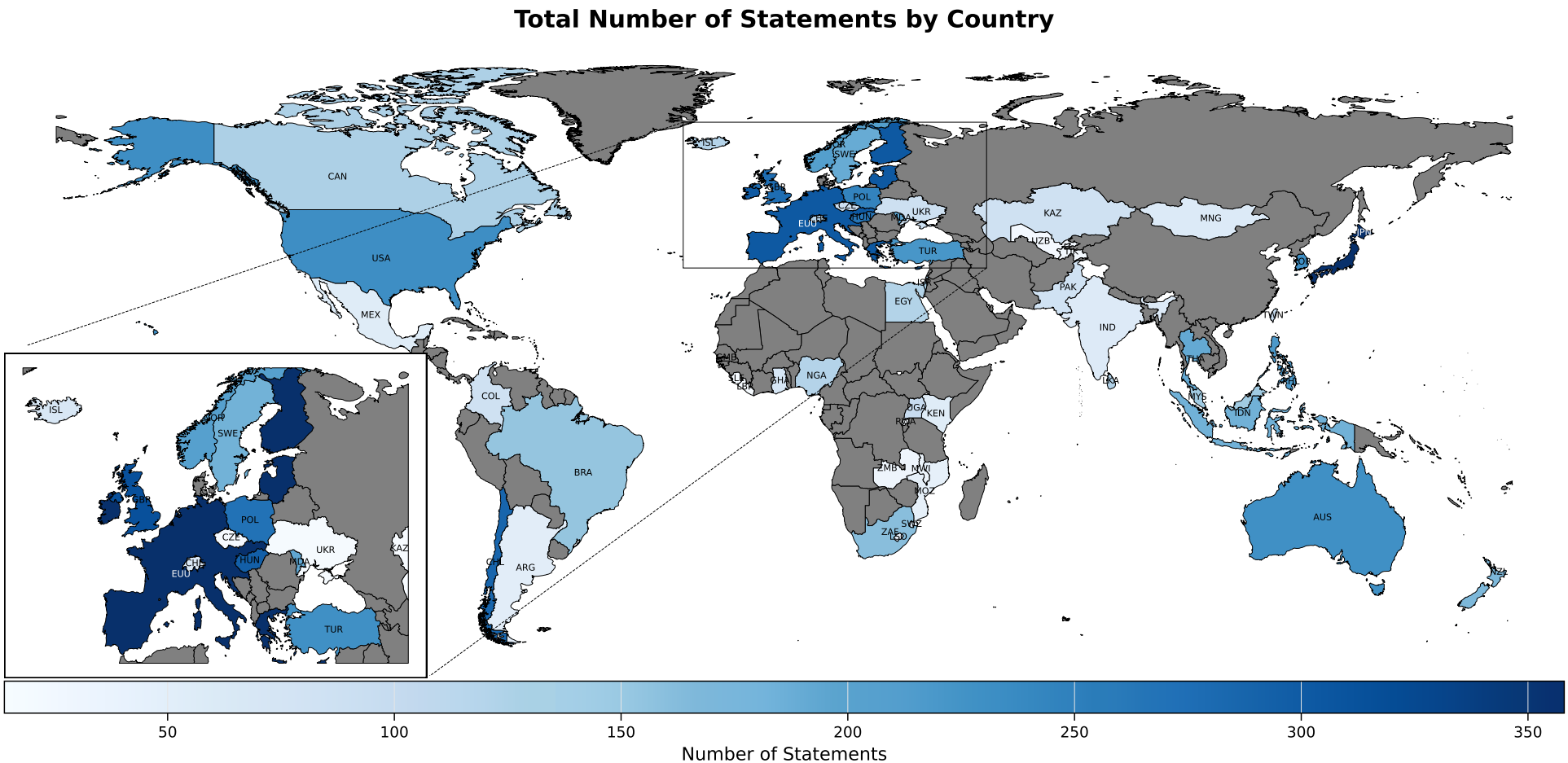


Figure 3: Countries and Monetary Policy Statements

Dataset Coverage

- The BoJ (Japan) and ECB emerge as the most prolific publishers of statements
- Central banks in middle-income countries exhibit more limited publication frequency and shorter historical coverage
- These cross-institutional variations reflect differences in institutional frameworks and communication strategies

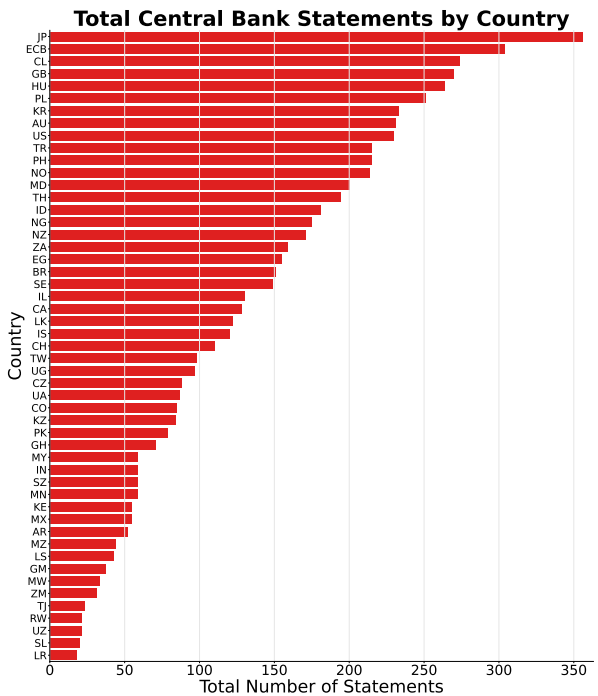


Figure 4: Statements by Country

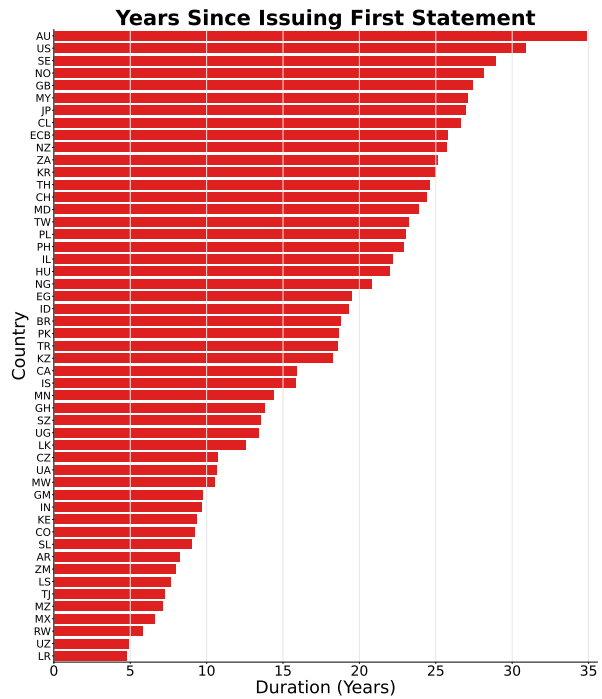


Figure 5: Years of Coverage by Country

Trends in Communication Over Time

- MPSs growing lengthier and more complex post-GFC
- MPSs becoming slightly more "readable" over time despite increased length

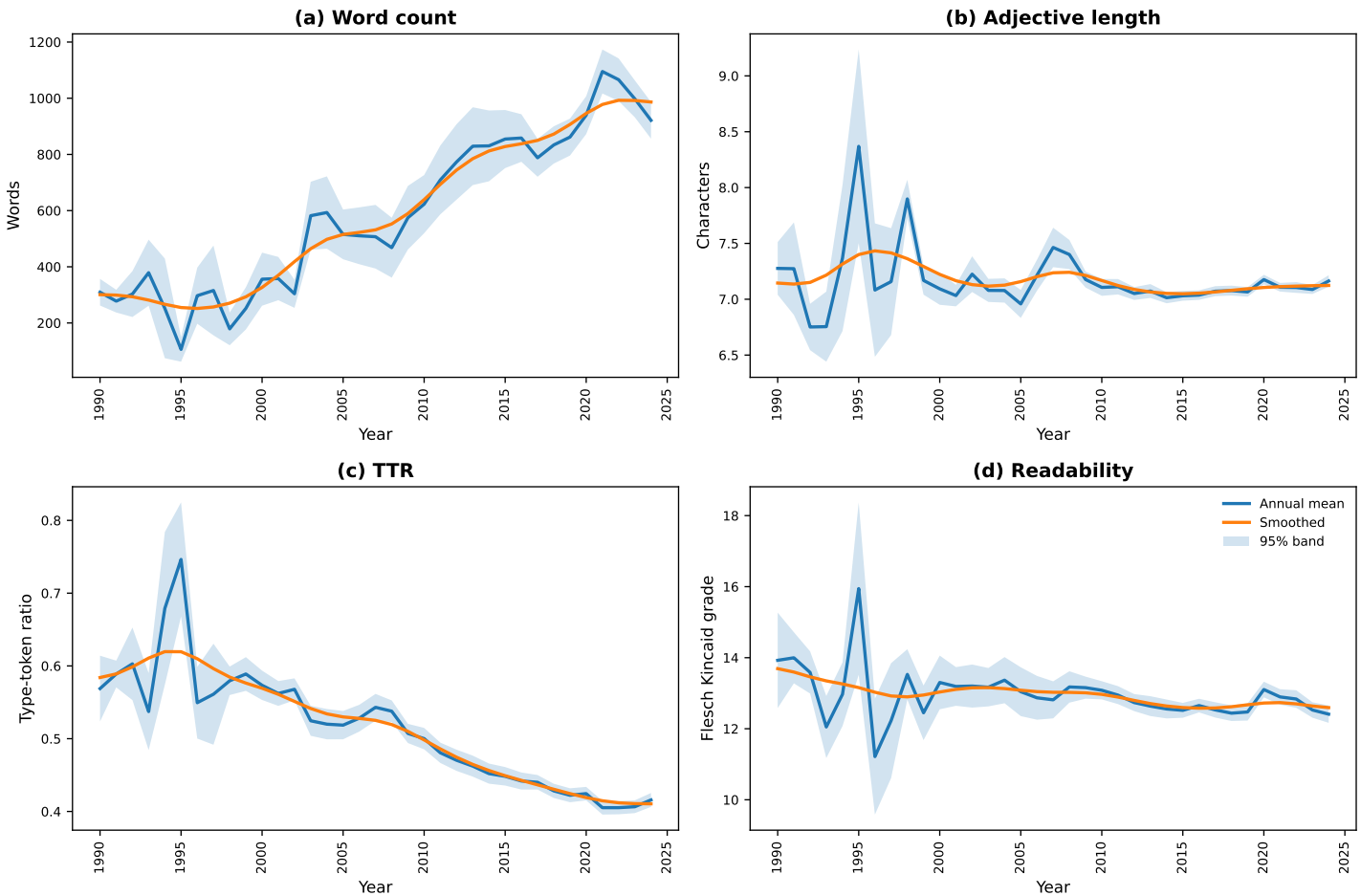


Figure 6: Linguistic Trends in CB Statements Over Time

Key Characteristics (Patterns) of the MPS

Economic Theme	**POS Pattern**	**Example**
Inflation Forecasts	inflation + AUX + VERB	inflation **will rise**; inflation **might persist**
Growth Projections	growth/GDP + AUX + VERB	growth **should recover**; GDP **may stall**
Inflation Dynamics	NOUN + ADP + inflation	risk **of** inflation; outlook **for** inflation
Rate Levels	rate + ADP + NUM	rate **at** 5.25%; rate **above** 2.0
Forward Guidance	policy/rate + AUX + VERB	policy **will remain**; rates **might stay**
Monetary Stance	policy + AUX + ADJ	policy **is restrictive**; stance **remains loose**
Fiscal Policy	NOUN + ADP + spending/deficit	impact **of** deficit; boost **from** spending
External Conditions	NOUN + ADP + trade/export	demand **for** exports; uncertainty **about** trade

Table 2: Linguistic Patterns in Monetary Policy Statements by Economic Theme

Most Common Sentence Patterns by Part-of-speech

- Common sentence patterns include:
  - `DET ADJ NOUN` (e.g., "The strong dollar," "The current rate")
  - `DET NOUN ADP` (e.g., "The impact of," "The level of")
  - `ADP DET NOUN` (e.g., "In the market," "On the economy")
- Sentences identified with spaCy's Universal POS tags and pattern matching.

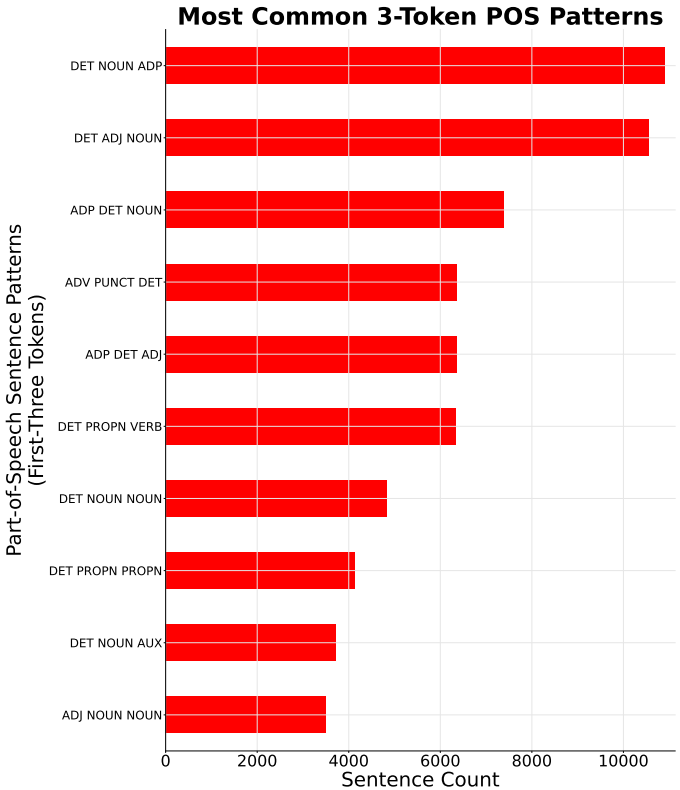


Figure 7: Most Common Patterns by Part-of-Speech (POS)

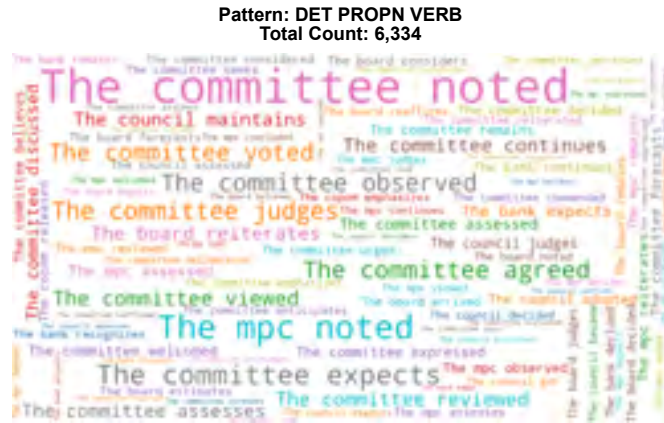


Figure 8: POS: The `DET PROPN VERB` pattern is associated with CB commentary/actions

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Addressing AI Challenges with the Corpus/Infrastructure

- **Challenges with LLMs**
  1. Irreproducibility: Insufficient transparency regarding model inputs and specifications makes replicating it difficult.
     - We provide open-source code, including LLM code, prompts, and Jupyter notebooks, ensuring full transparency and reproducibility.
  2. High computational costs and energy consumption.
     - We implement text filtering (keyword & part-of-speech) pipelines to reduce the computational burden.

- **Application to Econ: The Global Financial Cycle (GFC)**
  - Communication Channel: CB Communication $\Rightarrow$ Market Expectations $\Rightarrow$ Financial Conditions
  - Using our LLM/Data Pipeline we leverage LLMs for Question Answering (QA) & classification (topic, sentiment) tasks to explore origins and drivers of the Global Financial Cycle (Miranda-Agrippino and Rey, 2020, 2022)

Leveraging Our LLM/Data Pipeline

- Infrastructure is abstract for scalability
  - LLM: seamlessly add more models (ChatGPT, Huggingface, etc) and tasks (classification, summarization, Q&A)
  - Text retrieval: seamlessly add more central banks and types of communications (minutes, speeches, etc)

Modular LLM-based Data Processing Framework for Academic Research		
Data Acquisition Module (Abstract Class)	**LLM Processing Module (Abstract Class)**	**LLM Data Pipeline Workflow**
Methods: 1. Initialize (source parameters) 2. Scrape Statement Links 3. Download/Save Statements **Output:** Text Files CSV/Parquet Text Dataframes **Abstraction:** Extend abstract functions to accommodate new data sources (e.g., central bank speeches, minutes) and communication modes.	**Methods:** 1. Initialize (LLM Model) 2. Process/Load Prompts 3. Generate Response **Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) **Abstraction:** Implement model-agnostic functions to leverage diverse LLM architectures (e.g., Gemini, Claude, HuggingFace models).	**Methods:** 1. Initialize (Input Corpus) 2. Filter via Keywords or POS 3. LLM Inference 4. Save Response **Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) Figure for Research (PNG/PDF) **Use-Case:** Utilize concise, reusable scripts for seamless integration with LLM APIs (e.g., Gemini, Claude) by researchers.

Table 3: AI Infrastructure for Central Bank Communication Analysis

LLM Application: Exploring the Global Financial Cycle

- LLM Application: Question Answering (QA)
  - Prompts model to summarize drivers of global economy

    "The following texts come from central banks. [. . .] I want you to identify the main driver of global economic and financial conditions and the origin country driving these conditions. In the summary please note if any central banks identify the Federal Reserve as the driver of global economic or financial conditions."

- LLM Application: Aspect Based Sentiment[2](ABS)
  - Prompts model to extract sentiment for specific themes on a continuous scale (0.0–2.0)

    "Your task is to identify discussions related to the impact of global financial conditions on domestic economies. Focus on the following themes: 1. Economic growth, 2. Financial conditions, 3. Capital flows, 4. Monetary policy spillovers. For each theme explicitly discussed, assign a continuous sentiment score on a scale from 0.0 to 2.0, where 1.0 is neutral."

- Findings from LLM Analysis:
  - QA: Central Banks often cite the U.S. economy as a driver of global economic conditions rather than specifically the U.S. Federal Reserve.
  - ABS: Sentiment analysis shows that central banks expressed more concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period

Appendix: Full Prompt Example

[2]Aspect-based sentiment identifies sentiment of specific theme; see Appendix A.1–A.2 in the paper for full prompts

LLM Application: QA Results Differ by Model (2008)

Model	**Mechanism**	**Origin**	**Key Finding**
Claude	The Federal Reserve	United States	Names the Fed as key driver; US housing bust a recurring theme
Xiaomi	Global financial market turmoil / subprime crisis	United States	Fed acknowledged by multiple CBs as key driver
Mistral	US economy slowdown / subprime crisis	United States	No CB explicitly names the Fed as the driver
Arcee	US subprime crisis and broader US economic weakness	United States	Fed's policy actions cited as relevant but not the primary driver

LLM summaries of the same CB sentences differ in how prominently they identify the Federal Reserve

LLM Application: Aspect Based Sentiment

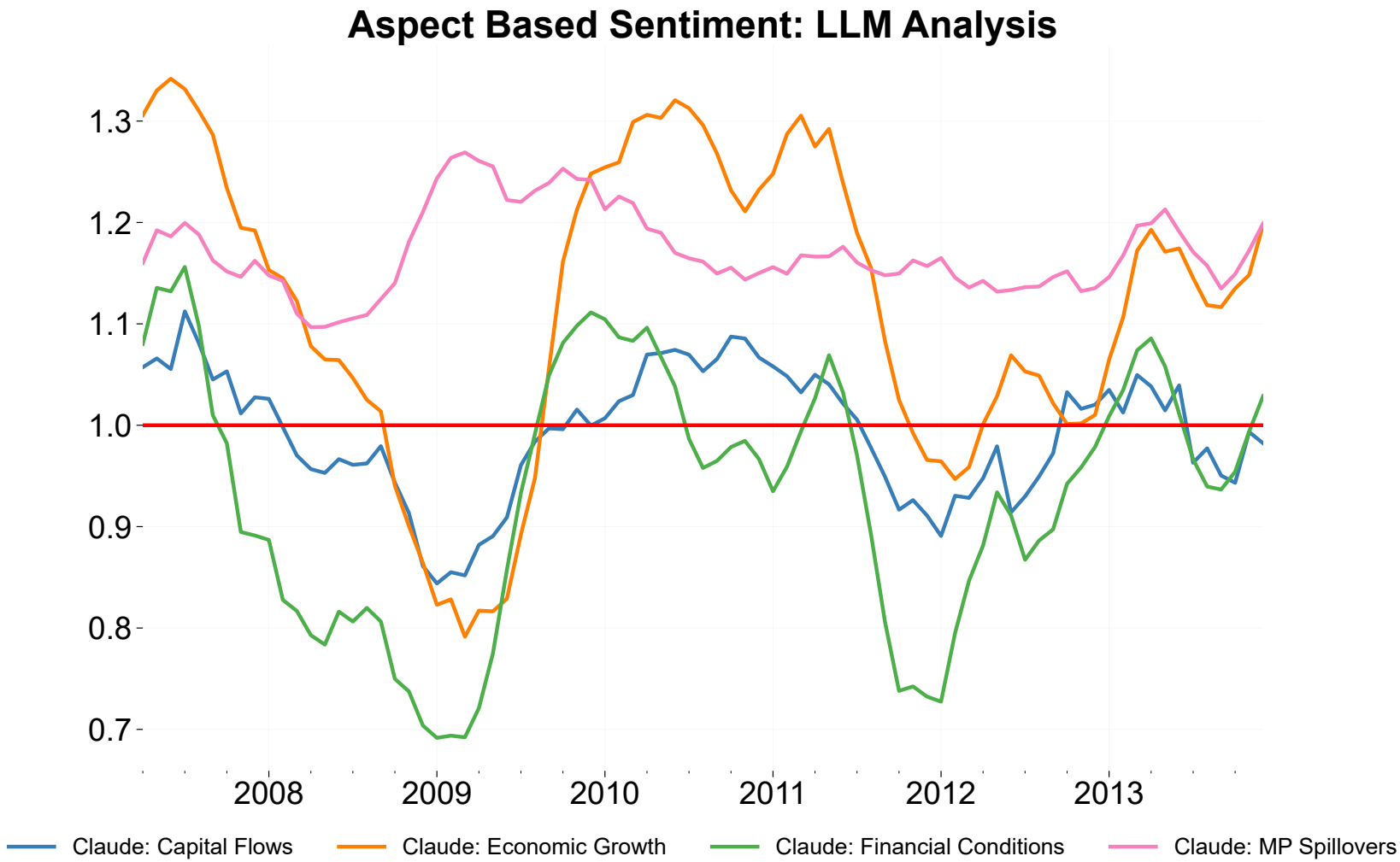


Figure 10: LLM can easily produce aspect based sentiment

Sentiment Validation: LLM Complements and Refines Lexical Methods

- LLM-based growth sentiment: correlation 0.52 with GFC, outperforming Correa (0.44) and Loughran-McDonald (0.46), comparable to Hubert (0.58)
- Cross-model LLM correlations of 0.90–0.93 for growth sentiment indicate robust consensus across architectures

Sentiment Validation: LLM Complements and Refines Lexical Methods

- LLMs complement established dictionaries by capturing structured, dimension-specific sentiment

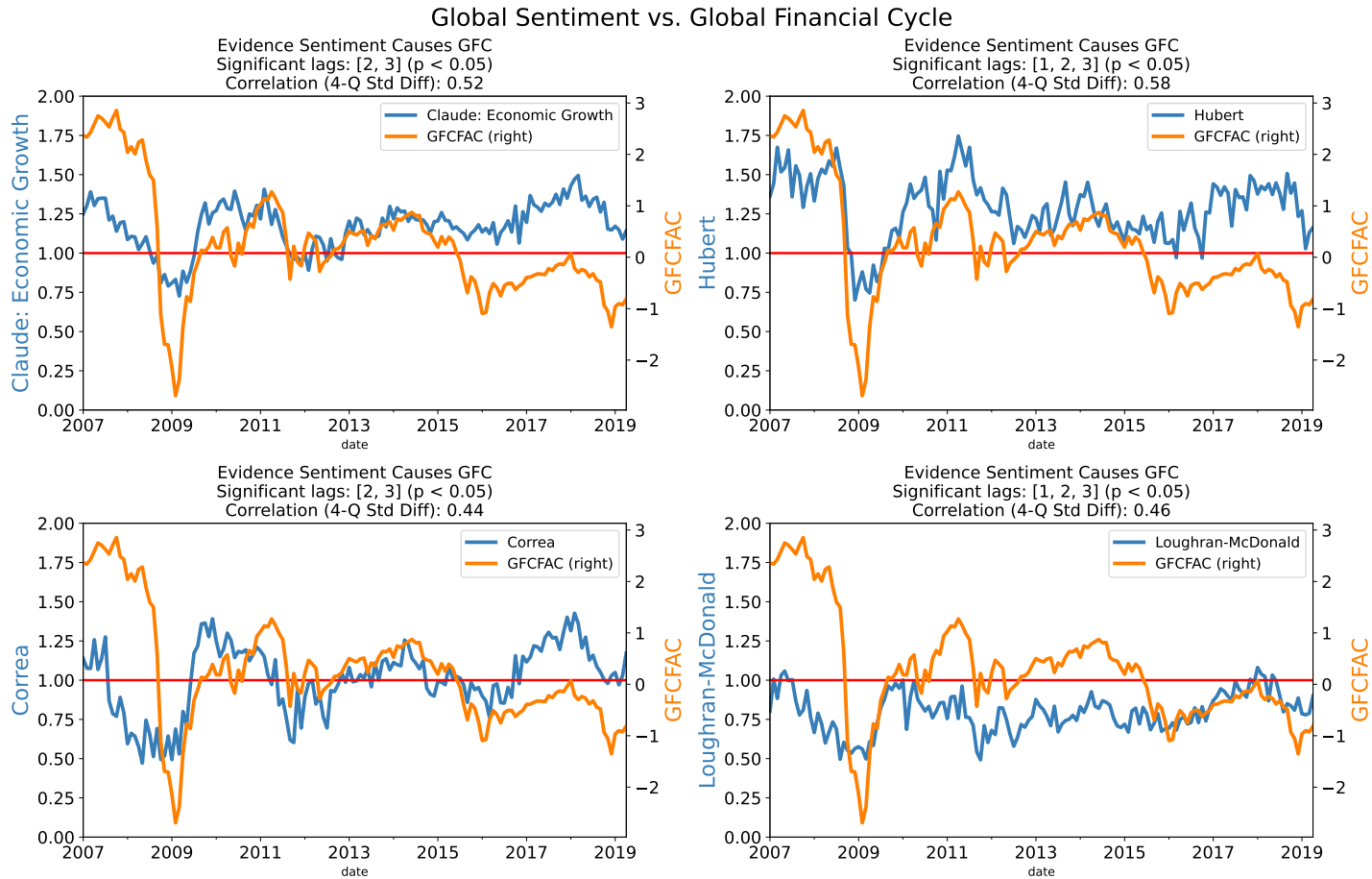
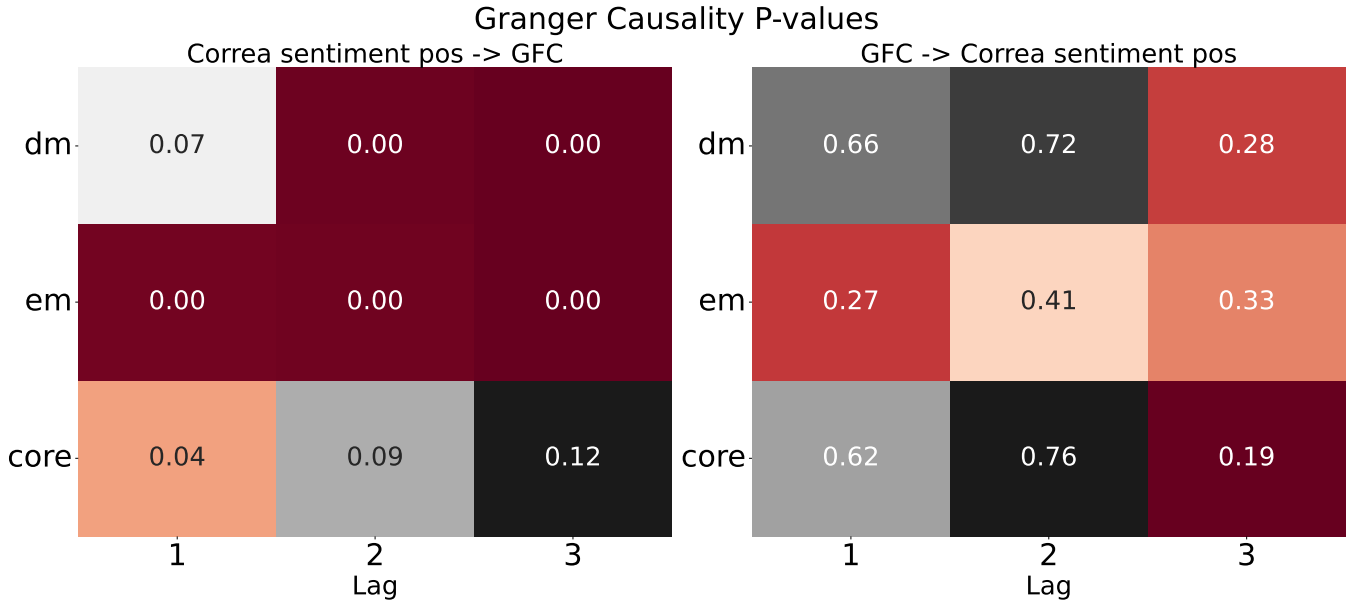


Figure 11: LLM-Generated Sentiment Indices vs. Lexical Approaches

Evidence that Sentiment Predicts the GFC

- Sentiment Granger-causes GFC ($p < 0.05$); GFC does *not* Granger-cause sentiment
- Granger causality captures **predictive precedence** rather than structural causation (Granger, 1969)
- Emerging-market CB sentiment appears to be a stronger predictor of the GFC than core CB sentiment
- Asymmetric pattern **consistent with** a communication transmission channel, though we cannot rule out omitted common factors or reverse causality at higher frequencies



Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Summary & Future Work

- We construct an open-source standardized database of CB communications designed as a benchmark for LLM tasks.
- We distribute all source code for our Data/LLM infrastructure to allow researchers to leverage AI tools in a scalable and reproducible manner.
- Using LLMs, we find that central banks often cite the U.S. economy broadly (rather than mention the U.S. Federal Reserve or monetary policy spillovers) as the primary driver of global conditions.
- Granger causality tests suggest that statement sentiment predicts the Global Financial Cycle rather than merely responding to it.
- We plan to scale our data to include more central banks and communication types.
- Our infrastructure includes major closed-source (Gemini, Claude, ChatGPT) and open-source (Hugging Face) LLM systems in an easy-to-use framework.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

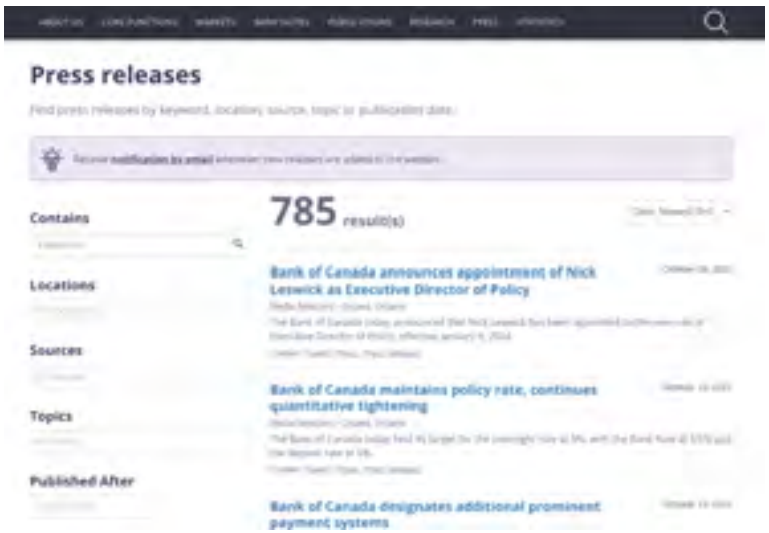
Summary & Future Work

Appendix

Monetary Policy Statements



(a) Brazil: Statements in press section



(b) Canada: Statements in press section

Figure 13: Examples of MPSs hosted on central bank websites

Key Words Time Series

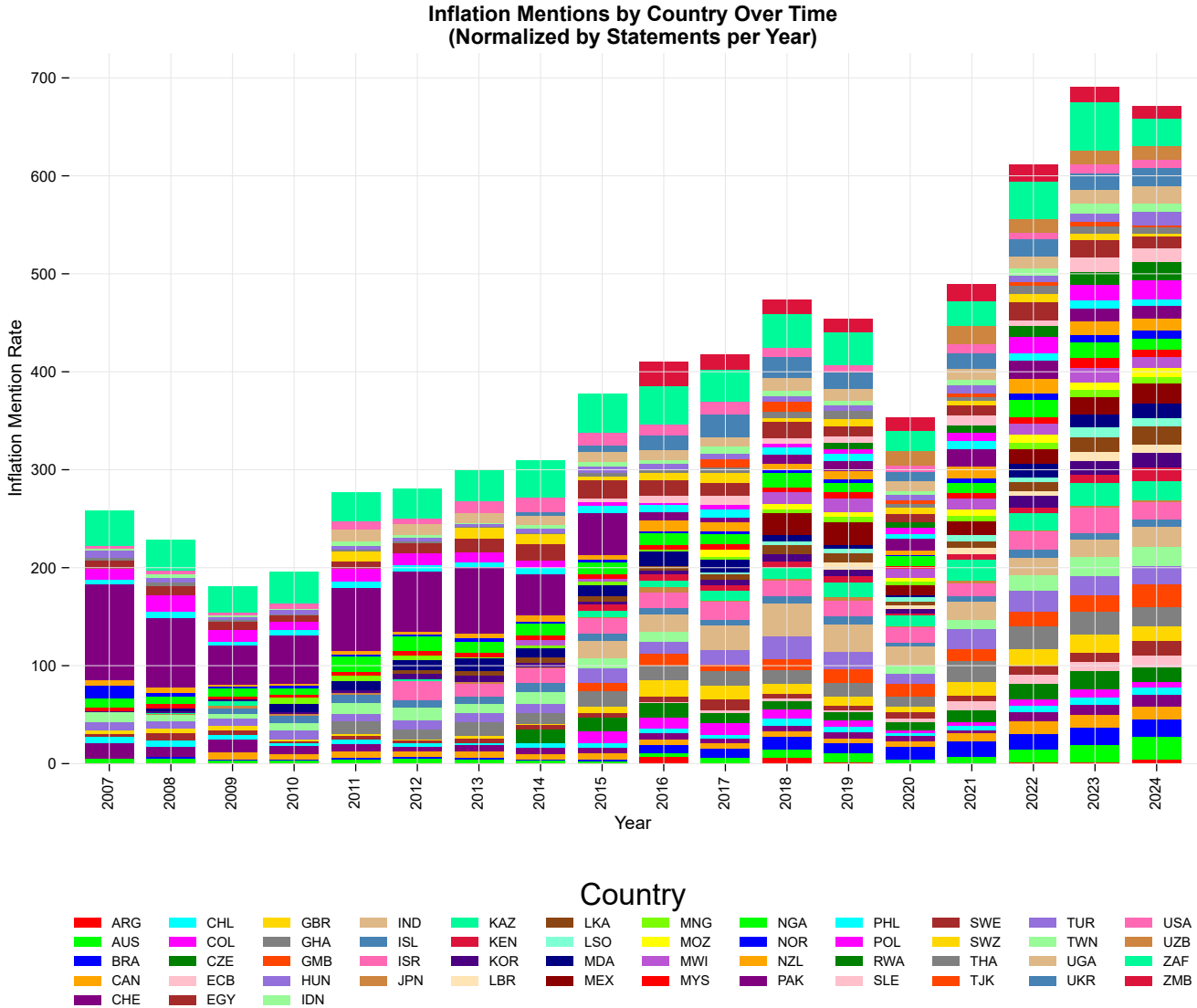


Figure 14: Inflation Mentions by Country

Regional Sentiment vs. Global Factor

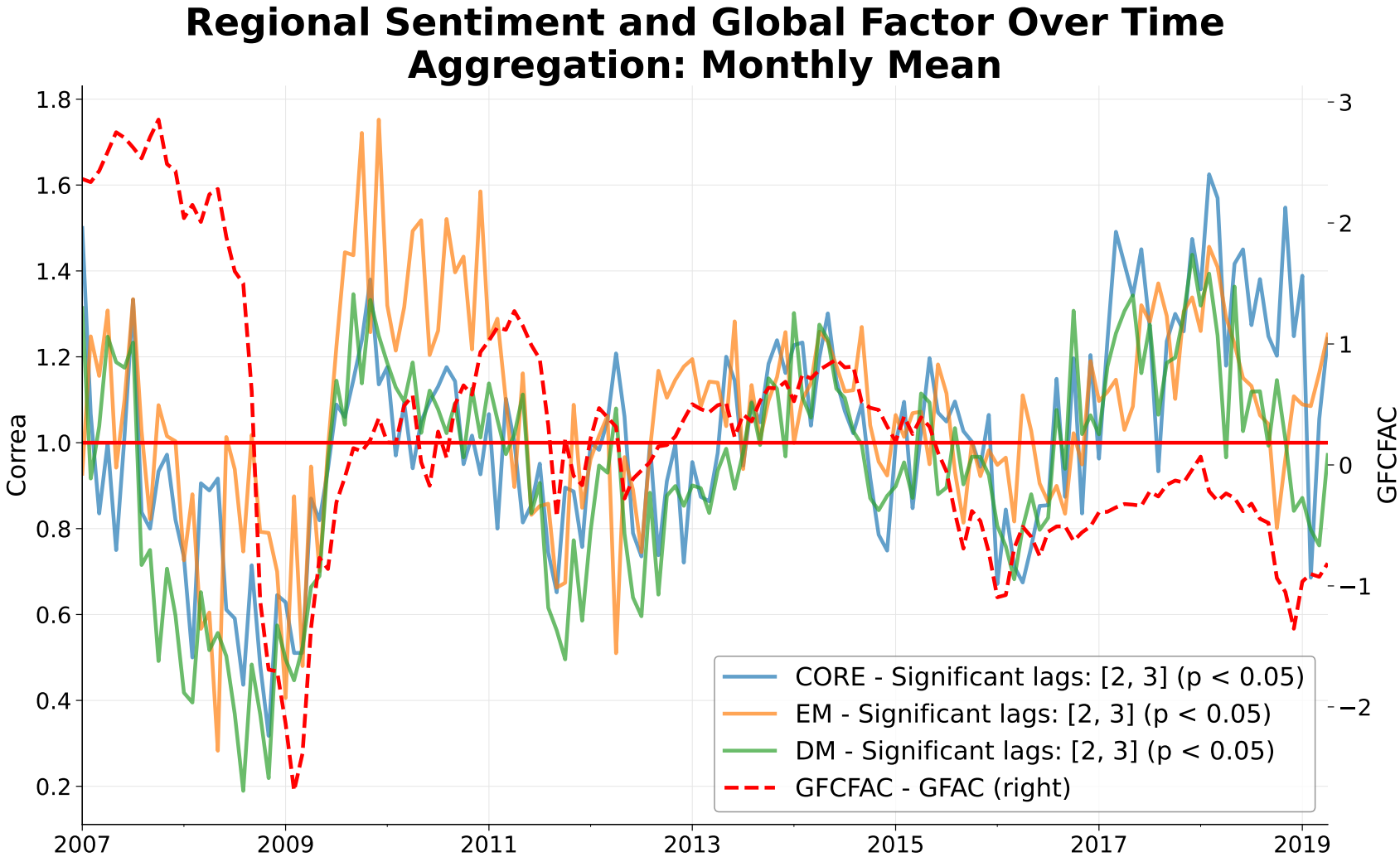


Figure 15: Regional Sentiment Evolution

References I

Born, B., Ehrmann, M., and Fratzscher, M. (2014). Central bank communication on financial stability. *Economic Journal*, 124(577):701–734.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Gonzalez, M. and Tadle, R. C. (2022). Monetary policy press releases: an international comparison. BIS Working Papers 1023, Bank for International Settlements.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(S1):114–133.

References II

Miranda-Agrippino, S. and Rey, H. (2020). U.S. monetary policy and the Global Financial Cycle. *Review of Economic Studies*, 87(6):2754–2776.

Miranda-Agrippino, S. and Rey, H. (2022). The global financial cycle. In *Handbook of International Economics*, volume 6, pages 1–43. Elsevier.

Picault, M. and Renault, T. (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance*, 79:136–156.

Rey, H. (2015). Dilemma not trilemma: the global financial cycle and monetary policy independence. NBER Working Papers 21162, National Bureau of Economic Research, Inc.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

The monetary policy statement database[1]

Cory Baird,
University of Tokyo

Jonathan Benchimol,
Bank of Israel

Vira Vyshnevska and Wook Sohn,
KDI School

Iegor Vyshnevskyi,
Sogang University

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

The Monetary Policy Statement Database: An LLM Application to Global Financial Conditions*

Cory Baird† Jonathan Benchimol‡ Wook Sohn§ Vira Vyshnevska§
Iegor Vyshnevskyi¶

March 11, 2026

Abstract

This study introduces the Monetary Policy Statement Database (MPSD), comprising 6,693 statements from 51 central banks worldwide (1990–2024). We develop a reproducible pipeline combining standard natural language preprocessing with large language model (LLM) tools for cross-country analysis. Four key findings emerge. First, statements lengthened substantially after the Global Financial Crisis while readability improved modestly. Second, inflation references comove across countries during global inflation episodes. Third, LLM-based question answering and aspect-based sentiment reveal that central banks attribute global financial conditions primarily to broad U.S. macroeconomic developments rather than to Federal Reserve policy actions specifically. Fourth, using a benchmark dictionary-based sentiment index and LLM-derived aspect-based sentiment indicators, Granger causality tests suggest that statement sentiment predicts the Global Financial Cycle rather than merely responding to it. The MPSD and accompanying codebase support reproducible research on monetary policy communication and international transmission.

Keywords: Central bank communication, Large language models, Text analysis, Generative database, Machine learning.

JEL Classification: C55, C63, E52, E58, G15.

*This paper does not necessarily reflect the views of the Bank of Israel or any other institution.

†University of Tokyo, Tokyo, Japan. Corresponding author. Email: cbaird@g.ecc.u-tokyo.ac.jp
‡Bank of Israel, Jerusalem, Israel.
§KDI School, Sejong, South Korea.
¶Sogang University, Seoul, South Korea.

1 Introduction

The evolution of monetary policy over recent decades has elevated central bank communication to a primary policy instrument, fundamentally transforming how monetary authorities influence economic outcomes (Woodford, 2005; Blinder, 2018). As former Federal Reserve Chairman Ben Bernanke observed, monetary policy has become "98 percent talk and 2 percent action," reflecting the critical role of communication in shaping market expectations and enhancing policy effectiveness (Blinder et al., 2008). This transformation has become particularly salient as central banks increasingly rely on forward guidance and other communication-based tools to influence both financial markets and household expectations (Hansen et al., 2019; Coibion et al., 2022).

The empirical analysis of central bank communications faces several critical constraints that limit our understanding of how monetary policy messages are crafted, transmitted, and interpreted across global markets. First, despite significant advances in communication strategies (de Haan and Hoogduin, 2024), the absence of a centralized, standardized database of monetary policy communications has restricted analysis to individual institutions (Born et al., 2014; Benchimol et al., 2025) and small groups of advanced or emerging economies (Cieslak and Schrimpf, 2019; Vyshnevskyi et al., 2024). Current research relies primarily on the Bank for International Settlements' repository of speeches, leaving a significant gap in the systematic analysis of monetary policy statements (Picault and Renault, 2017). Second, the lack of robust, open-source preprocessing pipelines has hindered the reproducibility of text-based monetary policy research, a methodological concern highlighted by Gentzkow et al. (2019) and Hansen and McMahon (2016). Third, while natural language processing (NLP) techniques have advanced rapidly, their application to monetary policy analysis has been constrained by insufficient computational infrastructure and domain-specific training data.

Recent studies confirm the complex and crucial influence of central bank communication. Hansen and McMahon (2016) demonstrate how linguistic analysis effectively reveals policy intentions and economic forecasts. Kryvtsov and Petersen (2021) provide experimental evidence on the effectiveness of different communication strategies, while Hayo and Neuenkirch (2015) document communication's crucial role during financial crises. These studies, alongside research on information effects by Nakamura and Steinsson (2018) and Bauer and Swanson (2023), highlight both the potential and challenges of systematically analyzing monetary policy communications.

This paper addresses these challenges by introducing the Monetary Policy Statement Database (MPSD), a comprehensive infrastructure for analyzing monetary policy statements across 51 central banks. Our contribution is fourfold.

First, we develop and publicly release a standardized database of monetary policy statements enabling rigorous cross-country analysis of central bank

communications. This resource empowers researchers to investigate policy transmission and coordination across diverse economies, including those beyond the traditional focus on major central banks (Armelius et al., 2020).

Second, we implement a sophisticated technical architecture that combines advanced NLP capabilities with robust code and data management tools. The infrastructure leverages data pipeline management (DVC), version control (Git), and experiment tracking while integrating cloud computing solutions through Amazon S3 buckets for storage and Hugging Face for open-source language model integration. This framework supports automated text summarization and question answering (QA) systems specifically designed for central bank applications, enabling efficient analysis of large-scale monetary policy communications. Following best practices in computational reproducibility (Gentzkow et al., 2019), our implementation uses object-oriented programming principles, with modular Python classes providing standardized tools for text preprocessing and advanced NLP analytics.

Third, we offer a practical framework enabling users to leverage large language models (LLMs), including Claude, Gemini, and ChatGPT. Building on methodological advances in text analysis, we develop tools for extracting policy-relevant information from complex monetary policy documents (Gonzalez and Tadle, 2022).

Fourth, we provide novel empirical evidence on the communication channel of international policy transmission. Our analysis yields four key findings with implications for understanding global monetary policy coordination: (i) Central banks worldwide exhibit remarkable narrative convergence during major economic shocks, with communications showing strong thematic comovement that precedes—rather than follows—observable financial market stress. (ii) When attributing global financial conditions to external factors, central banks systematically cite broad U.S. economic developments (housing markets, consumer spending, growth prospects) rather than Federal Reserve policy actions specifically. This distinction has important implications for modeling policy spillovers. (iii) Dictionary-based communication sentiment and LLM-derived sentiment indicators Granger-cause movements in the Global Financial Cycle (GFC), suggesting that central bank narratives actively shape global financial conditions rather than merely describing them. (iv) Substantial heterogeneity exists in how advanced-economy versus emerging-market central banks discuss global conditions, with emerging markets showing more idiosyncratic patterns, particularly at times of financial stress. These findings suggest that communication may be an underappreciated channel through which monetary policy transmits across borders.

Table 1 illustrates our approach to ensuring end-to-end transparency in artificial intelligence (AI) applications for monetary policy analysis. The pipeline shows how we maintain scientific rigor from raw data collection through preprocessing, model inference, and analysis—addressing concerns about

reproducibility in computational economic research, particularly when applying advanced language models to policy analysis.

Table 1: End-to-End Transparency for Replicable AI

Data Collection	**Featurization**	**Replicable AI Use**
Identify statement links	Cleaning	Replicable LLM infrastructure
Scrape statements	Sentence segmentation	Track inputs (prompts, texts)
Automate retrieval	Part-of-speech tagging	Track outputs (responses)

Source: Authors.

We leverage our data and AI infrastructure to investigate several key questions regarding monetary policy, the Global Financial Cycle (Miranda-Agrippino and Rey, 2020; Rey, 2015), and the transmission of policy narratives across borders (Armelius et al., 2020). We employ LLMs for QA and aspect-based sentiment (ABS) analysis. In QA, we prompt the model to summarize the drivers of the global economy, particularly focusing on instances where central banks identify the Federal Reserve as a key influence. For ABS, we use LLMs to extract sentiment for specific themes related to the impact of global financial conditions on domestic economies, such as economic growth, financial conditions, capital flows, and monetary policy spillovers.

Our findings from this LLM analysis reveal that central banks often cite the U.S. economy as a driver of global economic conditions rather than specifically attributing influence to the Federal Reserve. This nuanced distinction—where broader economic factors in the United States are perceived as more impactful than central bank actions alone—has important implications for modeling policy spillovers. Furthermore, our sentiment analysis shows that central banks expressed greater concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period, likely reflecting the lasting impact of the Global Financial Crisis and increased global economic interconnectedness.

While recent discussions of AI in economics have focused primarily on closed-source foundation models, we emphasize the fundamental importance of robust, open-source data science infrastructure. The MPSD provides comprehensive tools for text analysis specifically designed for central bank communications, enabling researchers to extract and analyze policy narratives systematically across institutions and time. This approach builds on seminal work in computational text analysis (Gentzkow et al., 2019) while addressing specific challenges in monetary policy research (Hansen and McMahon, 2016).

Our work connects to the growing literature on international monetary policy spillovers. Obstfeld (2015) highlights the challenges that global financial integration poses for monetary policy independence, Bruno and Shin (2015) demonstrate significant effects of U.S. monetary policy on capital flows to

emerging markets, and Kalemli-Özcan (2019) show that these spillovers primarily operate through global banks and risk premia in international asset markets. We extend this literature by focusing specifically on the communication channel of policy spillovers, providing direct evidence that central bank narratives—beyond policy actions themselves—significantly influence global financial conditions. Our findings complement Brusa et al. (2020), who document that ECB communications affect U.S. asset prices through signaling about global economic conditions.

Our work contributes to multiple strands of literature. First, we extend research applying NLP to central bank communications (Gonzalez and Tadle, 2022), moving beyond specific aspects like sentiment (Loughran and McDonald, 2011; Correa et al., 2021) or clarity (Bulír et al., 2013; Benchimol et al., 2023) to provide a comprehensive analytical framework. Second, we advance the literature on monetary policy transmission by enabling systematic cross-country analysis of communication strategies and their effects (Hansen et al., 2019; Picault and Renault, 2017). Third, we contribute to the literature on global financial cycles and policy spillovers (Miranda-Agrippino and Rey, 2020; Armelius et al., 2020). Fourth, we extend the methodological literature on text analysis in economics (Gentzkow et al., 2019; Benchimol et al., 2022) by developing reproducible procedures specifically for central bank communications. This project also complements recent large-scale data collection efforts in central bank communication research. In particular, Campiglio et al. (2025) introduce the cbspeeches.com corpus comprising over 35,000 speeches from 131 central banks (1986–2023). While their corpus focuses on speeches and covers a broader institutional range, our database concentrates specifically on monetary policy statements—the official communications released in direct connection with policy rate decisions. The two resources are therefore complementary and together support a comprehensive ecosystem for reproducible research on central bank communication.

Our approach also connects to recent surveys on central bank communication. Masciandaro et al. (2024) provide a comprehensive review of how central banks have evolved from silence to active social media engagement, documenting techniques used to analyze communication effectiveness. Our infrastructure complements their focus on social media by providing systematic analysis of formal policy statements. Additionally, our work relates to the broader literature on monetary policy transmission mechanisms. De Grauwe and Ji (2023) demonstrate that using forward-looking data in monetary policy affects policymaking quality differently across policy regimes, while Kumar et al. (2023) show that market volatility significantly influences the term premium through channels interacting with monetary policy. These findings underscore the importance of understanding how central bank communications shape expectations and market responses—precisely the mechanisms our database enables researchers to investigate.

The remainder of the paper is organized as follows. Section 2 details the

technical architecture and construction of the MPSD, including data collection procedures, preprocessing pipelines, and the LLM inference framework. Section 3 presents an exploratory analysis of the database, highlighting key patterns in central bank communications across time and institutions. Section 4 examines our application of generative AI to central bank communications, focusing on extracting economic narratives related to the GFC. Section 5 discusses policy implications and potential extensions, and Section 6 concludes.

2 Database Construction and Architecture

The MPSD constitutes a comprehensive research infrastructure designed to facilitate systematic analysis of monetary policy communications across global central banks. The database encompasses monetary policy statements from 51 central banks, with coverage varying by institution. Statement frequency ranges from monthly to quarterly depending on each central bank's policy framework, yielding 6,693 unique policy statements.

We define a monetary policy statement observation as a text communication released in direct connection with a policy rate decision event (whether a change or no change in rates). Because central banks employ varying naming conventions—including "decisions," "statements," "announcements," and "press releases"—we systematically map each institution's official document series to the corresponding decision dates. This mapping is fully documented in the accompanying metadata file. For example, the ECB's communication evolved from the "Introductory Statement" (1999–mid-2021) to the "Monetary Policy Statement" (since mid-2021), while short "Monetary Policy Decisions" press releases—which historically contained only rate decisions without supporting rationale—are not treated as equivalent to full policy statements in our corpus. Similarly, for the Reserve Bank of Australia, we include the post-meeting "Monetary Policy Decision" statements rather than the separate quarterly "Statement on Monetary Policy" inflation report. The corpus excludes ancillary documents such as inflation reports, minutes, and speeches, which serve different communicative functions. All documents were preprocessed to remove non-textual elements including tables, charts, and lists of meeting attendees, ensuring that linguistic analysis captures substantive policy communication content.

Monetary policy communications take different formal labels across central banks, including "decisions," "statements," and "press releases," as documented in Table 4. While these document types can be identified, they are not treated as separate analytical categories in this study. Despite differences in naming conventions and format, these communications are released in direct connection with policy rate decision events (change or no change) and serve a functionally comparable role: conveying the policy outcome together with varying degrees of economic assessment and forward-looking guidance. In many cases, a single

document simultaneously fulfills multiple functions, making a clean and objective partition by document type difficult. To preserve transparency and cross-country comparability, we therefore treat these texts as institution-specific realizations of a common policy-decision communication instrument, while explicitly accounting for observable sources of possible heterogeneity such as language of publication.

The technical architecture implements three fundamental design principles that address specific challenges in monetary policy research. The infrastructure enables scalable data collection and standardization through advanced web scraping and preprocessing pipelines, essential for maintaining consistency across the diverse spectrum of central bank websites and communication formats (PDF, DOC/DOCX, HTML). We also maintain data integrity through version control, addressing reproducibility challenges identified in computational social science research (Gentzkow et al., 2019).

2.1 Technical Infrastructure Implementation

The MPSD ensures integrity of economic research data by strictly enforcing lineage. Using Data Version Control (DVC), the system tracks the full data lifecycle—from web scraping to LLM-based analysis—via parameterized pipelines. To support full computational reproducibility, the architecture couples Git for code versioning with uv for unified Python version and dependency management (ensuring that data manipulation tools such as pandas work consistently across different environments).

Table 2 illustrates the LLM infrastructure we developed specifically for central bank communication analysis. This diagram shows the integration of multiple components (modules), including text retrieval, preprocessing, and LLM integration. The structure enables seamless incorporation of different AI models (such as ChatGPT and Gemini, as well as open-source models hosted on Hugging Face) and supports various analytical tasks including classification, summarization, and QA. This modular design ensures reproducibility while maintaining flexibility for users to adapt the infrastructure to their specific use cases.

Data storage relies on Amazon S3 buckets managed directly through DVC. The system segregates assets into distinct partitions—raw data, processed outputs, and model artifacts—ensuring clear data lineage and secure access control.

2.2 Code and Folder Infrastructure

The MPSD employs a modular design with discrete functional components for text retrieval, text cleaning, and LLM inference. As illustrated in Figure 1, the codebase follows a standardized folder structure based on the "cookiecutter" template for data science projects. This approach organizes the project into logical sections to facilitate maintenance and collaboration:

Table 2: AI Infrastructure for Central Bank Communication Analysis

Modular LLM-based Data Processing Framework for Academic Research		
Data Acquisition Module (Abstract Class)	**LLM Processing Module (Abstract Class)**	**LLM Data Pipeline Workflow**
Methods: 1. Initialize (source parameters) 2. Scrape Statement Links 3. Download/Save Statements	**Methods:** 1. Initialize (LLM Model) 2. Process/Load Prompts 3. Generate Response	**Methods:** 1. Initialize (Input Corpus) 2. Filter via Keywords or POS 3. LLM Inference 4. Save Response
Output: Text Files CSV/Parquet Text Dataframes	**Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML)	**Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) Figure for Research (PNG/PDF)
Abstraction: Extend abstract functions to accommodate new data sources (e.g., central bank speeches, minutes) and communication modes.	**Abstraction:** Implement model-agnostic functions to leverage diverse LLM architectures (e.g., Gemini, Claude, HuggingFace models).	**Use-Case:** Utilize concise, reusable scripts for seamless integration with LLM APIs (e.g., Gemini, Claude) by researchers.

Source: Authors.

- `src`: Contains Python scripts organized by functionality
- `data`: Stores raw and transformed text files and processed outputs
- `references`: Houses GenAI prompts and documentation

Complementing this structure, the system implements automated database updates through DVC pipelines. This enables researchers to access data and replicate the entire analysis—including the graphs and charts utilized in this paper—with a single command. This structure ensures end-to-end reproducibility from data manipulation through NLP analysis, addressing a critical need in text analysis.

2.3 Data Collection and Processing Framework

Our database houses two data types: statements retrieved programmatically via Python, stored in the 'API' folder, and statements downloaded manually, termed 'static downloads' (Vyshnevskyi et al., 2024). As illustrated in Table 2, the process for programmatically sourced data uses an agnostic class with three abstract methods: (1) downloading statement links, (2) iterating through these links to download each statement, and (3) parsing each statement and saving the output as Parquet and CSV files. The pipeline's final stage applies format-specific processing for PDF, Word, and HTML documents.



Figure 1: Folder Structure Based on "Cookiecutter" Template for Data Science Projects

Source: Authors.

The Python code employs a modular structure, maintaining an agnostic design while permitting targeted modifications. Base classes define abstract interfaces (including the three methods previously mentioned), complemented by specific

implementation classes tailored to individual central bank websites to handle their variations. We have decided not to release the scraping code due to concerns about potential misuse that could overload central bank web servers. This approach is consistent with established practices in empirical economics, where reproducibility is defined by the ability to replicate published results using shared data and code rather than by replicating the original data collection process itself. We hope our project highlights best practices for similar text database initiatives, demonstrating how abstract and agnostic designs can help optimally organize scraping code.

Continuous integration through GitHub Actions implements sophisticated workflows for database maintenance and updates. These workflows allow us to perform frequent checks for new policy statements, enabling us to cross-check any findings against the links stored in our system.

2.4 LLM Inference Framework

The MPSD incorporates a framework for leveraging language models in central bank applications. We created an abstract LLM class that seamlessly integrates Claude, ChatGPT, Gemini, and open-source models hosted on Hugging Face. This class provides methods to select models, process prompts from YAML files, and generate and save responses, enabling researchers to test various prompts and assess the robustness of their results across multiple models.

2.5 Future Development and Maintenance

The MPSD features a forward-looking architecture designed for scalability to incorporate other types of central bank communication, such as speeches, summaries, and minutes. The modular design also enables easy addition of new central banks. Furthermore, integration with GitHub and DVC enables automated workflows that ensure the database is regularly updated while preserving versioned, point-in-time snapshots of the data.

3 Database Content

3.1 Interest Rate Announcements

Over the last century, central bank communication has evolved from indirect and cryptic messaging to becoming an independent and significant monetary policy tool in both developed and developing economies (Blinder et al., 2024). This communication involves disseminating information about monetary policy objectives, economic outlook, and future policy trajectories (Blinder et al., 2008). Assenmacher et al. (2021) defines monetary policy communication as a central bank's transmission of its objectives, plans, instruments, policy decisions, and economic assessments.

Monetary policy statements—official communications explaining policy decisions and their underlying rationales (Dincer et al., 2022)—represent a crucial component of this communication framework. Their primary objectives include managing inflation expectations, enhancing policy predictability, ensuring transparency, and building institutional credibility with markets, governments, and the public (Kahn, 2007; Blinder et al., 2008; Issing, 2019; Lewis et al., 2020).

Central banks employ various nomenclature for these statements: the Reserve Bank of Australia (RBA) issues a post-meeting "Monetary Policy Decision," the Bank of Japan (BoJ) releases a "Statement on Monetary Policy," the Federal Reserve's primary monetary policymaking body (Federal Open Market Committee) produces a "FOMC Statement," and the European Central Bank (ECB) publishes an "Introductory Statement" (until mid-2021) or "Monetary Policy Statement" (since mid-2021)—documents that provide the detailed rationale accompanying each policy decision. This practice is relatively recent, with the RBA releasing one of the first official policy statements on January 23, 1990, following New Zealand's pioneering efforts in transparent communication during the mid-1990s (Blinder et al., 2001).

While the Australian media reported on the RBA's January 1990 monetary policy change, a notable article in a business-focused newspaper, potentially the first centered on this type of monetary policy statement, did not explicitly describe the statement itself as a new communication format.[1] Searches in other publications yielded similar findings.

The statement's novelty as a communication tool may have been overlooked because senior RBA officials were already actively communicating with markets and the public through other channels, such as speeches, prior to the January 23, 1990, statement, as indicated by news searches from that period.[2] This practice of announcing policy changes began formally in January 1990, marking a shift from previous operational methods where market participants inferred policy changes from the RBA's market actions.

The U.S. Federal Reserve issued its first press release on February 4, 1994, initially publishing statements only when modifying the federal funds rate. By May 1999, the FOMC began releasing statements after every meeting, and by May 2002, included voting details and dissenting vote explanations (Acosta and Meade, 2015). Other central banks followed similar patterns: the Reserve Bank of New Zealand in 1996, the BoJ in 1998, the Swiss National Bank in 2000, and the ECB in 2001. As Figure 2 illustrates, statement issuance accelerated significantly following the 2008 financial crisis.

While formatting varies across institutions, monetary policy statements typically include inflation and growth forecasts, economic assessments, and justifications for policy adjustments (Blinder, 2009). Federal Reserve statements

[1] Tim Dodd. (January 24, 1990 Wednesday). Timing Surprises Markets. Australian Financial Review.

[2] Tony Kaye. (April 12, 1989 Wednesday). Timing 'Crucial' in Monetary Policy.

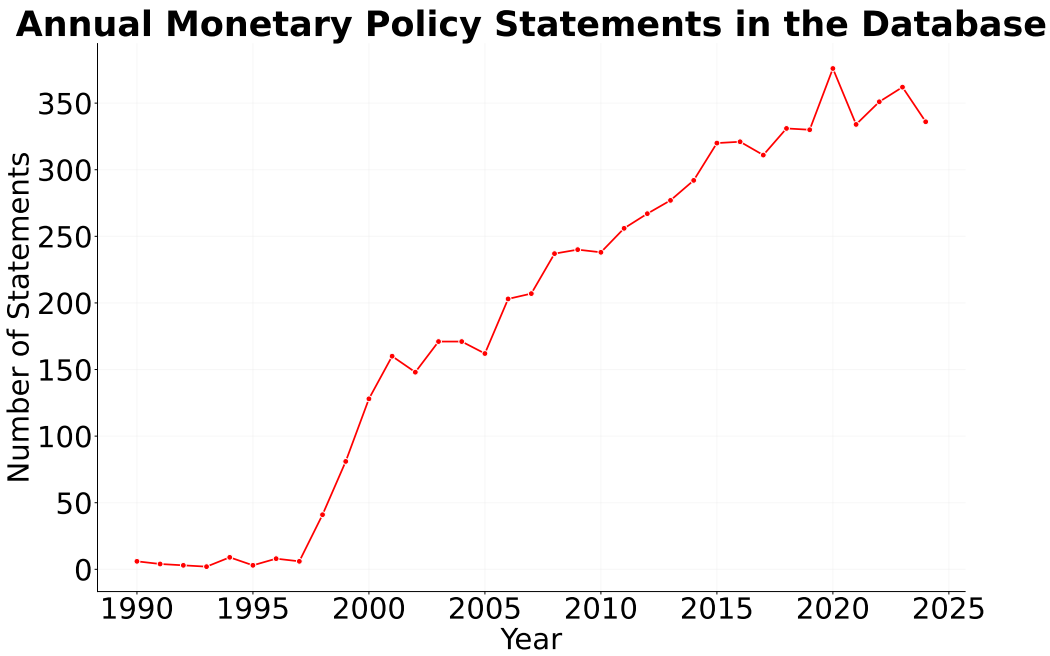


Figure 2: Annual Monetary Policy Statements in the Database

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

generally assess domestic economic conditions, reference the dual mandate, announce policy decisions, provide forward guidance, and report voting outcomes (Acosta and Meade, 2015). The ECB structures its statements around economic and monetary analyses, policy decisions, rationales, and forward guidance (Assenmacher et al., 2021). The Bank of Korea employs a six-paragraph framework covering policy decisions, global and domestic economic conditions, inflation trends, financial market developments, and future policy direction. These statements typically become more detailed during economic crises, such as the 2008 financial crisis or the COVID-19 pandemic (Macklem and Vardy, 2023).

Statement style has evolved toward greater accessibility and clarity (Blinder et al., 2024), increasingly incorporating visual elements to communicate complex concepts (Assenmacher et al., 2021). Most central banks have converged on approximately eight annual policy meetings, balancing the need for policy flexibility with regular updates. Notable exceptions include the Swiss National Bank (four annually) and the RBA (eleven annually, reduced to eight in 2024). As meeting frequency has declined at some institutions, statement length and detail have increased, particularly for the FOMC (Davis and Wynne, 2019), reflecting the rise of non-standard monetary policy tools such as forward guidance and asset purchases.

The monetary theory consensus now emphasizes open communication as essential for effective policy transmission (Eijffinger and Masciandaro, 2014). Research demonstrates communication's importance in improving monetary transmission and maintaining trust (Ehrmann and Wabitsch, 2022). Global

monetary policy has become increasingly transparent, with central banks—including those in emerging economies—publishing policy statements to enhance market stability and policy effectiveness (Dincer et al., 2022). Despite this expanded outreach, audience engagement remains uneven: firms and consumers demonstrate limited awareness, while financial market participants respond significantly to communications, including policy statements (Blinder et al., 2024). Semantic analyses of ECB, Federal Reserve, and other central bank statements reveal a trend toward increased precision and transparency following the Global Financial Crisis (Benchimol et al., 2020).

3.2 Database Overview

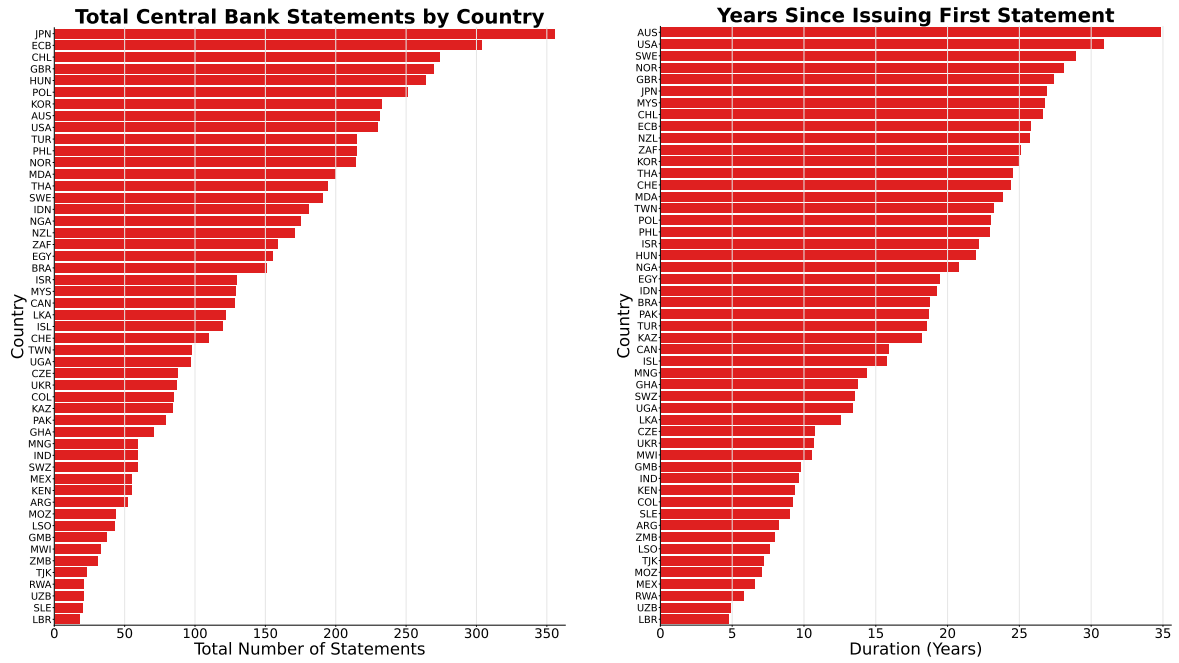
This section examines the evolution and characteristics of monetary policy statements (MPS). We define a *monetary policy statement observation* as a written statement released by a central bank on date t, associated with a policy-rate decision event (change or no change). To account for variations in naming conventions across central banks, we map each bank's official document series—released on or explicitly linked to a decision date—to the corresponding decision event t and document this mapping. We begin with an overview of our database, which encompasses approximately 6,693 unique monetary policy statements from 51 central banks globally, spanning the period from 1990 to 2024. As depicted in Figure 8, this standardized and diverse corpus provides an extensive dataset ideal for training and benchmarking AI systems.

A descriptive analysis of our dataset, as summarized in Tables 3 and 4, reveals substantial heterogeneity in central bank communication patterns across institutions.[3] The Bank of Japan (BoJ) and European Central Bank (ECB) emerge as prolific publishers of statements, while central banks in middle-income countries exhibit more limited publication frequency and shorter historical coverage. These cross-institutional variations reflect differences in institutional frameworks and communication strategies, as illustrated in Figure 3. The statistics on coverage periods, statement frequency, linguistic characteristics, and readability metrics demonstrate systematic differences in communication style and complexity, highlighting the evolution of central bank transparency practices.

Figure 3 provides comparative evidence of central bank communication practices across countries, with Panel (a) documenting the total number of monetary policy statements per country and Panel (b) illustrating the duration of coverage in years. The data reveal considerable heterogeneity in both dimensions. For instance, while the BoJ leads in total statement volume, it does not exhibit the longest historical coverage. This institutional variation reflects differences in

[3]Appendix A.4 documents that unusually long statements in the corpus are concentrated in countries where the policy decision is communicated through a broader analytical document rather than a concise press release, reflecting institutional differences in communication formats and document bundling.

transparency regimes, governance structures, and the evolution of communication strategies among monetary authorities.



(a) Statements per Country

(b) Coverage Period (Years)

Figure 3: Statement Volume and Coverage by Country

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

The map in Figure 4 illustrates the global reach of the MPSD, encompassing central banks from advanced economies, emerging markets, and developing countries across different continents. This geographical diversity is crucial for studying how monetary policy communication varies across different economic contexts and institutional settings, enabling more robust cross-country analyses than previously possible with existing datasets.

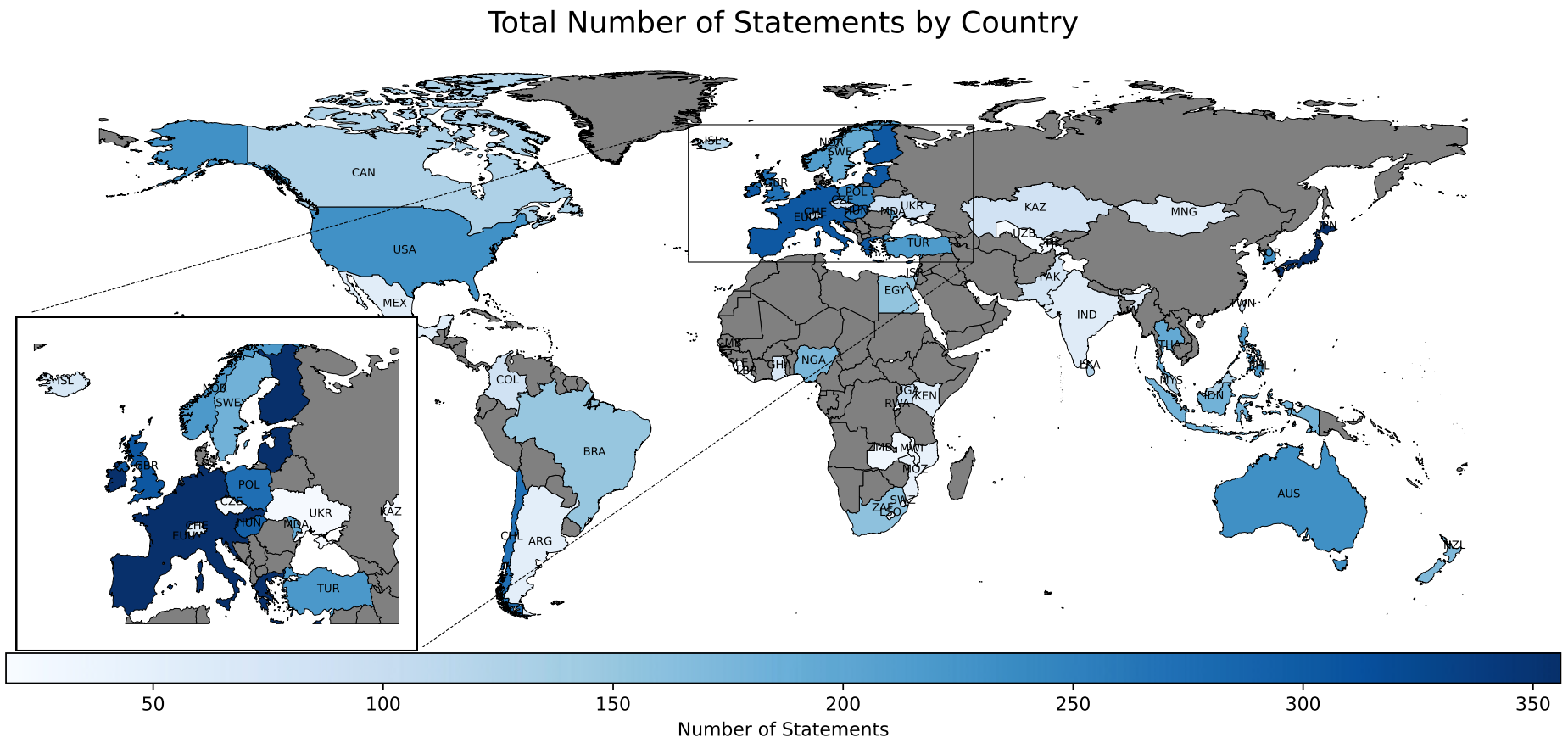


Figure 4: Number of Monetary Policy Statements by Country in the MPSD

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Table 3: Central Bank Monetary Policy Statements: Coverage and Communication Frequency

Country	Region	Income Group	Number of Statements	Avg. Statements per Year	Modal Statements per Year	Statements in 2023	Database
Japan (JPN)	East Asia & Pacific	High income	356	13	8	8	Programmatic
Euro area (EUU)	Europe & Central Asia	High income	304	11	12	8	Programmatic
Chile (CHL)	Latin America & Caribbean	High income	274	10	12	8	Programmatic
United Kingdom (GBR)	Europe & Central Asia	High income	271	9	12	8	Programmatic
Hungary (HUN)	Europe & Central Asia	High income	264	11	12	12	Programmatic
Poland (POL)	Europe & Central Asia	High income	251	10	11	11	Programmatic
South Korea (KOR)	East Asia & Pacific	High income	233	9	12	8	Programmatic
Australia (AUS)	East Asia & Pacific	High income	231	6	11	11	Programmatic
United States (USA)	North America	High income	230	7	8	8	Programmatic
Turkey (TUR)	Europe & Central Asia	Upper middle income	215	11	12	12	Programmatic
Philippines (PHL)	East Asia & Pacific	Lower middle income	215	9	8	9	Programmatic
Norway (NOR)	Europe & Central Asia	High income	214	7	8	8	Programmatic
Moldova (MDA)	Europe & Central Asia	Upper middle income	200	8	12	9	Programmatic
Thailand (THA)	East Asia & Pacific	Upper middle income	194	7	8	6	Programmatic
Sweden (SWE)	Europe & Central Asia	High income	191	6	6	5	Programmatic
Indonesia (IDN)	East Asia & Pacific	Upper middle income	181	9	12	12	Programmatic
Nigeria (NGA)	Sub-Saharan Africa	Lower middle income	175	8	11	9	Programmatic
New Zealand (NZL)	East Asia & Pacific	High income	171	6	8	7	Programmatic
South Africa (ZAF)	Sub-Saharan Africa	Upper middle income	159	6	6	6	Programmatic
Egypt (EGY)	Middle East & North Africa	Lower middle income	155	7	8	8	Programmatic
Brazil (BRA)	Latin America & Caribbean	Upper middle income	151	7	8	8	Programmatic
Israel (ISR)	Middle East & North Africa	High income	130	5	8	8	Programmatic
Malaysia (MYS)	East Asia & Pacific	Upper middle income	129	4	6	6	Programmatic
Canada (CAN)	North America	High income	128	8	8	8	Programmatic
Sri Lanka (LKA)	South Asia	Lower middle income	122	9	8	8	Manual

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD). Region and Income Group data are from the World Bank country classifications by income level for 2022–2023.

Notes: This table presents statistics on monetary policy statements for the 25 central banks with more than 120 statements in our database. Coverage period spans from January 1990 through December 2024. "Programmatic" indicates that data are collected via a Python script through an automated interface that updates regularly, while "Manual" indicates manually collected data. Modal statements per year represent the most frequently occurring number of statements issued annually. Income group classifications follow World Bank definitions. Table 3 reports this information for the illustrative subsample discussed in the text, while the complete classification for all countries in the database is provided in the accompanying *metadata.csv* file in the public repository.

Table 4: Central Bank Monetary Policy Statements: Complete Database

Country	Code	Number of Statements	First Statement	Last Statement	Coverage (Years)	Database Type	Document Type	English Language	Central Bank	Region
Argentina	ARG	52	2016-08-09	2024-11-01	9	Programmatic	MPC	Translated	Central Bank of Argentina	Latin America & Caribbean
Australia	AUS	231	1990-01-23	2024-12-10	35	Programmatic	MPD / IRD	Official	Reserve Bank of Australia	East Asia & Pacific
Brazil	BRA	151	2006-03-08	2024-12-11	19	Programmatic	CS	Translated	Central Bank of Brazil	Latin America & Caribbean
Canada	CAN	128	2009-01-20	2024-12-11	16	Programmatic	MPD	Official	Bank of Canada	North America
Chile	CHL	274	1998-01-08	2024-09-03	27	Programmatic	PR of RPM	Translated	Central Bank of Chile	Latin America & Caribbean
Colombia	COL	85	2015-08-21	2024-10-31	10	Programmatic	PR	Translated	Central Bank of Colombia	Latin America & Caribbean
Czechia	CZE	88	2014-02-06	2024-11-07	11	Programmatic	BD	Translated	Czech National Bank	Europe & Central Asia
Egypt	EGY	155	2005-06-02	2024-11-21	20	Programmatic	MPCD	Translated	Central Bank of Egypt	Middle East & North Africa
Eswatini	SWZ	59	2011-05-13	2024-11-22	14	Manual	MPS	Official	Central Bank of Eswatini	Sub-Saharan Africa
Euro area	EUU	304	1999-03-04	2024-12-12	26	Programmatic	IS / MPS	Official	European Central Bank	Europe & Central Asia
Gambia	GMB	37	2015-02-24	2024-11-26	10	Manual	MPC PR	Official	Central Bank of The Gambia	Sub-Saharan Africa
Ghana	GHA	71	2011-02-18	2024-11-29	14	Manual	MPC PR	Official	Bank of Ghana	Sub-Saharan Africa
Hungary	HUN	264	2002-12-16	2024-12-17	23	Programmatic	PR	Translated	Magyar Nemzeti Bank	Europe & Central Asia
Iceland	ISL	120	2009-01-29	2024-11-20	16	Programmatic	Statement of MPC	Official	Central Bank of Iceland	Europe & Central Asia
India	IND	59	2015-04-07	2024-12-06	10	Programmatic	MPS	Official	Reserve Bank of India	South Asia
Indonesia	IDN	181	2005-08-16	2024-11-20	20	Programmatic	PR	Translated	Bank Indonesia	East Asia & Pacific
Israel	ISR	130	2002-09-23	2024-11-25	23	Programmatic	IRA	Translated	Bank of Israel	Middle East & North Africa
Japan	JPN	356	1998-01-16	2024-12-19	27	Programmatic	Statement on MP	Translated	Bank of Japan	East Asia & Pacific
Kazakhstan	KAZ	84	2006-09-05	2024-11-29	19	Programmatic	PR	Translated	National Bank of Kazakhstan	Europe & Central Asia
Kenya	KEN	55	2015-08-05	2024-12-05	10	Manual	PR	Official	Central Bank of Kenya	Sub-Saharan Africa
South Korea	KOR	233	1999-12-02	2024-11-28	26	Programmatic	MPD	Translated	Bank of Korea	East Asia & Pacific
Lesotho	LSO	43	2017-04-04	2024-11-26	8	Manual	Statement of MPC	Official	Central Bank of Lesotho	Sub-Saharan Africa
Liberia	LBR	18	2019-11-20	2024-08-27	6	Manual	Communique	Official	Central Bank of Liberia	Sub-Saharan Africa
Malawi	MWI	33	2014-04-30	2024-11-04	11	Manual	Statement of MPC	Official	Reserve Bank of Malawi	Sub-Saharan Africa
Malaysia	MYS	129	1998-02-06	2024-11-06	26	Programmatic	MPS	Translated	Central Bank of Malaysia	East Asia & Pacific
Mexico	MEX	55	2018-05-17	2024-12-19	7	Programmatic	MPS	Translated	Bank of Mexico	Latin America & Caribbean
Moldova	MDA	200	2001-01-25	2024-12-05	24	Programmatic	MPD	Translated	National Bank of Moldova	Europe & Central Asia
Mongolia	MNG	59	2010-07-23	2024-12-13	15	Manual	MPS	Translated	Bank of Mongolia	East Asia & Pacific
Mozambique	MOZ	44	2017-10-26	2024-11-27	8	Manual	Communique	Translated	Bank of Mozambique	Sub-Saharan Africa
New Zealand	NZL	171	1999-03-17	2024-11-27	26	Programmatic	MPD	Official	Reserve Bank of New Zealand	East Asia & Pacific
Nigeria	NGA	175	2004-02-18	2024-11-26	21	Programmatic	MP Communique	Official	Central Bank of Nigeria	Sub-Saharan Africa
Norway	NOR	214	1996-11-05	2024-12-19	29	Programmatic	Rate decision	Translated	Central Bank of Norway	Europe & Central Asia
Pakistan	PAK	79	2005-11-27	2024-07-29	20	Programmatic	MPS	Official	State Bank of Pakistan	South Asia
Philippines	PHL	215	2002-01-17	2024-12-19	23	Programmatic	MPD	Official	Central Bank of the Philippines	East Asia & Pacific
Poland	POL	251	2001-11-28	2024-12-04	24	Programmatic	MPCo PR	Translated	Narodowy Bank Polski	Europe & Central Asia
Rwanda	RWA	21	2019-02-07	2024-11-20	6	Manual	PR	Official	National Bank of Rwanda	Sub-Saharan Africa
Sierra Leone	SLE	20	2015-12-14	2024-12-23	10	Manual	MPS	Official	Bank of Sierra Leone	Sub-Saharan Africa
South Africa	ZAF	159	1999-10-13	2024-11-21	26	Programmatic	Statement of MPC	Official	South African Reserve Bank	Sub-Saharan Africa
Sri Lanka	LKA	122	2012-05-11	2024-11-27	13	Manual	MP Review	Translated	Central Bank of Sri Lanka	South Asia
Sweden	SWE	191	1996-01-09	2024-12-19	29	Programmatic	MPD	Translated	Sveriges Riksbank	Europe & Central Asia
Switzerland	CHE	110	2000-01-20	2024-06-20	25	Programmatic	MPA	Translated	Swiss National Bank	Europe & Central Asia
Taiwan	TWN	98	2001-06-28	2024-09-19	24	Programmatic	MPD	Translated	Central Bank of the Republic of China (Taiwan)	East Asia & Pacific
Tajikistan	TJK	23	2017-01-31	2024-04-26	8	Manual	PR	Translated	National Bank of the Republic of Tajikistan	Europe & Central Asia
Thailand	THA	194	2000-05-23	2024-12-18	25	Programmatic	MPCD	Translated	Bank of Thailand	East Asia & Pacific
Turkey	TUR	215	2006-01-23	2024-08-20	19	Programmatic	PR on Interest Rates	Translated	Central Bank of the Republic of Turkey	Europe & Central Asia
Uganda	UGA	97	2011-07-01	2024-12-05	14	Manual	MPS	Official	Bank of Uganda	Sub-Saharan Africa
Ukraine	UKR	87	2014-04-15	2024-12-13	11	Programmatic	MPD	Translated	National Bank of Ukraine	Europe & Central Asia
United Kingdom	GBR	271	1997-06-06	2024-11-07	28	Programmatic	MPSm	Official	Bank of England	Europe & Central Asia
United States	USA	230	1994-02-04	2024-12-18	31	Programmatic	FOMC statement	Official	Board of Governors of the Federal Reserve System	North America
Uzbekistan	UZB	21	2020-01-18	2024-12-12	5	Manual	PR	Translated	Central Bank of the Republic of Uzbekistan	Europe & Central Asia
Zambia	ZMB	31	2016-11-16	2024-11-13	9	Manual	MPC Statement	Official	Bank of Zambia	Sub-Saharan Africa

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD). Region data are from the World Bank country classifications.

Notes: This table presents the complete coverage of the MPSD across all 51 central banks. Document Type abbreviations: MPC = Monetary Policy Committee; MPD = Monetary Policy Decision; MPS = Monetary Policy Statement; IS = Introductory Statement; PR = Press Release; CS = Copom Statement; BD = Board Decision; IRD = Interest Rate Decision; MPCD = Monetary Policy Committee Decision; IRA = Interest Rate Announcement; MPA = Monetary Policy Assessment; RPM = Rate-Setting Policy Meeting; MPCo = Monetary Policy Council; MP = Monetary Policy. Coverage years are calculated as the time between the first and last statements in our database.

3.3 Linguistic Patterns

We characterize the evolution of monetary policy statements using four complementary linguistic measures that capture distinct dimensions of central bank communication: scale, morphological complexity, lexical diversity, and readability (Figure 5).

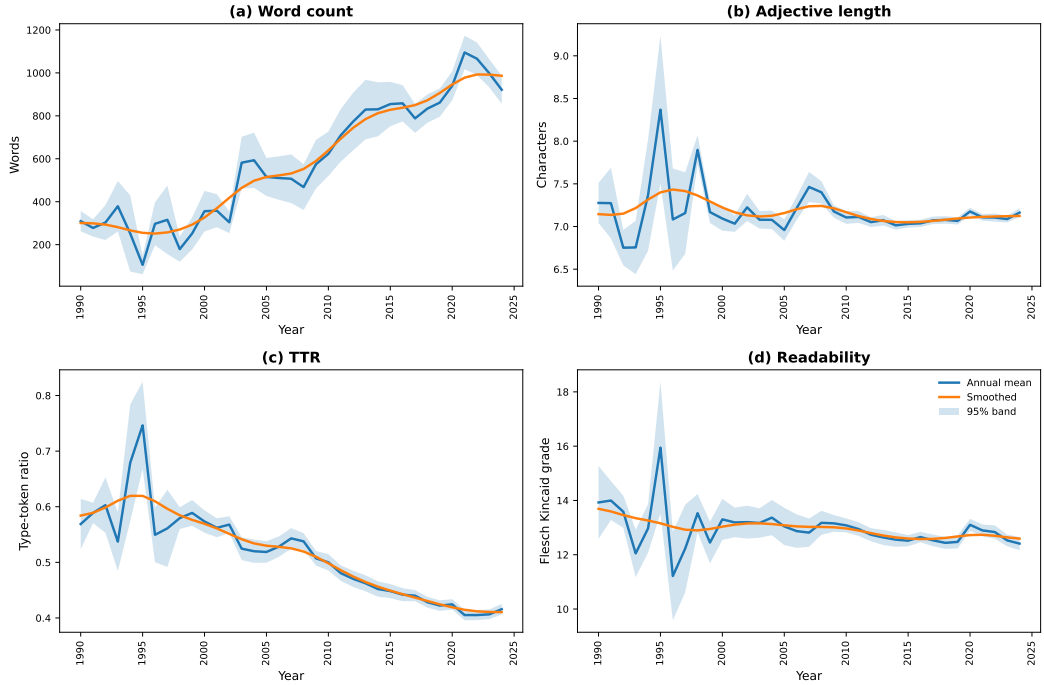


Figure 5: Evolution of Statement Length and Linguistic Characteristics

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: Four document-level linguistic indicators computed for each statement: (i) word count (total tokens); (ii) average adjective length (mean character count of adjectival tokens via part-of-speech tagging); (iii) type-token ratio (unique tokens divided by total tokens); and (iv) Flesch-Kincaid Grade Level. Statement-level measures are aggregated annually using cross-statement means. Shaded bands represent 95% confidence intervals based on the standard error of the annual mean. Smoothed trends obtained via Gaussian filtering.

Panel (a) documents a pronounced upward trend in average word count, with statements approximately tripling in length between 1990 and 2024. The expansion accelerates markedly following the Global Financial Crisis, consistent with the adoption of unconventional monetary policies and forward guidance, which require more extensive explanation. This finding aligns with Masciandaro et al. (2024)'s survey documenting the increasing complexity of central bank communication frameworks. Panel (b) reports average adjective length—a measure of morphological complexity in evaluative language. This indicator remains stable over the sample period, fluctuating modestly around 7.5 characters. The stability suggests that while statements have expanded substantially, central banks have not systematically adopted more complex descriptive vocabulary;

lengthening occurs primarily through additional content rather than denser prose. Panel (c) displays the type-token ratio (TTR), a standard measure of lexical diversity. The secular decline in TTR—from approximately 0.60 in the early 1990s to 0.40 by 2024—indicates increasing lexical repetition as statements lengthen. This pattern is consistent with the institutionalization of standardized policy language and templated communication structures documented in Hansen and McMahon (2016). Panel (d) presents Flesch-Kincaid Grade Level scores, where lower values indicate greater accessibility. Despite the substantial increase in length, readability improves modestly over time, suggesting that central banks have partially offset greater informational content with simpler sentence structures and more accessible word choices. This evolution reflects deliberate efforts to reach broader audiences beyond financial market specialists. Collectively, these patterns reveal a fundamental transformation in central bank communication: statements have become longer and more standardized while maintaining—and even improving—accessibility. This shift toward detailed yet readable policy communication represents a purposeful adaptation to the expanded role of transparency in modern monetary policy frameworks.[4]

A closer examination of monetary policy statements reveals a typical structure including economic forecasts, domestic and foreign economic analysis, policy announcements regarding interest rates and forward guidance, and comments on the policy mix. These statements exhibit recurring sentence patterns, often involving specific parts-of-speech sequences, as detailed in Table 5.

Table 5 illustrates the systematic relationship between economic themes and linguistic structures in monetary policy statements. By identifying these recurring syntactic patterns, we develop more effective NLP tools tailored to central bank communications. For instance, the sequence "policy + AUX + ADJ" is frequently associated with the Monetary Stance, as seen in examples like "policy is restrictive" or "stance remains loose," while the structure "NOUN + ADP + inflation" typically characterizes Inflation Dynamics. These patterns, identified using spaCy's Universal POS tags combined with lexical anchors, provide a robust framework for disentangling the stylistic conventions of central bank communication from simple word frequency.

Figure 6 highlights the most frequent part-of-speech patterns observed in monetary policy statements, providing further evidence of the structured nature of central bank communications.

As shown in Figure 6, common sentence patterns include `DET ADJ NOUN` (e.g., "The strong dollar," "The current rate"), `DET NOUN ADP` (e.g., "The impact of," "The level of"), and `ADP DET NOUN` (e.g., "In the market," "On the economy"). The prevalence of these patterns suggests a high degree of formalization in monetary policy discourse, with central banks employing consistent linguistic structures to

[4]Appendix A.3 presents these linguistic indicators separately for statements originally drafted in English and those translated into English. The similarity of trends across both groups suggests that the main results are not driven by translation-related measurement artifacts.

Table 5: Linguistic Patterns in Monetary Policy Statements by Economic Theme

Economic Theme	**Part-of-Speech Pattern**	**Representative Example**
Inflation Forecasts	inflation + AUX + VERB	inflation **will rise** inflation **might persist**
Growth Projections	growth/GDP + AUX + VERB	growth **should recover** GDP **may stall**
Inflation Dynamics	NOUN + ADP + inflation	risk **of** inflation outlook **for** inflation
Rate Levels	rate + ADP + NUM	rate **at** 5.25% rate **above** 2.0
Forward Guidance	policy/rate + AUX + VERB	policy **will remain** rates **might stay**
Monetary Stance	policy/stance + AUX + ADJ	policy **is restrictive** stance **remains loose**
Fiscal Policy	NOUN + ADP + spending/deficit	impact **of** deficit boost **from** spending
External Conditions	NOUN + ADP + trade/export	demand **for** exports uncertainty **about** trade

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: This table presents the simplified recurring linguistic patterns identified in our corpus. Pattern identification employs a structural approach combining Universal POS tags (AUX = Auxiliary, ADP = Adposition, ADJ = Adjective, NUM = Number) with single lemma-based lexical anchors. This structure allows for the capture of complex syntactic relationships (e.g., "risk of inflation") rather than simple word adjacency.

convey their assessments and decisions. Additionally, the `DET PROPN VERB` pattern is particularly associated with central bank commentary and actions, as illustrated in Appendix Figure 20.

Figure 7 presents a word cloud visualization of terms appearing in the `NOUN + ADP + inflation` pattern as shown in Table 5. This structure effectively filters for the context of inflation rather than mere mentions of the word. Dominant trigrams such as "risks to inflation," "outlook for inflation," and "decline of inflation" highlight that this syntactic pattern is primarily used to communicate uncertainty and future trajectory. The high frequency of terms like "rise," "pressures," and

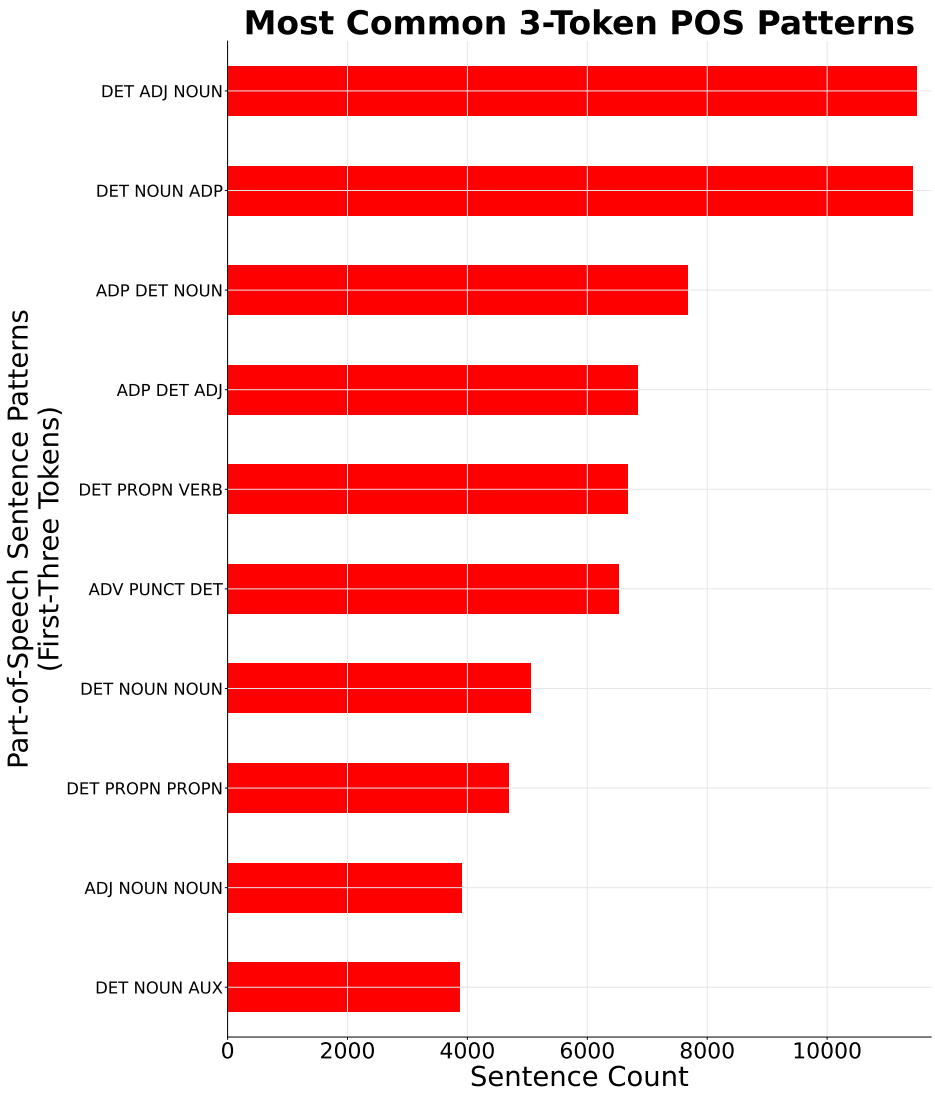


Figure 6: Most Common Patterns by Part-of-Speech (POS)

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

"convergence" confirms that this pattern acts as a linguistic marker for the central bank's assessment of price dynamics.

3.4 Corpus Characteristics and Word Frequency

Figure 8 presents a word cloud visualization of the most frequent terms in our corpus of central bank communications. The prominence of terms such as "inflation," "economic," "growth," "policy," and "financial" reflects the core concerns of monetary policymakers. This visualization not only illustrates the thematic focus of central bank communications but also demonstrates the richness and domain-specificity of our corpus, which is essential for training specialized language models that can accurately capture the nuances of monetary policy discourse.

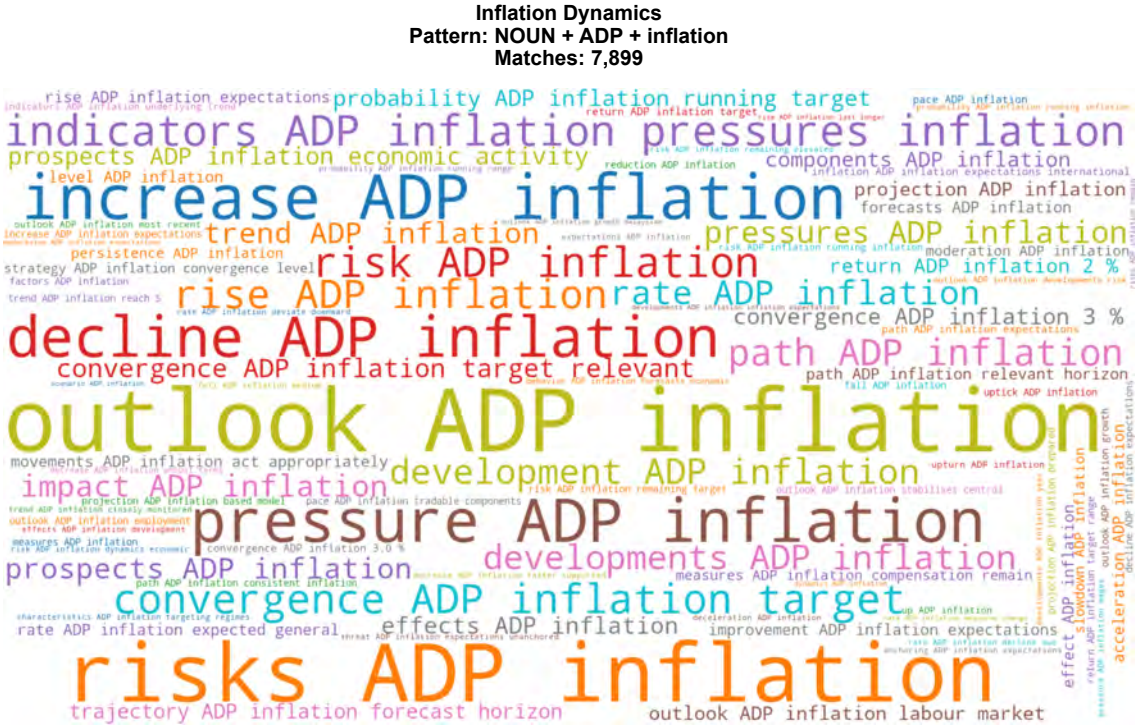


Figure 7: Most Common Trigrams by Part-of-Speech Pattern: `NOUN + ADP + inflation`. The `NOUN + ADP + inflation` pattern highlights contextual usage such as "risks to inflation" and "outlook for inflation."

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

3.5 Inflation Mentions Over Time

We analyze the evolution of key economic terms in monetary policy communications. Figure 9 tracks references to inflation across central banks over time, providing a measure of the prominence of price stability in policy discourse. To account for mechanical variation arising from changes in communication frequency and document length, we report two complementary normalized measures. Panel (a) of Figure 9 shows inflation mentions normalized by the number of statements issued per year, capturing the average intensity of inflation-related discussion per policy communication. Panel (b) normalizes inflation mentions by total word count, reporting mentions per 1,000 words and thus measuring the relative salience of inflation within the textual content of policy statements.

These normalizations ensure that observed changes in inflation-related communication are not mechanically driven by an increased number of policy meetings or longer statements, but instead reflect shifts in the emphasis placed on inflation by central banks. The resulting series reveal substantial comovement in inflation discourse across countries, with particularly pronounced increases during episodes of elevated global inflationary pressure. At the same time, persistent cross-country differences remain, reflecting heterogeneity in mandates,

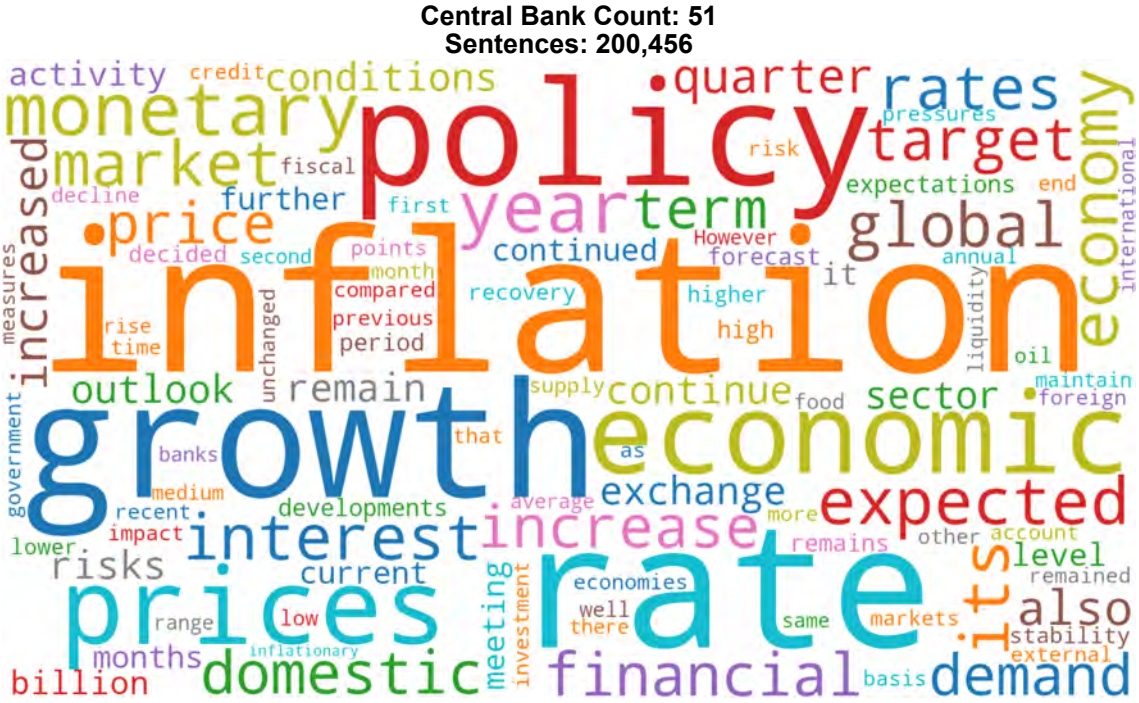
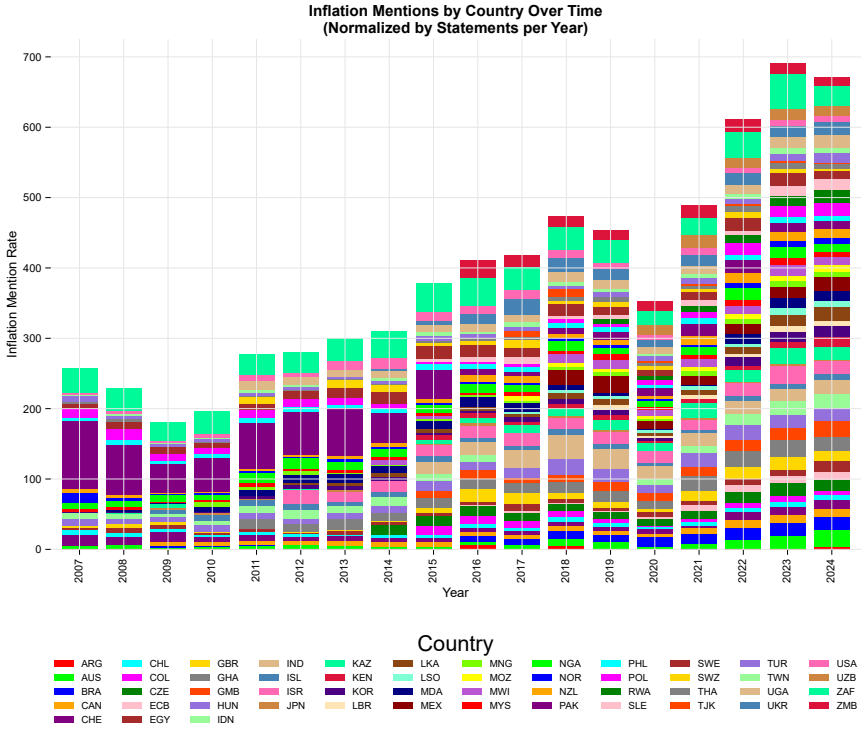


Figure 8: Standardized and Diverse Corpus for Training/Testing A.I.

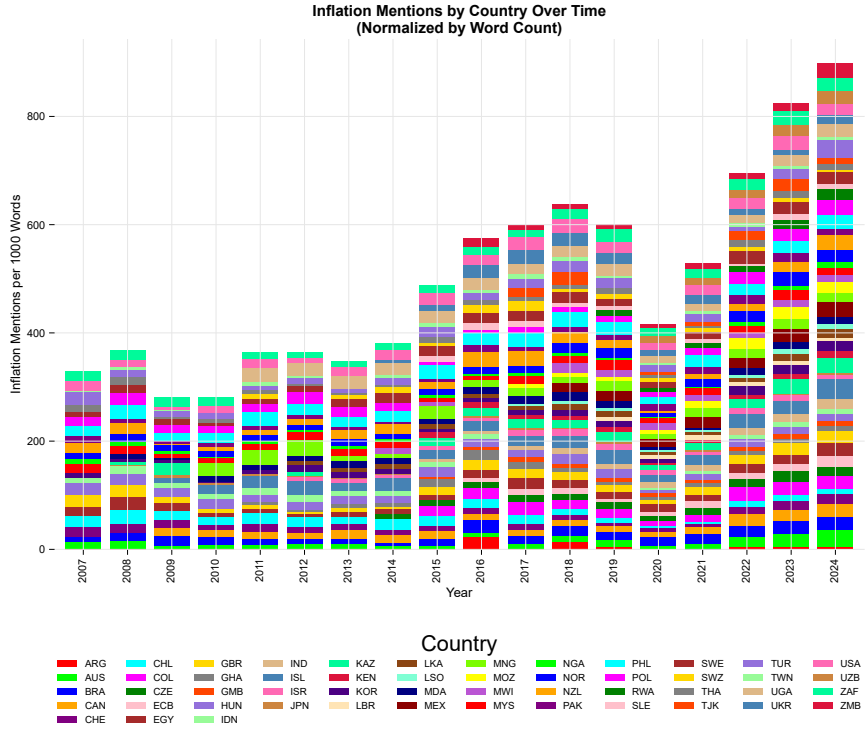
Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

inflation dynamics, and policy frameworks. Overall, the normalized measures provide a clearer interpretation of how central banks' communication priorities evolve in response to changing macroeconomic conditions, while facilitating meaningful comparison across institutions and over time.

This analysis of monetary policy statements contributes to a deeper understanding of how central banks communicate their policy stances and perspectives on economic developments. By dissecting the structure, content, and evolution of these statements, we gain valuable insights into the strategies and objectives of central banks.



(a) Normalized by statements per year



(b) Normalized by word count

Figure 9: Inflation Mentions by Country

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

4 A Generative AI Application to Central Bank Communications

This section illustrates the capabilities of generative AI, specifically LLMs, to analyze economic data and address pertinent challenges in economic research. Our application focuses on examining the GFC, a phenomenon characterized by the synchronized movement of capital, credit, and asset prices across countries (Rey, 2015; Miranda-Agrippino and Rey, 2020, 2022).

The applications presented here are illustrative and are intended to demonstrate how the database and infrastructure can be used, rather than to establish definitive causal claims.

4.1 Conceptual Framework

Central bank communications influence global financial conditions through multiple channels. First, statements from major central banks—particularly the Federal Reserve and ECB—serve as public signals about the state of the global economy, reducing information asymmetries among market participants worldwide (Morris and Shin, 2002). Second, policy communications coordinate expectations across borders, affecting asset prices and capital flows even in countries not directly targeted by the communication (Woodford, 2005). Third, the content and tone of central bank statements provide forward guidance that shapes expectations about future monetary policy paths globally (Blinder et al., 2008). These mechanisms suggest that systematically analyzing central bank communications can reveal how policymakers perceive and respond to global economic dynamics—precisely the objective our infrastructure enables.

4.2 Addressing Challenges in AI Applications to Economics

While LLMs offer immense analytical potential, they present significant methodological challenges for economic research. We identify and address two primary concerns:

1. **Reproducibility**: The lack of transparency surrounding the inputs and internal mechanisms of many LLMs hinders the replication of research findings, a cornerstone of the scientific method. To address this, our study prioritizes open-source code, providing access to LLM code inputs, prompts used for model querying, a clean code base, and user-friendly Jupyter notebooks. This approach ensures complete transparency and facilitates the reproduction of the research process, promoting rigorous scientific inquiry.

2. **Computational Efficiency**: High computational costs and energy consumption associated with LLMs present practical barriers to their

application in economic research. To mitigate this, we implement text-filtering pipelines that leverage keyword and part-of-speech analysis to refine the input data for the LLMs, effectively reducing the computational burden and enhancing efficiency.

While the internal mechanisms of LLMs are not fully transparent, reproducibility in empirical economics relies on the ability to replicate results using the same data, code, and documented procedures. We therefore ensure transparency at every stage under the researcher's control, including data collection, preprocessing, prompt design, model selection, and analytical workflows. This approach is consistent with current best practices in text-as-data research using modern NLP tools (Gentzkow et al., 2019). Importantly, scientific reproducibility does not require replicating the original data collection process; rather, it requires that published results can be exactly reproduced from the shared corpus and code. All figures, tables, and empirical findings in this paper satisfy this criterion.

Table 6 illustrates the output of our LLM-based analysis of global financial conditions over time. The visualization shows how central banks' discussion of global economic drivers has evolved through different economic cycles. This approach demonstrates how LLMs can systematically extract and analyze economic narratives from large corpora of policy communications, offering insights into how policymakers conceptualize and respond to global economic dynamics. The temporal patterns revealed in this analysis align with major economic events and shifts in the GFC, highlighting the value of applying advanced language models to economic research questions.

Table 6: LLM-Generated Analysis of Global Financial Conditions (2007–2013)

Year and Driver	Origin Country	Summary
2007 Global economic conditions, subprime mortgage crisis, global liquidity, U.S. economic slowdown/housing market, commodity prices.	U.S., China	Several central banks identify global economic conditions as a key driver, with concerns arising from the U.S. economic slowdown/housing market and the subprime mortgage crisis. Concerns regarding global liquidity and commodity prices (oil and food) are also noted. The U.S. Federal Reserve is implicitly referenced through discussions of the U.S. economy and monetary policy, but no central bank explicitly identifies the Federal Reserve as the main driver.
2008 U.S. subprime mortgage crisis, Global financial turmoil, Slowdown in the U.S. economy, Global economy slowdown.	United States	The main driver of global economic and financial conditions is the United States, specifically originating from the U.S. subprime mortgage crisis.
2009 Global Financial Crisis / Global Economic Slowdown / Global Recession.	United States (due to U.S. housing bust)	The main driver of global economic and financial conditions during this period was identified as the Global Financial Crisis, originating from the U.S. housing bust. Many central banks acknowledged the significant impact of the crisis on their domestic economies. No central banks directly identified the Federal Reserve as the driver, but the origin of the crisis is cited as the U.S. housing market.
2010 Global Economic Recovery.	Not explicitly one country (mentions of U.S., China, Asia)	The primary driver of global economic and financial conditions is identified as the global economic recovery, influenced by emerging market economies (especially China) and, to a lesser extent, advanced economies like the U.S. and Japan. Multiple texts mention fiscal stimulus measures as supporting the recovery.
2011 Sovereign debt crisis in Europe and global economic slowdown.	Euro Area	The main driver of global economic and financial conditions is the sovereign debt crisis in Europe, which is causing a global economic slowdown. Several central banks identify the European debt crisis as a significant risk. The Federal Reserve's Quantitative Easing (QE) is mentioned by Bank Indonesia as an event to monitor, but it is not portrayed as the primary driver of the global economy.
2012 Sovereign Debt Crisis in Europe / Weak Global Growth.	Euro Area / Multiple Countries	The main driver of global economic and financial conditions appears to be the sovereign debt crisis in Europe and the overall weak global growth. Many central banks across the globe cite the Eurozone crisis as a major source of risk and uncertainty. The central banks do not directly identify the Federal Reserve as the main driver of global economic and financial conditions, although some mention monetary policies of major central banks.
2013 Uncertainty regarding U.S. monetary policy (specifically tapering), fiscal policy, and debt ceiling issues.	United States (primarily), Euro Area and China	Several central banks identify the U.S. Federal Reserve's potential tapering of quantitative easing as a major driver of global financial market volatility and a downside risk to growth. The U.S. fiscal situation and debt ceiling debates also contribute to uncertainty. Concerns about the Euro Area's economic challenges and China's growth slowdown are also frequently mentioned.

Source: Structured summaries derived from the Monetary Policy Statement Database. We filtered 2,921 unique sentences containing the keywords "global" and "spillover" by year using spaCy (see Section 4.2) and generated summaries via Google's Gemini (gemini-2.0-flash) using the prompt in Appendix A.2. All input texts (sentences) are stored in JSON for replication and will be made available along with database metadata.

4.3 LLM Data Pipeline Infrastructure

Our methodological framework centers on a scalable LLM-based data pipeline architecture that integrates diverse language models—both proprietary (Claude, Gemini, ChatGPT) and open-source (hosted on Hugging Face)—within a unified computational environment. The pipeline's modular design permits seamless substitution of text corpora and models, enabling comparative analysis across central bank communications and addressing potential concerns regarding model-specific biases. This flexibility strengthens the external validity of our findings and ensures the analytical framework can seamlessly integrate emerging language models and additional data sources.

The architecture is designed for scalability along two dimensions:

- **LLM Agnosticism**: The system is compatible with any model (e.g., ChatGPT, Gemini, Claude) and task (e.g., sentiment analysis, QA)
- **Text/Prompt Flexibility**: The pipeline can be dynamically repurposed to evaluate alternative texts and prompts

4.4 Methodological Approach

To illustrate the application of LLMs to central bank communications, we explore the GFC through two key analytical approaches:

4.4.1 Question Answering

In the QA task, we prompt the LLM to provide summaries of the drivers of the global economy, with particular attention to the role of the Federal Reserve. The core directive of the prompt (the full prompt is available in Appendix A.2) is summarized as follows:

> "I want you to identify the main driver of global economic and financial conditions and the origin country driving these conditions. In the summary, please note if any central banks identify the Federal Reserve as the driver of global economic or financial conditions."

Table 7 presents a comparative analysis of how different LLMs interpret the *same* central bank text. This side-by-side comparison reveals significant variations in the summaries generated by different models, even when processing identical inputs and the same prompt. Notably, Claude identifies the U.S. Federal Reserve as the specific driver, whereas all other models—including Gemini (see Table 6)—attribute the influence to the U.S. economy more generally. These differences underscore the importance of transparency in LLM-based research and the need for careful model selection and evaluation when applying these tools to economic analysis. The figure illustrates why our pipeline's ability to integrate multiple

models is valuable—it allows researchers to compare outputs across different LLMs and assess the robustness of their findings.

Table 7: Comparative Analysis of LLM Interpretations of Global Economic Conditions (2008)

Model	Identified Mechanism	Origin Country	Summary Assessment
Claude	The Federal Reserve	United States	Based on central bank statements, the main driver is the Federal Reserve. Several central banks (e.g., Canada, New Zealand, Australia) identified the Fed's actions as a key factor impacting global growth. The global slowdown stemming from the U.S. housing bust was a recurring theme.
Xiaomi	Global financial market turmoil and economic slowdown	United States	The primary driver is global financial market turmoil stemming from the U.S. subprime mortgage crisis. The Federal Reserve is acknowledged by multiple central banks (e.g., PK, ZA, NG, GB, ECB, JP) as a key driver given its role in global stability.
Mistral	Slowdown in the U.S. economy / subprime mortgage crisis	United States	Multiple central banks cite the US economic slowdown—particularly the subprime crisis—as the primary driver. Notably, no central bank explicitly names the Federal Reserve as the driver in this interpretation.
Arcee	U.S. subprime mortgage crisis and broader U.S. economic weakness	United States	Multiple central banks explicitly cite the U.S. subprime crisis and economic slowdown as primary drivers. Several banks note that the Federal Reserve's policy actions are relevant to addressing these conditions.

Source: Structured summaries derived from the Monetary Policy Statement Database. We filtered 2,921 unique sentences containing the keywords "global" and "spillover" by year using spaCy (see Section 4.2) and generated summaries via Anthropic's Claude, mistralai/devstral-2512, arcee-ai/trinity-mini, xiaomi/mimo-v2-flash. All input texts (sentences) are stored in JSON for replication and will be made available along with database metadata.

4.4.2 Aspect-Based Sentiment Analysis

For the ABS task, we instruct the LLM to extract sentiment toward specific themes from central bank communications. The core directive of the prompt (the full prompt is available in Appendix A.1) is summarized as follows:

> "Analyze the provided central bank text to identify discussions regarding the impact of global financial conditions on domestic economies, focusing on themes such as economic growth, financial conditions, capital flows, and monetary policy spillovers. For each

explicitly discussed theme, assign a continuous sentiment score on a scale from 0.0 (negative) to 2.0 (positive), where 1.0 represents a neutral view."

This allows us to track sentiment toward specific economic factors across time and central banks, providing insights into how financial conditions and policy spillovers are perceived globally.

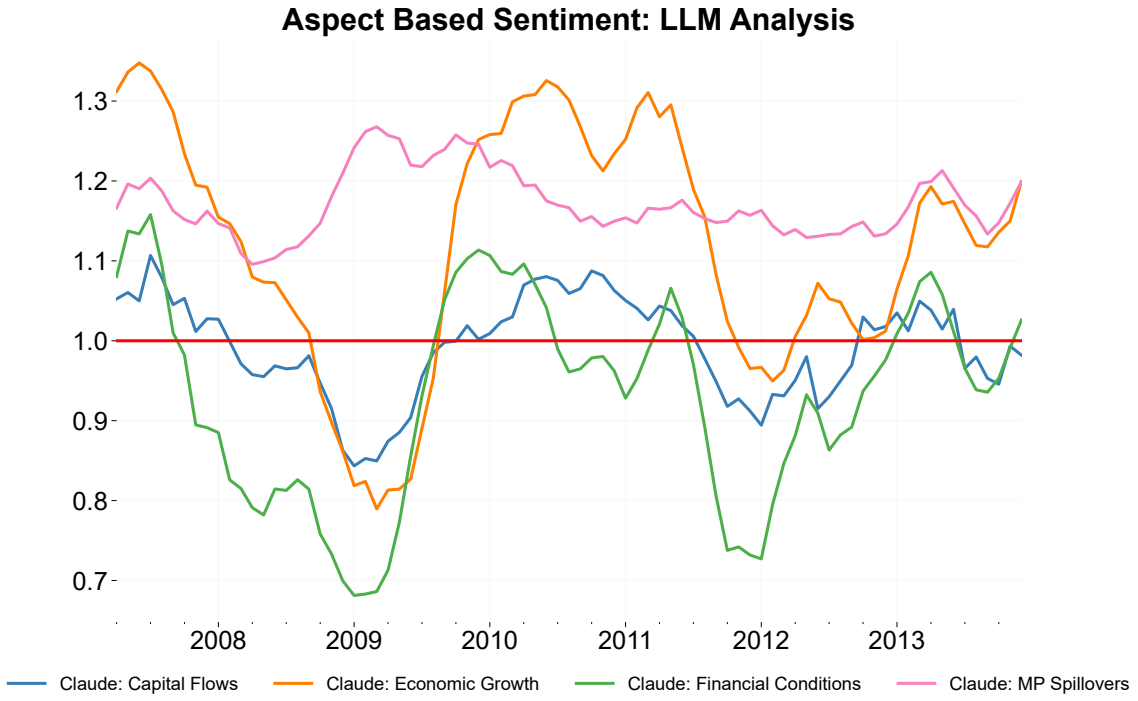


Figure 10: LLM-Generated ABS over Time

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLM tools.

Notes: Sentiment scores are generated using a version of Anthropic's Claude LLM (claude-3-haiku-20240307). The series capture sentiment related to capital flows, economic growth, financial conditions, and monetary policy spillovers (MP Spillovers), averaged across all countries.

Figure 10 illustrates the evolution of sentiment across different economic aspects over time, as extracted by our LLM-based analysis. The figure demonstrates distinct patterns in how central banks express sentiment toward different themes, with notable divergences during periods of economic stress. For instance, the sentiment toward economic growth and financial conditions shows greater volatility compared to monetary policy spillovers and, to a lesser extent, capital flows, particularly during and after the Global Financial Crisis. These temporal patterns offer valuable insights into the shifting priorities and concerns of monetary policymakers across different economic cycles.

4.5 Validation of LLM-Based Sentiment

To assess the relationship between traditional lexical sentiment measures and LLM-derived sentiment indices, Figure 11 reports the correlation matrix across all sentiment metrics used in the analysis (see the extended matrix in Appendix Figure 18). Several clear patterns emerge. First, established finance-oriented lexical measures exhibit moderate to strong similarity among themselves. The Correa and Loughran-McDonald indices are strongly correlated (0.76), indicating that both capture a similar notion of financial tone in monetary policy statements. The Hubert index shows weaker alignment with these measures (correlations of 0.28 with Correa and 0.33 with Loughran-McDonald), while the general-purpose Harvard IV-4 dictionary displays substantially lower and less stable correlations, including a negative association with the Hubert index (-0.30). This pattern underscores the importance of domain-specific sentiment tools when analyzing central bank communication.

Second, LLM-derived sentiment measures show consistently strong correlations with finance-specific lexical indices, particularly for the Economic Growth dimension. Across models, Economic Growth sentiment correlates between 0.74 and 0.88 with the Loughran-McDonald index and between 0.75 and 0.78 with the Correa index. In contrast, correlations with the Harvard IV-4 dictionary remain markedly lower, typically below 0.26 (with even 0.08). This divergence suggests that LLMs align more closely with financial-domain sentiment constructs rather than with generic polarity-based lexicons.

Third, correlations differ systematically across sentiment dimensions. Financial Conditions sentiment exhibits weaker and more heterogeneous correlations with lexical measures than Economic Growth sentiment, with values ranging from 0.44 to 0.76 against Loughran-McDonald and from 0.56 to 0.81 against Correa. At the same time, Financial Conditions sentiment shows very strong agreement across LLMs themselves, with correlations exceeding 0.75 in most cases. This pattern indicates that while LLMs consistently identify a shared financial-conditions signal, that signal is only partially captured by traditional dictionary-based approaches.

Finally, the high cross-model correlations among LLMs—reaching 0.90–0.93 for Economic Growth sentiment—point to a robust consensus across different architectures and training paradigms. Rather than producing idiosyncratic sentiment measures, LLMs converge on similar interpretations of economic tone, reinforcing their reliability as measurement tools. Taken together, these results suggest that LLM-based sentiment measures are neither orthogonal to nor simple replications of traditional lexical indices. Instead, they complement established finance-specific dictionaries by capturing structured, dimension-specific sentiment—particularly for growth and financial conditions—while diverging from broad, non-domain-specific sentiment constructs. This finding aligns with the view that monetary policy communication conveys multiple informational

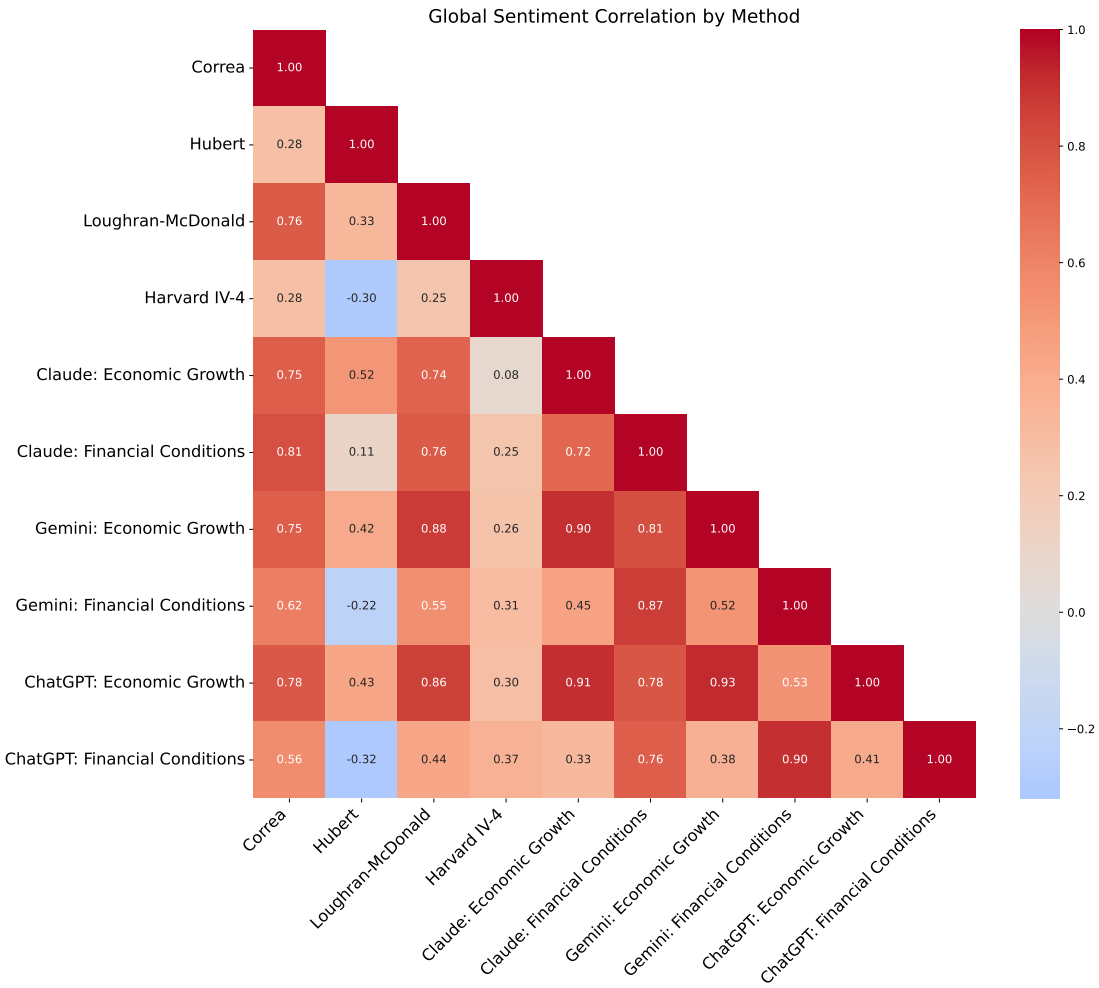


Figure 11: Correlation Matrix of Lexical and LLM-Based Sentiment Measures

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLM tools and Loughran-McDonald, Correa, Hubert, and Harvard IV-4 dictionaries.

Notes: Sentiment scores are generated using Loughran-McDonald, Correa, Hubert, and Harvard IV-4 dictionaries, and a version of Anthropic's Claude LLM (claude-3-haiku-20240307), OpenAI's ChatGPT (gpt-4o-mini) and Google's Gemini (gemini-2.0-flash). Correlations are computed based on the cross-sectional mean of the sentiment scores, aggregated across all sample countries.

dimensions and that modern LLMs are well-suited to disentangling these dimensions in a way that remains consistent with established financial text-analysis benchmarks.

These results contribute to the literature on text-as-data benchmarks. While Gentzkow et al. (2019) caution that different measures often target different estimands, our findings demonstrate reassuring convergence when the domain is strictly defined. The results support the premise of Loughran and McDonald (2011) regarding the necessity of domain-specific sentiment lexicons: the LLMs naturally emulate the financial-specific weighting of the Loughran–McDonald dictionary rather than the generic Harvard IV-4. Consequently, while Hansen and McMahon (2016) note that policy statements contain multiple informational

dimensions, our analysis suggests that modern LLMs are highly effective at identifying the specific dimensions of financial risk and economic growth that traditional financial dictionaries were designed to capture.

We identify three mechanisms through which LLMs outperform dictionary methods in this setting. First, LLMs capture contextual negation and qualification that bag-of-words approaches miss. When a central bank states that "inflation pressures have not materialized as expected," dictionary methods count "inflation" and "pressures" as negative, while LLMs correctly interpret the overall sentiment as positive (declining concern). Second, LLMs handle domain-specific terminology more flexibly. Central bank communications frequently use technical phrases ("quantitative easing," "forward guidance," "macro-prudential") whose sentiment depends heavily on context; LLMs infer meaning from surrounding text while dictionaries either ignore such terms or assign fixed scores. Third, LLMs can extract ABS—distinguishing concern about inflation from concern about growth—whereas most dictionary approaches produce a single aggregate score. This granularity is essential for understanding how central banks balance competing objectives in their communications. Our correlation analysis confirms these advantages: LLM-derived sentiment for specific dimensions (growth, financial conditions) correlates more strongly with the GFC than aggregate dictionary-based measures, suggesting that aspect-specific extraction captures economically meaningful variation that aggregate approaches obscure.

4.5.1 Validation Framework

Given the absence of labeled ground truth for ABS in central bank communication, we validate LLM-derived measures using two complementary strategies. First, we benchmark LLM-based sentiment indices against established dictionary-based sentiment measures commonly used in the economics literature. We interpret agreement as evidence of plausibility and divergence as reflecting differences in the underlying estimands rather than measurement error per se. Consistent with this framework, we find that our LLM-derived sentiment tracks finance-specific lexical measures closely, validating the models' ability to capture relevant economic tone. Figure 11 summarizes cross-method correlations between lexical sentiment measures and LLM-derived aspect sentiments, while Figure 12 compares their time-series behavior around major global episodes. Second, we assess the robustness of LLM-based sentiment by comparing outputs across multiple LLMs. Specifically, in addition to our baseline results based on Claude, we replicate the ABS extraction using alternative LLMs (Gemini and ChatGPT). Figure 17 reports the resulting sentiment series for key economic dimensions, including economic growth and financial conditions. Despite differences in scale and volatility across models, the series exhibit strong comovement and similar temporal dynamics, particularly around major macroeconomic events. This cross-model consistency provides an additional validation layer, indicating that

the extracted sentiment patterns are not driven by idiosyncratic behavior of a single LLM architecture. Finally, to ground these statistical aggregates in the underlying text, we conducted a manual review of a random sample of statements using our validation tool. While this qualitative inspection does not constitute a formal scientific audit, we observed substantial agreement between LLM outputs and human interpretation of the underlying text. An illustrative example of this validation exercise is provided in the Appendix (Figure 21). Taken together, benchmarking against transparent lexical measures, comparing across independent LLMs, and spot-checking against human judgment enhances confidence in the stability and interpretability of the proposed sentiment indicators, while preserving their ability to capture contextual and aspect-specific information that is difficult to obtain using dictionary-based approaches alone.

We acknowledge three limitations inherent to LLM-based sentiment extraction. First, prompt sensitivity may affect results; we mitigate this concern by documenting all prompts in the repository and comparing outputs across different model architectures. Second, model updates may affect reproducibility over time; we therefore report exact model versions (e.g., claude-3-haiku-20240307, gemini-2.0-flash, gpt-5-nano) and archive all outputs. Third, LLMs may exhibit biases inherited from training data; our cross-model comparison suggests that such biases do not dominate, as models trained on different corpora produce highly correlated sentiment series. These limitations are common to all LLM-based text analysis and do not invalidate our findings; rather, they underscore the importance of the validation framework we employ.

4.6 Empirical Implementation

Our implementation employs a multi-stage filtering process to extract relevant narrative elements from the corpus:

1. Search using keywords (global, spillover, flows): yielding 6,135 sentences
2. Truncate for specific time periods (e.g., 2019 data): reducing to 1,228 sentences

This approach balances computational efficiency with analytical rigor, addressing the computational challenges of LLM analysis of large text corpora while ensuring systematic sampling for robust results.

4.7 Results

Our analysis using LLMs produces several important insights into the GFC:

4.7.1 Origins of Global Financial Conditions

Our QA analysis, shown in Table 6, indicates that central banks often cite the U.S. economy as a driver of global economic conditions, and in some instances,

specifically mention the U.S. Federal Reserve as a factor influencing global economic and financial conditions, particularly during periods of high uncertainty such as 2007 and 2009. However, this is not a consistent trend, and in many years, broader economic factors in the U.S. and other regions, such as the Eurozone sovereign debt crisis and emerging market economies, are perceived as more impactful. This nuanced understanding of global economic influences, where the relative importance of different factors, including central bank actions, can vary over time, contrasts with some academic literature that tends to emphasize the direct and dominant influence of Fed policy on global financial conditions, suggesting a potential gap between academic models and policymakers' conceptualization of global economic linkages (Kalemli-Özcan, 2019).

4.7.2 Sentiment Analysis Results

The ABS analysis shows that central banks expressed greater concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period, likely reflecting the lasting impact of the Global Financial Crisis and increased global interconnectedness.

Our Granger causality analysis follows the formal specification:

$$\text{GFC}_t = \alpha + \sum_{k=1}^{p} \beta_k \text{GFC}_{t-k} + \sum_{k=1}^{p} \gamma_k \text{Sentiment}_{t-k} + \varepsilon_t \tag{1}$$

$$\text{Sentiment}_t = \alpha' + \sum_{k=1}^{p} \beta'_k \text{Sentiment}_{t-k} + \sum_{k=1}^{p} \gamma'_k \text{GFC}_{t-k} + \varepsilon'_t \tag{2}$$

where GFC_t represents the GFC factor at time t, Sentiment_t represents our measure of central bank communication sentiment, and p represents the lag order. The null hypothesis that sentiment does not Granger cause GFC corresponds to $\gamma_k = 0$ for all $k \in \{1, 2, ..., p\}$. Similarly, the null hypothesis that GFC does not Granger cause sentiment corresponds to $\gamma'_k = 0$ for all $k \in \{1, 2, ..., p\}$.

We conduct these tests using lag orders $p \in \{1, 2, 3\}$ and report the resulting p-values in Figure 13. The results consistently reject the null hypothesis that sentiment does not Granger cause GFC, while failing to reject the null hypothesis that GFC does not Granger cause sentiment. This asymmetric causality pattern provides robust evidence that central bank communication sentiment influences global financial conditions rather than merely responding to them.

We emphasize that Granger causality is a statistical concept capturing predictive precedence rather than structural causation in the economic sense (Granger, 1969). The asymmetric pattern we document—where sentiment predicts GFC but not vice versa—is consistent with communication serving as a transmission channel, though we cannot rule out omitted common factors or reverse causality operating at frequencies not captured by our quarterly specification. Nevertheless, the finding that communication sentiment leads the

GFC aligns with theoretical models in which central bank signals coordinate expectations and thereby influence asset prices (Morris and Shin, 2002; Woodford, 2005).

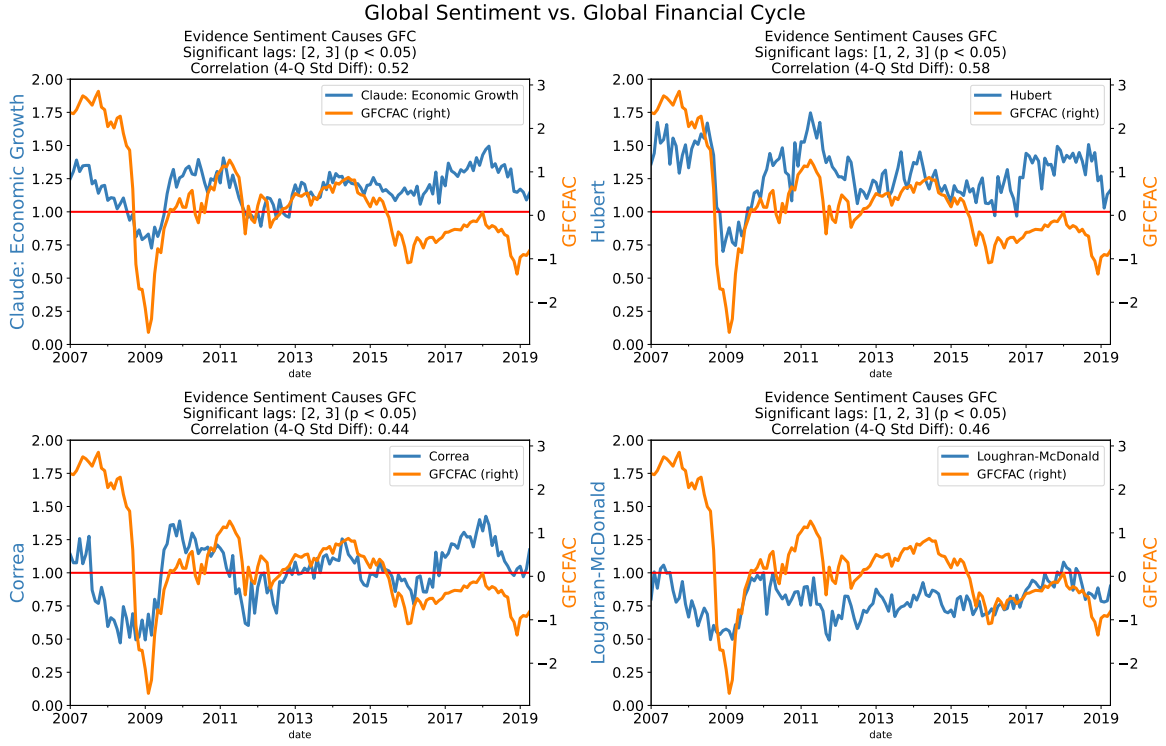


Figure 12: Comparison of LLM-Generated Sentiment Indices with Lexical Approaches

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD), dictionaries and LLM-based (Anthropic's Claude LLM claude-3-haiku-20240307) sentiment measures, and GFC factor (GFCFAC).

Notes: GFCFAC denotes the GFC factor. Blue lines show standardized global sentiment indices; orange lines show GFCFAC (right axis). Reported correlations refer to four-quarter standardized differences.

Figure 12 compares global sentiment measures derived from LLM-based and traditional lexical approaches with the GFC factor (GFCFAC). Contrary to the hypothesis that ABS measures might decouple from the aggregate cycle, the figure highlights a strong structural alignment between the methods. The LLM-based economic growth sentiment displays a robust correlation (0.52) with the GFC, effectively outperforming traditional finance-specific dictionaries such as Correa (0.44) and Loughran-McDonald (0.46), and performing comparably to the Hubert index (0.58). All measures exhibit significant predictive power at lags 2 and 3, confirming that the LLM-derived signal captures the same fundamental turning points as established benchmarks.

Taken together with Figures 10 and 11, these patterns indicate that LLM-based sentiment is highly consistent with the signal captured by domain-specific lexical dictionaries. Rather than diverging, the LLM-derived indices appear to refine the

signal found in traditional measures. The fact that the LLM-based growth sentiment tracks the GFC more closely than the Loughran–McDonald and Correa indices suggests that the model's contextual awareness allows it to filter out noise better than bag-of-words approaches. Consequently, the observed convergence between these methods serves as a robust cross-validation, demonstrating that LLMs can reliably recover key macroeconomic dynamics while offering the additional advantage of aspect-specific granularity.

4.7.3 Empirical Evidence for Sentiment Driving the GFC

Beyond descriptive analysis, we find empirical evidence that dictionary-based sentiment expressed in central bank communications may drive components of the GFC. Figure 13 presents results from Granger causality tests using the benchmark Correa sentiment index, showing that aggregated sentiment from both developed and emerging markets Granger-causes movements in the GFC (measured as a 4-quarter difference). Importantly, we do not find evidence of reverse causality, suggesting that central bank communications may play a causal role in shaping global financial conditions. For this analysis, we classify countries following World Bank income categories: high-income economies (per capita GNI above $13,845 in 2022, following the WDI/Atlas-method cutoff) are designated as developed markets, while upper-middle and lower-middle income economies are designated as emerging markets. A complete mapping of each central bank to its classification category is provided in the accompanying metadata file.

Notably, emerging market (EM) sentiment appears to be a more significant driver of the GFC than sentiment from core central banks (ECB, Australia, Canada, Switzerland, UK, Japan, and US). This finding may reflect that EM sentiment captures global investors' risk-on/risk-off preferences, which themselves influence the GFC. Thus, the relationship between EM central bank sentiment and the GFC likely operates through the former's capacity to reflect and amplify global investor risk appetite, rather than through direct causation. These findings extend the literature on international policy transmission by identifying communication as a significant channel through which central banks influence global financial markets.

To assess whether this predictive relationship extends beyond dictionary-based measures, Appendix Figure 19 reports analogous Granger causality tests using LLM-derived, ABS indicators for economic growth and financial conditions. Consistent with the benchmark results based on the Correa index, we find strong evidence that LLM-based sentiment Granger-causes movements in the GFC across multiple model architectures and lag specifications. In particular, sentiment related to economic growth exhibits robust predictive power for the GFC at horizons of one to three quarters, while financial-conditions sentiment also shows significant predictive content, albeit with greater heterogeneity across models. Importantly, mirroring the main results, we again find little evidence of reverse causality from the GFC to LLM-based sentiment. This asymmetry suggests that

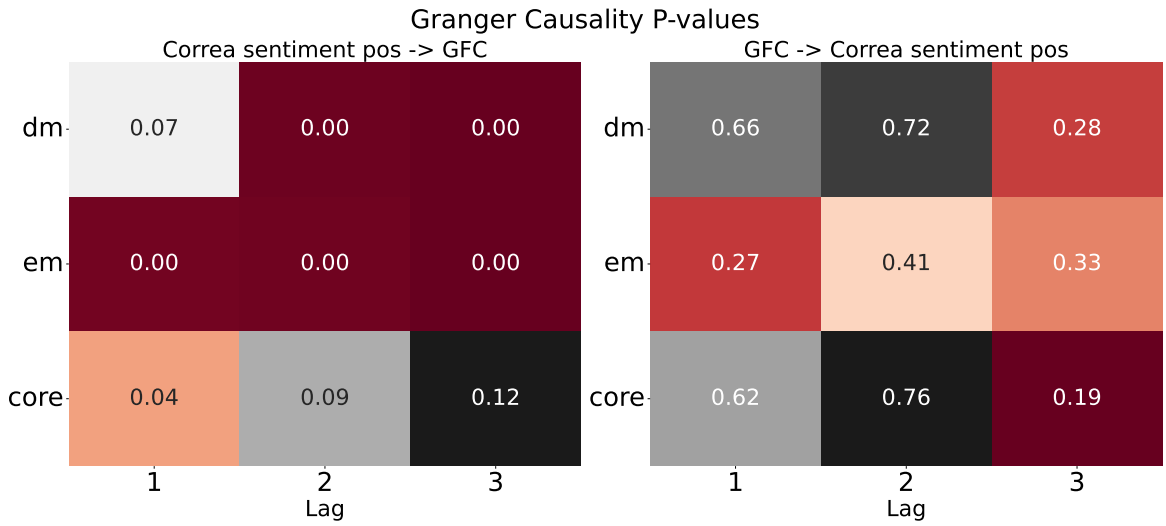


Figure 13: Dictionary-Based Sentiment Granger-Causes the GFC

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: Sentiment is the benchmark dictionary-based index from Correa et al. (2021) computed from MPSD statements. The GFC is from Miranda-Agrippino and Rey (2020). The figure reports Granger-causality test p-values (rounded to two decimal places) for lag orders $p \in \{1, 2, 3\}$ (see text). "core" denotes a subset of advanced-economy central banks comprising the ECB and the central banks of Australia, Canada, Switzerland, the United Kingdom, Japan, and the United States. "dm" (developed markets) includes Australia, Canada, Switzerland, the Czech Republic, the United Kingdom, Hungary, Israel, Iceland, Japan, South Korea, Norway, New Zealand, Poland, Sweden, Taiwan, and the United States. "em" (emerging markets) includes Argentina, Brazil, Chile, Colombia, Egypt, Ghana, Gambia, Indonesia, India, Kenya, Kazakhstan, Sri Lanka, Liberia, Lesotho, Moldova, Mongolia, Malawi, Mexico, Malaysia, Mozambique, Nigeria, the Philippines, Pakistan, Rwanda, Sierra Leone, Eswatini, Thailand, Tajikistan, Turkey, Ukraine, Uganda, Uzbekistan, South Africa, and Zambia. Country groupings follow World Bank income classifications: high-income economies are classified as developed markets, while upper-middle and lower-middle income economies are classified as emerging markets.

LLM-derived sentiment captures forward-looking information embedded in central bank communications, rather than merely responding mechanically to contemporaneous global financial conditions. Taken together, these appendix results reinforce the interpretation that the predictive content of central bank communication for the GFC is not an artifact of a particular sentiment construction, but persists across both traditional lexical indices and modern LLM-based approaches.

4.7.4 Regional Variation in Sentiment

Our analysis also shows important regional variations in how central banks perceive and respond to global financial conditions. Figure 14 illustrates how sentiment evolves differently across major economic regions, potentially reflecting heterogeneous exposure to global shocks and different policy responses.

Figure 14 displays distinct regional patterns, with periods of both convergence

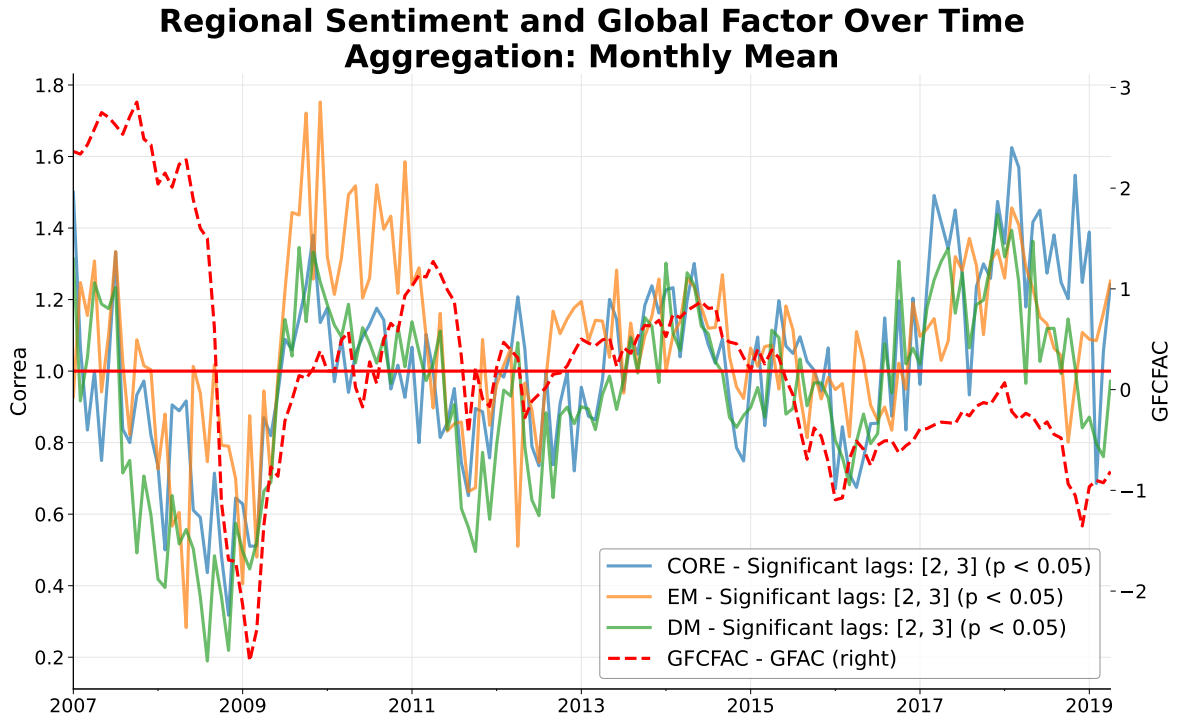


Figure 14: Regional Sentiment Evolution

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using Correa dictionary.

Notes: "Core" denotes a subset of advanced-economy central banks comprising the ECB and the central banks of Australia, Canada, Switzerland, the United Kingdom, Japan, and the United States. "DM" (developed markets) includes Australia, Canada, Switzerland, the Czech Republic, the United Kingdom, Hungary, Israel, Iceland, Japan, South Korea, Norway, New Zealand, Poland, Sweden, Taiwan, and the United States. "EM" (emerging markets) includes Argentina, Brazil, Chile, Colombia, Egypt, Ghana, Gambia, Indonesia, India, Kenya, Kazakhstan, Sri Lanka, Liberia, Lesotho, Moldova, Mongolia, Malawi, Mexico, Malaysia, Mozambique, Nigeria, the Philippines, Pakistan, Rwanda, Sierra Leone, Eswatini, Thailand, Tajikistan, Turkey, Ukraine, Uganda, Uzbekistan, South Africa, and Zambia. Country groupings follow World Bank income classifications: high-income economies are classified as developed markets, while upper-middle and lower-middle income economies are classified as emerging markets.

and divergence in sentiment. While advanced economies often show similar sentiment trends, emerging markets display more idiosyncratic patterns, particularly during periods of financial stress. These regional differences suggest that the GFC affects economies differently based on their level of financial integration, institutional characteristics, and economic fundamentals.

These findings contribute to the literature on the GFC by providing systematic evidence on how central banks perceive, discuss, and potentially influence global financial conditions through their communications.

5 Discussion

5.1 Policy Implications

Our findings have several implications for monetary policy implementation and coordination. First, the evidence that central bank sentiment Granger causes movements in the GFC (see Figures 13 and 19) suggests that communication strategies may have cross-border spillover effects beyond what is typically acknowledged. This raises questions regarding central banks' responsibility to consider the international implications of their communications, not just their policy actions, and implies that central banks should carefully calibrate their statements, recognizing that they may influence financial conditions well beyond their jurisdictions.

Second, the finding that central banks more frequently cite broad U.S. economic conditions rather than the Federal Reserve specifically as driving global conditions reveals a potential disconnect in how policy spillovers are conceptualized. While the academic literature often focuses on monetary policy transmission across borders (Miranda-Agrippino and Rey, 2020), central bankers themselves appear to emphasize broader economic linkages, suggesting that international policy coordination may need to encompass fiscal and regulatory dimensions as well. The distinction between attributing global influences to a country's economy versus its central bank has meaningful implications for how we model and address international policy spillovers.

Third, the observed regional variations in how sentiment responds to global conditions (as illustrated in Figure 14) highlight the potential for heterogeneous impacts of global shocks, which complicates the case for uniform policy approaches across different economies. This supports arguments for tailored policy frameworks that account for specific vulnerabilities and transmission channels in different economic contexts (Armelius et al., 2020). The substantial differences in how emerging market and advanced economy central banks discuss global conditions suggest that policy coordination efforts should account for these varying perspectives and constraints.

Our linguistic analysis of monetary policy statements (see Table 5 and Figure 6) also reveals how central banks have institutionalized certain communication patterns to convey policy intentions. These formalized linguistic structures may enhance the predictability of policy communications but could potentially constrain adaptability during rapidly evolving economic conditions. Policymakers might benefit from considering how these established communication patterns influence market interpretations and whether greater flexibility in communication styles could improve policy effectiveness in certain contexts.

Beyond these specific findings, our analysis suggests that if central bank narratives actively shape global financial conditions—as our Granger causality results suggest—then coordinating communication strategies may be as important

as coordinating policy rate decisions. The post-GFC trend toward longer, more detailed statements reflects a genuine tradeoff: while elaboration enhances transparency, it also increases the complexity of the signal that markets must interpret. Our readability analysis suggests that central banks have partially offset this complexity with simpler sentence structures, but the optimal balance between completeness and clarity remains an open question. The substantial cross-model agreement we document among LLMs suggests that these tools can reliably extract economically meaningful signals from text, but the systematic differences we observe (such as Claude's tendency to identify the Federal Reserve specifically while other models reference the U.S. economy more broadly) highlight the importance of methodological transparency when using these tools for policy analysis.

5.2 The MPSD as a Research Tool

The MPSD offers significant advantages for researchers in economics, finance, and related fields. By providing standardized access to monetary policy communications across multiple countries and time periods, it enables novel comparative analyses previously infeasible due to data limitations. For researchers interested in comparative economics, the database facilitates studies of how institutional differences across central banks manifest in their communication strategies. For those studying policy transmission, the coverage allows for analysis of how market reactions to policy communications vary across different economic environments and institutional frameworks. The significant variations in statement frequency, length, and readability documented in Tables 3 and 4 provide a rich set of institutional variations that can be exploited to identify the determinants and effects of different communication approaches.

The integrated technical infrastructure supporting the database (illustrated in Tables 1 and 2) is particularly valuable for researchers applying computational methods to economic questions. The modular design allows for customization to specific research needs while maintaining reproducibility, addressing a key challenge in computational social science (Gentzkow et al., 2019). Our standardized folder structure (Figure 1) and systematic approach to data processing ensure that analyses can be precisely replicated across different research contexts. The evolution of monetary policy statement characteristics over time, as documented in Figure 5, provides valuable insights into how central bank communication strategies have adapted to changing economic conditions and policy frameworks. Researchers can exploit this temporal variation to study how communication evolves in response to institutional learning, changes in monetary policy regimes, and shifts in the broader economic environment.

5.3 Future Extensions

Several promising directions exist for extending both the database and its applications. First, expanding the corpus to include speeches, minutes, and press conferences would provide a more complete view of central bank communication strategies and enable analysis of how different channels interact to shape market expectations.

Second, incorporating multimedia elements such as press conference transcripts and videos could enable multimodal analysis, including non-verbal cues that may convey policy signals. Advances in computer vision and audio processing make such analysis increasingly feasible.

Third, developing domain-specific pre-training approaches for language models—including continued pre-training on central bank corpora or specialized architectures that account for the hierarchical and temporal structure of policy communications—could improve performance on monetary policy tasks. Our analysis of part-of-speech patterns (Table 5 and Figure 6) provides a foundation for such specialized models.

Finally, integrating the database with high-frequency financial data could facilitate more granular analyses of market reactions to specific communication elements, advancing our understanding of the information effects of monetary policy communications (Nakamura and Steinsson, 2018; Bauer and Swanson, 2023). Building on our finding that sentiment Granger causes the GFC (Figures 13 and 19), such integration would enable more precise identification of which aspects of communication drive market responses across different asset classes.

6 Conclusion

This paper introduces the MPSD, a research infrastructure that addresses fundamental gaps in the empirical analysis of monetary policy communication. By compiling and standardizing 6,693 monetary policy statements from 51 central banks spanning 1990–2024, the database constitutes an unprecedented resource for systematic cross-country analysis, offering exceptional temporal and geographic coverage (Figures 2 and 4). Its technical architecture integrates version control, automated data collection, and standardized preprocessing pipelines (Table 1), ensuring precise replicability while leveraging advanced computational tools. The modular design facilitates extension to diverse research questions (Table 2), supporting the growing integration of NLP methods into economic analysis.

Our linguistic analysis uncovers systematic patterns in how central banks structure their communications (Table 5 and Figure 6), advancing the understanding of institutionalized discourse in monetary policy and informing the development of specialized NLP tools for this domain. Exploiting the breadth of the database, we document a fundamental transformation in central bank communication: statements have approximately tripled in length since 1990, with

the expansion accelerating markedly after the Global Financial Crisis, while readability has modestly improved as central banks offset greater informational content with simpler sentence structures (Figure 5). In addition, inflation references comove substantially across countries during episodes of elevated global inflationary pressure, with persistent cross-country differences reflecting heterogeneity in mandates, inflation dynamics, and policy frameworks (Figure 9).

Our empirical application investigates the communication channel of international policy transmission, focusing on the GFC. Using LLMs for question answering and aspect-based sentiment analysis, we document that central banks typically attribute global financial conditions to broad U.S. economic factors rather than specifically to Federal Reserve policy actions. Sentiment analysis reveals greater concern about growth and financial conditions than about monetary policy spillovers in the post-2008 period. Granger causality tests based on both a benchmark dictionary-based sentiment index and LLM-based indicators (Figures 13 and 19) suggest that statement sentiment may drive components of the GFC rather than merely respond to them.

Regional variations in central bank sentiment (Figure 14) reveal important heterogeneity in how central banks perceive and react to global financial conditions, with emerging-market central banks showing greater sensitivity to capital flow concerns and exchange rate pressures relative to their advanced-economy counterparts. These differences indicate that the transmission of global financial shocks depends significantly on institutional factors and the degree of financial integration, carrying important implications for international policy coordination. The comparative analysis of LLM-generated and lexical sentiment indices (Figure 12) illustrates both the potential and the limitations of advanced language models for economic research, underscoring the value of methodological pluralism in computational text analysis.

These findings contribute to the understanding of monetary policy transmission in an interconnected global economy and underscore the importance of the communication channel as a mechanism through which monetary policy transmits across borders. The systematic approach to text analysis developed here offers a replicable template for applying frontier NLP methods to economic research. By making the MPSD and all associated code publicly available, we aim to support future research on monetary policy communication, financial market reactions, and policy coordination. As central banks continue to refine their communication strategies, rigorous empirical analysis of these communications will remain essential for understanding how policy intentions translate into economic outcomes.

7 Acknowledgments

We are grateful to two anonymous referees and the Editor for their constructive comments and suggestions, which substantially improved this paper. We also thank participants at the 4th IFC and Bank of Italy Workshop on Data Science in Central Banking for helpful discussions. All remaining errors are our own.

8 Code and Data Availability

The full codebase used for data processing, analysis, and figure generation, together with the compiled Monetary Policy Statement Database, excluding the scraping code,[5] is publicly available at `https://github.com/CentralBankTexts/monetary-policy-statement-database`. The repository is released under the MIT License, allowing broad reuse while ensuring proper attribution. A permanent archived version of the repository, to ensure long-term accessibility and citability consistent with standard reproducibility practices, will be archived with a permanent DOI upon final publication.

A Appendix

A.1 LLM Sentiment Prompt

```
instruction: |
  You will be analyzing a text from a central bank. The text is provided
   ↪ below:

  <central_bank_text>
  {{CENTRAL_BANK_TEXT}}
  </central_bank_text>

  Your task is to identify discussions related to the impact of global
   ↪ financial conditions on domestic economies. Focus on the following
   ↪ themes:
  1. Economic growth
  2. Financial conditions
  3. Capital flows
  4. Monetary policy spillovers

  Carefully read through the text. For each theme explicitly discussed,
   ↪ assign a continuous sentiment score on a scale from 0.0 to 2.0.
```

[5] While institution-specific scrapers are not distributed, the compiled corpus and full analysis code enable reproduction of all reported results.

```
  Scoring Guidelines:
  * 0.0 to < 1.0 (Negative): The text suggests unfavorable impacts, risks,
   ↪  contraction, or tightening.
      * Closer to 0.0: Severe risks, crises, or strong negative shocks.
      * Closer to 1.0: Mild headwinds or slight downside risks.
  * 1.0 (Neutral): The text presents a balanced view, negligible impact,
   ↪  or uncertainty with no clear direction.
  * > 1.0 to 2.0 (Positive): The text suggests favorable impacts,
   ↪  expansion, loosening, or resilience.
      * Closer to 1.0: Mild tailwinds or slight improvements.
      * Closer to 2.0: Strong growth, robustness, or highly beneficial
       ↪  conditions.

  Output Rules:
  * Only include themes that are explicitly discussed.
  * Use floating-point numbers to capture nuance (e.g., 0.45, 1.2, 1.85).
  * Format the output as a Python list of lists, where each inner list
   ↪  contains the theme string and the score.

  Your final output should be a Python list containing only the
   ↪  theme-score pairs. Do not provide any additional commentary.

  Example Output:
  [["economic_growth", 1.8], ["financial_conditions", 0.45],
   ↪  ["capital_flows", 1.0]]

output_format: |
  A Python list of lists containing theme and float score pairs (e.g.,
   ↪  [["economic_growth", 1.25], ["financial_conditions", 0.8]]).
```

A.2 LLM Narrative Prompt

```
instruction: |
  The following texts come from central banks. The country is identified
   ↪  in the Country: and the Comment: is the text released by the central
   ↪  bank.
  I want you to identify the main driver of global economic and financial
   ↪  conditions and the origin country driving these conditions.
  In the summary please note if any central banks identify the Federal
   ↪  Reserve as the driver of global economic or financial conditions.

output_format: |
  Please respond in the following format:
  Driver: [Identified Driver]
  Origin Country: [Origin Country]
```

```
Summary: [brief explanation]
```

A.3 Language Groups Analysis

We distinguish between monetary policy statements originally drafted in English and those translated into English from other languages. Figure 15 reports the evolution of linguistic indicators separately for these two groups, allowing us to assess whether observed patterns are driven by language or translation effects. For statements originally drafted in English, the results closely mirror the full-sample trends. Average word counts increase markedly over time, while lexical diversity, as measured by the type-token ratio, declines. At the same time, readability exhibits only a modest downward trend, indicating that although statements have become substantially longer, they have not become proportionally more complex in terms of sentence structure and word choice. Statements translated into English display broadly similar dynamics. Despite differences in levels, the time-series trends closely track those observed for native-English statements: translated texts become longer over time, show declining lexical diversity, and maintain relatively stable readability. Average adjective length remains largely unchanged in both groups, suggesting that the expansion of communication occurs primarily through increased volume rather than more complex evaluative wording. The similarity of trends across native-English and translated-English statements suggests that the main results are not driven by translation-related measurement artifacts. Instead, they reflect systematic changes in central bank communication practices over time. Persistent level differences across groups are more plausibly attributed to institutional drafting conventions and communication styles rather than language effects per se.

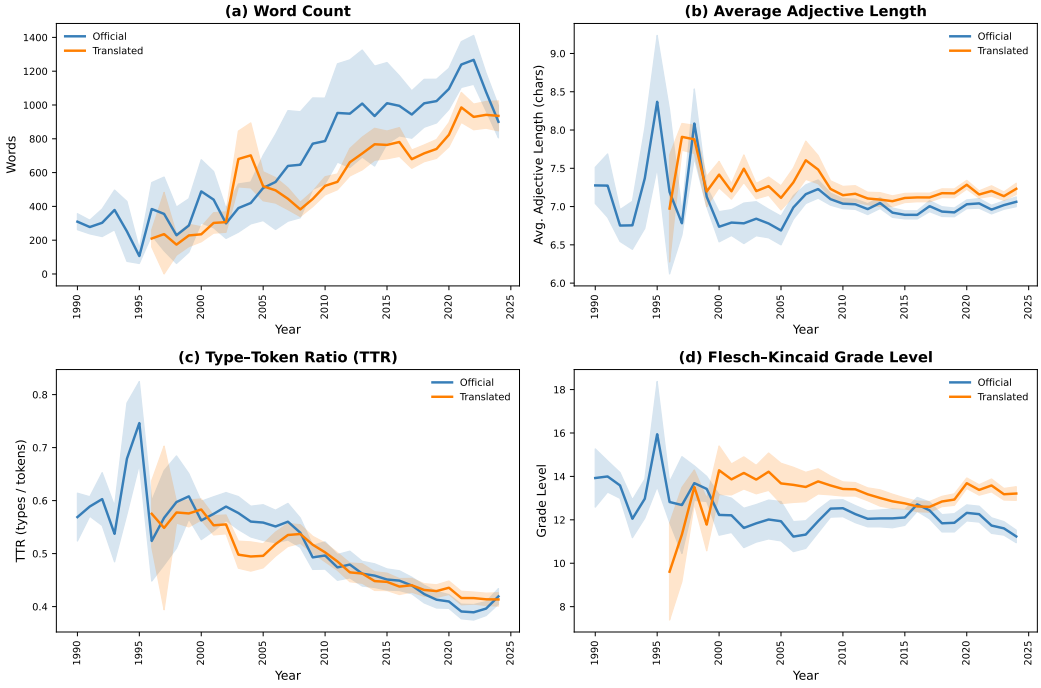


Figure 15: Average Statement Length and Readability Measure by Language Groups

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: For each central bank statement, we compute four document-level linguistic indicators for countries with English as official language and English as translated language: (i) word count, defined as the total number of tokens; (ii) average adjective length, measured as the mean character length of adjectival tokens identified using part-of-speech tagging; (iii) the type–token ratio (TTR), defined as the number of unique tokens divided by total tokens; and (iv) readability, captured using standard readability indices (e.g., Flesch-Kincaid Grade Level). Each indicator is first computed at the statement level and then aggregated to the annual frequency by taking the cross-statement mean. To capture uncertainty and heterogeneity across institutions and over time, we report confidence bands constructed from the standard error of the annual mean.

A.4 Statement Sentence Count

We observe that unusually long "statements" in the corpus are concentrated in countries where the policy-rate decision is communicated through a broader analytical document rather than a short press release (see Figure 16). For example, in Pakistan (PK) the rate decision has often been published within, or together with, a Monetary Policy Statement/Report that includes detailed discussion of global and domestic economic developments and the outlook. A similar pattern appears in several other countries as well, such as Nigeria (NG) or Ghana (GH), for example, where MPC statements typically combine the decision with an extended review of macroeconomic conditions and risks. In India (IN), statement length is amplified by the release of a policy package around each meeting (e.g., MPC resolution together with accompanying policy statements), which can be captured as a single text. Large sentence-count variation primarily reflects institutional differences in communication formats and document bundling.

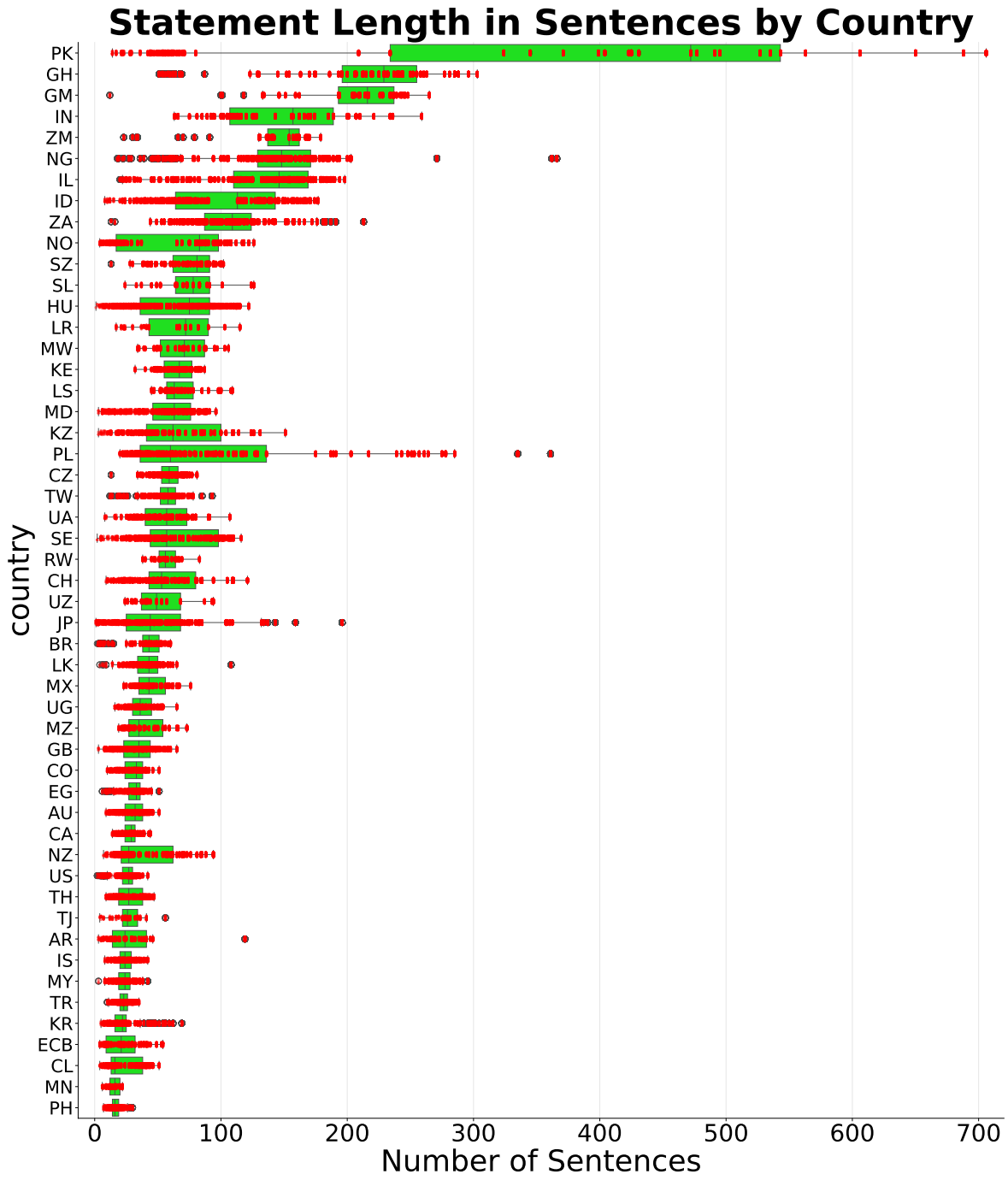


Figure 16: Statement Sentence Count

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.

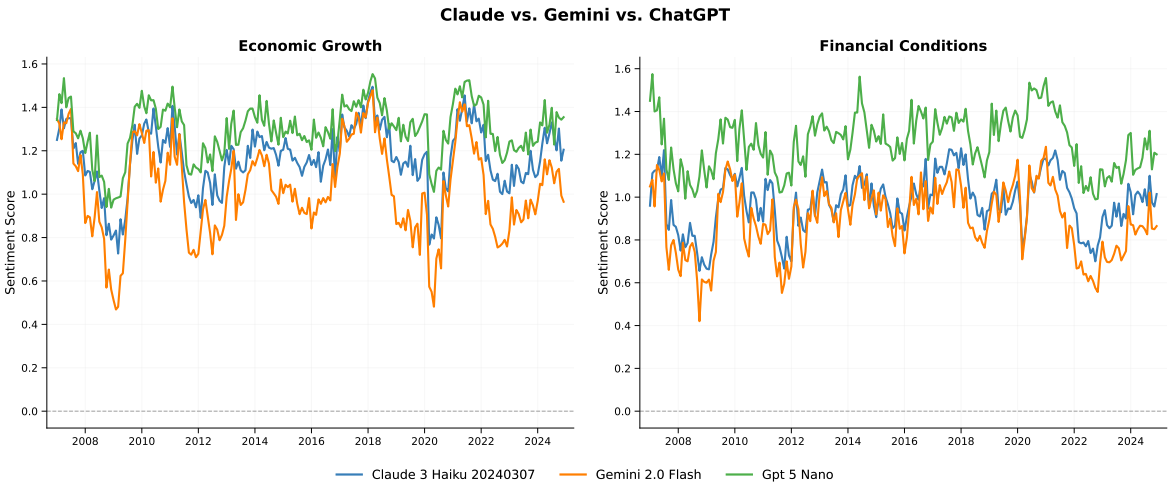


Figure 17: Cross-Model LLM Sentiment Comparison

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.

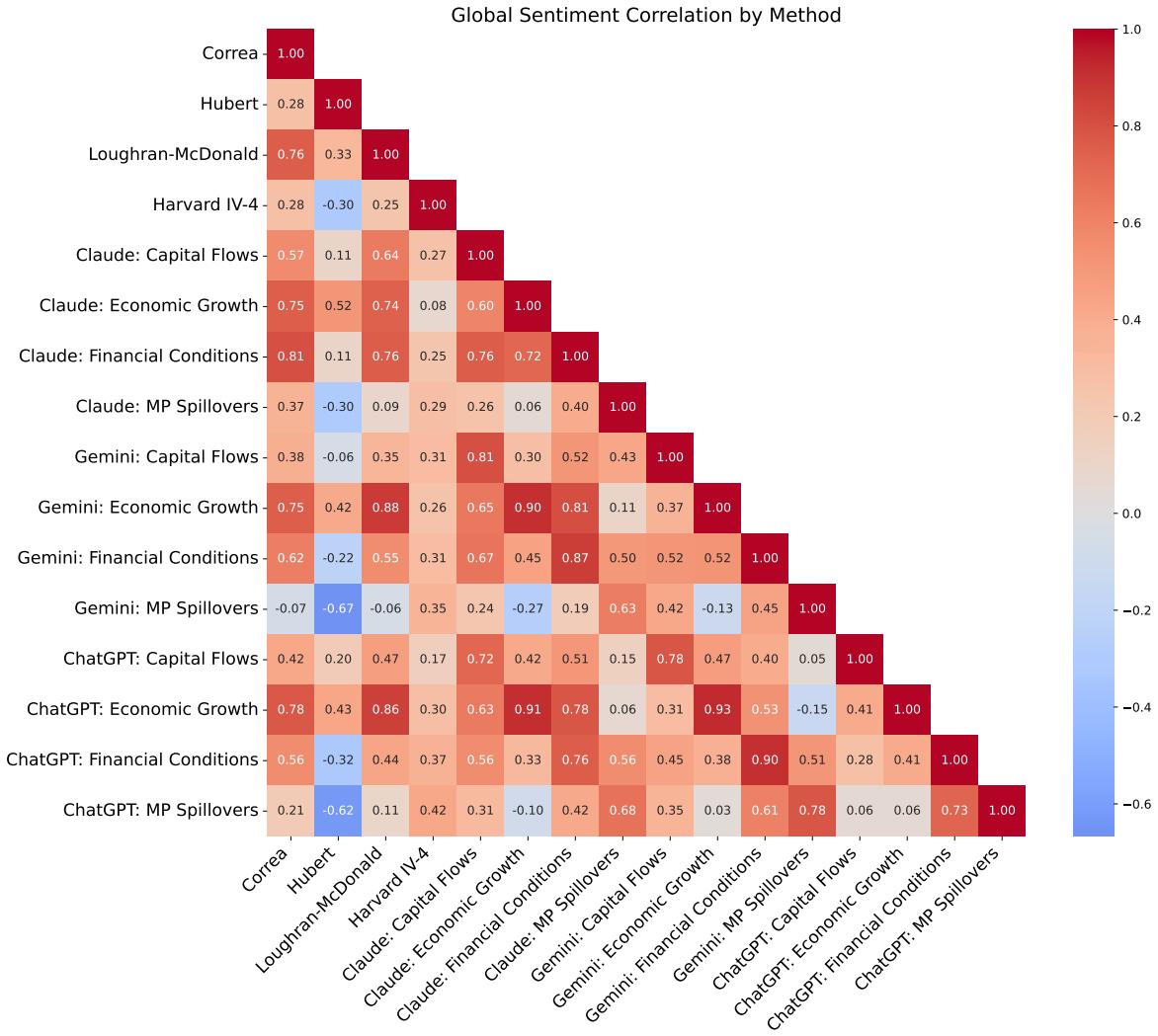


Figure 18: Extended Correlation Matrix of Sentiment Measures

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD) and using LLMs.

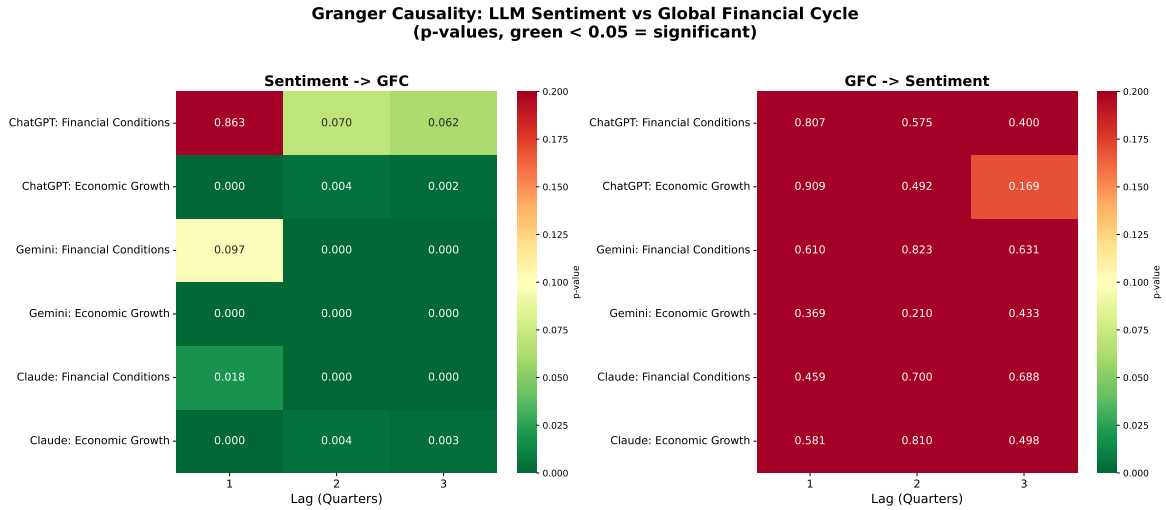


Figure 19: LLM-Based Sentiment Granger-Causes the GFC

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: Sentiment measures are derived from MPSD statements using multiple LLMs (ChatGPT, Gemini, and Claude) and capture ABS related to economic growth and financial conditions. The GFC is from Miranda-Agrippino and Rey (2020). The figure reports p-values (rounded to three decimal places)
from Granger causality tests at quarterly frequency. Green cells indicate statistical significance at the 5% level. Tests are conducted for lags of one to three quarters.

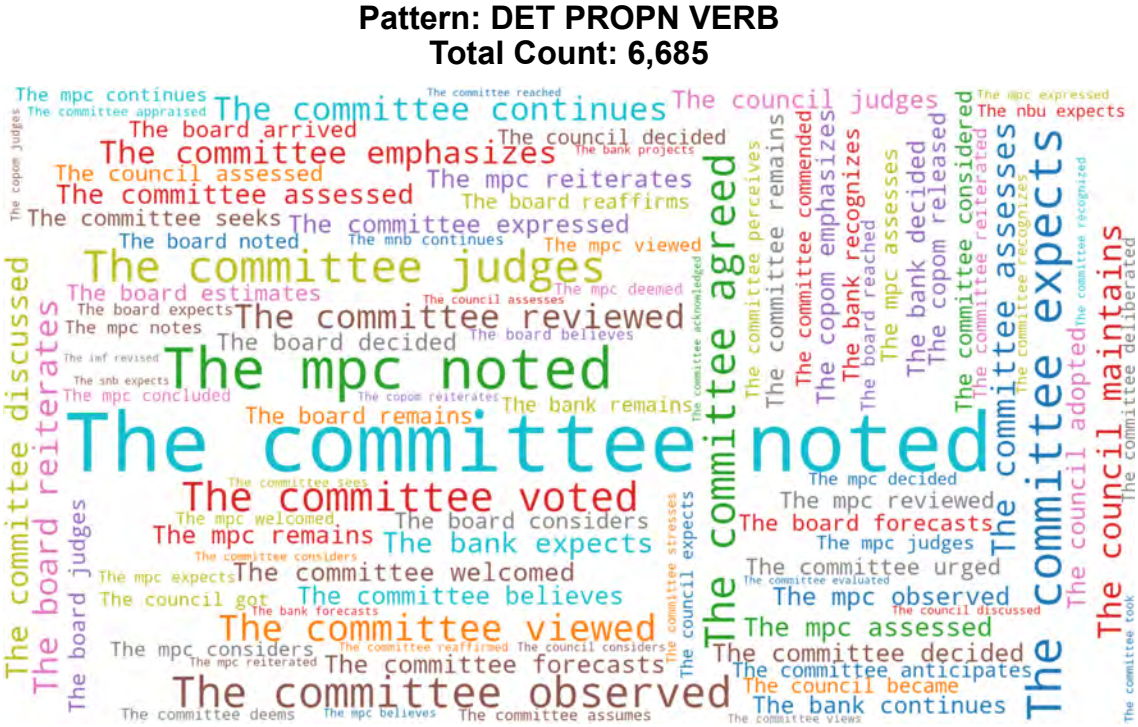


Figure 20: Most Common Three-Token Sentences by Part-of-Speech Pattern: `DET PROPN VERB`. The `DET PROPN VERB` pattern often forms the root of sentences associated with CB.

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

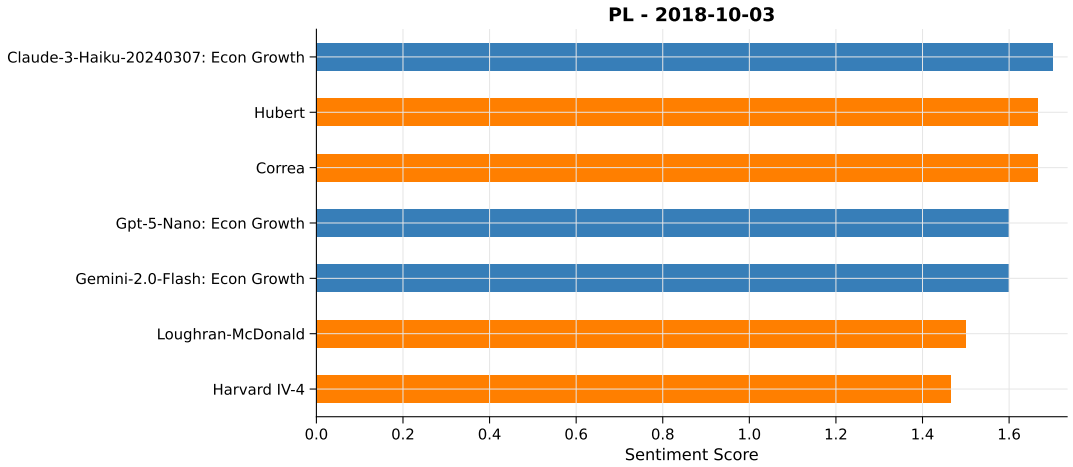


Figure 21: Sentiment Variation Across Example Statement, Poland 2018-10-03

Source: Authors' calculations based on the Monetary Policy Statement Database (MPSD).

Notes: This figure illustrates the qualitative validation approach by displaying sentiment variation across methods for the same policy statement

References

Acosta, M. and E. E. Meade (2015). Hanging on every word: semantic analysis of the FOMC's postmeeting statement. FEDS Notes 2015-09-30, Board of Governors of the Federal Reserve System.

Armelius, H., C. Bertsch, I. Hull, and X. Zhang (2020). Spread the Word: International spillovers from central bank communication. *Journal of International Money and Finance 103*, 102116.

Assenmacher, K., G. Glöckler, S. Holton, P. Trautmann, D. Ioannou, S. Mee, K. Bakk-Simon, S. Bergbauer, M. Catenaro, E. Charalampakis, M. Ehrmann, G. Ferrero, D. Georgarakos, P. Gertler, A. Giovannini, R. Grandia, N. Hernborg, N. Herrala, D. Kedan, G. Kenny, T. Linzert, M. Manrique, R. Mestre, E. Mönch, S. Nardelli, N. Nomm, L. Pavlova, G. Schultefrankenfeld, M. Silgoner, I. Skotida, B. Winkler, M. T. Bitterlich, E. Argiri, C. Alonso, J. Byron, F. Arigoni, M. Deroose, M. Gardt, K. Istrefi, O. Goldfayn-Frank, J. Hellström, P. Huertgen, K. Kalnberzina, G. Kocharkov, V. Márquez, C. Ruhe, M. Šanta, R. Reedik, G. Sciot, A. Stylianou, and E. Taylor (2021). Clear, consistent and engaging: ECB monetary policy communication in a changing world. Technical Report 274, ECB Occasional Paper.

Bauer, M. D. and E. T. Swanson (2023). An alternative explanation for the "Fed Information Effect". *American Economic Review 113*(3), 664–700.

Benchimol, J., I. Caspi, and S. Kazinnik (2023). Measuring communication quality of interest rate announcements. *Economists' Voice 20*(1), 43–53.

Benchimol, J., S. Kazinnik, and Y. Saadon (2020). Communication and transparency through central bank texts. Paper presented at the 132nd Annual Meeting of the American Economic Association, January 3-5, 2020, San Diego, CA, United States.

Benchimol, J., S. Kazinnik, and Y. Saadon (2022). Text mining methodologies with R: an application to central bank texts. *Machine Learning with Applications 8*, 100286.

Benchimol, J., S. Kazinnik, and Y. Saadon (2025). Federal Reserve Communication and the COVID-19 Pandemic. *Manchester School 93*(5), 464–484.

Blinder, A. (2009). Talking about monetary policy: the virtues (and vice?) of central bank communication. BIS Working Papers 274, Bank for International Settlements.

Blinder, A., C. Goodhart, P. Hildebrand, D. Lipton, and C. Wyplosz (2001). *How Do Central Banks Talk?* Number 3 in Geneva Reports on the World Economy. London, UK: Centre for Economic Policy Research.

Blinder, A. S. (2018). Through a crystal ball darkly: The future of monetary policy communication. *AEA Papers and Proceedings 108*, 567–571.

Blinder, A. S., M. Ehrmann, J. de Haan, and D.-J. Jansen (2024). Central bank communication with the general public: promise or false hope? *Journal of Economic Literature 62*(2), 425–457.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. D. Haan, and D.-J. Jansen (2008). Central bank communication and monetary policy: a survey of theory and evidence. *Journal of Economic Literature 46*(4), 910–945.

Born, B., M. Ehrmann, and M. Fratzscher (2014). Central bank communication on financial stability. *Economic Journal 124*(577), 701–734.

Bruno, V. and H. S. Shin (2015). Capital flows and the risk-taking channel of monetary policy. *Journal of Monetary Economics 71*, 119–132.

Brusa, F., P. Savor, and M. Wilson (2020). One Central Bank to Rule Them All. *Review of Finance 24*(2), 263–304.

Bulír, A., M. Cihák, and D.-J. Jansen (2013). What drives clarity of central bank communication about inflation? *Open Economies Review 24*(1), 125–145.

Campiglio, E., J. Deyris, D. Romelli, and G. Scalisi (2025). Warning words in a warming world: Central bank communication and climate change. *European Economic Review 178*, 105101.

Cieslak, A. and A. Schrimpf (2019). Non-monetary news in central bank communication. *Journal of International Economics 118*, 293–315.

Coibion, O., Y. Gorodnichenko, and M. Weber (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy 130*(6), 1537–1584.

Correa, R., K. Garud, J. M. Londono, and N. Mislang (2021). Sentiment in central banks' financial stability reports. *Review of Finance 25*(1), 85–120.

Davis, J. S. and M. A. Wynne (2019). Central Bank Communications: A Case Study. In *Oxford Handbook of the Economics of Central Banking*. Oxford University Press.

De Grauwe, P. and Y. Ji (2023). On the use of current and forward-looking data in monetary policy: a behavioural macroeconomic approach. *Oxford Economic Papers 75*(2), 526–552.

de Haan, J. and L. Hoogduin (2024). ECB communication policies: An overview and comparison with the Federal Reserve. *Journal of International Money and Finance 142*, 103050.

Dincer, N., B. Eichengreen, and P. Geraats (2022). Trends in monetary policy transparency: further updates. *International Journal of Central Banking 18*(1), 331–348.

Ehrmann, M. and A. Wabitsch (2022). Central bank communication with non-experts - A road to nowhere? *Journal of Monetary Economics 127*, 69–85.

Eijffinger, S. and D. Masciandaro (2014). *Modern Monetary Policy and Central Bank Governance*. Cheltenham, UK: Edward Elgar Publishing.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gonzalez, M. and R. C. Tadle (2022). Monetary policy press releases: an international comparison. BIS Working Papers 1023, Bank for International Settlements.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Hansen, S. and M. McMahon (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics 99*(S1), 114–133.

Hansen, S., M. McMahon, and M. Tong (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics 108*(C), 185–202.

Hayo, B. and M. Neuenkirch (2015). Central bank communication in the financial crisis: Evidence from a survey of financial market participants. *Journal of International Money and Finance 59*(C), 166–181.

Issing, O. (2019). *The long journey of central bank communication*. Cambridge, MA: MIT Press.

Kahn, G. A. (2007). Communicating a policy path: the next frontier in central bank transparency? *Economic Review-Federal Reserve Bank of Kansas City 92*(1), 25.

Kalemli-Özcan, Ş. (2019). U.s. monetary policy and international risk spillovers. In *Proceedings - Jackson Hole Economic Symposium*, Jackson Hole, WY, pp. 95–191.

Kryvtsov, O. and L. Petersen (2021). Central bank communication that works: Lessons from lab experiments. *Journal of Monetary Economics 117*(C), 760–780.

Kumar, A., S. K. Mallick, M. Mohanty, and F. Zampolli (2023). Market volatility, monetary policy and the term premium. *Oxford Bulletin of Economics and Statistics 85*(1), 208–237.

Lewis, D., C. Makridis, and K. Mertens (2020). Do monetary policy announcements shift household expectations? CEPR Discussion Papers 14360, Centre for Economic Policy Research.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance 66*(1), 35–65.

Macklem, T. and J. Vardy (2023). 20 years of central bank communications, and lessons for the future. In *Central banking in the Americas: Lessons from two decades*, pp. 55–67. Bank for International Settlements.

Masciandaro, D., O. Peia, and D. Romelli (2024). Central bank communication and social media: from silence to Twitter. *Journal of Economic Surveys 38*(2), 365–388.

Miranda-Agrippino, S. and H. Rey (2020). U.S. monetary policy and the Global Financial Cycle. *Review of Economic Studies 87*(6), 2754–2776.

Miranda-Agrippino, S. and H. Rey (2022). The global financial cycle. In *Handbook of International Economics*, Volume 6, pp. 1–43. Elsevier.

Morris, S. and H. S. Shin (2002). Social value of public information. *American Economic Review 92*(5), 1521–1534.

Nakamura, E. and J. Steinsson (2018). High-frequency identification of monetary non-neutrality: the information effect. *Quarterly Journal of Economics 133*(3), 1283–1330.

Obstfeld, M. (2015). Trilemmas and tradeoffs: living with financial globalization. In C. Raddatz, D. Saravia, and J. Ventura (Eds.), *Global Liquidity, Spillovers to Emerging Markets and Policy Responses*, Volume 20, Chapter 2, pp. 013–078. Central Bank of Chile.

Picault, M. and T. Renault (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance 79*, 136–156.

Rey, H. (2015). Dilemma not trilemma: the global financial cycle and monetary policy independence. NBER Working Papers 21162, National Bureau of Economic Research, Inc.

Vyshnevskyi, I., W. Jombo, and W. Sohn (2024). The clarity of monetary policy communication and financial market volatility in developing economies. *Emerging Markets Review 59*, 101121.

Woodford, M. (2005). Central bank communication and policy effectiveness. *Proceedings - Economic Policy Symposium - Jackson Hole Aug*, 399–474.

The Monetary Policy Statement Database

An LLM Application to Global Financial Conditions*

Cory Baird[1] Jonathan Benchimol[2] Wook Sohn[3] Vira Vyshnevska[3] Iegor Vyshnevskyi[4]

[1]University of Tokyo, Tokyo, Japan
Corresponding author. Email: cbaird@g.ecc.u-tokyo.ac.jp

[2]Bank of Israel, Jerusalem, Israel

[3]KDI School, Sejong, South Korea

[4]Sogang University, Seoul, South Korea

February 18, 2025

This presentation does not necessarily reflect the views of the Bank of Israel or any other institutions.

*Code and data: `github.com/CentralBankTexts/monetary-policy-statement-database`

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Not Just a Database: Reproducible AI for Macroeconomics

- **Scientific Rigor**
  Infrastructure to maintain scientific standards in text analysis.

- **Unprecedented Scale**
  6,693 policy rate statements from 51 central banks worldwide.

- **Cutting-Edge Infrastructure**
  Transparent, automated, and scalable pipelines.

- **Applying AI to Macroeconomics**
  Leverages AI to identify economic narratives and sentiment over time and across countries.
  - Global Financial Cycle origins and drivers (Rey, 2015).

Current Literature

Key Limitations in CB Communication Research:

- **Data & Standardization:** Existing research focuses on a few countries or uses only BIS speeches (Born et al., 2014; Picault and Renault, 2017).
  - Contribution: standardized dataset of 51 CBs
    Larger than existing non-public databases, e.g., Gonzalez and Tadle (2022)

- **Reproducibility:** Lack of open-source pipelines impedes reproducible research (Gentzkow et al., 2019; Hansen and McMahon, 2016).
  - Contribution: transparent and accessible text cleaning, and Large Language Model (LLM) pipelines

Applying the Scientific Method to "AI" and NLP

- Failure to share data/algorithms hinders reproducibility and efficiency
  - **Can reproduce** lexical/dictionary methods at immense cost
  - **Cannot reproduce** black box LLMs (Claude, Gemini) because data, prompts, and models must be **exactly** the same

- **Irreproducible NLP methods** hinder scientific progress and data reliability in downstream modeling

Data Collection	**Featurization**	**Replicable AI Use**
Identify statement links	Cleaning	Replicable LLM infrastructure
Scrape statements	Sentence segmentation	Track inputs (prompts, texts)
Automate retrieval	Part-of-speech tagging	Track outputs (responses)

Table 1: End-to-End Transparency for Replicable AI

Source: Authors.

Infrastructure Overview

- **Transparency**
  - All code/scripts made public on GitHub[1]
  - Technologies: GitHub, Pyenv, Poetry
- **Automation**
  - Scripts executed regularly via a Data Version Control (DVC) pipeline
  - Technologies: DVC, GitHub Actions.
- **Scalability**
  - Modular scripts easily adaptable for other CB texts
  - Technologies: Python Object Oriented Programming, GitHub
- **Reproducibility**
  - Transparent text analysis
  - Reproducible pipelines for cleaning, featurization, and LLMs
  - Technologies: UV, Pyenv, Poetry, S3, DVC

[1] `github.com/CentralBankTexts/monetary-policy-statement-database`

MPSD Overview

- Monetary Policy Statement Database (MPSD)
  - **Scope**
    51 CBs (35 API and 16 static currently cleaned)
  - **Data**
    Approximately 6,693 unique MPSs
  - **Temporal Coverage**
    1990–2024

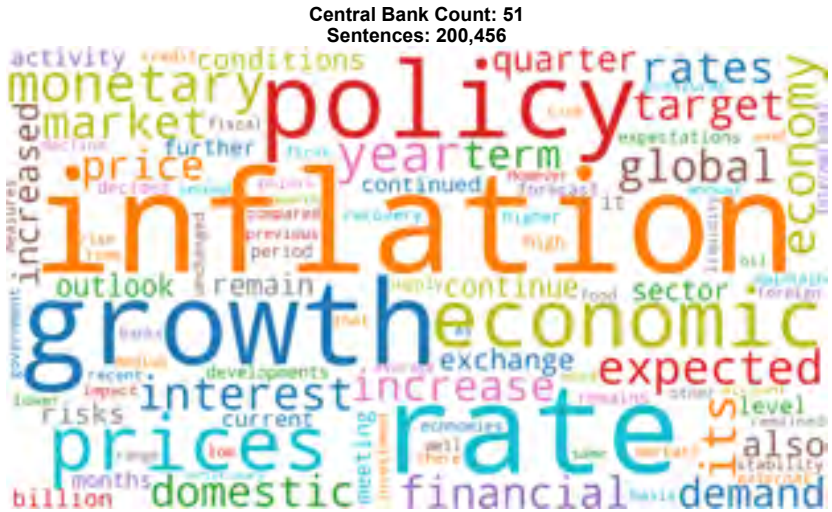


Figure 1: Standardized and diverse corpus for training/testing A.I.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Technical Infrastructure: Code & Data Management

- **Data integrity and reproducibility**
  - Data Version Control (DVC) tracks full lifecycle: web scraping $\rightarrow$ LLM analysis
  - Parameterized pipelines enforce data lineage at every stage
- **Code and environment management**
  - Code versioning: Git and GitHub
  - Python version and dependency management: uv
  - Cloud storage: AWS S3 (raw, processed, and artifact partitions)
- **Automated updates**
  - GitHub Actions: daily checks for new CB policy statements
  - Single-command replication of data, models, and outputs
- **LLM infrastructure**
  - Abstract LLM class integrating Claude, ChatGPT, Gemini, and Hugging Face
  - Methods: model selection, YAML prompt processing, response generation/saving

Code/Folder Infrastructure: Automated & Scalable

- **Modular Design:** Programs/scripts are modularized for:
  1. Text retrieval
  2. Text cleaning
  3. LLM inference
- **Standardized Folder Structure:**

  `src` : Python scripts
  `data` : Text files
  `references` : GenAI prompts
- **MPSD Updates:** Automated updates using DVC.
- **End-to-End Reproducibility:** A full pipeline ensures reproducible results from data acquisition to NLP analysis.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

The Rise of Central Bank Communication

- CB communication dramatically changed since the 1990s
  - First official MPS: Reserve Bank of Australia in 1990
  - Rapid growth in MPSs after the year 2000
- MPSs are a crucial tool for:
  - Managing inflation expectations
  - Enhancing policy predictability
  - Building credibility and trust

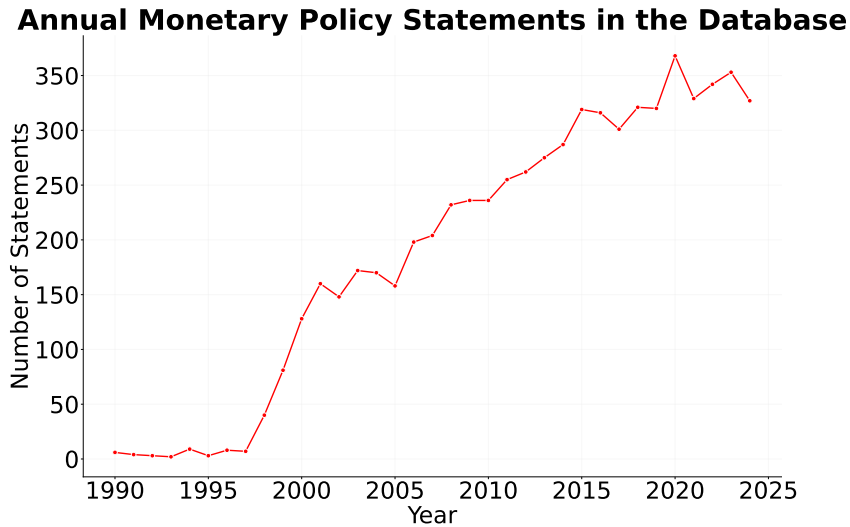


Figure 2: Annual MPSs in the MPSD

Dataset Coverage

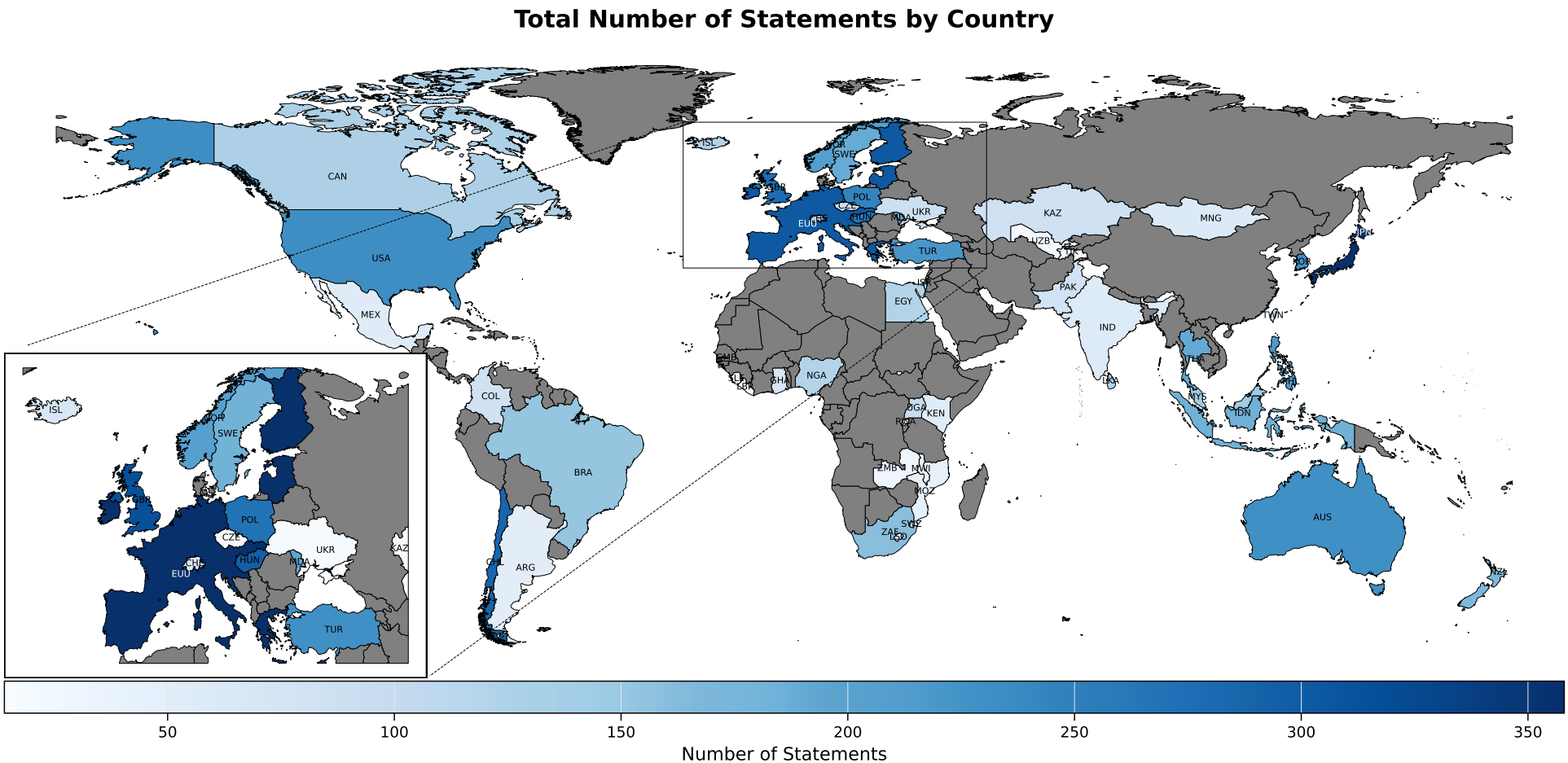


Figure 3: Countries and Monetary Policy Statements

Dataset Coverage

- The BoJ (Japan) and ECB emerge as the most prolific publishers of statements
- Central banks in middle-income countries exhibit more limited publication frequency and shorter historical coverage
- These cross-institutional variations reflect differences in institutional frameworks and communication strategies

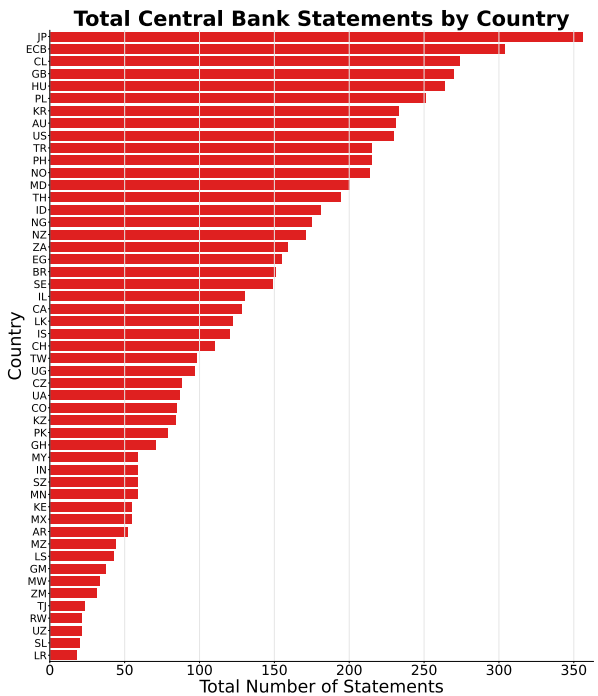


Figure 4: Statements by Country

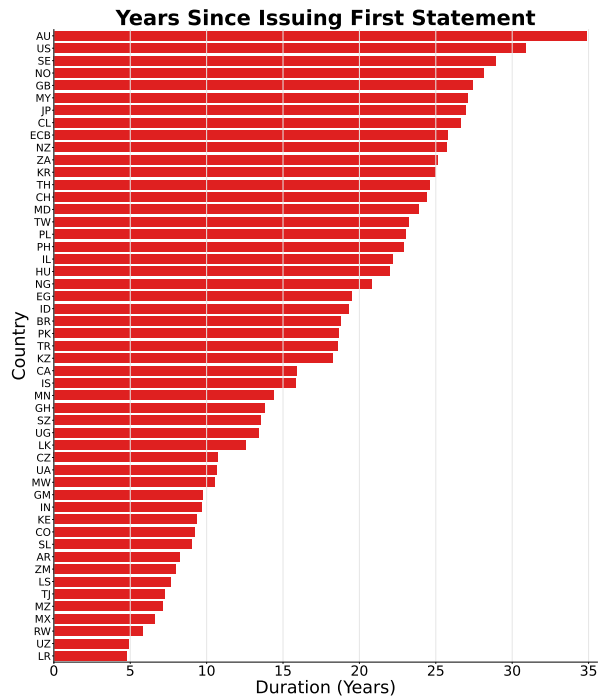


Figure 5: Years of Coverage by Country

Trends in Communication Over Time

- MPSs growing lengthier and more complex post-GFC
- MPSs becoming slightly more "readable" over time despite increased length

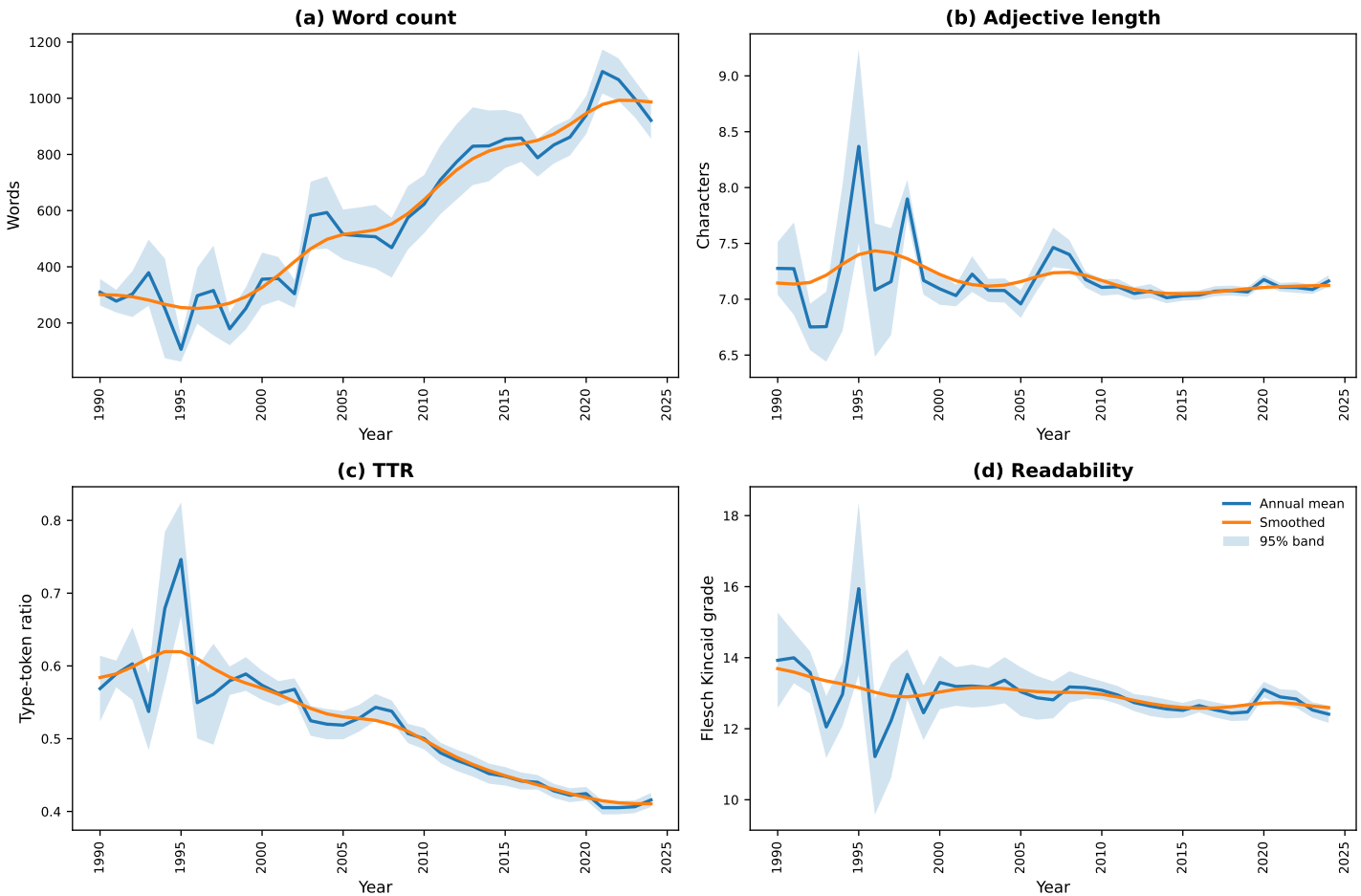


Figure 6: Linguistic Trends in CB Statements Over Time

Key Characteristics (Patterns) of the MPS

Economic Theme	**POS Pattern**	**Example**
Inflation Forecasts	inflation + AUX + VERB	inflation **will rise**; inflation **might persist**
Growth Projections	growth/GDP + AUX + VERB	growth **should recover**; GDP **may stall**
Inflation Dynamics	NOUN + ADP + inflation	risk **of** inflation; outlook **for** inflation
Rate Levels	rate + ADP + NUM	rate **at** 5.25%; rate **above** 2.0
Forward Guidance	policy/rate + AUX + VERB	policy **will remain**; rates **might stay**
Monetary Stance	policy + AUX + ADJ	policy **is restrictive**; stance **remains loose**
Fiscal Policy	NOUN + ADP + spending/deficit	impact **of** deficit; boost **from** spending
External Conditions	NOUN + ADP + trade/export	demand **for** exports; uncertainty **about** trade

Table 2: Linguistic Patterns in Monetary Policy Statements by Economic Theme

Most Common Sentence Patterns by Part-of-speech

- Common sentence patterns include:
  - `DET ADJ NOUN` (e.g., "The strong dollar," "The current rate")
  - `DET NOUN ADP` (e.g., "The impact of," "The level of")
  - `ADP DET NOUN` (e.g., "In the market," "On the economy")
- Sentences identified with spaCy's Universal POS tags and pattern matching.

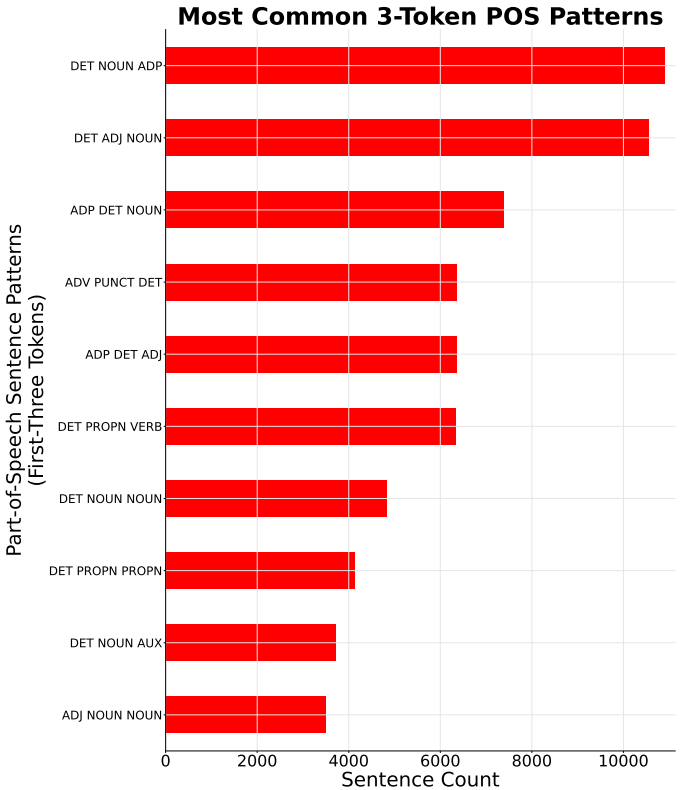


Figure 7: Most Common Patterns by Part-of-Speech (POS)

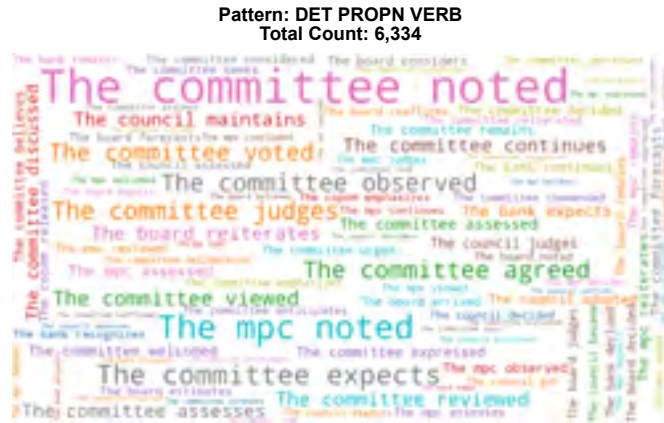


Figure 8: POS: The `DET PROPN VERB` pattern is associated with CB commentary/actions

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Addressing AI Challenges with the Corpus/Infrastructure

- **Challenges with LLMs**
  1. Irreproducibility: Insufficient transparency regarding model inputs and specifications makes replicating it difficult.
     - We provide open-source code, including LLM code, prompts, and Jupyter notebooks, ensuring full transparency and reproducibility.
  2. High computational costs and energy consumption.
     - We implement text filtering (keyword & part-of-speech) pipelines to reduce the computational burden.

- **Application to Econ: The Global Financial Cycle (GFC)**
  - Communication Channel: CB Communication $\Rightarrow$ Market Expectations $\Rightarrow$ Financial Conditions
  - Using our LLM/Data Pipeline we leverage LLMs for Question Answering (QA) & classification (topic, sentiment) tasks to explore origins and drivers of the Global Financial Cycle (Miranda-Agrippino and Rey, 2020, 2022)

Leveraging Our LLM/Data Pipeline

- Infrastructure is abstract for scalability
  - LLM: seamlessly add more models (ChatGPT, Huggingface, etc) and tasks (classification, summarization, Q&A)
  - Text retrieval: seamlessly add more central banks and types of communications (minutes, speeches, etc)

Modular LLM-based Data Processing Framework for Academic Research		
Data Acquisition Module (Abstract Class)	**LLM Processing Module (Abstract Class)**	**LLM Data Pipeline Workflow**
Methods: 1. Initialize (source parameters) 2. Scrape Statement Links 3. Download/Save Statements **Output:** Text Files CSV/Parquet Text Dataframes **Abstraction:** Extend abstract functions to accommodate new data sources (e.g., central bank speeches, minutes) and communication modes.	**Methods:** 1. Initialize (LLM Model) 2. Process/Load Prompts 3. Generate Response **Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) **Abstraction:** Implement model-agnostic functions to leverage diverse LLM architectures (e.g., Gemini, Claude, HuggingFace models).	**Methods:** 1. Initialize (Input Corpus) 2. Filter via Keywords or POS 3. LLM Inference 4. Save Response **Output:** Response (Text/JSON) Prompt (Text/JSON) Model Parameters (YAML) Figure for Research (PNG/PDF) **Use-Case:** Utilize concise, reusable scripts for seamless integration with LLM APIs (e.g., Gemini, Claude) by researchers.

Table 3: AI Infrastructure for Central Bank Communication Analysis

LLM Application: Exploring the Global Financial Cycle

- LLM Application: Question Answering (QA)
  - Prompts model to summarize drivers of global economy

    "The following texts come from central banks. [. . .] I want you to identify the main driver of global economic and financial conditions and the origin country driving these conditions. In the summary please note if any central banks identify the Federal Reserve as the driver of global economic or financial conditions."

- LLM Application: Aspect Based Sentiment[2](ABS)
  - Prompts model to extract sentiment for specific themes on a continuous scale (0.0–2.0)

    "Your task is to identify discussions related to the impact of global financial conditions on domestic economies. Focus on the following themes: 1. Economic growth, 2. Financial conditions, 3. Capital flows, 4. Monetary policy spillovers. For each theme explicitly discussed, assign a continuous sentiment score on a scale from 0.0 to 2.0, where 1.0 is neutral."

- Findings from LLM Analysis:
  - QA: Central Banks often cite the U.S. economy as a driver of global economic conditions rather than specifically the U.S. Federal Reserve.
  - ABS: Sentiment analysis shows that central banks expressed more concern about growth and financial conditions than monetary policy and capital flows in the post-2008 period

Appendix: Full Prompt Example

[2]Aspect-based sentiment identifies sentiment of specific theme; see Appendix A.1–A.2 in the paper for full prompts

LLM Application: QA Results Differ by Model (2008)

Model	**Mechanism**	**Origin**	**Key Finding**
Claude	The Federal Reserve	United States	Names the Fed as key driver; US housing bust a recurring theme
Xiaomi	Global financial market turmoil / subprime crisis	United States	Fed acknowledged by multiple CBs as key driver
Mistral	US economy slowdown / subprime crisis	United States	No CB explicitly names the Fed as the driver
Arcee	US subprime crisis and broader US economic weakness	United States	Fed's policy actions cited as relevant but not the primary driver

LLM summaries of the same CB sentences differ in how prominently they identify the Federal Reserve

LLM Application: Aspect Based Sentiment

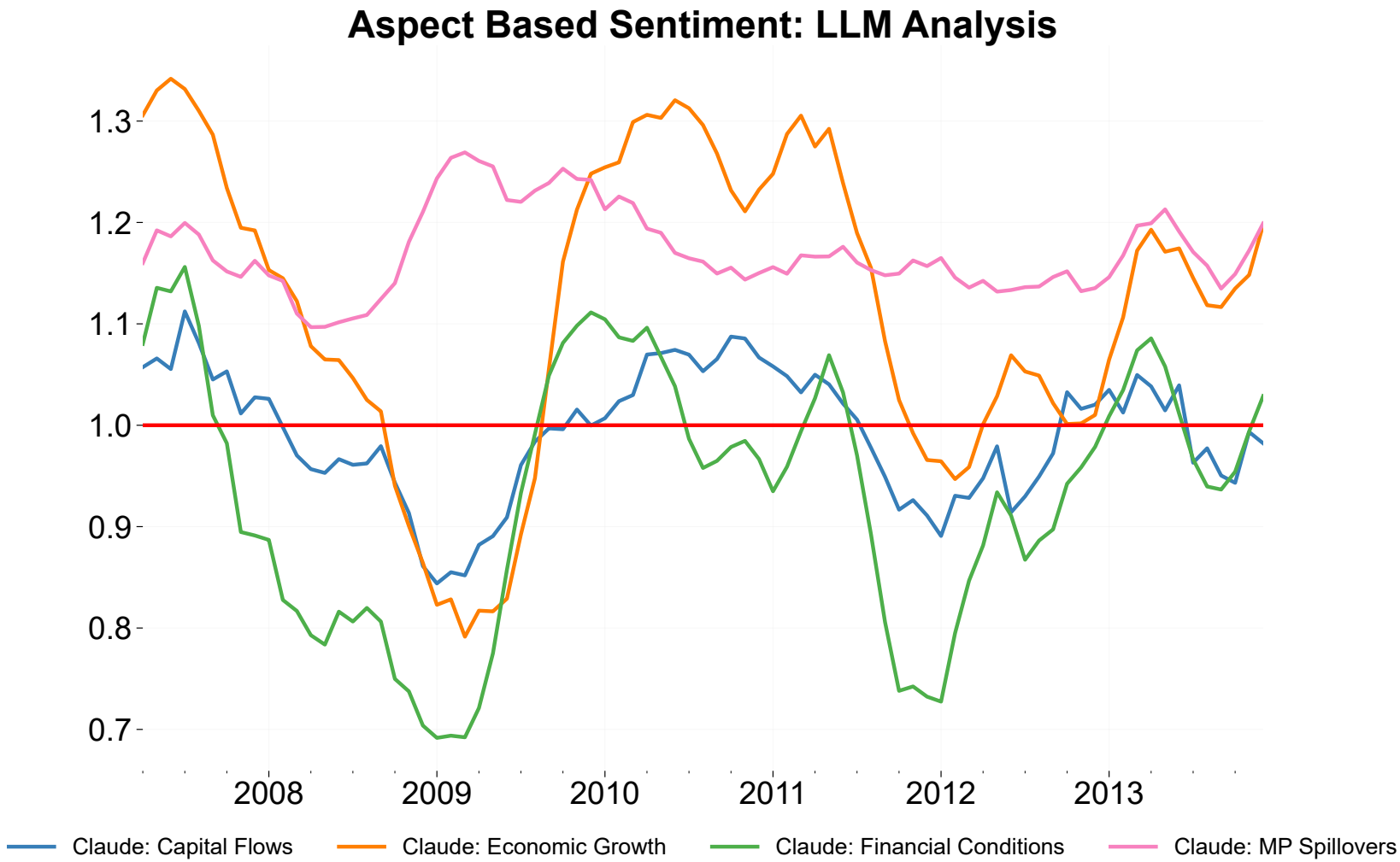


Figure 10: LLM can easily produce aspect based sentiment

Sentiment Validation: LLM Complements and Refines Lexical Methods

- LLM-based growth sentiment: correlation 0.52 with GFC, outperforming Correa (0.44) and Loughran-McDonald (0.46), comparable to Hubert (0.58)
- Cross-model LLM correlations of 0.90–0.93 for growth sentiment indicate robust consensus across architectures

Sentiment Validation: LLM Complements and Refines Lexical Methods

- LLMs complement established dictionaries by capturing structured, dimension-specific sentiment

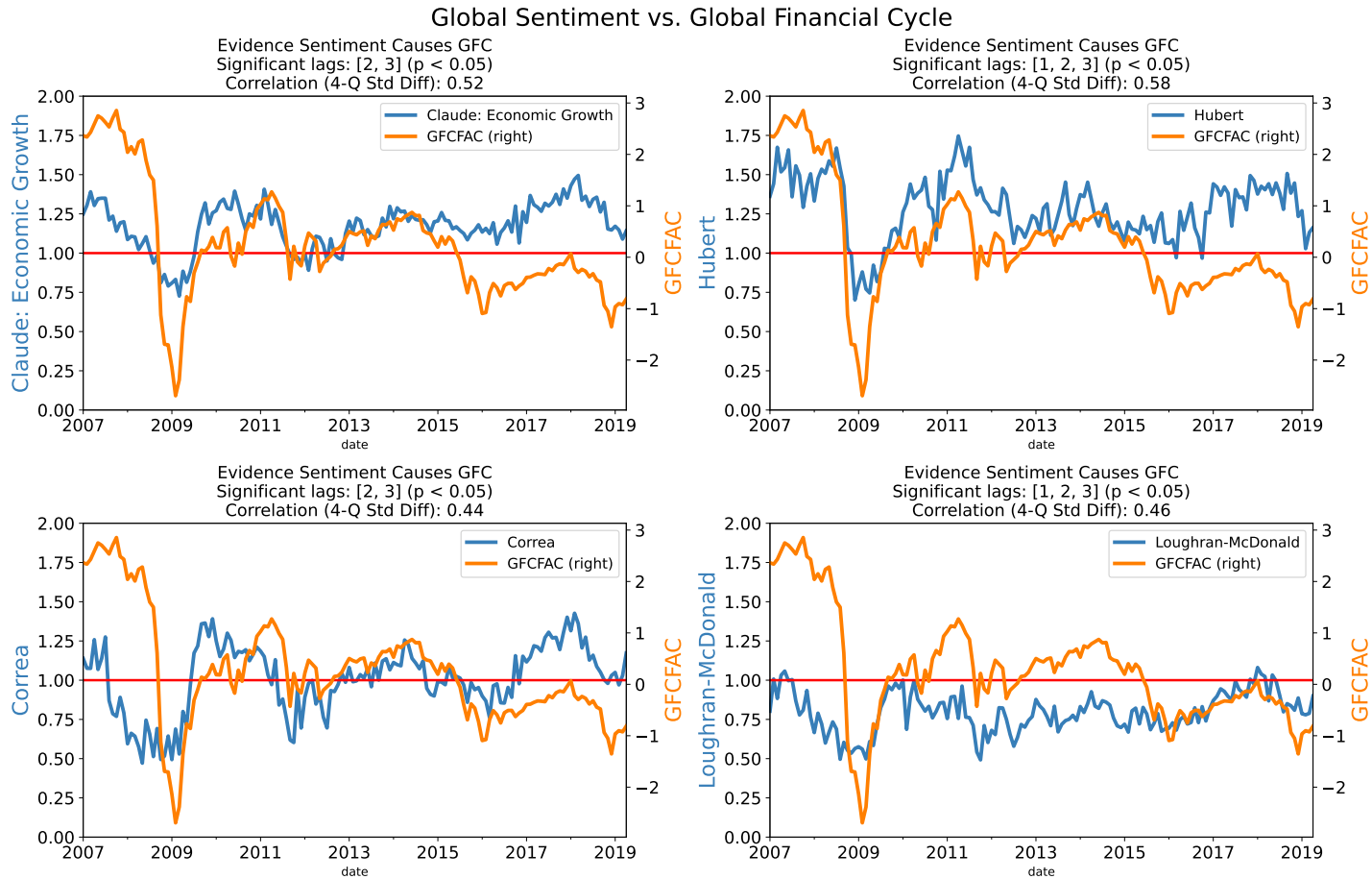
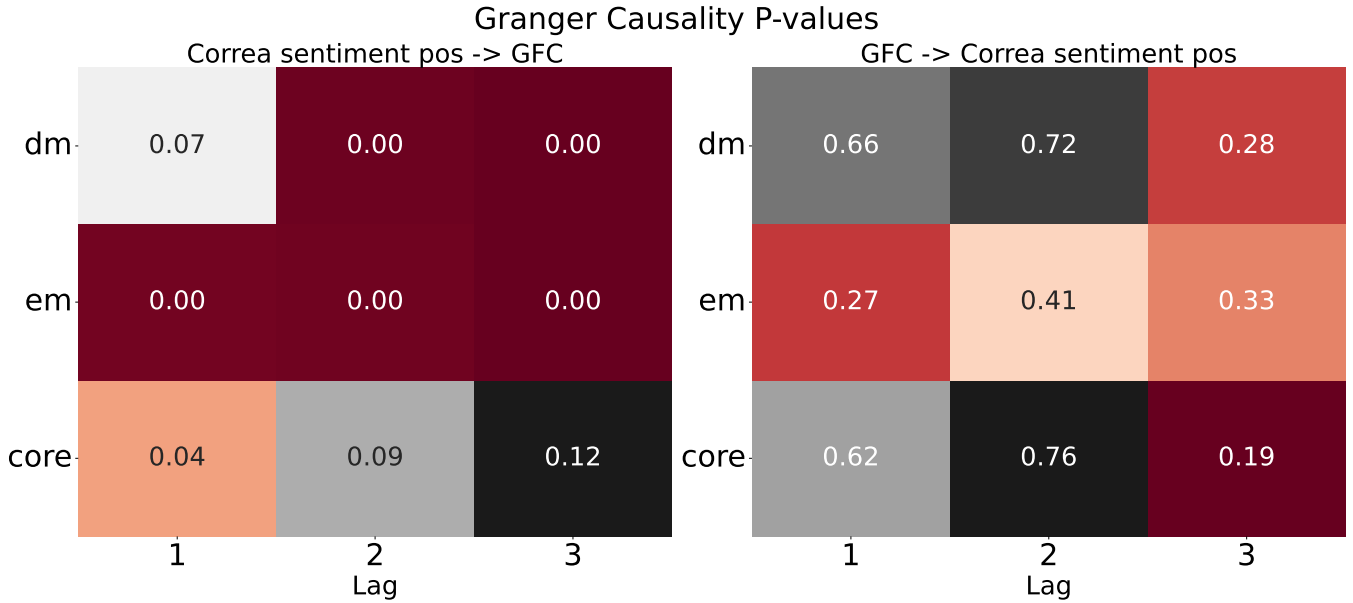


Figure 11: LLM-Generated Sentiment Indices vs. Lexical Approaches

Evidence that Sentiment Predicts the GFC

- Sentiment Granger-causes GFC ($p < 0.05$); GFC does *not* Granger-cause sentiment
- Granger causality captures **predictive precedence** rather than structural causation (Granger, 1969)
- Emerging-market CB sentiment appears to be a stronger predictor of the GFC than core CB sentiment
- Asymmetric pattern **consistent with** a communication transmission channel, though we cannot rule out omitted common factors or reverse causality at higher frequencies



Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

Summary & Future Work

Appendix

Summary & Future Work

- We construct an open-source standardized database of CB communications designed as a benchmark for LLM tasks.
- We distribute all source code for our Data/LLM infrastructure to allow researchers to leverage AI tools in a scalable and reproducible manner.
- Using LLMs, we find that central banks often cite the U.S. economy broadly (rather than mention the U.S. Federal Reserve or monetary policy spillovers) as the primary driver of global conditions.
- Granger causality tests suggest that statement sentiment predicts the Global Financial Cycle rather than merely responding to it.
- We plan to scale our data to include more central banks and communication types.
- Our infrastructure includes major closed-source (Gemini, Claude, ChatGPT) and open-source (Hugging Face) LLM systems in an easy-to-use framework.

Outline

Introduction

Technical Infrastructure

Dissecting CB Statements

Generative AI Application

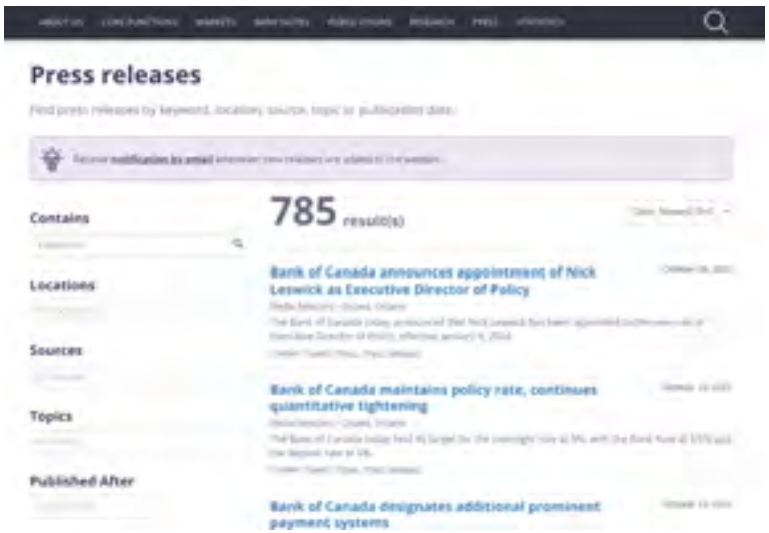
Summary & Future Work

Appendix

Monetary Policy Statements



(a) Brazil: Statements in press section



(b) Canada: Statements in press section

Figure 13: Examples of MPSs hosted on central bank websites

Key Words Time Series

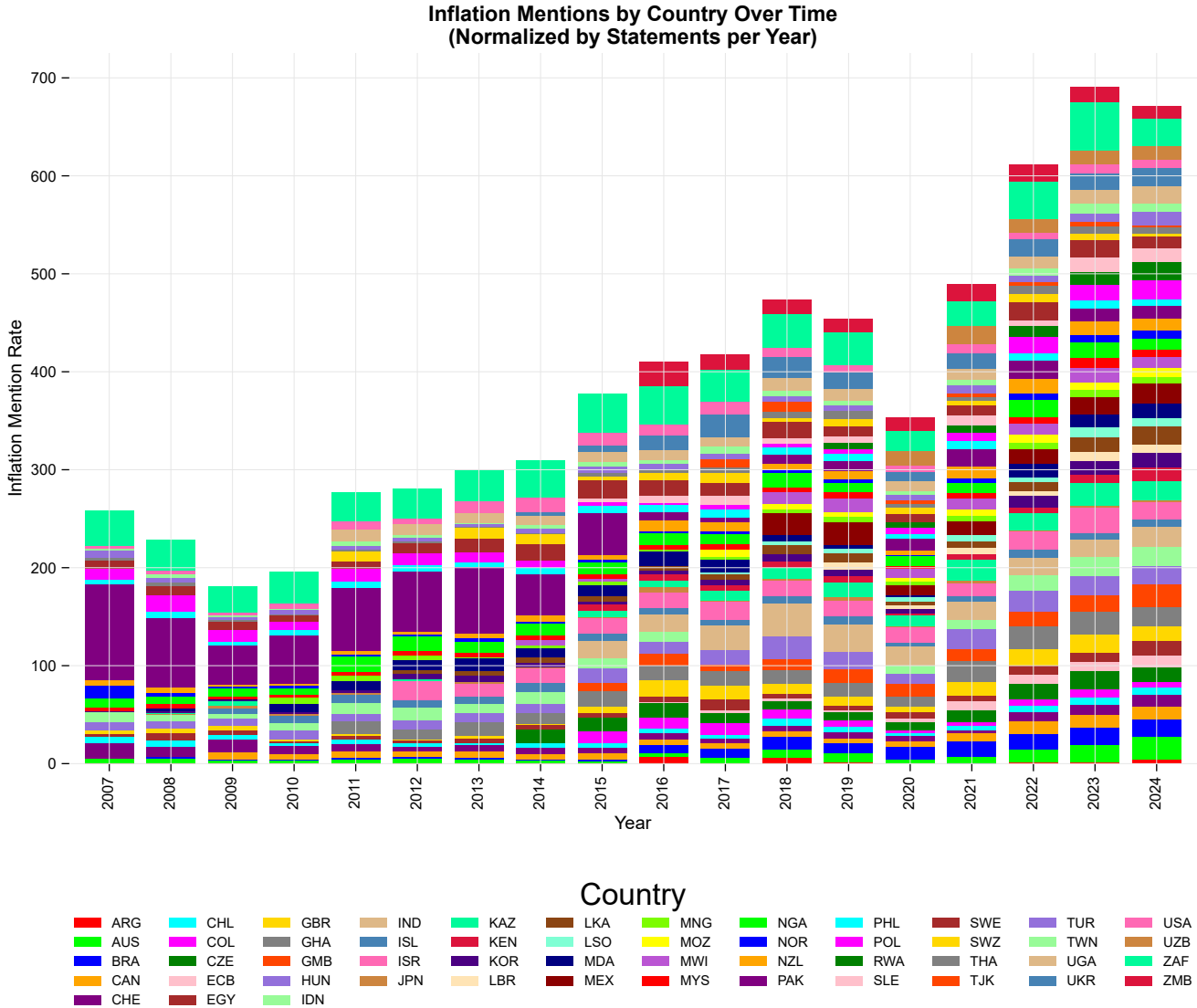


Figure 14: Inflation Mentions by Country

Regional Sentiment vs. Global Factor

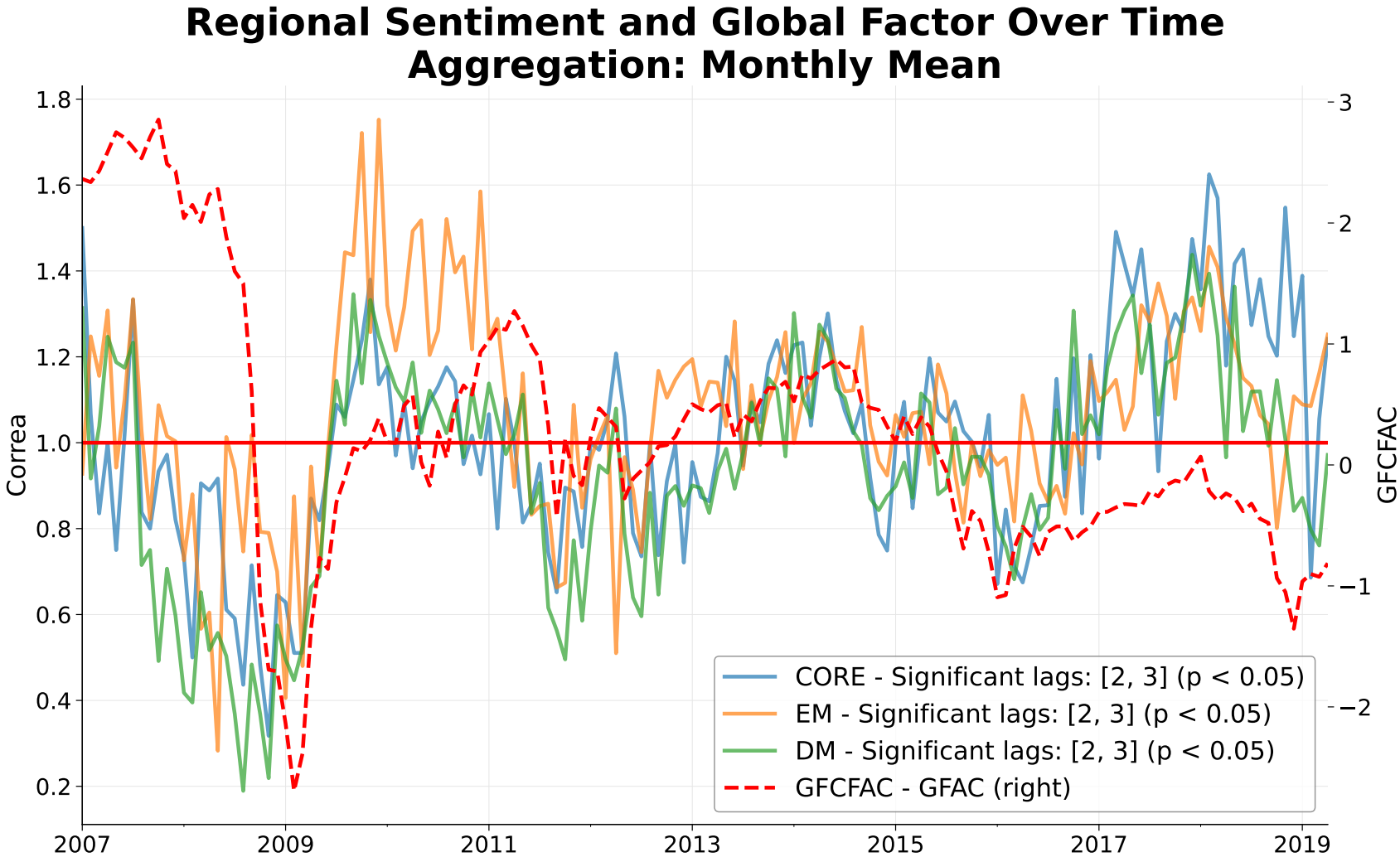


Figure 15: Regional Sentiment Evolution

References I

Born, B., Ehrmann, M., and Fratzscher, M. (2014). Central bank communication on financial stability. *Economic Journal*, 124(577):701–734.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Gonzalez, M. and Tadle, R. C. (2022). Monetary policy press releases: an international comparison. BIS Working Papers 1023, Bank for International Settlements.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(S1):114–133.

References II

Miranda-Agrippino, S. and Rey, H. (2020). U.S. monetary policy and the Global Financial Cycle. *Review of Economic Studies*, 87(6):2754–2776.

Miranda-Agrippino, S. and Rey, H. (2022). The global financial cycle. In *Handbook of International Economics*, volume 6, pages 1–43. Elsevier.

Picault, M. and Renault, T. (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance*, 79:136–156.

Rey, H. (2015). Dilemma not trilemma: the global financial cycle and monetary policy independence. NBER Working Papers 21162, National Bureau of Economic Research, Inc.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Read between the headlines: can news data predict inflation?[1]

Alan Chester Arcin, Ma. Ellysah Joy Guliman,
Genna Paola Centeno, Jacqueline Margaux Herbo,
Sanjeev Parmanand and Cherrie Mapa,
Bangko Sentral ng Pilipinas

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Read between the headlines: can news data predict inflation?

Alan Chester Arcin, Ma. Ellysah Joy Guliman, Genna Paola Centeno, Jacqueline Margaux Herbo, Sanjeev Parmanand, Cherrie Mapa[1]

Abstract

Big data and machine learning applications are increasingly gaining traction in central bank operations. Among others, central banks have tapped big data and advanced techniques in their nowcasting exercises. In this study, we construct inflation news indices and examine if such indices can help predict inflation. The indices are developed using lexicon-based sentiment analysis refined using reinforcement learning and supervised machine learning method, particularly long short-term memory model. These indices are then used as additional feature variables in time series and machine learning models for nowcasting regional and nationwide inflation in the Philippines. We find empirical evidence that our constructed inflation news indices can improve the predictive capability of these forecasting models.

Keywords: inflation; news; nowcasting; text analytics; machine learning

JEL classification: C45, E31, E37

[1] All authors are from the Department of Economic Research, Bangko Sentral ng Pilipinas. The views expressed in this study are those of the authors and do not necessarily represent the official position of the Bangko Sentral ng Pilipinas.

1. Introduction

Big data and machine learning (ML) based applications are increasingly gaining traction in central bank operations. The 2020 survey conducted by the Irving Fisher Committee on Central Bank Statistics of the Bank for International Settlements indicates that 80 percent of the central banks surveyed utilize big data for their operations, up from just 30 percent in 2015. Among various types of applications, the survey showed that central banks have used big data in their nowcasting exercises. This study adds to this existing body of empirical studies by tapping novel data sources alongside advanced computational techniques for nowcasting regional and nationwide inflation in the Philippines. Inflation nowcasting and forecasting is important for central banks like the Bangko Sentral ng Pilipinas (BSP) that operates an inflation targeting framework for monetary policy formulation. The use of big data and ML models could further support inflation surveillance and forecasting tasks of a central bank.

While Beck et al (2023) and Macias and Stelmasiak (2019) used online price data to nowcast inflation, our research examines if textual data such as online news articles could be useful for inflation nowcasting. We opt to explore online news data as an indicator for inflation trends as major developments that are relevant for inflation are likely captured by news reports. This study also builds upon an earlier work by Gabriel et al (2020) for regional inflation nowcasting in the Philippines which employed machine learning techniques but did not incorporate big data.

To answer our research question, we employ a two-step approach. First, we build inflation news indices which provide information on whether there are more reports of increasing/higher inflation vis-à-vis decreasing/lower inflation in the news media. We take a comprehensive approach and explore two methods of text analytics: (a) lexicon-based sentiment analysis refined by reinforcement learning and (b) ML-based approach using long short-term memory model. Second, we employ these indices as additional feature variables in time series and machine learning (ML) models for inflation nowcasting.

Overall, we find empirical evidence that our indices are useful for nowcasting regional and nationwide inflation in the Philippines. The news-based indices have information content that help improve the forecasting accuracy of models that utilize the said indices as additional feature variables relative to models that did not.

2. Constructing the Inflation News Indices

Our proposed inflation news indices (INIs) harness the power of online news media to capture near real-time information on price trends of goods and services as well as other factors that could affect overall inflation. Essentially, the INIs are quantitative summaries of digital news coverage on inflation and the prices of goods and services. The INIs are not sentiment indices. They do not associate inflation with either positive or negative sentiment. Instead, these indices are meant to provide indication on the prevailing information on inflation trends as reported in online news data published

by media outlets in the Philippines. This section outlines the steps in the construction of our own lexicon- and ML-based INIs.

2.1 Data sources and annotation

News articles for this study are sourced from media outlets in the Philippines that allowed us to web scrape the data from their websites' business, finance, or economy sections.[2] Our news database contains articles from different time periods but the common period when we have data from all media outlets is from January 2018 to present. We further filtered our data to news articles containing any of the words: "inflation", "price" and "prices". A total of 3,000 randomly selected sentences from our news database are annotated. These are manually labelled as 1, -1, or 0, indicating increase, decrease or no change in inflation, respectively. The annotation of sample sentences is a crucial first step to generate and to implement further enhancements of the lexicon- and ML-based INIs.

2.2 Lexicon with reinforcement learning method

The lexicon-based method involves creating a list of keywords associated with increasing or decreasing inflation based on our annotated dataset.[3] We also incorporated the negation rule but was tweaked to indicate *no change*.[4]

Following Church and Hanks (1990), pointwise mutual information (PMI) was leveraged to identify which words can be classified as *increase/decrease*. PMI is a measure of association which utilizes co-occurrence with each of the class. The overall score $Score_{PMI}$ for a given word *w* is the difference between PMI score of the word with respect to *increase* and *decrease*. Words having $Score_{PMI} > 0$ and $Score_{PMI} < 0$ are therefore assigned under *increase* and *decrease* wordlists, respectively.

The resulting wordlists is manually evaluated for suitability and reinforcement learning is used to further refine the wordlists. In particular, we use Q-learning (a type of reinforcement learning (Watkins and Dayan, 1992)) to retrieve the optimal set of words under *increase* and *decrease* that maximizes the Macro-F1 score of the test set within a reasonable amount of time.

[2] These media outlets include Business Mirror, Business World, Manila Bulletin, Manila Standard and Philippine Daily Inquirer.

[3] Standard pre-processing procedures were applied to the news dataset to reduce noise, remove unwanted parts of the data and subsequently improve the accuracy of the models used in this study. These include removal of extra spaces, numbers, punctuation marks and html tags as well as case normalization and tokenization. Furthermore, additional steps are applied such as the use of Part-of-Speech (POS) tagging and Named Entity Recognition (NER) to ensure that the dictionaries only include verbs, verb-nouns and adverbs.

[4] Negation rule applies when a word is preceded by a negation word (e.g., "not" or "never").

2.3 ML-based method using neural networks and long short-term memory models

This method relies on supervised ML models trained using our manually annotated sentences.[5] The sentences are transformed using Word2Vec and are fed into artificial neural networks (ANNs) such as multilayer perceptron (MLP) and Bidirectional Long Short-Term Memory (BiLSTM) models. We test several model configurations and explore the use of Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al, 2002) to create synthetic entries and remedy data scarcity concerns.[6] For the purposes of this study, all ANNs are designed to classify three output classes: increase, decrease, and no change. We then compare the Macro-F1 scores of the different ANNs.

2.4 Index construction and evaluation

To construct the INIs, we evaluate the lexicon and trained neural networks based on their accuracy and Macro-F1 scores. Table 1 presents a summary of evaluation of different lexicons and ANNs. The results suggest that INIs derived using lexicon refined by reinforcement learning (hereafter referred to as INI-RL) and using BiLSTM (32) (hereafter referred to as INI-BiLSTM) have the best accuracy and Macro-F1 scores and are therefore preferred methods for index construction. These are then used to score the sentences in the news articles with scores aggregated per article and indexed for the month. Afterwards, the INIs are normalized to a mean of 100. Thus, any estimate above 100 suggests that relative to average, news about increasing inflation outnumber reports of declining inflation. Meanwhile, any estimate below 100 suggests otherwise.

INI methods: Evaluation metrics for test set

Table 1

	Accuracy (In percent)	Macro-F1 score (In percent)
Initial lexicon	64.3	41.8
Lexicon with reinforcement learning	89.3	69.3
MLP	68.9	45.7
BiLSTM (16)	78.4	56.2
BiLSTM (32)	79.8	57.9

Note: Number in parenthesis for ANNs refers to number of hidden units

Sources: Authors' estimates

[5] Similar to the lexicon approach, standard pre-processing procedures were applied.

[6] A common challenge with annotated datasets is having imbalanced datasets or having unequal number of entries per class. To remedy this, we explored the use of SMOTE. This technique synthesizes new samples of the minority class based on the nearest neighbors in the feature space. However, caution is advised in using this technique for text-based data as the synthetic samples may not be able to have a complete context.

Analyses of the resulting series indicate that the INI-RL and INI-BiLSTM tend to move together. This is reassuring as two different methodologies point to similar results. More importantly, we also find that both INIs tend to co-move with actual headline inflation. This hints of the ability of news data to mirror prevailing inflation trends.

3. Forecasting inflation using news data

Our main research objective is to determine the usefulness of news data in nowcasting both regional and nationwide inflation in the Philippines. Following Gabriel et al (2020), we first estimate time series and ML models without the INIs. These serve as our baseline models. We then augment these models with the INIs and perform forecast evaluation (using mean absolute error (MAE)) to determine if the inclusion of INIs leads to improved nowcasting capability. We use the Diebold-Mariano test to determine if there is indeed statistically significant improvement in forecast errors.

The target variable in these models is the year-on-year inflation rate.[7] We build individual models for each of the 16 regions and a separate model for nationwide inflation.

For the baseline time series model, we used auto ARIMA to help identify the optimal lag, integration, moving average and seasonal order that will yield the least error in nowcasting. We included additional feature variables such as the month-on-month difference in inflation and a dummy variable that takes a value of 1 if inflation is on an uptrend in the last two months and -1 if inflation is on a downtrend in the last two months to help improve nowcasting performance. The model was trained for the period January 2018 to December 2022. Out-of-sample nowcasts for regional and nationwide inflation were generated for the test period January to December 2023.

For the baseline ML models, we tried a number of models including Support Vector Regression (SVR), Gradient Boosted Machines (GBM) and Extremely Randomized Trees (EXT). We utilized similar feature variables as those in the baseline ARIMA models. We also employed the same training and test periods. We find that SVR produces the lowest nowcasting error among these ML models. We therefore only report the results for the SVRs.

For the INI-augmented models, we add another variable S^{INI} defined in Equation (1+:

$$S_t^{INI} = \begin{cases} 1, & if\ INI_t > INI_{t-1} \\ -1, & if\ INI_t < INI_{t-1} \\ 0, & otherwise \end{cases} \qquad (1)$$

where INI_t refers to either INI-RL or INI-BiLSTM. These augmented models were trained for the same period with regional and nationwide inflation nowcasts generated for the period January – December 2023.

[7] Data on regional and nationwide inflation for the Philippines are sourced from the Philippine Statistics Authority.

4. Analysis of Results

This section presents the results of our inflation nowcasting exercises. As we are concerned mainly with out-of-sample forecasting performance, we only present the MAE from the test set for the regional and nationwide models.

Table 2 shows the forecast evaluation of the baseline traditional time series and ML models vis-à-vis those that included INI-RL and INI-BiLSTM as additional regressor or feature variable. On average, forecast errors have declined in models augmented with INIs compared to baseline models. This holds true for either INI-RL or INI-BiLSTM. Particularly, for the SVR model, we find that the addition of INI-RL led to statistically lower MAEs based on the Diebold-Mariano test in 7 out of 17 regions and in the nationwide model. For the SVR model augmented with INI-BiLSTM, forecast errors were statistically lower for 8 regions and nationwide level.

Nowcast Evaluation
January – December 2023

Table 2

	Traditional Time Series Model (Auto-ARIMA)			Machine Learning Model (SVR)		
	Baseline	With INI-RL	With INI-BiLSTM	Baseline	With INI-RL	With INI-BiLSTM
NCR	0.24	0.24	0.27	0.29	0.28	0.02*
CAR	0.28	0.26	0.26	0.28	0.04*	0.24*
Region1	0.27	0.24	0.24*	0.46	0.33*	0.22*
Region 2	0.36	0.33	0.33*	0.47	0.32	0.39
Region 3	0.40	0.34*	0.32*	0.47	0.08*	0.03*
Region 4a	0.33	0.30	0.29	0.36	0.14*	0.02*
Region 4b	0.30	0.24*	0.23*	0.30	0.06*	0.02*
Region 5	0.31	0.28	0.27	0.38	0.22	0.12*
Region 6	0.41	0.37	0.37	0.41	0.40	0.45
Region 7	0.30	0.28	0.30	0.28	0.22	0.24
Region 8	0.22	0.21	0.21	0.27	0.30	0.29
Region 9	0.39	0.40	0.38	0.61	0.59	0.60
Region 10	0.26	0.29	0.27	0.30	0.21	0.23
Region 11	0.29	0.27	0.29	0.29	0.29	0.28
Region 12	0.38	0.38	0.37	0.33	0.35	0.33
BARMM	0.38	0.23*	0.25*	0.37	0.13*	0.25*
Region 13	0.16	0.15	0.16	0.20	0.14*	0.22
Philippines	0.30	0.31	0.29	0.37	0.03*	0.18*
Average	0.31	0.28	0.28	0.36	0.23	0.23

Note: * -passed Diebold-Mariano Test

NCR – National Capital Region; CAR – Cordillera Administrative Region; BARMM – Bangsamoro Autonomous Region in Muslim Mindanao

Sources: Authors' estimates

It is important to note a few caveats in the interpretation of the results. First, the evaluation period is relatively short owing to data limitations. This warrants further assessment when more data become available. Second, there is need to fine tune the news indices to the regional level. News data came from nationwide newspapers. There is need to acknowledge that certain news reports may not be relevant for all regions. For example, a typhoon in a particular region that led to agricultural crop damage in the area and consequently, higher regional food prices would likely be covered in nationwide newspapers and would therefore be part of our dataset. However, this may not be as relevant for regions not affected by the typhoon. Fine tuning the news reports to the regional level (i.e., by tapping regional newspapers and tagging reports based on regions) could further improve regional inflation surveillance and nowcasting. Third, 2023 is a relatively volatile year for inflation in the Philippines, leading to higher forecast errors. The results of the forecast evaluation of the workhorse models of the BSP, published in the May 2024 Monetary Policy Report, showed an increase in month-ahead MAE to 0.28 percentage point in 2023 from 0.17 percentage point in 2022 (BSP, 2024). For the same period, we estimated errors of the INI-augmented models at 0.23 – 0.28 percentage point, broadly similar to the nowcast errors of the BSP's workhorse models. This is encouraging as the INI-augmented models relied solely on textual data and time series properties of the inflation series. It did not incorporate any traditional economic data as regressor. Furthermore, the MAEs of the INI-augmented models are also lower compared to the 0.41 percentage point MAE of the month-ahead consensus forecasts from Bloomberg's monthly survey of private sector analysts (BSP, 2024). Thus, available evidence suggests that the INIs contain information that could be useful for nowcasting even as we acknowledge the need to fine tune the indices and undertake further evaluation in the future.

5. Conclusion and future directions

In this study, we develop news-based inflation indices. These indices rely on online news articles and are derived using two approaches: (1) lexicon refined by reinforcement learning and (2) machine learning model, particularly BiLSTM. The indices capture prevailing information on inflation trends as reported in digital news by Philippine media outlets. We find that augmenting ARIMA and SVR models for regional and nationwide inflation nowcasting with these indices help in lowering their nowcasting errors. Thus, these indices could be useful for inflation surveillance. Our findings contribute to the existing literature on the relevance of big data for nowcasting macroeconomic variables. Future plans to improve the indices include fine tuning the indices to the regional level, exploring additional keywords that could pave the way for the creation of sub-indices (i.e., food INI, energy INI) and further evaluation of nowcasting capability.

References

Bangko Sentral ng Pilipinas (2024): "An evaluation of the BSP's 2023 Forecasting Performance (Box Article no. 2)'', *Monetary Policy Report* May 2024.

Beck, G. W., Carstensen, K., Menz, J. O., Schnorrenberger, R., & E. Wieland (2023): "Nowcasting consumer price inflation using high-frequency scanner data: Evidence from Germany", *Deutsche Bundesbank Discussion Paper* No. 34/2023.

Chawla, N., K. Bowyer, W., L. Hall, & W.P. Kegelmeyer (2002): "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, vol 16, pp 321-357.

Church, K., & P. Hanks (1990): "Word association norms, mutual information, and lexicography", *Computational Linguistics*, vo.l 16, no. 1, pp. 22-29.

Cortes, C., & V. Vapnik, (1995): "Support-vector networks", *Machine Learning*, vol 20, pp 273-297.

Friedman, J. (2001): "Greedy function approximation: a gradient boosting machine", *Annals of Statistics*, vol. 29, no. 5, pp 1189-1232.

Gabriel M., D. Bautista & C. Mapa (2020): "Forecasting regional inflation in the Philippines using machine learning techniques: A new approach", *Bangko Sentral ng Pilipinas Working Paper Series* No. 2020-10.

Geurts, P., D. Ernst & L. Wehenkel (2006): "Extremely randomized trees", *Machine Learning*, vol 63, pp3-42.

Macias, P., & D. Stelmasiak, (2019): "Food inflation nowcasting with web scraped data", *Narodowy Bank Polski Working Paper* no. 302.

Mahadevaswamy, U. B., & P. Swathi (2023): "Sentiment analysis using bidirectional LSTM network", *Procedia Computer Science*, vol 218, pp 45-56.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean (2013): "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, October.

Irving Fisher Committee on Central Bank Statistics (2021): "Use of big data sources and applications at central banks", *IFC Report* no. 13.

Watkins, C. J., & P. Dayan (1992): "Q-learning", *Machine Learning*, vol 8, pp 279-292.

Read between the headlines: Can news data predict inflation?

ALAN CHESTER ARCIN, MA. ELLYSAH JOY GULIMAN, GENNA PAOLA CENTENO, JACQUELINE MARGAUX HERBO, SANJEEV PARMANAND, **CHERRIE MAPA**
*Department of Economic Research, Bangko Sentral ng Pilipinas**

4th BIS-IFC– Bank of Italy Workshop on Data Science in Central Banking
18-20 February 2025

* The usual institutional disclaimer applies.

Classification: GENERAL

Motivation



Leverage nontraditional data sources



Support inflation nowcasting

Classification: GENERAL

Inflation News Index (INI)





The INI is a quantitative summary of digital news coverage on inflation and prices of goods and services

Classification: GENERAL

Constructing the INIs: An overview

Web scraping



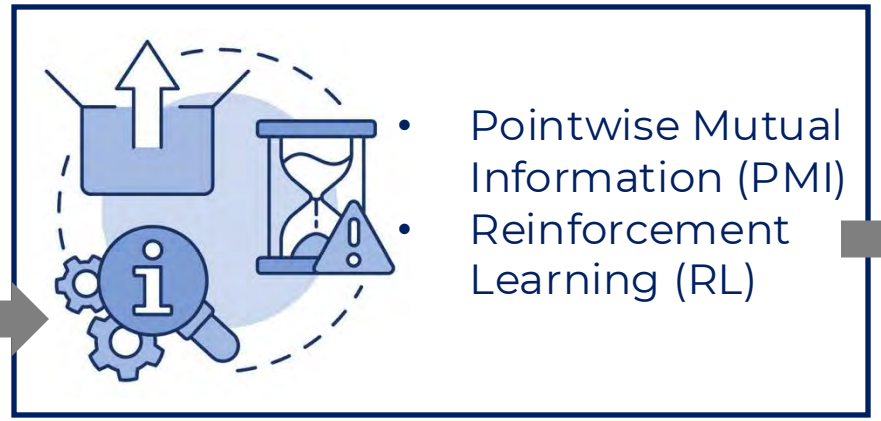


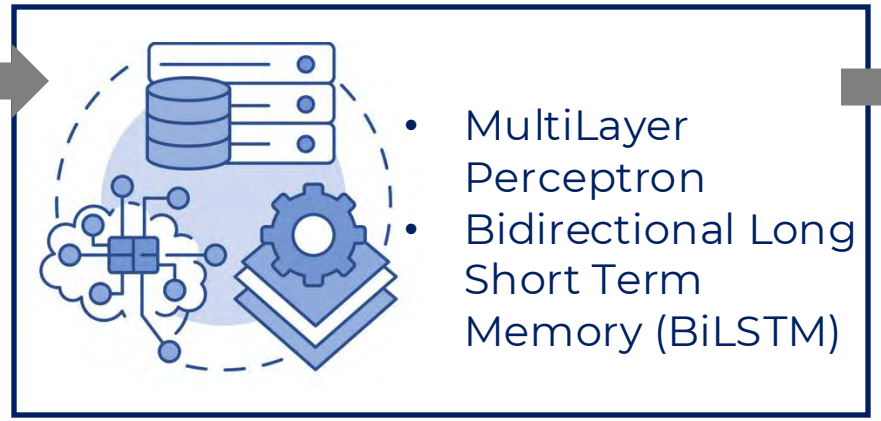
Keyword Filtering



Manual Annotation

Deriving Lexicons





Machine Learning

INI-RL





INI-BiLSTM

Classification: GENERAL

Data sources and annotation

- ✓ Data were collected from multiple local news sources. [Sections of interest: economy, banking, finance].
- ✓ Articles are then filtered with the keywords: “inflation”, “price”, “prices”
- ✓ A random sample of 3,000 sentences were annotated

Sample annotated sentences and their classes/labels[

News Text	Label
Aside from flour, other ingredients also went up including LPG, and sugar which substantially increase prices.	1
This was a slight decline from last weeks price level of P37.83/kg.	-1
Some will not pass on the fuel price increases on passengers so they can keep their ticket prices low and haul in more customers.	0

Source: Various media websites, authors’ estimates

Classification: GENERAL

Data pre-processing

Pre-processing steps include:

- Removal of extra spaces in the text
- Case normalization
- Removal of punctuations marks
- Unicode conversion
- Tokenization
- Vectorization

Additional steps include:

- Part of speech tagging
- Named entity recognition

Classification: GENERAL

Methodology I: Lexicon refined with reinforcement learning

Dictionary based method involves a creating a list of keywords. Pointwise mutual information (PMI) was used to infer word associations with increasing and decreasing inflation

Overall PMI score for a word w:

$$Score(w) = PMI(w|increase) - PMI(w|decrease)$$

Initial lexicon: Words with Score > 0 are classified as increase and words with Score < 0 are classified as decrease

Reinforcement learning was used to improve the initial lexicon

Classification: GENERAL

Methodology II: Machine Learning models

Machine learning based method utilizes two artificial neural networks to predict association of text

- MultiLayer Perceptron (MLP)
  - A fully connected feedforward neural network, notable for finding nonlinear relationships.

- Bidirectional Long Short Term Memory (BiLSTM)
  - Contains two LSTM layers in opposite directions. LSTMs are used for finding long term dependencies in sequential data such as time series, text and audio.

Classification: GENERAL

Evaluation of methodologies

Both dictionaries and machine learning models were evaluated against test set of the manually-labelled sentences sampled from online news articles

Lexicons	Accuracy (in percent)	Macro F1 (in percent)
Initial lexicon	64.3	41.8
Lexicon with RL	89.3	69.3
Machine Learning Models	**Accuracy (in percent)**	**Macro F1 (in percent)**
MLP	68.9	45.7
BiLSTM(16)	78.4	56.2
BiLSTM(32)	79.8	57.9

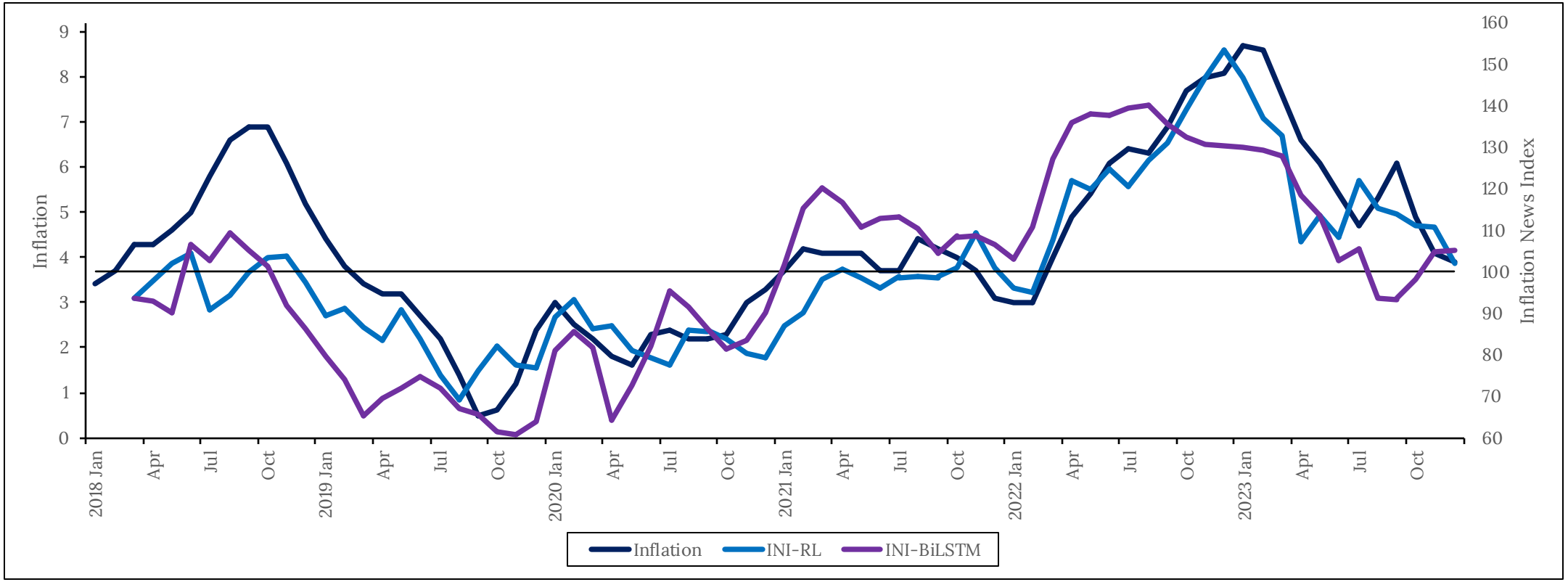
Note: Number in parenthesis for BiLSTM refers to number of hidden units.
Source: Authors' estimates

Classification: GENERAL

Results

Estimates of INIs strongly co-move with inflation.

INIs vis-à-vis inflation
Q1 2018 to Q4 2023



Source: Authors' estimates

Classification: GENERAL

Nowcast evaluation

On average, forecast errors have declined in models augmented with INI-RL compared to baseline models

Nowcast evaluation: Mean Absolute Error

January-December 2023

	Baseline AR	ARX+INI
NCR	0.24	0.24
CAR	0.28	0.26
R1	0.27	0.24
R2	0.36	0.33
R3	0.40	0.34*
R4A	0.33	0.30
R4B	0.30	0.24*
R5	0.31	0.28
R6	0.41	0.37
R7	0.30	0.28
R8	0.22	0.21
R9	0.39	0.40
R10	0.26	0.29
R11	0.29	0.27
R12	0.38	0.38
BARMM	0.38	0.23*
R13	0.16	0.15
Philippines	0.30	0.31
Average	**0.31**	**0.28**

	Baseline SVR	SVR+INI
NCR	0.29	0.28
CAR	0.28	0.04*
R1	0.46	0.33*
R2	0.47	0.32
R3	0.47	0.08*
R4A	0.36	0.14*
R4B	0.30	0.06*
R5	0.38	0.22
R6	0.41	0.40
R7	0.28	0.22
R8	0.27	0.30
R9	0.61	0.59
R10	0.30	0.21
R11	0.29	0.29
R12	0.33	0.35
BARMM	0.37	0.13*
R13	0.20	0.14*
Philippines	0.37	0.03*
Average	**0.36**	**0.23**

Source: Authors' estimates

*Passed Diebold-Mariano test

Classification: GENERAL

Nowcast evaluation

On average, forecast errors have declined in models augmented with INI-BiLSTM compared to baseline models

Nowcast evaluation: Mean Absolute Error

January-December 2023

	Baseline AR	ARX+INI
NCR	0.24	0.27
CAR	0.28	0.26
R1	0.27	0.24*
R2	0.36	0.33*
R3	0.40	0.32*
R4A	0.33	0.29
R4B	0.30	0.23*
R5	0.31	0.27
R6	0.41	0.37
R7	0.30	0.30
R8	0.22	0.21
R9	0.39	0.38
R10	0.26	0.27
R11	0.29	0.29
R12	0.38	0.37
BARMM	0.38	0.25*
R13	0.16	0.16
Philippines	0.30	0.29
Average	**0.31**	**0.28**

	Baseline SVR	SVR+INI
NCR	0.29	0.02*
CAR	0.28	0.24*
R1	0.46	0.22*
R2	0.47	0.39
R3	0.47	0.03*
R4A	0.36	0.02*
R4B	0.30	0.02*
R5	0.38	0.12*
R6	0.41	0.45
R7	0.28	0.24
R8	0.27	0.29
R9	0.61	0.60
R10	0.30	0.23
R11	0.29	0.28
R12	0.33	0.33
BARMM	0.37	0.25*
R13	0.20	0.22
Philippines	0.37	0.18*
Average	**0.36**	**0.23**

Source: Authors' estimates

*Passed Diebold-Mariano test

Classification: GENERAL

Key findings

- INIs are quantitative summary of digital news coverage on inflation and prices of goods and services
- INIs are found to co-move with actual inflation
- INIs exhibit strong correlation with survey-based inflation expectations of firms and professional forecasters
- INIs contain information that can help nowcast inflation

Classification: GENERAL

Sources

- Beck, G. W., Carstensen, K., Menz, J. O., Schnorrenberger, R., & Wieland, E. (2023). Nowcasting consumer price inflation using high-frequency scanner data: Evidence from Germany (No. 34/2023). Deutsche Bundesbank Discussion Paper.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
- Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- Gabriel M., Bautista D., & Mapa C. (2020). Forecasting regional inflation in the Philippines using machine learning techniques: A new approach. Bangko Sentral ng Pilipinas Working Paper Series No. 2020-10.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63, 3-42
- Macias, P., & Stelmasiak, D. (2019). Food inflation nowcasting with web scraped data (p. 302). Warsaw: Narodowy Bank Polski, Education & Publishing Department.
- Mahadevaswamy, U. B., & Swathi, P. (2023). Sentiment analysis using bidirectional LSTM network. Procedia Computer Science, 218, 45-56
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
- Serena, J. M., Tissot, B., Doerr, S., & Gambacorta, L. (2021). Use of big data sources and applications at central banks (No. 13). Bank for International Settlements.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8, 279-292

Classification: GENERAL

Thank you!

Classification: GENERAL

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

The impact of monetary surprises on exchange rates: results from textual and high-frequency analysis[1]

Jean-Charles Bricongne,
Bank of France

Louis Marolleau,
Lille University

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

The impact of monetary surprises on exchange rates: results from textual and high-frequency analysis[1]

Jean-Charles Bricongne (2)*, Louis Marolleau (3)

Abstract

This study examines the impact of monetary policy surprises on exchange rates using a text indicator. It uses a textual analysis method to analyze 746 business press articles and creates a database of 510 monetary policy decisions between 2018 and 2023 for 11 countries with floating exchange rates. The study compares the amplitudes of exchange rate variations for surprises with expected decisions. Results show that monetary surprises increase the amplitudes of minute-to-minute variations by 0.19 percentage point to 0.39 percentage point, with amplitudes 100 to 1,000 times greater than the median over a year for the currencies in the database. The study suggests that appreciation and depreciation of currencies are immediate in the currency market following monetary announcements.

Keywords: Monetary policy surprises, Text Analysis, Exchange Rates, Signal Processing, High-Frequency data

JEL classification: E52, E58, E43

1. Introduction

Central banks aim to maintain credibility and influence expectations by making their decisions predictable (Blattner et al., 2008). Surprising decisions, such as interest rate setting or communication on future monetary policy, can impact financial variables and exchange rates (Jarociński & Karadi, 2020). The Uncovered Interest Rate Parity theory suggests that risk-neutral investors invest in foreign currency when foreign expected returns are higher (McCallum, 1994; Keynes, 1923). Exchange rates react strongly and immediately when communicating unexpected decisions (Kearns &

[1] This work has been published as an article in the Journal of International Money and Finance (JIMF). See Bricongne and Marolleau, "The impact of monetary surprises on exchange rates: results from textual and high-frequency analysis", *Journal of International Money and Finance,* vol 158, article 103401, September 2025, https://doi.org/10.1016/j.jimonfin.2025.103401.

[2] Banque de France, 115 rue Réaumur, 75002 Paris, Paris I University and LEO (Laboratory of Economics of Orléans). * Corresponding author: jean-charles.bricongne@banque-france.fr

[3] Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France.

Manners, 2006). To identify surprises, interest rate futures data (Kuttner, 2001) (Gürkaynak et al., 2021) or consensus forecasts, among others, are used, but these methods have limitations. This paper uses a textual analysis method to identify monetary surprises in business press articles, creating a database of monetary policy decisions with or without surprises. The impact of these decisions on exchange rates is then analyzed, focusing on the amplitudes of exchange rate variations and the times when the central bank begins its monetary policy communication. This method can be applied to many countries over a long period, particularly those with limited financial data access.

The study reveals that monetary surprises lead to greater exchange rate variations in the first few minutes after a decision is communicated, compared to expected decisions. Indeed, a surprising decision increases the amplitude of variation at the time of the decision from 0.19 p.p. to 0.39 p.p. on a one-minute scale, depending on the period considered or whether Turkey, which is atypical, is present. This effect is significant as it is calculated from one minute to the next. The amplitude of a surprise is 100 to 1000 times greater than the median amplitudes of minute-to-minute variation for currencies in the database. The COVID-19 crisis also contributed more to exchange rate variation due to uncertainty and market volatility.

2. Literature Review

Surprising monetary policy decisions can impact monetary policy transmission, making it crucial for central banks. Kuttner (2001) and Gürkaynak et al. (2005) found that US monetary policy decisions impact asset prices through a "current federal funds rate target factor" and a "future path of policy factor" linked to central bank communication. Rosa & Verga (2008) examined the European Central Bank's communication effect on Euribor's future prices using high-frequency intraday data. They found that the unanticipated component of monetary policy decisions affects future prices.

Building on Rosa & Verga's (2008) paper, we highlight surprises using an indicator in the regression analysis to deduce the impact on exchange rate variations.

To estimate the impact of surprises linked to monetary policy decisions, researchers can use derivatives market data (Gürkaynak et al., 2021), Fed fund futures rates (Kuttner, 2001), and Overnight Index Swap (OIS) (Enders et al., 2019) to analyze changes in interest rates. However, obtaining data on derivative contract pricing variations can be costly. Textual analysis methods can be used to identify uncertainty shocks in press articles, such as Baker et al. (2016) study on press articles, which found increased stock price volatility and reduced investment and employment in sectors like defense, healthcare, finance, and infrastructure construction.

The current study's added value lies in increasing the basis for data recovery from more easily accessible data sources, which enables to cover a panel of developed and emerging countries.

Most papers on monetary surprises have focused on a few developed countries, but Deb et al. (2023) compared short-term interest rates with forecasted interest rates for 33 countries to identify monetary policy surprises in a more diverse panel. Furceri

et al. (2018) construct annual monetary policy shocks for advanced and emerging countries using unanticipated changes in policy rates. Our study proposes studying surprises using a textual analysis method for 11 countries with floating exchange rates.

3. Data and Methodology

a) Monetary policy decision calendar and database of press articles

A database of monetary policy decisions is created using online communication calendars from *centralbanknews.info* and *investing.com*. Monetary policy decisions are announced the day after meetings and communicated through press conferences and website press releases. *Factiva* and *Focus Economics* websites are used to extract and analyze articles related to monetary policy decisions, providing summaries of decisions on most days when they are taken.

b) Text identification of surprises

An automatic search algorithm is used to identify monetary surprises in press articles. The algorithm checks if the article contains at least one of the following keywords related to surprises: ["surpris"], ["unexpect"], ["unanticipat"], ["than expect"], ["than anticipat"], ["not expect"], ["not anticipat"]. If found, it is read in the paragraph where the keyword appears and the previous and following paragraphs, creating a context window. This process checks 75% of calendar days for monetary policy decisions, with extraordinary and unexpected meetings during the COVID-19 crisis considered as surprises.

The study investigates surprises in interest rate setting and monetary policy. It uses a "surprise" indicator from economic press articles to enrich a database of monetary policy decisions. The robustness of the tagging was checked using ChatGPT-4, resulting in 94% of initial human tagging being confirmed. In cases of disparity, 5% of tagged decisions were corrected, ensuring accuracy.

c) Dependent variable

The Uncovered Interest Rate Parity mechanism (UIP) helps explain exchange rate fluctuations by implying that a rise in a foreign country's key interest rate implies an appreciation of the foreign currency, which works for countries with floating exchange rates. If we take the example of US investors investing abroad, we have the Uncovered Interest rate Parity:

$$(2)\ 1 + i = \frac{\mathbb{E}[S_{USD/X}](1+i*)}{S_{USD/X}}$$

With i the domestic interest rate, i^* the external interest rate, $S_{USD/X}$ the exchange rate expressed in US dollars per unit of foreign currency, $\mathbb{E}[s_{USD/X}]$ the expected exchange rate.

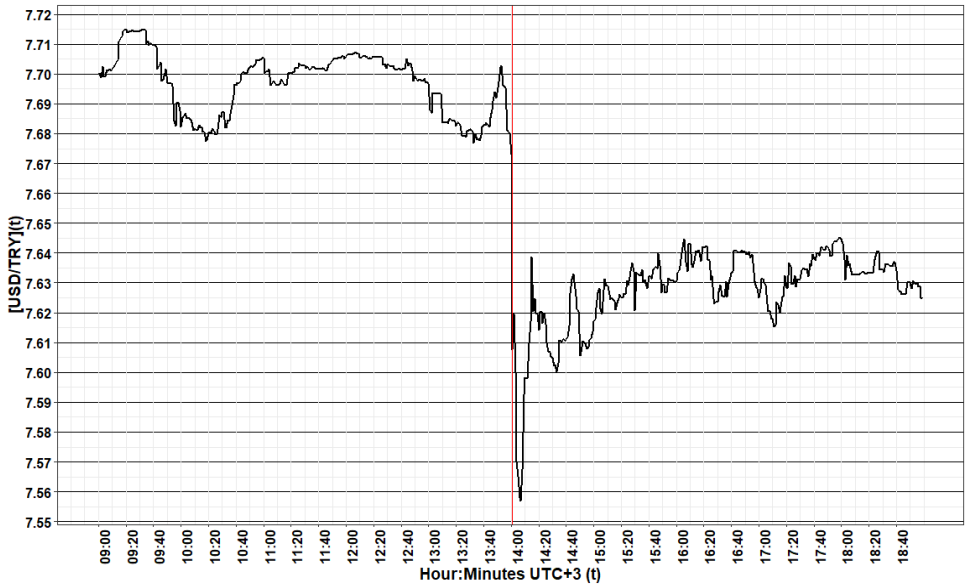
In equilibrium, US investors are indifferent between domestic investment and foreign currency conversion. They will be indifferent between investing in the US and earning interest, or converting their assets into foreign currency and investing in the foreign country at the interest rate i*. Financial markets expect the USD to depreciate if the foreign interest rate rises.

Exchange rates are expressed as relative variations from one minute to the next, with minimal variation on a scale of one minute. This indicator is used in market finance to study the high-frequency return on an asset. By analyzing relative variations, we can normalize bilateral exchange rates on a percentage basis and make comparisons. In this case, currencies are expressed in US dollars (USD) per foreign currency, as USD is the reference currency in international markets. The bilateral exchange rate is expressed vis-à-vis the euro when considering the United States.

Relative variations of a per thousand or even a per cent in exchange rates are rare on a one-minute scale when considering several years, as shown in **Table 1** in the Appendix. The mean, median, quantiles, minimum, and maximum relative variations in exchange rates for currencies the panel covers from January 2018 to November 2023 are calculated minute-by-minute. Therefore, if relative variations are measured when a monetary policy decision is communicated, the decision strongly impacts the exchange rate.

The UIP effect is illustrated in Turkey's case. The Central Bank's Monetary Policy Committee raised the one-week repo rate by 200 basis points to 10.25%, surprising market analysts and affecting the USD/TRY exchange rate, as shown in **Figure 1**.

US dollar against Turkish lira (1 TRY x USD) on 24 September 2020 Figure 1



Note: The monetary policy decision of the Central Bank of the Republic of Turkey contained a surprise. Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction

The Turkish lira appreciated rapidly after the announcement of a monetary policy decision on 24 September 2020, with a significant price jump from 2:00 pm to 2:01 pm in local time. This significant variation, which corresponds to the announcement of the policy, resulted in a 1.8% appreciation in just five minutes.We retain a measure of relative variation from one minute to the next, taken at absolute value to remove the signs of variation. These are the amplitudes of variation, and these amplitudes will be the dependent variable in the econometric models.

d) Database creation

The daily decision-making process matches surprising and non-surprising decisions with exchange rate valuation minutes from *histdata.com*, an open-source Forex data source. The exchange rate level is defined at the opening of each minute and expressed in EST time.

We explore the signal of central bank communication by tracing the moment of the decision day when the amplitude was highest. The hypothesis is that the moment of the decision day when the amplitude was highest corresponds to the start of the central bank communication. The likelihood of other events explaining the high amplitudes of variation at the communication time appears very low. We test this hypothesis by identifying the time of the day when the amplitude was maximum, ensuring consistency in timing.

The timetable obtained experimentally from the central bank's monetary policy schedule is compared, showing a distribution mode that matches the moment the decision impacts financial markets (**Figures 5 to 15** in the Appendix). For each decision day for each country, only the minute when the maximum amplitude of variation was greatest is kept. This allows for more effective comparison between decision days, as only one minute is kept.

We obtain a synthetic database of decision communications matched with exchange rate variation amplitudes at the minute when the amplitude is maximum. This panel of monetary policy decisions indexed to the country and month in which the decision is made, measured to the minute, contains two distinct surprise/non-surprise groups. The statistical unit is an amplitude of variation for one minute of communication to one day of monetary policy decision for a currency pair.

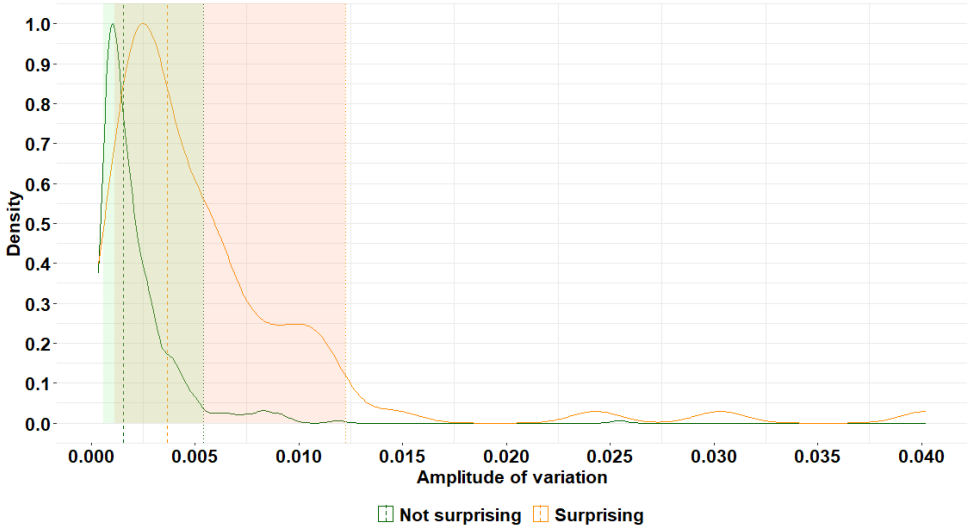
Ultimately, the panel database is unbalanced, with 510 days of decisions, 11 countries, and 78 days with surprises (see **Table 2** in the Appendix). The panel countries are Australia, Canada, the euro area, the Czech Republic, Hungary, Mexico, New Zealand, Sweden, Turkey, the United Kingdom, and the United States of America, and the period extends from January 2018 to November 2023.

e) Econometric model

We seek to measure the causal impact of a surprise on the amplitude of the exchange rate variation when monetary policy decisions are communicated. The exchange rate amplitudes of variation for surprises are more significant than for expected decisions, as shown in **Figures 2 and 3**. The groups of surprises and expected decisions are pretty distinct by design.

Density of exchange rates amplitude of variations (one-minute scale) for surprising and non-surprising decisions

Figure 2



Note: The dashed lines represent the medians, the dotted lines and coloured areas represent 95 percentiles.

Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023.

Boxplots and Wilcoxon test for surprising and non-surprising decision groups, exchange rates amplitude of variations (one-minute scale)

Figure 3



Note: The boxplot lines represent the medians and interquartile ranges.

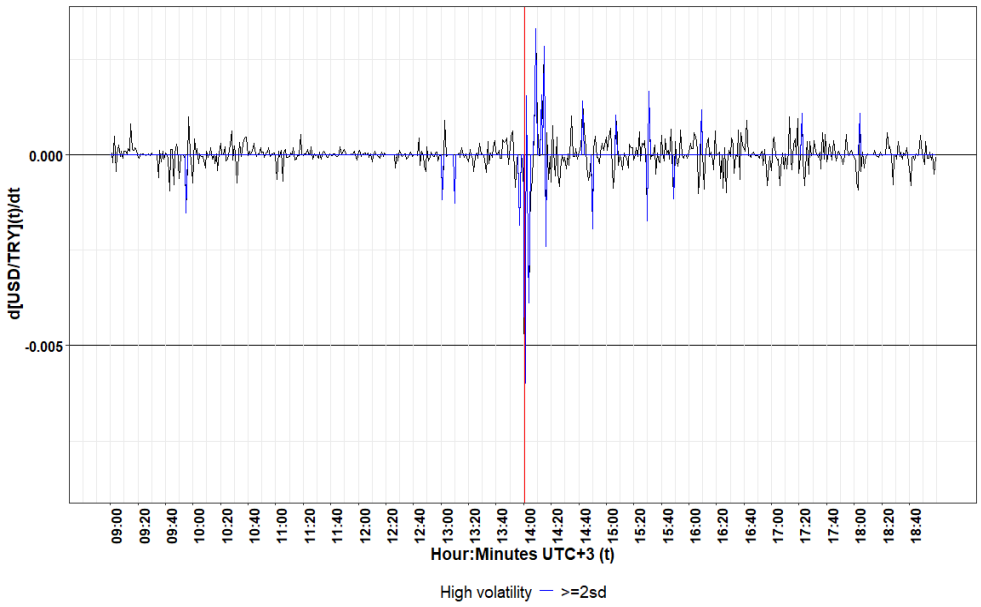
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023.

We examine the impact of the COVID-19 crisis on exchange rate variation, focusing on three sub-periods: before the crisis (2018-19), during the crisis (2020-21), and after the crisis (2022-23). The 2018-19 period saw low rates and inflation, with central banks implementing unconventional monetary policies. The 2020 crisis saw

exceptional central bank interventions, while 2021 saw faster global demand than supply due to Asian value chain disruptions. The 2022-23 period saw high post-COVID inflation, amplified by imported inflation from the Ukraine war, leading to a period of interest rate hikes.

The model incorporates a volatility variable to track market movements before the central bank's announcement. This variable is calculated from three hours before the announcement to the minute just before the announcement. If the decision is an exogenous surprise, volatility may be lower before the announcement. However, a price jump may occur at the time of the surprise, followed by high volatility. **Figure 4** shows the relative minute-to-minute changes in the USD/TRY exchange rate on 24 September 2020. Before the announcement, volatility was moderate, but after the announcement, volatility increased significantly. The logarithm of this volatility variable is kept to determine the impact of a 1% increase in volatility before the price jump on the amplitude at the time of communication.

Derivated signal of USD/TRY on 24 September 2020: the monetary policy decision of the Central Bank of the Republic of Turkey contained a surprise Figure 3



Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction.

The autoregressive process of order two (AR(2)) is considered, with the first amplitude lag added to account for the correlation between two successive amplitudes of variation for anticipated decisions. Surprise extreme amplitudes should be isolated from the amplitudes of previous and following expected decisions. In an example where a decision is surprising, the second lagged term considers the correlation between two anticipated decisions in t-1 and t+1.

For all these reasons, the econometric specification is as follows:

$$(4)\ Y_{i,d} = \alpha_i + \alpha_d + \beta_1 \mathbb{1}_{SURPRISE_{i,d}} + \beta_2 Y_{i,d-1} + \beta_3 Y_{i,d-2} + \beta_3 log(\sigma(Y))_{i,t<t_d} + \epsilon_{i,d}$$

With i a bilateral exchange rate; d the monetary policy decision day, t_d the minute for which the amplitude of variation was at its maximum during the day (corresponds

to the minute when monetary policy communication begins); t a minute of exchange rate valuation before t_d; $Y_{i,d}$ the maximum amplitude of variation on the decision day d; $Y_{i,d-1}$ and $Y_{i,d-2}$ the maximum amplitudes of variation in the two previous decision days; $\mathbb{1}_{SURPRISE_{i,d}}$ a dummy variable equal to 1 if the decision contains a surprise, and 0 otherwise; $log(\sigma(Y))_{i,t<t_d}$ a logarithmic measure of the standard deviation of amplitudes on the decision day before t_d; α_i and α_d the spatial and temporal fixed effects; $\epsilon_{i,d}$ the error term.

Turkey exhibits extreme variation in amplitudes compared with the other countries in the panel. We thus run regressions with and without Turkey to check that the surprise parameter is robust after Turkey is removed.

4. Results

Tables 3, 4 and 5 show the results of regressions for model (4) with heteroscedasticity treatment, showing highly significant regressions over the entire period and the 2018-19 and 2020-21 periods. The surprise indicator parameters range from 0.0019 to 0.0039, with a significant effect from one minute to the next. A surprising decision increases the amplitude of variation at the time of the decision from 0.19 p.p. to 0.39 p.p., depending on the period or Turkey's inclusion. The amplitude of a surprise is 100 times greater than the median amplitudes of minute-to-minute variation for the currencies in the database.

By taking a step of 15 minutes rather than one minute, we obtain close results for the parameters. Thus, over the whole period with Turkey, surprises increase the amplitude of exchange rate variation by 1.6‰ (p-value <1%) and 1.7‰ without Turkey (p-value< 1%). These results are close to those obtained for 1-minute steps, so we deduce that our identification method by exploring the signal and identifying the moments when price jumps appear is robust.

Impact of surprises on the amplitude of exchange rate variation (one-minute scale) at the time of the decision (within models) Table 3

	(1) Whole period	(3) Before Covid19 (2018-2019)	(4) During Covid19 (2020-2021)	(5) After Covid19 (2022-2023)
$\mathbb{1}_{SURPRISE_{i,d}}$	0.0024***	0.0039**	0.0021***	0.0009
	(0.0005)	(0.0012)	(0.0004)	(0.0005)
$log(\sigma(Y)_{i,t<t_d}$	0.0016*	0.0020	0.0017*	0.0005
	(0.0008)	(0.0013)	(0.0007)	(0.0004)
$Y_{i,d-1}$	0.2766***	0.4531***	0.0944	0.0730
	(0.0614)	(0.0850)	(0.0550)	(0.0548)
$Y_{i,d-2}$	0.0850**	-0.0706	0.2933*	-0.0491
	(0.0326)	(0.0571)	(0.1221)	(0.0725)
Observations	476	150	183	143
R^2	26%	37%	28%	4%
Adjusted R^2	24%	31%	22%	-6%
Statistics	137.99***	68.47***	8.87***	2.09

Time and country fixed effects are included in all estimations. Heteroskedasticity-consistent robust standard errors are reported within parentheses. Statistical significance: *p < 0.10; **p < 0.05; ***p < 0.01.

Impact of surprises on the amplitude of exchange rate variation (one-minute scale) at the time of the decision (within models), without Turkey Table 4

	(1) Whole period	(3) Before Covid19 (2018-2019)	(4) During Covid19 (2020-2021)	(5) After Covid19 (2022-2023)
$\mathbb{1}_{SURPRISE_{i,d}}$	0.0019***	0.0023**	0.0019***	0.0009
	(0.0003)	(0.0008)	(0.0003)	(0.0006)
$log(\sigma(Y)_{i,t<t_d}$	0.0008*	0.0003	0.0010**	0.0006
	(0.0003)	(0.0002)	(0.0004)	(0.0004)
$Y_{i,d-1}$	0.0930**	-0.0465	-0.0451	-0.0113
	(0.0328)	(0.0560)	(0.0588)	(0.0651)
$Y_{i,d-2}$	0.1420***	0.0395	-0.0135	0.0153
	(0.0330)	(0.0827)	(0.0513)	(0.1139)
Observations	435	135	160	140
R^2	21%	22%	30%	4%
Adjusted R^2	18%	14%	24%	-6%
Statistics	30.79***	4.31**	10.29***	1.50

Time and country fixed effects are included in all estimations. Heteroskedasticity-consistent robust standard errors are reported within parentheses. Statistical significance: *p < 0.10; **p < 0.05; ***p < 0.01.

Summary of the parameters obtained for the surprise variable (within models) Table 5

	With Turkey	Without Turkey
Whole period	0.0024***	0.0019***
	(0.0005)	(0.0003)
Before Covid19 (2018-2019)	0.0039**	0.0023**
	(0.0012)	(0.0008)
During Covid19 (2020-2021)	0.0021***	0.0019***
	(0.0004)	(0.0003)
After Covid19 (2022-2023)	0.0009	0.0009
	(0.0005)	(0.0006)

Statistical significance: *p < 0.10; **p < 0.05; ***p < 0.01

The regression model is highly significant over the whole period with Turkey, with an adjusted R^2 of 24% (see Table 3). According to the model, a surprising decision increases the amplitude of variation at the time of the decision by 0.24 p.p.

Once Turkey is removed from the sample, monetary surprise increases the amplitude of variation during communication by 0.19 p.p. The parameter is slightly lower than with Turkey, but it remains robust.

The regression model with Turkey is highly significant for 2018-19, with an adjusted R^2 of 31%. The only significant parameters are the surprise indicator and one lag parameter. The surprise parameter is stronger for 2018-19 than the whole period, as Turkey made surprise monetary policy decisions from April to September 2018, resulting in extremely high amplitudes. The Bank of Turkey implemented massive rate hikes to support its depreciating currency and counter potential imported inflation. If Turkey is removed from the 2018-19 sub-period database, the surprise parameter is highly significant at 0.23 p.p., confirming a surprise effect close to 0.2 p.p.

Two regressions with and without Turkey examine the 2020-21 period, marked by the COVID-19 crisis. The adjusted R^2 is 22% and 24%, respectively. Both models are highly significant, with robust surprise parameters and significant volatility variables due to the high uncertainty resulting from the crisis, indicating high market volatility.

Finally, the models for 2022-23 with and without Turkey are insignificant, with adjusted R^2 values of -6% for both regressions. However, the results must be qualified because, as shown in **Table 6**, there were fewer surprises over this period than in the other two.

Distribution of surprising and unsurprising decisions — Table 6

	2018	2019	2020	2021	2022	2023
Expected Decisions	77	76	70	81	68	60
Surprising Decisions	13	13	22	12	12	6

Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction Period: January 2018 - November 2023

Uncovered Interest Rate Parity (UIRP) measures the uncertainty in a currency value based on exchange rate variation over time. However, when the sign aligns with UIRP, the effect on interest rate variation is weaker than expected. The credibility of a central bank and the depth of the capital market significantly influence decision communications. A less credible central bank is more likely to make surprise decisions, while a deeper market limits variation after a surprise. The forex market's efficiency is crucial, as exchange rate adjustment is immediate after a surprise.

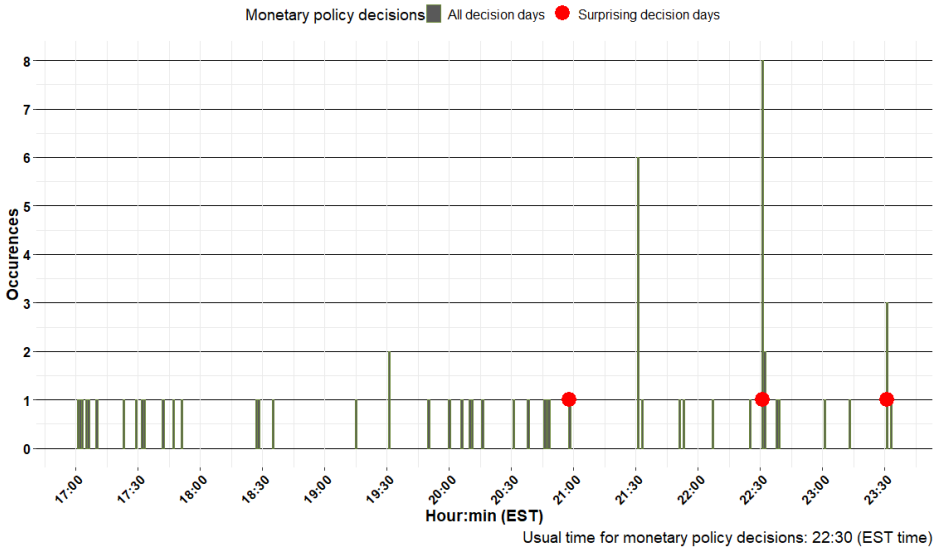
5. Conclusion

This article examines the impact of surprising monetary policy decisions on exchange rate variation when communicated. It analyses 746 economic press articles to identify monetary surprises. A foreign central bank's change in key interest rates leads to a significant currency evolution. The paper reveals that a surprise decision increases the

amplitude of exchange rate variation from 0.19 percentage point to 0.39 percentage point. The magnitude of a price jump from one minute to the next can be one per thousand or even one percent, which is 100 to 1,000 times greater than the median magnitude order of the minute-to-minute relative variation of exchange rates calculated over a year of valuation. The article aims to help central banks and financial markets professionals understand the impact of monetary surprises on exchange rates, which are essential instruments for international trade. A large language model (LLM) was used to check article tagging, improving the database's quality.

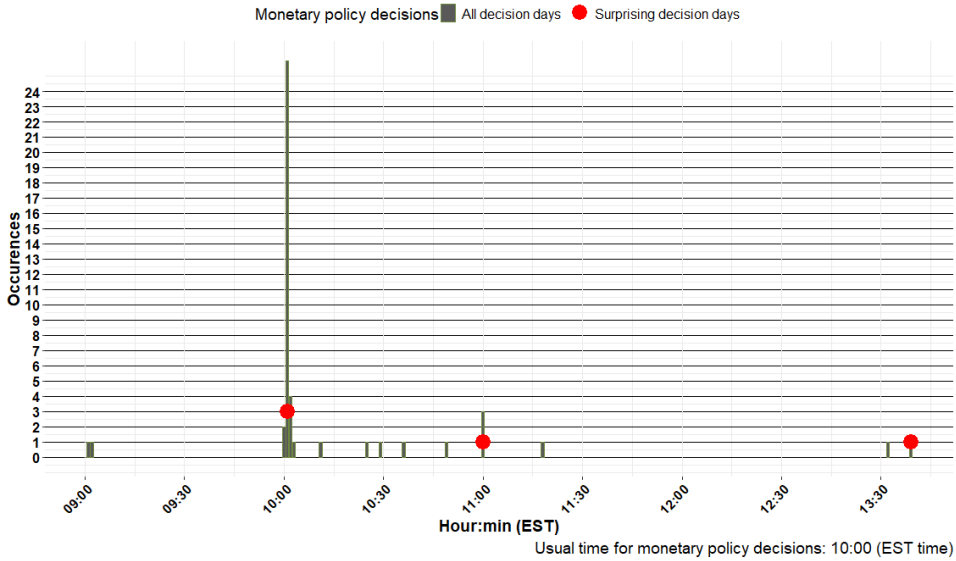
6. Appendix: A

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Australia** Figure 5



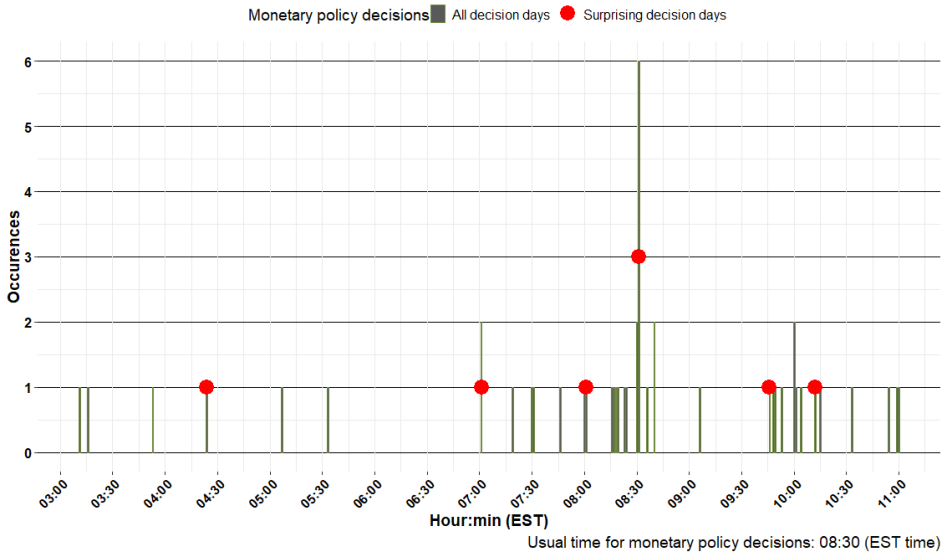
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Canada** Figure 6



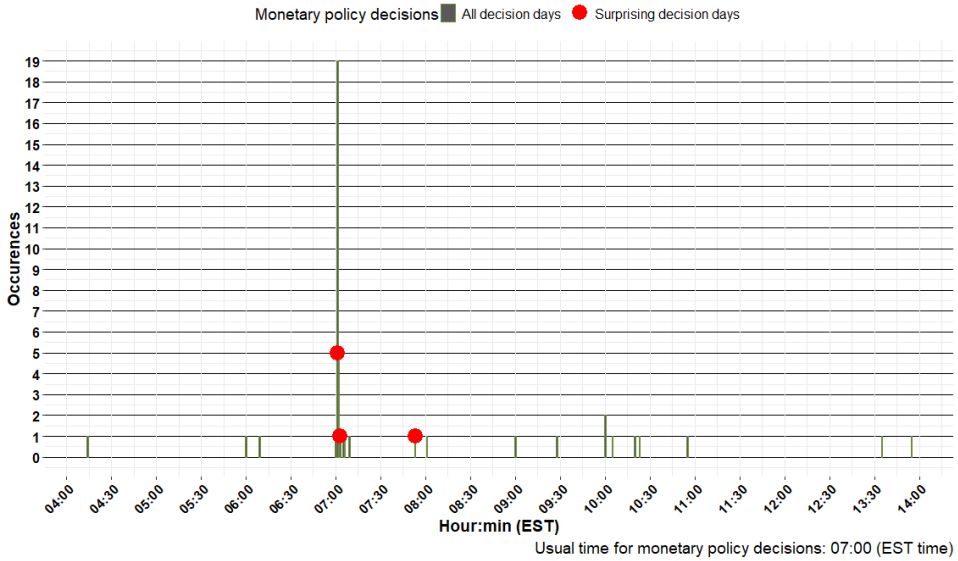
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Czech Republic** Figure 7



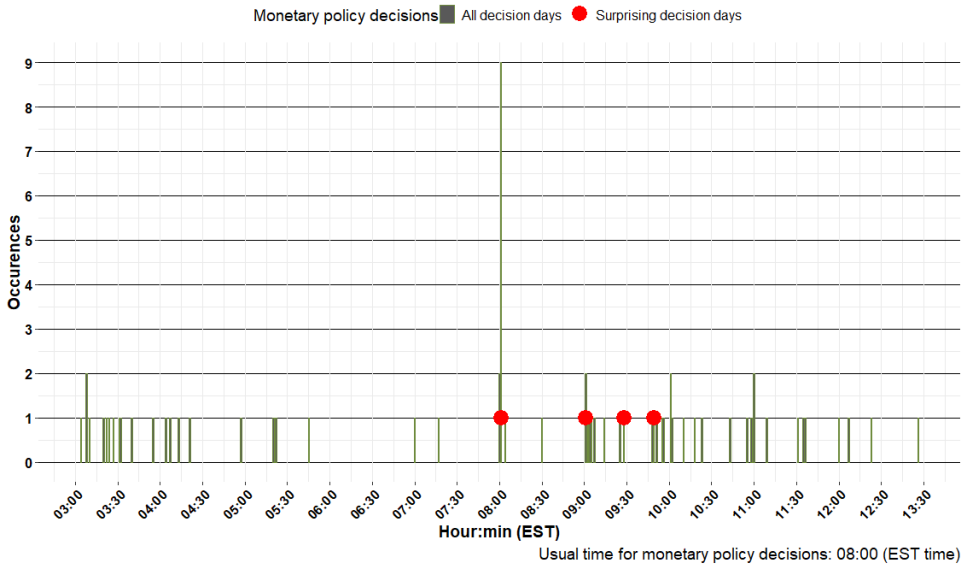
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **United Kingdom** Figure 8



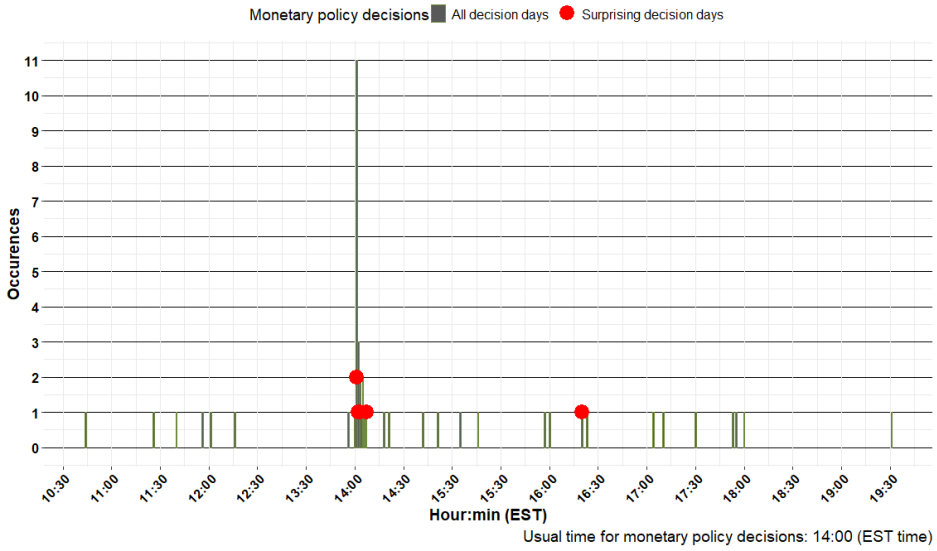
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Hungary** Figure 9



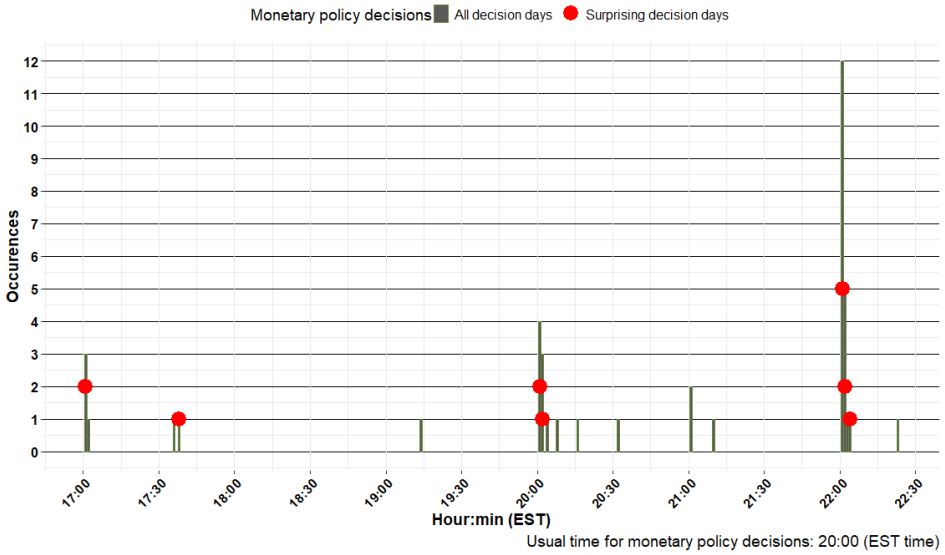
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Mexico** Figure 10



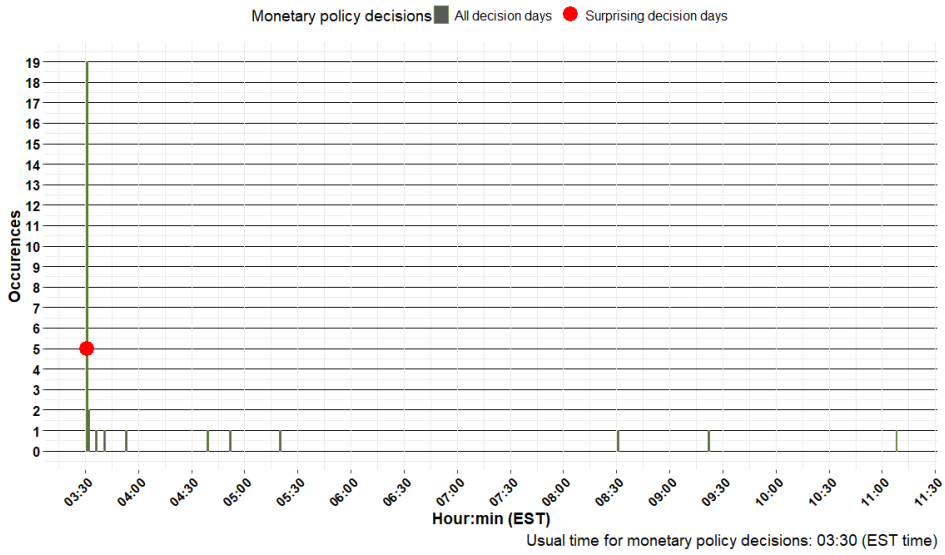
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **New-Zealand** Figure 11



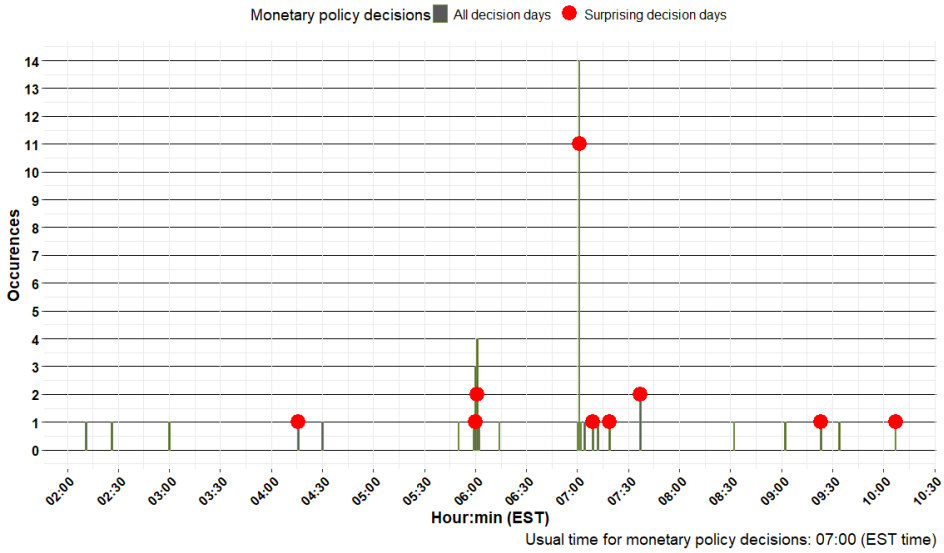
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Sweden** Figure 12



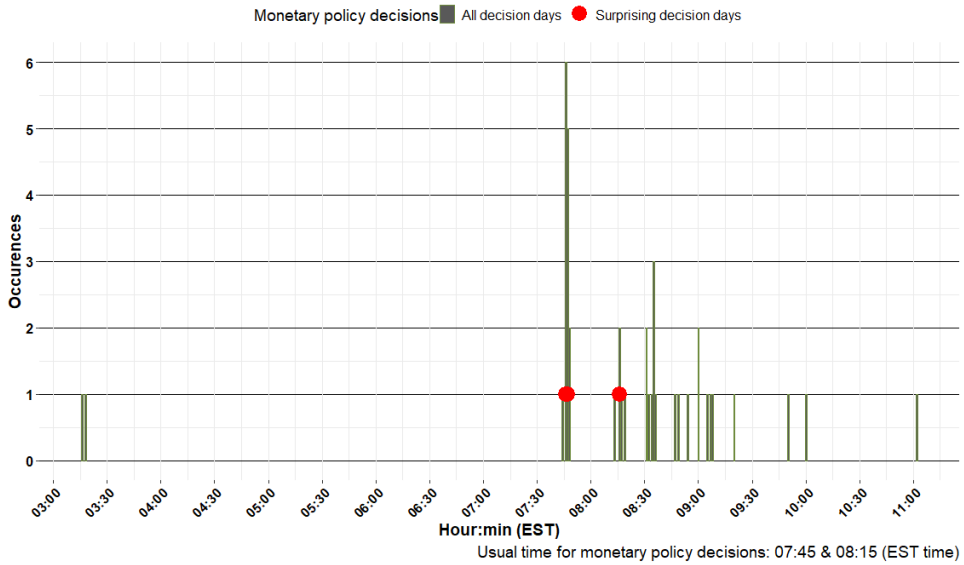
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Turkey** Figure 13



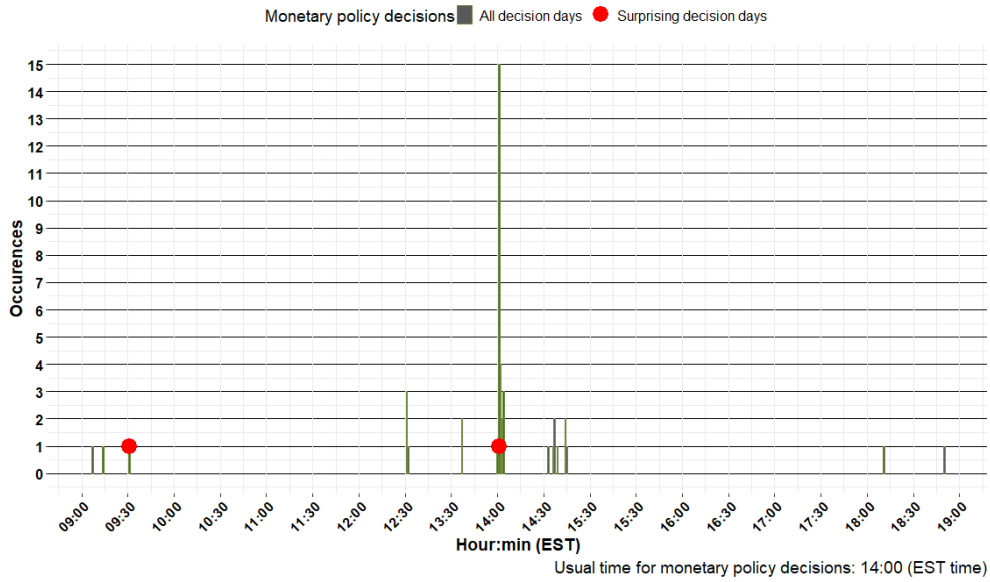
Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **Euro Area** Figure 14



Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Hours and minutes of decision days for which variation amplitudes were maximums (one-minute scale): **United States of America** Figure 15



Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Distribution of relative variations on a one-minute scale from January 2018 to November 2023 for the different pairs of the database: quantiles, number of observations

Table 1

Pair	Minimum	0.1-percentile	1-percentile	Median	99-percentile	99.9-percentile	Maximum	nb obs.
USD/AUD	$-1.4*10^{-2}$	$-1.1*10^{-3}$	$-5.0*10^{-4}$	$7.5*10^{-5}$	$5.0*10^{-4}$	$1.1*10^{-3}$	$1.8*10^{-2}$	2 134 534
USD/CAD	$-7.6*10^{-3}$	$-7.1*10^{-4}$	$-3.4*10^{-4}$	$5.5*10^{-5}$	$3.4*10^{-4}$	$6.9*10^{-4}$	$9.1*10^{-3}$	2 132 656
USD/CZK	$-1.1*10^{-2}$	$-1.4*10^{-3}$	$-5.7*10^{-4}$	$7.0*10^{-5}$	$5.5*10^{-4}$	$1.1*10^{-3}$	$2.2*10^{-2}$	2 090 922
USD/GBP	$-1.2*10^{-2}$	$-9.9*10^{-4}$	$-4.5*10^{-4}$	$7.1*10^{-5}$	$4.5*10^{-4}$	$9.8*10^{-4}$	$1.6*10^{-2}$	2 140 865
USD/HUF	$-1.4*10^{-2}$	$-1.6*10^{-3}$	$-6.6*10^{-4}$	$6.8*10^{-5}$	$6.4*10^{-4}$	$1.4*10^{-3}$	$2.1*10^{-2}$	1 952 172
USD/MXN	$-1.9*10^{-2}$	$-1.5*10^{-3}$	$-6.7*10^{-4}$	$7.7*10^{-5}$	$6.6*10^{-4}$	$1.4*10^{-3}$	$2.2*10^{-2}$	2 085 470
USD/NZD	$-1.6*10^{-2}$	$-1.1*10^{-3}$	$-5.0*10^{-4}$	$8.9*10^{-5}$	$5.0*10^{-4}$	$1.1*10^{-3}$	$1.5*10^{-2}$	2 131 854
USD/SEK	$-1.3*10^{-2}$	$-1.4*10^{-3}$	$-5.9*10^{-4}$	$7.0*10^{-5}$	$5.8*10^{-4}$	$1.2*10^{-3}$	$2.2*10^{-2}$	2 122 857
USD/TRY	$-9.1*10^{-2}$	$-3.5*10^{-3}$	$-1.1*10^{-4}$	$6.0*10^{-5}$	$1.0*10^{-3}$	$3.2*10^{-3}$	$4.8*10^{-2}$	1 796 598
USD/EUR	$-1.2*10^{-2}$	$-9.9*10^{-4}$	$-4.5*10^{-4}$	$5.4*10^{-5}$	$4.5*10^{-4}$	$9.8*10^{-4}$	$1.6*10^{-2}$	2 140 865

Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

Number of decisions and average amplitude of variation (one-minute scale) for decisions with and without surprises by country

Table 2

Country	Decisions without surprises		Decisions with surprises		All decisions	
	Average amplitude of variation	Number of decisions	Average amplitude of variation	Number of decisions	Average amplitude of variation	Number of decisions
Australia	$1.2*10^{-3}$	57	$6.3*10^{-3}$	3	$1.4*10^{-3}$	60
Canada	$2.2*10^{-3}$	41	$4.3*10^{-3}$	5	$2.5*10^{-3}$	46
Czech Republic	$1.5*10^{-3}$	38	$3.8*10^{-3}$	8	$1.9*10^{-3}$	46
Euro Zone	$1.9*10^{-3}$	38	$2.7*10^{-3}$	3	$1.8*10^{-3}$	41
Hungary	$2.1*10^{-3}$	65	$3.1*10^{-3}$	4	$2.1*10^{-3}$	69
Mexico	$1.6*10^{-3}$	40	$2.8*10^{-3}$	6	$1.7*10^{-3}$	46
New-Zealand	$3.4*10^{-3}$	27	$5.6*10^{-3}$	14	$4.2*10^{-3}$	41
Sweden	$3.1*10^{-3}$	25	$5.8*10^{-3}$	5	$3.5*10^{-3}$	30
Turkey	$4.3*10^{-3}$	22	$9.4*10^{-3}$	21	$6.8*10^{-3}$	43
United-Kingdom	$2.4*10^{-3}$	37	$2.9*10^{-3}$	7	$2.5*10^{-3}$	44
United States of America	$2.1*10^{-3}$	42	$3.7*10^{-3}$	2	$2.2*10^{-3}$	44
Total	$2.4*10^{-3}$	432	$5.6*10^{-3}$	78	$2.7*10^{-3}$	510

Sources: Free Forex Data (HistData.com), Central bank websites, Focus Economics, Factiva Extraction, Period: January 2018 - November 2023

7. Acknowledgements

We want to thank **Olivier de Bandt, Davide Furceri, Refet S. Gürkaynak, Christophe Hurlin, Matthieu Picault, Jean-Paul Renne, Antoine Sigwalt, and Xin Zhang, for their comments on previous versions of the paper, as well as participants in seminars and conferences (e.g., Banque de France, European Central Bank, University of Orléans, and Irving Fisher Committee conference organized in Rome by Banca d'Italia and the Bank for International Settlements (BIS)).**

We also thank **Marwan Menaa** for his assistance with data processing.

8. References

Aubrun C., Morel R., Benzaquen M. & Bouchaud J.-P. (2024), Riding Wavelets: A Method to Discover New Classes of Price Jumps, arXiv.

Bacchetta M., Bekkers E., Piermartini R., Rubinova S., Stolzenburg V. & Xu A. (2021), COVID-19 and global value chains: A discussion of arguments on value chain organisation and the role of WTO, WTO Staff Working Paper No. ERSD-2021-3.

Baker S. R., Bloom N. & Davis S. J. (2016), Measuring economic policy uncertainty, The Quarterly Journal of Economics, Oxford University Press, Vol. 131, 1593-1636, March.

Blattner T., Catenaro M., Erhmann M., Strauch R. & Turunen J. (2008), The predictability of monetary policy, European Central Bank Occasional Paper Series, No 83, March.

Bollerslev T. (1986), Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics, Vol. 31, Issue 3, 307-327, April.

Caldara D., Conlisk S., Iacoviello M. & Penn M. (2022), The Effect of the War in Ukraine on Global Activity and Inflation, FEDS Notes. Washington: Board of Governors of the Federal Reserve System, May 27.

Camanho N., Hau H. & Rey H. (2022), Global Portfolio Rebalancing and Exchange Rates, The Review of Financial Studies, Vol. 35, Issue 11, 5228–5274, November.

Champagne J. & Sekkel R. (2018), Changes in monetary regimes and the identification of monetary policy shocks: Narrative evidence from Canada, Journal of Monetary Economics, 99, 72-87.

Choi S., Willems T. & Yong Yoo S. (2024), Revisiting the monetary transmission mechanism through an industry-level differential approach, Journal of Monetary Economics, 145.

Deb P., Estefania-Flores J., Firat M., Furceri D. & Kothari S. (2023), Monetary Policy Transmission Heterogeneity: Cross-Country Evidence, IMF Working Paper, No. WP 23/204.

Duarte F. M. & Rosa C. (2013), A Way With Words: The Economics of the Fed's Press Conference, Liberty Street Economics, Federal Reserve Bank of New York, November 25.

Enders Z., Hünnekes F. & Müller G. J. (2019), Monetary policy announcements and expectations: Evidence from German firms, Journal of Monetary Economics, Vol. 108, 45-63.

Engle R. (2001), GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics, Journal of Economic Perspectives, Vol. 15, No. 4, Fall, 157-168.

Evstigneeva A., Shchadilova Y. & Sidorovskiy M. (2022), The Role of Communication and Information in the Emergence of Surprises in Bank of Russia Monetary Policy, Bank of Russia Working Paper, No. 99, August.

Fama E. (1970), Efficient capital markets: a review of theory and empirical work, The Journal of Finance, Vol. 25, No. 2, May.

Furceri D., Loungani P. & Zdzienicka A. (2018), The effects of monetary policy shocks on inequality, Journal of International Money and Finance, 85, 168-186.

Gürkaynak R. S., Sack B. & Swanson E. (2005), Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements, International Journal of Central Banking, Vol.1, May.

Gürkaynak R. S., Kara A. K., Kısacıkoğlu B. & Lee S. S. (2021), Monetary Policy Surprises and Exchange Rate Behavior, Journal of International Economics, Vol. 130, May.

Jarociński, M. & Karadi P. (2020), Deconstructing monetary policy surprises—the role of information shocks, American Economic Journal: Macroeconomics, Vol. 12(2), 1-43.

Kandil M. E. & Mirzaie I. A. (2003), The Effects of Exchange Rate Fluctuations on Output and Prices: Evidence From Developing Countries, IMF Working Paper, 03/200, October.

Kearns J. & Manners P. (2006), The Impact of Monetary Policy on the Exchange Rate: A study Using Intraday Data, International Journal of Central Banking, Vol. 2, Issue 4.

Keynes, J.M. (1923), A Tract on Monetary Reform, MacMillan and Co., Ltd.

Kuttner K. N. (2001), Monetary policy surprises and interest rates: Evidence from the Fed funds futures market, Journal of Monetary Economics, Vol. 47, Issue 3, 523-544, June.

Leombroni M., Vedolin A., Venter G. & Whelan P. (2021), Central bank communication and the yield curve, Journal of Financial Economics, Vol. 141, Issue 3, 860-880, September.

McCallum B.T. (1994), A reconsideration of the uncovered interest parity relationship, Journal of Monetary Economics, Vol. 33, Issue 1, 105-132, February.

Miranda-Agrippino S. & Ricco G. (2021), The transmission of monetary policy shocks, American Economic Journal: Macroeconomics, 13, No. 3, 74-107.

Rosa C. & Verga G. (2008), The Impact of Central Bank Announcements on Asset Prices in Real Time, International Journal of Central Banking, Vol.4, 175-217, June.

Romer C. D. & Romer D. H. (2004), A new measure of monetary shocks: Derivation and implications, American Economic Review, No. 4, 1055-1084.

The Impact of Monetary Surprises on Exchange Rates: Insights from a Textual Analysis Approach on a Panel of Countries

Working Paper

Jean-Charles Bricongne [1], Louis Marolleau [2]

[1]Banque de France, Paris I University, Laboratory of Economics of Orléans, France. jean-charles.bricongne@banque-france.fr
[2]Univ. Lille, CNRS, IESEG School of Management, Lille Économie Management, France. louis.marolleau@univ-lille.fr

Bank of International Settlements: 4th IFC Workshop on Data Science in Central Banking - Rome, Bank of Italy - February 2025

1. Introduction
2. Literature Review
3. Data
4. Econometric modelling
5. Results
6. Discussion
7. Conclusion

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Introduction



Monetary surprise : a monetary policy decision that contradicts the expectations of the financial markets regarding this decision.

图 1: Article of Financial Times about a surprise rate hike of The Central Bank of the Republic of Turkey on 24 September 2020

Introduction



图 2: FocusEconomics article on 24 September 2020 monetary policy communication of The Central Bank of the Republic of Turkey: the decision contained a surprise

Introduction

- Exchange rates react instantly to surprising monetary policy announcements
- Transmission of monetary policy: exchange rate channel
- Uncovered Interest Rate Parity describes an absence of arbitrage opportunities: a rise (fall) in domestic key interest rates is accompanied by an appreciation (depreciation) of the domestic currency through immediate purchases (sales) on Forex market.

Introduction

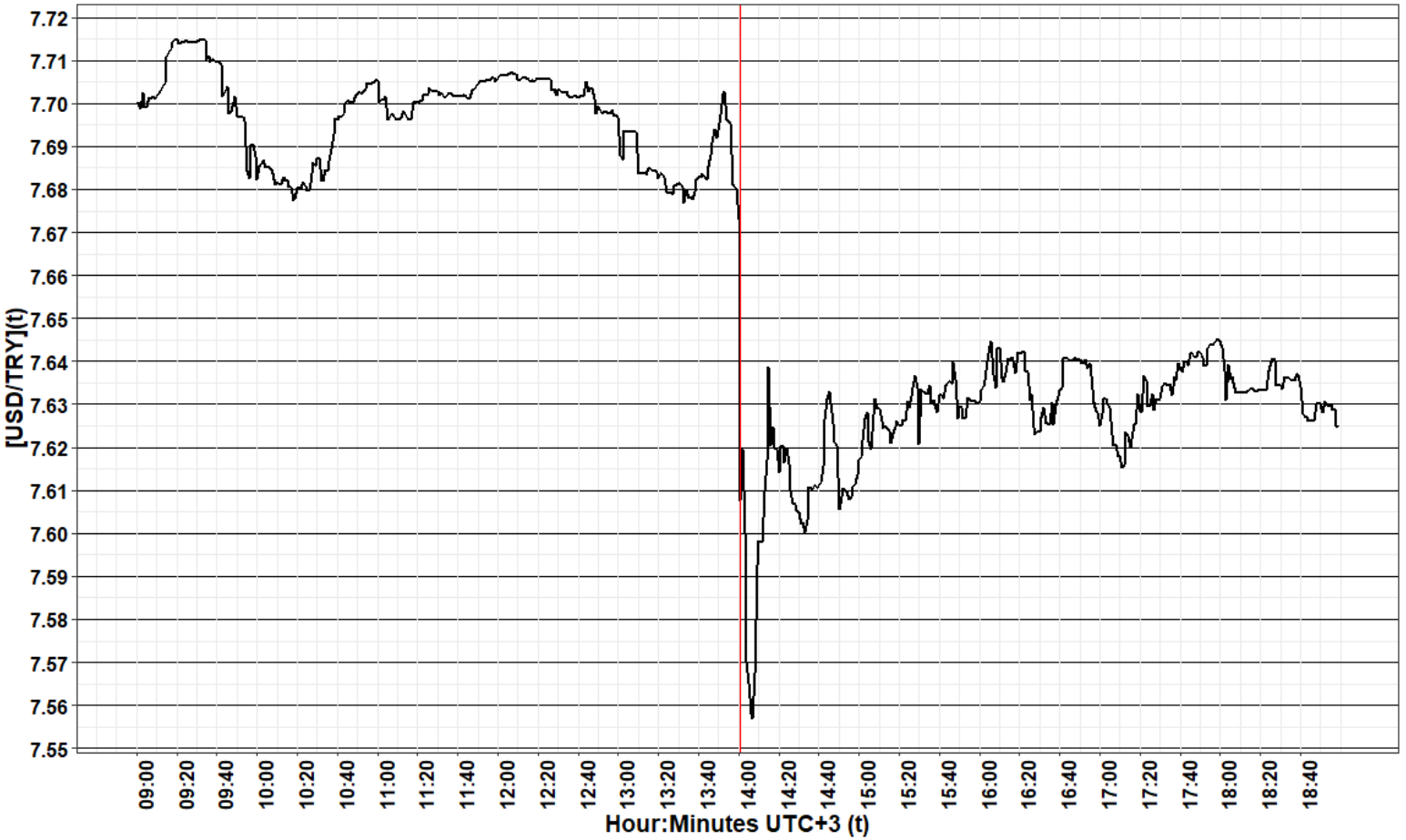


图 3: US Dollar against 1 Turkish Lira on 24 September 2020 : the Turkish Central Bank's monetary policy decision contained a surprise

Introduction

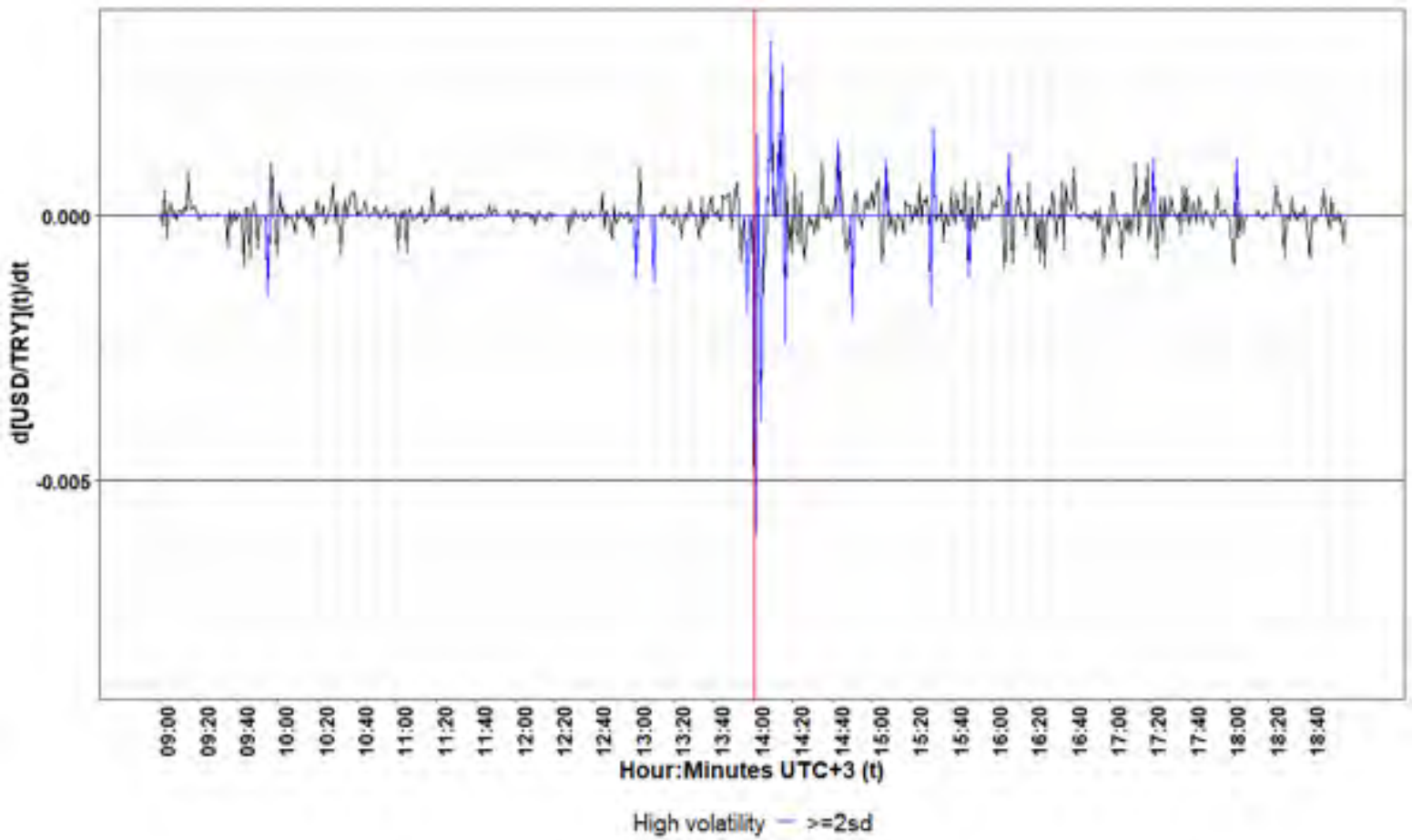


图 4: 24 September 2020 Central Bank of the Republic of Turkey surprise : derived signal

Introduction

1. Several methods used in the litterature to identify surprises :
   - Prices of option contracts on fed funds futures: used to determine financial market expectations regarding future changes in the Fed's monetary policy
   - Overnight Index Swaps
2. These methods have their limitations

Introduction

Research question: How can a textual analysis method applied to press articles identify surprises and measure their causal effect on exchange rates?

Introduction

Method to answer that :

- Database of press articles
- Calendar of monetary policy decisions
- Measuring the up-to-the-minute effect of monetary surprises on the exchange rate by comparing two groups : suprising (*treatment*) and not surprising (*control*) decisions

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Litterature Review I

[1] Davis S.J. Baker S.R., Bloom N.
Measuring economic policy uncertainty.
The Quarterly Journal of Economics, Oxford University Press, 131:1593–1636, March 2016.

[2] Sack B. Swanson E. Gürkanyak R. S.
Do actions speak louder than words? the response of asset prices to monetary policy actions and statements.
International Journal of Central Banking, 1, May 2005.

[3] Kuttner K. N.
Monetary policy surprises and interest rates: Evidence from the fed funds futures market.
Journal of Monetary Economics, 47(Issue 3):523–544, June 2001.

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Data

Press articles from:

- 2 data sources for press articles referring to monetary policy decisions: *Factiva* and *FocusEconomics*. 746 articles.
- Search by keywords in the articles : ["surpris"], ["unexpect"], ["unanticipat"], ["than expect"], ["than anticipat"], ["not expect"], ["not anticipat"]
- Context window around each keyword to check that it is a monetary surprise linked to the decision in question
- Checks on the dates of monetary policy announcements in press articles.

Data

- Creation of a calendar of monetary policy decisions supplemented by a surprise indicator variable worth one if the decision contains a surprise, zero otherwise.
- 2 distinct groups of monetary policy decisions: with or without surprises
- Exchange rate data:
  - *histdata.com* : exchange rates on a one-minute scale
  - 11 pairs: AUD/USD, CAD/USD, GBP/USD, CZK/USD, HUF/USD, MXN/USD, NZD/USD, SEK/USD, TRY/USD, EUR/USD.

Data

Press articles database is matched to exchange rate data for getting the following database:

- 11 countries from January 2018 to November 2023, 510 monetary decisions including 78 surprises
- Statistical unit: a maximum amplitude of variation for one minute of monetary policy communication to one day of monetary policy decision for a currency pair

Data

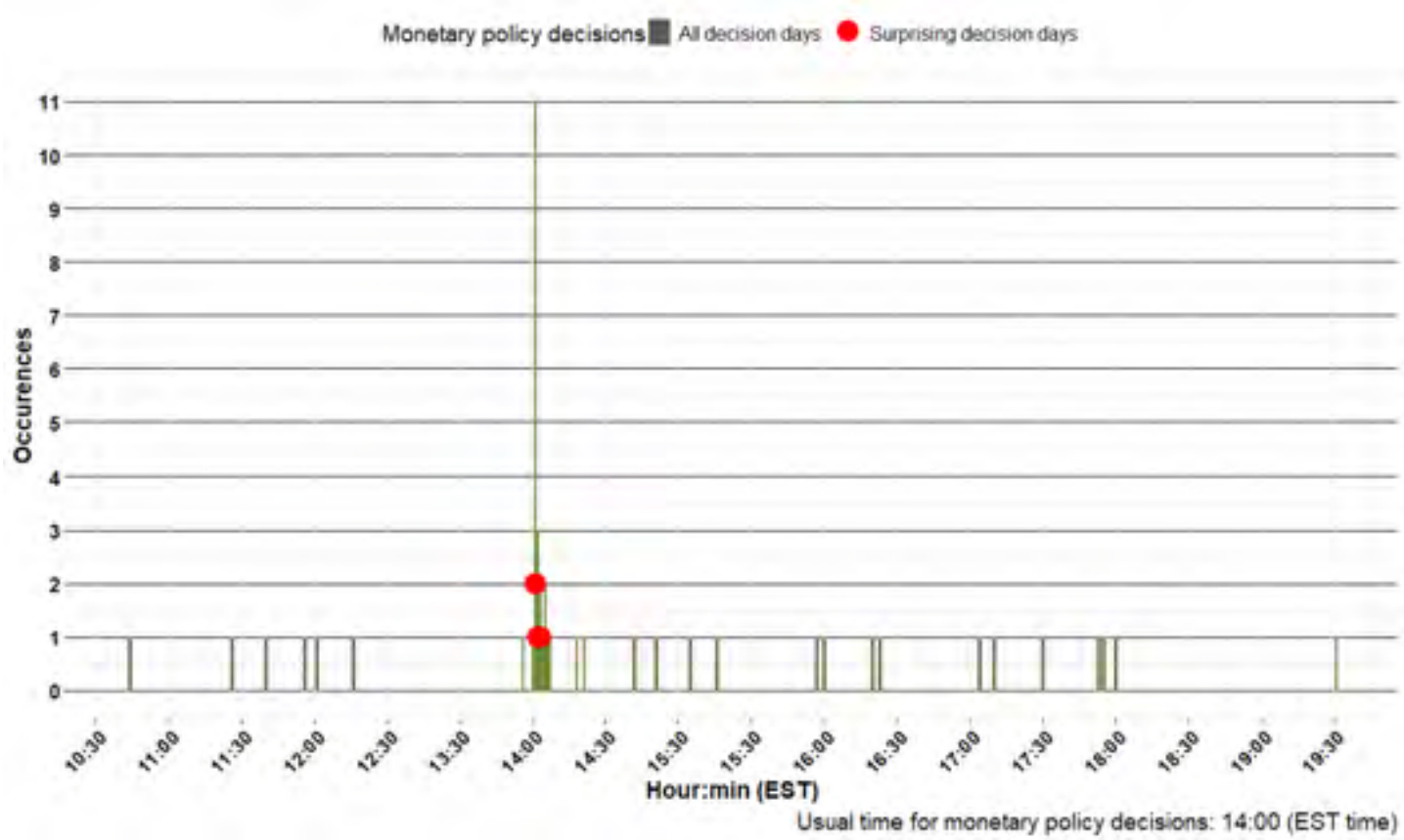


图 5: Distribution of hour:minutes for which amplitudes of variation on decision days were maximum : Mexico

Data

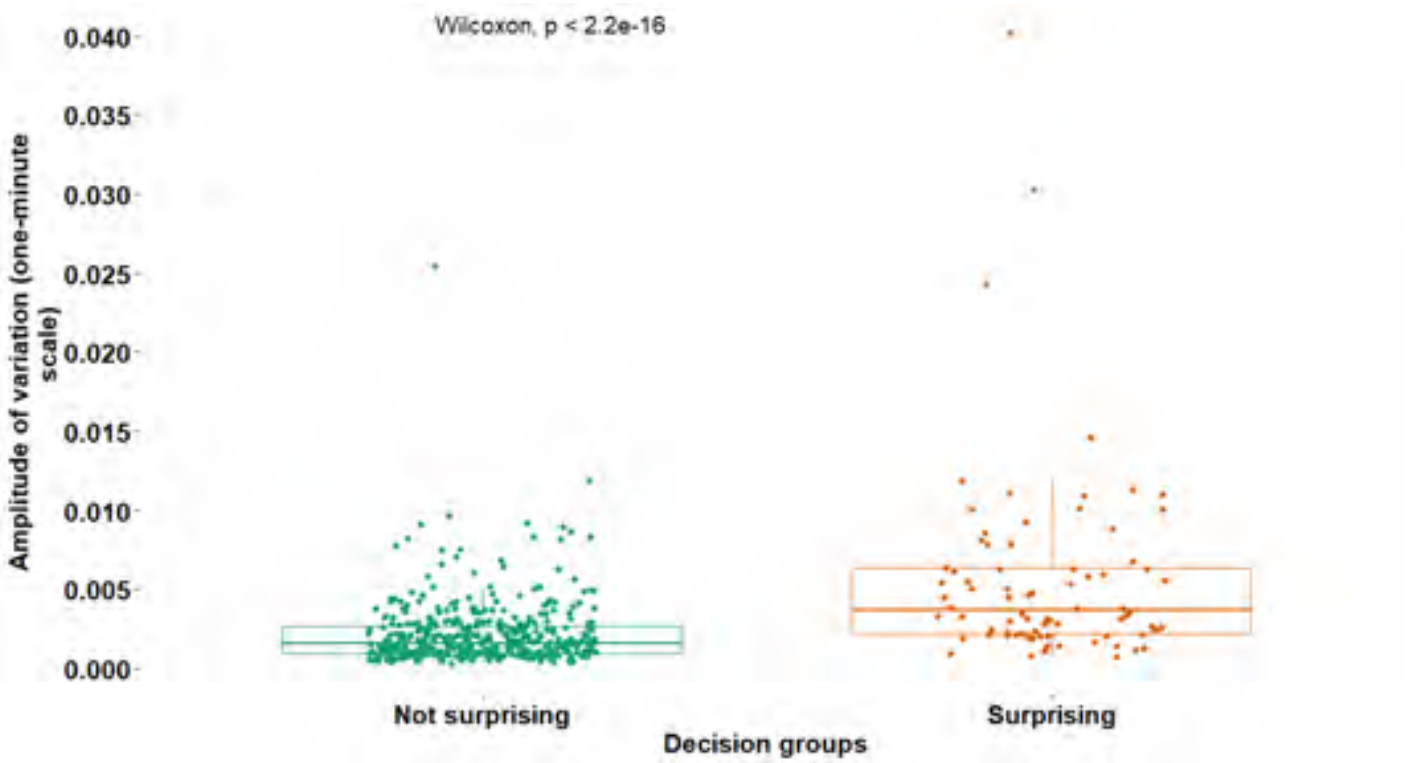


图 6: Boxplots and Wilcoxon test for surprising and non-surprising decision groups, exchange rates amplitudes of variation (one-minute scale)

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Econometric modelling

$$Y_{i,t^*} = \alpha_i + \alpha_t + \beta_1 * 1_{SURPRISE_{i,t^*}} + \beta_2 * Y_{i,t^*_{[-1]}} + \beta_3 * Y_{i,t^*_{[-2]}} + \beta_4 * log(\sigma)_{i,t<t^*} + \epsilon_{i,t^*} \quad (1)$$

With :

- i a currencies pair, t^* a minute of monetary policy decision communication
- Y_{i,t^*} the amplitude of variation of exchange rate at the minute the decision is communicated
- $1_{SURPRISE_{i,t^*}}$ an indicator variable worth one if the decision contains a surprise, zero otherwise
- $Y_{i,t^*_{[-1]}}$ and $Y_{i,t^*_{[-2]}}$ the lagged amplitude of variation at the 2 previous decisions at moments of central bank communication
- $log(\sigma)_{i,t<t^*}$ the volatility of the amplitude of variation over the day of decision before the communication minute

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Results

	With Turkey	Without Turkey
Whole period	0.0024***	0.0019***
	(0.0005)	(0.0003)
Before Covid19 (2018-2019)	0.0039**	0.0023**
	(0.0012)	(0.0008)
During Covid19 (2020-2021)	0.0021***	0.0019***
	(0.0004)	(0.0003)
After Covid19 (2022-2023)	0.0009	0.0009
	(0.0005)	(0.0006)

图 7: Summary of the parameters obtained for the surprise variable (within models). Fixed effects of time and country are included in all estimations. HC robust standard errors are reported within parentheses. Statistical significance: *p $<$ 0.10; **p $<$ 0.05; ***p $<$ 0.01.

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

New method: Natural Language Processing

- The study can also be carried out using Natural Language Processing (Large Language Models) to tag surprises more efficiently and quickly
- We have tested this method by using ChatGPT4 to classify the articles we had used
- 450 FocusEconomics articles have been tagged by ChatGPT4
- 94% of our initial tagging on these 450 days of decisions are correlated with the ChatGPT4 one

Improvements

- Natural Language Processing: use several LLMs to tag surprises efficiently, have a larger database, and add robustness:
  - Increasing the number of years and countries
  - Number of criteria for classifying surprises: Interest rate Hike/Cut/Quantitative Easing/Surprise Announcement on Forward Guidance or Monetary Normalisation/Inflation Expectation/Intensity of Surprise. These criteria can be added to the model
- Different time pace for the calculation of amplitudes:2, 5, 10 minutes

1 Introduction

2 Literature Review

3 Data

4 Econometric modelling

5 Results

6 Discussion

7 Conclusion

Conclusion

- Impact of monetary surprises on exchange rates at the time of monetary policy announcements
- Uncovered Interest Rate Parity
- Textual analysis method on business press articles. Database of monetary policy decisions: 11 countries, 2018-2023, 78 surprises
- Surprising and unexpected decisions are compared

Conclusion

- The amplitudes of variation for surprises are [0.19p.p.;0.39p.p.] higher compared to the expected decisions on a one minute scale.
- Other events in addition to monetary surprises can imply high amplitudes of variation.
- Large Language Models can be used. Other dependent variables can be studied (bond spreads, OIS, stock market indexes, etc.).

Conclusion

Thank you for your attention.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Turbulent times: capturing social unrest in the news[1]

Sydney Eck, Flora Haberkorn, April Patel, Zach Proom
and Abhinuv Uppal,
Federal Reserve Board of Governors

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

TURBULENT TIMES: CAPTURING SOCIAL UNREST IN THE NEWS

SYDNEY ECK*, FLORA HABERKORN*, APRIL PATEL*, ZACH PROOM*, ABHINUV UPPAL*

*FEDERAL RESERVE BOARD OF GOVERNORS

THE VIEWS EXPRESSED HEREIN ARE THOSE OF THE AUTHORS' AND DO NOT NECESSARILY REFLECT THE OPINIONS OF THE BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM OR ANYONE ELSE IN THE FEDERAL RESERVE SYSTEM.

LITERATURE: QUANTIFYING UNREST

- Caldara & Iacoviello 2022: Constructed a geopolitical risk (GPR) index using the coverage of major English language papers in the US, UK, and Canada.
- Barrett 2022: Constructed the Reported Social Unrest Index using the Dow Jones' Factiva news aggregator, focusing on unrest within a country and filtering out foreign policy protests.
- Redl & Hlatshwayo 2021: Constructed a social unrest risk index, where the risk of social unrest in a given year is based on the conditions of the previous year. The key drivers of risk appear to be previous unrest, food price inflation, and the popularity of cell phones.

CALDARA & IACOVIELLO 2022

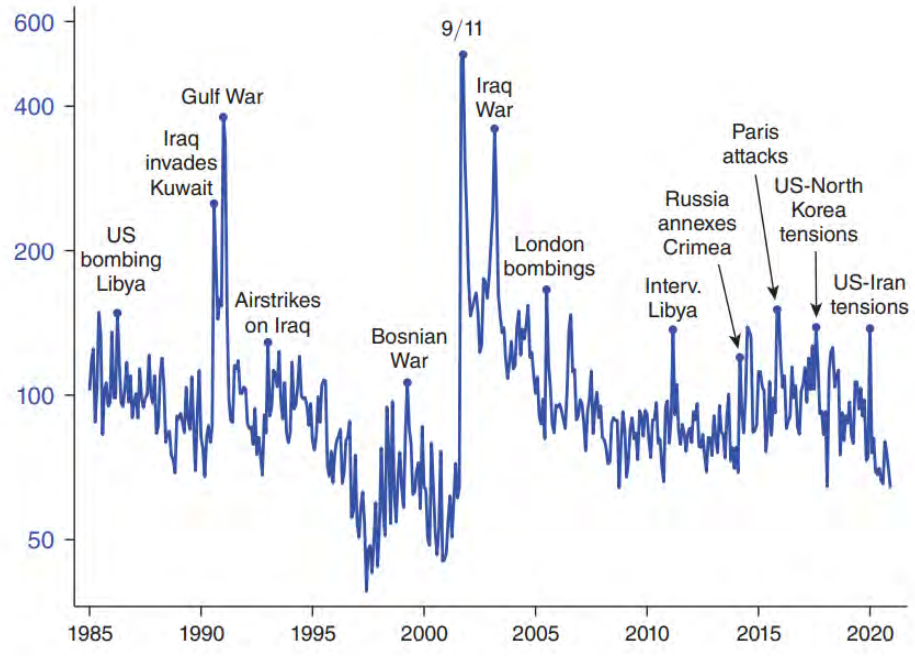


FIGURE 1. RECENT GPR INDEX FROM 1985

- Constructed a geopolitical risk (GPR) index and used that index to assess the impact of geopolitical risk on macro-economic outcomes
- The GPR is constructed by taking the share of articles that mention the threat, realization, or escalation of "adverse events" (such as war or terrorism) out of the total number of articles published in major US, Canada, and UK papers. A "mention" is defined as the presence of specific words in proximity to each other that are associated with certain event types.
  - Performed checks to verify the share of non-social unrest events (e.g. natural disasters) to ensure their index was not crowded out by other events
- They demonstrated GPR's association with lower investment, higher unemployment, larger downside risks, and a higher probability of economic disaster.

BARRETT 2022

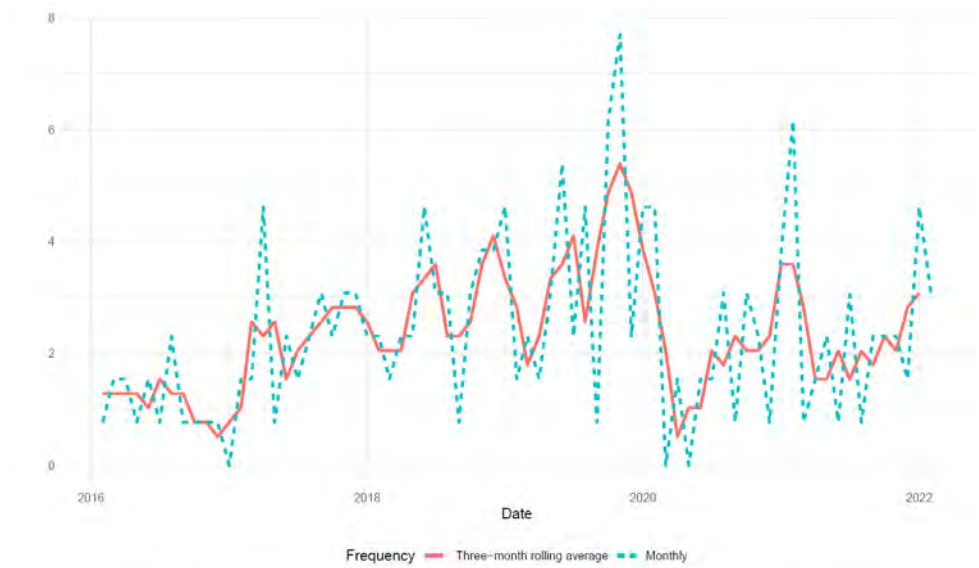


Figure 1: Fraction of countries with RSUI-identified events, 2016-2022

- Developed the Reported Social Unrest Index in order to track global unrest over time.
- The index utilizes the Dow Jones' Factiva news aggregator, focusing on 130 countries. This index is primarily focused on unrest that occurs within a given country and attempts to filter out foreign policy protests.
- To address potential bias from the news outlets selected, Barrett filters out certain articles, such as those with clear country specific bias. And to avoid retrospective rather than current data, he filters for articles that have the word "today."
- Barrett's index is primarily focused on unrest that occurs within a given country and attempts to filter out foreign policy protests.
- His index has since been used in a wide body of literature that seeks to tease out the relationship between unrest and various outcomes

LITERATURE: UNREST & ECONOMIC CONSEQUENCES

- Cerra & Saxena 2008: Civil wars and a deterioration in the quality of governance (fewer constraints on the powers of the political executive) are associated with loss of output, though only civil wars were associated with a partial rebound in output.
- Novta & Pugacheva 2020: Conflict (defined as number of deaths relative to the size of a population) is associated with drops in consumption and trade.
- Glick & Taylor 2010: International wars are associated with heavy costs including decreased volume of international trade, per capita income, and general economic welfare.
- Caldara & Iacoviello 2022: Geopolitical risk is associated with lower investment, higher unemployment, larger downside risks, and higher probability of economic disaster (used their own GPR index).
- Baker, Bloom, & Davis 2015: Economic policy uncertainty (EPU) spiked during periods of known social unrest, indicating that social unrest might be associated with negative economic outcomes (used their own EPU index).
- Hadzi-Vaskov, Pienknagura, & Ricci 2023: Unrest is associated with contractions in manufacturing and services, consumption, GDP, and trade (used Barrett's index).
- Kollias & Tzeremes 2022: Economic turmoil and unrest were often bidirectional, causing vicious cycles (used Barrett's index).

SURI: SOCIAL UNREST RISK INDEX

- SURI is a novel index for social unrest constructed using the Global Data of Events, Tone, and Language (GDELT) Project Events dataset and the Goldstein scale.
- SURI is bidirectional, meaning it captures social unrest originating in one country, conglomerate, or group and geared towards another.
  - For example, we can capture unrest triggered by civilians of a country and targeted at the military of another country or even a supranatural organization (e.g., NATO).
- The goal of this index is to provide a real-time transparent metric to visualize the risk of social unrest between two parties.

DATA GDELT

- The GDELT Events database is a large (terabyte-scale) database of global events, each categorized using the CAMEO system. The database is an output of sophisticated NLP models on news data that factors global events into fields such as
  - **who** was involved (each event contains exactly 2 actors),
  - **what** type of event happened (protest, aggression, etc.),
  - **when** the event occurred,
  - **how many** news articles reported it,
  - and much more.

- A row in the dataset could tell us "In **February 2022**, the **United States placed economic sanctions** on **Russia**, reported by **250 articles.**"
- In the above case **United States -> Russia** the United States is Actor1 and Russia is Actor2.
- **Actor1 perpetrates a given action against Actor2.** If the Russian government were to retaliate against the United States, Russia would be Actor1, as the perpetrator of the action. This allows us to filter and construct subsets of the data that focus on government/government unrest (e.g., economic sanctions), domestic government/civilian unrest (e.g., insurrections), and international government/civilian unrest (e.g., foreign policy protests).

- We specifically examine events related to **protests, demands**, and **threats.** We look at events containing both **non-government** actors (civilian groups, educational organizations, etc.) and **government** actors, defined as social unrest. We associate actors with their **country of origin.** Using these fields, we construct the bi-directional Social Unrest Risk Index (SURI) between countries.

GOLDSTEIN SCALE

- First published by Goldstein in 1992, the Goldstein scale assigns a numeric value between –10 to 8.3 to various events, scaling them from conflict (negatives), to neutral (0), to cooperation (positives)
- In order to construct the scale, Goldstein used McClelland's (1971) World Events Interaction Survey (WEIS)
  - For each WEIS event there were two actors and an action: Country A toward Country B.
  - For example, Country A gives warning to Country B
- A panel of international affairs researchers were asked to rate each WEIS event along two scales (one for conflict, one for cooperation) from 0 to 10. The Goldstein scale is an average of their responses.

TABLE 1
New Weights for WEIS Events

Event Type	*Weight*	SD
223 Military attack; clash; assault	–10.0	0.0
211 Seize position or possessions	–9.2	0.7
222 Nonmilitary destruction/injury	–8.7	0.5
221 Noninjury destructive action	–8.3	0.6
182 Armed force mobilization, exercise, display; military buildup	–7.6	1.2
195 Break diplomatic relations	–7.0	1.3
173 Threat with force specified	–7.0	1.1
174 Ultimatum; threat with negative sanction and time limit	–6.9	1.4
172 Threat with specific negative nonmilitary sanction	–5.8	1.9
193 Reduce or cut off aid or assistance; act to punish/deprive	–5.6	1.4
181 Nonmilitary demonstration, walk out on	–5.2	2.1
201 Order person or personnel out of country	–5.0	1.7
202 Expel organization or group	–4.9	1.4
150 Issue order or command, insist, demand compliance	–4.9	1.7
171 Threat without specific negative sanction stated	–4.4	1.5
212 Detain or arrest person(s)	–4.4	2.3
192 Reduce routine international activity; recall officials	–4.1	1.2
112 Refuse; oppose; refuse to allow	–4.0	1.5
111 Turn down proposal; reject protest, demand, threat	–4.0	1.5
194 Halt negotiation	–3.8	0.9
122 Denounce; denigrate; abuse	–3.4	1.1
160 Give warning	–3.0	1.3
132 Issue formal complaint or protest	–2.4	0.9
121 Charge; criticize; blame; disapprove	–2.2	1.3
191 Cancel or postpone planned event	–2.2	1.5
131 Make complaint (not formal)	–1.9	0.6
063 Grant asylum	–1.1	2.5
142 Deny an attributed policy, action, role or position	–1.1	1.0

CONSTRUCTING SURI: METHODOLOGY I

$$SURI_{i,j} = Geopolitical\ Unrest\ Score_{i,j} * Political\ Involvement\ Score_{i,j}$$

$$Geopolitical\ Unrest\ Score_{i,j} = \sum Demand, Threaten, Protest\ Events_{i,j}$$

$$Political\ Involvement\ Score_{i,j} = \left(\frac{\sum Gov, Gov_{i,j}}{\sum Events_{i,j}} + \frac{\sum NonGov, Gov_{i,j}}{\sum Events_{i,j}} + \frac{\sum Gov, NonGov_{i,j}}{\sum Events_{i,j}}\right)$$

$$i = date$$

$$j = actor1countrycode, actor2countrycode$$

CONSTRUCTING SURI: METHODOLOGY II

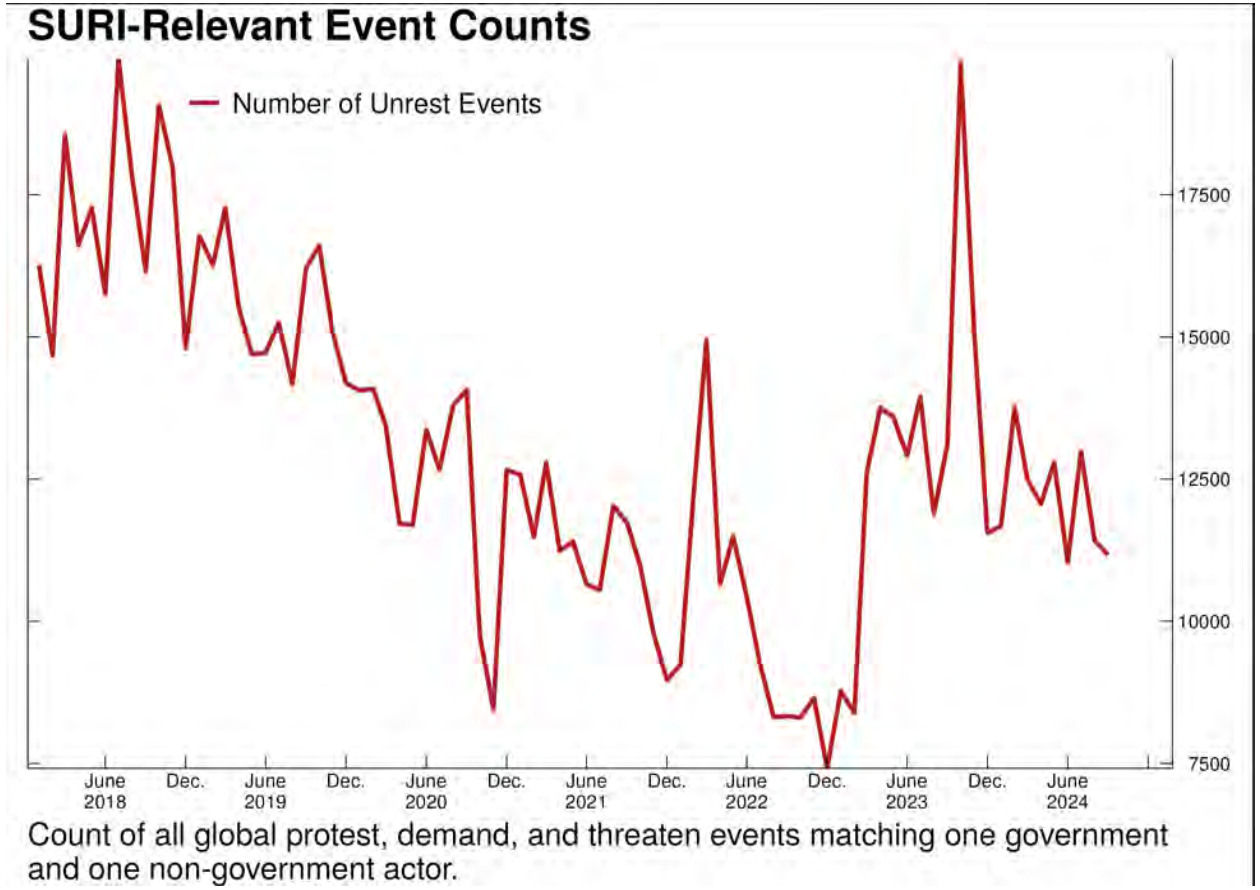
$$SURI_{i,j} = ImportanceScore_{i,j} * IntensityScore_{i,j}$$

$$ImportanceScore_{i,j} = \sum News\ Articles_{i,j}$$

$$IntensityScore_{i,j} = \frac{\sum Goldstein\ Scores_{i,j}}{\sum Events_{i,j}}$$

$$j = actor1countrycode, actor2countrycode$$

SURI EVENT COUNTS

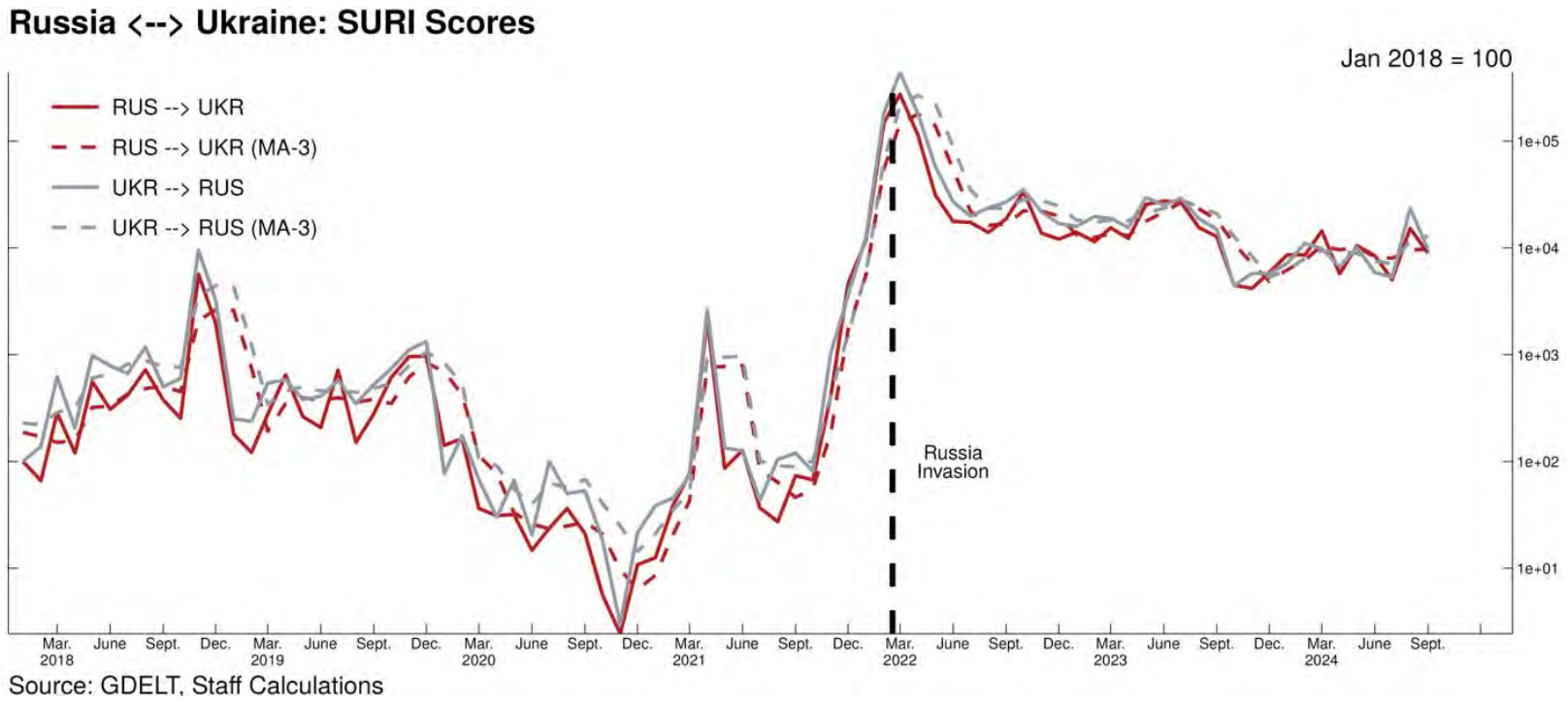


Events in each country
United States: 772
Israel: 277
United Kingdom: 217
Nigeria: 154
Russia: 149
China: 131
Iran: 105
Sudan: 93
India: 91
Kenya: 76

CASE STUDY: RUSSIAN INVASION OF UKRAINE

SURI scores between Russia and Ukraine peaked starting at Russia's invasion and have remained elevated since.

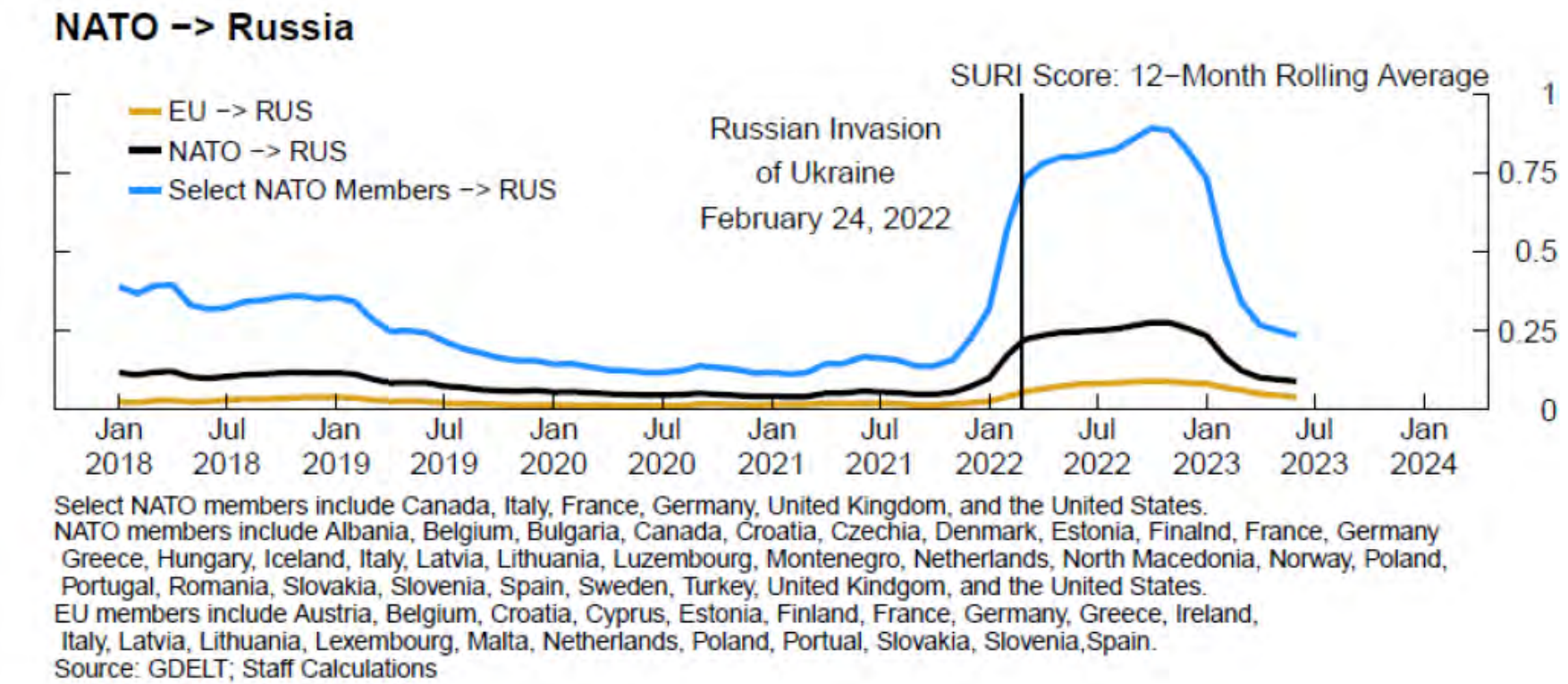
This chart is created using Methodology 1.



CASE STUDY: RUSSIAN INVASION OF UKRAINE

In this chart, the SURI is plotted for multiple groups of countries including the EU, NATO, and select NATO members

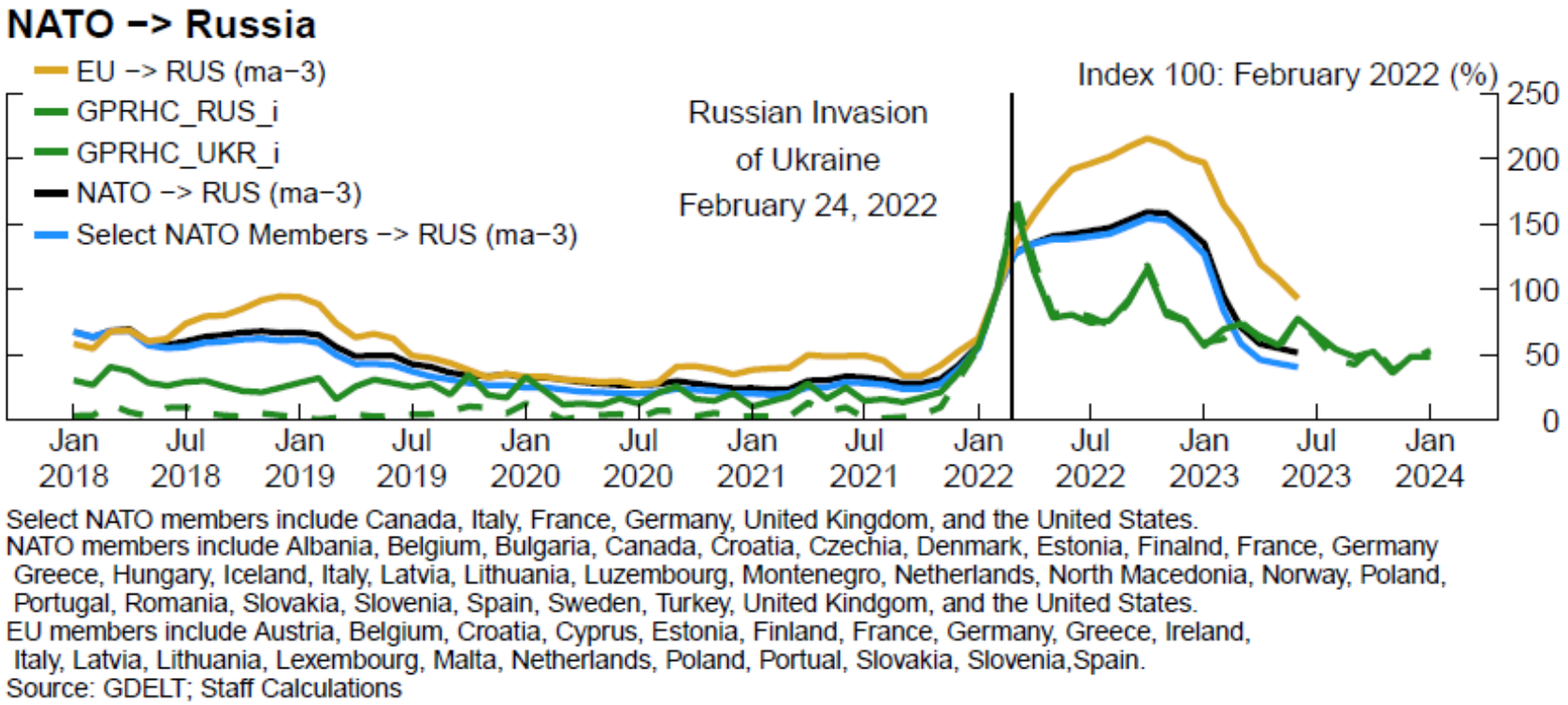
This chart is created using Methodology 1.



CASE STUDY: RUSSIAN INVASION OF UKRAINE

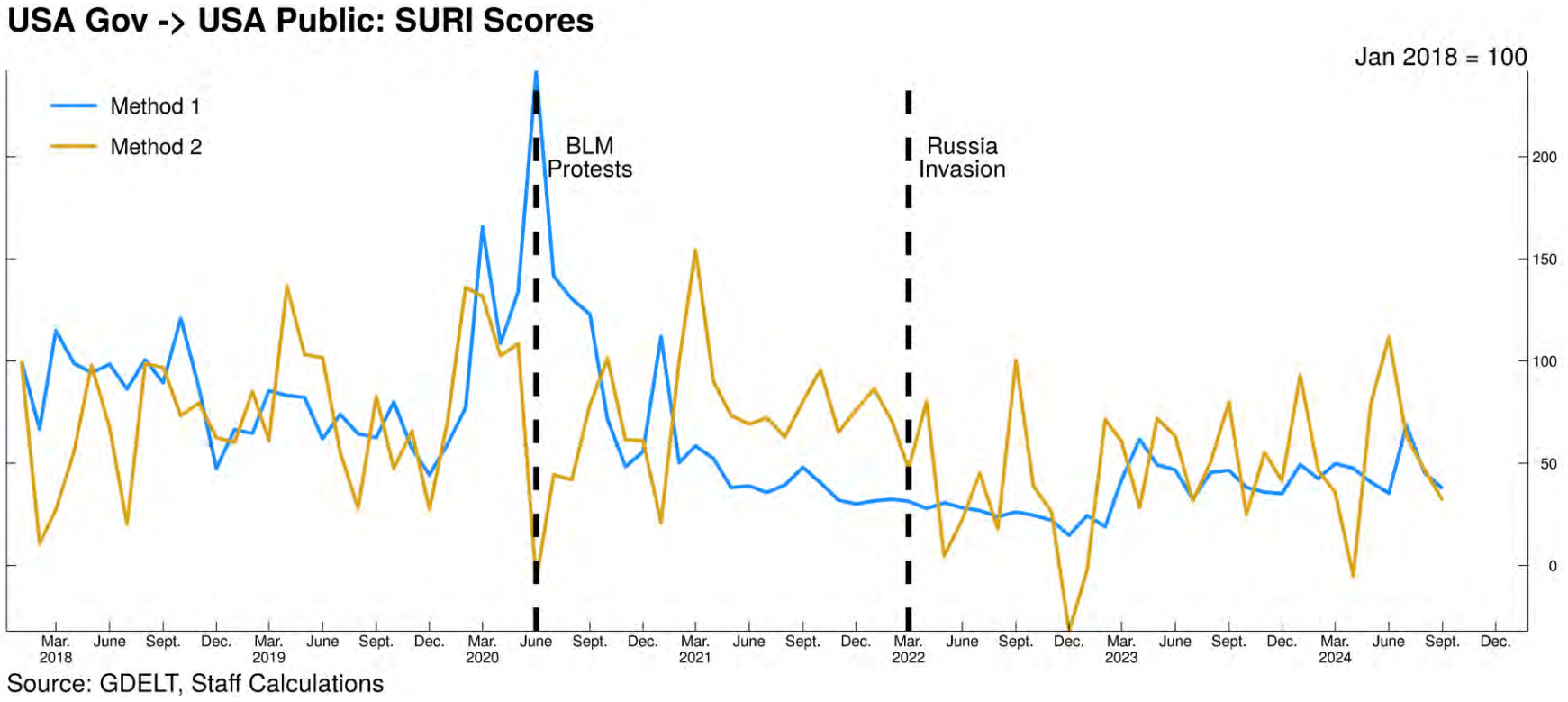
Indexed at the onset of the Russian invasion of Ukraine, we see around the 9-month mark, unrest risk starts to decrease.

This chart is created using Methodology 1.



Select NATO members include Canada, Italy, France, Germany, United Kingdom, and the United States.
NATO members include Albania, Belgium, Bulgaria, Canada, Croatia, Czechia, Denmark, Estonia, Finalnd, France, Germany Greece, Hungary, Iceland, Italy, Latvia, Lithuania, Luzembourg, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Turkey, United Kindgom, and the United States.
EU members include Austria, Belgium, Croatia, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Lithuania, Lexembourg, Malta, Netherlands, Poland, Portual, Slovakia, Slovenia,Spain.
Source: GDELT; Staff Calculations

CASE STUDY: BLACK LIVES MATTER PROTESTS IN THE US



Source: GDELT, Staff Calculations

REFLECTIONS ON EACH METHODOLOGY

Methodology 2 appears to be more sensitive (and noisy) than Methodology 1. However, both appear effective in terms of having large swings around large domestic unrest events. While it's unlikely that these indices would lead to someone discovering a "new" or "overlooked" unrest event, a weighted aggregation of these scores across many countries could serve as a quantitative estimate of the level of global unrest. This measure could then be used in lower-frequency (e.g., monthly) fore/nowcasting models to succinctly capture media information on social unrest. It is worth noting that we heavily restrict the data in this study; relaxing these restrictions could allow higher-frequency analysis. Similar measures could be used to capture labor strikes, hostility levels, and other forms of geopolitical risk.

NEXT STEPS

Moving forward, we are refining and executing a series of regression analyses. These will tease out potential applications of SURI and hopefully provide meaningful information about the relationship between unrest and various elements of the economy.

We are excited about the richness and nuance within SURI, and we believe some of its most fruitful applications will be in targeted analyses of specific international relationships over time. Due to its bidirectionality, diverse country representation, and ability to capture both threatened and realized events, its applications are nearly infinite.

PROPOSED ANALYSIS

The basic structure of our analysis would follow the below equation:

$$Y_{i,j,t} = \alpha + \beta\, SURI_{i,j,t} + \gamma_t + C_{i,j,t} + \varepsilon_{i,j,t}$$

$$Y_{i,j,t} = \alpha + \beta\, SURI_{i,j,t-1} + \gamma_{t-1} + C_{i,j,t-1} + \varepsilon_{i,j,t-1}$$

$$Y_{i,j,t} = \alpha + \beta\, SURI_{i,j,t+1} + \gamma_{t+1} + C_{i,j,t+1} + \varepsilon_{i,j,t+1}$$

Where $Y_{i,j,t}$ is a macroeconomic variable value at time t

i is the country of origin for Actor1 and j is the country of origin for Actor2

SURI is our unrest index, γ are the time fixed effects, C are the other control variables, and ε is the error term

We plan to evaluate these relationships contemporaneously and with leads and lags by one period

PROPOSED ANALYSIS

$$Y_{i,j,t} = \alpha + \beta\, SURI_{i,j,t} + \gamma_t + C_{i,j,t} + \varepsilon_{i,j,t}$$

For any of these permutations, the countries of origin for Actor 1 and Actor 2 may be the same **or** two different countries.

If the same, one can examine directly the economic state of the country at hand.

If different, one can exploit or construct a number of bilateral variables in order to understand the international economic dynamics at play. For example:

- $Y_{i,j,t} = (GDP_{i,t} - GDP_{j,t})$
- $Y_{i,j,t} = (\text{Portfolio Flows } i \rightarrow j)_t$
- $Y_{i,j,t} = (\text{Trade Flows } i \rightarrow j)_t + (\text{Trade Flows } j \rightarrow i)_t$
- $Y_{i,j,t} = (Unemployment_{i,t} - Unemployment_{j,t})$

There are many subsets for which the above equation could be used.

To list a few:

- Actor 1: Government Organization, Actor 2: Civilian Organization
- Actor 1: Civilian Organization, Actor 2: Government Organization
- Actor 1: Government Organization, Actor 2: Government Organization
- Actor 1: Civilian Organization, Actor 2: Civilian Organization
- Actor 1: Any actor type, Actor 2: Government Organization
- Actor 1: Government Organization, Actor 2: Any actor type
- Actor 1: Any actor type, Actor 2: Any actor type

BEYOND SURI

- The methodologies for SURI have the potential to be expanded beyond GDELT as a source. Richer but messier data from social media could also be used to identify unrest events and intensity.
  - TikTok (captions, audio, textual content, hashtags)
  - Instagram (captions, hashtags)
  - Instagram Reels (captions, audio, textual content, hashtags)
  - Twitter / X (text, hashtags)

WELCOME YOUR QUESTIONS AND FEEDBACK

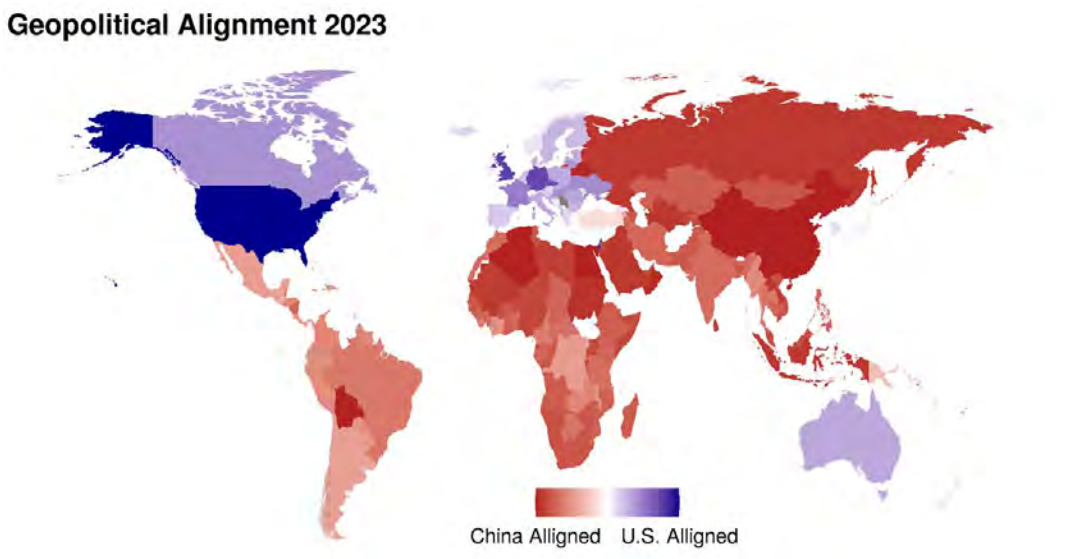


APPENDIX

REFERENCES

- Airaudo, Florencia, and Keith Richards. “Geopolitical Distance Influence on Economic Policy and Outcomes.” International Finance Division, Federal Reserve Board of Governors. 19 Nov. 2024, Washington, DC.
- **Alex Braithwaite, Jeffrey Kucik, Jessica Maves, The Costs of Domestic Political Unrest, *International Studies Quarterly*, Volume 58, Issue 3, September 2014, Pages 489–500, https://doi.org/10.1111/isqu.12061**
- **Albert Saiz, Uri Simonsohn, Proxying For Unobservable Variables With Internet Document-Frequency, *Journal of the European Economic Association*, Volume 11, Issue 1, 1 February 2013, Pages 137–165, https://doi.org/10.1111/j.1542-4774.2012.01110.x**
- **Caldara, Dario, and Matteo Iacoviello. 2022. "Measuring Geopolitical Risk." *American Economic Review*, 112 (4): 1194–1225.DOI: 10.1257/aer.20191823**
- **Cerra, Valerie, and Sweta Chaman Saxena. 2008. "Growth Dynamics: The Myth of Economic Recovery." *American Economic Review*, 98 (1): 439–57.DOI: 10.1257/aer.98.1.439**
- *Gerner, Deborah J.; Schrodt, Philip A.; Abu-Jabr, Rajaa; Yilmaz, Omur (2012). "Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions"*
- Goldstein, J. S. (1992). A Conflict-Cooperation Scale for WEIS Events Data. *The Journal of Conflict Resolution*, *36*(2), 369–385. http://www.jstor.org/stable/174480
- **Hadzi-Vaskov, Metodij, Pienknagura, Samuel and Ricci, Luca Antonio. "The Macroeconomic Impact of Social Unrest" *The B.E. Journal of Macroeconomics*, vol. 23, no. 2, 2023, pp. 917-958. https://doi.org/10.1515/bejm-2022-0039**
- **Kollias, C. and Tzeremes, P. (2022), "The nexus between social unrest and economic growth in Middle East and Central Asia countries", *Review of Economics and Political Science*, Vol. 7 No. 2, pp. 74-86. https://doi.org/10.1108/REPS-06-2021-0059**
- **Martin Beraja, Andrew Kao, David Y Yang, Noam Yuchtman, AI-tocracy, *The Quarterly Journal of Economics*, Volume 138, Issue 3, August 2023, Pages 1349–1402, https://doi.org/10.1093/qje/qjad012**
- **Natalija Novta & Evgenia Pugachev**a, 2020. "The Macroeconomic Costs of **Conflict," IMF Working Papers 2020/110, International Monetary Fund.**
- **Philip Barrett. "Reported Social Unrest Index: September 2024 Update", IMF Working Papers 2024, 225 (2024), accessed December 6, 2024, https://doi.org/10.5089/9798400290893.001**
- **Reuven Glick, Alan M Taylor; Collateral Damage: Trade Disruption and the Economic Impact of War. *The Review of Economics and Statistics* 2010; 92 (1): 102–127. doi: https://doi.org/10.1162/rest.2009.12023**
- **Schneider, G., & Troeger, V. E. (2006). War and the World Economy: Stock Market Reactions to International Conflicts. *Journal of Conflict Resolution, 50*(5), 623–645. https://doi.org/10.1177/0022002706290430**
- **Scott Baker, Nicholas Bloom, Steven Davis. “Measuring Economic Policy Uncertainty.” *National Bureau of Economic Research*, Oct. 2015, https://doi.org/10.3386/w21633**
- **Tarek A Hassan, Stephan Hollander, Laurence van Lent, Ahmed Tahoun, Firm-Level Political Risk: Measurement and Effects, *The Quarterly Journal of Economics*, Volume 134, Issue 4, November 2019, Pages 2135–2202, https://doi.org/10.1093/qje/qjz021**
- **Redl, Chris and Hlatshwayo, Sandile, Forecasting Social Unrest: A Machine Learning Approach (November 1, 2021). IMF Working Paper No. 2021/263, Available at SSRN: https://ssrn.com/abstract=4026493**
- **Rodrik, D. Where Did All the Growth Go? External Shocks, Social Conflict, and Growth Collapses. *Journal of Economic Growth* 4, 385–412 (1999). https://doi.org/10.1023/A:1009863208706**

CONTEXT: INCREASED POLARIZATION, TENSION, AND UNREST



2024, Airudo & Richards

- Increased polarization is a global phenomenon, shedding light on dissatisfaction with the status quo and discontent across national boarders.
- From trade fragmentation and diversion to geopolitical aggression, unrest is not only costly but is increasingly relevant to our understanding of international dynamics.
- The ability to examine the actions between citizens, governments, and non-governmental organizations has never been more relevant.

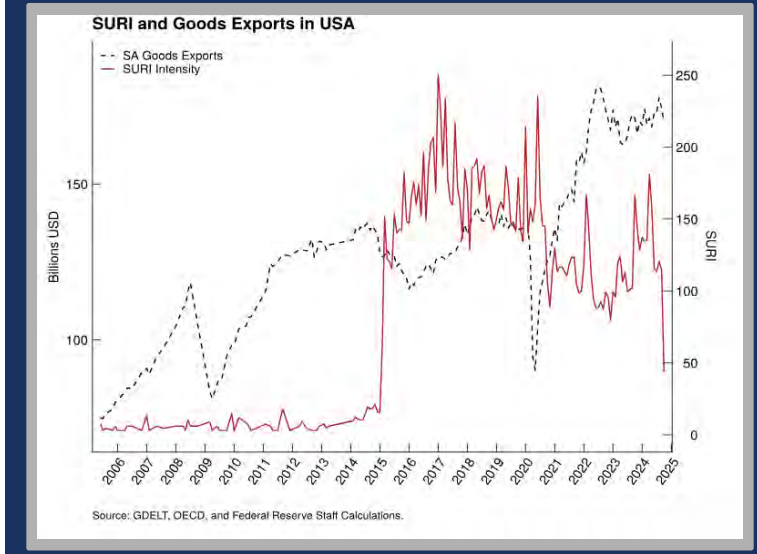
SURI AND MACROECONOMIC INDICATORS

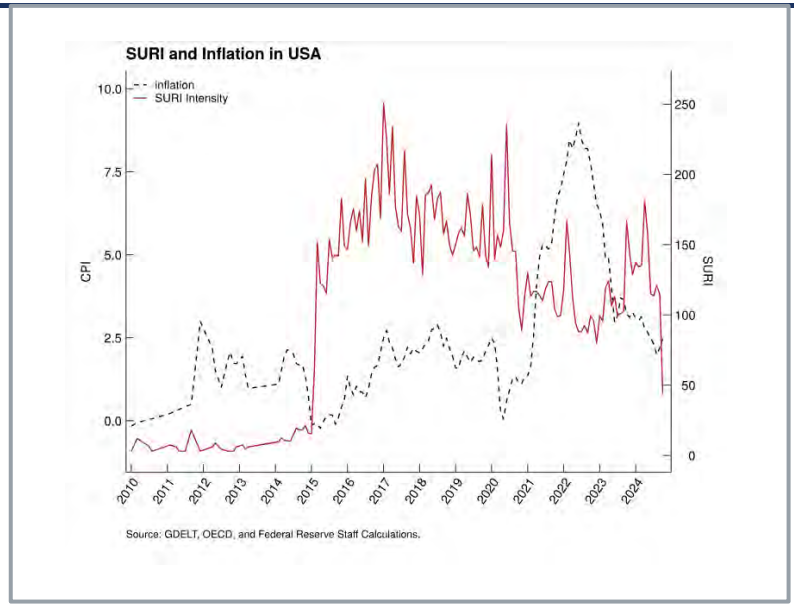
Previous literature has shown compelling evidence of a relationship between of unrest and macroeconomic indicators. In order to assess, at a preliminary level, the potential applications of SURI in macroeconomic research, we have begun to examine its associations with the following:

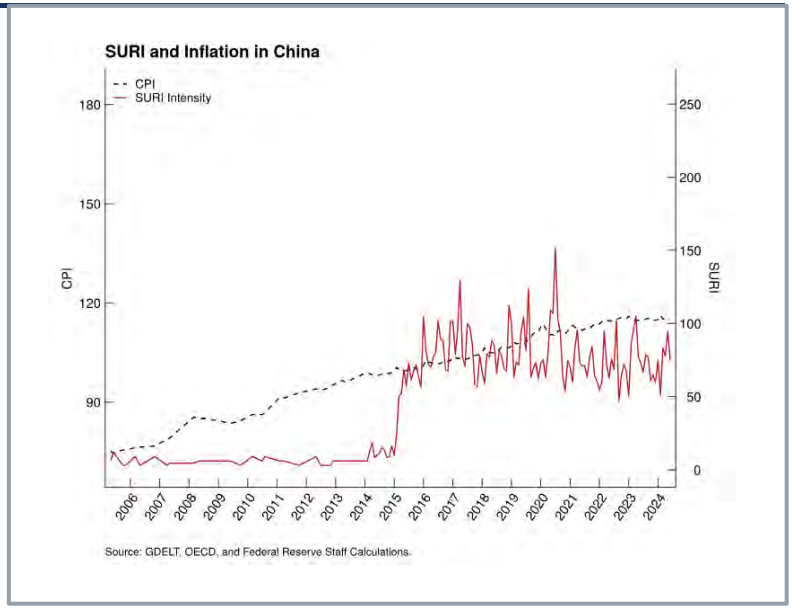
- Foreign Direct Investment
- Portfolio Flows
- Unemployment
- GDP & GDP Growth
- Trade Dynamics (Imports, Exports, Trade Openness)
- Inflation

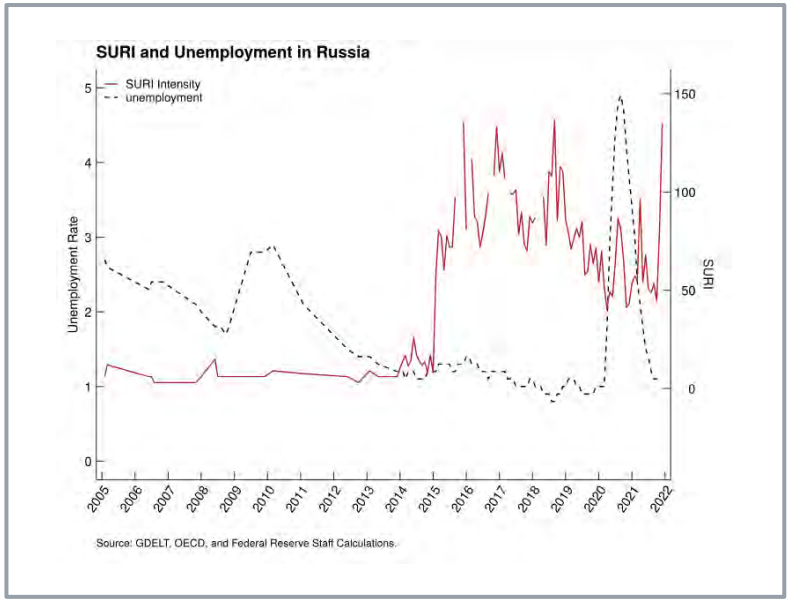
PRELIMINARY ASSOCIATIONS METHOD 1

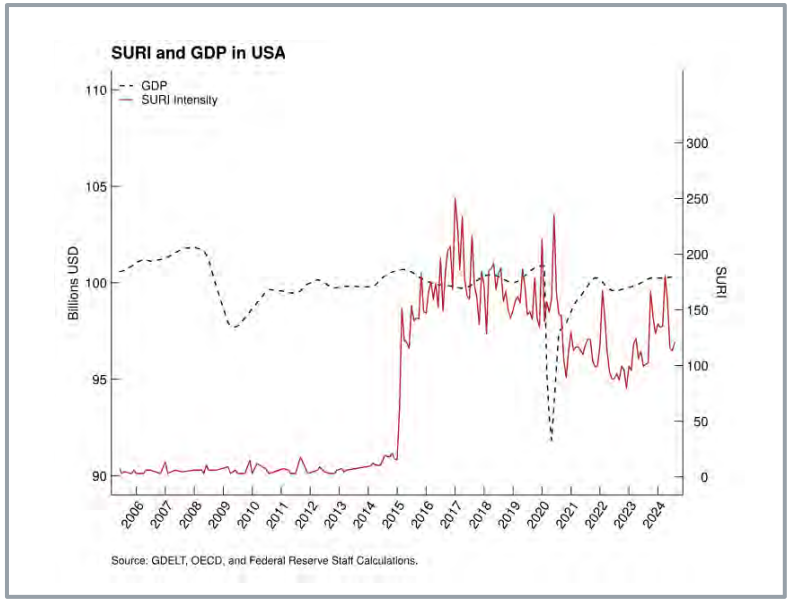
DOMESTIC ISSUES: COUNTRY OF ORIGIN FOR ACTOR1 AND ACTOR2 ARE THE SAME





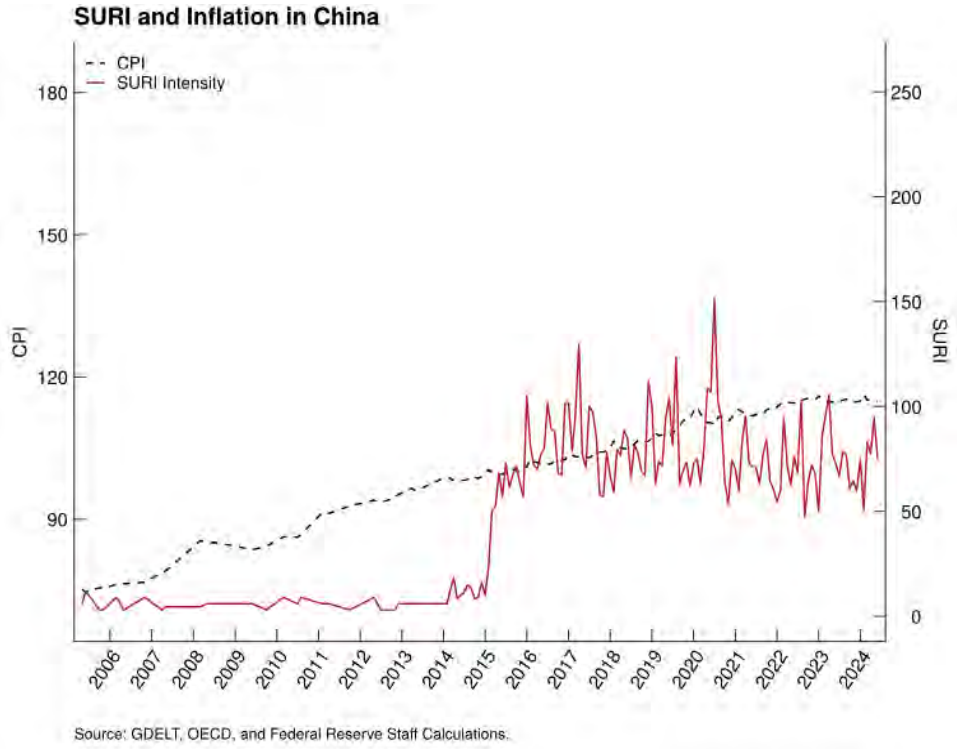






PRELIMINARY ASSOCIATIONS: INFLATION IN CHINA

Method I
Country of Origin for
Actor1 and Actor2 are the same

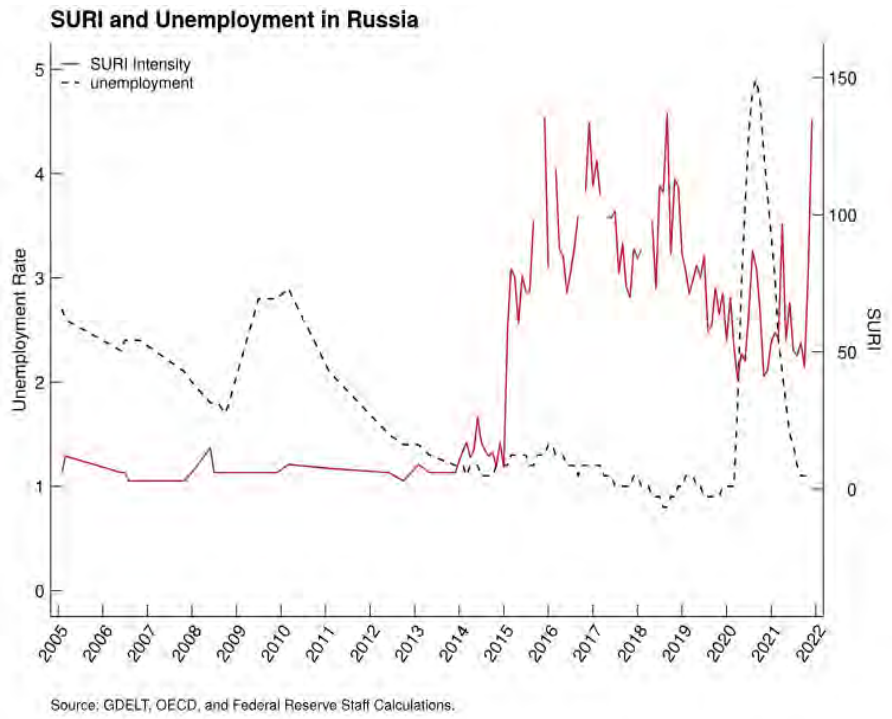


Method II
Either Actor1 or Actor2 is China

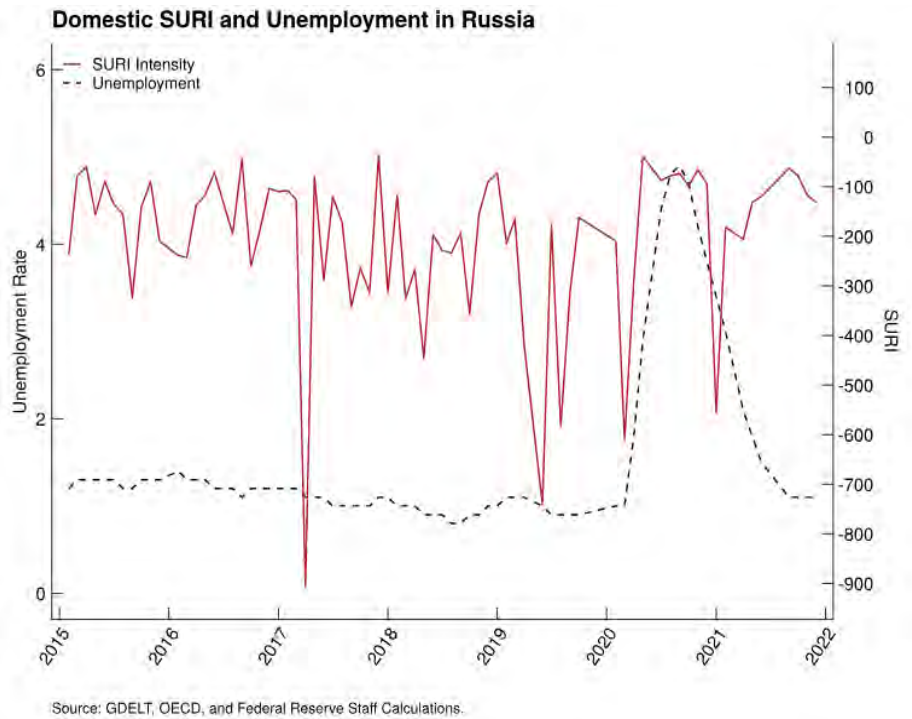


PRELIMINARY ASSOCIATIONS: UNEMPLOYMENT IN RUSSIA

Method I
Country of Origin for
Actor1 and Actor2 are the same

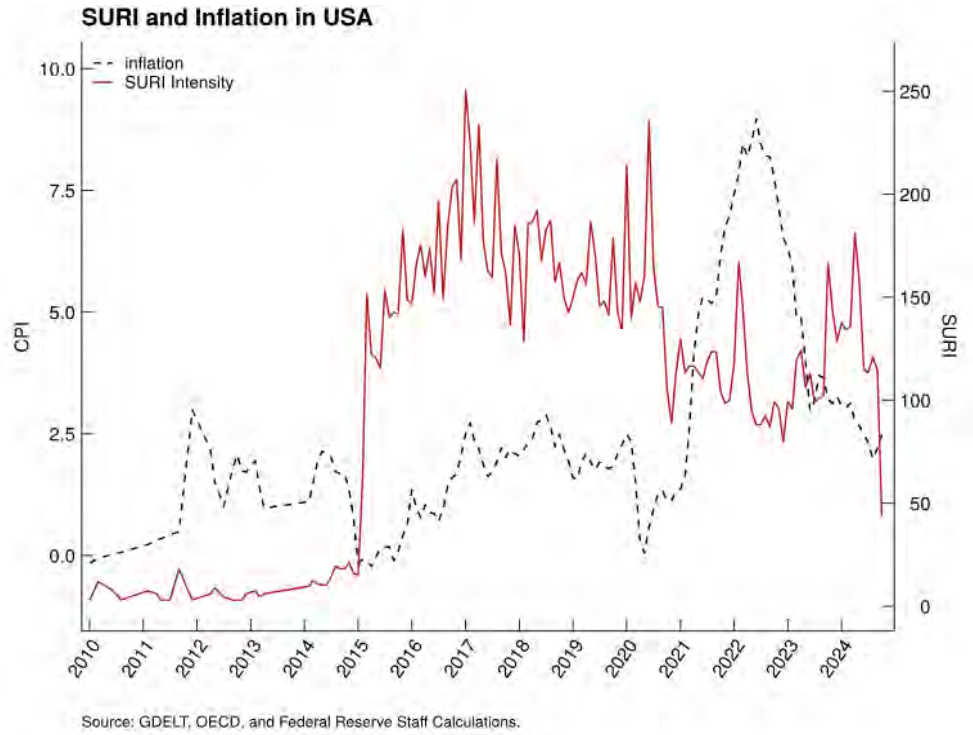


Method II
Both Actor1 and Actor2 are Russia

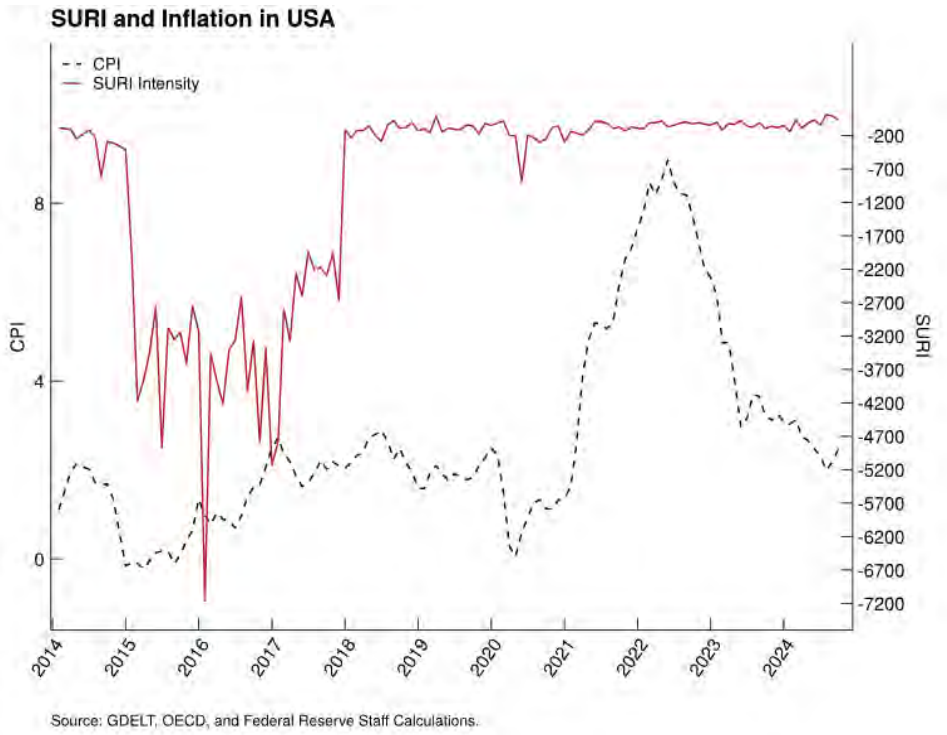


PRELIMINARY ASSOCIATIONS: INFLATION IN THE USA

Method I
Country of Origin for
Actor1 and Actor2 are the same



Method II
Either Actor1 or Actor2 is the USA

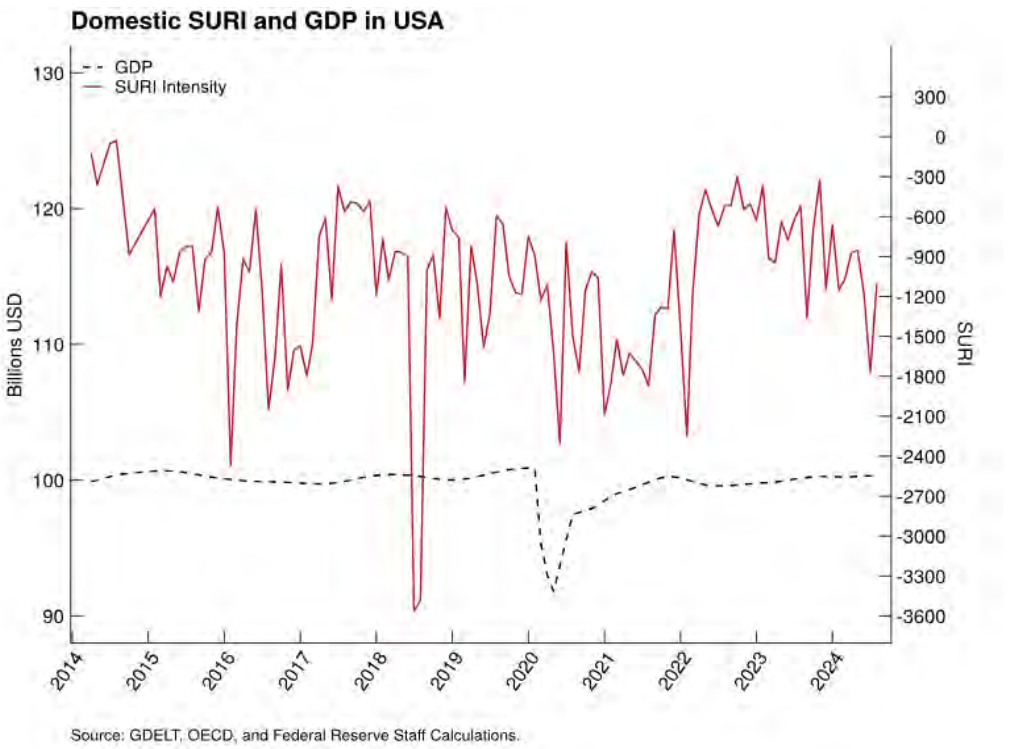


PRELIMINARY ASSOCIATIONS: GDP IN THE USA

Method I
Country of Origin for
Actor1 and Actor2 are the same

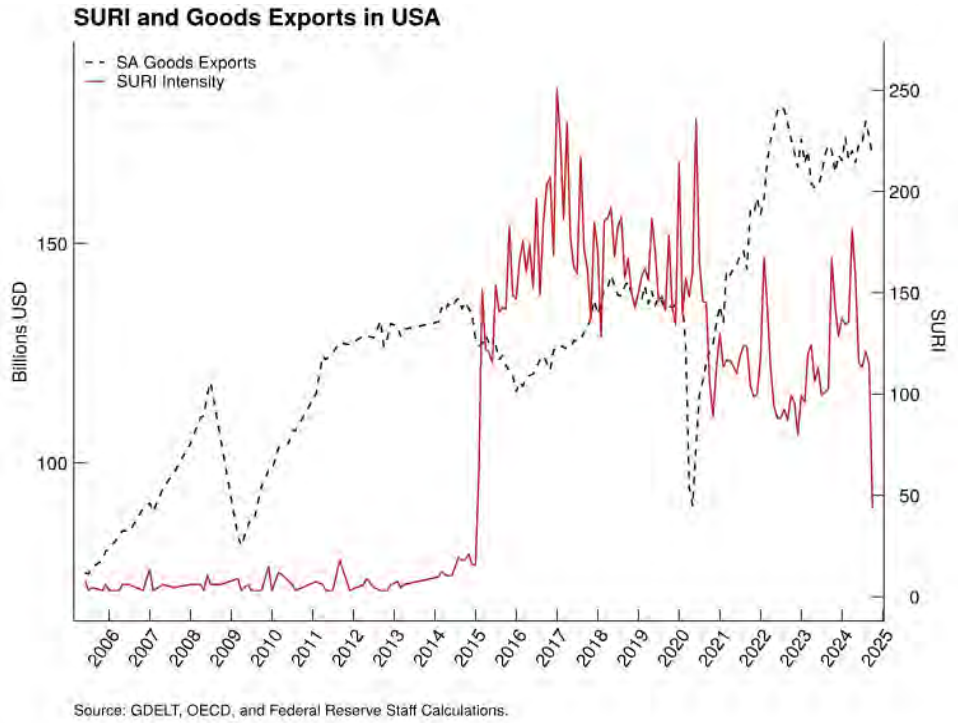


Method II
Both Actor1 and Actor2 are the USA

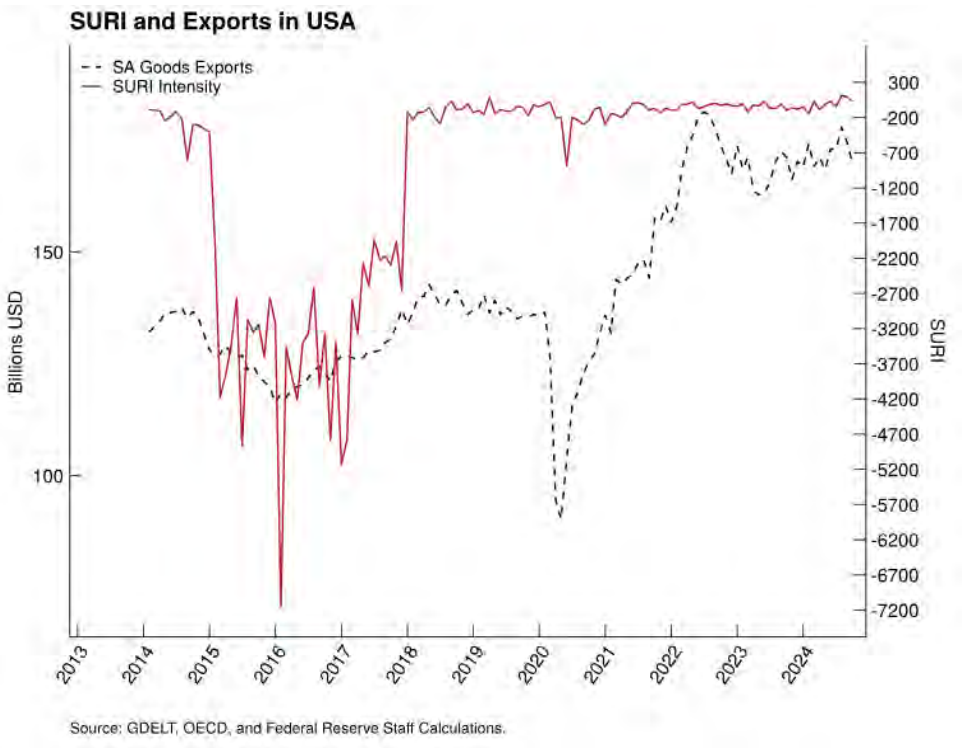


PRELIMINARY ASSOCIATIONS: EXPORTS IN THE USA

Method I
Country of Origin for
Actor1 and Actor2 are the same



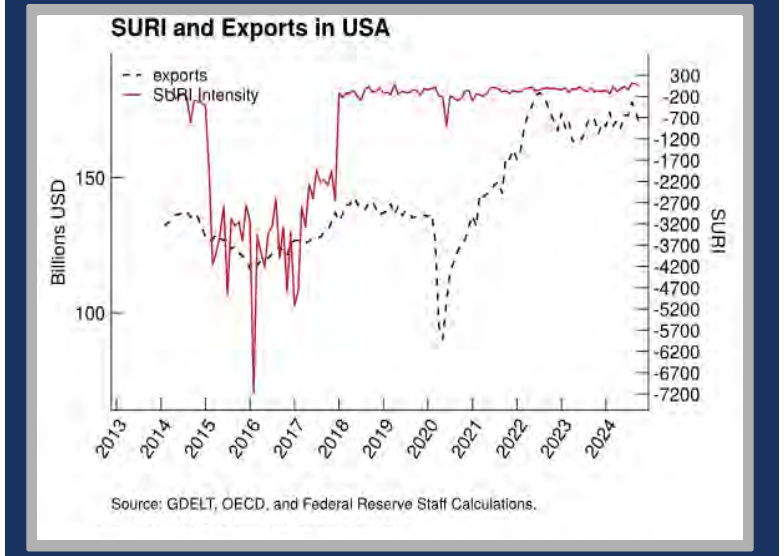
Method II
Either Actor1 or Actor2 is the USA

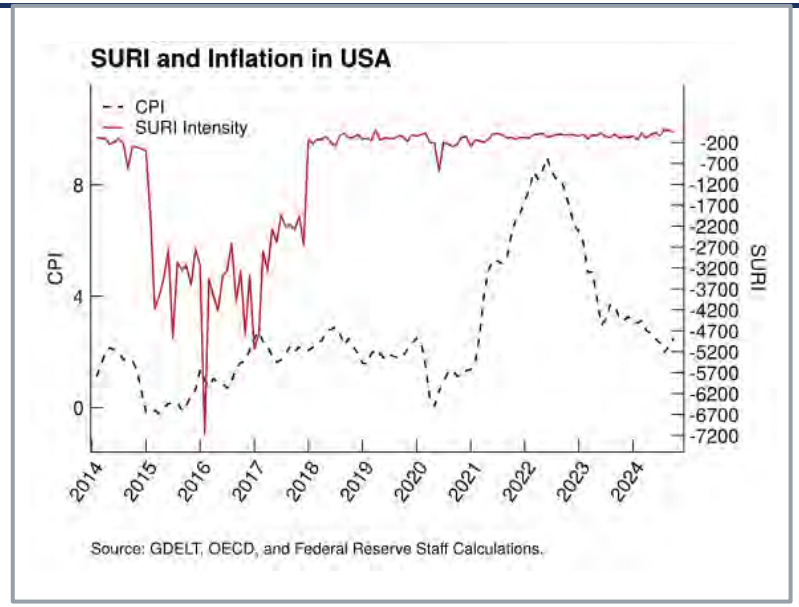


PRELIMINARY ASSOCIATIONS METHOD 2

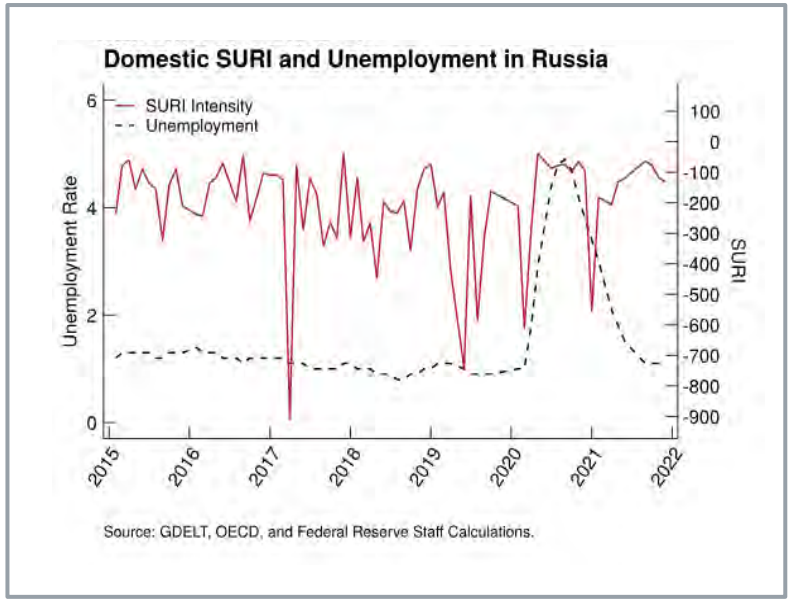
LABELED "DOMESTIC": COUNTRY OF ORIGIN FOR ACTOR1 AND ACTOR2 ARE THE SAME

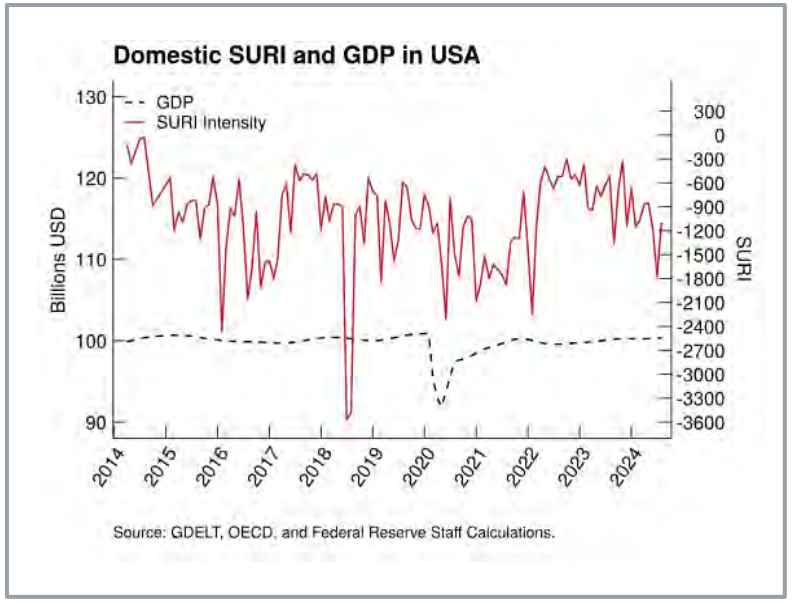
NOT LABELED "DOMESTIC": THE RELEVANT COUNTRY IS THE COUNTRY OF ORIGIN FOR EITHER ACTOR1 OR ACTOR2











Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Indexing and visualization of climate narratives using BERT and causal extraction[1]

Noriyasu Kaneda,
Bank of Japan

Hiroki Sakaji,
Hokkaido University

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

BANK OF JAPAN

Indexing and Visualization of Climate Narratives Using BERT and Causal Extraction

4th IFC and Bank of Italy Workshop
on “Data Science in Central Banking”

2025/2/19

Bank of Japan Noriyasu Kaneda
The Hokkaido university Hiroki Sakaji

*The opinions expressed in this document are those of the authors alone and do not represent the official views of the Bank of Japan. Any possible errors belong to the authors personally.

Background

- *"Climate change is a global challenge and could have a broad impact on our society and economic activity into the future."*

  - BOJ's Strategy on Climate Change

- Economists in central banks are focusing on how climate change will affect financial markets and macro-economy.

- Some researches used traditional text analysis to capture climate-related risk from newspaper and creating climate index using the NLP methods.

  Engle et al. [2020], Pastor et al. [2022], Faccini et al. [2023], Hiraki et al. [2025]

Overview and Contribution

- We proposed a method to analyze cause-effect relationship in text data (we call it "economic narrative"), using two NLP techniques, (1) BERT and (2) causal extraction.

(1)BERT (Bidirectional Encoder Representations from Transformers) is a Deep Learning based language model by Google, and it can learn the sentence structure and context in a text.

(2)We also use another NLP, "Causal Extraction" (Sakaji et al. [2008]). This method is an algorithm based on linguistic knowledge.

- This research focuses on climate narratives. We analyze the newspapers on climate addressing by using the NLP methods above, and index/visualize climate narratives.

Data

【 Nikkei newspaper (morning edition, weekdays only) 】

- Period: January 2000 - November 2021
- 17,000 climate change-related articles* in Japanese

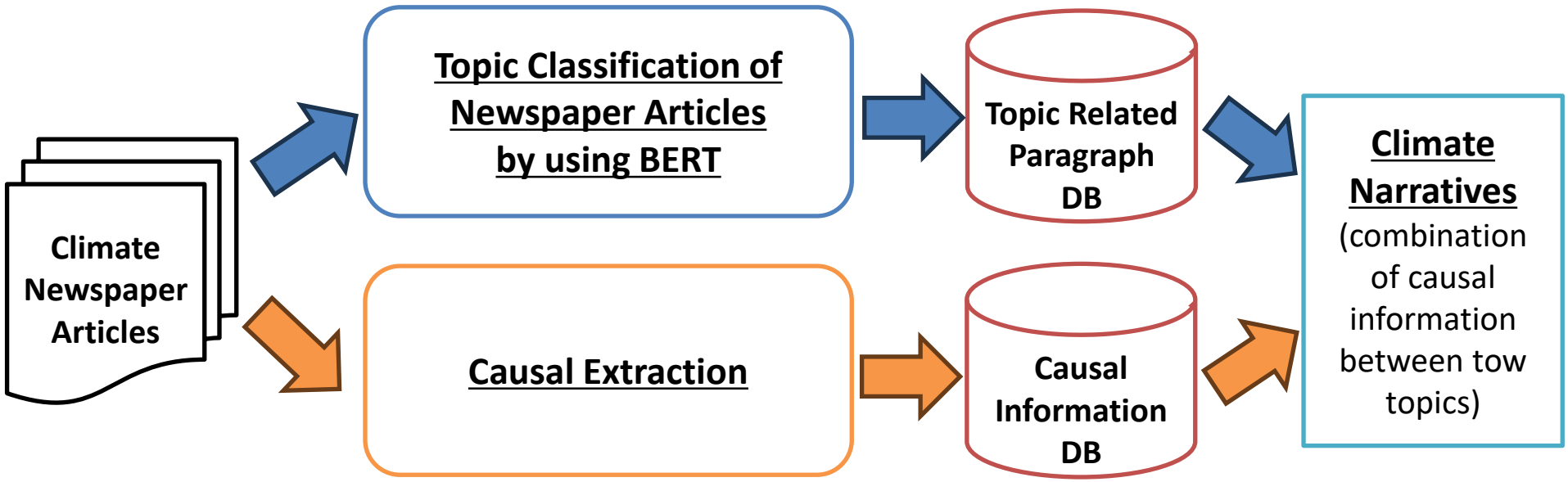
*containing at least one of the words "climate change," "global warming," or "greenhouse effect.

■ 40 topics (classification tags) are the target of this analysis.

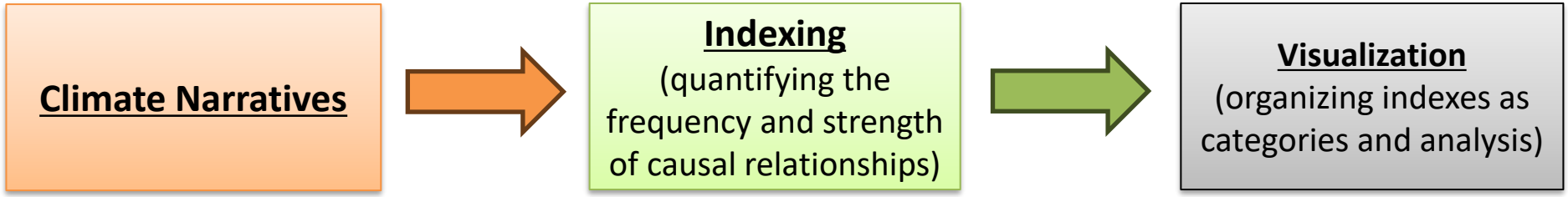
【Corporate】		【Politics】	【Economics】	【Society】
1 New business	11 Governance	21 Party	28 Monetary policy	35 Weapon
2 M&A	12 Labour	22 Regulation	29 Inflation	36 Disaster
3 Business strategy	13 Name change	23 Fiscal policy	30 Business cycle	37 Trial
4 Price strategy	14 Wage	24 Energy policy	31Finacial market	38 Energy probrem
5 Production strategy	15 Finance	25 Seculity	32 Foreign Exchange	39 Environmental probrem
6 Cost reduction	16 Performance	26 Summit	33 Bond	40 Consumer trend
7 Supply chain	17 Sales	27 Social seculity	34 Renewable energy	
8 Patent	18 Price			
9 R&D	19 Market share			
10 Investment	20 Hot selling			

Methodology for Extracting Climate Narratives

Methodology for Extracting Climate Narratives

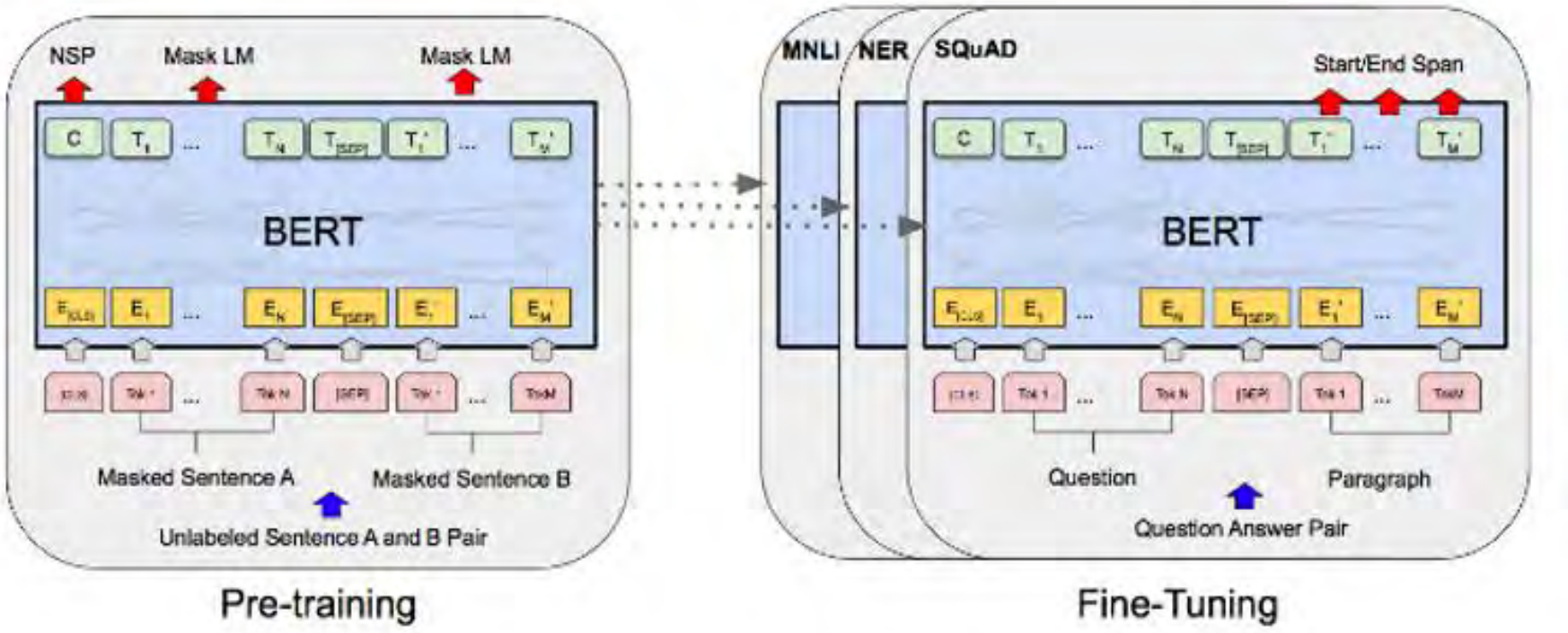


Methodology for Indexing and Visualization of Climate Narratives (for empirical research)



BERT (Bidirectional Encoder Representations from Transformers)

Devlin et al. [2019]



Trained and fine-tuned by Google for treating basic linguistic problems (public source)

More Fine-tuned for classification of climate news by topics (specific purpose)

Results of the “OLYMPIC” Classification Model

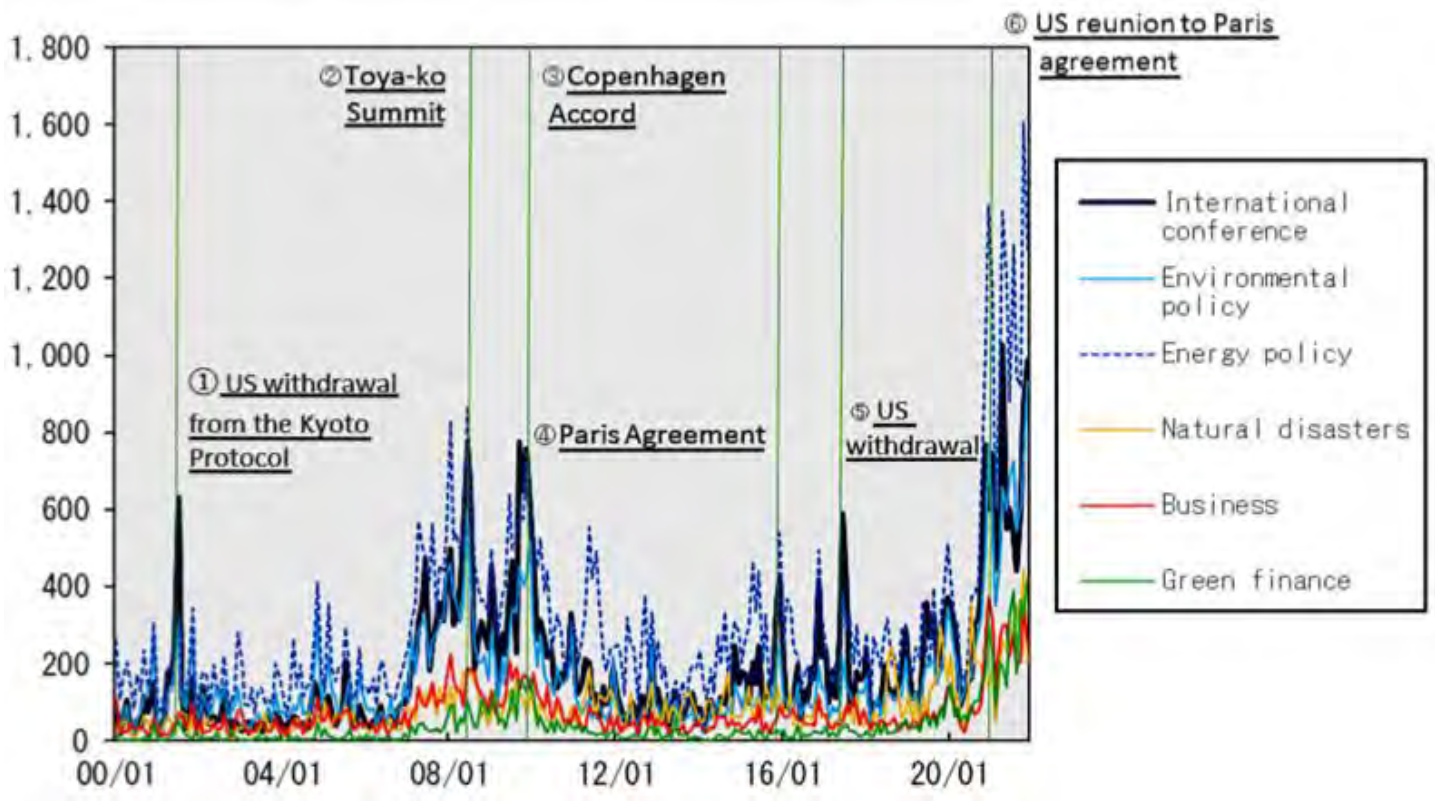
	Accuracy rate	Precision rate	Recall rate	F1 score
Linear Regression	0.958	0.958	0.958	0.958
Random Forest	0.950	0.951	0.950	0.950
Support Vector Machine	0.954	0.954	0.954	0.954
BERT	0.960	0.960	0.960	0.960

- In the classification for Olympic article as an example, BERT is slightly better than other models.

Basic analysis (Topic based index using only BERT)

- **Topic based index** increased through 2008 to 2010 (in Toya-ko Summit, Copenhagen Accord) and in 2015(the Paris Agreement).

- In recent years (since 2018), many topics show a surge, including micro developments such as corporate strategy.



⇒ High correlation among topics. It is not clear how climate narratives are linked across topics.

Causal extraction: The second NLP for finding causality

- We focus on linguistic causal-effect relationship and consider the causal relationships in newspaper as economic narratives by using causal extraction.

Step.1 Causality Determination [Sakaji et al. 2017]

- Determine whether or not the input sentence contains a causal relationship (a causal sentence).

Step.2 Causality Extraction [Sakaji et al. 2008]

- Extract “cause and effect expressions” from causal sentences that contain causal relations picked up in Step1.

- This method analyzes syntactic patterns based on dependency parsing in linguistic knowledge.

Example of Climate Narrative (from Regulation to Corporate topic)

Causal sentence 1 on Environmental Regulation in Jan 2016

With Paris agreement agreed to in 2015, climate change policies are being discussed around the word.
As a result, the Japanese government plans to strengthen the environment to promote decarbonization policies.

High cosine similarity (similar information)

Causal sentence 2 on Corporate Strategy in Feb 2017

In response to the government's move to tighten regulations, corporates are begging to incorporate climate change addressing into their corporate strategies.

We observe a climate narrative from Regulation to Corporate in Feb 2017.
If there are many narratives in Feb 2017, this narrative is a hot topic at this point.

Generalization

Causal Event (Topic A)

Causal expression ⇒ ***Effect expression***

Prior causal information
(past)

High cosine similarity

Combine a pair of "Causal event's *effect expression*" and "*effect expression*'s *causal expression*": *A result relates Topic B is caused by some event relates Topic A.*

Effect Event (Topic B)

Causal expression ⇒ Effect expression

Subsequent
causal information
(present)

Indexing Economic Narratives

- Links between Cause $(\vec{i}_{t-d})$ <past information from t time to d days ago> and Effect $(\vec{j}_t)$ <information as of time t> are calculated based on the cosine similarity (how close the textual information is in content) of the causal chain and sum it by month.

$$Index_monthly_m = \sum_{j=0}^{M} \sum_{i=0}^{L(j)} \frac{1}{1+ae^{bd}} cos(\vec{i}_{t-d} \cdot \vec{j}_t) \quad (1)$$

$$cos(\vec{i}_{t-d} \cdot \vec{j}_t) = \frac{\vec{i}_{t-d} \cdot \vec{j}_t}{|\vec{i}_{t-d}||\vec{j}_t|} \quad (2)$$

Here,

M : set of causal chains included in month m.

$L(j)$: set of cause event $\vec{i}_{t-d}$ connected to result event $\vec{j}_t$.

$t-d$: observation point of cause event leading to result event ($d > 0$).

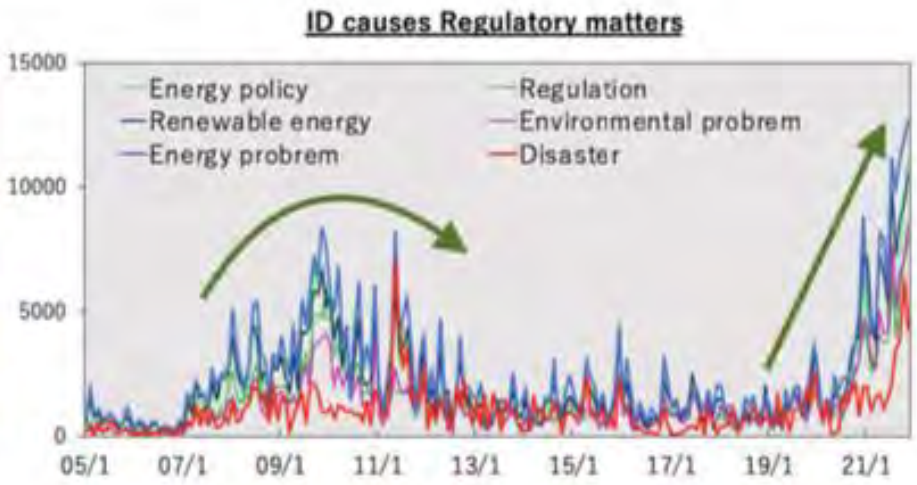
t : observation point of result event included in month m.

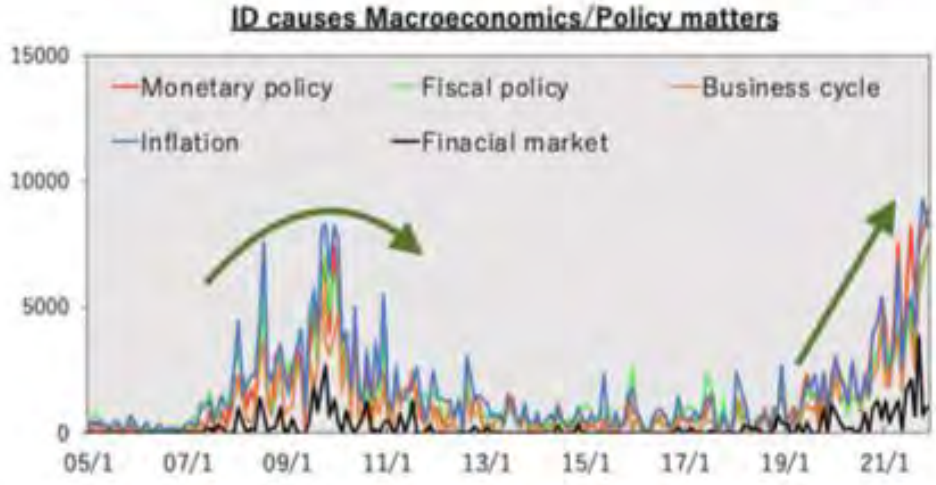
d : time difference (days) between cause event and result event.

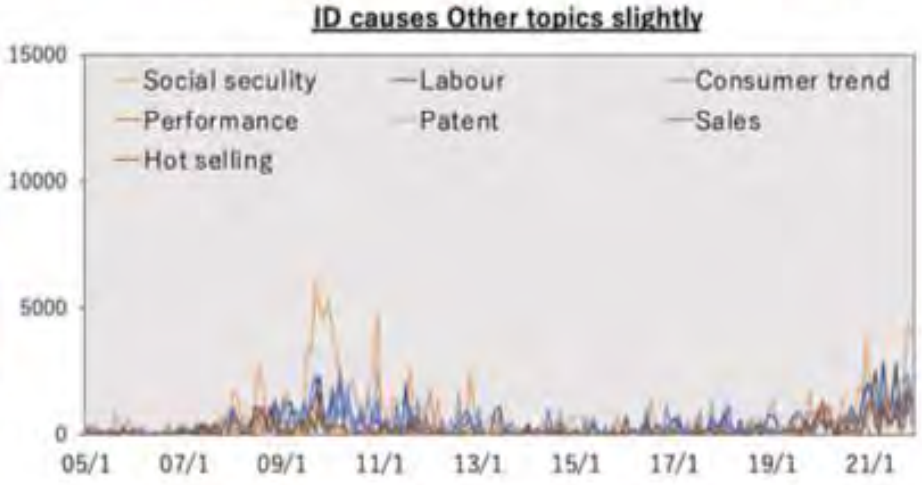
*Old causal chains are depreciated according to elapsed time to eliminate the bias that the latest news has more links to the past (weights are halved after 5 years based on the logistic function) in (1). **Count all links with $cos(\vec{i}_{t-d} \cdot \vec{j}_t) \geq 0.7$** (strong causality) in (2).

Climate indexes (International discussion ⇒ each topics)









Mapping all narratives from 2001 to 2017

- Strong links to many topics from international conferences on climate change, environmental policy, and energy policy.

Effect Topic

Cause Topic

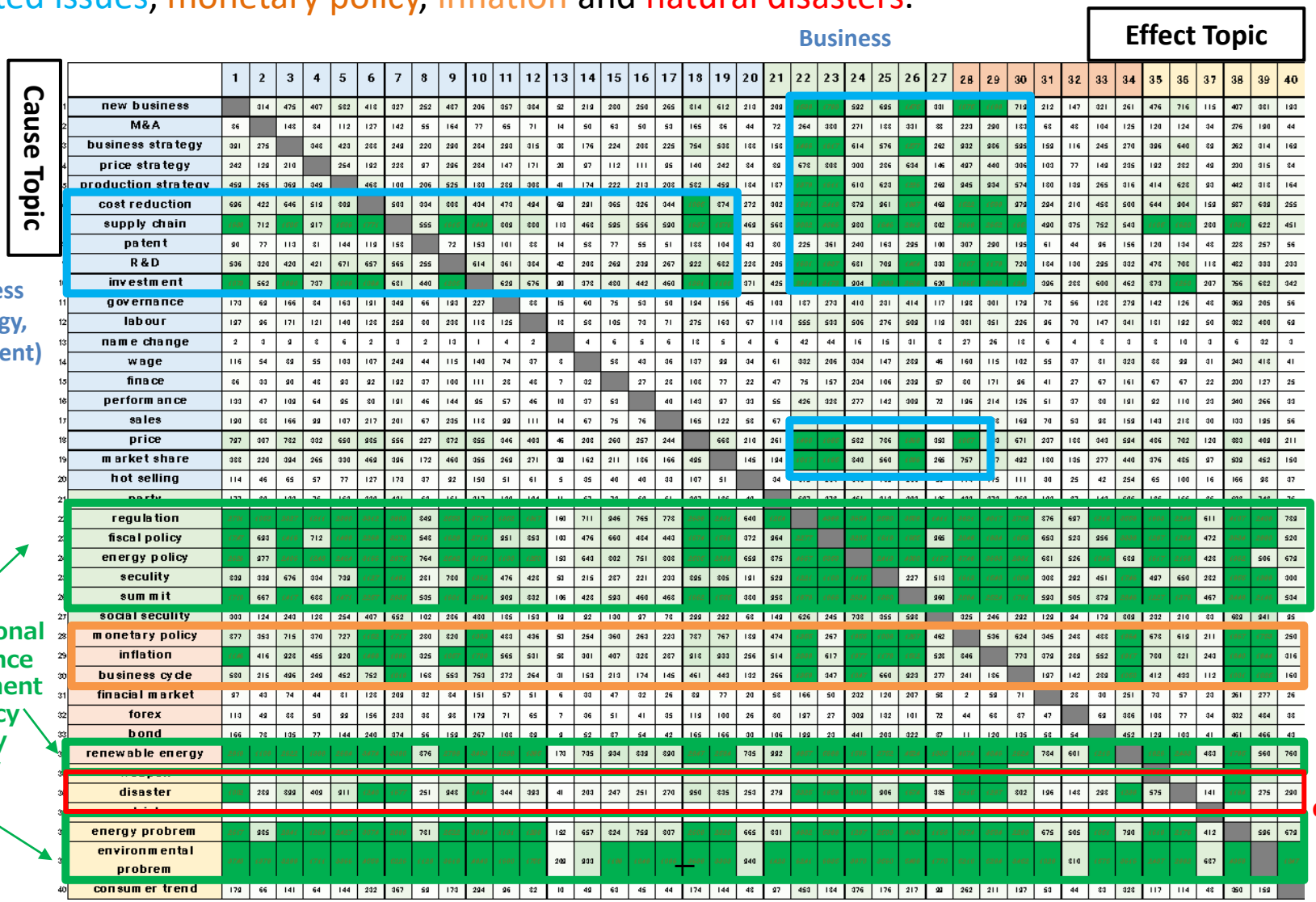
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	new business		116	203	257	282	272	192	133	237	192	100	147	55	143	93	91	201	402	268	118	190	768	836	464	566	385	265	377	666	279	70	97	84	181	364	430	70	233	198	94
2	M&A	12		32	35	32	29	30	19	45	19	17	18	9	20	18	14	26	58	27	14	32	77	120	93	82	115	39	58	101	45	14	20	18	51	57	52	14	91	59	15
3	business strategy	229	112		247	267	214	185	133	198	234	93	145	55	141	93	94	200	429	275	124	172	859	836	530	521	868	248	364	654	269	62	89	74	171	352	441	65	168	173	91
4	price strategy	135	51	103		132	132	156	61	172	173	43	69	25	65	46	47	96	96	124	51	85	389	397	280	200	392	116	182	251	127	35	48	41	176	159	183	34	168	204	40
5	production strategy	223	102	184	219		233	92	118	272	110	84	125	49	123	93	83	159	305	222	103	162	784	775	439	489	738	224	363	530	239	66	94	73	172	321	382	54	272	135	30
6	cost reduction	435	172	367	375	440		418	208	510	355	150	220	80	222	157	134	284	618	437	168	320			704	836		434	588	960	415	113	151	133	396	558	606	116	407	371	132
7	supply chain	700	272	706	584	740	888		317	881	901	228	332	114	318	241	199	441	852	704	257	483			761		1172	624	920		637	168	240	194	314	829	976	167	623	274	217
8	patent	47	32	62	62	78	85	73		47	92	34	80	15	38	35	23	44	103	57	27	67	106	214	163	145	196	76	110	171	73	21	33	27	94	100	95	28	152	105	28
9	R&D	316	130	262	300	385	446	323	153		437	109	160	58	155	121	100	231	498	344	145	205	900	981	454	602	958	286	404	726	295	68	101	84	195	415	489	79	250	102	119
10	investment	503	200	485	448	486	502	452	233	594		167	249	91	238	174	157	325	698	496	193	345			623	950		462	648		458	124	171	146	296	625	727	115	383	336	159
11	governance	32	11	31	24	30	28	54	16	38	41		12	5	14	10	8	16	41	24	9	31	58	75	109	70	96	33	39	71	28	11	15	14	33	46	36	12	32	76	12
12	labour	60	27	57	57	46	38	94	36	85	42	27		14	29	33	24	37	105	53	25	59	239	188	231	148	210	68	102	147	68	25	34	29	152	100	89	22	172	209	25
13	name change	1	2	4	7	3	1	4	4	7	1	3	2		4	3	3	6	11	4	3	6	21	20	17	14	22	7	11	17	9	4	3	3	4	10	9	3	10	14	3
14	wage	77	24	62	49	72	72	172	33	81	96	26	26	11		26	21	35	86	54	17	58	237	151	254	129	195	45	85	85	53	22	28	31	220	81	78	24	154	244	22
15	finace	28	13	32	31	48	44	81	19	46	67	15	21	7	22		12	28	45	37	12	34	25	114	148	81	132	44	47	102	42	14	18	16	100	48	41	14	116	49	12
16	performance	32	12	27	24	20	17	53	17	46	20	12	13	8	16	14		18	44	25	13	25	142	109	105	69	112	37	52	82	33	12	17	14	75	43	41	9	75	111	12
17	sales	112	36	82	72	70	108	181	43	142	71	31	47	19	46	33	35		119	70	33	62	379	302	194	180	293	82	137	185	86	27	32	33	122	115	135	20	81	108	29
18	price	344	105	360	121	293	507	339	117	412	507	85	141	42	127	90	85	178		280	100	192	655	781	474	515	714	244	353	433	250	79	105	86	378	321	373	63	482	246	88
19	market share	164	82	160	165	162	230	193	94	206	183	74	104	37	98	77	70	119	246		71	135	567	533	545	391	553	187	258	383	182	59	85	65	237	257	277	50	306	245	67
20	hot selling	55	19	26	37	60	73	91	24	55	86	16	28	13	23	20	19	29	62	26		30	203	148	240	87	149	42	67	101	49	12	18	15	120	55	71	12	79	60	18
21	party	125	41	101	74	111	209	313	64	127	242	52	96	19	58	46	29	52	133	107	26		364	175	409	219	290	124	194	221	125	47	61	57	498	182	138	58	452	445	50
22	regulation		416		846				502			407	524	147	512	374	287	616			338	872												322	431	362				341			367
23	fiscal policy	731	228	652	414	669			301	788		247	293	85	287	205	159	304	708	606	146	599					912	620	718	843	559	227	294	240		840	699	214			224
24	energy policy		402		899				503			360	513	157	485	344	295	675			393	946												267	384	315	502			283	970	280	371
25	seculity	401	131	348	203	369	613	773	181	421	706	136	152	48	131	102	79	157	393	346	80	376	640	570	306		152	371	536	717	346	109	160	122	920	372	389	135		527	134
26	summit	727	232	654	392	669			296	786		235	279	83	244	195	150	297	692	603	149	636	931	964		807		649	827		586	200	271	219		785	714	208		823	221
27	social security	181	52	154	110	159	238	438	78	190	334	68	77	22	81	64	42	75	183	151	36	141	395	121	665	334	465		176	156	138	58	73	71	660	219	159	65	540	671	55
28	monetary policy	278	95	250	175	278	437	671	125	295	494	103	133	35	124	84	76	131	273	240	61	246	560	84	968	907	485	237		167	173	102	126	103	930	357	268	86	886	851	92
29	inflation	540	151	488	304	471	791		196	577	900	158	208	59	203	139	120	228	441	415	117	369		392		826	925	371	384		321	150	188	160		575	487	124		962	146
30	business cycle	193	64	183	123	171	291	416	77	212	294	65	87	25	77	58	54	87	175	143	46	146	557	155	561	347	375	152	80	61		65	82	63	511	228	193	47	487	511	61
31	finacial market	30	12	26	24	33	50	76	14	30	59	18	18	6	16	12	10	16	28	26	7	33	69	32	120	92	64	35	6	20	29		11	11	122	39	27	10	121	149	11
32	forex	51	20	45	36	51	90	125	24	53	94	22	28	8	23	19	17	26	51	45	12	52	109	27	216	92	67	55	14	32	35	17		21	218	68	48	17	306	260	20
33	bond	38	16	31	27	40	62	104	18	39	80	18	22	6	20	15	12	19	43	38	9	44	59	12	159	87	82	39	9	27	25	14	17		164	54	35	14	154	178	14
34	renewable energy		440		975				548			326	561	175	533	371	326	732			426	901												294	421	345				293		295	338
35	weapon	263	84	250	143	243	406	414	111	285	415	81	100	32	80	67	54	109	246	226	56	220	548	440	411	156	222	220	336	466	216	66	97	74	470		277	78	567	360	88
36	disaster	392	85	390	210	362	532	743	109	419	602	70	103	35	100	66	67	162	380	288	112	181		725	477	481	747	201	293	498	217	48	77	61	525	351		62	486	126	117
37	trial																																								
38	energy probrem		401		915				502			356	515	162	499	343	300	682			401	883												265	261	372	311	534			270		297	361
39	environmental probrem		533						677			433	685	203	651	456	389	868			505													331	544	435				390			502
40	consumer trend	52	15	48	26	47	76	115	19	61	96	16	18	6	18	14	10	21	49	36	15	37	141	66	154	72	91	38	48	60	39	14	19	14	104	52	49	14	107	50	

International Conference Environment -al policy Energy Policy

Note: Dark green highlight means a high level of economic narrative index between each topic.

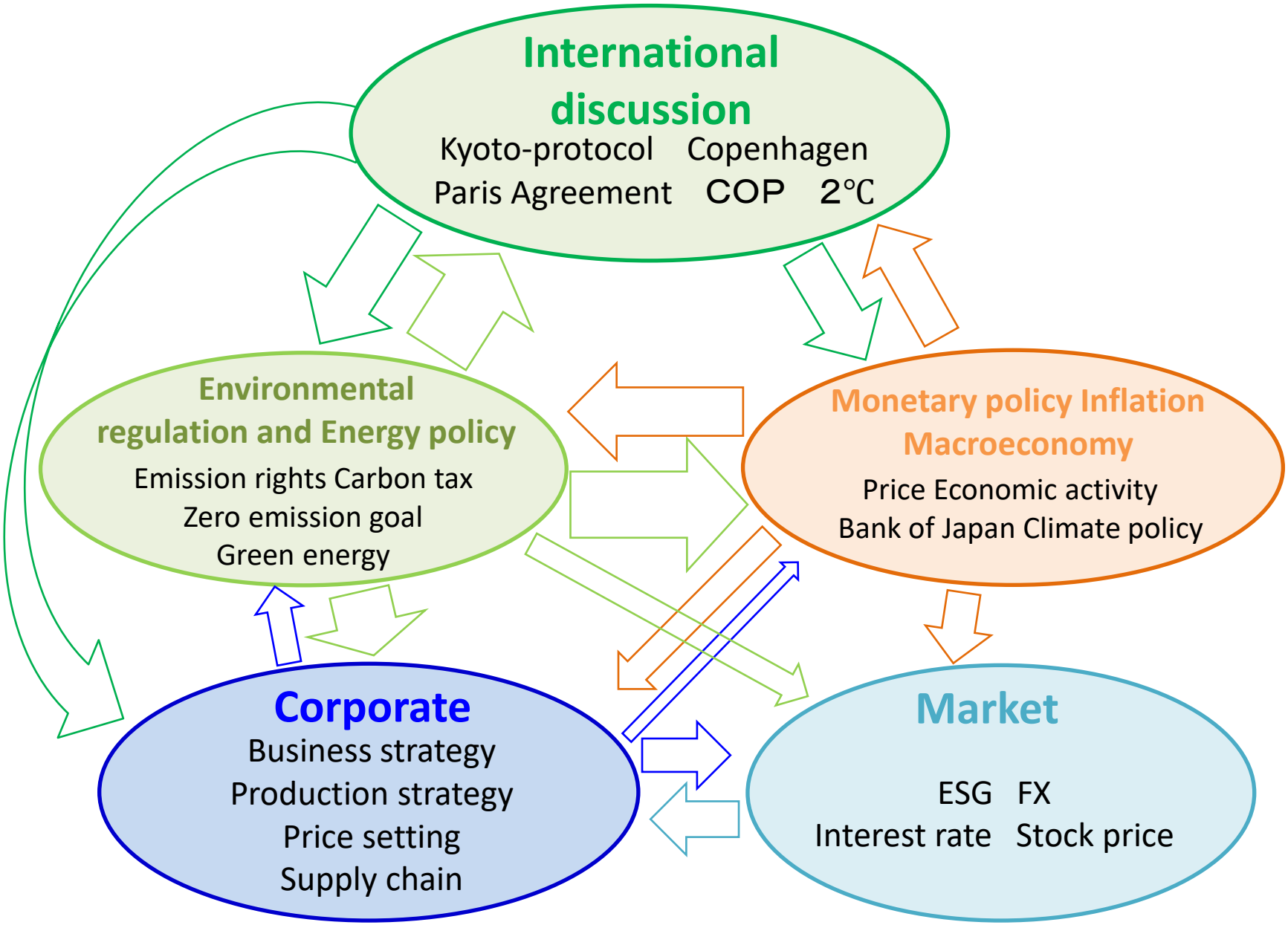
Mapping all narratives from 2018 to 2021

- Strong connections on a wide range of topics from international conferences on climate change, environmental policy and energy policy. Connections have emerged in business-related issues, monetary policy, inflation and natural disasters.



Note: Dark green highlight means a high level of economic narrative index between each topic.

Visualization of Climate Narratives



Note: The size of an arrow means a level of economic narrative index between each topic categories.

Main Result

- The climate narratives suggest that only governments played a major role in the discussion of international regulations for carbon neutrality in the post-Kyoto period, or before the Paris Agreement.

- In recent years, climate change has become a hot topic not only in international and domestic regulations, but also in corporate strategy, green finance and big issues, and as a major issue for central banks since 2018.

- These developments could be related to the growing awareness of transition risk and physical risk in Japan.

Discussions

- Is climate narrative from causal sentence statistical causality?

- ✓ Newspaper could contain noisy or incorrect information due to the misunderstanding of the article writers.

- ✓ Empirical analysis of market data using climate narrative indexes is the next step (like Engle et al. [2020], Faccini et al. [2023]).

- A possible next topic: how central bank narratives about climate risk have affected the behavior of other economic entities?

"*An economic narrative is a contagious story that can change how people make economic decisions*" (Shiller [2020]).

Reference

- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre- training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
- R. F. Engle, S. Giglio, B. Kelly, H. Lee, and J. Stroebel, “Hedging climate change news,” The Review of Financial Studies, vol. 33, no. 3, pp. 1184–1216, 2020.
- R. Faccini, R. Matin, and G. Skiadopoulos, “Dissecting climate risks: Are they reflected in stock prices?” Journal of Banking & Finance, vol. 155, p. 106948, 2023.
- K. Hiraki, N. Kaneda, T. Kimata, and T. Mtsue, “Climate change news indices: Are they reflected in japanese stock prices?” , IMES Discussion Paper Series, 2025, forthcoming.
- R. J. Shiller, *Narrative economics: How stories go viral and drive major economic events.* Princeton University Press, 2020.
- L. Pastor, R. F. Stambaugh, and L. A. Taylor, “Dissecting green returns,” Journal of Financial Economics, vol. 146, no. 2, pp. 403–424, 2022.
- H. Sakaji, S. Sekine, and S. Masuyama, “Extracting causal knowledge using clue phrases and syntactic patterns,” in Practical Aspects of Knowledge Management. Springer Berlin Heidelberg, 2008, pp. 111–122.
- H. Sakaji, R. Murono, H. Sakai, J. Bennett, and K. Izumi, “Discovery of rare causal knowledge from financial statement summaries,” in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 2017, pp. 602–608.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Gen AI in action: transforming data use in suptech[1]

Patrick Hoffmann, Maha El Dimachki
and Sivasubramanian Ramanathan,
BIS Innovation Hub Singapore Centre

Kenneth Gay and Harry Lee,
Monetary Authority of Singapore

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Gen AI in action: transforming data use in suptech

Patrick Hoffmann,[1] Maha El Dimachki,[2] Kenneth Gay,[3] Sivasubramanian Ramanathan,[4] Harry Lee[5]

Abstract

Supervisory activities are fundamentally dependent on data; however, much of these data tend to be based on backward-looking templates, limiting their effectiveness and relevance. To tackle these challenges, central banks and regulatory authorities are increasingly embracing generative artificial intelligence (gen AI) as a transformative solution. For instance, gen AI can streamline the process of summarising financial reports and news articles, as well as evaluating banks' compliance with existing regulations, thereby enhancing the efficiency and accuracy of supervisory activities. This paper delves into a collaborative initiative spearheaded by the Bank for International Settlements Innovation Hub (BISIH) Singapore Centre, in partnership with the Monetary Authority of Singapore (MAS). This initiative empowers central banks and financial authorities to explore and develop shareable code for gen AI applications within the realm of supervisory technology (suptech). Additionally, we present the current use cases being investigated by this community and outline the potential future applications of gen AI in enhancing supervisory practices. Through these efforts, the initiative seeks to modernise supervisory methodologies, ensuring that they remain relevant and effective in an ever-evolving financial landscape.

1. Introduction

Central banks play a crucial role in maintaining financial stability and ensuring sound banking practices. To effectively supervise banks, they need access to accurate, timely and comprehensive data. These data serve as the backbone for assessing risk, monitoring compliance and identifying potential vulnerabilities within the financial system (Araujo et al (2023), Duijm and van Lelyveld (2025)). A major problem that central banks face in banking supervision is that traditional data collection methods have often relied on backward-looking, template-based systems that can limit responsiveness and hinder the ability to make informed decisions. Generative artificial intelligence (gen AI) offers a transformative means for central banks to process new and vast volumes of data more efficiently and effectively (BIS (2024), El Dimachki (2024)). With gen AI, central banks can harness advanced analytics to derive deeper insights from available information, allowing for real-time monitoring of banking activity and more dynamic risk assessments. With its potential to enhance predictive

[1] Corresponding author; Adviser, Bank for International Settlements Innovation Hub Singapore Centre
[2] Head, Bank for International Settlements Innovation Hub Singapore Centre
[3] Chief FinTech Officer, Monetary Authority of Singapore
[4] Junior Product Owner, Bank for International Settlements Innovation Hub Singapore Centre
[5] Deputy Director and Head, Data Governance and Transformation, Monetary Authority of Singapore

capabilities, gen AI can help central banks navigate the complexities of modern financial systems across both advanced and emerging markets. Embracing gen AI promises not only to improve operational efficiency but also to strengthen financial oversight, and it promises to enhance overall stability in the banking sector (Ashley and Papenbrock (2023), Avramovic (2023)).

However, emerging technologies such as gen AI necessitate enhanced collaboration among central banks to remain at the forefront of developments (BIS (2024)). Collaboration is also needed to ensure knowledge-sharing of the risks and experiences of using these technologies, and joint experimentation will enhance and accelerate upskilling and learning on how to deploy these tools safely (Crisanto et al (2024)). The Ellipse Data and Knowledge Platform Collaboration Community (EDKP CC) addresses this need by providing a platform for collective innovation. Facilitated by the BIS Innovation Hub (BISIH) Singapore Centre and the Monetary Authority of Singapore (MAS), the community builds on the modular Ellipse Data and Knowledge Platform (EDKP) developed under Project Ellipse – an integrated regulatory data and analytics platform (BIS (2022)). The community offers a unique avenue for central banks and financial regulators to share source code, pool expertise and jointly advance suptech innovation. By collaborating on knowledge, resources (like development capacity) and solutions, the EDKP CC aims to speed up collective progress in these fields. The more impactful use cases discussed in the community may also add to the BISIH's project pipeline (for example, using the digital twin technology for climate risk stress testing).

2. The EDKP CC – innovation through collaboration

In 2021, the BISIH Singapore Centre and the MAS embarked on a new initiative known as Project Ellipse (BIS (2022)). This collaborative venture, which also included the Bank of England, the International Swaps and Derivatives Association, Financial Network Analytics and Accenture, was designed to foster a more proactive, insight-driven and data-centric approach to regulatory supervision. At its core, Project Ellipse focused on the development of an integrated regulatory data and analytics platform. A standout feature of the Ellipse prototype is its innovative capability to seamlessly integrate both structured and unstructured data sources. This integration is crucial for incorporating into supervision real-time information concerning current events that may affect financial stability and market dynamics. Through the application of advanced analytics on these consolidated data sets, regulators gain valuable early warning indicators that empower supervisors to act pre-emptively in response to potential risks (Hoffmann et al (2025)).

Building on the initial achievements of Project Ellipse, the BISIH Singapore Centre and MAS have taken significant strides to enhance this modular platform. This evolution has transformed Ellipse into a foundational suptech tool, positioned to support central banks and financial authorities in their experiments with emerging technologies, particularly in the realm of gen AI. The EDKP CC has since flourished, with membership growing to 19 institutions, including notable entities such as Bank Indonesia, the European Central Bank and the Office of the Superintendent of Financial Institutions in Canada (BIS (2023)). The EDKP CC serves as a forum where members convene regularly to exchange trends in financial supervision and insights

on other contemporary issues. The community's collaborative nature is evidenced by regular hackathons that focus on developing innovative solutions to shared challenges. A dedicated GitHub environment facilitates the sharing of code among member central banks and financial authorities, enabling collaborative development and promoting best practices. The community's collaboration is structured around several key venues:

1. **Virtual meetings**, held bimonthly for one hour, provide members with the opportunity to present their ongoing work in domains like suptech, gen AI and green finance/climate risk. Guest speakers are occasionally invited to share insights and expertise.
2. **An annual virtual workshop** encourages deep exploration of specific topics, such as the limitations and ethical considerations of gen AI. Topics for discussion are determined through member polls, reflecting a participative and inclusive decision-making process to capture the collective interests of the community.
3. **Hackathons** – the community's flagship event – are in-person and held annually. Preparation begins well in advance, from brainstorming potential use cases, to forming diverse teams across various authorities, to establishing the necessary technical infrastructure. All code produced during the hackathon is shared with members through the EDKP GitHub.[6] Having a common GitHub among the community members facilitates iterative improvement and collective learning, ensuring that innovations are accessible to the entire community. The insights and innovations generated during hackathons extend beyond the community's own interests, as the BISIH Singapore Centre frequently receives inquiries to showcase successful use cases from the hackathon, illustrating how central banks can collaboratively drive innovation.

Through its initiatives, the EDKP CC not only fosters a culture of collaboration and innovation but also underscores the vital role that advanced technologies can play in enhancing regulatory practices and maintaining financial stability in an ever-evolving global landscape.

3. How gen AI transforms data use in suptech

The most recent EDKP CC hackathon was held in June 2024, focusing on harnessing the capabilities of large language models (LLMs) to address prevalent suptech needs. This three-and-a-half-day event drew participation from over 50 individuals representing 15 different central banks and financial institutions, fostering a collaborative environment rich with diverse expertise. Participants worked in mixed teams to tackle four specific gen AI-related use cases that had been proposed in advance by the members of the collaboration community. These use cases were carefully selected based on their relevance and potential impact on financial supervision and regulation. The organisations involved included notable central

[6] The EDKP GitHub is an access-controlled repository where participants can collaborate on code, track version histories, document best practices and experiment with emerging tools.

banks and regulatory bodies, reflecting a broad spectrum of perspectives and challenges in the field.

To enhance collaboration and outcomes, the facilitators of the hackathon, the BISIH Singapore Centre and the MAS, matched participants with cross-institutional teams based on their individual skills, interests and experience. This approach not only promoted teamwork but also ensured that each group was equipped with the right mix of talents to effectively address the challenges posed by each use case. The use cases were as follows:

1. **AI-powered regulatory insight:** This use case, undertaken by participants from Labuan Financial Services Authority and Bank Indonesia, explored the application of gen AI in summarising and analysing regulatory reports and disclosures. The team aimed to develop a system capable of extracting insights regarding a bank's financial condition and generating automated reports that summarise financial statement assessments.
2. **AI-driven financial newsletters:** A collaborative effort between participants from the Bank of Japan, Swiss Financial Market Supervisory Authority and BISIH London Centre investigated the potential of gen AI in scanning an information feed for anything of interest. The team focused on developing a system that automatically generates and distributes newsletters containing only articles that covered topics of interest (eg credit risk), thereby reducing manual effort and mitigating information overload for recipients.
3. **Navigating supervisory frameworks with AI:** The Bangko Sentral ng Pilipinas team developed the Artificial Intelligence Hybrid Analysis Supervision (AIHAS) system. This system functions as an intelligent assistant for bank supervisors, leveraging supervisory and regulatory data to provide rapid responses to queries and generate comprehensive assessments. The AIHAS system also aims to automate the evaluation of bank documents for compliance with relevant policies and regulations, identifying potential risks associated with banking operations and enhancing productivity by automating repetitive tasks.
4. **Enhancing supervisory reporting:** This use case involved a multi-institutional team comprising members from the Deutsche Bundesbank, Office of the Superintendent of Financial Institutions, Central Bank of Malaysia and Bank of Italy. The team focused on several aspects of supervisory reporting, including the development of systems for answering questions, as well as retrieving and summarising official documents. Additionally, the team explored the application of AI in conducting market analysis of commercial real estate reports.

At the conclusion of the EDKP CC hackathon, the code developed during the event was shared with all community members, allowing them to build on the insights and innovations generated during the sessions.

A prominent theme emerges from the diverse use cases identified for the hackathon: the urgent need to efficiently and accurately extract actionable insights from vast amounts of data. Whether the focus is on analysing complex financial reports, pinpointing relevant information or automating compliance checks, the ability to effectively process and interpret data is crucial for regulatory bodies tasked

with maintaining the integrity of financial systems. Participating institutions represent a global cross-section of regulatory and supervisory authorities, and they are increasingly turning to gen AI as a promising solution to address this pressing challenge. Gen AI's innovative capabilities, especially in analysing unstructured data and generating human-like text, offer significant potential for streamlining a variety of processes. This technology can automate routine tasks, thereby saving valuable time and resources. Furthermore, gen AI can enhance decision-making by providing supervisors with advanced analytical tools to derive insights from complex data sets, improving predictive analytics for risk assessment. As the financial landscape becomes increasingly complex, embracing gen AI could mark a pivotal step towards achieving a more robust and responsive financial supervision and regulatory framework. The EDKP CC looks forward to further exploring these use cases and continuing to innovate in the suptech landscape.

To deepen collaboration, the EDKP CC can build further on its existing mechanisms for collaboration (eg knowledge exchange, common codebase). Putting in place a sandbox environment for members to test solutions quickly can both speed up prototyping and level the playing field for members which may face different technical infrastructural challenges. Having in place a common codebase and common technical infrastructure will also pave the way for common upskilling initiatives, as well as set up a community of mentors within the EDKP CC that will accelerate capacity building. There is also scope to develop central services that support the wider EDKP CC, such as in vulnerabilities management where a central resource manages the codebase to, for example, address unearthed zero-day vulnerabilities, replace deprecated packages or develop standardised technical documentation.

Beyond developments that are primarily technical in nature, it is also essential for the members of the EDKP CC to consider frameworks that guide the responsible use of AI and data analytics. For example, the MAS introduced the Fairness, Ethics, Accountability, and Transparency (FEAT) Principles to ensure that AI-driven solutions in the financial sector maintain robust standards of governance and trust (MAS (2018)). Additionally, MAS partnered with the financial industry on the Veritas Initiative, providing a framework and toolkit to assess AI and data analytics solutions against the FEAT criteria (MAS (2019)). Although the EDKP CC does not focus on formulating policy, it discusses these established frameworks and toolkits to inform the development and deployment of emerging technologies.

4. The path forward: opportunities and challenges

The EDKP CC will continue its discussion on how to benefit from gen AI, with a particular focus on banking supervision. During the hackathon, the members of the EDKP CC expressed a strong interest in further exploring the potential use of gen AI in the following areas:

1. **Data analysis and reporting:** GenAI can automate the analysis of large data sets from financial institutions, helping supervisors quickly identify trends,

anomalies and potential risks. This can streamline the preparation of reports and improve the speed and accuracy of data interpretation.

2. **Risk assessment:** Gen AI can model complex risk scenarios by simulating various economic conditions and their impact on banks. This can help supervisors assess the resilience of financial institutions under stress and evaluate capital adequacy.
3. **Fraud detection:** Gen AI could play a significant role in combating financial crime, including anti-money laundering (AML), countering the financing of terrorism (CFT), fraud detection and scam prevention. For example, gen AI can assist in identifying suspicious activities that may indicate fraud or money laundering by analysing transaction patterns and behaviours. This can improve the effectiveness of AML efforts in real time.
4. **Enhanced communication:** Gen AI can assist in generating clear and concise communication materials for stakeholders, including regulatory guidelines, policy updates and supervision outcomes, ensuring clarity and consistency.
5. **Supervisory tools development:** Central banks can use gen AI to develop predictive tools that provide early warnings about potential failures within institutions. These tools can assess financial health indicators and trigger alerts if certain thresholds are breached.
6. **Stress testing:** Gen AI can enhance stress testing frameworks by creating more sophisticated economic scenarios and modelling how banks would respond to various shocks, thus improving the reliability of these exercises.
7. **Sentiment analysis:** By analysing news articles, social media and market reports, gen AI can help supervisors gauge the sentiment and perception of financial institutions, which can serve as an additional indicator of stability or risk.
8. **Regulatory compliance monitoring:** Gen AI can assist in monitoring compliance with regulations by analysing transaction data, enabling proactive identification of non-compliance issues. Moreover, gen AI could be used to enhance compliance monitoring by automatically checking internal guidelines against sanctions lists, regulatory requirements and legal frameworks. Additionally, gen AI could be deployed to analyse transaction-level data and related textual content to identify suspicious patterns and potential illicit activities (Araujo et al (2024)).
9. **Training and education:** Gen AI can be used to create training simulations and educational content for bank supervisors, helping them stay up to date on best practices, emerging risks and regulatory changes.
10. **Scenario analysis:** By generating various hypothetical scenarios, gen AI can help supervisors test the resilience of banks against a range of economic conditions and events, aiding future-proofing efforts.
11. **Climate risk:** Climate change has severe implications for central banks and financial authorities (Ramanathan et al (2024)). However, gen AI offers powerful tools for analysing and visualising climate risk. For example, multi-modal foundation models could be used to process large data sets, including satellite data, to assess climate-related financial risks and inform the

development of climate-focused suptech tools, such as digital twins (Hoffmann et al (2025)).

By adopting gen AI in these ways, central banks can enhance their supervisory capabilities and knowledge, making the oversight of banking institutions more responsive to rapid changes in the financial landscape. While the potential of gen AI in suptech is significant, it is crucial to acknowledge the inherent challenges and risks. Ensuring the accuracy, reliability and explainability of AI-driven systems, addressing potential biases and navigating the ethical implications of AI use in financial supervision are all critical considerations. Further joint research and development, coupled with robust governance frameworks, will be essential to harness the power of gen AI while mitigating potential risks. By collaborating with international partners and tapping into global knowledge networks, central banks and supervisors can accelerate their gen AI learning and adoption while ensuring these solutions have incorporated best practices to mitigate risks.

5. References

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2023): "Machine learning applications in central banking", *Journal of AI, Robotics & Workplace Automation*, vol 2, no 3, pp 271–93.

Araujo, D, S Doerr, L Gambacorta and B Tissot (2024): "Artificial intelligence in central banking", *BIS Bulletin*, no 84, January.

Ashley, J and J Papenbrock (2023): "Modern computing platforms as key technology for central banks, financial supervisors, and regulators", *IFC Bulletin,* no 59, October.

Avramovic, P (2023): "Digital transformation of financial regulators and the emergence of supervisory technologies (suptech): a case study of the UK Financial Conduct Authority", *Harvard Data Science Review,* no 5.2, spring.

Bank for International Settlements (BIS) (2022): *Project Ellipse: an integrated regulatory data and analytics platform*, March, www.bis.org/publ/othp48.pdf.

——— (2023): "Ellipse and DKP Collaboration Community: terms of reference", 31 May, www.bis.org/pdf/EDKP_License_Agreement.pdf.

——— (2024): *Annual Economic Report*, June. www.bis.org/publ/arpdf/ar2024e.pdf.

Crisanto, J C, C B Leuterio, J Prenio and J Yong (2024): "Regulating AI in the financial sector: recent developments and main challenges", *FSI Insights*, no 63, December.

Duijm, P and I van Lelyveld (2025): "Data science for central banks and supervisors: how to make it work, actually", *Harvard Data Science Review*, no 7.1, winter.

El Dimachki, M (2024): *Fintech regulation in practice: navigate the complexities and opportunities of regulation*, Kogan Page.

Hoffmann, P, M El Dimachki, K Gay, L Grisey and Z Bossert (2025): "Introducing the work program of the BIS Innovation Hub Singapore Centre on green finance and climate risk data", *IFC Bulletin*, forthcoming.

Monetary Authority of Singapore (MAS) (2018): "MAS introduces new FEAT Principles to promote responsible use of AI and data analytics", press release, 12 November, www.mas.gov.sg/news/media-releases/2018/mas-introduces-new-feat-principles-to-promote-responsible-use-of-ai-and-data-analytics.

——— (2019): "MAS partners financial industry to create framework for responsible use of AI", press release, 13 November, www.mas.gov.sg/news/media-releases/2019/mas-partners-financial-industry-to-create-framework-for-responsible-use-of-ai.

Ramanathan, S, P Hoffmann, K Gay and H Lee (2024): "Project Viridis: a blueprint for managing climate-related financial risk", OMFIF *SPI Journal*, Q3.

BIS Innovation Hub

MAS Monetary Authority of Singapore



GenAI in Action: Transforming Data Use in SupTech

Dr Patrick Hoffmann, Maha El Dimachki, Kenneth Gay, Sivasubramanian Ramanathan, Harry Lee

February 2025

EDKP CC: Innovation through collaboration

- Supervisory activities heavily rely on data, but this data is often template-based and outdated.
- Therefore, central banks and regulatory authorities are increasingly turning to **Generative Artificial Intelligence (GenAI)** to address these challenges.
- The EDKP Collaboration Community, facilitated by the **BISIH Singapore Centre** and the **Monetary Authority of Singapore**, offers a unique way to jointly experiment with new technologies, for instance through knowledge-sharing and pooling of expertise.



Introducing the EKPD Collaboration Community

…started as a pilot in 2022

…was extended until 2026

…has grown to 19 members and many observers

…should be extended beyond 2026

BIS Innovation Hub

BANK INDONESIA

DeNederlandscheBank
EUROSYSTEEM

BANK OF MAURITIUS

PIDM
Perbadanan Insurans Deposit Malaysia
Protecting Your Insurance And Deposits In Malaysia

Superintendencia Financiera de Colombia

finma

Labuan FSA
Financial Services Authority

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

SUPERINTENDENCIA
DE BANCA, SEGUROS Y AFP
República del Perú

DEUTSCHE BUNDESBANK
EUROSYSTEM

MAS
Monetary Authority of Singapore

South African Reserve Bank

BANCA D'ITALIA

OSFI
BSIF

MAGYAR NEMZETI BANK

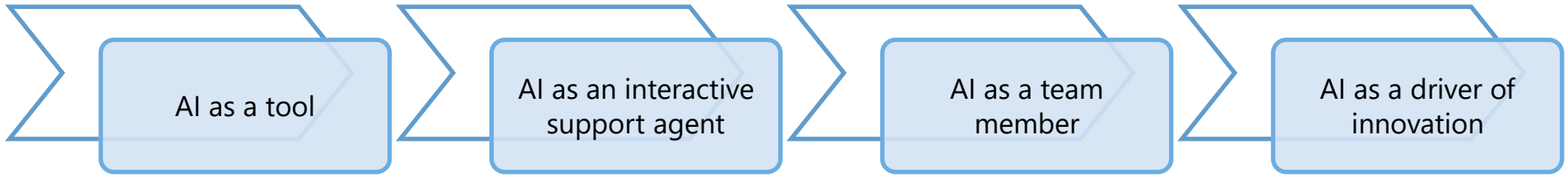
BANK OF JAPAN

BANGKO SENTRAL NG PILIPINAS

Currently, most central banks use AI…

- … as an innovation tool, or
- … as an interactive support agent.

The role of AI in central banking will develop over time with increasing capabilities of AI.



In June 2024, the EDKP CC met for a hackathon in Singapore…

- … Round about 50 people participated in cross-institutional teams to work on six use cases, four of them on GenAI.

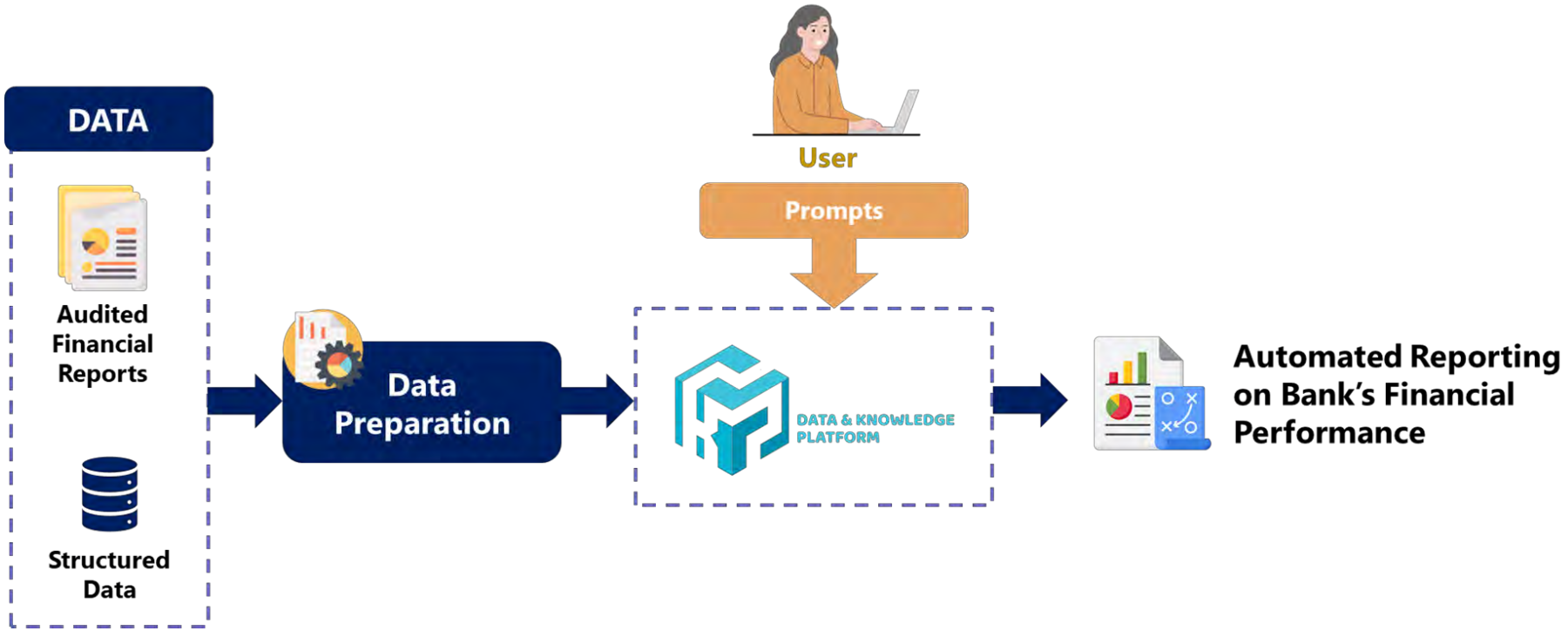






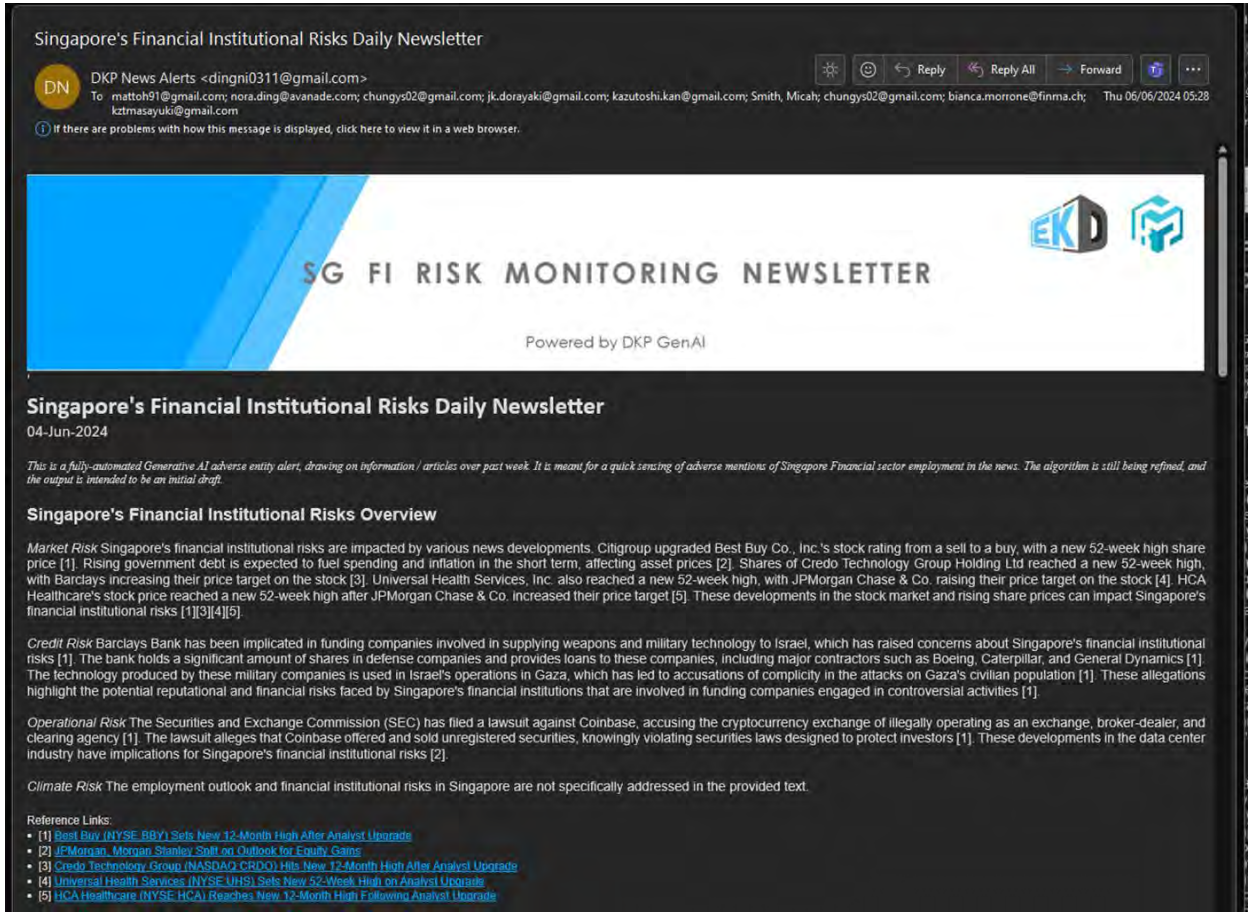
EDKP CC Hackathon - GenAI use case 1

- **AI-Powered Regulatory Insight:** Team 1 explored the application of GenAI in summarizing and analyzing regulatory reports and disclosures. The team aimed to develop a system capable of extracting insights regarding a bank's financial condition and generating automated reports that summarize financial statement assessments.



EDKP CC Hackathon - GenAI use case 2

- **AI-Driven Financial Newsletters:** Team 2 investigated the potential of GenAI in scanning information feed for information of interest. The team focused on developing a system that automatically generates and distributes newsletters containing only articles that covered topics of interests (e.g. credit risks), thereby reducing manual effort and mitigating information overload for recipients.



Singapore's Financial Institutional Risks Daily Newsletter

DKP News Alerts <dingni0311@gmail.com>
To mattoh91@gmail.com; nora.ding@avanade.com; chungys02@gmail.com; jk.dorayaki@gmail.com; kazutoshi.kan@gmail.com; Smith, Micah; chungys02@gmail.com; bianca.morrone@finma.ch; kztmasayuki@gmail.com
Thu 06/06/2024 05:28

If there are problems with how this message is displayed, click here to view it in a web browser.

SG FI RISK MONITORING NEWSLETTER

Powered by DKP GenAI

Singapore's Financial Institutional Risks Daily Newsletter

04-Jun-2024

This is a fully-automated Generative AI adverse entity alert, drawing on information / articles over past week. It is meant for a quick sensing of adverse mentions of Singapore Financial sector employment in the news. The algorithm is still being refined, and the output is intended to be an initial draft.

Singapore's Financial Institutional Risks Overview

Market Risk Singapore's financial institutional risks are impacted by various news developments. Citigroup upgraded Best Buy Co., Inc.'s stock rating from a sell to a buy, with a new 52-week high share price [1]. Rising government debt is expected to fuel spending and inflation in the short term, affecting asset prices [2]. Shares of Credo Technology Group Holding Ltd reached a new 52-week high, with Barclays increasing their price target on the stock [3]. Universal Health Services, Inc. also reached a new 52-week high, with JPMorgan Chase & Co. raising their price target on the stock [4]. HCA Healthcare's stock price reached a new 52-week high after JPMorgan Chase & Co. increased their price target [5]. These developments in the stock market and rising share prices can impact Singapore's financial institutional risks [1][3][4][5].

Credit Risk Barclays Bank has been implicated in funding companies involved in supplying weapons and military technology to Israel, which has raised concerns about Singapore's financial institutional risks [1]. The bank holds a significant amount of shares in defense companies and provides loans to these companies, including major contractors such as Boeing, Caterpillar, and General Dynamics [1]. The technology produced by these military companies is used in Israel's operations in Gaza, which has led to accusations of complicity in the attacks on Gaza's civilian population [1]. These allegations highlight the potential reputational and financial risks faced by Singapore's financial institutions that are involved in funding companies engaged in controversial activities [1].

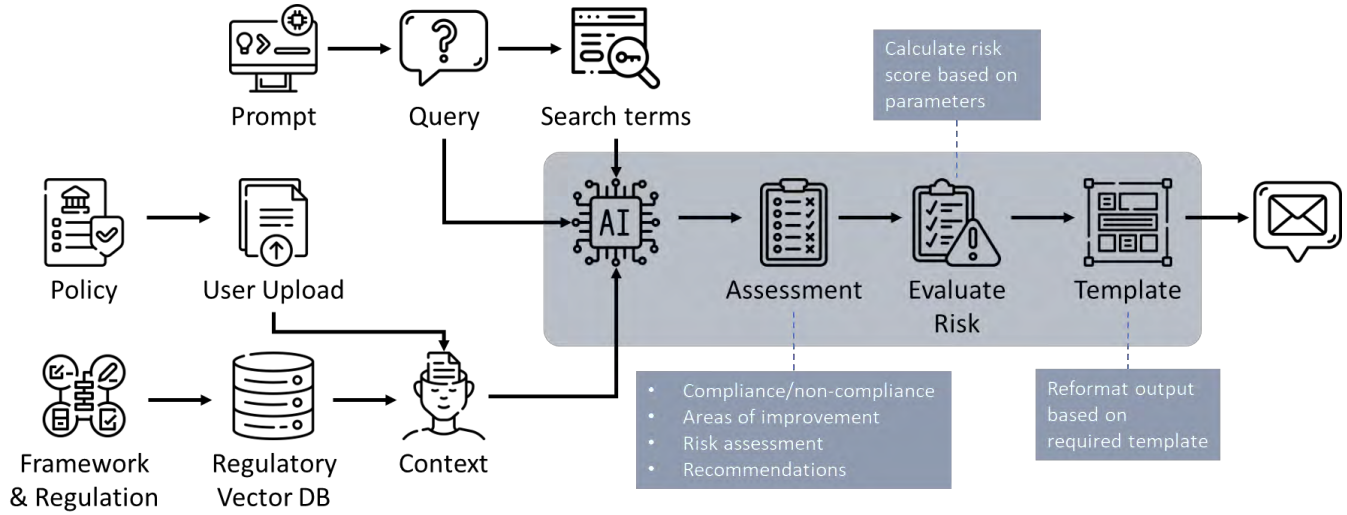
Operational Risk The Securities and Exchange Commission (SEC) has filed a lawsuit against Coinbase, accusing the cryptocurrency exchange of illegally operating as an exchange, broker-dealer, and clearing agency [1]. The lawsuit alleges that Coinbase offered and sold unregistered securities, knowingly violating securities laws designed to protect investors [1]. These developments in the data center industry have implications for Singapore's financial institutional risks [2].

Climate Risk The employment outlook and financial institutional risks in Singapore are not specifically addressed in the provided text.

Reference Links:
- [1] Best Buy (NYSE:BBY) Sets New 12-Month High After Analyst Upgrade
- [2] JPMorgan, Morgan Stanley Split on Outlook for Equity Gains
- [3] Credo Technology Group (NASDAQ:CRDO) Hits New 12-Month High After Analyst Upgrade
- [4] Universal Health Services (NYSE:UHS) Sets New 52-Week High on Analyst Upgrade
- [5] HCA Healthcare (NYSE:HCA) Reaches New 12-Month High Following Analyst Upgrade

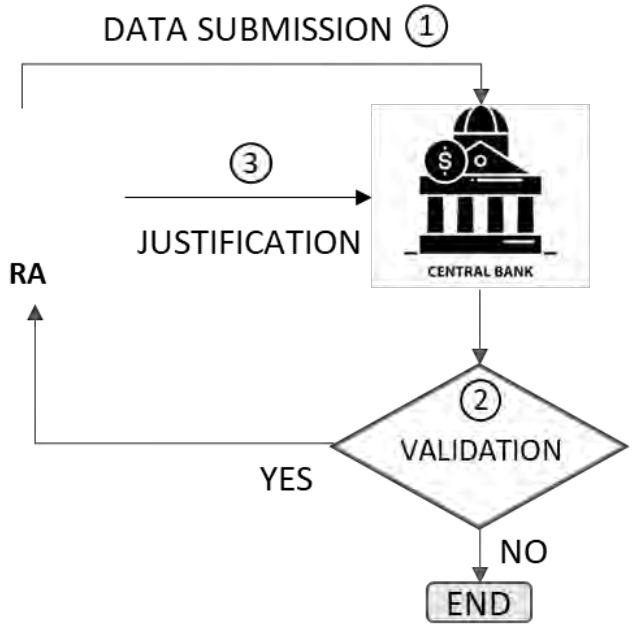
EDKP CC Hackathon - GenAI use case 3

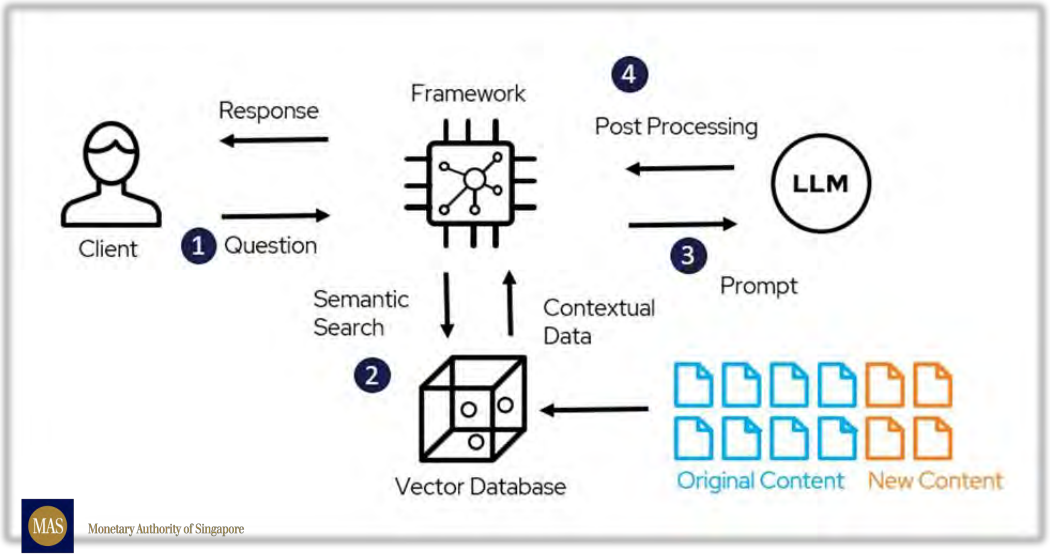
- **Navigating Supervisory Frameworks with AI:** Team 3 developed the "Artificial Intelligence Hybrid Analysis Supervision" (AIHAS) system. This system functions as an intelligent assistant for bank supervisors, leveraging supervisory and regulatory data to provide rapid responses to queries and generate comprehensive assessments. AIHAS also aims to automate the evaluation of bank documents for compliance with relevant policies and regulations, identifying potential risks associated with banking operations, and enhancing productivity by automating repetitive tasks.

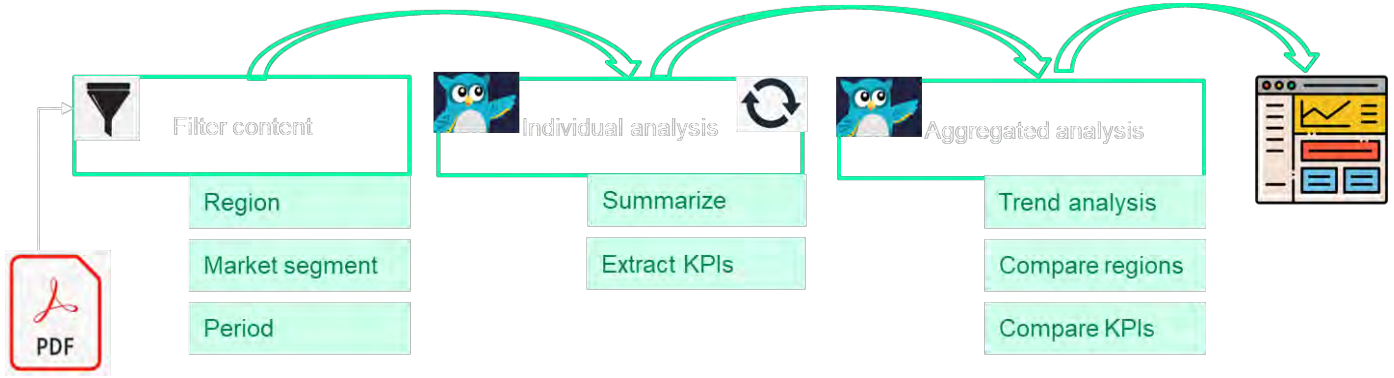


EDKP CC Hackathon - GenAI use case 4

- **Enhancing Supervisory Reporting:** Team 4 focused on several aspects of supervisory reporting, including the development of systems for question answering, retrieval and summarization of official documents, and automation and scaling of text similarity assessments in regulatory reporting to streamline compliance analysis. Additionally, the team explored the application of AI in conducting market analysis of commercial real estate reports.



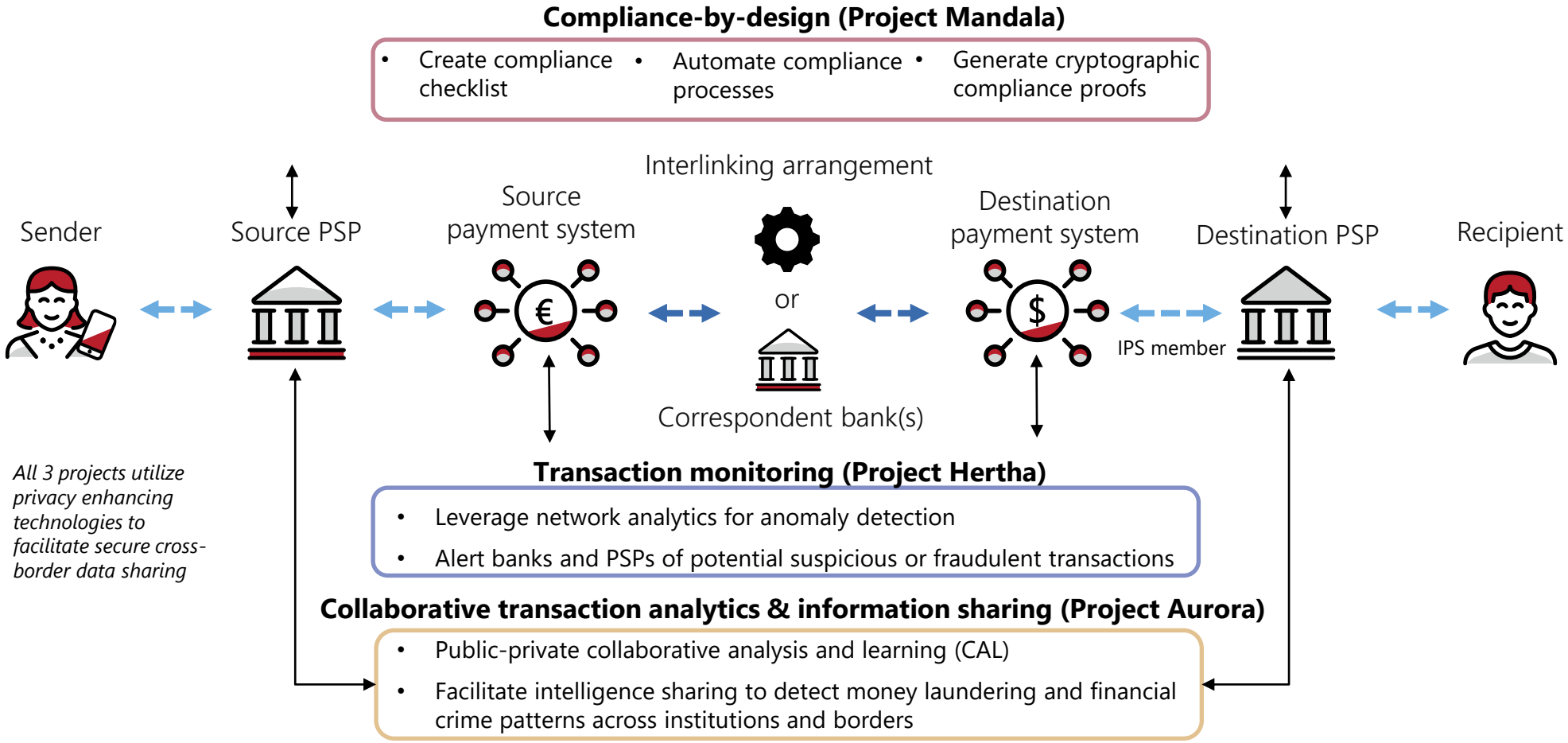




EDKP CC Hackathon - Lessons learned

- **A common theme emerges from these diverse use cases:** the need to efficiently and accurately extract meaningful insights from large volumes of data.
- Whether it is analyzing financial reports, identifying information of interest, or automating compliance checks, the ability to process and understand data is paramount in financial supervision.
- The participating institutions, representing a global cross-section of regulatory bodies, are increasingly looking towards GenAI as a potential solution to this challenge.
- GenAI, with its capacity to analyze unstructured data, and generate human-like text, holds promise for automating tasks, improving decision-making, and ultimately enhancing the effectiveness of financial supervision.

The path forward: Potential and challenges for GenAI in SupTech – Combating financial crime



The path forward: Potential and challenges for GenAI in SupTech – Analyzing and visualizing climate risk



- Supervisors face challenges accessing and analyzing unstructured, underutilized environmental data impacting financial institutions.
- **How AI Helps:**
  - Processes unstructured environmental data at scale
  - Identifies patterns and correlations in complex datasets
  - Generates initial risk assessments and automated reports
  - Provides actionable insights as a foundation for deeper analysis
- **BISIH's Goal:**
  Develop a Digital Twin MVP that leverages AI to enable supervisors to interactively assess physical climate risks using the most up-to-date data, enhancing decision-making.

BIS Innovation Hub

MAS Monetary Authority of Singapore



Thank you very much for your attention.
Any questions?

Dr Patrick Hoffmann, Maha El Dimachki, Kenneth Gay, Sivasubramanian Ramanathan, Harry Lee

February 2025

References

- Alonso-Robisco, A., Marques, J. M. (2024). Finding the balance with technology. OMFIF SPI Journal.
- Araujo, D., Doerr, S., Gambacorta, L., Tissot, B. (2024). Artificial intelligence in central banking. BIS Bulletin (84).
- Ashley, J., Papenbrock, J. (2023). Modern computing platforms as key technology for central banks, financial supervisors, and regulators. IFC Bulletin (59).
- Bank for International Settlements (BIS) (2022): Project Ellipse: An integrated regulatory data and analytics platform. www.bis.org/publ/othp48.pdf.
- Bank for International Settlements (BIS) (2023): Ellipse and DKP collaboration community – Terms of reference. www.bis.org/pdf/EDKP_License_Agreement.pdf.
- Bank for International Settlements (BIS) (2024): Annual economic report. https://www.bis.org/publ/arpdf/ar2024e.pdf
- El Dimachki, M. (2024). Fintech regulation in practice: Navigate the complexities and opportunities of regulation. Kogan Page.
- Hoffmann, P.; El Dimachki, M.; Gay, K.; Grisey, L.; Bossert, Z. (forthcoming). Introducing the work program of the BIS Innovation Hub Singapore Centre on green finance and climate risk data. IFC Bulletin.
- Ramanathan, S., Hoffmann, P., Gay, K., Lee, H. (2024). Project Viridis: A blueprint for managing climate-related financial risk. OMFIF SPI Journal.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

AI-driven legal summarization for financial regulation monitoring[1]

Diana López Avilés and Isidora Díaz Jorquera,
Central Bank of Chile

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

AI-Driven Legal Summarization for Financial Regulation Monitoring

Diana López Avilés[1], Isidora Díaz Jorquera[2]

Abstract

In Chile, financial regulation is becoming increasingly intricate due to the rapid expansion of the digital payments ecosystem. This growth has brought a surge of new market participants, evolving business models, and changing competitive dynamics — all of which contribute to more frequent and complex legal disputes. As a result, regulatory bodies, including the Central Bank of Chile, must navigate a growing volume of nuanced legal material to monitor compliance and actively oversee emerging market trends. To address this challenge, this article proposes a generative AI application that automates the summarization of legal arguments in cases involving payment and banking disputes, using public daily status data from Chile's Court of Defense of Free Competition. By fine-tuning mBART-50, a multilingual transformer model, the tool produces specialized one-page summaries that reduce the Central Bank's manual workload, supporting the early identification of market pressure points and providing insights into potential shifts in the regulatory environment — thereby strengthening financial oversight and enabling more agile policy responses.

Keywords: Legal summarization, GenAI, Financial regulation, Digital payments

JEL classification: C45, G28, K22

Content

1. Introduction 2

2. Data 4

3. Methodology 6

Data Preprocessing 6

Model Selection 6

Fine-Tuning Strategy 7

Evaluation 8

[1] Central Bank of Chile (dlopez@bcentral.cl)

[2] Central Bank of Chile (idiaz@bcentral.cl)

The views are those of the authors and do not necessarily reflect the views of the Central Bank of Chile or its board members. We thank Rosario Celedón, Gabriel Aparici and Pablo Furche for their valuable comments and suggestions.

4. Results 8

5. Concluding remarks 11

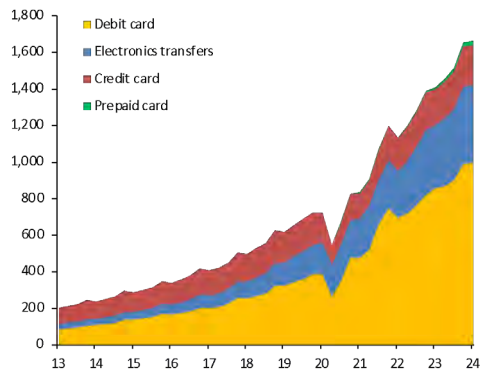
6. References 12

1. Introduction

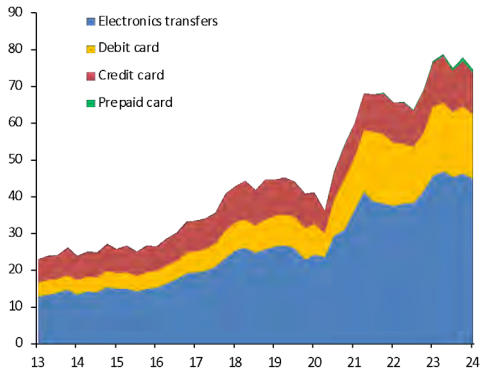
The landscape of digital payments in Chile has undergone significant transformation in the past decade, with rapid growth in both number and amount of transactions. According to the Central Bank of Chile's *Informe de Sistemas de Pago* 2024 [1], digital payment activity—including both card-based payments and electronic fund transfers—has grown steadily, reflecting a nationwide move away from cash-based transactions. As of early 2024, the annual volume of digital payments surpassed six billion transactions, marking a substantial increase from the previous year (Graph 1). This equates to an average of over 300 digital payments per person, underscoring the maturity of Chile's transition to a more integrated and accessible digital financial infrastructure [1].

Digital payments in Chile — Graph 1

Number of transactions (millions of transactions, quarterly) — 1.1



Amount of transactions (billions of dollars, quarterly) — 1.2



* Credit and debit card data include both businesses and individuals. TEF (Electronic Funds Transfers) data only account for individuals and include both intrabank and interbank transactions.

Sources: Central Bank of Chile based on data from the Financial Market Commission (CMF) [1]

Multiple structural factors have contributed to this shift. The broadening availability of payment instruments and the continued rollout of point-of-sale (POS) systems by both traditional and emerging actors have significantly improved the user experience. Meanwhile, the ecosystem of electronic transfers has enabled seamless peer-to-peer (P2P) payments, with these transfers accounting for a disproportionately large share of transaction value despite representing a smaller portion of total transactions [1]. By contrast, person-to-business (P2B) transfers via EFT channels remain less prevalent, pointing to ongoing gaps in merchant integration.

Payment card networks have generally led the way in terms of interoperability and merchant reach. These schemes benefit from legacy arrangements among issuers, acquirers, and network providers, which allow for coordinated rule-setting and the early adoption of new technologies such as contactless payments, QR code integration, and mobile wallets. This structural cohesion contrasts with the more fragmented development of EFT systems in the P2B segment, where industry-wide standards and real-time merchant interfaces are still evolving.

However, the rapid expansion of the digital payments sector has introduced a new layer of institutional and legal complexity. The growing involvement of fintechs, cross-border acquirers, and payment service providers (PSPs) has led to disputes over fee structures, network participation rules, and operational limitations. These tensions have increasingly been brought before Chile's Tribunal for the Defense of Free Competition (TDLC) and other oversight bodies, as participants seek clarity on competitive conduct and market access. Recent decisions by the Supreme Court have addressed these concerns by establishing clearer guidelines on cross-border acquiring, pricing transparency, and non-discriminatory card acceptance. In doing so, the Court has effectively shifted some aspects of private network governance into the public regulatory domain [2].

Regulators have responded accordingly. The Central Bank of Chile recently expanded its regulatory framework to oversee PSPs and cross-border acquiring arrangements, including bringing certain actors under the early supervision of the Financial Market Commission (CMF) to mitigate operational risks to merchants. Simultaneously, the CMF has launched implementation of the Open Finance System (SFA), which will —over a multi-year horizon— integrate third-party providers such as payment initiators into the financial ecosystem, enabling them to initiate transfers on behalf of users with proper authorization [3]. These reforms represent a step toward greater modularity in financial services, as well as more active oversight of previously under-regulated business models.

In parallel, these regulatory shifts have increased the informational demands placed on oversight institutions. The BCCh, in particular, is tasked under its Organic Constitutional Law with safeguarding the integrity of both domestic and international payments. Fulfilling this mandate requires real-time awareness of legal and institutional developments in the market. Yet the volume and complexity of legal proceedings —particularly those involving competition law or the introduction of novel financial technologies— pose challenges for timely analysis. Documents submitted to the TDLC and similar bodies are frequently lengthy, domain-specific, and difficult for non-specialist audiences to fully grasp.

Moreover, given the underdevelopment of certain market sectors, it is likely that frictions between incumbent actors and new entrants will re-emerge as activity expands into these areas. While some disputes have been addressed through recent regulatory and judicial decisions, evolving frameworks and diversified business models may reactivate tensions around access, pricing, and compliance. This prospect reinforces the critical role of systematic market monitoring in financial regulation.

To address this challenge, the remainder of the article presents a natural language processing tool designed to automate the summarization of legal arguments in disputes related to means of payment. The tool, based on a fine-tuned multilingual translation model (mBART-50), ingests raw TDLC case documents, and

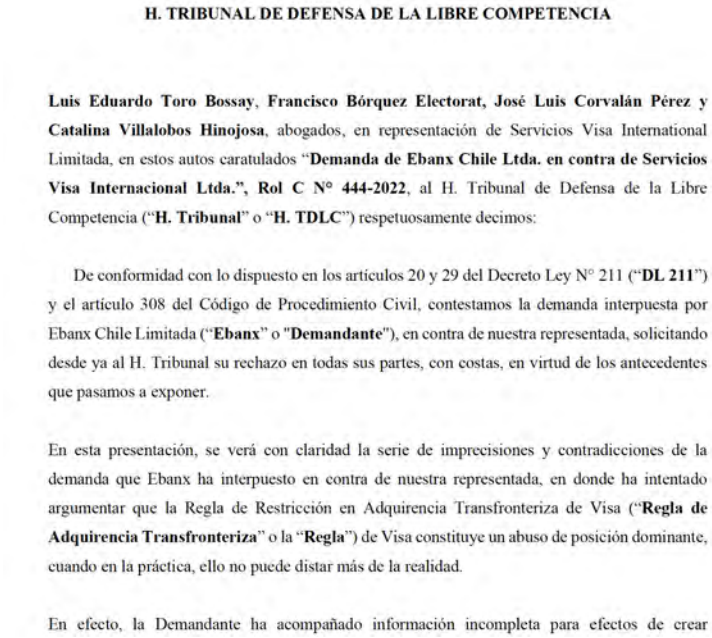
generates concise, one-page summaries that highlight key claims, involved parties, and procedural context, with the primary objective to streamline regulatory review processes, reduce manual workload, and strengthen the Central Bank's capacity to proactively identify market stress points and emerging regulatory needs.

2. Data

The dataset encompasses 19 cases related to payments brought before the TDLC, covering the period from 2016 to 2024. Each case includes a minimum of two documents (typically a claim and counterclaim) and an average of four, each ranging from 40 to 60 pages in length. The dataset covers both contentious cases—involving adversarial disputes and active judicial intervention—and non-contentious cases, which consist of petitions submitted by individual entities where judicial review occurs only as mandated by law.

The complexity of each document varies considerably, shaped primarily by the number and type of entities involved, as well as the depth and intricacy of their claims, responses, and counterclaims. These documents often contain dense legal arguments, detailed business model descriptions, economic justifications, and references to regulatory frameworks, with considerable variation in scope and emphasis depending on the specifics of each case (Graph 2.1).

Original document example (extract) Graph 2.1



H. TRIBUNAL DE DEFENSA DE LA LIBRE COMPETENCIA

Luis Eduardo Toro Bossay, **Francisco Bórquez Electorat, José Luis Corvalán Pérez y Catalina Villalobos Hinojosa**, abogados, en representación de Servicios Visa International Limitada, en estos autos caratulados "**Demanda de Ebanx Chile Ltda. en contra de Servicios Visa Internacional Ltda.", Rol C N° 444-2022**, al H. Tribunal de Defensa de la Libre Competencia ("**H. Tribunal**" o "**H. TDLC**") respetuosamente decimos:

De conformidad con lo dispuesto en los artículos 20 y 29 del Decreto Ley N° 211 ("**DL 211**") y el artículo 308 del Código de Procedimiento Civil, contestamos la demanda interpuesta por Ebanx Chile Limitada ("**Ebanx**" o "**Demandante**"), en contra de nuestra representada, solicitando desde ya al H. Tribunal su rechazo en todas sus partes, con costas, en virtud de los antecedentes que pasamos a exponer.

En esta presentación, se verá con claridad la serie de imprecisiones y contradicciones de la demanda que Ebanx ha interpuesto en contra de nuestra representada, en donde ha intentado argumentar que la Regla de Restricción en Adquirencia Transfronteriza de Visa ("**Regla de Adquirencia Transfronteriza**" o la "**Regla**") de Visa constituye un abuso de posición dominante, cuando en la práctica, ello no puede distar más de la realidad.

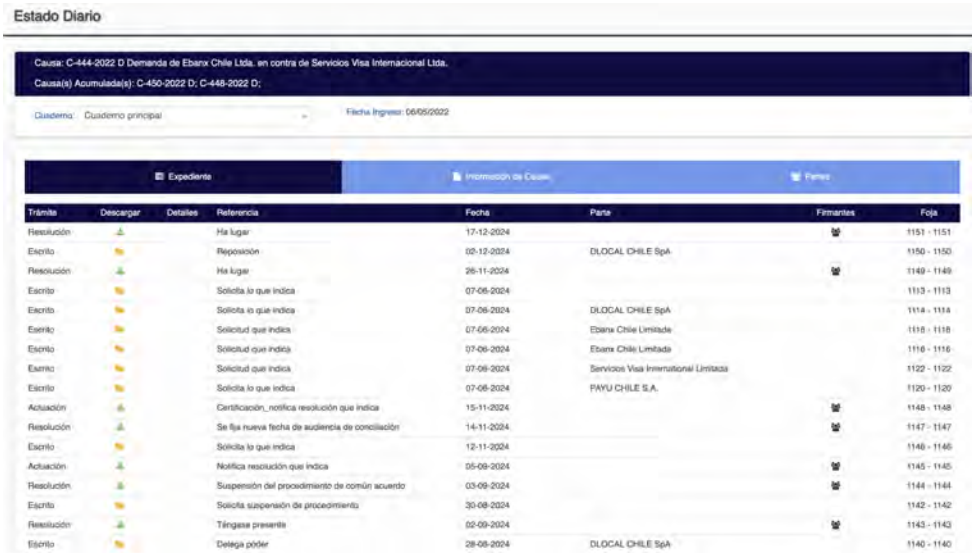
En efecto, la Demandante ha acompañado información incompleta para efectos de crear

Source: TDLC

The dataset, while substantial, presents some inherent limitations. First, since the analysis focuses solely on cases publicly available through the TDLC platform, it may not capture the full spectrum of disputes within the payments ecosystem. For

Daily status platform example

Graph 2.3



Source: TDLC

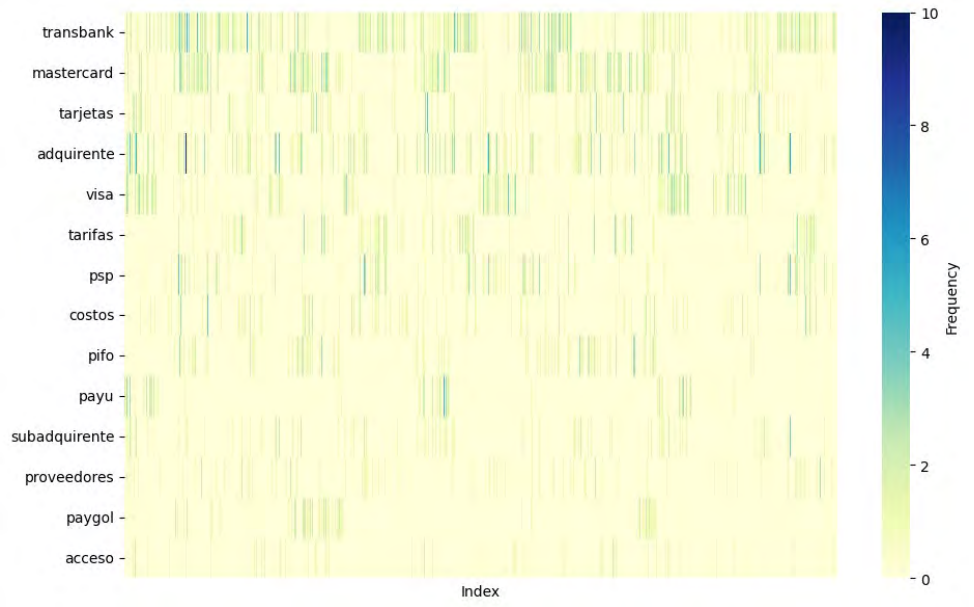
example, over the past decade, most cases have involved interchange fees and acquirers' costs (Graph 2.2).

Second, the diversity in document formats—ranging from scanned PDFs to structured digital filings—introduces variability in data quality and processing reliability (Graph 2.3). To mitigate the risk of overfitting on such a small and heterogeneous dataset, we prioritized model selection over extensive fine-tuning, relying on a fully pre-trained multilingual model capable of handling varied input formats and generalizing across cases.

The dataset remains dynamic and continues to evolve, with new documents frequently uploaded to the TDLC platform as long as cases remain open. These ongoing updates reflect the need for continuous monitoring to remain aligned with real-time judicial developments.

Top TF-IDF Words Across Paragraphs

Graph 2.2



Source: Authors' own elaboration based on selected cases from the TDLC

3. Methodology

Data Preprocessing

Each PDF document undergoes a structured preprocessing pipeline aimed at creating high-quality, informative units suitable for supervised fine-tuning. Initially, the text is extracted and segmented into individual paragraphs based on structural cues such as spacing, indentation, and formatting patterns. Subsequently, shorter paragraphs are merged into coherent textual chunks with a minimum length of 600 characters to ensure sufficient content for a stand-alone summary[3]. Overall, the final dataset contains 3,524 chunk-summary pairs, with an average of 50 words each.

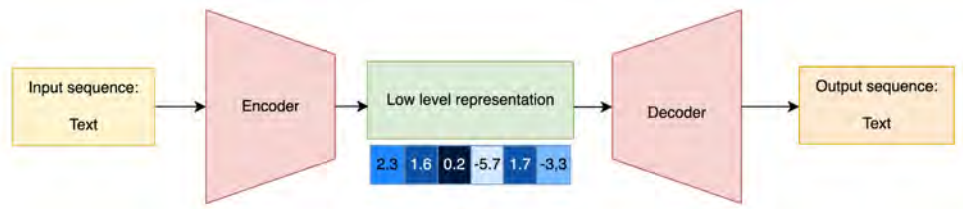
This segmentation ensures that each resulting chunk is contextually meaningful and sufficiently informative, optimizing the process for manual summarization. By creating adequately sized textual units, annotators could more reliably produce high-quality summaries, improving the overall relevance of each chunk-summary pair used for training. The granularity achieved through this preprocessing facilitates detailed scrutiny, enabling downstream models to effectively focus only on critical arguments, content and identify nuanced differences between perspectives.

Model Selection

A multilingual transformer architecture, specifically Meta's mbart50-large model [4], was selected to perform fine-tuned summarization of legal texts in Spanish. Built on the BART architecture, this sequence-to-sequence (seq2seq) Transformer model employs an encoder-decoder framework originally optimized for machine translation tasks (Graph 3.1).

Graph 3.1

Encoder-Decoder Model



Source: Authors' own elaboration

The decision to leverage a translation-oriented model, rather than one trained exclusively for summarization, was based on the structural similarity between the two tasks: both involve extracting semantic content from a source text and generating

[3] Different minimum length thresholds were tested by randomly selecting text chunks ranging from 300 to 1,500 characters. Human annotators found that a minimum of 600 characters struck a balance between context and content.

coherent, linguistically accurate output. The model's multilingual capacity—achieved through pre-training on a broad corpus of languages including Spanish—offers robust linguistic representations crucial for accurately capturing the legal and syntactic complexity of legal documents.

Pre-training for this model involved a denoising autoencoding objective, where the system learns to reconstruct original sentences from corrupted inputs (e.g., masked tokens or shuffled segments). This objective encourages the acquisition of deep contextual embeddings and an understanding of syntax, which is particularly valuable for summarizing dense, domain-specific language. Importantly, this strategy reduces the amount of annotated data required for effective fine-tuning—a critical advantage given the limited size of the training dataset in this application.

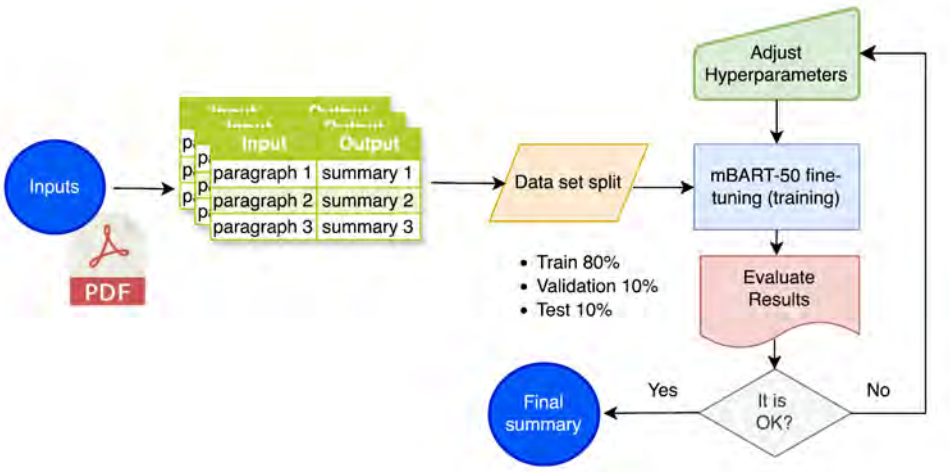
Ultimately, the model was chosen for its strong performance in multilingual contexts, its ability to generalize from small datasets, and its effectiveness in capturing the linguistic and semantic features of legal Spanish. Fine-tuning on a curated set of legal text chunks and corresponding expert-generated summaries allows us to exploit these strengths for accurate and context-aware summarization.

Fine-Tuning Strategy

During fine-tuning, the model receives the segmented legal text chunks produced during preprocessing and is trained to generate concise summaries that closely mirror human-generated references (Graph 3.2). The dataset was split into training (80%), validation (10%), and test (10%) subsets. Training was conducted over 8 epochs using a learning rate of 2e-5 and a batch size of 4.

Approximately 46.7% of the model's weights were updated during training, while the rest were kept frozen to stabilize training [5]. This selective fine-tuning approach mitigates overfitting and preserves the robust language representations learned during pre-training. It is particularly effective given the limited volume of annotated data, enabling the model to adapt to task-specific nuances while maintaining general linguistic accuracy.

Implementation Graph 3.2



Source: Authors' own elaboration

Evaluation

Model performance was quantitatively evaluated using ROUGE-1 and ROUGE-L metrics. The ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation) is widely recognized as an automated evaluation tool for summarization tasks. Specifically, ROUGE-1 evaluates performance by measuring unigram overlap, reflecting the inclusion and relevance of key terms and essential concepts. ROUGE-L, on the other hand, assesses the structural coherence and syntactic similarity by identifying the longest common subsequence between generated and reference summaries, emphasizing readability and contextual coherence.

Metrics such as ROUGE-2 and ROUGE-SU were omitted due to their limited incremental value over ROUGE-1 and ROUGE-L in capturing conceptual accuracy and readability, particularly given the lower relevance of exact phrasing and skip-bigram matches.

Loss metrics recorded during training and validation phases provided complementary insights into the model's learning behaviour, aiding in the early detection and mitigation of overfitting or underfitting and supporting optimal fine-tuning outcomes.

4. Results

The best-performing model checkpoint was obtained at epoch 4, where training loss reached its lowest point and validation loss remained stable (Graph 4.1), indicating good generalization and minimal overfitting. At this stage, the model achieved a ROUGE-1 score of approximately 35% and a ROUGE-L score of 25% on the validation set (Graph 4.2). These results suggest that the model was able to reliably capture key terms and preserve a moderate degree of structural coherence in its summaries compared to the original reference, despite the complexity and length of the source legal documents.

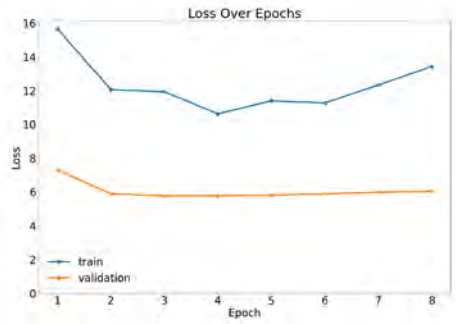
While a ROUGE-1 score of 35% may appear modest in isolation, it aligns with the range of results reported in prior legal summarization studies, as surveyed by Steffes et al. [6]. For example, Polsley et al.'s CaseSummarizer achieved a ROUGE-1 score of approximately 19%, while Zhong et al.'s iterative masking algorithm reached around 27%. More advanced methods, such as those by Furniturewala et al. and Yoon et al., report higher ROUGE-1 scores (≈49% and ≈70%, respectively), though these typically rely on extensive training data. As Steffes et al. emphasize, ROUGE scores—while widely adopted—correlate only partially with human judgment in legal domains, where precision, structure, and argumentative clarity often outweigh simple lexical overlap. Within this context, our results reflect a competitive performance given the dataset size and domain complexity.

This trade-off between generality and domain precision is further illustrated in Graph 4.3, which compares summaries generated by ChatGPT (GPT-4o) and our fine-

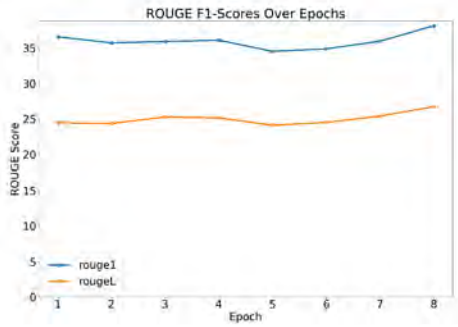
tuned model for a representative legal text chunk. While ChatGPT produces a fluent summary, it introduces elaborations and paraphrases that risk blurring legal definitions and market distinctions central to the case. In contrast, the fine-tuned model offers a more targeted, compressed rendering that mirrors the legal framing of the original text—demonstrating its alignment with the structural and argumentative style of TDLC proceedings. This qualitative difference underscores the value of fine-tuning over zero-shot prompting for tasks that require legal precision and domain adherence.

Training and validation metrics — Graph 4

4.1 Loss over epochs



4.2 Validation ROUGE scores (F1)



To operationalize the model during inference, we adopt a two-stage summarization process that mirrors the structure of the training pipeline. In the first stage, individual summaries are generated for each preprocessed text chunk, allowing the model to handle long documents by breaking them into manageable, semantically coherent segments. In the second stage, these chunk-level summaries are concatenated and passed once more through the model to generate a final, global summary. In addition, a set of descriptive statistics (such as entity count and bigram frequency) is computed to provide contextual information about the case. This hierarchical approach supports coherence at both the local and document-wide levels, ensuring that key legal arguments are preserved while reducing redundancy across sections (Graph 4.4).

Chunk summary example — Graph 4.3

Original Chunk	22.- The advantage of cards associated with international brands -such as VISA- lies in the fact that they are accepted by the entire universe of merchants subscribed to the brand, regardless of the issuer, by virtue of the rule known as \Honor All Cards." This contrasts with cards tied to a proprietary brand, which do not enjoy universal acceptance and whose use is restricted exclusively to merchants affiliated with that proprietary brand.⁵⁰ Consequently, the relevant market affected by the anticompetitive conduct carried out or entered into by VISA, Transbank, and Klap comprises the market of transactions conducted at electronic merchants using credit, debit, and prepaid cards of universal acceptance in Chile, since these types of cards enable cardholders to connect with the e-commerce platforms associated with dlocal, such that proprietary-

	brand cards are not capable of serving as substitutes in this type of transaction.
ChatGPT - GPT 4o	International brand cards like VISA are accepted at all merchants affiliated with that brand, regardless of which bank issues them, thanks to the "Honor All Cards" rule. In contrast, proprietary brand cards can only be used at merchants that accept that specific brand. This means that the market affected by anticompetitive practices by VISA, Transbank, and Klap in Chile is the market for transactions at electronic merchants using universally accepted credit, debit, and prepaid cards. These cards allow users to connect to platforms like dLocal, while proprietary brand cards cannot serve as substitutes in these types of transactions.
Fine-tuned model	The market affected by anticompetitive practices by Visa, Transbank, and Klap is the market for transactions at electronic merchants using universally accepted cards, which enable cardholders to connect to the commerce platforms associated with dlocal.

* Translated from Spanish

Final one-pager summary example Graph 4.4



Resumen

Demanda

Ebanx Chile demandó a Visa International por supuestas prácticas anticompetitivas, acusándola de abuso de posición dominante en el mercado chileno de medios de pago. Ebanx, que facilita pagos internacionales sin tarjetas internacionales, alega que Visa busca excluirlo del mercado mediante reglas restrictivas que limitan su acceso a la red de tarjetahabientes. Esto afecta gravemente su modelo de negocio, encarece las transacciones y restringe opciones para consumidores y comercios.

Respuesta a la demanda

Visa International responde a la demanda de Ebanx Chile, argumentando que la Regla de Adquirencia Transfronteriza no constituye abuso de posición dominante, sino que asegura la seguridad y funcionalidad de la cadena de pagos, cumpliendo con normativas regulatorias, negando competir directamente con Ebanx, señalando que operan en mercados distintos.

Además refuta las acusaciones de exclusión de mercado, alegando que Ebanx ha presentado información incompleta, incumplido contratos de licenciamiento y distorsionado hechos. Sumado al hecho que su regla se aplica uniformemente a nivel global y garantiza prevención de fraude y cumplimiento normativo.

Participantes

- **Ebanx Chile**: Empresa demandante, agente de recaudo internacional.
- **Visa International**: Empresa demandada, acusada de prácticas anticompetitivas.
- **Transbank S.A., iSwitch S.A., Klap**: Adquirentes locales implicados en la implementación de las reglas de Visa.
- **PayU**: Competidor de Ebanx en pagos internacionales.

Highlights

Demanda

tarjetas visa

acceso red

red tarjetahabientes

Contra Demanda

proveedores servicios

comercios extranjeros

servicios psp

5. Concluding remarks

This tool demonstrates the feasibility of developing a specialized summarization model tailored to legal documents within a narrowly defined domain. By generating concise, targeted summaries of complex legal texts, the model offers practical support for financial regulation monitoring. In particular, it facilitates the early identification of emerging conflicts in the payments industry and can assist in shaping regulatory frameworks that both reflect market dynamics and promote orderly payment system functioning.

Recognizing the inherent subjectivity in both summarization and regulatory interpretation, the model is designed to incorporate targeted preferences and thematic priorities. Through the use of expert-generated training summaries, domain-specific fine-tuning, and adjustable inference parameters, the system allows regulatory departments to tailor outputs to their analytical needs and strategic goals. As such, the model delivers not only technical utility but also institutional adaptability.

Looking ahead, future work will extend the model's application to market inquiries submitted to the Central Bank, which often involve non-public information and require strict confidentiality safeguards. Because the tool can be securely deployed within the BCCh's internal infrastructure, it eliminates reliance on external data processing and supports full compliance with internal data protection protocols. This secure, on-premises deployment capability makes the model especially well-

suited for analysing sensitive regulatory materials and serves as a robust foundation for broader applications across internal textual datasets.

6. References

[1] Banco Central de Chile. Informe de Sistemas de Pago (ISiP), 2024

[2] Corte Suprema de Chile. Sentencias relevantes sobre adquirencia y sistemas de pago digital. 2024.

[3] Comisión para el Mercado Financiero. Normativa sobre el Sistema de Finanzas Abiertas, 2024.

[4] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," arXiv preprint arXiv:2008.00401, 2020.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," arXiv preprint arXiv:1910.03771, 2019.

[6] B. Steffes, P. Rataj, L. Burger, and L. Roth, "On evaluating legal summaries with ROUGE," in *Proc. 19th Int. Conf. Artif. Intell. Law (ICAIL)*, Braga, Portugal, Jun. 2023, pp. 457–461.

AI-Driven Legal Summarization for Regulation Monitoring

Central Bank of Chile

4th IFC Workshop on Data Science in Central Banking

Authors: Isidora Díaz J. & Diana López A.

Wednesday 19th February, 2025

Disclaimer

The views are those of the authors and do not necessarily reflect the views of the Central Bank of Chile or its board members.

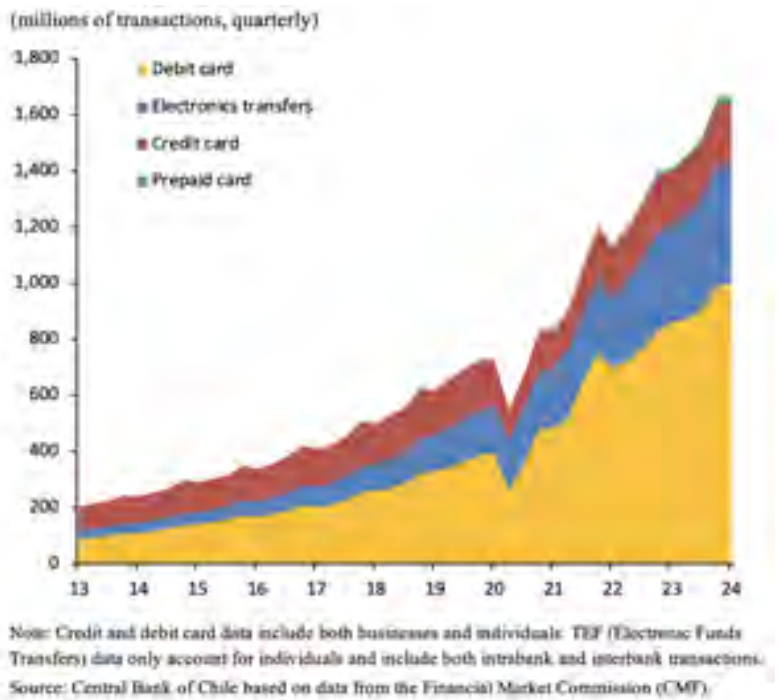
Agenda

1. Context and motivation
2. Data
3. Model and implementation details
4. Results
5. Final comments

Context

- The Chilean payment systems industry has experienced significant growth in recent years, particularly in digital payments, which continue to show rapid increases in transaction volume, with new participants entering areas such as payment issuance and acquiring [1]

Figure: Digital Payment Transactions [2]



Context

- The consolidation of this industry's development has created frictions among market participants, particularly concerning fees and the enablement of new business models [1].
- This has increasingly led some entities to seek resolution from institutions like the Tribunal for the Defense of Free Competition (TDLC) or the National Economic Prosecutor's Office (FNE), which provide legal solutions to these conflicts.
- These dynamics pose regulatory challenges for the Central Bank of Chile (CBCh), as one of its objectives under the Organic Constitutional Law (LOC) is to "ensure the proper functioning of domestic and international payments."
- Consequently, it is **essential to understand and monitor the cases brought to the TDLC by various stakeholders in the payment systems industry.**

Motivation

- **Problem:**
  - Legal documents are lengthy, complex, and use terminology that is not accessible to non-experts, hence the need to have a comprehensive and timely overview of each case.
  - Relevant information (issues, entities, and context) is often buried within extensive legal texts.
- **What we do:**
  1. Customize a Large Language Model (LLM) to process original documents.
  2. Generate one-page summaries highlighting key content in an accessible, high-level manner.

Data

- Includes TDLC cases from 2016 to 2022 related to payment systems.
- Covers both contentious and non-contentious cases (19 in total, avg. 4 documents per case, 40-60 pages per document):
  - Contentious: Judicial disputes with judge intervention.
  - Non-Contentious: Petitioner-only cases; judge intervenes if required by law.
- Analyzed documents that focused on main claims and responses (counterclaims), which vary based on the number of entities involved.
- Each document is then split into chunks, with approximately 150 chunks per document

TDLC legal cases

- **Case Files:**
  - May include numerous documents per case, in diverse formats with extended file lengths.
  - New documents are uploaded regularly (as long as the case remains open).

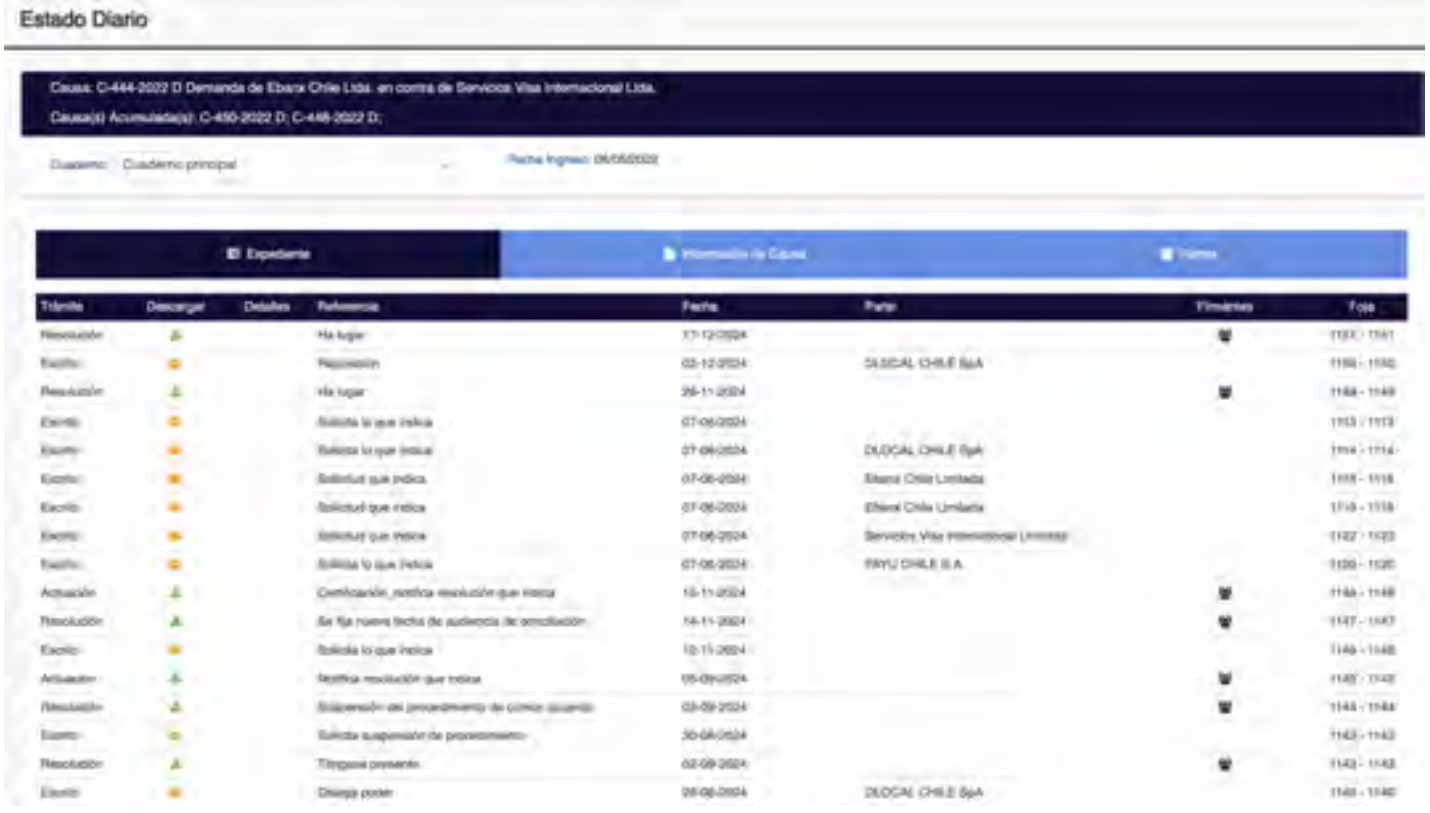
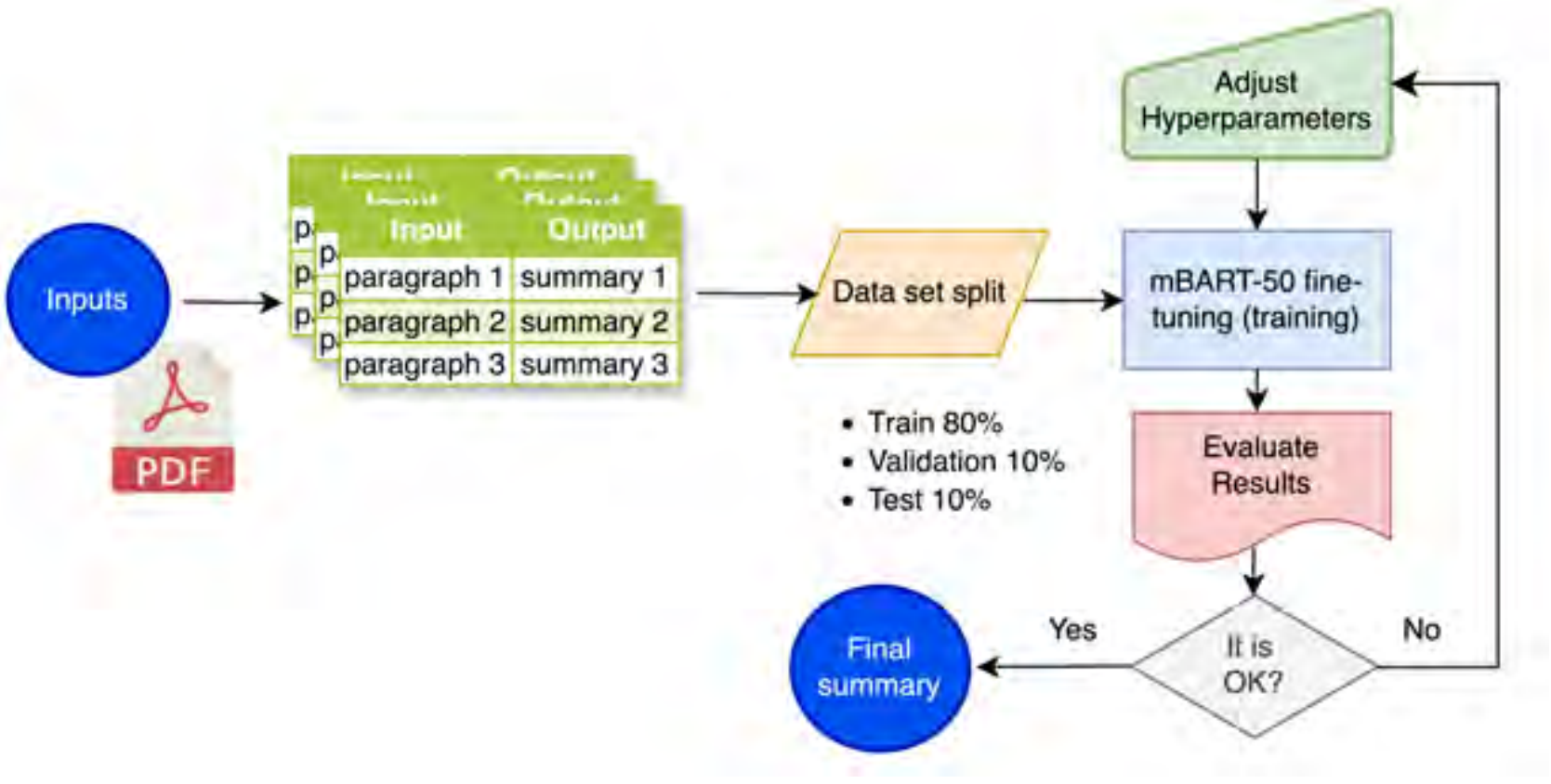


Figure: Example of TDLC case and platform

Implementation



- Fine-tuning adapts a pre-trained language model to a specific task (such as summarization) using a smaller, task-specific dataset.
- It updates the model's weights to optimize performance on the new task while leveraging knowledge from pre-training [3].

Implementation details

- Multilingual transformer model: `facebook/mbart50-large` [4]
- Input: Legal text chunks.
- Target: Human-made summaries.
- Dataset split: 80% training , 10% validation, 10% test
- Training Configuration:
  - Learning rate: 2e-5
  - Batch size: 4
  - Epochs: 8
  - Trainable parameters: 46.7%
- Evaluation: loss & ROUGE
- During inference, summaries are generated in chunks to simplify large documents into manageable extracts. A global summary is then produced in 2 stages, by aggregating these chunks and prompting the model again.

Results: Metrics

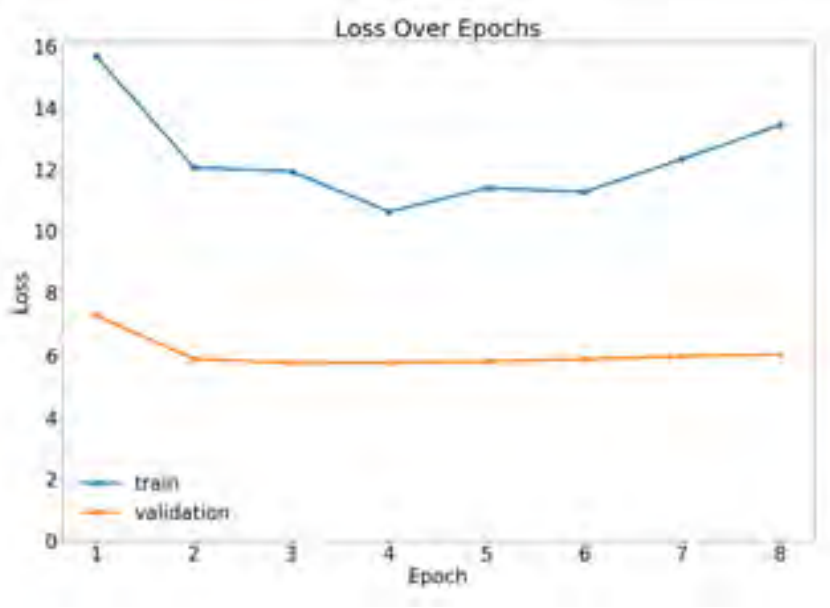


Figure: Loss over Epochs

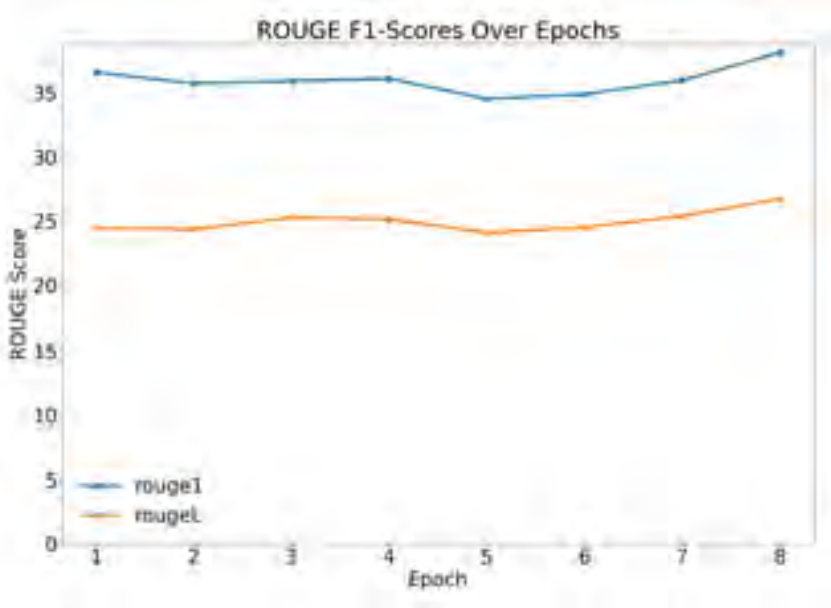


Figure: Validation ROUGE Scores (F1)

Results: Example (1/2)

Original chunk

22.- The advantage of cards associated with international brands -such as VISA- lies in the fact that they are accepted by the entire universe of merchants subscribed to the brand, regardless of the issuer, by virtue of the rule known as \Honor All Cards." This contrasts with cards tied to a proprietary brand, which do not enjoy universal acceptance and whose use is restricted exclusively to merchants affiliated with that proprietary brand.⁵⁰ Consequently, the relevant market affected by the anticompetitive conduct carried out or entered into by VISA, Transbank, and Klap comprises the market of transactions conducted at electronic merchants using credit, debit, and prepaid cards of universal acceptance in Chile, since these types of cards enable cardholders to connect with the e-commerce platforms associated with dlocal, such that proprietary-brand cards are not capable of serving as substitutes in this type of transaction.

Results: Example (1/2)

ChatGPT - GPT 4o

International brand cards like VISA are accepted at all merchants affiliated with that brand, regardless of which bank issues them, thanks to the "Honor All Cards" rule. In contrast, proprietary brand cards can only be used at merchants that accept that specific brand. This means that the market affected by anticompetitive practices by VISA, Transbank, and Klap in Chile is the market for transactions at electronic merchants using universally accepted credit, debit, and prepaid cards. These cards allow users to connect to platforms like dLocal, while proprietary brand cards cannot serve as substitutes in these types of transactions.

Fine-tuned model

The market affected by anticompetitive practices by Visa, Transbank, and Klap is the market for transactions at electronic merchants using universally accepted cards, which enable cardholders to connect to the commerce platforms associated with dlocal.

Results: Example



H. TRIBUNAL DE DEFENSA DE LA LIBRE COMPETENCIA

Luis Eduardo Toro Bossay, **Francisco Bórquez Electorat**, **José Luis Corvalán Pérez** y **Catalina Villalobos Hinojosa**, abogados, en representación de Servicios Visa International Limitada, en estos autos caratulados "**Demanda de Ebanx Chile Ltda. en contra de Servicios Visa Internacional Ltda.**", **Rol C N° 444-2022**, al H. Tribunal de Defensa de la Libre Competencia ("**H. Tribunal**" o "**H. TDLC**") respetuosamente decimos:

De conformidad con lo dispuesto en los artículos 20 y 29 del Decreto Ley N° 211 ("**DL 211**") y el artículo 308 del Código de Procedimiento Civil, contestamos la demanda interpuesta por Ebanx Chile Limitada ("**Ebanx**" o "**Demandante**"), en contra de nuestra representada, solicitando desde ya al H. Tribunal su rechazo en todas sus partes, con costas, en virtud de los antecedentes que pasamos a exponer.

En esta presentación, se verá con claridad la serie de imprecisiones y contradicciones de la demanda que Ebanx ha interpuesto en contra de nuestra representada, en donde ha intentado argumentar que la Regla de Restricción en Adquirencia Transfronteriza de Visa ("**Regla de Adquirencia Transfronteriza**" o la "**Regla**") de Visa constituye un abuso de posición dominante, cuando en la práctica, ello no puede distar más de la realidad.

En efecto, la Demandante ha acompañado información incompleta para efectos de crear segmentaciones artificiosas del mercado relevante y confusiones acerca del verdadero efecto de

Figure: Original document



Resumen

Demanda

Ebanx Chile demandó a Visa International por supuestas prácticas anticompetitivas, acusándola de abuso de posición dominante en el mercado chileno de medios de pago. Ebanx, que facilita pagos internacionales sin tarjetas internacionales, alega que Visa busca excluirlo del mercado mediante reglas restrictivas que limitan su acceso a la red de tarjetahabientes. Esto afecta gravemente su modelo de negocio, encarece las transacciones y restringe opciones para consumidores y comercios.

Respuesta a la demanda

Visa Internacional responde a la demanda de Ebanx Chile, argumentando que la Regla de Adquirencia Transfronteriza no constituye abuso de posición dominante, sino que asegura la seguridad y funcionalidad de la cadena de pagos, cumpliendo con normativas regulatorias, negando competir directamente con Ebanx, señalando que operan en mercados distintos.

Además refuta las acusaciones de exclusión de mercado, alegando que Ebanx ha presentado información incompleta, incumplido contratos de licenciamiento y distorsionado hechos. Sumado al hecho que su regla se aplica uniformemente a nivel global y garantiza prevención de fraude y cumplimiento normativo.

Participantes

- **Ebanx Chile**: Empresa demandante, agente de procesamiento internacional.
- **Visa International**: Empresa demandada, acusada de prácticas anticompetitivas.
- **Transbank S.A., iSwitch S.A., Klap**: Adquirentes locales implicados en la implementación de las reglas de Visa.
- **PayU**: Competidor de Ebanx en pagos internacionales.

Highlights

Demanda

Contra Demanda

Figure: Generated summary (one-pager)

Final comments

- This application makes it possible to create a model capable of producing specialized summaries on a very specific topic, thus helping our unit identify new business models in the payment industry and develop regulations that align with industry growth.
- There is the potential to use this tool for confidential documents, as it can ensure that the information never leaves CBCh's network and servers.
- Finally, this application is flexible enough to be tailored to the preferences and emphasis of a regulatory department manager, considering the subjective nature involved in training a model for this use case.

Bibliography

[1] C. B. of Chile, "Payment systems report," August 2023. https://www.bcentral.cl/documents/33528/4387742/Informe-de-Sistemas-de-Pago-agosto-2023.pdf/ec39444d-4623-05ee-aa4c-263cbcd72b11.

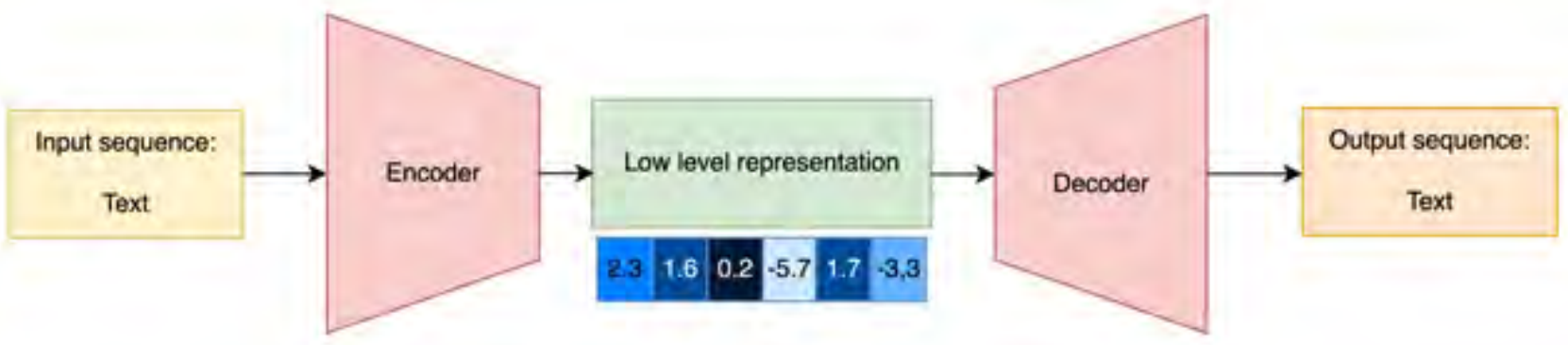
[2] C. B. of Chile, "Payment systems report," August 2024. https://www.bcentral.cl/documents/33528/6165052/ISIP_2024.pdf/880f1337-7904-2386-c815-077cfcb7896b.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[4] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

Appendix: Model mBART-50

- mBART-50 is built on BART's sequence-to-sequence (seq2seq) Transformer framework, including an encoder and a decoder built for translation.
- It is pre-trained to reconstruct sentences from noisy inputs (masked or shuffled text), enabling it to learn deep linguistic representations.



Back

Appendix: Model evaluation

- **Loss function**: Cross-entropy loss, matching predicted and actual token sequences.
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation): used to evaluate the quality of generated summaries, by measuring n-gram overlap between generated and reference summaries.
- However, note that ROUGE ignores semantic meaning and focuses only on surface-level word overlap.
- Human feedback is still valuable to improve the model, as summary quality is subjective and context-dependent.

Back

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

A novel multi-step prompt approach for LLM-based Q&As on banking supervisory regulations[1]

Daniele Licari, Canio Benedetto, Daniele Bovi,
Alessandro De Gregorio and Marco De Leonardis,
Bank of Italy

Praveen Bushipaka and Tommaso Cucinotta,
Sant'Anna School of Advanced Studies

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

A novel multi-step prompt approach for LLM-based Q&As on banking supervisory regulations

Daniele Licari*, Canio Benedetto*, Daniele Bovi*, Praveen Bushipaka**, Alessandro De Gregorio*, Marco De Leonardis* and Tommaso Cucinotta**

* Bank of Italy
** Sant'Anna School of Advanced Studies

Abstract

This paper investigates the use of large language models (LLMs) in analysing and answering questions related to banking supervisory regulations. We propose a multi-step prompt approach that enriches the context provided to the LLM with relevant articles from the Capital Requirements Regulation (CRR). We compare our method against standard "zero-shot" prompting, where the LLM answers are solely based on its pre-trained knowledge, and a standard "few-shot" prompting, where the LLM is given every time only a limited number of examples of questions and answers to draw on. To assess the quality of the generated answers, we also build an "LLM evaluator" which, for each question, compares the correctness and completeness of the answer provided by our approach and the two standard methods with the available official EBA's answer used as benchmark. Our findings on inquiries concerning Liquidity Risk rules indicate that our multi-step approach significantly improves the quality of LLM-generated answers, offering the analyst a valuable starting point to formulate appropriate answers to particularly complex questions.

Keywords: Regulatory Q&A, Banking Supervisory Regulation, Artificial Intelligence, GenAI, GPT-4o, RAG, LLM Evaluator

JEL Classification: C45, C63, G21, G28

1. Introduction[1]

The advent of generative AI (GenAI), and specifically of large language models (LLMs), offers significant opportunities to implement innovative solutions to improve the efficiency and effectiveness of various domains of activities (Wu et al., 2023; Lai et al., 2023; Biancotti and Camassa, 2023; Horton, 2023; Homoki and Ződi, 2024).

One of the most promising applications is the support to the navigation and analysis of complex regulatory documents (Wiratunga et al., 2024; Louis et al., 2023; Zhang et al., 2023; Abdallah et al., 2023), which can be particularly valuable for compliance officers, legal teams, and other professionals, who need to have a clear and timely understanding of the regulations and the consequent obligations. This is, for example, the case of the EU harmonised banking supervisory reporting obligations, i.e. the set of rules defined by the European Banking Authority (EBA) about the information that the banking system is obliged to produce and transmit to the national competent authorities (NCAs). It is an articulated set of rules including the Capital Requirements Regulation[2] (CRR) and several delegated and implementing acts, technical standards, guidelines, and recommendations. In our work, however, for the sake of simplicity, we will only consider the CRR.

The complexity of such regulatory documents, with their dense network of cross-referenced texts and specialised content, is a challenge also for the same NCAs whose experts always necessitate a careful and deep analysis across the various acts to retrieve the needed information (Prenio, 2024). While LLMs offer advantages for this purpose, they also pose risks like bias and inaccuracies (Huang et al., 2023). Therefore, it is essential to establish strong verification procedures and retain human supervision to limit these risks.

This study introduces a novel methodology to automate and expedite the "question & answer" (Q&A) process in regulatory compliance, leveraging advanced LLMs to provide accurate, complete and timely responses to inquiries[3] usually presented by the banking sector about the EBA's supervisory regulations. Our multi-step approach aligns with Retrieval-Augmented Generation principles (Lewis et al., 2020) through the implementation of mechanisms like explicit extraction of CRR references, implicit reference analysis, and a dedicated cross-encoder for precise regulatory text retrieval. Our work finds application within the domain of EBA

[1] The authors would like to express their sincere gratitude to Manuela Biagi, Vincenzo Capone, Francesca Monacelli, Roberto Sabbatini and Fabio Zambuto (Bank of Italy) for their insightful comments and stimulating discussions which greatly enhanced the clarity and focus of our work. A special thanks goes to Pamela Maggiori (Bank of Italy), whose careful analysis has significantly contributed to the development of the LLM evaluator.

The authors would also like to thank the anonymous reviewers of the CLiC-it 2024 Conference and the participants of the 4th IFC Workshop on Data Science in Central Banking for their invaluable suggestions and constructive feedback.
The views expressed in the paper are those of the authors and do not necessarily reflect those of the Bank of Italy.

[2] Regulation (EU) 2013/575: eur-lex.europa.eu/legal-content/en/ALL/?uri=celex%3A32013R0575

[3] For the purpose of this paper, we equally use the words "question" and "inquiry" to mean a request for clarification on the supervisory regulations presented by a domain expert.

regulatory rules, that are characterised by a large and complex set of interrelated documents, although in this first research we will limit the focus the CRR.

Compared to standard inquiries of LLM engines, our approach improves the suitability of the response generation for the regulatory compliance context where precise and comprehensive answers are crucial, but the complex articulation of the legal acts makes the cross analysis very complicated and time consuming for the human expert. The business case is therefore both challenging and rewarding.

Of all the possible topics covered in the CRR, in this work, we focus on inquiries related to Liquidity Risk as a first use case to evaluate the potential benefit of enriched context for an accurate response generation. The main reason for this choice is that this topic is regulated by a relatively limited number of documents, predominantly within the CRR. This manageable scope facilitated the initial development and evaluation of our methodology, especially considering the time-intensive nature of pre-processing regulatory information.

We retrieve actual questions presented by banks and official EBA answers on the topic of Liquidity Risk from the "EBA Q&As database"[4] as the foundation for developing a novel system capable of generating automated accurate and complete (i.e. contextually relevant) responses to inquiries. It is important to emphasise that the questions contained in the "EBA Q&A database" are characterized to be all particularly complex and challenging by nature, as they have proven to be difficult issues to answer for the industry experts who submitted them, requiring an authentic interpretation of the regulations by EBA specialists. This means that we aim to build an LLM-based Q&A tool able to provide the results of high-quality analyses of supervisory regulations. That said, given the relevance in the field of legal acts interpretation of producing a correct and complete answer, in formal and practical terms we consider the output of the tool as a draft text for refinement for the experts and not a replacement of their analysis. The ambition of course is to obtain a draft text that needs as little adjustment as possible. Lastly, to assess the quality of the results obtained with the tool, we leverage recent studies on the potential of LLMs for qualitative assessment (Ye et al., 2024; Zheng et al., 2023; Liu et al., 2023; Chan et al., 2023) and, in this work, we also propose the use of an "LLM Evaluator" to automate the validation process.

The structure of this paper is the following. Chapter 2 introduces the methodology and provides a description of the approach adopted; it explains the dataset used and the retrieval techniques employed to identify the relevant regulatory documents necessary to address the questions on Liquidity Risk. Chapter 3 presents the LLM Evaluator and the evaluation criteria. Chapter 4 reports the experimental results and presents the main outcomes of the study. Chapter 5 discusses the challenges as well as the potential areas for future developments.

[4] Stakeholders can submit questions to the EBA on the practical application or implementation of the banking, payment services, AML/CFT and other legislation that falls within the EBA's remit. This includes the associated delegated and implementing acts, RTS, ITS, guidelines and recommendations.

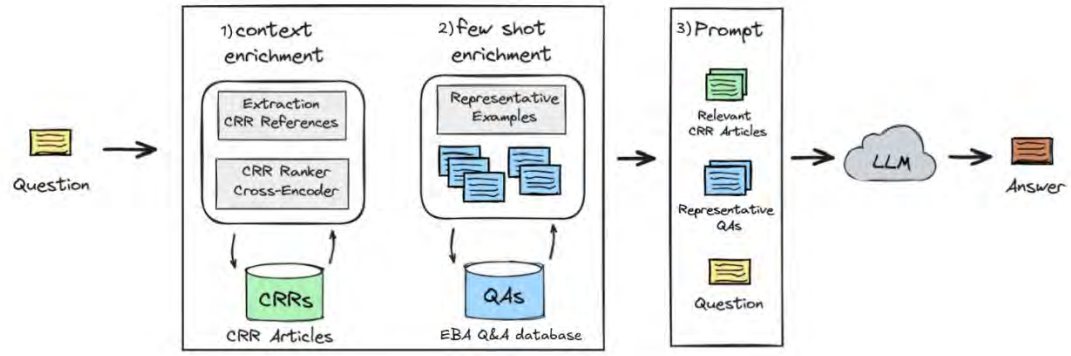
EBA Single Rulebook Q&A: www.eba.europa.eu/single-rule-book-qa

2. Methodology

This research employs a multi-step methodology to construct a comprehensive prompt for the GPT-4 omni 2024-11-20 version (GPT-4o) language model (OpenAI et al., 2024), to appropriately enrich questions on EBA's banking supervisory rules concerning Liquidity Risk, as contained in the EBA Q&A database, in order to obtain an answer resembling the official EBA answer (also contained in the EBA Q&A database) as much as possible.

Our stepwise approach focuses on enriching the context provided with the question in three steps: first, it identifies the relevant CRR references specified in the inquiry; second, it incorporates examples of other Q&As to ensure that the LLM's output format is aligned with the EBA's (e.g. lexicon, style, structure of the answer); third, this enriched context is then leveraged by GPT-4o to generate the answer (details in Figure 1).

Multi-Step Approach for Answer Generation — Figure 1



2.1. Dataset construction

As a preliminary step, we extract from the EBA Q&A database, all the pairs of question and answer submitted between 2013 and 2020 (to consider Q&As referring to a more stable regulatory framework, before the introduction of CRR2). We focused on the following variables: question ID, question, submission date, status, topic, legal act, article, background information, final answer, submission date and status (details in Table 1). Given our specific focus on CRR articles, we excluded variables such as 'COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations' and the related and more granular 'Article/Paragraph'. While 'Subject matter' offers a concise headline for the question-answer, we prioritised the richer semantic context provided by 'question' and 'background information' for our retrieval and answer generation process. Also 'Final publishing date', 'Type of submitter', and 'Answer prepared by' were deemed less relevant for this purpose and were therefore excluded."

EBA Q&As dataset. For this research, we focused on the fields highlighted in yellow

Table 1

VARIABLE NAME	DESCRIPTION
Question ID	The unique identifier for each question.
Topic	The general topic or category under which the question falls.
Subject matter	The specific subject matter of the question.
Legal act	The specific legal act to which the question relates. (e.g., CRR)
Article	The specific article of the legal to which the question relates.
COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations	Other legislation, standards, guidelines or recommendations to which the question relates.
Article/Paragraph	The specific article or paragraph within the above-mentioned
Question	The actual question asked.
Background on the question	Any additional information or context provided by the question submitter.
Final answer	The official answer provided to the question.
Submission date	The date when the question was submitted.
Final publishing date	The date when the final answer to the question was published.
Status	The current status of the question (e.g. Final, rejected, etc.).
Type of submitter	The type of entity that submitted the question (e.g. Credit institution, investment firm, etc.).
Answer prepared by	The entity that prepared the answer to the question.

Secondly, we implemented a two-step filtering process aimed at strengthening the model efficacy: by excluding non-English questions, and by focusing on CRR-related questions.

This resulted in a final dataset of 1,597 CRR-related pairs of question and answer, which was then split into training (50%), validation (10%), and test set (40%) for evaluation.

The distribution of samples for the dataset is summarized in Table 2.

Distribution across training, validation, and test sets for CRR-related Q&A and the subset of only Liquidity Risk Q&A

Table 2

Set	CRR-related Q&A	Liquidity Risk Q&A	(percentage)
Training	798	58	50%
Validation	162	12	10%
Test	637	46	40%
Total	1597	116	100%

2.2. Context Enrichment

The context enrichment process is a three-step approach designed to identify, within the dataset built as indicated in Paragraph 2.1, the most relevant CRR references to provide an appropriate background to formulate the answer to the inquiry. The first step simply involves extracting explicit CRR references, if directly mentioned in the question (content of the variable "Article" in Table 1). The second step leverages on the capabilities of GPT-4o (see "Prompt 1" in Appendix) to analyse the content of the variables "Question" and "Background information" to identify other CRR references that are not explicitly stated in the question. The last step of the process makes use of our "CRR Ranker" model, i.e. a cross-encoder architecture that has been trained to identify and retrieve pertinent references from the CRR in response to specific inquiries. This 3-steps comprehensive approach provides a broader and more accurate context to what is included in the original question, thus improving the possibility for the model to "understand" the question and identify the specific CRR references deemed to be relevant.

2.2.1. CRR Ranker training

With regard to the last step of the context enrichment described in the paragraph above, we employ a specifically trained cross-encoder model (Chen et al., 2024) to identify the most relevant CRR references to be considered. We use a dedicated dataset of "question-article" pairs derived from our EBA Q&A Train Database, from which we exclude questions related to CRR Article 99 being this irrelevant for the purpose of our study. Each data point consists of a question (including the related background information - see Table 1), the associated CRR article, and a binary label indicating relevance.

We constructed the training dataset by selecting positive samples, consisting of pairs of questions and articles where the article is deemed relevant to the

corresponding question, and negative samples, consisting of pairs where the article is unrelated to the question:

- Positive samples comprise question-article pairs where the article is explicitly linked to the question (see variable Article in Table 1). Additionally, we include pairs formed by questions and other CRR references implicitly contained in the text, context information, and EBA's final answer, that we extract using GPT-4o (see "Prompt 1" in Appendix);
- Negative training samples are mined using the BAAI bge-large-en-v1.5 pre-trained language model (Xiao et al., 2023) employing a two-phase process. First, all CRR articles are encoded using the bge-large-en-v1.5 model, and cosine similarity is used to rank them relative to a question; second, a subset of 20 negative examples was randomly chosen from a pre-defined ranking interval[5] (250-300). The choice of 20 negative samples provides a good balance between computational efficiency and the availability of a sufficient number of training data. This approach aimed to balance the representation of relevant and irrelevant information within the training data, ensuring the model learns to select articles that are actually related to the question and discard those that, although semantically similar, are ultimately off-topic (Xuan et al., 2021).

The final dataset comprises 12,533 unique "question-article" pairs with positive and negative labels. This dataset is then split into training (10,179 pairs) and development (2,354 pairs) sets for model fine-tuning to learn robust semantic representations for questions and CRR articles, enabling the model to effectively identify relevant CRR references for enriching the question's context.

We select the BAAI BGE Reranker v2 m3 model (Chen et al., 2024) as the basis for our cross-encoder, owing to its task-specific aptness and its demonstrated superior performance relative to the BGE Reranker Large (Xiao et al., 2023), as reported in Section 4. We adopt the Cross-Entropy Binary Classification loss function, following the approach suggested in the BGE Rerank Git repository (Xiao et al., 2024). To promote stable convergence, we incorporate a warmup schedule (with a number of steps 0.1 × len(train_data) × num_epochs step) that gradually increases the learning rate during the initial phase of training. The entire fine-tuning process was conducted over 4 epochs. We employed an evaluation interval of 800 steps during training and saved the model that achieved the highest F1 score on the development set.

Finally, we evaluate the model's ability to retrieve CRR articles for a given question present in the Q&A Test Database. This evaluation employes recall metrics at various retrieval cutoffs, including recall@5, recall@10, recall@20, and recall@30 (see the results in Section 4).

2.3. Enriching the prompt with examples

To improve the model's understanding of the desired format to be used to formulate the answer, we adopt a few-shot prompting approach (Brown et al., 2020). This

5 Of the overall ~500 articles, we chose the 250-300 range to select negative examples that are not obviously irrelevant but also not highly similar to the question. This "middle ground" provides challenging negative examples that help the model learn subtle differences between relevant and irrelevant content.

involves extracting 5 relevant examples of "question and answer" pairs from the EBA Q&A Train Database concerning other similar issues within the topic of liquidity risk. These examples serve as demonstrations for the tool, showcasing the ideal structure, language style, and level of detail expected in the final answers. Notably, the selection process ensures variety within the chosen topic, meaning the examples cover various aspects to promote a broader understanding of the expected communication standards. Limiting the number of examples to five strikes a balance between providing diverse demonstrations and maintaining cost-efficiency during inference, as the LLM's input token length is limited.

2.4. Improving the prompt to enhance answer generation

To generate accurate and contextually appropriate answers, we employ an advanced prompt engineering strategy leveraging GPT-4o's capabilities as detailed in Figure 1. Our approach integrates three techniques to enhance answer generation within a regulatory compliance framework (see "Prompt 2" in Appendix).

Firstly, we implement a role-based contextualization method that aligns the model to specific regulatory requirements and professional communication standards. Secondly, we use step-by-step reasoning to decompose complex queries into logical, hierarchical components. Finally, we employ few-shot prompting by carefully selecting and integrating examples from our training dataset as a contextual guide (as detailed in Paragraph 2.3).

3. LLM Evaluator

To assess the quality of the answers generated by the tool (following the steps illustrated in Section 2), we developed an LLM Evaluator that compares the answers produced by the tool with the official ones of the EBA "Q&A" database. Recent research highlights that employing an LLM Evaluator offers significant advantages in terms of cost-effectiveness and efficiency compared to the traditional human evaluation method when dealing with large-scale natural language evaluation tasks (Liu, Iter et al., 2023; Dubois et al., 2024; Fu et al., 2023).

The evaluation process uses a scale from 1 to 4, based on two evaluation criteria: correctness and completeness. A generated answer is considered "correct" if its content aligns with the information presented in the official answer. Additionally, a response is deemed "complete" if it incorporates all relevant regulatory references provided in the official answer.

The following scoring rubric outlines the evaluation criteria:

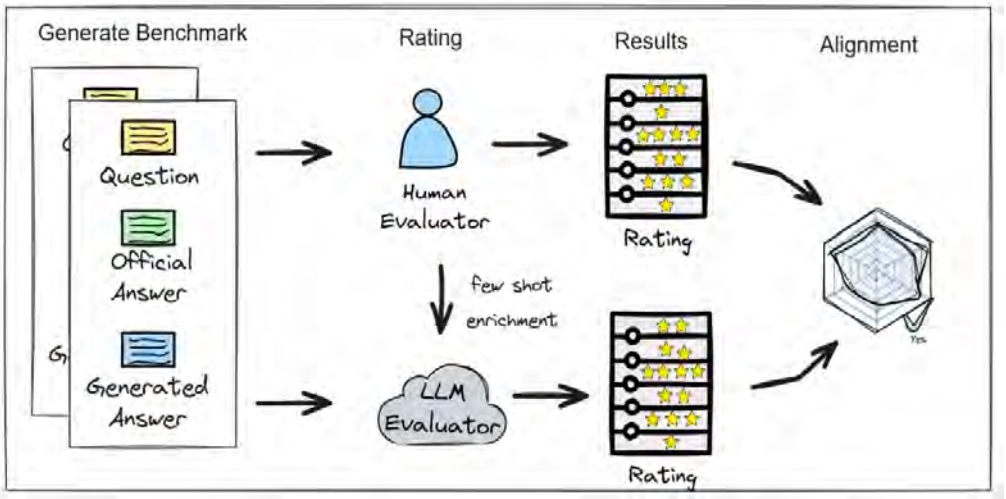
- Score 1: The generated answer is completely incorrect and incomplete compared to the official answer.
- Score 2: The generated answer is incorrect but either complete or partially complete compared to the official answer. It contains some useful information found in the official answer, but the main statement is incorrect.

- Score 3: The generated answer is correct but only partially complete. The main statement matches the official answer, but some information from the official answer is missing.
- Score 4: The generated answer is fully correct and complete. It is essentially a rephrased version of the official answer with no significant differences.

To preliminary validate the effectiveness of our LLM Evaluator, we conducted an experiment using a synthetic dataset. This dataset was carefully designed to test various aspects of language generation and was evaluated by both a human expert and the LLM. The alignment between the human expert's assessments and that of the LLM was then analysed (see "Prompt 3" in Appendix for the complete details of the final prompt used for the LLM evaluator).

The synthetic dataset comprises 60 Q&A pairs, balanced across the four score categories. For each category, two pairs were excluded as they were used as examples for the prompt for the LLM evaluator, resulting in a final dataset of 52 Q&A pairs to measure the alignment between the human and the LLM Evaluator. Using GPT-4o, we obtained a Kendall-tau coefficient of 0.86, with a p-value of 6.23x10−11. These results justified the adoption of the LLM Evaluator, especially for tasks involving prompt optimization. Figure 3 illustrates the complete process of evaluating the alignment between the LLM Evaluator and the human expert.

Evaluating the alignment between the LLM Evaluator and the human expert Figure 3



4. Experiments and Results

This section describes the results obtained by measuring retrieval effectiveness and answer quality. The first is measured by the number of relevant articles retrieved (i.e. the recall) using different encoder models, the second is then evaluated by the LLM Evaluator. We compare the multi-step prompt approach with a few-shot and a zero-shot one focusing on a single topic within the EBA Q&A framework, specifically

Liquidity Risk. Finally, we benchmark our multi-step pipeline using other LLM models, such as Google Gemini Flash 1.5 and Llama 3.1 70B.

4.1. Measuring the relevance of the CRR articles retrieved by the tool

We employed the "recall" as the primary metric to assess the performance of bi- and cross-encoder models in retrieving relevant CRR articles based on the information submitted with the question. The recall at N (r@N) is defined as the proportion of relevant CRR articles over the first N retrieved by the model (Manning et al., 2008). In the context of legal information retrieval, prioritizing the retrieval of all regulatory information pertinent to the question makes the recall a particularly relevant metric. Our primary objective was to identify a model that delivers high retrieval accuracy while maintaining computational efficiency. This potentially excluded models with an extremely large number of parameters, as they can be computationally expensive to run.

We conducted an effectiveness comparison between our fine-tuned CRR Ranker and several other pre-trained models:

- Bi-encoders: all-MiniLM-L6-v2 (Lewis et al., 2021), gte-large-en-v1.5 (Li et al., 2023), and bge-large-en-v1.5 (Xiao et al., 2023).
- Cross-encoders: bge-reranker-large (Xiao et al., 2023), bge-reranker-v2-m3 (Li, Zhang et al., 2023; Chen et al., 2024).

The detailed results on the EBA Q&As Test Database are presented in Table 3. Our fine-tuned CRR Ranker significantly outperformed all other models, achieving a more than 20% improvement compared to the best pre-trained model tested (bge-large-en-v1.5).

Recall scores on EBA Q&As Test Dataset — Table 3

Model	r@5	r@10	r@20	r@30
bge-reranker-large	0.17	0.23	0.31	0.38
bge-reranker-v2-m3	0.24	0.31	0.39	0.44
all-MiniLM	0.37	0.46	0.55	0.59
gte-large	0.39	0.48	0.57	0.63
bge-large-en-v1.5	0.41	0.52	0.62	0.67
CRR Ranker (ours)	0.51	0.67	0.81	0.86

4.2. Comparing our multi-step approach with standard ones

This paragraph presents the comparison of our multi-step approach against a zero-shot baseline for answering EBA liquidity risk questions. Our evaluation focuses on a subset of 46 liquidity risk-specific Q&As taken from the EBA Q&A Test database. We compared the answers produced by the multi-step approach described in the Section 2 (where the LLM receives a prompt containing the query, the enriched context and the examples) with the ones produced by a:

- Zero-Shot approach: the LLM receives a standard prompt containing just the query;
- Few-Shot approach: the LLM receives a prompt containing the query and some examples.

The quality of the answers produced by the three approaches is assessed by our LLM Evaluator, which scores each answer based on correctness and completeness relative to the official EBA one, using a scale of 1 (fully incorrect and incomplete) to 4 (fully correct and comprehensive), as described in Section 3.

To optimize the multi-step approach, we conduct experiments focusing on:

- Optimizing the top-k parameter: we investigate the impact of the number of retrieved articles (top-k) on answer quality. Using the LLM Evaluator, we experiment with different values of k to identify the optimal number for achieving the best performance.
- LLM Example filtering and re-ranking: we explore incorporating LLM-based filtering and re-ranking techniques (Lee and Roh, 2024; Chang et al., 2024) to further refine the retrieved information and the example selection.

Table 4 summarizes the evaluation results for answers generated by different approaches. Our findings demonstrate the significant superiority of the multi-step approach over "zero-shot" and "few-shot" methods in terms of answer correctness. Specifically, the "Multi Step - top 5" method, which retrieves the top 5 most relevant articles using the CRR Ranker, achieves the highest average LLM score (2.63 ± 0.90) and the highest percentage of correct answers (56.5%). This represents a significant improvement over the "zero-shot" and "few-shot" approaches, with average scores increasing by 14.4% and 3.5% respectively (from 2.30 and 2.54 to 2.63); more importantly, the percentage of correct answers increases by 23.8% and 30% respectively (from 45.7 and 43.5 to 56.5).

This underscores the importance of combining advanced retrieval techniques with a well-designed answer generation prompt.

Evaluation results for responses generated by zero-shot, few-shot and multi-step Table 4

Approach with GPT4o	AVG Score	# Correct (score>2)	% Correct
Zero Shot	2.30 (±0.89)	21	45.7
Few Shot	2.54 (±1.00)	20	43.5
Multi Step - top 5	2.63 (±0.90)	26	56.5
Multi Step - top 10	2.50(±0.94)	23	50
Multi Step - top 20	2.48(±0.86)	23	50
+ LLM-Reranker top 5	2.50(±0.89)	23	50
+ LLM Examples filter	2.48(±0.96)	24	53.3

It is important to highlight that the potential negative impact of incorrect information in this critical domain makes "correctness" the primary dimension for an immediate usage of the response. On the other hand, we argue that the "completeness" of the response serves as a key indicator of the tool's overall usefulness, as it ensures the provision of a comprehensive set of relevant references

that support the analyst's assessment. In addition, given the way the evaluation scale is defined, achieving a score of 4 is extremely difficult, as it requires the tool's response to be nearly identical the one provided by the EBA or its paraphrase. This is very challenging, as there may well be different equally valid variations to the same official answer, each varying depending on the specific sensitivity and expertise of the analyst producing the answer.

Our results also reveal a crucial caveat: enriching the context with too much information can introduce irrelevant or weak references, negatively impacting the quality of the answer. This emphasises the need to balance context enrichment with noise reduction. That is why, while it is fundamental to retrieve relevant information, ensuring its quality is equally important to prevent hindering the LLM's reasoning and answer generation capabilities. In this regard, employing an LLM reranker to select the most relevant articles (see "Prompt 4" in Appendix) and filtering out irrelevant examples (see "Prompt 5" in Appendix) improves the correctness of the answer compared to the "Multi Step - top 20" baseline.

These findings suggest that the context enrichment of our multi-step prompts effectively guides the LLM toward generating more comprehensive and informative answers.

4.2.1. Other LLMs

Our multi-step pipeline is also tested using other large language models (LLMs), specifically Google Gemini Flash 1.5 and Llama 3.1 70B. The first is widely recognized for its high-speed processing capabilities and efficiency in response generation, making it a suitable benchmark for comparative performance analysis. Conversely, the second is noted for its robustness in handling complex queries while maintaining moderate computational demands, providing an interesting contrast in terms of performance and resource efficiency.

Our experimental results indicate that the average evaluation score achieved by Google Gemini Flash 1.5 was 2.0, whereas Llama 3.1 70B achieved an average score of 2.2. Notably, these scores did not surpass the performance of the GPT-4o zero-shot approach, which underscores the advanced capabilities of GPT-4o in addressing the complexities of the inquiries in our dataset.

Future research will focus optimizing each step of the multi-step pipeline in a model-specific manner to further enhance the overall accuracy and reliability of the generated answers.

5. Challenges and Advancements

In our work we were confronted with several challenges. One of the main issues was the limited size of our test dataset since, as already mentioned, we focused on the single topic of Liquidity Risk, to limit the very time-consuming preprocessing phase of regulatory documents.

However, to achieve a robust alignment with the results obtained by the human expert and ensure that the tool addresses diverse inquiries across other EBA Q&A topics, future efforts should prioritize dataset expansion and human evaluation

integration. The study also emphasizes the need to retrieve relevant CRR articles to properly enrich the context but warned against the risk of gathering too much information.

Future research could also investigate methods to further refine the format of the generated answers by incorporating legal reasoning and argumentation capabilities into the LLM (Yu et al., 2023; Lu and Kao, 2024), and Q&As selected case by case as relevant examples for few-shot prompting (Wiratunga et al., 2024).

It is also crucial to underscore the importance of optimizing prompts for this kind of application, and we plan to address this moving forward. Our future research will focus on investigating automatic prompt engineering techniques (Ye et al., 2024). Moreover, we intend to extend our study to test other models that have demonstrated similar performance levels as GPT-4o in the field of open question answering (Huang et al., 2024). This will help us identify the most effective model for our application (Panickssery et al., 2024).

Similarly, in the context of LLM evaluators, we also intend to explore additional models, including open-source options (Kim, Suk et al., 2024; Kim, Suk, Cho et al., 2024), that have shown strong performance in assessing the quality of outputs of other LLMs. This approach could furtherly increase the correlation between human and LLM evaluations, thereby enhancing the tool's overall accuracy and reliability. The scientific community is very active in this area to better understand the limitations of the different types of models considered as evaluators (Huang, Qu et al., 2024).

By addressing the identified limitations through increased human involvement, expanded data coverage, and domain-specific evaluation methods, we believe it is possible to enhance the tool's effectiveness and generalizability across a wide range of regulatory domains.

6. Conclusion

This study explores a novel approach for generating answers to questions on the Regulation (EU) 2013/575, specifically on the liquidity risk topic. We propose a multi-step prompt construction method that enriches the context provided to an LLM, enabling it to generate more accurate and informative answers. An LLM Evaluator, which demonstrated strong agreement with human experts, is employed to compare our multi-step approach with standard zero-shot and few-shot methods that lack context enrichment. The quality of the generated answers is assessed, and our findings indicate that the multi-step approach significantly outperforms both the other methods, resulting in more comprehensive and accurate answers. These results suggest that the multi-step prompt construction is a promising approach for enhancing LLM performance in legal information retrieval tasks, particularly within domains with complex regulatory frameworks. Specifically, it is worth noting that the CRR Ranker enables the retrieval of most relevant documents in a fast and accurate manner, significantly aiding in the process of addressing a question. Even at this early stage, the tool has demonstrated its ability to make the work of the human analyst more efficient. We believe that answers with a quality score of at least 2 already provide a useful starting point for addressing complex regulatory questions. The results obtained with our tool provide a solid basis for the human expert (who should

always stay "in the loop" when dealing with sensitive issues such as legal requirements), reducing the time needed to analyse the issue behind the questions.

The development of a more precise metric to quantify usefulness would require the involvement of additional experts in supervisory regulation, which would allow a more accurate assessment of the correlation between the score and perceived usefulness in improving process efficiency. Specifically, this could be achieved through a dedicated internal survey among the domain experts to quantitatively assess the "perceived usefulness" of the tool, for instance, by requesting to evaluate, for each pair of question and answer the estimated reduction in the time required to analyse the question and produce the answer with and without the tool.

Finally, our research highlights the advantages of using public cloud infrastructures for LLM research and development. This approach enables efficient testing of diverse LLMs, resource and cost optimization, proving particularly suitable for projects utilizing public data. On the other hand, even when less desirable in terms of cost, powerful on-premises solutions remain preferable, if not the only ones, for managing sensitive data.

References

Abdallah, A., B. Piryani, A. Jatowt, Exploring the state of the art in legal QA systems, Journal of Big Data 10 (2023) 127. URL: https://doi.org/10.1186/s40537-023-00802-8. doi:10.1186/s40537-023-00802-8.

Biancotti, C., C. Camassa, Loquacity and Visible Emotion: ChatGPT as a Policy Advisor, 2023. URL: https://papers.ssrn.com/abstract=4533699. doi:10.2139/ssrn.4533699.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

Chan, C.-M., W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate, 2023. URL: http://arxiv.org/abs/2308.07201. doi:10.48550/arXiv.2308.07201, arXiv:2308.07201 [cs].

Chang, C.-Y., Z. Jiang, V. Rakesh, M. Pan, C.-C. M. Yeh, G. Wang, M. Hu, Z. Xu, Y. Zheng, M. Das, N. Zou, Main-rag: Multi-agent filtering retrieval-augmented generation, 2024. URL: https://arxiv.org/abs/2501.00332. arXiv:2501.00332.

Chen, J., S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

Dubois, Y., X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, T. B. Hashimoto, Alpacafarm: A simulation framework for methods that learn from human feedback, 2024. URL: https://arxiv.org/abs/2305.14387. arXiv:2305.14387.

Fu, J., S.-K. Ng, Z. Jiang, P. Liu, Gptscore: Evaluate as you desire, 2023. URL: https://arxiv.org/abs/2302.04166. arXiv:2302.04166.

Homoki, P., Z. Ződi, Large language models and their possible uses in law, Hungarian Journal of Legal Studies 64 (2024) 435–455. URL: https://akjournals.com/view/journals/2052/64/3/article-p435.xml. doi:10.1556/2052.2023.00475, publisher: Akadémiai Kiadó Section: Hungarian Journal of Legal Studies.

Horton, J. J., Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, 2023. URL: https://arxiv.org/abs/2301.07543v1.

Huang, H., Y. Qu, H. Zhou, J. Liu, M. Yang, B. Xu, T. Zhao, On the limitations of fine-tuned judge models for llm evaluation, 2024. URL: https://arxiv.org/abs/2403.02839. arXiv:2403.02839.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL: https://arxiv.org/abs/2311.05232. arXiv:2311.05232.

Huang, Z., Z. Wang, S. Xia, P. Liu, Olympicarena medal ranks: Who is the most intelligent ai so far?, 2024. URL: https://arxiv.org/abs/2406.16772. arXiv:2406.16772.

Kim, S., J. Suk, J. Y. Cho, S. Longpre, C. Kim, D. Yoon, G. Son, Y. Cho, S. Shafayat, J. Baek, S. H. Park, H. Hwang, J. Jo, H. Cho, H. Shin, S. Lee, H. Oh, N. Lee, N. Ho, S. J. Joo, M. Ko, Y. Lee, H. Chae, J. Shin, J. Jang, S. Ye, B. Y. Lin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models, 2024. URL: https://arxiv.org/abs/2406.05761. arXiv:2406.05761.

Kim, S., J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, Prometheus 2: An open source language model specialized in evaluating other language models, 2024. URL: https://arxiv.org/abs/2405.01535. arXiv:2405.01535.

Lai, J., W. Gan, J. Wu, Z. Qi, P. S. Yu, Large Language Models in Law: A Survey, 2023. URL: http://arxiv.org/abs/2312.03718. doi:10.48550/arXiv.2312.03718, arXiv:2312.03718 [cs].

Lee, J., M. Roh, Multi-reranker: Maximizing performance of retrieval-augmented generation in the financerag challenge, 2024. URL: https://arxiv.org/abs/2411.16732. arXiv:2411.16732.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for

knowledge-intensive NLP tasks, 2020. In Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada). URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Lewis, P. S. H., Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, S. Riedel, PAQ: 65 million probably-asked questions and what you can do with them, CoRR abs/2102.07033 (2021). URL: https://arxiv.org/abs/2102.07033. arXiv:2102.07033.

Li, C., Z. Liu, S. Xiao, Y. Shao, Making large language models a better foundation for dense retrieval, 2023. arXiv:2312.15503.

Li, Z., X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153. doi:10.18653/v1/2023.emnlp-main.153.

Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL: https://arxiv.org/abs/2303.16634. arXiv:2303.16634.

Louis, A., G. van Dijck, G. Spanakis, Interpretable Long-Form Legal Question Answering with Retrieval Augmented Large Language Models, 2023. URL: http://arxiv.org/abs/2309.17050. doi:10.48550/arXiv.2309.17050, arXiv:2309.17050 [cs].

Lu, Y. an, H. yu Kao, 0x.yuan at semeval-2024 task 5: Enhancing legal argument reasoning with structured prompts, in: International Workshop on Semantic Evaluation, 2024. URL: https://api.semanticscholar.org/CorpusID:270765544.

Manning, C. D., P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, USA, 2008.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J.

Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, L. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774, arXiv:2303.08774 [cs].

Panickssery, A., S. R. Bowman, S. Feng, Llm evaluators recognize and favor their own generations, 2024. URL: https://arxiv.org/abs/2404.13076. arXiv:2404.13076.

Prenio, J., Peering through the hype - assessing suptech tools' transition from experimentation to supervision (2024). URL: https://www.bis.org/fsi/publ/insights58.htm.

Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A Large Language Model for Finance, 2023. URL: http://arxiv.org/abs/2303.17564, arXiv:2303.17564 [cs, q-fin].

Wiratunga, N., R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering, 2024. URL: https://arxiv.org/abs/2404.04302. arXiv:2404.04302.

Xiao, S., Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

Xiao, S., Z. Liu, P. Zhang, N. Muennighoff, FlagEmbedding/FlagEmbedding/reranker at master · FlagOpen/FlagEmbedding, 2024. URL: https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/reranker .

Xuan, H., A. Stylianou, X. Liu, R. Pless, Hard negative examples are hard, but useful, 2021. URL: https://arxiv.org/abs/2007.12749. arXiv:2007.12749.

Ye, Q., M. Axmed, R. Pryzant, F. Khani, Prompt engineering a prompt engineer, 2024. URL: https://arxiv.org/abs/2311.05661. arXiv:2311.05661.

Ye, S., D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, M. Seo, FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets, 2024. URL: http://arxiv.org/abs/2307.10928. doi:10.48550/arXiv.2307.10928, arXiv:2307.10928 [cs].

Yu, F., L. Quartey, F. Schilder, Exploring the effectiveness of prompt engineering for legal reasoning tasks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13582–13596. URL: https://aclanthology.org/2023.findings-acl.858. doi:10.18653/v1/2023.findings-acl.858.

Zhang, W., H. Shen, T. Lei, Q. Wang, D. Peng, X. Wang, GLQA: A Generation-based Method for Legal Question Answering, in: 2023 International Joint Conference on Neural Networks (IJCNN), 2023, pp. 1–8. URL: https://ieeexplore.ieee.org/document/10191483?denied=. doi:10.1109/IJCNN54540.2023.10191483, iSSN: 2161-4407.

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023. URL: http://arxiv.org/abs/2306.05685. doi:10.48550/arXiv.2306.05685, arXiv:2306.05685 [cs].

Appendix

Prompt 1 - Extracting Law References

Gpt4-omni Prompt

#task
Extract from the text (#text) any reference to regulatory documents contained in it and insert them into a list (e.g. ["regulatory document name": ["article 1","article 2",...]]). I will provide you an example (#text (example)) and the expected output (#output (example)):

#text (example) "In accordance with Article 425 (1) of Regulation (EU) No. 575/2013 (CRR) institutions may exempt contractual liquidity inflows from borrowers and bond investors arising from mortgage lending funded by covered bonds eligible for preferential treatment as set out in Article 129b (4-6) of CRR or by bonds as referred to in Article 52(4) of Directive 2009/65/EC from the 75% inflow cap."

#output (example) "["Regulation (EU) No. 575/2013 (CRR)": ["425","129b"], "Directive 2009/65/EC" : ["52"]]"
#text
> *text_to_extract*

#output (list only)

This prompt was used to extract any reference to regulatory documents from the provided text_to_extract) (placeholder to input text)

Prompt 2 - Answer Generation

Gpt4-omni Prompt

" #system
You are a virtual assistant for the European Banking Authority (EBA), handling user inquiries related to Liquidity Risk regulations. The user's query specifically pertains to Regulation (EU) No. 575/2013 (CRR) or Delegated Regulation (EU) No. 2015/61 (LCR DA)."""

#task
Answer the question based on the instructions below.
1. Analyze the User's Question (#question):
- Identify the central topic and relevant keywords related to Liquidity Risk and the specified EBA regulations.
2. Leverage the Provided Context (#context):
- Incorporate the context (including CRR articles and additional information) to tailor the answer to the user's specific scenario.
3. Liquidity Risk Topic
- Reference relevant articles from provided context (#context) that address the specific aspect of Liquidity Risk raised in the question. 4. Desired Answer (#answer):
- Use only the information provided in the context and examples (if provided) to answer the question.
- Craft a well-reasoned and informative response that covers all aspects of the user's query.
- Clearly articulate the regulatory implications while considering the provided context.
- Maintain a professional and informative tone suitable for the EBA.

#examples:
Example 1: > *example_1*
Example 2: > *example_2*
Example 3: > *example_3*
Example 4: > *example_4*
Example 5: > *example_5*

#question:
> *question*

#context:
> *context*
> *enhanced_context*

#answer:

This prompt was used to generate answer given a question and context. #examples section (placeholder to include 5 examples) and enhanced_context (placeholder to include CRR articles), highlighted in yellow, were used only for multi-step approach.

Prompt 3 - LLM as Evaluator

Gpt4-omni Prompt

I will provide you with two answers to a question.One is the #official answer, which serves as the benchmark. The other is the #generated answer, which needs to be evaluated against the #official answer. You must compare the answers step by step.

Consider the following definitions for this evaluation:

- Correctness: A #generated answer is correct if its content aligns with that of the #official answer.
- Completeness: A #generated answer is complete if it includes all the information present in the #official answer.

Your task is to act as an evaluator and rate the #generated answer according to the following scale:

RATING 1: The #generated answer is completely incorrect and incomplete compared to the #official answer.
RATING 2: The #generated answer is incorrect but either complete or partially complete compared to the #official answer. It contains some useful information found in the #official answer but the main statement is incorrect.
RATING 3: The #generated answer is correct but only partially complete. The main statement matches the #official answer, but some information from the #official answer is missing.
RATING 4: The #generated answer is fully correct and complete. It is essentially a rephrased version of the #official answer with no significant differences.
Please provide a single numerical rating (1-4) followed by a brief explanation for your rating

<EXAMPLE 1>
...
<EXAMPLE 8>

Compute the score in the following case:

#question
> question

#background
> background

#official answer
> answer

#generated answer
generated answer

Output:

This prompt was used to compare an AI-generated answer (#generated answer) to an official one (#official answer), rating its correctness, completeness, and providing an explanation.

Prompt 4 – Rerank

Gpt4-omni Prompt

You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

I will provide you with num passages, each indicated by number identifier [].

Rank the passages based on their relevance to query: {query}. Search
Query: {query}

Rank the {num} passages above based on their relevance to the search query.

The passages should be listed in descending order using identifiers. The most relevant passages should be listed first.
The output format should be [] > [] > [] > [] > ..., e.g., [1] > [2] > [3] > [4] > ...

Only response the ranking results, do not say any word or explain.

Prompt 5 Examples Filter Prompt

Gpt4-omni Prompt

You are a virtual assistant for the European Banking Authority (EBA), responsible for analyzing inquiries related to Liquidity Risk regulations under Regulation (EU) No. 575/2013 (CRR) and Delegated Regulation (EU) No. 2015/61 (LCR DA).

Your task is to filter out irrelevant examples provided by the user.

Follow these instructions to determine which examples are not useful for addressing the user's specific question.

1. Understand the user's question (#question) by identifying its core topic, keywords, and references to relevant regulations or concepts.
2. Analyze the provided context (#context), including the operational details and CRR articles referenced, toclarify the regulatory framework applicable to the question.
3. Review the examples (#examples), which are numbered from 1 to 5 and contain separate Q&A entries,each with its own context and answer.
4. Evaluate the relevance of each example by checking if it directly contributes to answering the user'squestion based on: - Relevance to the regulatory topic or specific articles mentioned in the question. - Applicability of the example's context to the user's scenario. - Alignment with the CRR or LCR DA framework relevant to the question.
5. For each example, determine if it is irrelevant and briefly justify why it does not provide useful informationfor the specific question.

6 Output a list of the relevant examples by their number (do not provide any short justification but only the list of number).

#question:
question

#context:
context

#examples
examples

A Novel Multi-Step Prompt Approach for LLM-based Q&As on Banking Supervisory Regulation

Daniele Licari[1], Canio Benedetto[1], Daniele Bovi[1], Praveen Bushipaka[2], Alessandro De Gregorio[1], Marco De Leonardis[1] and Tommaso Cucinotta[2]

[1]Banca d'Italia, via nazionale, 91, Rome, 00184, Italy - (name.surname@bancaditalia.it)

[2]Scuola Superiore Sant'Anna, P.zza dei Martiri della Libertà, 33, Pisa, 56100, Italy – (name.surname@santannapisa.it)

[1]The views and opinions expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of the Bank of Italy.

Outline

Introduction

1. Motivation
2. EBA Single Rulebook Q&A
3. Dataset
4. Challenges in Q&A System Dev
5. Methodology Overview

Methodology

1. 3-Step CRR Articles Retrieval
2. Answer Generation Process
3. LLM Evaluator
4. Results and Analysis
5. Conclusion and Future Work

Motivation

- **Challenges in Regulatory Compliance**
  Regulatory documents are complex, with dense cross-references and specialized content, making manual analysis time-consuming and error-prone.

- **Need for Efficient Solutions**
  Compliance professionals and supervisory authorities require tools to streamline the navigation and interpretation of these regulations.

- **Explore the potential of LLMs**
  Large Language Models (LLMs) can serve as powerful assistants, helping users understand regulatory frameworks through appropriate tools while simplifying and streamlining work-related tasks as **Question & Answer** (Q&A) systems.

- **Proposed Methodology**
  This study uses a **Retrieval-Augmented Generation (RAG)** like framework using LLMs to automate and enhance the Q&A process for regulatory compliance using the **European Banking Authority (EBA) Single Rulebook Q&A**.

European Banking Authority (EBA) Single Rulebook Q&A

- 1,597 Q&A pairs covering regulatory topics in European banking (2013-2020).
- Provides clarification on the application of complex banking regulations.
- Supports compliance professionals in understanding legal obligations.
- Each Q&A contains extensive information. For the development of our system, the following fields proved particularly useful:
  - **Background of the question**: any additional information or context provided by the question submitter.
  - **Question**: the actual question being asked.
  - **Answer**: the official answer provided to the question.
  - **Submission date**: the date the question was submitted, used to determine the regulatory documents valid at that time.
  - Specific references provided by the user, including **Article**, **Paragraph**, **Subparagraph**, **COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations**, or other relevant legislation, standards, guidelines, or recommendations related to the question.

Legal act	Regulation (EU) No 575/2013 (CRR)
Topic	Transparency and Pillar 3
Article	449a
Paragraph	-
Subparagraph	-
COM Delegated or Implementing Acts/RTS/ITS/GLs/Recommendations	Regulation (EU) 2022/2453 - ITS on ESG disclosures
Article/Paragraph	1
Date of submission	30/01/2023
Published as Final Q&A	25/08/2023
Disclose name of institution / entity	No
Type of submitter	Competent authority
Subject matter	ESG P3 - Reg 2453/22 - Scope of application
Question	What is the scope of application of Regulation (EU) 2022/2453?
Background on the question	Article 449a CRR states that: "large institutions which have issued securities that are admitted to trading on a regulated market of any Member State, as defined in point (21) of Article 4(1) of Directive 2014/65/EU, shall disclose information on ESG risks". In the EBA/ITS/2022/01 we found the following clarification: "Institutions should disclose the information at the highest level of consolidation consideration in the EU, as regulated in Article 13 CRR." (page 52) Due to the above, there is different understanding of the scope of application of the ITS on Pillar 3 ESG disclosures, both between banks as well as the regulator. Shall all large listed institutions in a jurisdiction disclose the P3 ESG information (this means disclosure on subconsolidated basis for some banks) or only large listed institutions which are the EU highest parent institutions (this significantly reduces the number of banks subject to the Reg. 2022/2453 in PL).
Final answer	As clarified by the EBA/ITS/2022/01, institutions should disclose the information at the highest level of consolidation in the EU, as regulated in Article 13 CRR. For the detailed information on scope of application of Article 449a CRR please refer to QA2022_6652.
Link	https://www.eba.europa.eu/single-rule-book-qa/qna/view/publicId/2023_6708

Q&A Dataset

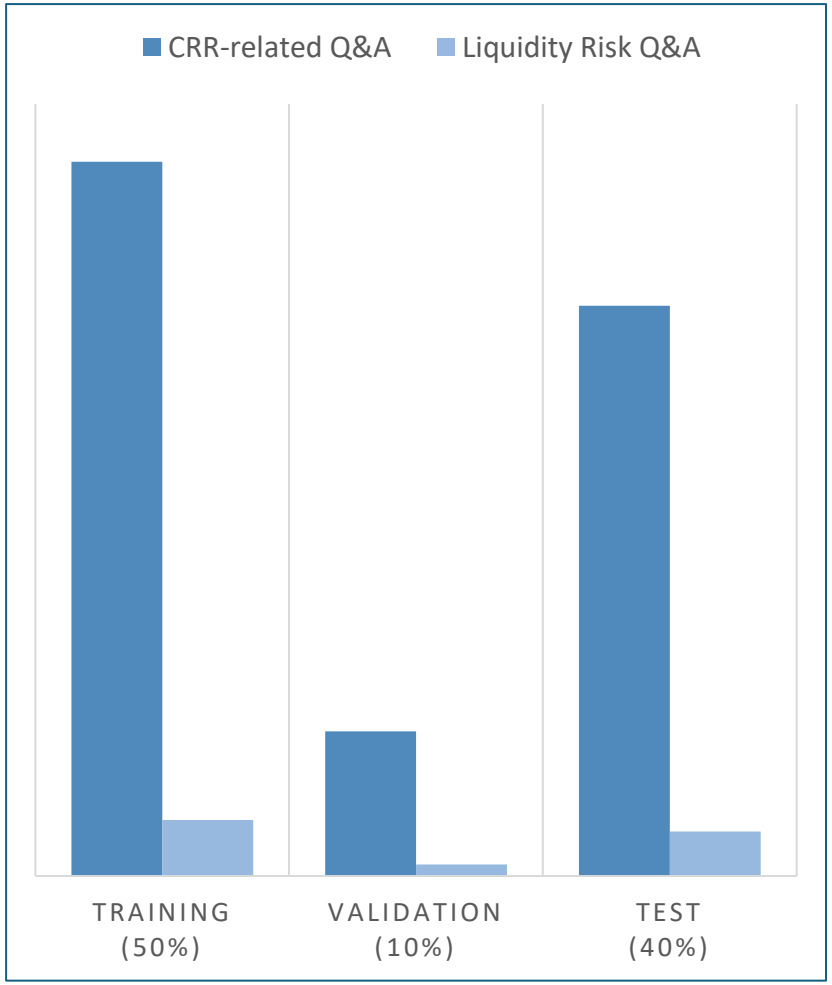
- Focus Topic: **Liquidity Risk (**112 Q&A).

- Dataset split:

  - **Training** Set: used to train and optimize the RAG phase.
  - **Validation** Set utilized for hyperparameter optimization during RAG development.
  - **Test** Set: used to evaluate the final model performance.

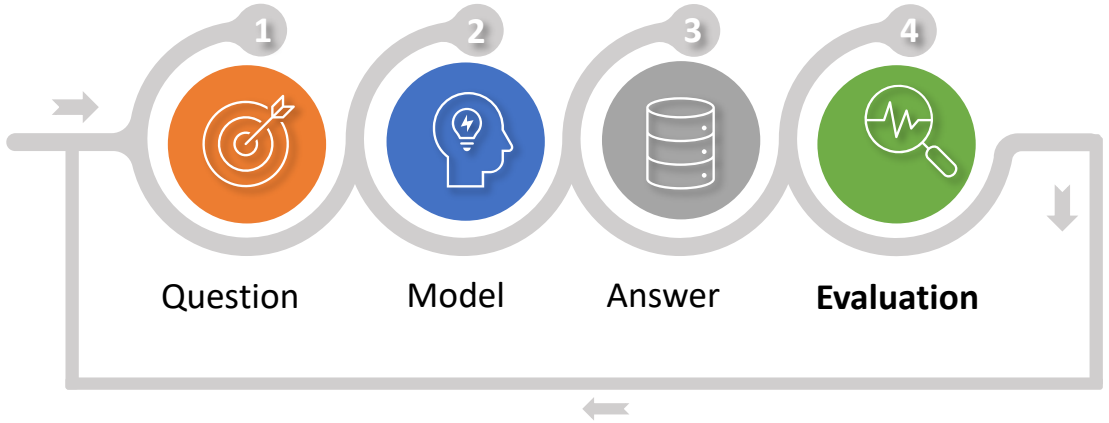
- Additional relevant documents: **Capital Requirements Regulation (CRR)** EU No. 575/2013: core banking framework counting around 500 articles.

- Planned additional documents: relevant materials to be included, such as
  - **Liquidity Coverage Requirement (LCR) Delegated Regulation**
  - **Implementing Technical Standards on Supervisory Reporting**



Challenges in Q&A System Development

- **Development Challenge**: Testing and developing a Q&A system requires repeated testing to identify the best approach.



- **Evaluation Phase**: Accurate evaluation is essential but often very **time-consuming** and **tedious**.
- **LLMs as evaluators**: LLMs act as automated evaluators, enabling **scalable**, **interpretable**, and **cost-efficient** assessments. By automating repetitive tasks, they save resources while ensuring consistent quality evaluation.
- **Challenges to Consider**: LLM judgments must align with human reasoning, which is ensured through an initial calibration and verification phase.

Methodology Overview

3-Step Retrieval CRR Articles

Specific references provided by the user

Extract **explicit** CRR references

Question (Q) + Background (B)

Prompt + [Q + B]

GPT – 4o

Extract **implicit** CRR references

CRR Ranker (cross-encoder)

Extract **top-5** CRR References

Q + B + CRR Articles
(Explicit, Implicit, top 5)

Few shot

Enrichment with examples

Generation

GPT – 4o

Instructions Prompt

Question/Background

CRR Articles

Examples

Evaluation

GPT – 4o

LLM Evaluator Scores [1-4]

Outline

Introduction

1. Motivation
2. Challenges in Q&A System Dev
3. EBA Single Rulebook Q&A
4. Dataset
5. Methodology Overview

Methodology

1. 3-Step CRR Articles Retrieval
2. Answer Generation Process
3. LLM Evaluator
4. Results and Analysis
5. Conclusion and Future Work

3-Step Retrieval CRR Articles: Context Enrichment

Q&A	
Question ID	2016_2731
Legal act	Regulation (EU) No 575/2013 (CRR)
Article	411
Question	To be eligible for Article 153(4) CRR, do firms need to have externally verified annual sales numbers, and if so how frequently should firms assess eligibility?
Background	This question is on the application of a definition in the context of the LCR.

1

A: 411

Q + B

2

LLM-Based Extraction of CRR References

#task
Extract from the text (#text) any reference to regulatory documents contained in it and insert them into a list (e.g. ["regulatory document name": ["article 1","article 2",...]]). I will provide you an example (#text (example)) and the expected output (#output (example)):

#text (example) "In accordance with Article 425 (1) of Regulation (EU) No. 575/2013 (CRR) institutions may exempt contractual liquidity inflows from borrowers and bond investors arising from mortgage lending funded by covered bonds eligible for preferential treatment as set out in Article 129b (4-6) of CRR or by bonds as referred to in Article 52(4) of Directive 2009/65/EC from the 75% inflow cap."

#output (example) "["Regulation (EU) No. 575/2013 (CRR)": ["425","129b"], "Directive 2009/65/EC" : ["52"]]"
#text
> *text_to_extract*

#output (list only)

A: 153



A: [153, 411]

3

Ranking Model

CRR Ranker

Top K



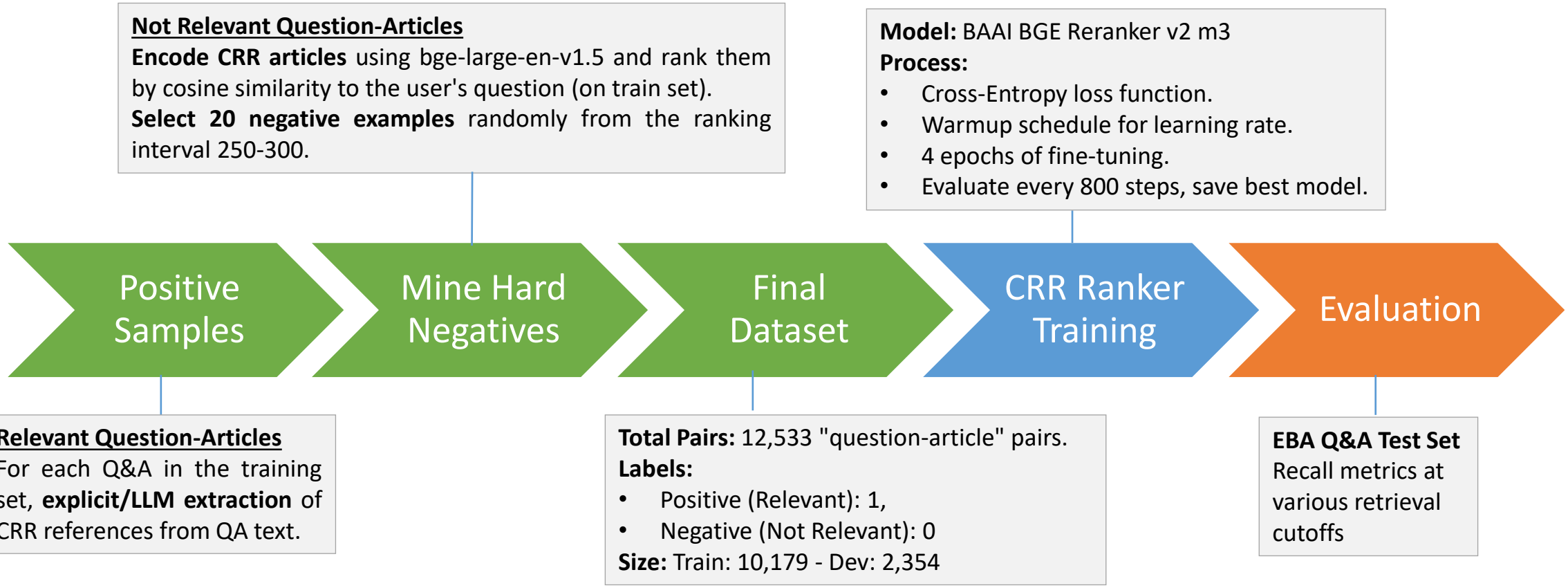
Relevant CRR Articles

Three-Step Approach:

1. **Explicit Extraction**: Directly extract CRR references mentioned in the Q&A.
2. **LLM Extraction**: identify other relevant CRR references from the *Question* and *Background* information using GPT-4o .
3. **Ranking Model**: retrieve the **top *K*** relevant CRR articles based on their similarity to the *Question* and *Background*

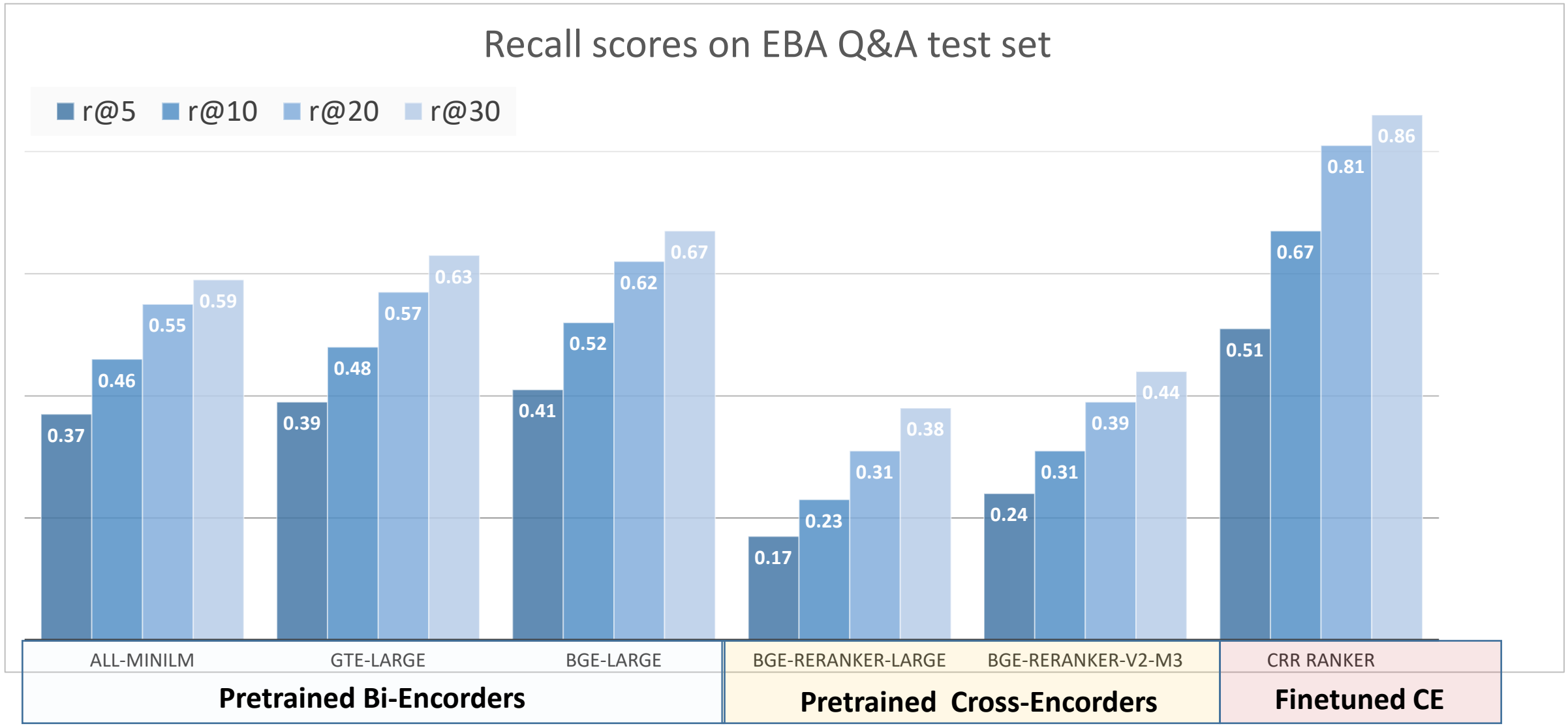
CRR Ranker Training

CRR Ranker model was trained **to retrieve pertinent articles from the CRR** in response to specific inquiries



H. Xuan et al., "Hard negative examples are hard, but useful," 2021. arXiv:2007.12749.
Chen et al., "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," 2024. arXiv:2402.03216.

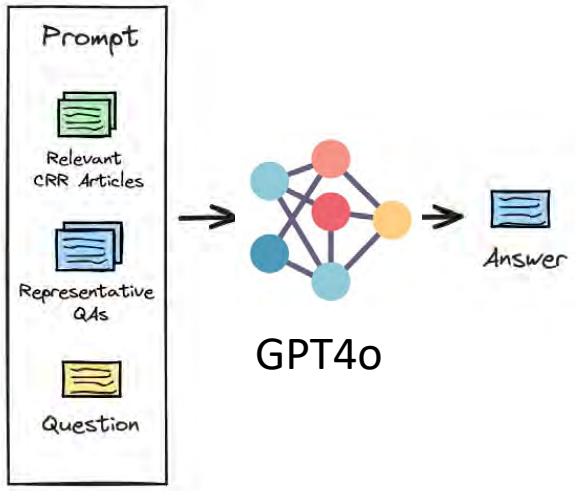
Results: CRR Retrieval



* Weiwei Sun,et al (2023). Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent

Answer Generation

- GPT 4o* for Generation.
- Role-based CoT few–shot approach includes 5 examples from the train set.
- Provides the desired tone, structure, language style, and level of detail expected.



Prompt for Answer Generation

" #system
You are a virtual assistant for the European Banking Authority (EBA), handling user inquiries related to Liquidity Risk regulations. The user's query specifically pertains to Regulation (EU) No. 575/2013 (CRR) or Delegated Regulation (EU) No. 2015/61 (LCR DA)."""

#task
Answer the question based on the instructions below.
1. Analyze the User's Question (#question):
- Identify the central topic and relevant keywords related to Liquidity Risk and the specified EBA regulations.
2. Leverage the Provided Context (#context):
- Incorporate the context (including CRR articles and additional information) to tailor the answer to the user's specific scenario.
3. Liquidity Risk Topic:
- Reference relevant articles from provided context (#context) that address the specific aspect of Liquidity Risk raised in the question. 4. Desired Answer (#answer):
- Use only the information provided in the context and examples (if provided) to answer the question.
- Craft a well-reasoned and informative response that covers all aspects of the user's query.
- Clearly articulate the regulatory implications while considering the provided context.
- Maintain a professional and informative tone suitable for the EBA.

#examples:
Example 1: > *example_1*
Example 2: > *example_2*
Example 3: > *example_3*
Example 4: > *example_4*
Example 5: > *example_5*

Question ID:
1. 2013_192
2. 2018_3730
3. 2013_301
4. 2019_4705
5. 2014_783

#question:
> *question*

#context:
> *context*
> *enhanced_context*

Retrieved CRR Articles from our 3-Step Approach

#answer:

Role-based

Task
Step by Step Reasoning

Few-shot Guidance

Question

Context Enrichment

*2024-11-20 version

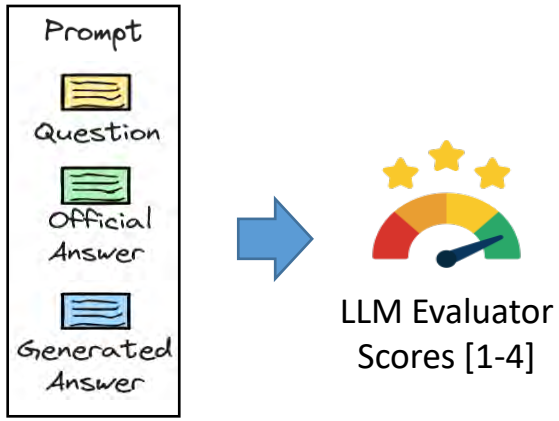
LLM Evaluator

Evaluation Scale (1-4):

- **Correctness**: Alignment with the official answer.
- **Completeness**: Inclusion of all relevant regulatory references.

Scores:

1. Completely incorrect.
2. Incorrect but complete or partially complete.
3. Correct but partially complete
4. Fully correct and complete.



Prompt for Answer Evaluation

I will provide you with two answers to a question: the #official answer (benchmark) and the #generated answer (to be evaluated). Compare them step by step based on:

Correctness: Does the #generated answer align with the #official answer?
Completeness: Does the #generated answer include all relevant information from the #official answer?
Rating Scale (1-4):

1. Completely incorrect and incomplete.
2. Incorrect but complete or partially complete.
3. Correct but partially complete.
4. Fully correct and complete.

Provide a numerical rating (1-4) followed by a brief explanation. Format your output as follows: Output: {score} Motivation: {motivation}

Examples
<EXAMPLE 1> ... <EXAMPLE 8>

Compute the score:
Question:
> question
> background
Official Answer:
> answer
Generated Answer:
> generated answer
Output:

Task

Criteria

Few-shot

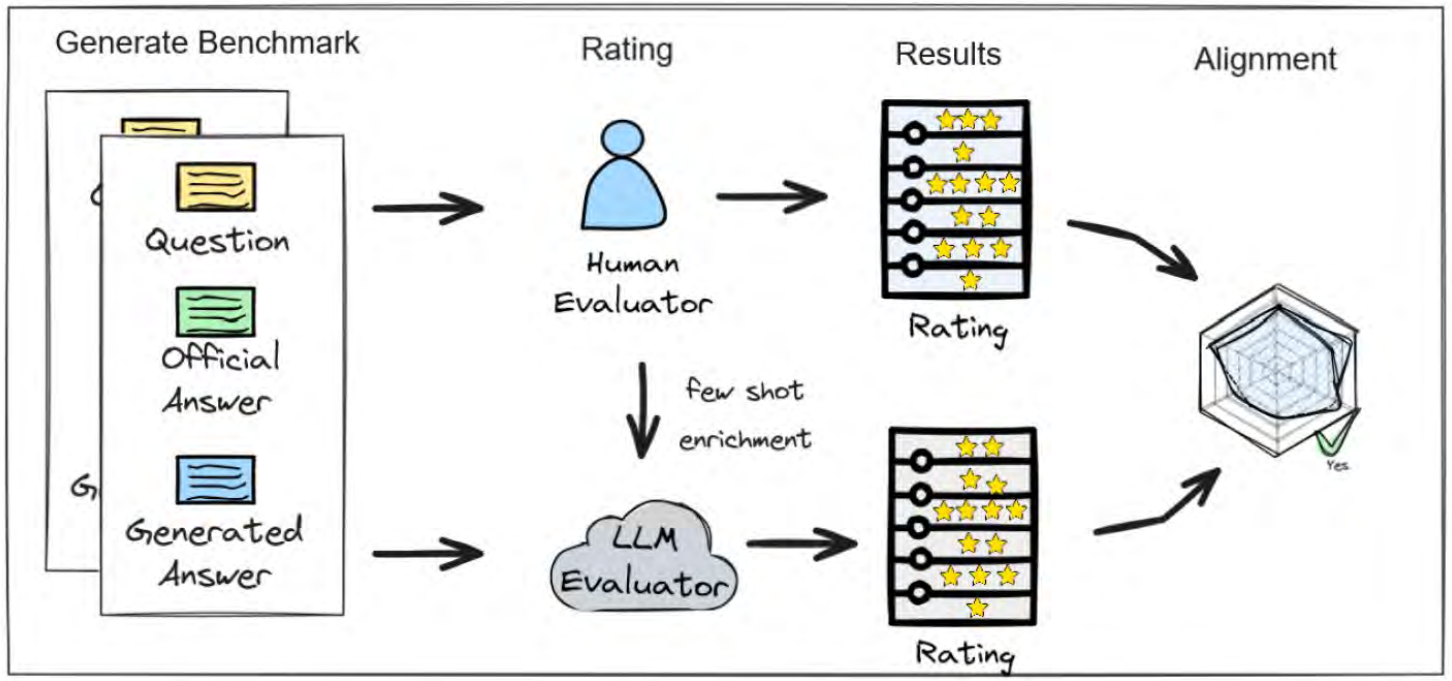
Question

Official Answer

Generated Answer

LLM Evaluator: Alignment with Human Expert

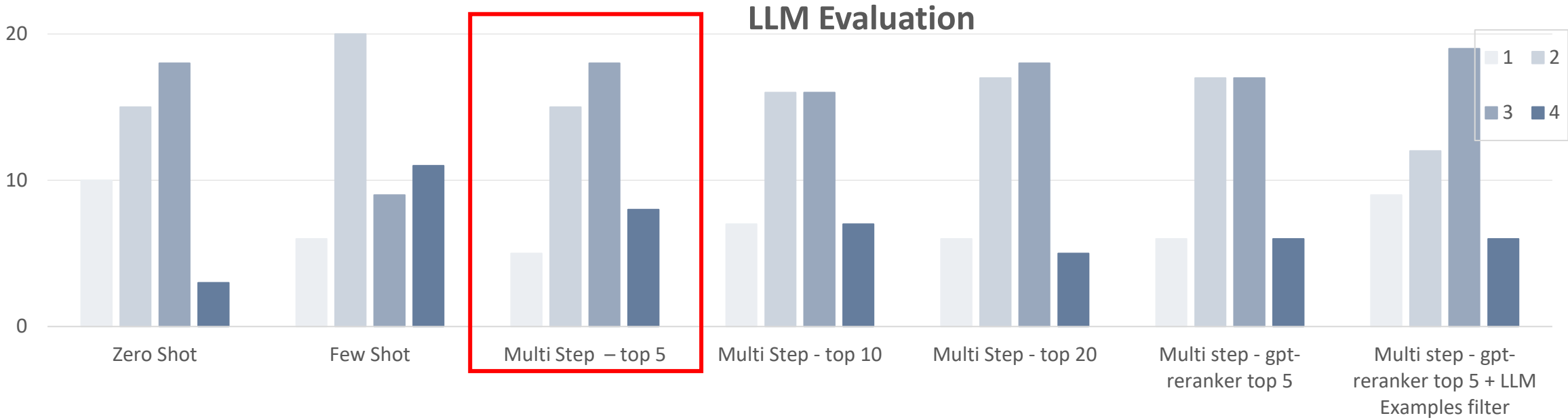
- Preliminary verification on a synthetic human-validated dataset.
- 60 Q&A pairs on Liquidity Risk topic balanced across the four categories.
- With GPT -4o**, Kendall-tau coefficient: 0.86** ,and p-value : 6.23^{-11}



Model	T	p-value
GPT4o (May '24)	0.77	6.10e-11
GPT4o (Nov '24)	**0.86**	6.23e-15
O1-mini (Sep '24)	0.64	9.44e-11

Kendall-tau: Agreement between the LLM Evaluator and human experts

Results: Answer Generation on Liquidity Risk Test Set (46 Q&As)



Approach with GPT4o (Nov'24)	AVG LLM Score	# Correct (score>2)	% Correct
Zero Shot	2.30 (±0.89)	21	45.7
Few Shot	2.54 (±1.00)	20	43.5
Multi Step – top 5	**2.63 (±0.90)**	**26**	**56.5**
Multi Step - top 10	2.50(±0.94)	23	50
Multi Step - top 20	2.48(±0.86)	23	50
+ LLM-Reranker top 5	2.50(±0.89)	23	50
+ LLM Examples filter	2.48(±0.96)	24	53.3

Multi-Step: Uses LLMs for extracting and the fine-tuned model for retrieving **top K** CRR articles.

Conclusion

Contributions:

- **Multi-Step Prompt Construction**: Enhances context for LLMs, improving answer precision and informativenes.
- **Context Enrichment**: Uses explicit and implicit CRR references, LLM capabilities, and a cross-encoder for precise retrieval.
- **LLM Evaluator**: Automates validation, ensuring response quality in terms of correctness and completeness.
- **Dataset Development**: Creates a comprehensive dataset from EBA's Single Rulebook Q&A for training and evaluation.
- **Performance Improvement**: Multi-step approach outperforms zero-shot and few-shot methods, providing better responses.

Future Directions:

- Increase the dataset size and generalization to other domains.
- Self-reflection and human feedback integration
- Explores different LLM architectures and fine tuning.

Thank you for your attention!

daniele.licari@bancaditalia.it
alessandro.degregorio@bancaditalia.it

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

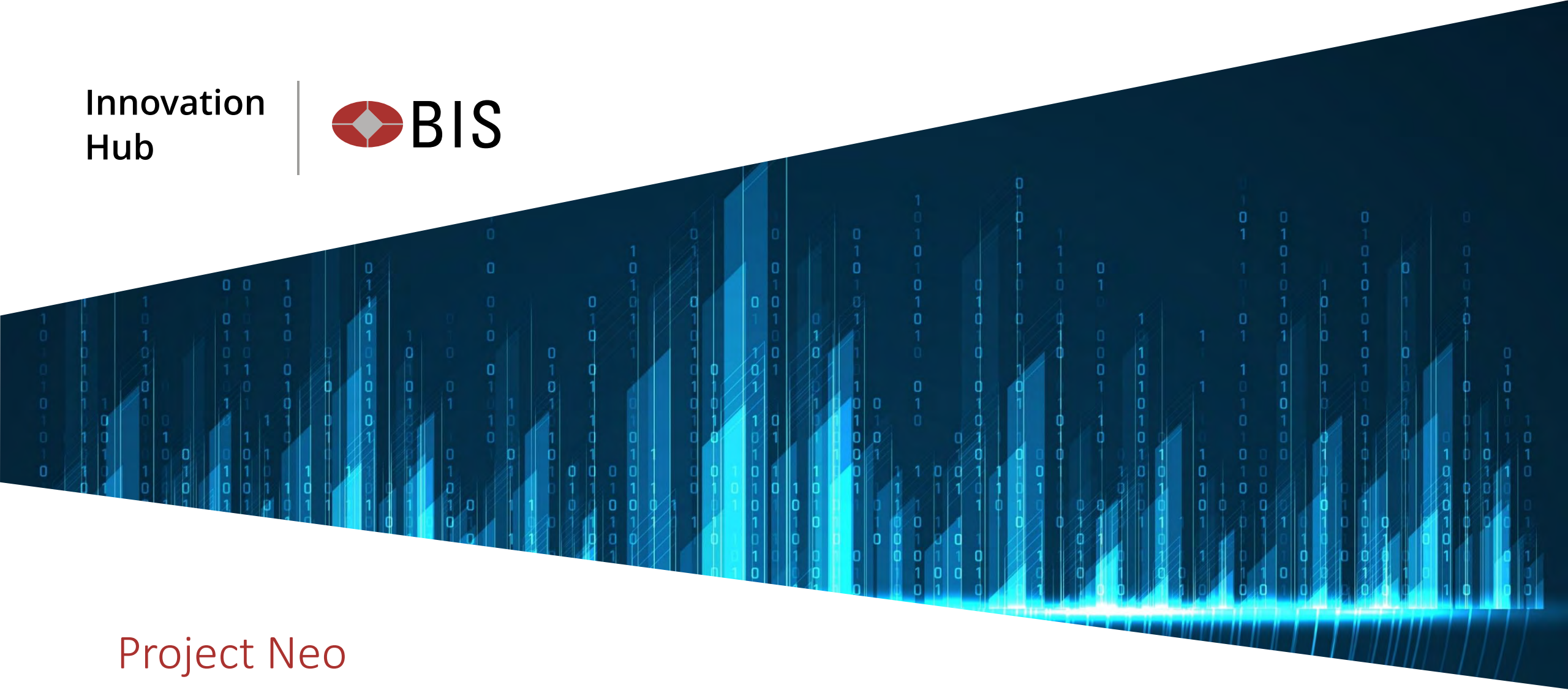
Project Neo: gaining economic insights with AI and novel data[1]

Rachel Arulraj-Cordonier, Marko Nanut Petric
and Flurina Strasser,
BIS Innovation Hub

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Innovation Hub

BIS



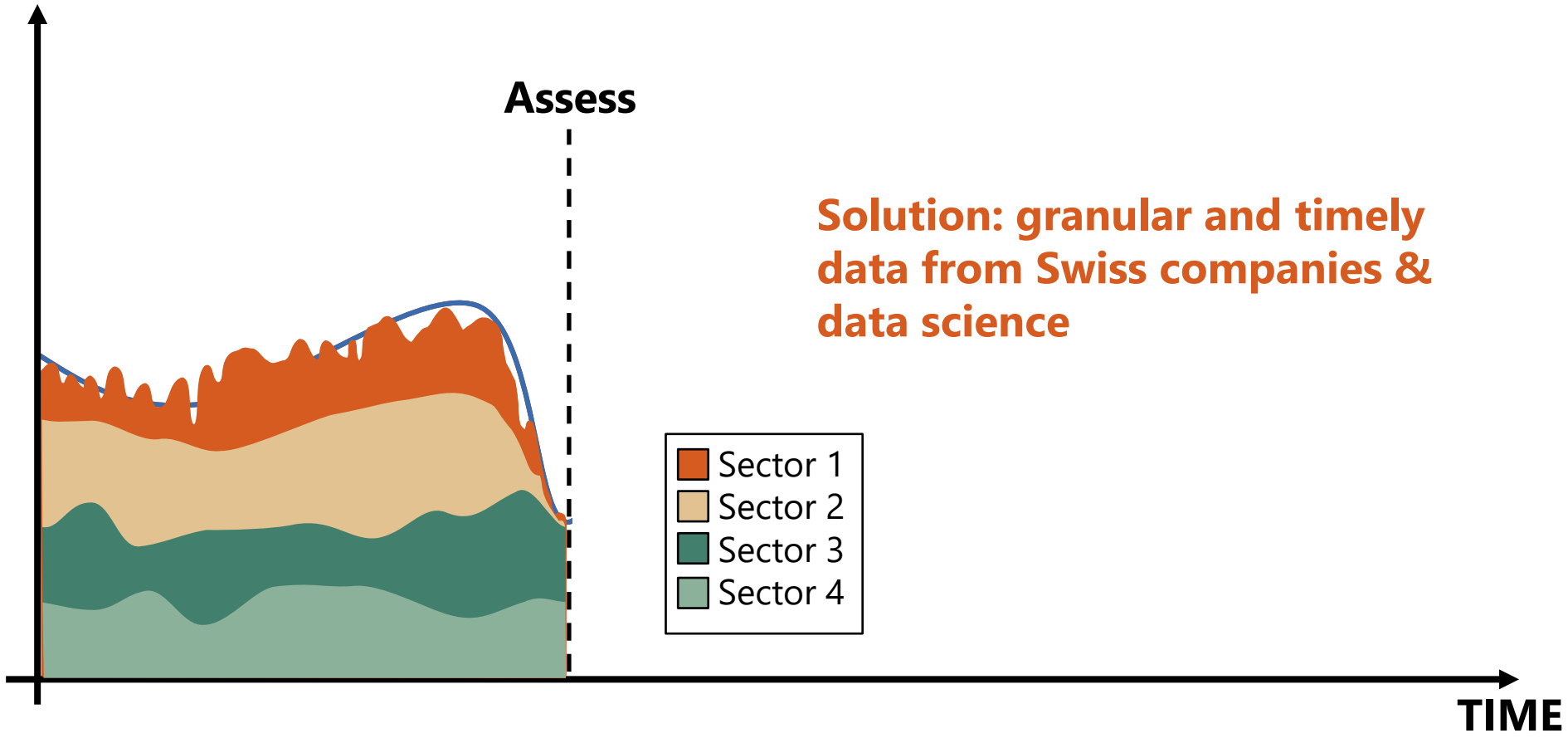
Project Neo

4th IFC-BoI workshop on "Data science in central banking", 18-20 Feb 2025

Rachel Arulraj-Cordonier

DISCLAIMER: The views are those of the presenter and not necessarily those of the BIS nor those of the SNB.

Problem statement



Solution: granular and timely data from Swiss companies & data science

Objectives

Extract value from granular data

Comparable to official statistics

To monitor individually and use in models

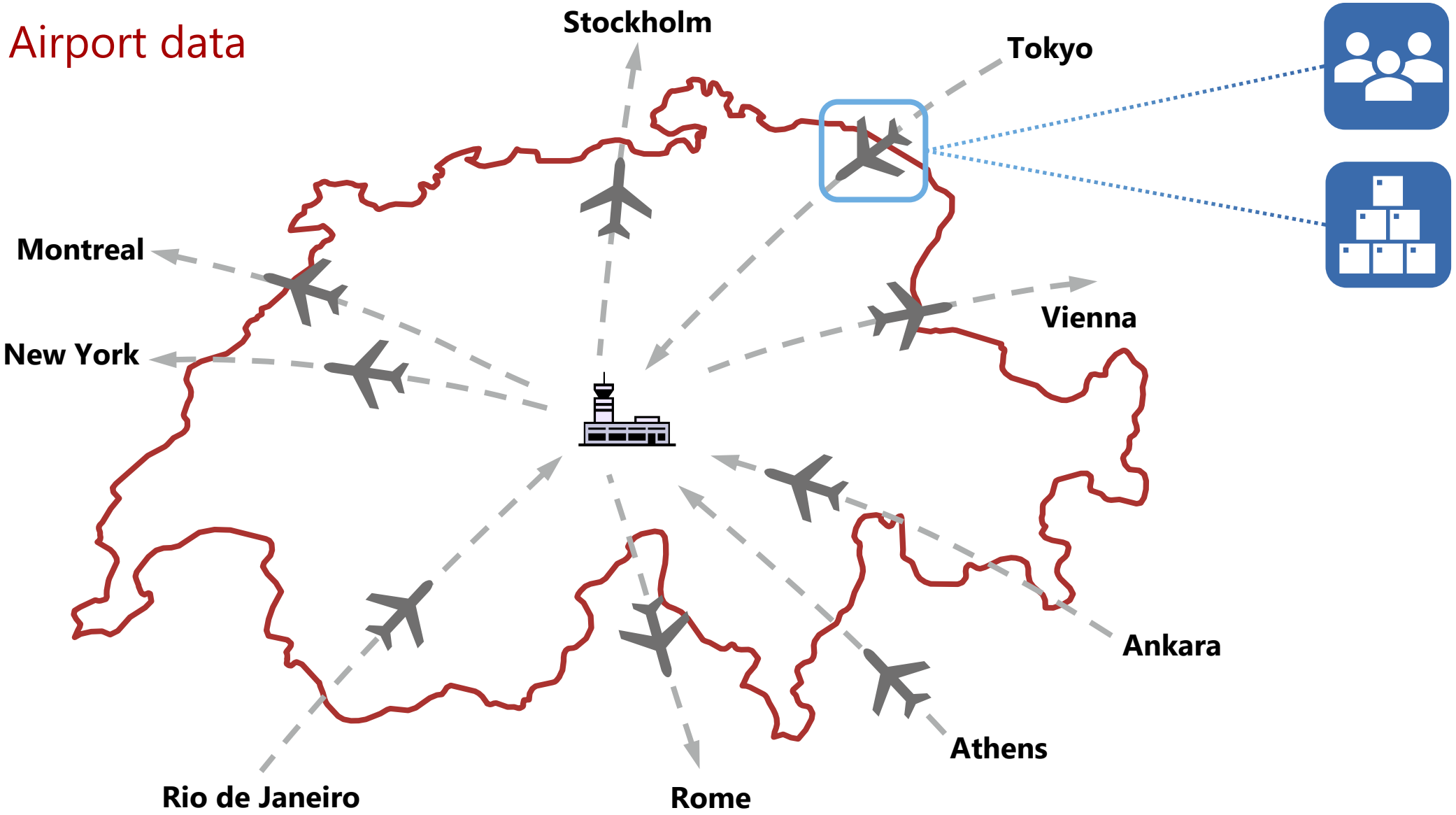


Economic monitoring

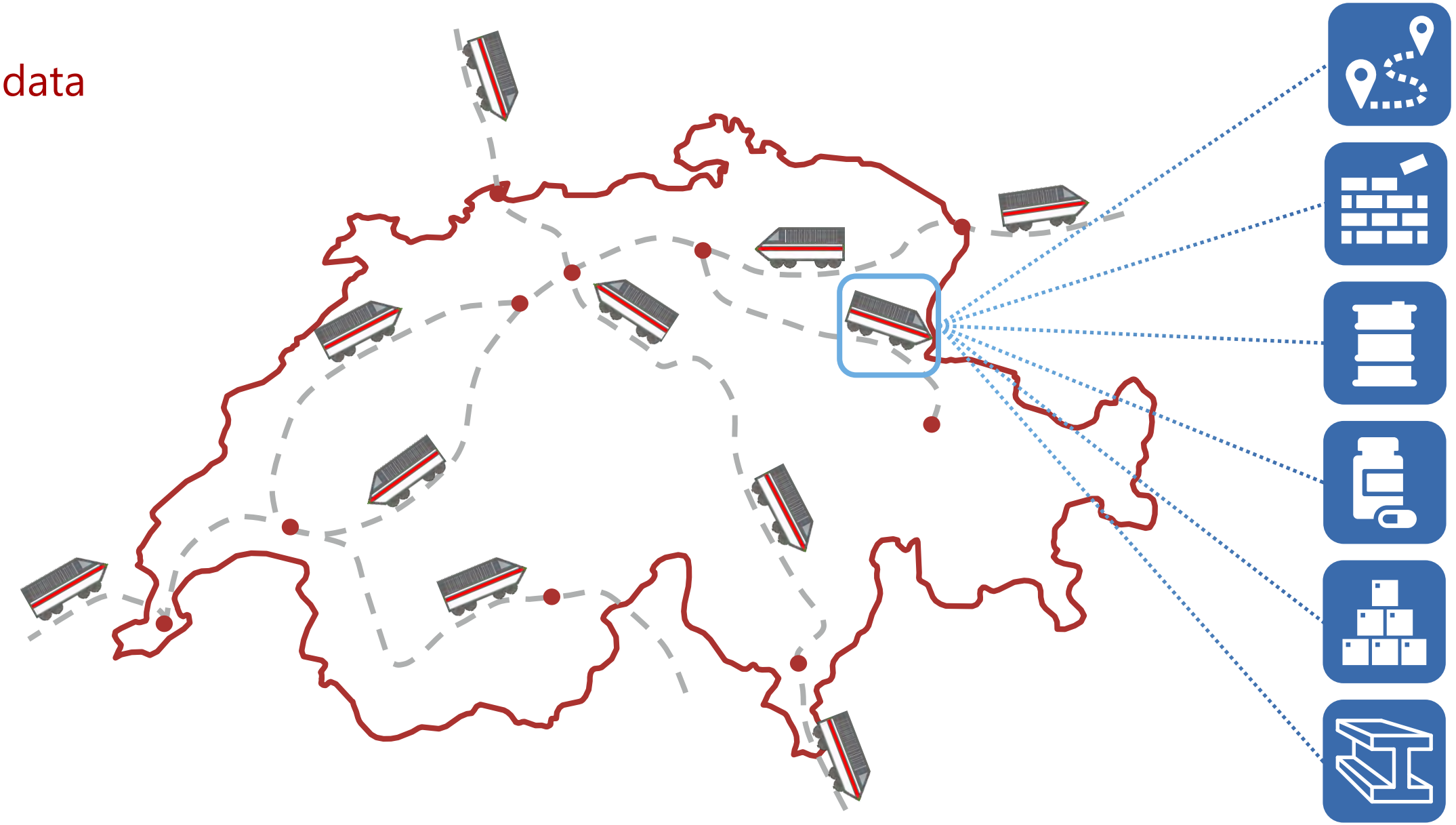
Can we create indicators directly interpretable?

1

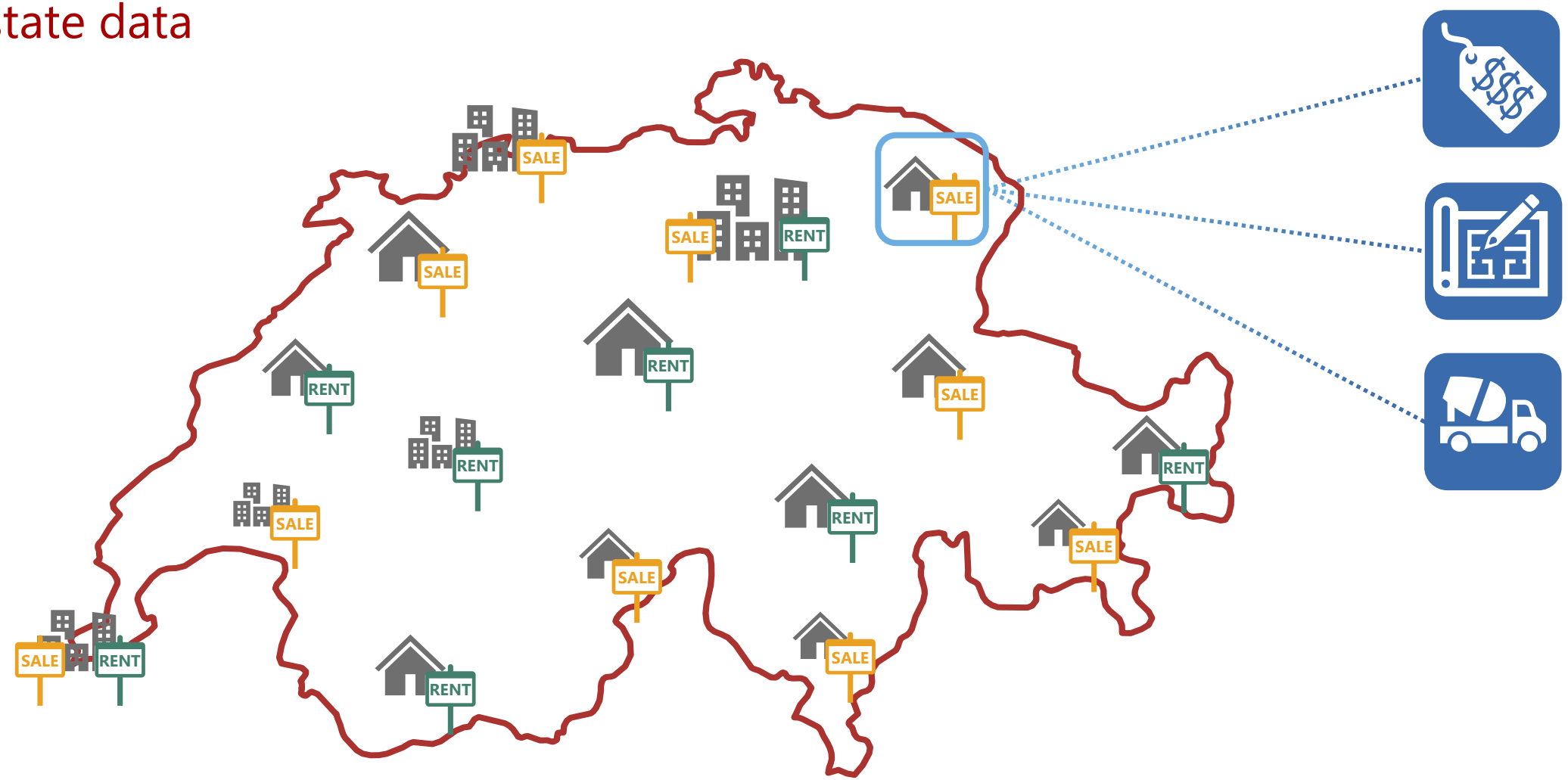
Zürich Airport data



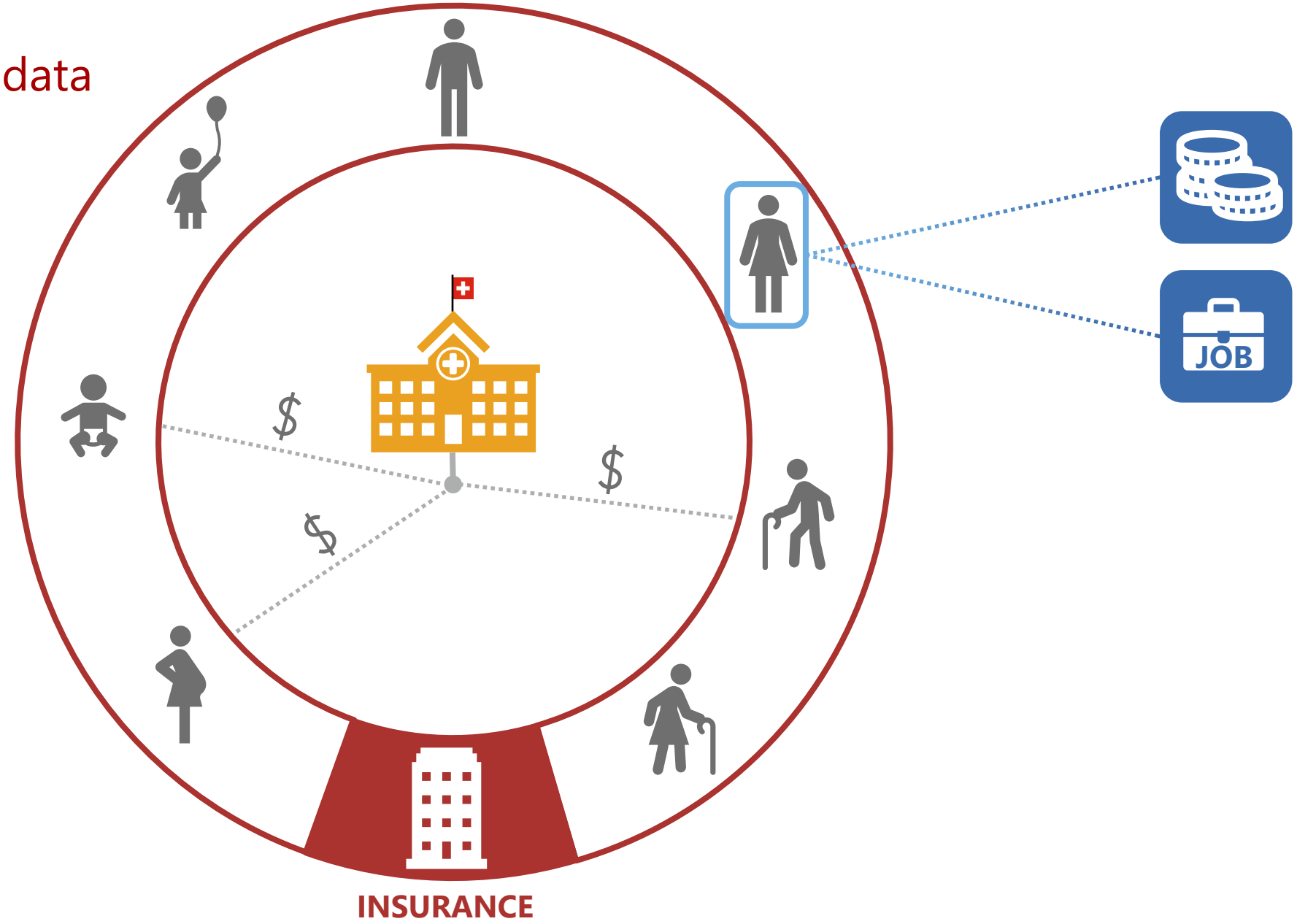
Rail data



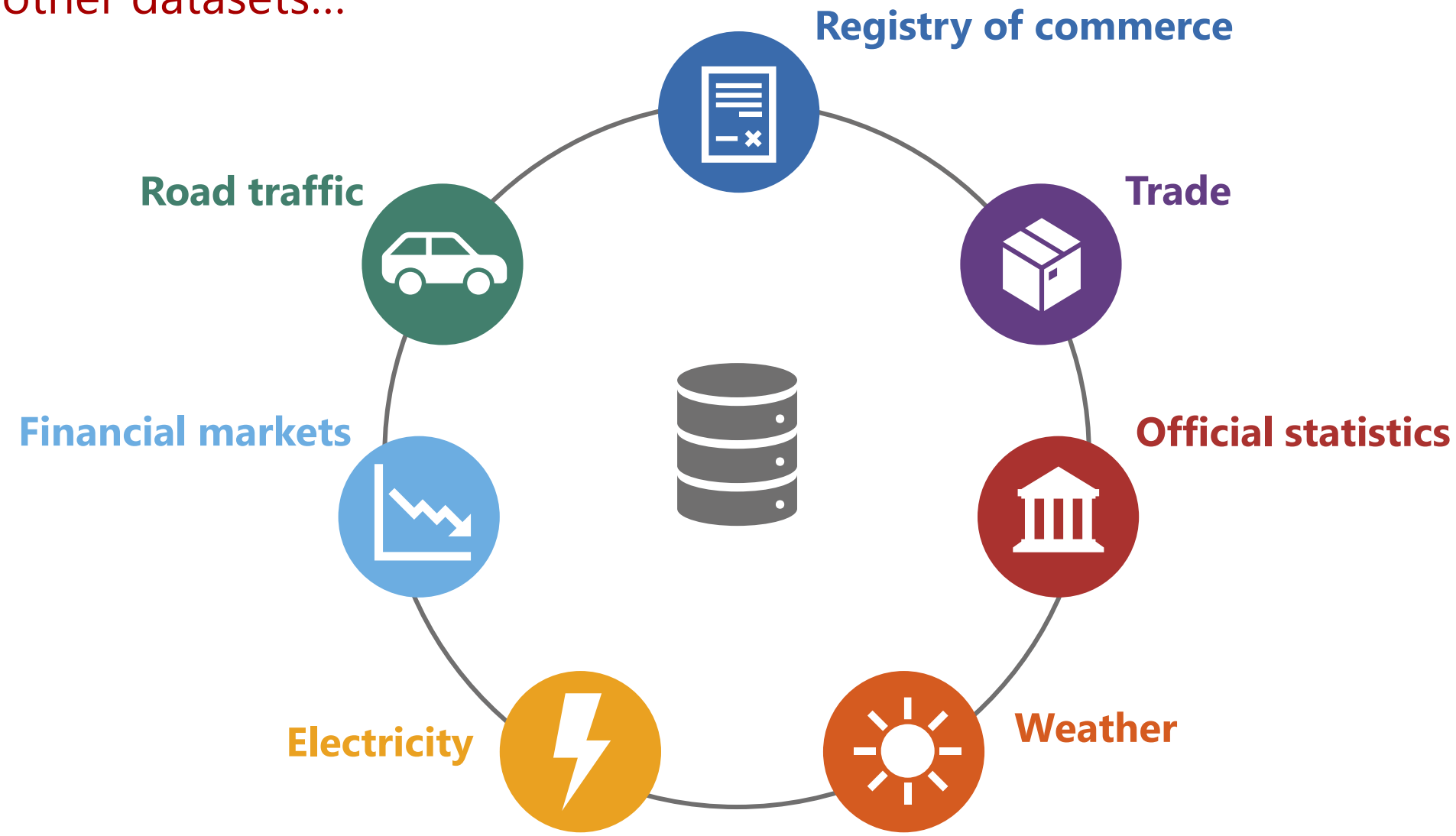
Real estate data



Health insurance data



And other datasets…



BIS Innovation Hub

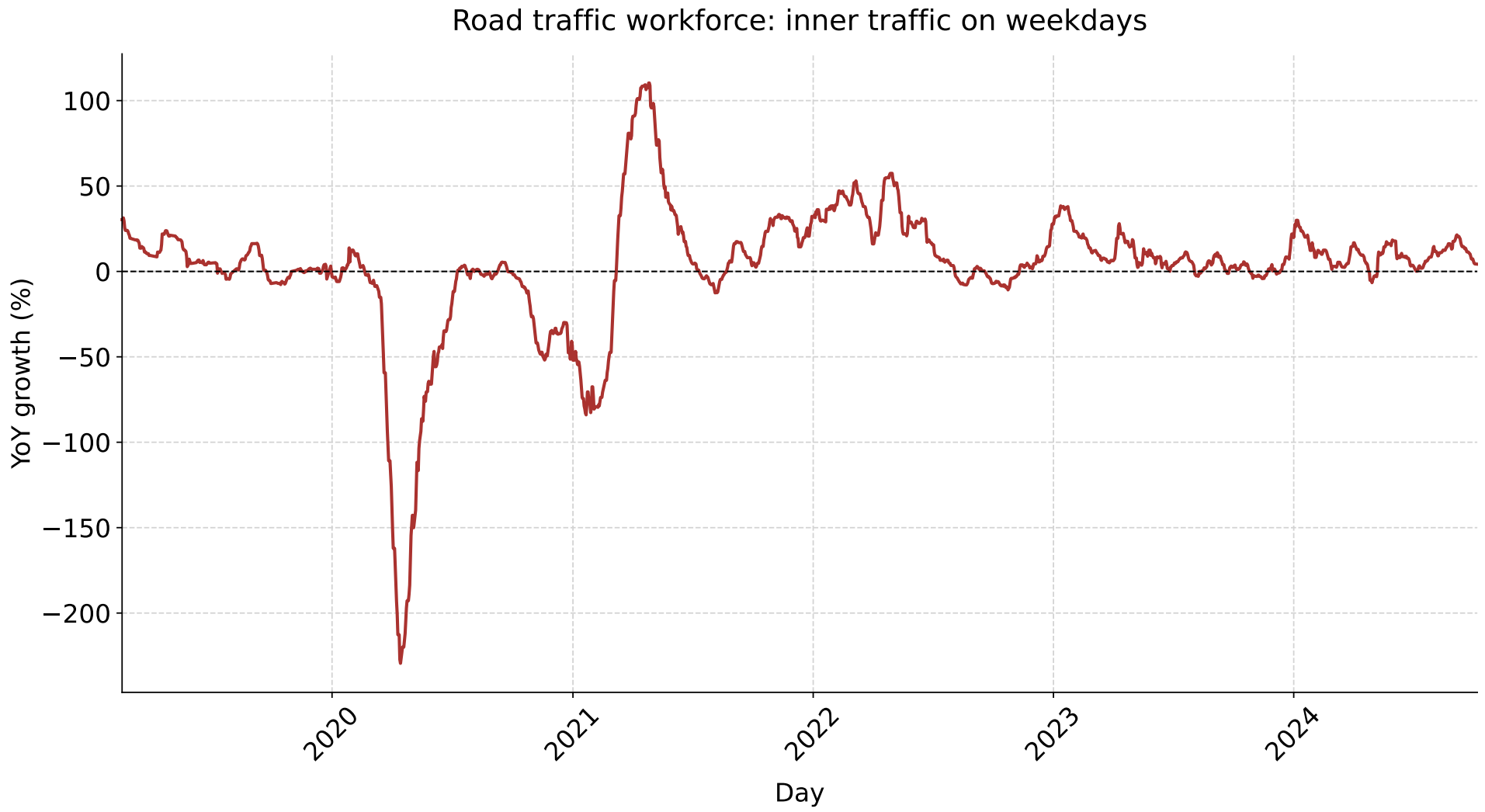


Demo

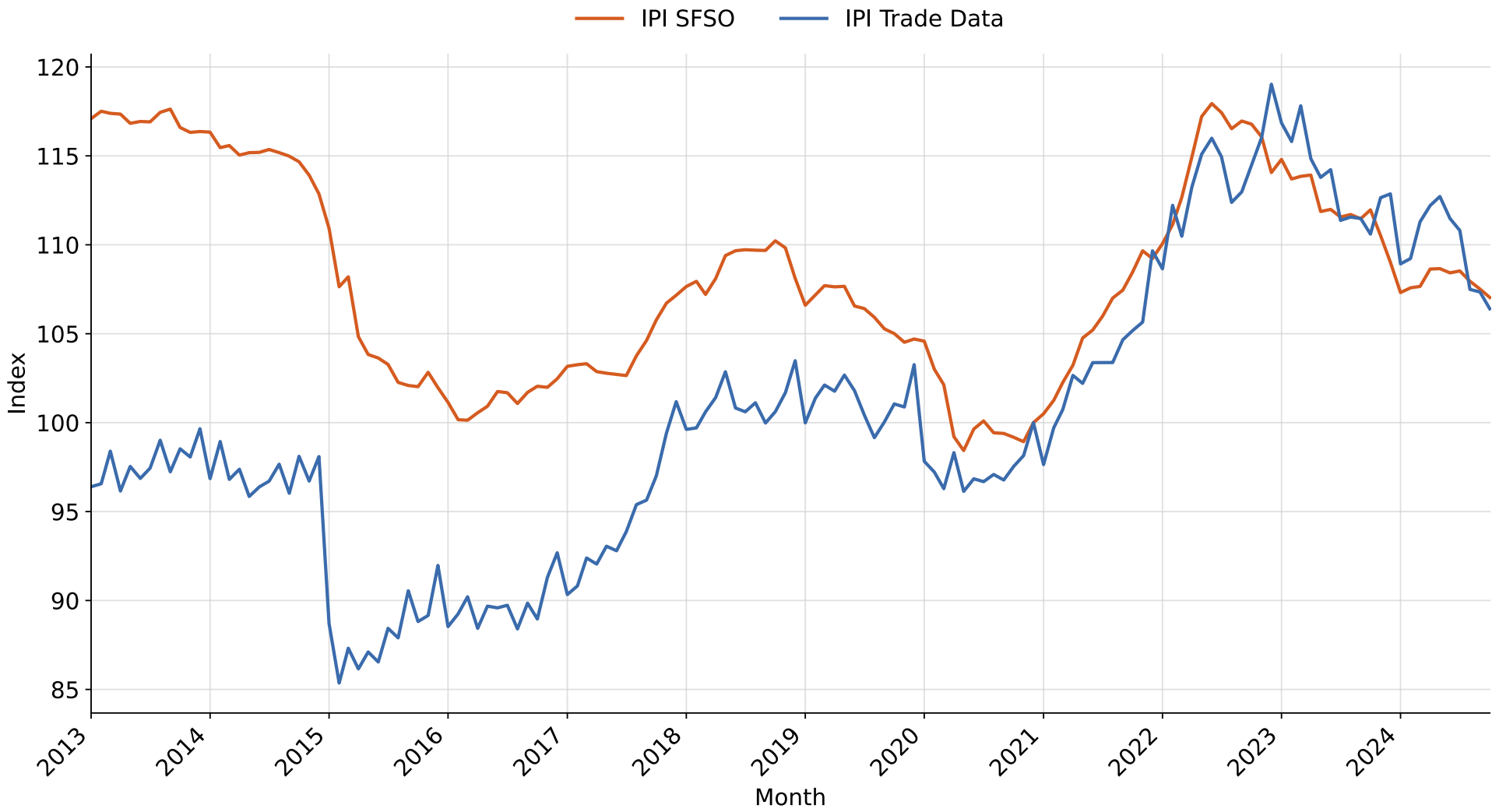
BIS Innovation Hub

Appendix

Example of features engineering not suitable for economic indicators



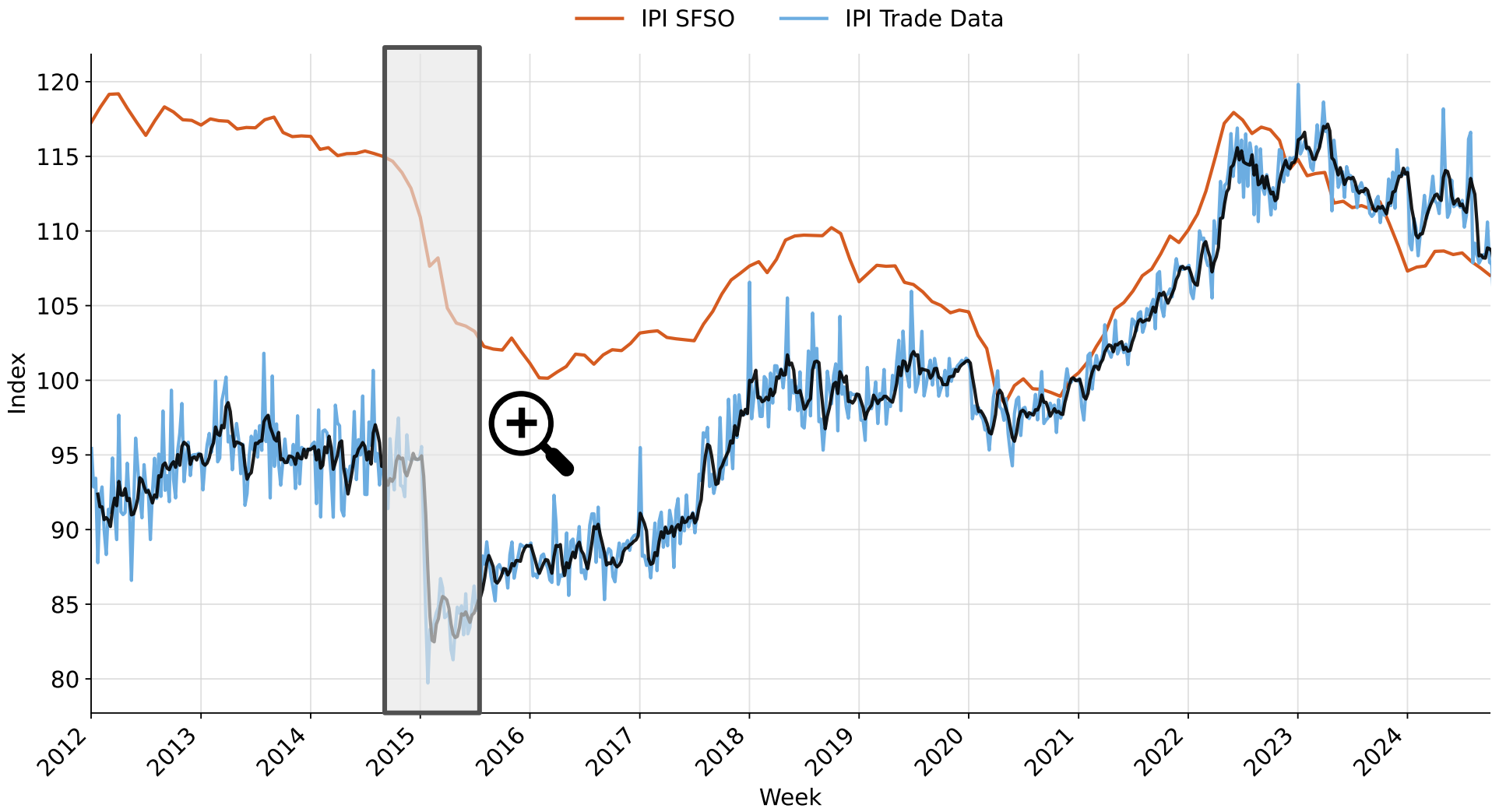
Economic indicator: Import Price Index (IPI) using trade data



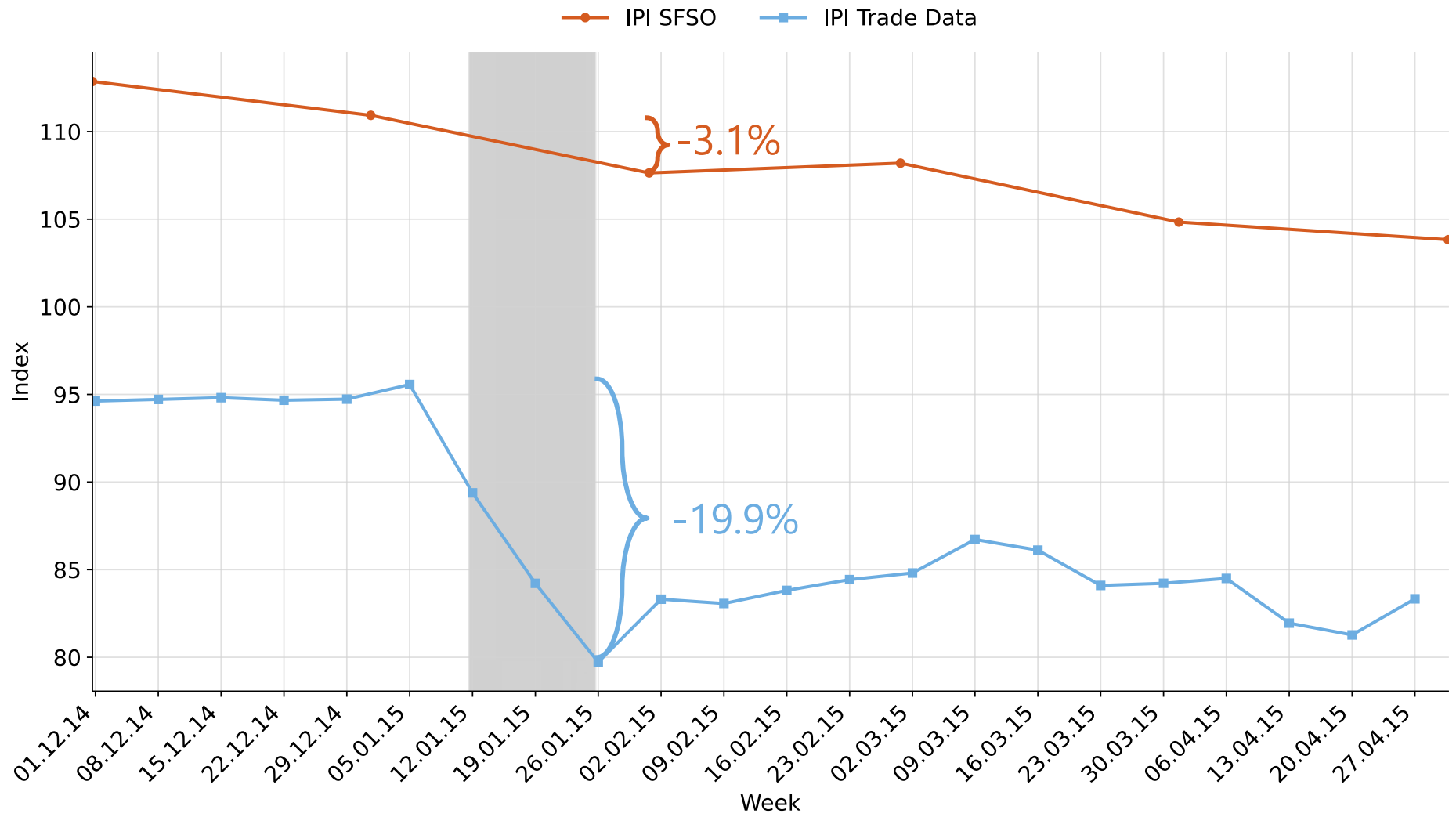
Why is trade IPI useful?

Official IPI		Trade IPI
Measure the prices of carefully chosen imported items, constant over time		Official IPI is the "truth" → **Noisy signal** of official IPI
Monthly frequency and published on the 3rd week of following month		Weekly frequency with 2-week lag → **Available** up to **5 weeks earlier**
Use surveys and other estimations to measure prices that takes time		Use directly the full population of paid imported prices → **Leading** official IPI **by ~3 months**
Some sub-indices are published and no particular link to CPI		Any granularity sub-indices or export indices can be built → related to and **leading CPI**

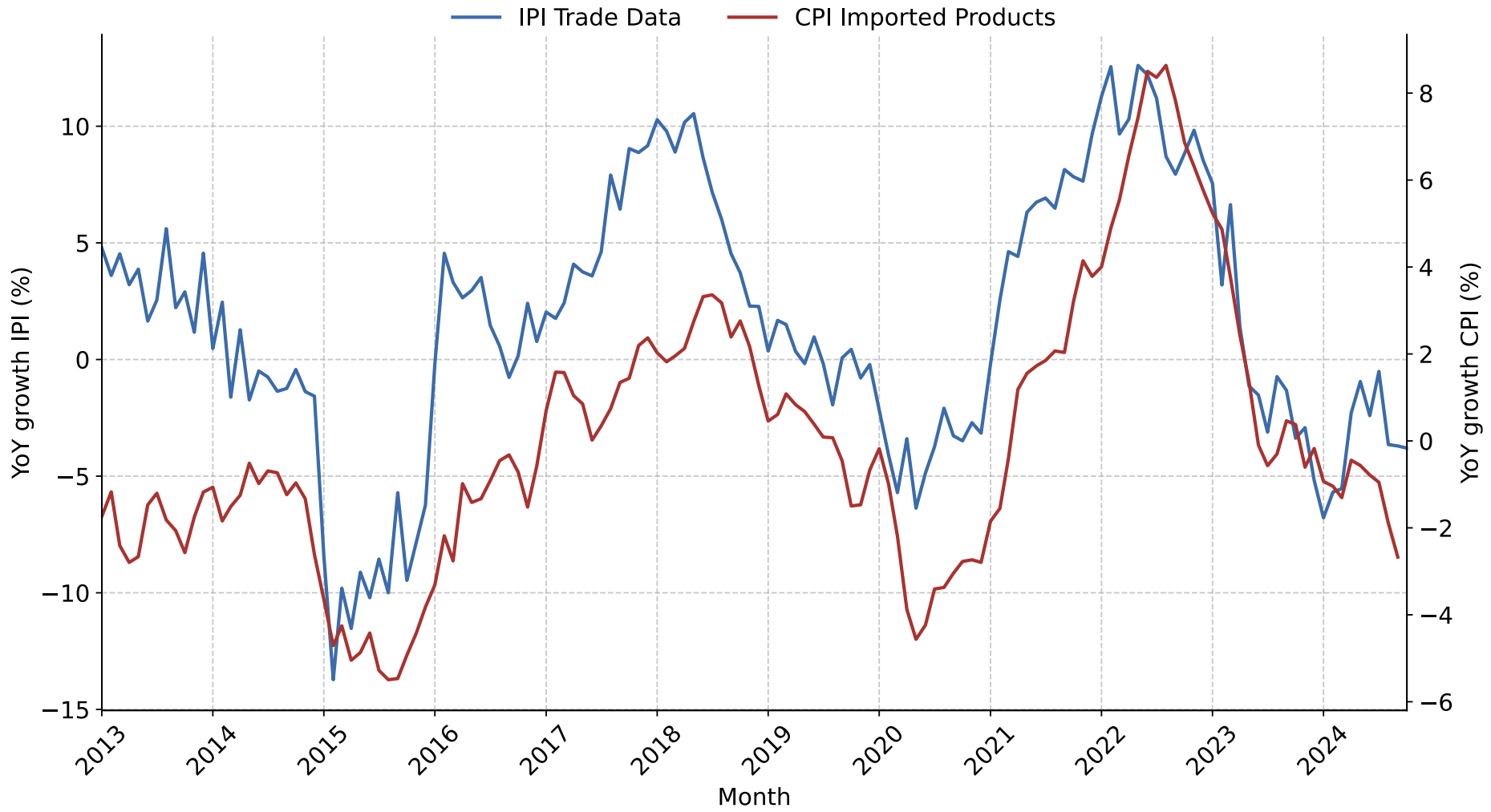
Trade IPI: early weekly noisy signal leading official IPI



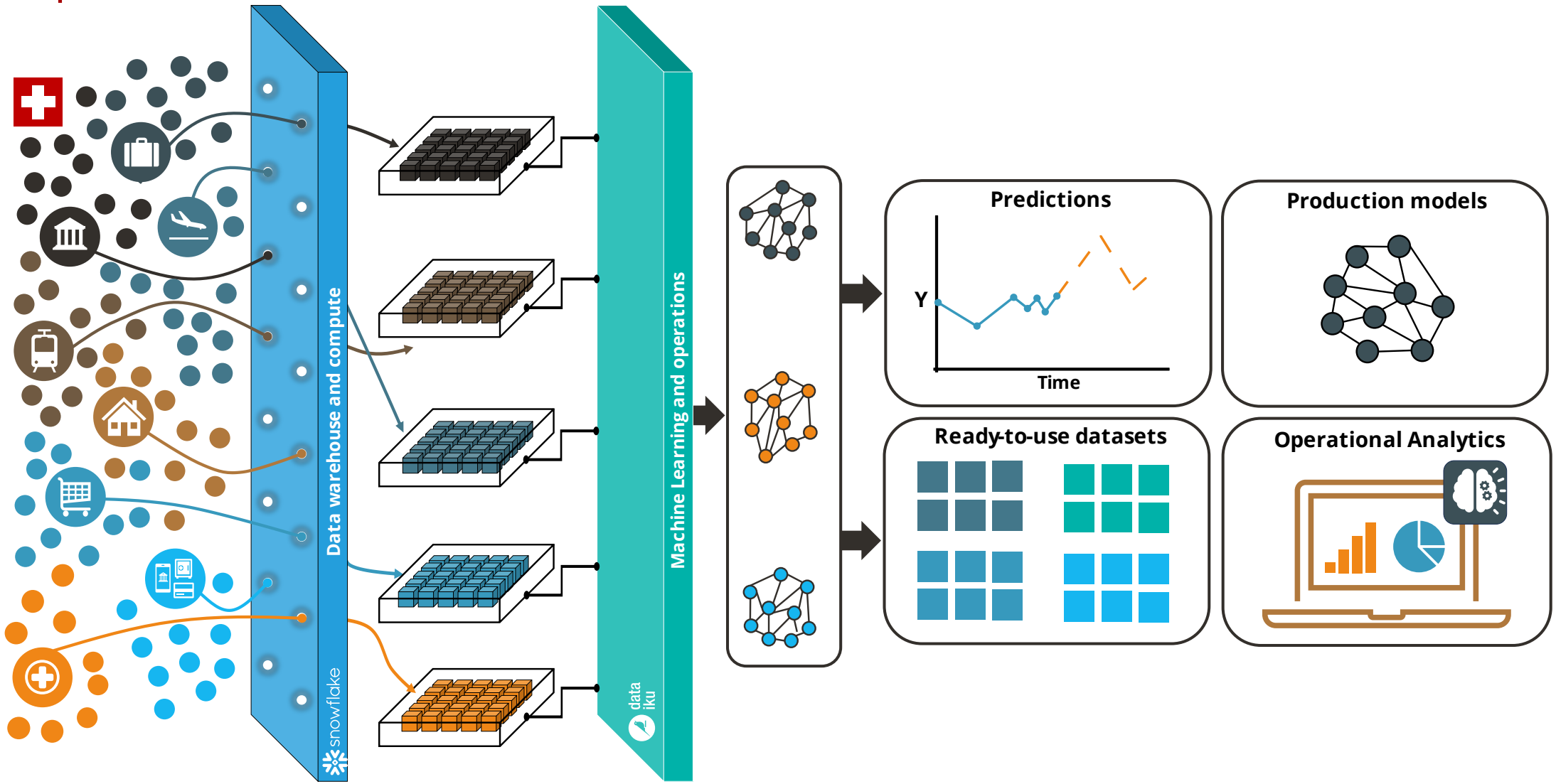
Zooming in on the discontinuation of the minimum CHF/EUR exchange rate



Leading imported goods CPI by 2 months & 0.8 correlation



Pipeline


Data warehouse and compute
snowflake
Machine Learning and operations
data iku
Predictions
Y
Time
Production models
Ready-to-use datasets
Operational Analytics

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Enhancing external debt nowcasting: a data-driven approach using advanced analytics[1]

Candra Puspita Dewi, Aditya Wisnugraha Sugiyarto and Sony Panji Wicaksono,
Bank Indonesia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Enhancing external debt nowcasting: a data-driven approach using advanced analytics

Candra Puspita Dewi[1], Aditya Wisnugraha Sugiyarto[2], Sony Panji Wicaksono[3]

Abstract

External debt is a critical component of a country's Balance of Payments (BoP), reflecting the financial transactions between residents and non-residents that are vital for economic activities and policymaking. In emerging economies like Indonesia, external debt plays a crucial role in supplementing domestic resources for investment and growth. Therefore, accurate forecasting of debt disbursement and repayment is essential for effective economic planning and management. However, current methods face challenges due to insufficient data sources and the unreliability of projections for future transactions. This paper introduces an advanced analytics approach to nowcasting external debt by integrating time series analysis with machine learning techniques. By leveraging historical debt transaction data, a granular nowcast model is developed, which identifies patterns in disbursement and repayment across various debt types, creditors, and corporate sectors. Time series analysis allows for precise future predictions, while machine learning groups transactions based on their historical patterns and characteristics. This combined approach significantly improves the accuracy of short- and medium-term external debt forecasts, offering a more reliable alternative to conventional methods. Our initial findings show that this approach can effectively project patterns in external debt transactions, especially in trade and other loan. The enhanced nowcasting model contributes to more informed decision-making in external sector management and strengthens the accuracy of Indonesia's BoP analysis.

Keywords: balance of payments, time series analysis, external debt, nowcasting, machine learning

JEL classification: C82, E58, F14, O19

[1] Department of Statistics, Bank Indonesia, email: candra_pd@bi.go.id

[2] Department of Statistics, Bank Indonesia, email: aditya_wisnugraha@bi.go.id

[3] Department of Statistics, Bank Indonesia, email: sony_pw@bi.go.id

Contents

1. Background 3

2. Literature Review 5

2.1. External Debt Concept 5

2.2. External Debt Statistics in Indonesia 6

2.3. Forecasting of External Debt 7

2.4. Clustering and Time Series 8

2.4.1. Clustering in Time Series Data 8

2.4.2. Challenges in Time Series Clustering 9

2.4.3. Applications and Developments in Time Series Clustering 9

2.5. Hybrid Model of Time Series 10

2.5.1. Advantages and Challenges of Hybrid Models 10

2.5.2. Applications of Hybrid Models 10

3. Methodology 11

3.1. Data Source 11

3.2. Data Processing 12

4. Result and Discussion 16

4.1. Data Analysis 16

4.2. Data Preprocessing 17

4.2.1. Data Preprocessing for Clustering-based Hybrid Model 17

4.2.2. Data Preprocessing for Decomposition-based Hybrid Model 19

4.3. Data Modelling and Evaluation 20

4.3.1. Clustering-based Hybrid Model on Trade Credit Loans 20

4.3.2. Decomposition-based Hybrid Model on Other Loans 24

4.3.3. Comparison to Existing Method 26

5. Conclusion and Future Works 27

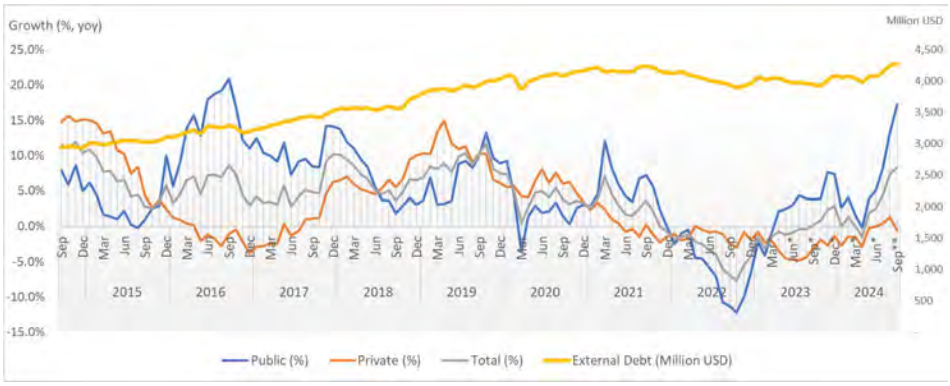
References 28

1. Background

External debt (ED) is one of the primary sources of financing frequently used by emerging countries. Naturally, emerging countries prioritize economic growth, which requires substantial capital investments. However, emerging countries often face challenges such as current account deficits, insufficient domestic savings to meet investment needs (saving-investment gap), and fiscal imbalances. These conditions contribute to the increase of external debt as a vital financing option (Dawood, 2021).

Indonesia, as an emerging country, experiences similar conditions. The increased of external debt over the past decade prove this condition. Indonesia's external debt grew from USD 293.8 billion at the end of 2014 to USD 427.8 billion by the end of third quarter 2024 (Fig. 1). This growth stems from increases in both public and private external debt, with an average annual growth rate of 4.0% (*y-o-y*). After a contraction period during the COVID-19 (2020-2022), Indonesia's external debt has resumed growth post-pandemic. In terms of proportion, the proportions of public and private external debt remain relatively balanced.

Indonesian External Debt Growth (yoy) Figure 1



Source: Indonesian External Debt Publication (September 2024)

External debt also plays an essential role in the Balance of Payments (BoP). External debt transactions are part of the financial account liabilities, besides equity. These transactions appear in the form of direct investments, portfolio investments, or other investment. In the BoP framework, external debt disbursement often offset current account deficits, thereby maintaining the overall balance of payments stability (Bulut, 2011).

Indonesia's BoP have similar patterns. From 2014 to 2020, Indonesia's current account deficit was offset by financial account surplus (Fig. 2). This trend reversed slightly during the pandemic, as the current account recorded a surplus driven by reduced imports due to the decline in economic activity. Additionally, the rise in commodity prices bolstered export values, further contributing to the current account surplus. Conversely, the financial account experienced a deficit because of capital outflows in line with the global economics uncertainties. This anomaly ended post-pandemic, with current account returning to deficit, balanced by surplus in financial account. In addition, external debt forms 55% of financial account liabilities, surpassing equity (Fig.3).

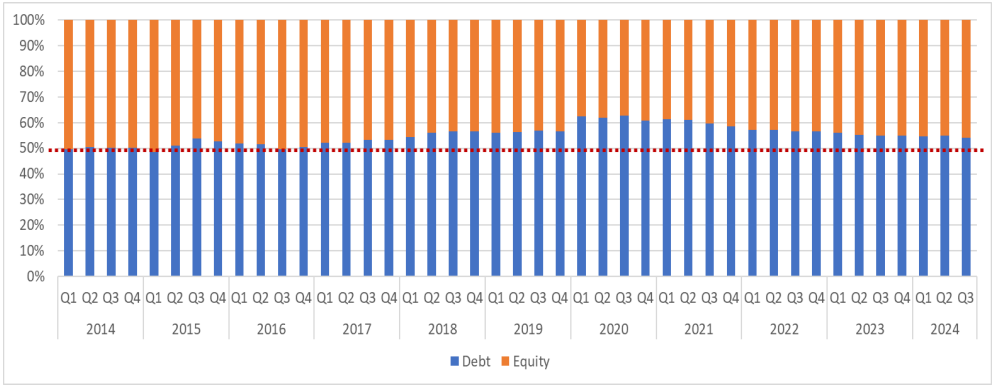
Indonesian Balance of Payment Figure 2



Source: Indonesian Balance of Payment (Q3-2024)

The substantial share of external debt in financing strengthens the needs of debt sustainability monitoring. This not only monitoring external debt transaction realization but also projecting future disbursements and repayments. Such projections are crucial for support preventive policies to ensure external debt sustainability and for estimating foreign exchange demands.

Indonesian Liability on International Investment Position Figure 3



Source: International Investment Position Publication (Q3-2024)

Forecasting foreign exchange demands aligns with the role of Bank Indonesia as the central bank, whose primary objective is to achieve and maintain Rupiah stability. To achieve this objective, Bank Indonesia conduct various strategies that require reliable information, including projections of foreign exchange demands.

Bank Indonesia projections for external debt disbursements and repayments is a part of foreign exchange demands forecasting. For public external debt, the method is by doing coordination with the Ministry of Finance to estimate disbursements and repayments for government loans and bonds. For private external debt, projections rely on reports submitted by corporations. While projections for public external debt are considered reliable due to the dependable data source from Ministry of Finance, private external debt projections are considered less reliable because of dependency on corporate reports. Therefore, there is a need to enhance methodologies for

projecting private external debt disbursements and repayments using alternative approaches in addition to corporate reports.

This paper attempts to adopt a data-driven approach to forecast private external debt disbursements and repayments. Using historical data, we categorize corporations based on their behavior. Clustering is also conducted by grouping corporate external debt types (e.g., loans, debt securities, trade payables, and other debts). Subsequently, time series modeling is applied to each cluster to forecast future external debt disbursements and repayments, a process referred to as a cluster-based hybrid model. This clustering-based approach ensures that similar patterns within each group are captured effectively, enhancing the model's predictive accuracy.

The study focuses on Trade Credits (TC) and Other Loan, as these loan types exhibit clearer and more consistent patterns compared to other categories like loans and debt securities, which are often influenced by irregularities or market volatility. TC and Other Loan are typically tied to operational and trade cycles, making their behaviors more predictable and suitable for hybrid modeling. Concentrating on these loans enhances model accuracy and provides actionable insights for decision-making.

Additionally, when data does not meet the required granularity for clustering, the hybrid framework is adapted to include time series decomposition. This method breaks down the data into Trend, Seasonality, and Residual components, enabling a component-wise analysis to further refine predictions. In cases where decomposition or clustering is not feasible due to data limitations, the analysis is performed at an aggregate level to maintain forecast reliability. Hybrid modeling, including the cluster-based approach and decomposition techniques, which has been extensively studied and shown to outperform classical methods, forms the backbone of this study, ensuring robust and flexible predictions across varying data conditions.

This paper consists of five sections. In the next section, we review relevant literature on data-driven approaches, clustering, and time series analysis for projection. The third section elaborates on the methodology, data sources, and data processing techniques. The results are presented in section four, while the final section provides the conclusions drawn from this study.

2. Literature Review

2.1. External Debt Concept

Gross external debt is outstanding amount of those actual current account, and not contingent, liabilities that require payment(s) of principal and/or interest by debtor at some point(s) in the future and that are owed to non-residents by resident of an economy. Principle amount is the amount that debtor owed to creditor, while principal amount payment is all payment by debtor to creditor that reduce the principal amount outstanding. To be categorised as external debt, the debtor must be a residence and the creditor must be a non-resident. Residence is an economic unit, either a Person or an Entity domiciled or plans to reside in a country for at least 1 (one) year, whose main centre of economic activity is in the country of domicile (IMF, 2014).

External debt statistics presentation is based on the institutional sector. The institutional sector presentations are consistent with the Systems of National Account (SNA) 2008 and Balance of Payments and International Investment Position Manual (BPM) 6th Edition. There are 4 institutional sectors:

1. General government

The general government units of a country consist of those authorities and their agencies that are units established by political processes and exercise legislative, judicial, and executive authority over institutional units within a given territorial area.

2. Central bank

The central bank is a financial institution (or institutions) that exercise control over key aspects of financial system.

3. Deposit taking corporation except central bank

This includes all resident units engaging in financial intermediation as principal activity.

4. Other sectors

This last category consists of other financial corporation (other from deposit taking corporation), non-financial corporation, households, and nonprofit institution serving households.

Other statistics presentation is based on financial instruments. There are 5 financial instruments categories, which is debt securities, loans, currency and deposits, trade credit and advances, and special drawing rights (SDRs).

2.2. External Debt Statistics in Indonesia

Indonesian External Debt Statistics is a joint publication between Bank Indonesia (BI) as Indonesian Central Bank and Ministry of Finance (MoF). While MoF responsible for government external debt, BI responsible for central bank and private sector external debt. For institutional presentation, Indonesia External Debt Statistics consists of government, central bank and private sector external debt.

Government external debt refers to the debt held by the central government, consisting of bilateral/multilateral debt, export credit facilities, commercial debt, and leasing, including government securities (SBN) (issued in Indonesia and abroad) held by non-residents. SBNs include Government Debt Securities (SUN) and Government Sharia Debt Securities (SBSN). SUN consists of government bonds lasting more than 12 months and Treasury Bills (SPN) with up to 12 months of tenures. SBSN consists of long-term SBSN (Ijarah Fixed Rate – IFR) and Global Sukuk.

Central Bank external debt refers to the obligations held by Bank Indonesia to support the balance of payments and reserve assets. The debt includes liabilities in the form of securities issued by Bank Indonesia and held by non-residents and non-residents deposits at Bank Indonesia.

Private external debt refers to the obligations of residents (excluding the government and central bank) to non-residents in a foreign currency and/or the rupiah based on loan agreements or other agreements, deposits, and other liabilities. Private external debt includes liabilities in the form of debt securities issued by

residents and held by non-residents. The private sector comprises banks and non-banks. The private non-bank consists of Non-Bank Financial Institutions and Non-Financial Corporations (NFC).

Subsequently, Indonesia employs a comprehensive approach to categorizing its external debt by instruments, which includes debt securities, loans, trade credit, and other loans. This classification system allows for a nuanced analysis of the country's financial obligations, facilitating more effective risk management and policy development. By breaking down external debt into these specific categories, policymakers can better assess the nature and structure of the country's liabilities, enabling them to devise targeted strategies to manage and mitigate associated risks.

Furthermore, Indonesia has disclosed Special Drawing Rights (SDRs) as part of the central bank's external debt. SDRs are international reserve assets created by the IMF and allocated to members to supplement existing official reserves. SDR allocations are considered long-term liabilities, with repayment occurring if a member country decides to withdraw from the IMF or if the IMF's SDR Department is liquidated.

2.3. Forecasting of External Debt

Forecasting external debt is a critical component of economic planning and financial stability for any country. An accurate forecasting involves analysing historical data, current economic conditions, and potential future scenarios. The primary goal of forecasting external debt is to predict the future levels of debt and assess the sustainability of current borrowing practices. This process helps policymakers make informed decisions about fiscal policies, debt management strategies, and economic reforms to ensure long-term financial stability (Cuddington, 1997; Reinhart & Rogoff, 2010).

One key methodology in forecasting external debt is econometric modelling, which uses indicators like GDP growth rates, interest rates, exchange rates, and trade balances. Advanced techniques such as vector autoregression (VAR) and structural equation modelling (SEM) capture dynamic relationships between these variables. Scenario analysis further evaluates the impact of different economic conditions on external debt, helping policymakers understand potential risks and develop contingency plans (Chowdhury, 2001).

Nowcasting enhances real-time assessment by using high-frequency data and advanced statistical methods to provide near real-time estimates of economic indicators. This approach allows for timely adjustments to debt management strategies in response to rapidly changing conditions. By integrating nowcasting with traditional forecasting models, central banks can better predict and manage external debt (Banbura et al., 2013). The nowcasting process involves the following steps:

1. Data Collection: Gather high-frequency data from various sources, such as financial markets, trade statistics, and other economic indicators.
2. Modelling: Use statistical models, such as dynamic factor models (DFMs) or Bayesian vector autoregressions (BVARs), to process and analyse the data.
3. Estimation: Apply these models to estimate current economic conditions and predict near-term future trends.

4. Updating: Continuously update the models with new data as it becomes available to refine and improve the accuracy of the forecasts.

By integrating nowcasting techniques with traditional forecasting methods, policymakers can make more informed decisions and respond more swiftly to changing economic conditions.

2.4. Clustering and Time Series

Clustering is a fundamental technique in data analysis, designed to group similar objects into homogeneous clusters without the need for prior labels. Clustering in the realm of time series analysis, is employed to organize complex temporal data, enabling the identification and utilization of hidden patterns. Time series data possess unique characteristics, such as temporal order, high dimensionality, and varying lengths, which pose significant challenges for conventional analytical methods (Chen, 2020; Aghabozorgi et al., 2015; Jorge & Ruben, 2024).

Additionally, clustering serves as an invaluable tool to simplify the complexity when dealing with granular data where each observation exhibits its own patterns and uniqueness, and where it is difficult to discern aggregate time series patterns. This process facilitates the grouping of similar observations, providing an initial structure that is advantageous for subsequent analyses, including the development of predictive models and classification tasks.

2.4.1. Clustering in Time Series Data

Time series clustering can be classified in various approaches:

1. Shape-Based Clustering

This approach directly compares the shapes of time series without additional modifications. It relies on similarity metrics such as:

- Dynamic Time Warping (DTW), which is highly effective in handling differences in speed and temporal deformations between time series, although it requires high computational costs (Javed et al., 2020; Jorge & Ruben, 2024).
- Euclidean Distance, used in methods like k-Means for data that do not have significant time scale differences (Huang, 2016)

Shape-based clustering is often used in applications with simple time series patterns, such as seasonal patterns in stock prices or climate data.

2. Feature-Based Clustering

This approach transforms time series data into a feature space with more concise and relevant representations, such as:

- Autocorrelation and Spectral Transformation, to identify periodic patterns.
- Wavelet Transformation, which captures high and low-frequency details.

The extracted features are then used by conventional clustering algorithms like K-Means or Hierarchical Clustering. This approach is more robust at handling noise and effective for large-scale datasets.

3. Model-Based Clustering

This model relies on parametric approaches to represent time series, such as Hidden Markov Models (HMM) and ARIMA, which are suitable for data with complex statistical patterns. These parametric models allow for trend-based analysis and long-term predictions, although they have limitations in handling high-dimensional or noisy data (Jorge & Ruben, 2024).

4. Hybrid Approaches

This approach combines shape-based, feature-based, and model-based techniques to leverage the strengths of each method. An example is the R-Clustering algorithm, which uses random convolutional kernels to improve efficiency in handling high-dimensional time series (Jorge & Ruben, 2024).

2.4.2. Challenges in Time Series Clustering

Despite its potential, time series clustering faces the following challenges:

1. High Dimensionality and Large Data Volume. Time series data often consist of hundreds of thousands of data points, requiring representation approaches such as dimensionality reduction to speed up the process (Aghabozorgi et al., 2015; Huang et al., 2016).
2. Length Variability. Differences in the length of time series necessitate approaches like normalization or the use of robust distance metrics such as DTW (Javed et al., 2020).
3. Noise and Outliers. Data imperfections require preprocessing techniques such as smoothing or robust feature extraction (Aghabozorgi et al., 2015).

2.4.3. Applications and Developments in Time Series Clustering

Time series clustering has wide applications across various domains, such as:

1. Finance: Analysis of seasonal patterns, volatility, and stock price predictions (Huang et al., 2016; Jorge & Ruben, 2023).
2. Healthcare: Identification of patterns in ECG signals or breathing patterns (Chen, 2020).
3. Environmental Science: Analysis of weather pattern changes or pollutant concentrations (Aghabozorgi et al., 2015).
4. Technology: Anomaly detection in internet traffic or user behaviour (Jorge & Ruben, 2024).

Clustering serves as an essential preliminary step before modelling, especially when datasets have granular characteristics that make aggregate pattern analysis difficult. Through clustering, the initial structure of the data can be identified to produce more targeted information.

2.5. Hybrid Model of Time Series

The hybrid model in time series analysis is an increasingly popular approach to overcoming the limitations of single models. By combining the strengths of various models, the hybrid approach can enhance the accuracy and robustness of forecasting, especially for data with complex, non-linear, or non-stationary patterns (Xu, et al., 2019).

Hybrid models integrate two or more approaches to address the shortcomings of single models. According to Xu (2019), hybrid models can be divided into three main structures:

1. Parallel Structure: Data is processed simultaneously through multiple models, and the results are combined using specific methods. This structure is suitable for data with many parallel patterns.
2. Series Structure: The first model produces an output that becomes the input for the next model, allowing for sequential processing of linear and non-linear patterns. For example, ARIMA is used to capture linear patterns, while its residuals are processed by a neural network for non-linear patterns.
3. Parallel-Series Structure: A combination of parallel and series approaches to leverage the advantages of both methods.

2.5.1. Advantages and Challenges of Hybrid Models

Hybrid models have proven to be superior to single models in various aspects:

- Higher Accuracy: By combining complementary methods, hybrid models can capture complex patterns, including long-term trends and short-term fluctuations.
- Adaptation to Non-Stationary Data: Models such as Empirical Mode Decomposition (EMD) and Recursive Empirical Mode Decomposition (REMD) are often used to decompose data into simpler components, each of which can be modelled more effectively.
- Multidomain Flexibility: Hybrid models have been used in various fields, including finance, energy, health, and meteorology, demonstrating their adaptability to different types of data and scenarios.

However, despite these advantages, hybrid models also face challenges, including design complexity. It's because the models combining various models requires a complex architectural design and a deep understanding of data characteristics.

2.5.2. Applications of Hybrid Models

These hybrid models have been applied in various fields, including:

- Financial Sector: Stock market and currency price predictions often use a combination of ARIMA with deep learning methods such as LSTM or Deep Belief Network (DBN) to capture complex patterns, including market volatility (Rahman & Pujiati, 2021; Xu et al., 2019), and public sector external debt predictions (Razinkova et al., 2023).

- Energy and Environmental Sector: Forecasting electricity load and renewable energy using hybrid models that combine statistical techniques with neural networks or ensemble methods to capture seasonal fluctuations and long-term trends (Hajirahimi & Khashei, 2019).
- Health Sector: Hybrid models are applied to forecast disease spread, such as COVID-19, using ARIMA for linear patterns and LSTM for non-linear patterns. This integration has helped in better health policy planning (Kontopoulou et al., 2023).
- Meteorological Sector: In weather prediction, hybrid models utilize decomposition methods such as wavelet transform combined with LSTM to capture complex patterns of wind, rain, or temperature (Yang et al., 2021).

3. Methodology

3.1. Data Source

Indonesian External Debt Statistics use various data source, 1) government external debt is compiled using MoF internal data, which consist of outstanding, disbursement, repayment and projection/scheduled external debt; 2) central bank external debt is compiled using BI internal data; 3) private external debt is compiled mainly using data form reporting system which is External Debt Reporting System (EDIS), in addition with data from bank report, data provider such as C-Bonds, and other anecdotal information. In accordance with Bank Indonesia's regulations on the requirement to report external debt, all corporations that have external debt must report their debt to Bank Indonesia through EDIS.

While government and central bank external debt projection data is more reliable, as the result of using institutional data, private external debt projection is more difficult. Currently, EDIS as private external debt main data sources have function to differentiated between scheduled and actual debt repayment. However, EDIS purposes is for data collection, so it doesn't have estimating function. In addition, disbursements and repayments of debt is very related with the company's financial condition, so there can be difference between scheduled and actual debt repayments. This presents a challenge for Bank Indonesia to ensure the accuracy and completeness of the data needed for effective analysis and policy making. In line with that condition, this paper is mainly focus is on the projection of private external debt.

For compiling private external debt, data from EDIS is transform into database named the "realization cube" and the "projection cube". Realization cube contains detailed information on the position values, as well as the realization of principal and interest payments, which are updated monthly. The dimensions of this database are quite complex, including the individual names of corporations, names of creditors, maturity periods, types of debt, ownership status, economic sectors, and original currencies. This comprehensive data structure allows for precise tracking and analysis of private external debt, facilitating effective monitoring and management.

Meanwhile, a projection cube accommodates planned values for disbursements and repayments, as well as the position of debt until maturity. The "projection cube"

dimensions are similar with the "realization cube," ensuring consistency and comparability between planned and actual data. This dual-database approach enhances our ability to forecast and manage external debt, supporting informed decision-making and strategic planning to maintain financial stability. However as mentioned above, the accuracy for planned disbursements and repayments data is still not reliable.

In this paper, we use several fields or variables, namely:

- Company name, as a standard name use by each corporation.
- Loan type, consist of loan agreement, debt securities, trade credit and other loan.
- Outstanding, is the principal debt amount.
- Disbursement, is the amount that corporation withdraw in certain period.
- Repayment, is the amount of payment in certain period.
- Investment status, which categorize into direct investment and non-direct investment.

Furthermore, in this paper we utilize data starting from January 2022 until September 2024. We divide this data into 2 categories, first is data for develop the model or data training. For these purposes we use data from January 2022 until March 2024. Meanwhile for testing or out of sample data, we use data from April 2024 until September 2024.

3.2. Data Processing

This research uses a hybrid model. To perform this data processing, we undertake four stages, that are data analysis, data preprocessing, data modelling, and evaluation.

a) Data Analysis

This step includes exploring the collected external debt data to identify patterns and relationships that can be used in the development of the nowcasting model. For example, this analysis includes examining historical data regarding debt disbursements and repayments based on loan type, investment status and issuer. The primary objective is to understand the fundamental behaviour of external debt transactions. This includes identifying any monthly patterns in disbursements and repayments within each company, as well as determining which entities exhibit similar repayment behaviours to their disbursements within the same month.

b) Data Preprocessing

This stage focuses on preparing raw data before modelling. In the context of this paper, these steps include cleaning up external debt data, handling missing or inconsistent values, and normalizing the dataset.

Additionally, this stage involves selecting key entities based on the Pareto 80/20 principle. Only corporations contributing to 80% of the total external debt (as determined from historical data) are prioritized for further analysis. This approach not only reduces computational complexity but also optimizes the results by concentrating on corporations with the most significant impact on the overall external

debt data. By focusing on these key contributors, the analysis remains both efficient and impactful.

In addition, data normalization is also used to align variable scales to equalize, improve model performance, reduce bias, and facilitate analysis. Min-Max scaling is used as a normalization function in this study with the formula:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The final step in this process is clustering. These techniques are used to group companies based on their external debt behaviour patterns, so that the subsequent analysis and modelling becomes more structured. The clustering technique used is based on Dynamic Time Wrapping (DTW) in accordance with what has been discussed by Aghabozorgi, et. al (2015) with code implementation by Tavenard, et. al (2020).

Once cluster memberships are determined, we apply the "window and horizon" technique to model the time series data. This approach divides the time series into overlapping windows of a fixed length, which are used to train the model. The horizon represents the forecasting period beyond the observed window, where the model predicts future values. This method captures both local and global patterns within the time series, enabling it to handle complex and non-linear relationships.

This technique is chosen because, as supported by previous research (e.g. Sugiyarto & Abadi (2019), Chou & Ngo (2016), Zhou et. al (2018)), it has demonstrated superior performance compared to classical time series analysis. By focusing on smaller windows, the model can adapt to localized trends and changes, while the overlapping structure ensures continuity and consistency in the predictions. This combination enhances the model's accuracy and robustness, particularly when dealing with the variability and complexity inherent in external debt data. Through this approach, we aim to achieve more reliable and precise forecasting outcomes.

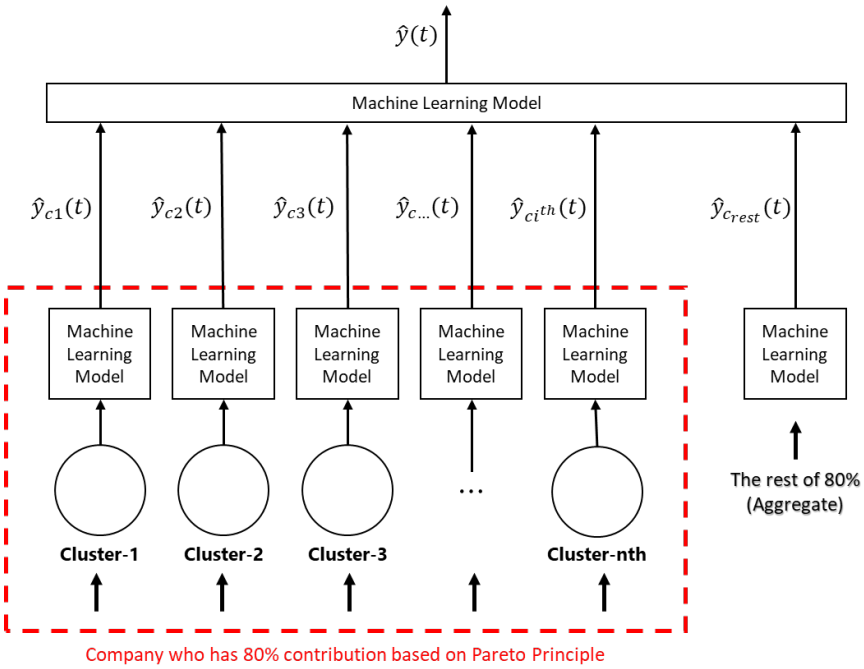
c) Data Modelling

In this step, the clustering method and time series analysis are combined to develop a hybrid nowcasting model. There are 2 (two) types of hybrid model in this paper: 1) Cluster-based hybrid model and 2) Decomposition-based hybrid model.

In cluster-based hybrid model, clustering is used to organize debt data into homogeneous groups, thus allowing the application of a specific time series model for each group. This hybrid approach aims to improve prediction accuracy by leveraging the power of both methods. The scheme of this hybrid model is explained in (Fig. 4):

Clustering-based Hybrid Model

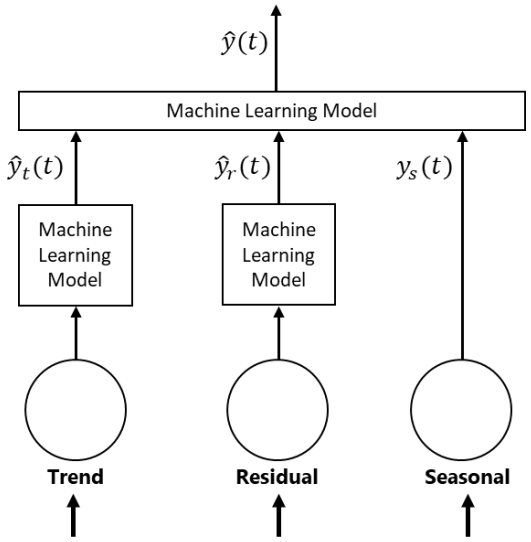
Figure 4



Modelling using machine learning consists of two levels (Fig. 4). At the first level, machine learning is used to model aggregate data from each cluster, which has been grouped based on the characteristics of similar corporation. This ensures that each corporation is modelled specifically according to its unique behaviour. In the second level, machine learning is used to process the prediction results of each first-level model. The goal is to produce more accurate final predictions by reducing errors, which could potentially arise if the results of the first level are simply summed up without further processing.

While the decomposition-based hybrid model schemes presented in (Fig. 5):

Decomposition-based Hybrid Model Figure 5



In the (Fig. 5) illustrates a hybrid modelling framework using decomposition of time series. The time series is decomposed into 3 (three) components: (i) trend, (ii) seasonality, and (iii) residual. Each component is modelled separately using specialized machine learning models tailored to capture its unique characteristics. The outputs of these component-specific models are then combined by a unified machine learning model in the top layer to produce the final prediction. This modular approach enhances prediction accuracy by addressing the distinct behaviours of each time series component.

Additionally, during the modelling process, we employ Grid Search optimization for hyperparameter tuning to enhance the performance of our models. This approach ensures that the most suitable parameter combinations are identified, optimizing the accuracy of the forecasts. In this study, we utilize a range of algorithms, including Random Forest (RF), Support Vector Regression (SVR), and various linear models such as Lasso Regression, Ridge Regression, and Linear Regression, to forecast and nowcast external debt disbursements and repayments effectively. This diverse selection of algorithms enables the model to capture complex patterns and relationships within the data, ensuring robust and reliable predictions.

d) Evaluation

The last stage is evaluation. This stage assesses the performance of the nowcasting model using metrics such as mean absolute percentage error (MAPE). This evaluation ensures that the model can provide accurate short-term and medium-term predictions for external debt disbursement and repayment. The formula for the metric is as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

$\hat{y}_i$: predicted value for i-th data point

y_i : actual value for the i-th data point

n : number of observations

To conclude whether the model is reliable or not, Lewis (1982) divide MAPE criteria as follows (Table 1):

MAPE Criteria for Model Evaluation (Lewis, 1982) Table 1

MAPE	Forecasting Power
<10%	Highly accurate forecasting
10% - 20%	Good forecasting
20% - 50%	Reasonable forecasting
>50%	Weak and inaccurate forecasting

In this paper, we exclusively use MAPE as the evaluation metric because it provides several advantages over other metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Unlike MSE and RMSE, which are sensitive to the range and scale of the data, MAPE expresses the error as a percentage, offering a scale-independent and more interpretable measure. This makes MAPE particularly useful for comparing models across datasets with varying scales and for communicating results to non-technical stakeholders, as it provides a clearer understanding of the model's performance (Myttenaere, B., Grand, & Rossi, 2016).

While MSE and RMSE emphasize larger errors due to their squaring nature, making them more sensitive to outliers, MAPE provides a balanced view by treating all errors proportionally to their actual values. This makes it better suited for cases where understanding relative performance is crucial. Additionally, MAPE is more intuitive, as it reflects the average percentage error, enabling easier interpretation for stakeholders who may not have a technical background.

4. Result and Discussion

4.1. Data Analysis

In order to enhance external debt nowcasting, the first step involves analysing the values of external debt payments and disbursement to determine how many debtors exhibit disbursement and repayment patterns that align with the “projection cube” dataset. This external debt analysis will be conducted based on different instruments, namely debt securities (DS), loans, trade credit (TC), and other loans.

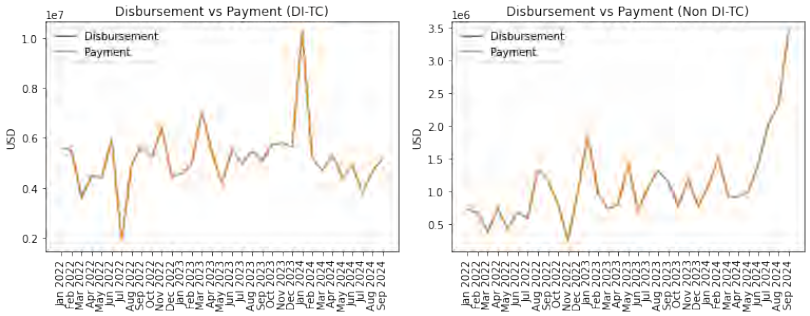
The results indicate that trade credit and other loans have consistent repayment and disbursement patterns, whereas debt securities and loans show inconsistent patterns. Considering these findings, this paper will focus on analysing the instruments of trade credit and other loans.

In analysing TC, there are debtors with identical repayment patterns to their disbursement patterns. This pattern suggests several interesting possibilities for further examination. First, the similarity in values may reflect consistency in the repayment patterns of the debtors, indicating financial stability in fulfilling liabilities on schedule. Second, it may also indicate that the transactions are short-term, specifically to meet temporary operational liquidity needs. Additionally, the

characteristics of loan types such as TC tend to follow trade cycles. Such patterns have important implications for the development of external debt projection models, particularly in predicting future repayment and disbursement needs. Therefore, identifying entities with similar characteristics can support the accuracy of the nowcasting process and more effective external debt management. In this analysis, there are 4 (four) debtors from the DI-TC group and 9 (nine) debtors from the Non-DI-TC group with such patterns. The examples of these patterns are presented in (Fig. 6):

Examples of Identical Disbursement and Repayment Patterns

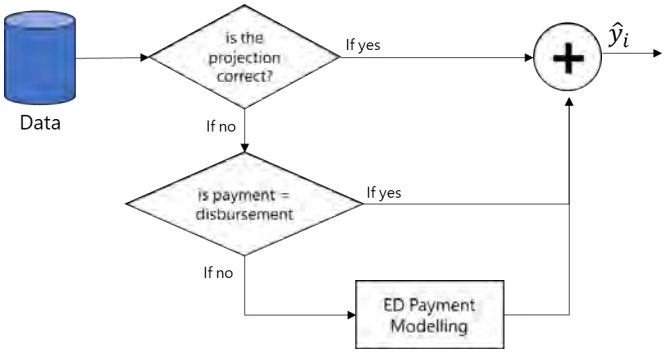
Figure 6



Based on the results of the data analysis, the modelling process will be carried out according to the following scheme (Fig. 7):

Modelling scheme (a) Repayments and (b) Disbursement based on data analysis results

Figure 7



1.2. Data Preprocessing

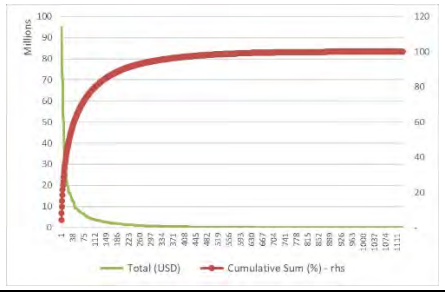
1.2.1. Data Preprocessing for Clustering-based Hybrid Model

At the data preprocessing stage, an analysis based on Pareto principle was conducted, resulting the following results:

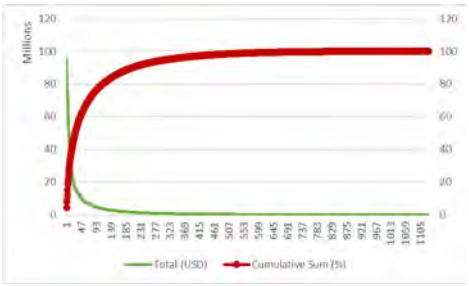
Example of Pareto Calculation Results (a) DI-TC Disbursements (b) DI-TC Repayments

Figure 8

(a)



(b)

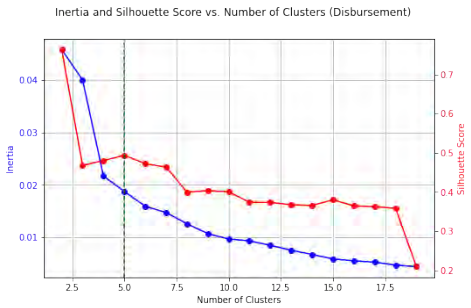


Based on Figure 8, according to the Pareto principle, the data analysis will focus on debtors who contribute 80% of the total external debt. The debtors contributing 80% for the DI-TC loan type (disbursement) are 113 debtors, Non-DI-TC (disbursement) are 115 debtors, DI-TC (payment) are 114 debtors, and Non-DI-TC (payment) are 115 debtors. After that, data normalization is performed to align variable scales to equalize, improve model performance, reduce bias, and facilitate analysis. The next step is the clustering process using DTW, selected based on the Within-Cluster Sum of Squares (WCSS) and Silhouette Score values. The results for the WCSS and Silhouette Score TC are as follows:

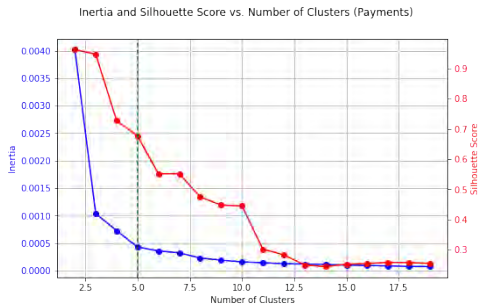
Example of Selecting the Optimal Number of Clusters for DI-TC (a) DI-TC Disbursements (b) DI-TC Repayments

Figure 9

(a)



(b)



In determining the best cluster, we use the WCSS and Silhouette score. The Silhouette score measures how well each point in a cluster fit with its own cluster compared to other clusters, with a value range of [-1,1]. Higher scores indicate better clusters. Meanwhile, the WCSS value measures the total squared distance between points within a cluster and the cluster centre. For selecting the best cluster with WCSS, the Elbow Method is used, where the elbow point on the WCSS graph indicates the optimal number of clusters. Based on data processing, the best clusters for TC payments and disbursements are presented on (Table 2):

Best Cluster Search Results

Table 2

		# Best Clusters	Silhouette Score	WCSS Score
Disbursement	DI-TC	5	0.4936	0.0187
	Non-DI-TC	4	0.7000	0.0007
Repayment	DI-TC	5	0.4206	0.0202
	Non-DI-TC	6	0.4668	0.0137

Based on Table 2, the optimal number of clusters is around 5. Next, aggregate modelling will be performed for each cluster and then aggregated overall.

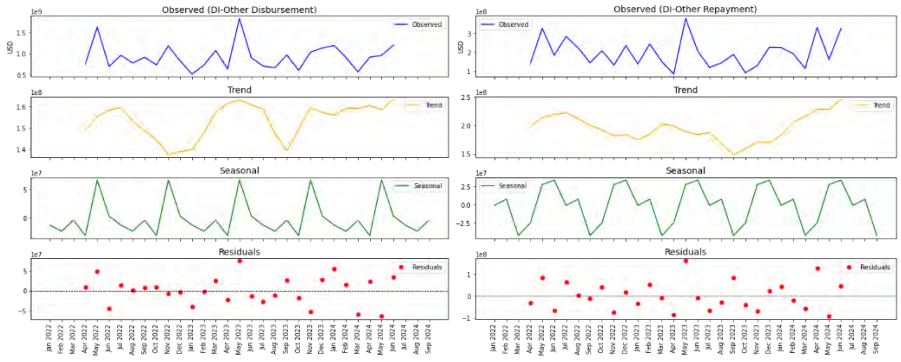
1.2.2. Data Preprocessing for Decomposition-based Hybrid Model

Next, for other types of loans, modelling is conducted using a hybrid model based on time series decomposition. This approach is chosen because, for other loan types, it is not feasible to group debtors into clusters, as indicated by low Silhouette Scores, WCSS values and the MAPE results. During the data preprocessing stage, the raw data is decomposed using the additive method, resulting in three components: Trend, Seasonality, and Residual. The Trend component reflects the overall long-term movement in the data, Seasonality captures recurring patterns or cycles over a fixed period, and Residual represents the random noise or unexplained variability. The results obtained from this decomposition process are presented on (Fig. 10):

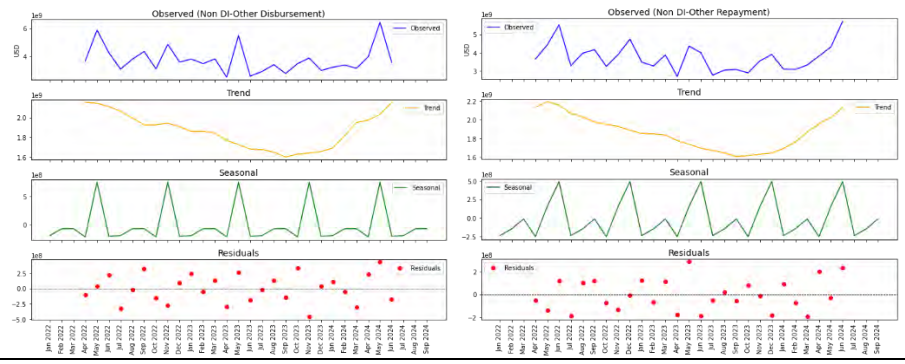
Decomposition Results of (a) DI-Other Loan (b) Non-DI-Other Loan

Figure 10

(a)



(b)



Based on Figure 10, due to the impact of the additive method on the decomposition process, the initial data of 33 data points (Jan 2022 – Sep 2024) was reduced to 27 data points (Apr 2022 – Jun 2024). Each component formed (trend, seasonal, and residual) will undergo different treatments. The trend and residual components will be modelled separately, while the seasonal component will be studied for its seasonal pattern and directly added to the predictions of the trend and residual components. This approach is taken because the seasonal patterns formed through additive decomposition are already very clear, as explained in Section 3.2.c, subsection 2.

1.3. Data Modelling and Evaluation

1.3.1. Clustering-based Hybrid Model on Trade Credit Loans

In this approach, we use a clustering-based hybrid model to forecast Trade Credit (TC) Loans. This method involves grouping time series data using clustering algorithms to form clusters with similar characteristics. The clustering process aims to identify localized patterns within each cluster, enabling a more focused and customized analysis. In this modelling approach, we separate the data based on investment status (Direct Investment (DI) and Non-Direct Investment (Non-DI)) categories for further analysis. The clustering results for the time series disbursement realization data of the DI-TC loan group are presented in (Fig. 11).

Clustering Results from DI-TC Disbursement — Figure 11



Five clusters were generated, each with distinct pattern characteristics:

- Cluster 0, exhibits a simple pattern where most values remain flat with a significant spike toward the end.
- Cluster 1 displays a more complex and fluctuating pattern, with values rising and falling variably over time.
- Cluster 2, shows highly volatile and seemingly random patterns, indicating significant variation within this cluster.
- Cluster 3 stands out with its unique pattern, where values remain stable for most of the period before experiencing a sharp increase at the end.

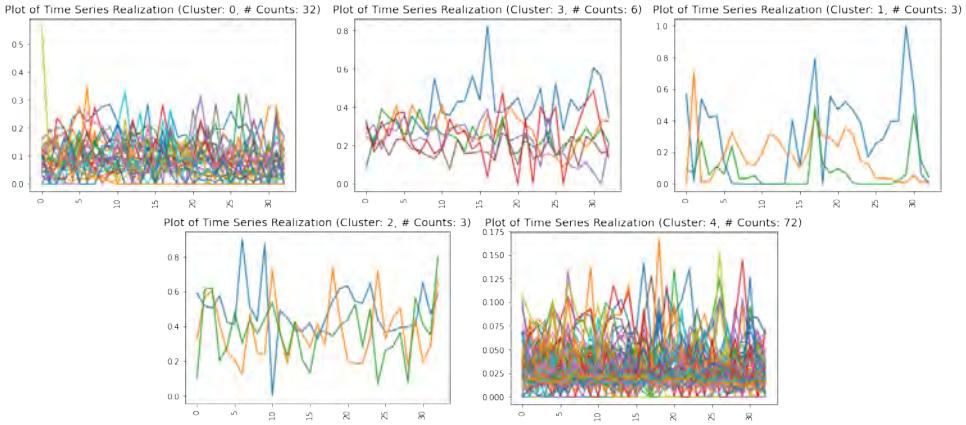
- Cluster 4 demonstrates relatively more organized fluctuations compared to Cluster 2, although variations are still present.

In summary, this clustering result highlights that the time series data for DI-TC Disbursement can be grouped based on similar behavioural patterns. This analysis provides valuable insight into the underlying trends and patterns within the data, which can be further explored for decision-making and interpretation.

Meanwhile, the results for the Non-DI-TC presented in (Fig. 12):

Clustering Results from DI-TC Repayment

Figure 12



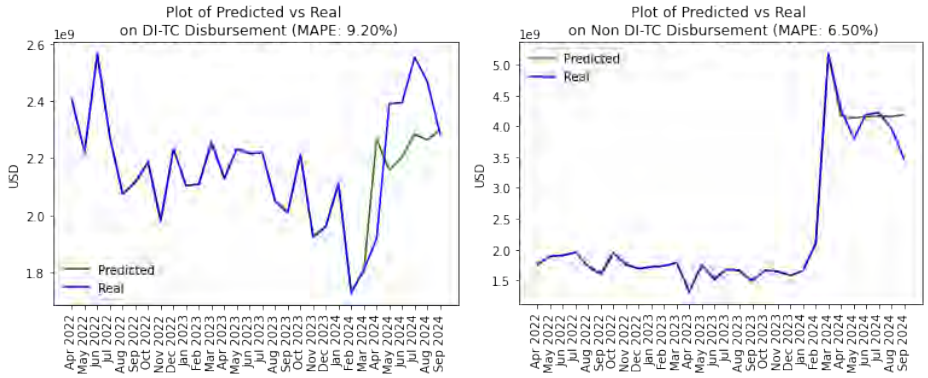
For the Non-DI-TC, five clusters were generated with different characteristics:

- Cluster 0, the lines are densely packed, indicating that the data points share relatively consistent characteristics but with some variations.
- Cluster 1, the patterns are more distinct, suggesting unique characteristics compared to other clusters.
- Cluster 2, the data exhibit different trends, potentially indicating unique underlying behaviours.
- Cluster 3, the patterns show moderate similarity, with slight variations in trends over time.
- Cluster 4, the lines are densely packed and overlap significantly, suggesting that the data points in this cluster share broadly similar trends with minor deviations.

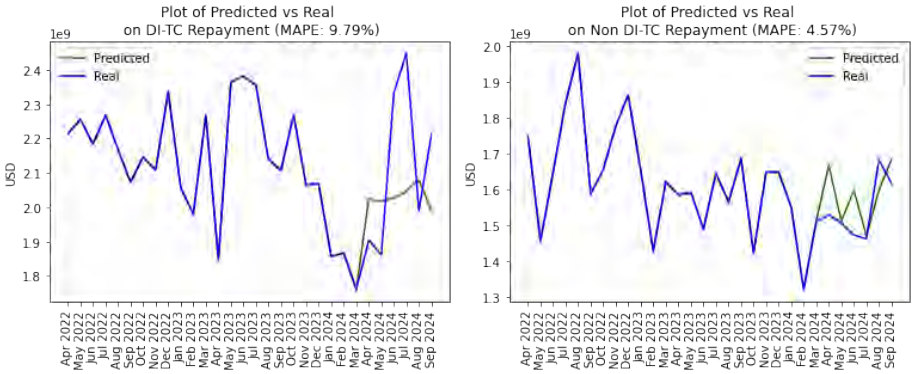
Next, data modelling is performed based on the results from the previous data preprocessing. The following results of the TC modelling are presented in (Fig. 13):

Trade Credit Modelling Results (a) Disbursement (b) Repayment Figure 13

(a)



(b)



Based on Figure 13, the predicted versus actual values of disbursement and repayment for Trade Credit (TC) loans are compared, categorized into Direct Investment (DI) and Non-Direct Investment (Non-DI). In summary, the DI-TC disbursement model achieved a MAPE of 9.2% in medium-term nowcasting (6 months) and 13.9% in short-term nowcasting (2 months), while the repayment model resulted in a MAPE of 9.79% in medium-term and 7.35% in short-term. For Non-DI-TC, the disbursement model demonstrated a MAPE of 6.5% in medium-term and 5.6% in short-term, and the repayment model achieved a MAPE of 4.57% in medium-term and 4.65% in short-term. In this modelling, we perform modelling in each cluster. In Table 3, the details of MAPE and the best model in each cluster are presented.

Comparison of Medium-Term Nowcasting MAPE in Each Cluster

Table 3

		i-th Clusters	Best Model[4]	Best MAPE Cluster (%)	MAPE Hybrid Model (%)
Disbursement	DI-TC	1	RF	206.43	9.2 (RF)
		2	RF	8.28	
		3	Ridge	9.07	
		4	RF	87.76	
		5	Ridge	8.02	
	Non-DI-TC	1	Linear	9.76	6.5 (RF)
		2	RF	15.86	
		3	SVM	85.36	
		4	RF	22.96	
Repayment	DI-TC	1	RF	37.52	9.79 (Lasso)
		2	RF	15.21	
		3	SVM	14.85	
		4	RF	87.73	
		5	Linear	9.04	
	Non-DI-TC	1	RF	23.49	4.57 (Lasso)
		2	SVM	9.86	
		3	Linear	14.88	
		4	Lasso	5.54	
		5	Lasso	8.94	
		6	SVM	39.23	

Based on Table 3, the hybrid modelling approach significantly improves prediction accuracy by leveraging the strengths of different models across clusters, tailoring predictions to the unique characteristics of each group. This flexibility is crucial for handling complex datasets like DI-TC and Non-DI-TC, where variability across clusters can be substantial. For instance, the hybrid model achieves lower overall MAPEs (9.2% for DI-TC disbursement and 6.5% for Non-DI-TC disbursement) compared to individual models. Similarly, for repayment, the hybrid model reduces errors to 9.79% for DI-TC and 4.57% for Non-DI-TC by selecting models like Lasso regression that best fit each cluster. By addressing cluster-specific behaviours, the hybrid approach outperforms single-model strategies, making it a powerful tool for predicting diverse financial data accurately. The DI-TC results show a higher MAPE value compared to Non-DI-TC in terms of both disbursements and repayments. This is consistent with its nature, considering that for external debt, DI-TC is provided by affiliates or shareholders who, at certain periods, will make large debt disbursements or repayments in line with the Company's business needs.

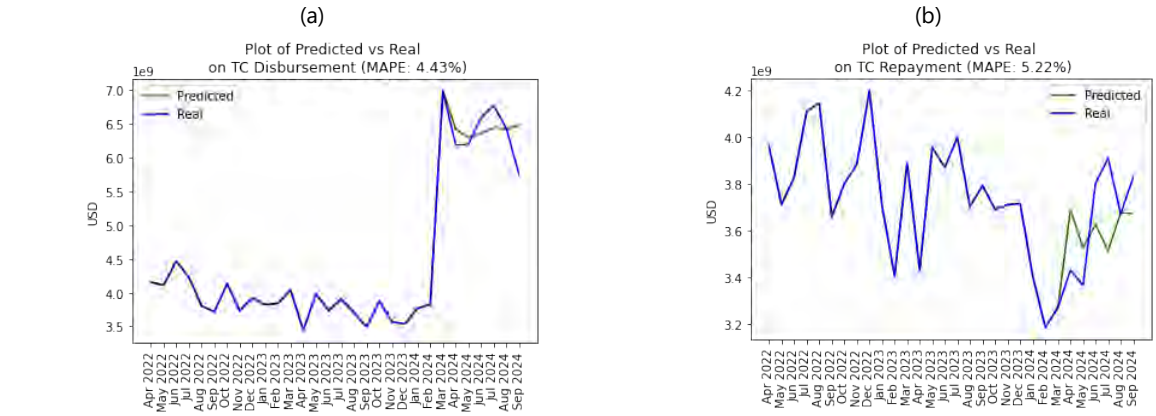
After obtaining the prediction results for DI and Non-DI separately, the results are combined to generate overall predictions for TC Disbursement and TC Repayment. The combined predictions achieve a MAPE of 4.43% (medium-term) and

[4] We utilize Random Forest (RF), Support Vector Regression (SVR), and various linear models such as Lasso Regression, Ridge Regression, and Linear Regression to identify the best model

2.75% (short-term) for TC Disbursement, also for TC Repayment achieve 5.22% (medium-term) and 6.14% (short-term), as illustrated in (Fig. 14):

Results of TC Prediction

Figure 14



1.3.2. Decomposition-based Hybrid Model on Other Loans

For modelling external debt with the "Other" loan type, we applied a time series decomposition-based modelling approach. Following the data preprocessing step, which involved decomposing the time series, each component was subsequently modelled individually. The results of this modelling are presented in Table 4:

Results of Medium-term Nowcasting using Decomposition-based Modelling

Table 4

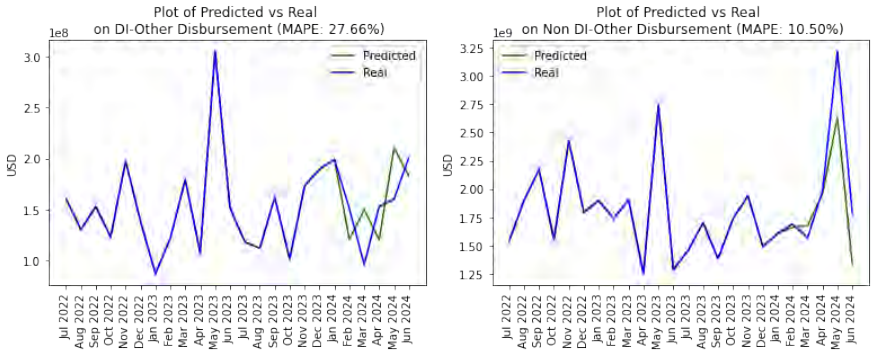
		Time Series Component	Best Model	Best MAPE (%)	MAPE Hybrid Model (%)
Disbursement	DI-Other	Trend	Ridge	2.81	27.66 (RF)
		Residual	Ridge	66.84	
	Non-DI-Other	Trend	Lasso	5.43	10.5 (Ridge)
		Residual	Lasso	88.96	
Repayment	DI-Other	Trend	Ridge	7.39	20.8 (RF)
		Residual	Ridge	67.36	
	Non-DI-Other	Trend	Ridge	4.74	7.78 (Linear)
		Residual	SVR	91.15	

The hybrid model demonstrates strong performance by effectively combining the strengths of different models for the Trend and Residual components, significantly improving prediction accuracy for both Disbursement and Repayment. For non-DI loans, the hybrid model achieves low MAPEs of 7.78% and 10.5%, indicating its robustness in handling predictable trends and variable residuals. While the model also enhances accuracy for DI loans, achieving a MAPE of 27.66% for DI-Other Disbursement, the high MAPE values for Residuals (e.g., 67.36% for DI-Other Repayment) highlight the challenges of modelling unpredictable components. Overall, the hybrid model effectively balances accuracy across components,

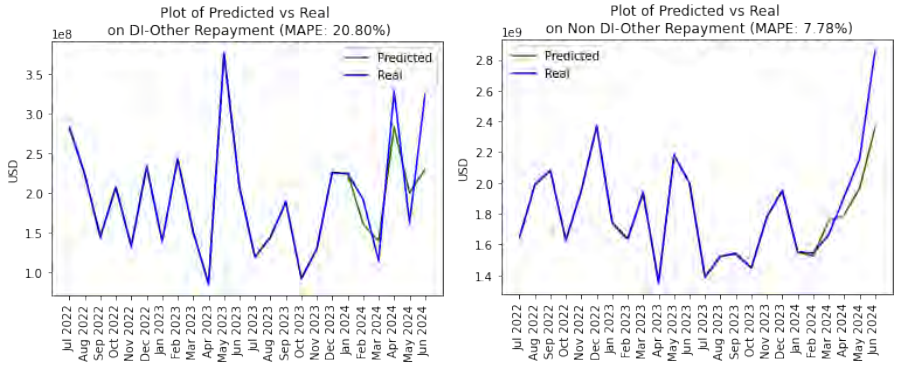
particularly excelling in non-DI loan predictions, though refinements in handling Residual components could further improve its performance for DI loans. Visualization of the result of decomposition-based hybrid model presented in Figure 15:

Other Liabilities Modelling Results (a) Disbursement (b) Repayment Figure 15

(a)



(b)



After obtaining the results for each investment status, all results are combined to produce the final predictions for disbursement and repayment of external debt for the “Other” loan type. The combined result gives better evaluation score, since DI investment type give worse result than non-DI. This condition is due to the behaviour of DI company, that have more flexibility on disbursement and repayment from its parent company. The modelling results for the "Other" loan type are illustrated in Figure 16:

Results of Other Loan Prediction Figure 16

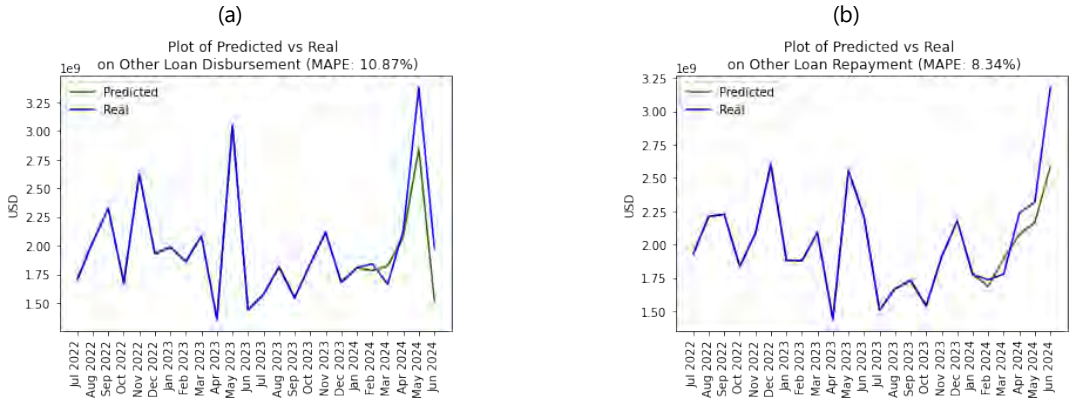


Figure 16 shows the prediction results for disbursement and repayment of other loans, with disbursement MAPE are 10.87% (medium-term) and 6.41% (short-term) also for repayment MAPE are 8.34% (medium-term) and 4.64% (short-term). The disbursement predictions align well with actual values but show slight deviations during periods of sharp spikes, while repayment predictions demonstrate higher accuracy and a closer fit across the time series. These results highlight the hybrid model's effectiveness in capturing overall trends and patterns for other loans, though slight improvements could be made to handle volatility and residual noise more effectively.

1.3.3. Comparison to Existing Method

Currently, in preparing the Balance of Payments for the liabilities section of the Financial Account, the projections/nowcasting of disbursements are solely based on transaction projections submitted by debtors. For repayments, we apply a cleansing technique (e.g., correcting the transaction values based on their behaviours) in the reporting. The comparison of prediction results with the existing method is presented in Table 5:

Comparison of Proposed vs Existing Method Table 5

MAPE (%)		Short-term (2 months)		Medium-term (6 months)	
		Proposed	Existing	Proposed	Existing
Disbursement	TC	**2.75**	14.33	**4.43**	24.54
	Other	**6.41**	17.89	**10.87**	16.95
Repayment	TC	**6.14**	21.33	**5.22**	11.18
	Other	**4.64**	18.01	**8.34**	20.78

Based on Table 5, the proposed method achieves remarkable accuracy improvements for disbursement. In the case of Trade Credit (TC) disbursement, the proposed method reduces the short-term MAPE to 2.75% compared to the existing method's 14.33%, and in the medium term, the proposed method records a MAPE of

4.43%, far better than 24.54% with the existing method. Similarly, for other loans, the proposed method achieves a MAPE of 6.41% in the short term and 10.87% in the medium term, compared to the consistently high MAPE of 17.89% and 16.95%, respectively, for the existing method.

For repayment, the proposed method also demonstrates significant improvements in accuracy. For TC repayment, the short-term MAPE is reduced to 6.14% compared to 21.33% with the existing method, and in the medium term, the proposed method achieves a MAPE of 5.22%, outperforming the existing method's 11.18%. Similarly, for other loans, the proposed method achieves a MAPE of 4.64% in the short term and 8.34% in the medium term, significantly better than the existing method's 18.01% and 20.78%, respectively.

Overall, the proposed method exhibits consistent superiority in accuracy, particularly for disbursement predictions, where the existing method struggles with extremely high MAPE values. The improvements in repayment predictions further highlight the robustness and reliability of the proposed approach for both short-term and medium-term forecasting. These results demonstrate that the proposed method is far more effective and reliable for predicting external debt disbursement and repayment compared to the existing method.

2. Conclusion and Future Works

Based on this discussion, several conclusions from this paper are as follows:

a) Advanced analytics, particularly specifically clustering-decomposition and time series methodologies, can be used for projecting private external debt (trade credit and other loan) disbursement and repayment.

b) The implementation of clustering-decomposition and time series methodologies enhances the reliability of projections. This is proven by projection evaluation using MAPE, as shown in the Table 6.

MAPE Score Result – TC

Table 6

		MAPE Score	Description
Disbursement	DI-TC	9,20%	<10%, highly accurate forecasting
	Non-DI TC	6,50%	
	TC Disb. (in total)	**4,43%**	
Repayment	DI-TC	9,76%	
	Non-DI-TC	4,57%	
	TC Repay. (in total)	**5,22%**	

MAPE Score Result – Other Liabilities

Table 7

		MAPE Score	Description
Disbursement	DI-Other Liab.	27,66%	20-50%, Reasonable forecasting
	Non-DI Other Liab.	10,5%	10-20%, Good forecasting
	Other Liab Disb. (in total)	**10,87%**	**10-20%, Good forecasting**
Repayment	DI-Other Liab.	20,8%	20-50%, Reasonable forecasting
	Non-DI-Other Liab.	7,78%	<10%, highly accurate forecasting
	Other Liab. Repay. (in total)	**8,34%**	**<10%, highly accurate forecasting**

c) Implementing advanced analytics for projecting private external debt, can enhance work process efficiency. This is achieved by reducing dependencies on corporate reporting data, thereby eliminating coordination processes such as data confirmation.

Nevertheless, in this paper, the advanced analytics methodology (clustering-decomposition and time series) is limited to the projection of trade credit and other loans. The use of this methodology for other types of external debt, such as loan agreements, still requires further research. In the future, BI will develop this methodology and explore other data-driven methodologies for external debt projection, while also considering several external indicators that may influence external debt. Advanced analytics methodologies can also be further explored to project other components in the Balance of Payments, such as investment, trade, or income payment/receipt projections.

References

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering – A decade review. *Information Systems, 53*, 16-38.

Awan, R. U., Anjum, A., & Rahim, S. (2015). An econometric analysis of determinants of external debt in Pakistan. *British Journal of Economics, Management & Trade, 5*(4), 382-391.

Banbura, M., Giannone, D., Mondugno, M., & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow. *Centre for Economic Policy Research*.

Biqiong, Z., Al-Asfour, A., & Nilofar, M. (2020). External Debt and Economic Growth: A Heterogeneous Static Panel Study in Asian Developing and Transition Economies. *Studies of Applied Economics*.

Bulut, L. (2011). External Debts and Current Account Adjustments. *The B.E. Journal of Macroeconomics*.

Chen, M. (2020). Time series clustering and classification. *Journal of the American Statistical Association, 115*(531), 1558.

Chou, J., & Ngo, N. (2016). Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy

consumption patterns. *Applied Energy*, 751-770. doi:10.1016/j.apenergy.2016.05.074

Chowdhury, A. (2001). External Debt and Growth in Developing Countries: A Sensitivity and Causal Analysis. *World Institute for Development Economic Research (UNU-WIDER)*.

Cuddington, J. (1997). Analyzing the Sustainability of Fiscal Deficits in Developing Countries.

Dawood, M., Baidoo, S. T., & Shah, S. M. (2021). An empirical investigation into the determinants of external debt in Asian developing and transitioning economies. *Development Studies Research*.

Hajirahimi, Z., & Khashei, M. (2019). Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence, 86*, 83-106.

Huang, X. Y. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences, 367-368*, 1-13.

IMF. (2014). *External Debt Statistics, Guideline for Compilers and Users.* Washington D. C.: International Monetary Fund.

Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications, 1*.

Jorge, M.-B., & Ruben, C. (2024). Time series clustering with random convolutional kernels. *Data Mining and Knowledge Discovery, 38*, 1862-1888.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data-driven networks. *Future Internet, 15*(255), 1-31.

Lewis, C. (1982). *Industrial and business forecasting methods.* London: Butterworth Scientific.

Myttenaere, A., B., G., Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing, 192*, 38-48. doi:10.1016/j.neucom.2015.12.114

Rahman, Y. A., & Pujiati, A. (2021). Dynamic forecasting of government foreign debt: Case of Indonesia. *JEJAK: Journal of Economics and Policy, 14*(1), 52-60.

Razinkova, M., Filatova, H., Pylypenko, A., Nebaba, N., Firsov, O., & Zhuravka, F. (2023). Forecasting Ukraine's external public debt under uncertainty conditions. *Financial and Credit Activity: Problems of Theory and Practice, 6*(53), 202-213.

Reinhart, C., & Rogoff, K. (2010). Growth in a Time of Debt. *American Economic Review*.

Sugiyarto, A., & Abadi, A. (2019). Prediction of Indonesian Palm Oil Production Using Long Short-Term Memory Recurrent Neural Network (LSTM-RNN). *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 53-57). Ipoh, Malaysia: IEEE Xplore. doi:10.1109/AiDAS47888.2019.8970735

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., . . . Woods, E. (2020). Tslearn, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research, 21*, 1-6.

Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X., & Peng, X. (2019). A hybrid modeling method for time series forecasting based on a linear regression model and deep learning. *Applied Intelligence, 51*(4), 1-20.

Yang, Y., Fan, C., & Xiong, H. (2021). A novel general-purpose hybrid model for time series forecasting. *Applied Intelligence, 51*(8), 1-17.

Zhao, G., & Zhao, Z. (2024). EMD-DL: A hybrid forecasting approach using empirical mode decomposition and deep learning for time series prediction. *Preprint*.

Zhou, W., Wang, D., Li, H., & Song, W. (2018). Long-Term Forecasting of Time Series Based on Sliding Window Information Granules and Fuzzy Inference System. *2018 5th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)* (pp. 375-380). Hangzhou, China: IEEE Xplore. doi:10.1109/ICCSS.2018.8572415

Enhancing External Debt Nowcasting: A Data-Driven Approach Using Advanced Analytics

Bank Indonesia, Statistics Department

February

2025



Table of Contents

1 Background

2 Research Question

3 Methods

4 Results

5 Conclusion

Background

Why external debt projection important?



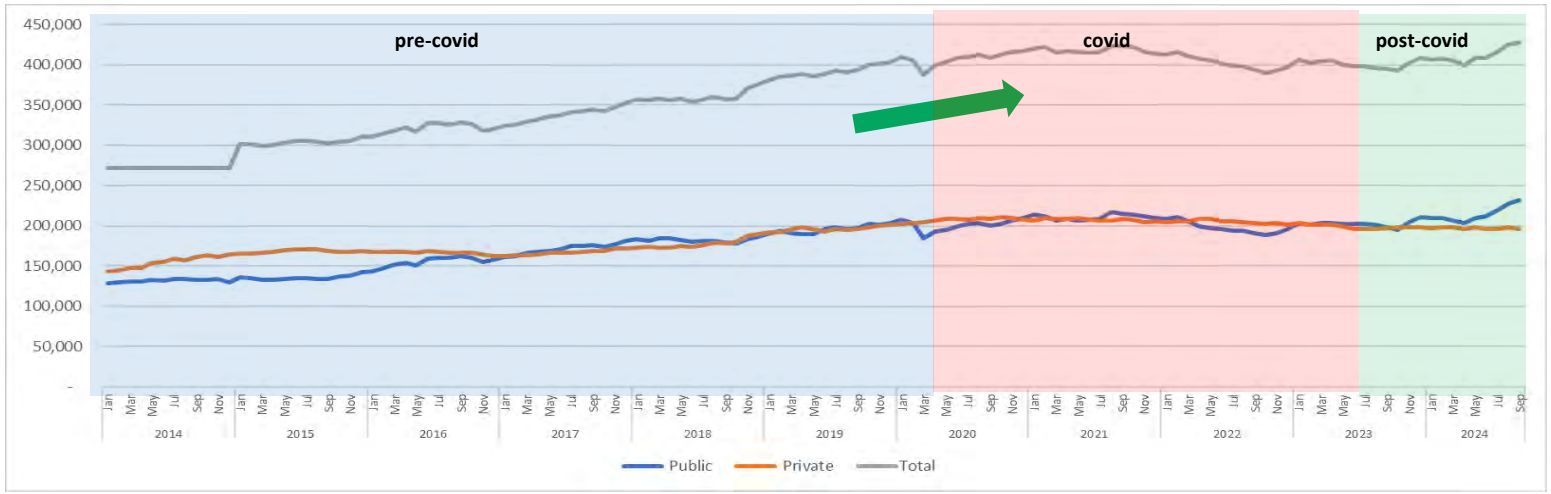
1. External debt is one of the Indonesian primary financing sources.
2. The substantial share of external debt in financing highlights the importance of debt sustainability monitoring (including accurate projection).
3. Projections for external debt disbursements and repayments is a part of foreign exchange supply-demands projection, which are critical for Bank Indonesia's objective of achieving and maintaining Rupiah exchange rate stability.
4. Due to the dependency on corporate report which still lack of good disbursement dan repayment projection, private external debt projection needs improvement in methodology.
5. This research attempt to use data driven approaches (clustering-decomposition, and time series analysis) for private external debt disbursements and repayments projection.

Background

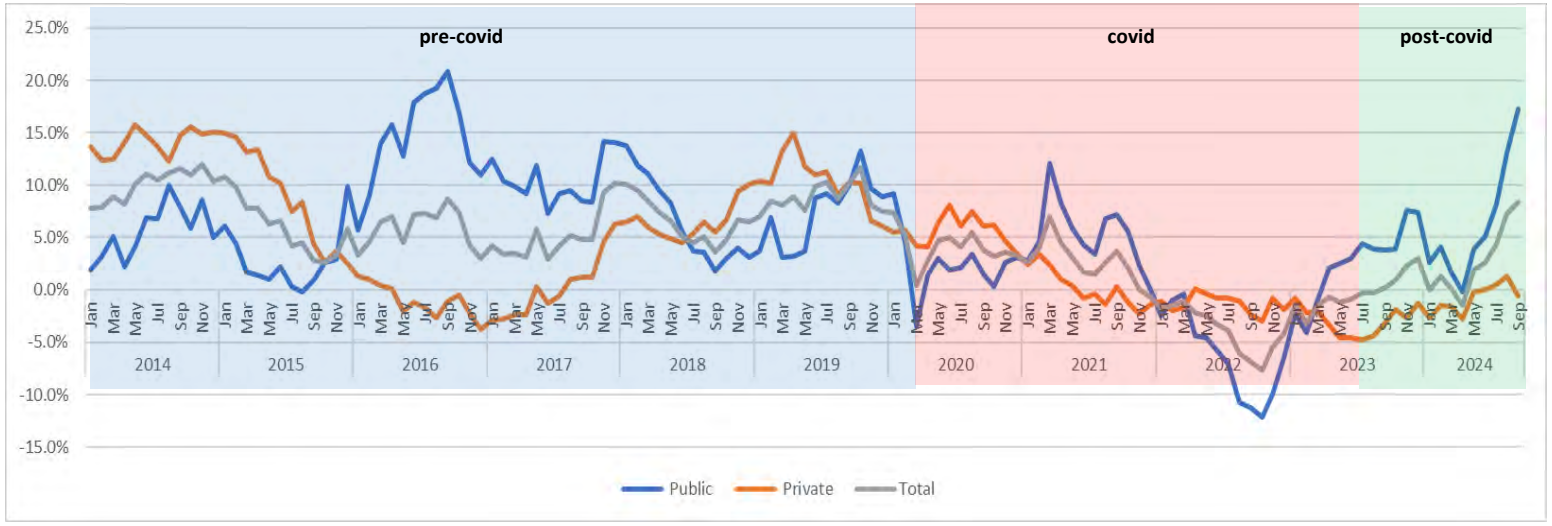
Indonesia's external debt has grown over the past decade

Indonesian External Debt

million of USD



% (yoy)



Source: Indonesian External Debt Publication period: September 2024

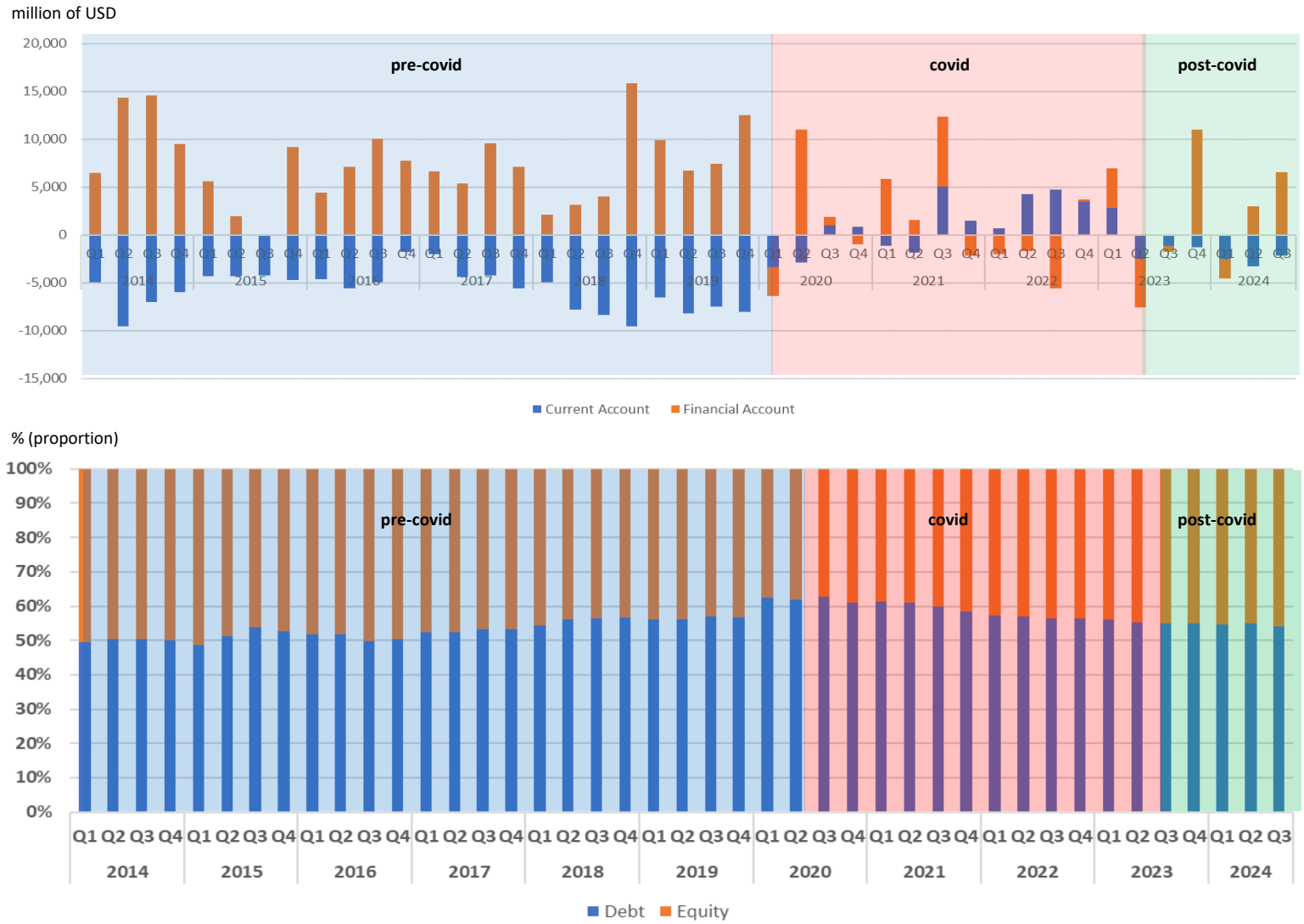
Indonesia's external debt grew from USD 293.8 billion at the end of 2014 to USD 427.8 billion by the end of third quarter 2024. This growth stems from increases in both public and private external debt.

This growth stems from increases in both public and private external debt, with an average annual growth rate of 4.4% (y-o-y). After a contraction during the COVID-19 period (Q1 2020 - Q2 2022), Indonesia's external debt has resumed growth post-pandemic. However, private external debt growth still under pre-pandemic growth, this affected by post-pandemic economic growth that has not yet fully recovered.

Background

External debt played important role as part of Indonesia Balance of Payment (BoP)

Indonesian Balance of Payment (BoP) and International Investment Position (IIP)



Indonesia's current account deficit was offset by financial account surplus. This trend reversed slightly during the pandemic, as the current account recorded a surplus driven by reduced imports due to the decline in economic activity. This anomaly ended post-pandemic, with current account returning to deficit, balanced by surplus in financial account.

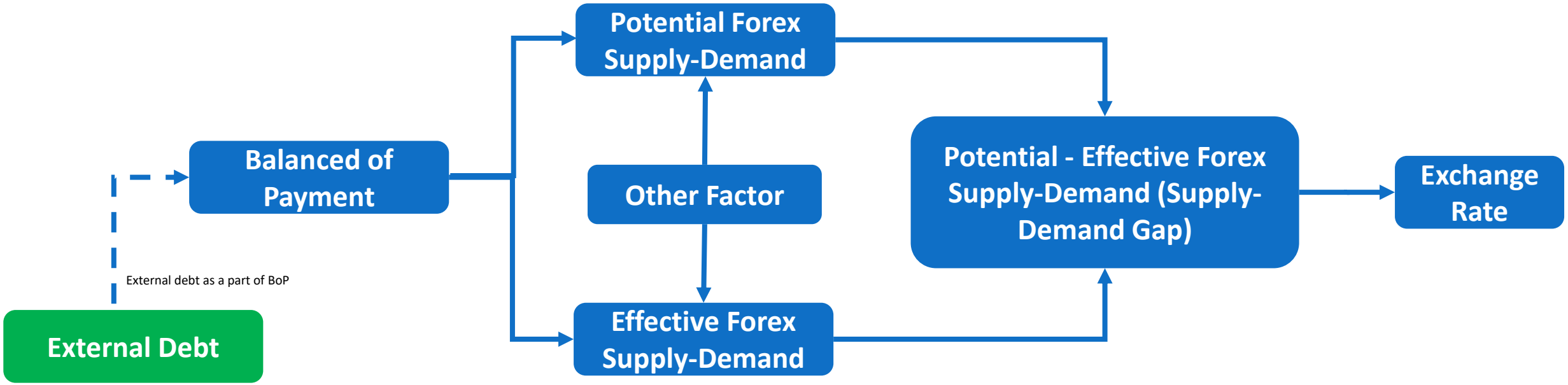
Indonesian external debt also plays an essential role in the Balance of Payments (BoP) and International Investment Position (IIP), especially in financial liability. External debt forms approximately 55% of financial account liabilities position.

Source: Indonesian Balance of Payment and International Investment Position Publication, period: Q3-2024

Background

External Debt can affect Foreign Exchange (Forex) Supply and Demand

- Forecasting foreign exchange demands aligns with the role of Bank Indonesia as the central bank, whose primary objective is to achieve and maintain Rupiah stability. Rupiah stability implies price stability of goods and services as well as rupiah exchange rate stability. To achieve the objective of maintaining rupiah exchange rate stability, Bank Indonesia conduct various strategies that require reliable information, including projections of foreign exchange supply-demands.
- Bank Indonesia projections for external debt disbursements and repayments is an important part of foreign exchange supply-demands projection.



Framework: Working Paper (Piter, 2016)

Research Question

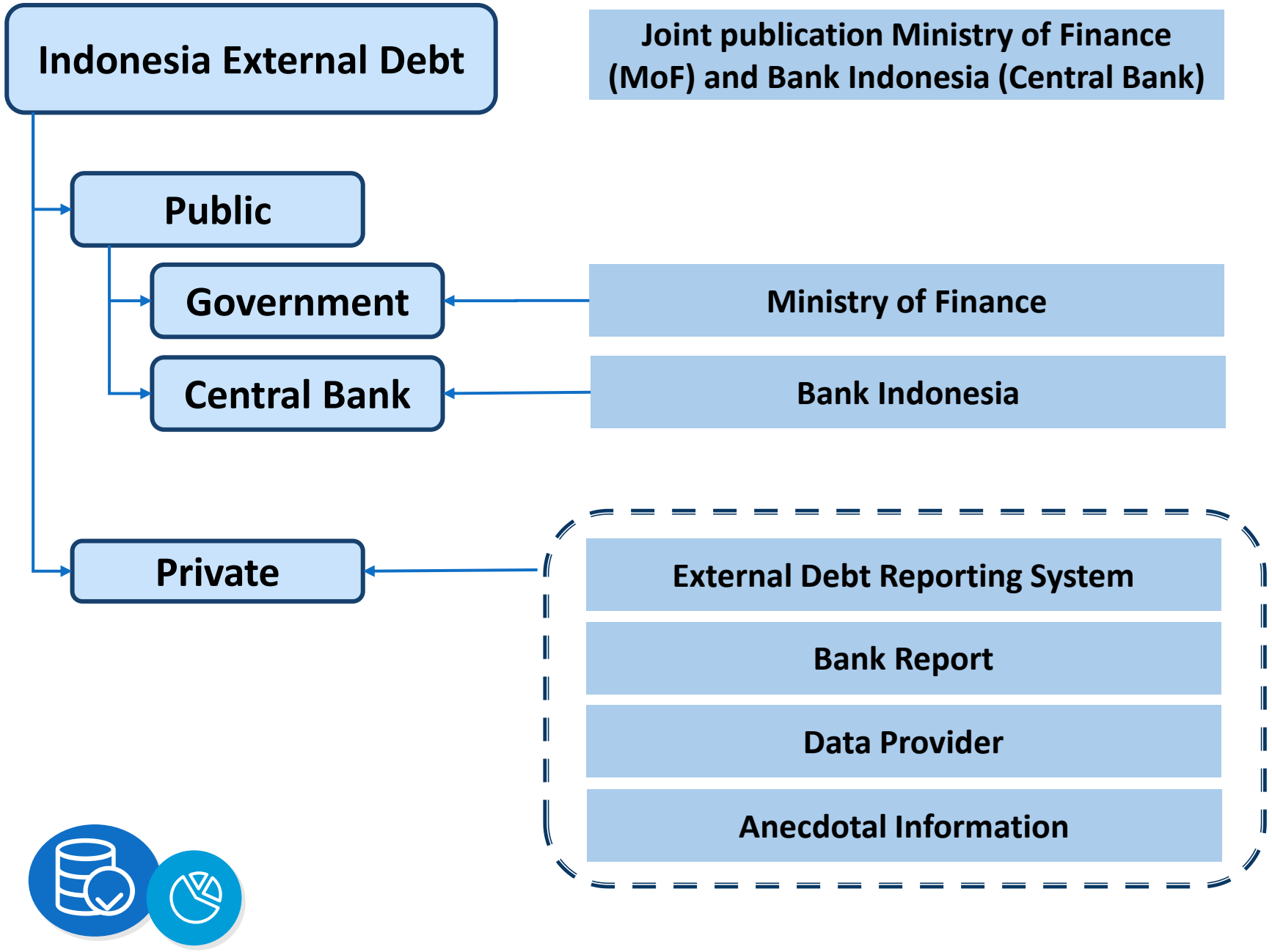
- Is alternative methodology using data-driven approach is suitable for private external debt projection?
- How accurate is data-driven approach (clustering and time series) in projecting private external debt disbursement and repayment?
- What are the advantages in using data-driven approach (clustering-decomposition and time series) for private external debt projection?

Limitation

- This paper is limited to the projection of private external debt, specifically for trade credit and other loan. The other loan type, such as loan agreement and debt securities, needs more deep research on its projection method, including adding external variable such as economic growth, external rate, etc.



Methods: Data Sources



Methods:
Research Stages

Data Analysis

This step includes exploring the collected external debt data to identify patterns and relationships that can be used in the development of the nowcasting model. The main objective is to understand the basic behavior of external debt transactions.

Data Preprocessing

This stage focuses on preparing raw data before modelling. In the context of this paper, these steps include cleaning up external debt data, handling missing or inconsistent values, and normalizing the dataset. This stage also select perpetrators based on the Pareto 80/20 principle. Only actors who account for 80% of the total contribution (based on historical data) are prioritized for further analysis.

Data Modelling

In this step, the clustering-decomposition method and time series analysis are combined to develop a hybrid nowcasting model. Clustering-decomposition is used to organize debt data into homogeneous groups, thus allowing the application of a specific time series model for each group. This hybrid approach aims to improve prediction accuracy by leveraging the power of both methods.

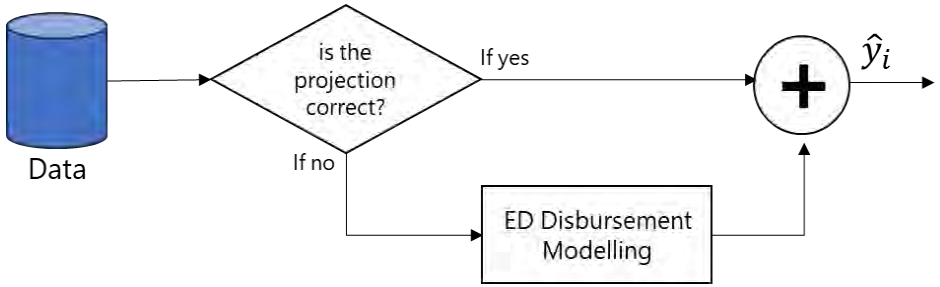
Evaluation

The last stage is evaluation. This stage assesses the performance of the nowcasting model using metrics such as mean absolute percentage error (MAPE). This evaluation ensures that the model is able to provide accurate short-term and medium-term predictions for external debt disbursement and repayment.

Methods

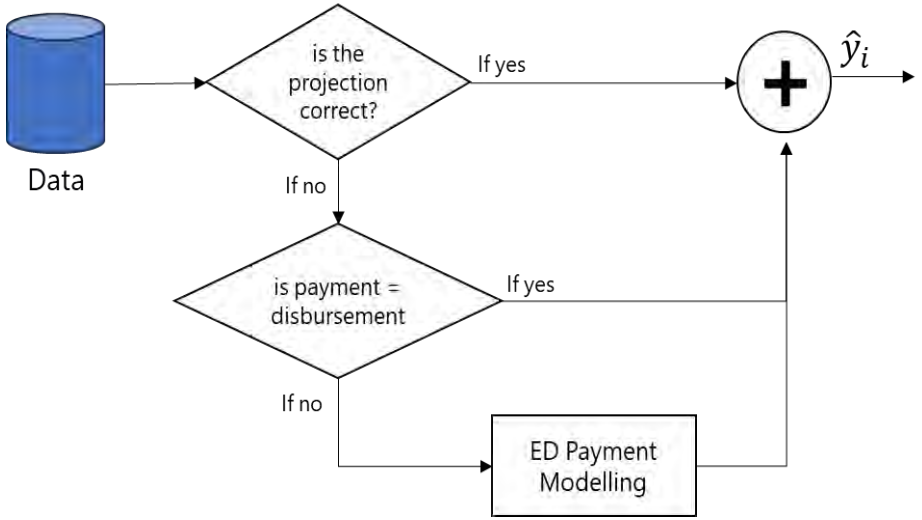
Data analysis:
company selection

Disbursement



- First step is to decide whether the "projection/plan" data reported by company are correct.
- If correct, then use the "projection". If not, then the company is selected to be used in modeling stage.

Repayment

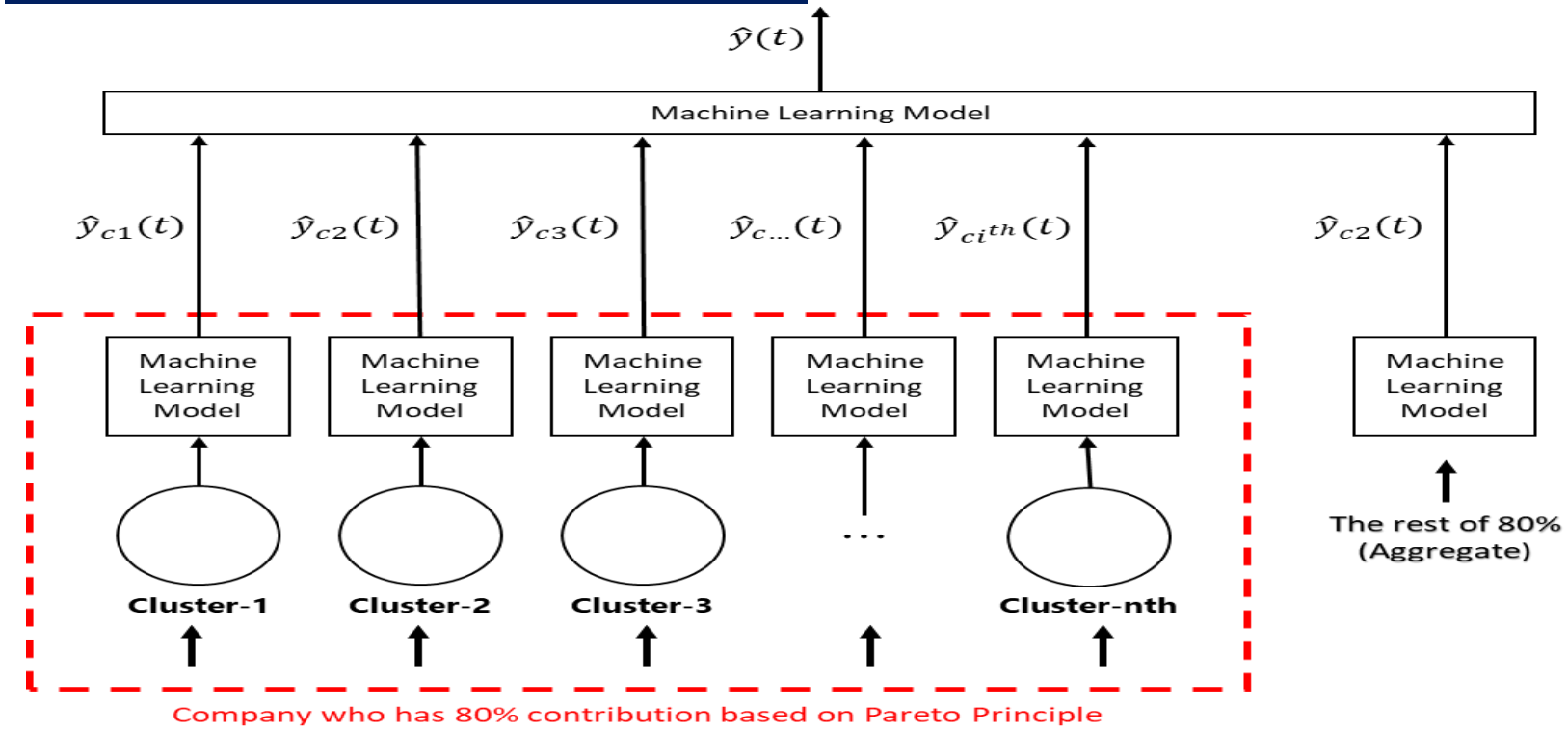


- First step is to decide whether the "projection/plan" data reported by company are correct.
- If correct, then use the "projection". If not, then decide whether the payment is mirroring the disbursement in the same period.
- If correct, then use the disbursement data for repayment. If not, then the company is selected to be used in modeling stage.

Methods

Data preprocessing and modelling:
clustering and time series analysis

Clustering and time-series analysis

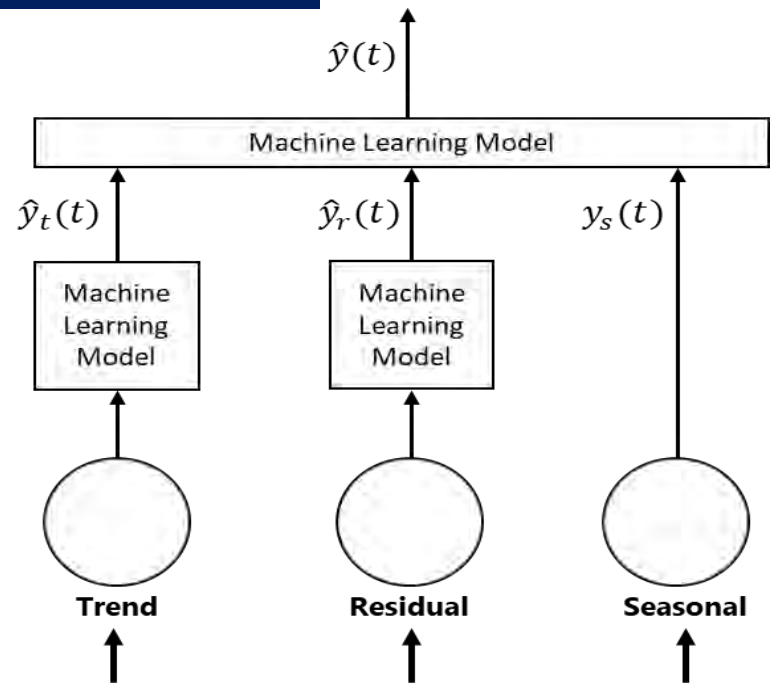


1. Select list of company that have 80% contribution based on nominal value.
2. Using machine learning to divide group of company into several cluster, based on their behaviour on external debt disbursement and repayment.
3. Construct machine learning model (time-series model) for each cluster.
4. Combining result from each model to produce a result.

Methods

Data preprocessing and modelling: decomposition and time series analysis

Decomposition and time-series analysis



1. Divide the company population into 3 categories based on their behaviour in debt disbursement and repayment. The category are:
   - Trend, if there is a pattern on company disbursement and repayment
   - Seasonal, if company make disbursement and repayment in certain period.
   - Residual, other company that doesn't have pattern on disbursement and repayment.
2. Construct machine learning model (time-series model) for each category.
3. Combining result from each model to produce a result.

Methods:

Evaluation

Mean Absolute Percentage Error (MAPE)

For evaluation we use mean absolute percentage error (MAPE). This evaluation ensures that the model is able to provide accurate short-term and medium-term predictions for external debt disbursement and repayment. The formula for the metric is as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

$\hat{y}_i$: predicted value for i-th data point

y_i : actual value for the i-th data point

n : number of observations

In order to conclude whether the model is reliable or not, Lewis (1982) divide MAPE criteria as follows:

MAPE	Forecasting Power
<10%	Highly accurate forecasting
10% - 20%	Good forecasting
20% - 50%	Reasonable forecasting
>50%	Weak and inaccurate forecasting

Result

This research show that data-driven approach using clustering-decomposition and time series can project external debt disbursement and repayment specifically for trade credit and other loan





Result : Projection on Trade Credit

1. Data Analysis

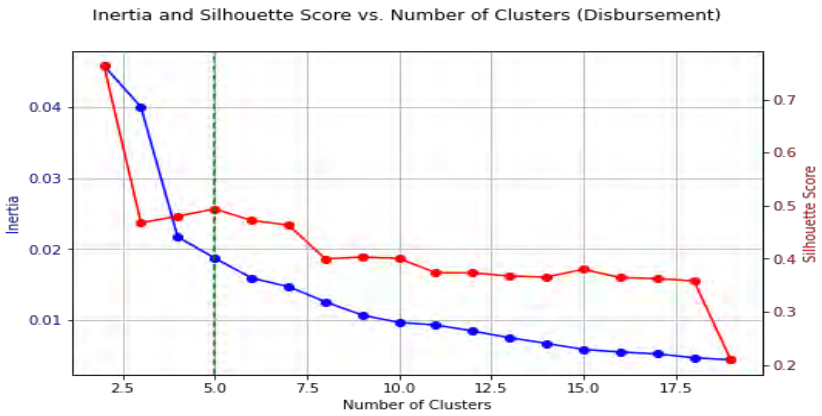
For the trade credit, we use clustering and time-series analysis. The first step is to identify if the projection from company reporting is correct. We found that for all company that have trade credit external debt, the projection reporting is not reliable. Next step, for repayment, we identified 13 company have identical repayment and disbursement patters. There are 4 (four) debtors from the Direct Investment-Trade Credit (DI-TC) group and 9 (nine) debtors from the Non-DI-TC group with such patterns.

2. Data Preprocessing



Example of DI-TC Disbursements

The first step of data preprocessing is to determine company who contribute 80% of the total external debt. The company contributing 80% for the DI-TC loan type (disbursement) are 113 debtors, Non-DI-TC (disbursement) are 115 debtors, DI-TC (payment) are 114 debtors, and Non-DI-TC (payment) are 115 debtors. After that, data normalization is performed to align variable scales to equalize, improve model performance, reduce bias, and facilitate analysis.



Example of Selecting the Optimal Number of Clusters for DI-TC Disbursement

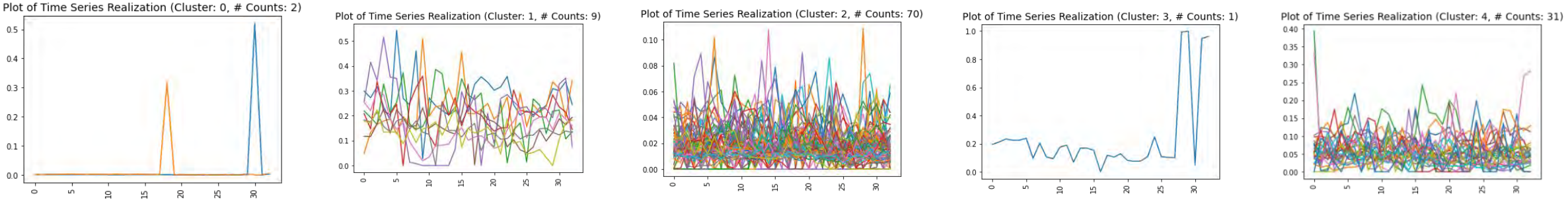
The next step is the clustering process using DTW, selected based on the Within-Cluster Sum of Squares (WCSS) and Silhouette Score values. The results for the WCSS and Silhouette Score TC are as follows:

		# Best Clusters	Silhouette Score	WCSS Score
Disbursement	DI-TC	5	0.4936	0.0187
	Non-DI-TC	4	0.7000	0.0007
Repayment	DI-TC	5	0.4206	0.0202
	Non-DI-TC	6	0.4668	0.0137

Result : Projection on Trade Credit

3. Data Modelling

In this stages, we use a clustering-based hybrid model to forecast Trade Credit (TC) Loans. This method involves grouping time series data using clustering algorithms to form clusters with similar characteristics. The clustering process aims to identify localized patterns within each cluster, enabling a more focused and customized analysis. In this modelling approach, we separate the data based on investment status (Direct Investment (DI) and Non-Direct Investment (Non-DI)) categories for further analysis.



Example of Clustering Results from DI-TC Disbursement

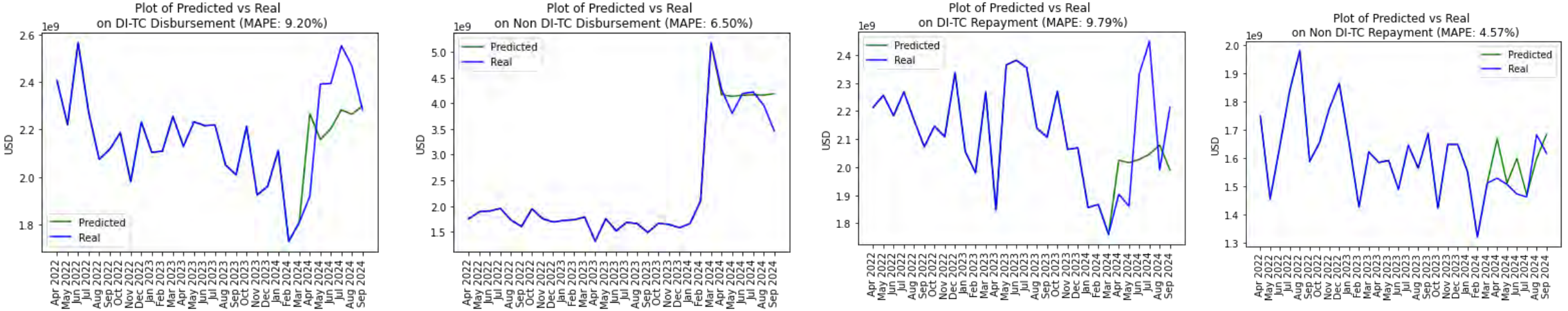
Five clusters were generated for DI-TC Disbursement, each with distinct pattern characteristics:

1. Cluster 0, exhibits a simple pattern where most values remain flat with a significant spike toward the end.
2. Cluster 1 displays a more complex and fluctuating pattern, with values rising and falling variably over time.
3. Cluster 2, shows highly volatile and seemingly random patterns, indicating significant variation within this cluster.
4. Cluster 3 stands out with its unique pattern, where values remain stable for most of the period before experiencing a sharp increase at the end.
5. Cluster 4 demonstrates relatively more organized fluctuations compared to Cluster 2, although variations are still present.

Result : Projection on Trade Credit

4. Data Modelling and Evaluation

Next, data modelling is performed for all cluster based on the results from the previous steps. Next, we combined result form each cluster for: 1) DI TC Disbursement; 2) Non-DI TC Disbursement; 3) DI TC Repayment; and 4) Non-DI TC Repayment.



Result of Projection on Trade Credit

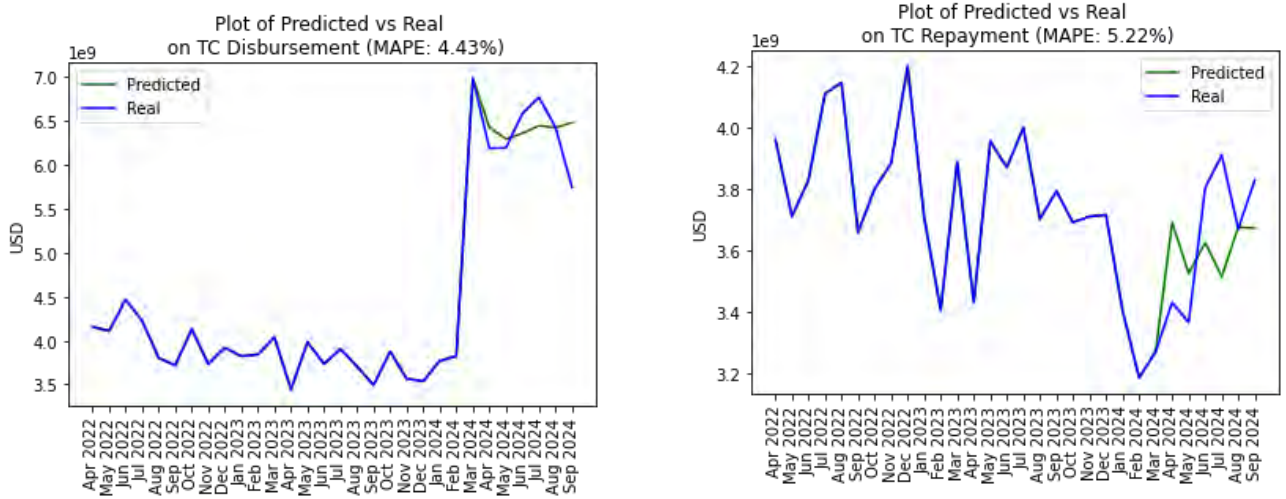
Final step is evaluation, this process use mean absolute percentage error (MAPE). The result is as follows:

No	Loan Type	MAPE score	Description
1	DI TC Disbursement	9,20%	<10%, highly accurate forecasting
2	Non-DI TC Disbursement	6,50%	<10%, highly accurate forecasting
3	DI TC Repayment	9,79%	<10%, highly accurate forecasting
4	Non-DI TC Repayment	4,57%	<10%, highly accurate forecasting

Result : Projection on Trade Credit

5. Data Modelling and Evaluation

Furthermore, we tried to combined result form DI and Non-DI trade credit, both disbursement and repayment.



Result of Projection on Trade Credit

The combined result give better evaluation score. The evaluation use mean absolute percentage error (MAPE), as follows:

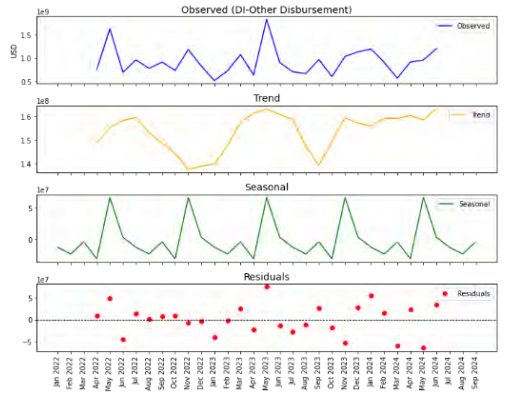
No	Loan Type	MAPE score	Description
1	TC Disbursement	4,43%	<10%, highly accurate forecasting
2	TC Repayment	5,22%	<10%, highly accurate forecasting

Result : Projection on Other Loan

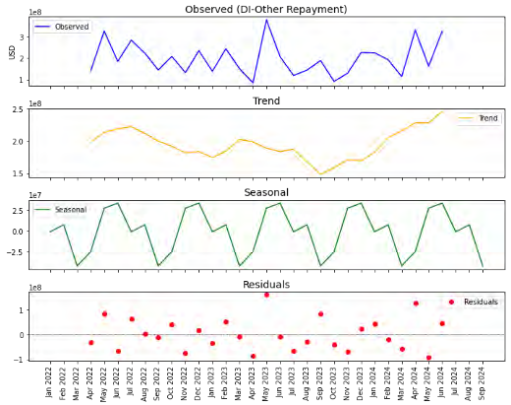
1. Data Analysis

For the other loan, we use decomposition and time-series analysis. Clustering is not used in this type of loan, because the behaviour of company in disbursement and repayment external debt often change over time. The first step is to identify if the projection from company reporting is correct. We found that for all company that have other loan external debt, the projection reporting is not reliable.

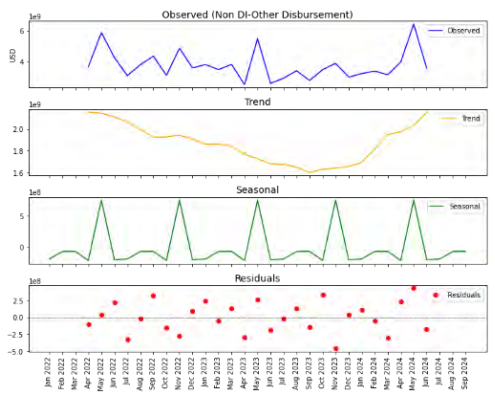
2. Data Preprocessing



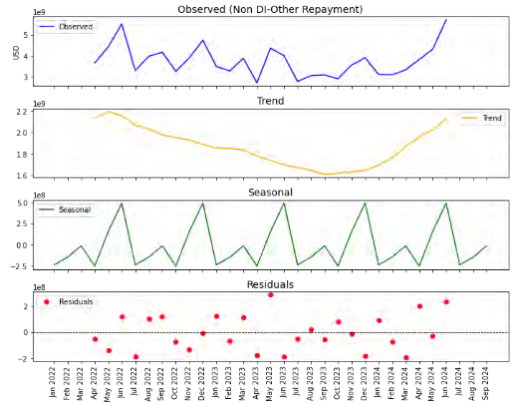
Decomposition results of DI Other Loan Disbursement



Decomposition results of DI Other Loan Repayment



Decomposition results of Non-DI Other Loan Disbursement



Decomposition results of Non-DI Other Loan Repayment

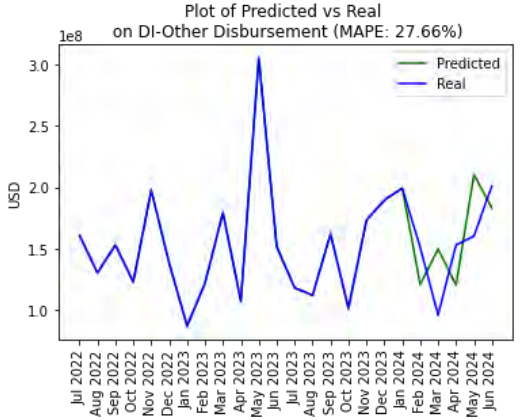
For each other loan type component (DI disbursement, DI repayment, non-DI disbursement and non-DI repayment), we formed 3 models which is trend, seasonal, and residual. The trend and residual components will be modelled separately, while the seasonal component will be studied for its seasonal pattern and directly added to the predictions of the trend and residual components. This approach is taken because the seasonal patterns formed through additive decomposition are already clear.

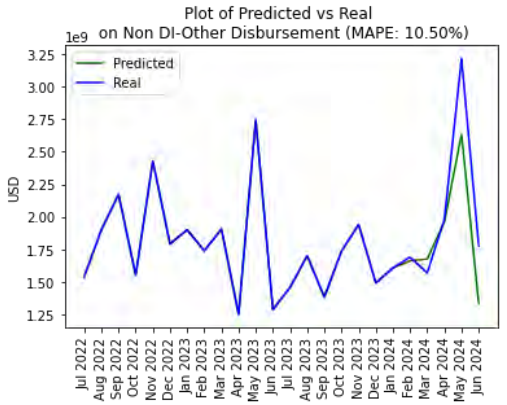
Example of Selecting the Optimal Number of Clusters for DI-TC Disbursement

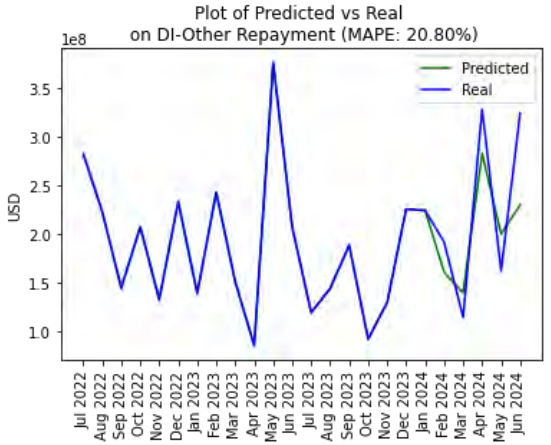
Result : Projection on Other Loan

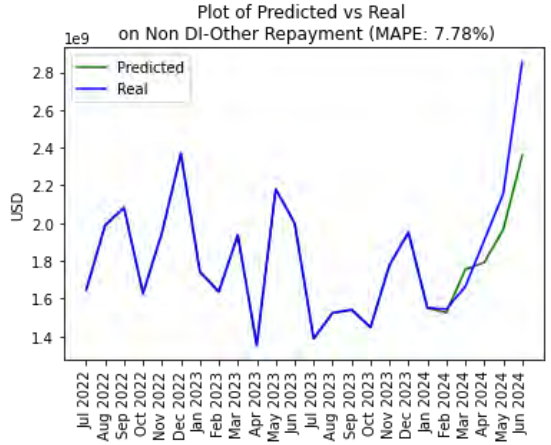
3. Data Modelling and Evaluation

Next, data modelling is performed for all category based on the results from the previous steps. Next, we combined result form each cluster for: 1) DI OL Disbursement; 2) Non-DI OL Disbursement; 3) DI OL Repayment; and 4) Non-DI OL Repayment.









Result of Projection on Trade Credit

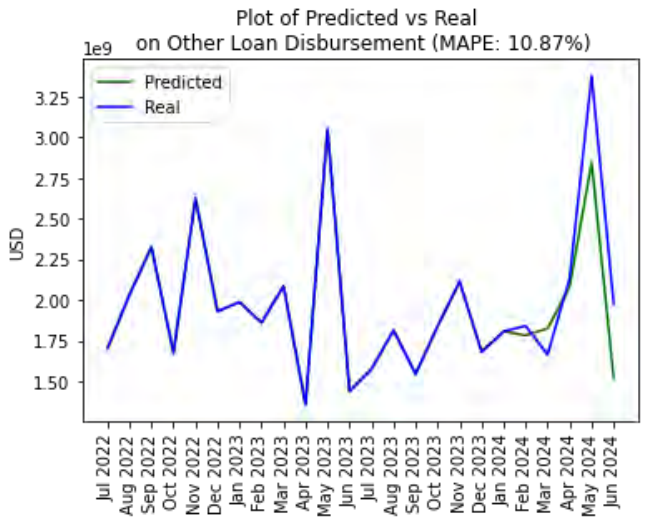
Final step is evaluation, this process use mean absolute percentage error (MAPE). The result is as follows:

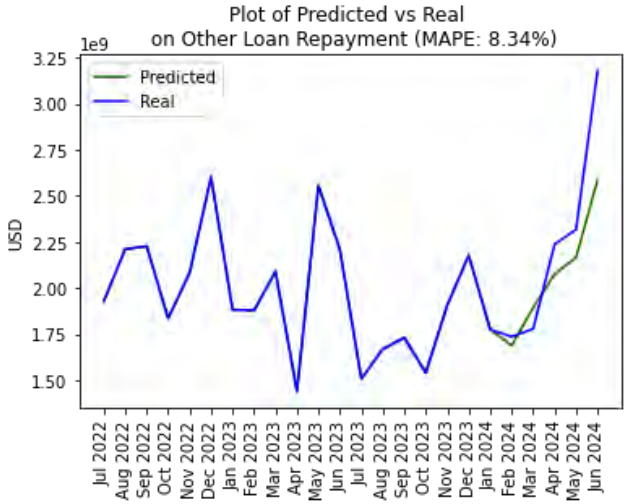
No	Loan Type	MAPE score	Description
1	DI OL Disbursement	27,66%	20-50%, Reseonable forecasting
2	Non-DI OL Disbursement	10,50%	10-20%, Good forecasting
3	DI OL Repayment	20,8%	20-50%, Reasonable forecasting
4	Non-DI OL Repayment	7,78%	<10%, highly accurate forecasting

Result : Projection on Other Loan

4. Data Modelling and Evaluation

Furthermore, we tried to combined result form DI and Non-DI other loan, both disbursement and repayment.





Result of Projection on Other Loan

The combined result give better evaluation score. The evaluation use mean absolute percentage error (MAPE), as follows:

No	Loan Type	MAPE score	Description
1	OL Disbursement	10,87%	10-20%, Good forecasting
2	OL Repayment	8,34%	<10%, highly accurate forecasting



Conclusion

- Advanced analytics, specifically clustering-decomposition and time series methodology **can be used for projecting** private external debt (trade credit and other loan) disbursement and repayment.
- Implementing clustering-decomposition and time series methodology increase **the reliability of projection**. This proves by projection evaluation using Mean Absolute Percentage Error (MAPE).
- Implementing advanced analytics for projection private external debt, **can increase work process efficiency**. This achieved by reducing dependencies to company reporting data, thus eliminating coordination process such as the data confirmation


THANK
YOU

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Generalised weighted framework for synthetic data evaluation[1]

Chiung Ching Ho,
Central Bank of Malaysia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Generalised weighted framework for synthetic data evaluation

Chiung Ching Ho (BNM)

The views expressed in this paper are of the author and do not necessarily represent the views of the associated institutions.

Abstract

The application of synthetic data in central banking has been on the rise, particularly as a means of data sharing securely sharing data between central banks and governmental agencies or research institutions. The need to securely share data is driven by increasing legislation worldwide concerning data protection and privacy. Another emerging reason for using synthetic data in central banks is to increase data availability, in situations where the original data is scarce or difficult to collect

As synthetic data sharing increases, it becomes imperative that the correctness and quality of the synthetic is assessed by data providers prior to publication. In this paper, a generalised weighted framework for synthetic data evaluation is proposed to enable this objective. This framework aims to combine measures for four-dimensions:

i. Statistical distance
ii. Structural similarity
iii. Machine learning task performance
iv. Privacy

The composite final weighted score = wSD + wSS +wML + wP , where SD denotes statistical distance, SS denotes structural similarities, ML denotes machine learning task performance and P denotes privacy, with w being the corresponding weight for each dimension. Recent works for synthetic data evaluation has focused on the statistical distance of the synthetic data from the original data and its structure, which intuitively impacts machine learning outcomes or privacy concerns but not all these dimensions together.

This framework was evaluated on a credit-card default dataset with identifying information and content that is suitable for classification, clustering, and time-series forecasting. A synthetic dataset was created from the original dataset, using a variety of methods including statistical and deep learning methods, as well as differential privacy methods. Subsequently, the synthetic datasets were compared to the original dataset using the four dimensions mentioned above during experiments for classification, clustering and time-series forecast. Ablative experiments were further conducted to evaluate the framework's ability to cater to different scenarios.

In the ablative experiments, the framework was adjusted for different scenarios e.g. different prioritisation for different analytical outcomes, or privacy-first analysis ahead of analytics. This generalised approach will allow users to conduct differentiated evaluation on synthetic data. This in-turn will help data providers

provide high-quality synthetic data, which is well suited to the task on hand, leading to efficiency gains.

JEL classification: C18, C45, C63

Contents

1. Introduction 2

2. Synthetic data 4

2.1 Synthetic data and central bank 4

2.2 Models for synthetic data generation 5

2.3 Assessing synthetic data 5

2.3.1 Privacy 6

2.3.2 Utility 6

2.3.3 Fidelity 6

3. Methods 7

3.1 Framework 7

3.2 Experiment 10

3.2.1 Data set 11

3.2.2 Experimental Results 11

4. Discussion 17

5. Extension 18

6. Conclusion 19

References 20

1. Introduction

Data collected by central banks serve as essential resources that have gained considerable value with the increasing application of artificial intelligence (AI). This data is especially valuable because it is typically not published as open data, due to data governance policies and legislation (European Union, 2024, p. 10) designed to protect privacy and confidentiality of data. Consequently, this data can be utilized by central banks to enhance existing large language models (LLMs), which are experiencing a scarcity of data (Robinson, 2024).

Simultaneously, data sharing conducted by central banks remains influenced by global events such as the Great Financial Crisis (GFC) of 2007-2009 and, more recently, the Covid-19 pandemic. Following the GFC, seven key recommendations have been

identified to enhance data-sharing efforts (Irving Fischer Committee on Central Bank Statistics, 2023):

1. Promoting the use of common statistical identifiers.
2. Promoting the exchange of experience on statistical work with granular data and improving transparency
3. Balancing confidentiality and users' needs.
4. Linking different datasets.
5. Provision of data at the international level.
6. Consideration of ways of improved data-sharing of granular data; and
7. Collection of data only once.

To balance the need for confidentiality with users' requirements, central banks have adopted the practice of sharing data at an aggregated level and providing granular data in a depersonalised format. While the former method maintains confidentiality and privacy, it is less effective than granular data for conducting timely and thorough risk analysis, as well as formulating targeted policy measures (Ling, 2013).

Recent developments have seen the usage of synthetic data as a solution for sharing data whilst adhering to privacy and confidentiality concerns. Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s) (Jordon et al., 2022). Here, synthetic data is contrasted with real data. Real data is data generated not by a model but by a real-world system, e.g. data that is submitted by a central bank's regulatees.

As synthetic data is statistically or semantically similar to the originally data used in its creation, an argument is made that the synthetic data will then be able to be used in the same way for machine learning applications, and with reasonable expectations that the outcome would be similar to outcomes using the original data. At the same time, the synthetic data set would be able to preserve privacy and confidentiality. As central banks have a mission to manage reputational risk, it is imperative that any data synthetic data set is assessed for utility, fidelity and privacy prior to being shared.

In this paper, a framework is proposed to address this problem and apply it on an open-sourced credit card dataset that has been widely used for machine learning experiments. The framework proposes a way to assess synthetic data sets from the perspective of utility, fidelity and privacy. Utility is measured as fitness-of-purpose for the synthetic data to be used for machine learning applications, while fidelity is measured as the statistical distance and structural similarities between the synthetic data and the original data. Finally, privacy is measured as the statistical distance between the synthetic data and the original data. This combination fulfils the conflicting requirements for synthetic data to be meaningful, in that it both sufficiently similar and different from the original data. This framework also proposes a weighted average of the different measures for utility, fidelity and privacy for greater flexibility. In terms of originality, as far as it is known, this framework is the first framework which introduces a weighting mechanism to fine-tune the assessment of the synthetic dataset for privacy, utility or fidelity.

2. Synthetic data

Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s). Here, synthetic data is contrasted with real data. Real data is data generated not by a model but by a real-world system, e.g. data that is submitted by a central bank's regulatees.

The synthetic data generator may include models such as gaussian mixed models (GMM), Variational Auto-encoders (VAE), Copula models, agent based or econometric models or even stochastic differential equations. Models which are used in this paper are mentioned in Section 3.2.

Synthetic data has proven to be valuable in scenarios where real data is limited, for instance, when generating synthetic data derived from original datasets that were costly to obtain or within financial systems that have only recently commenced data collection. In such cases, the synthetic data serves as training data for machine learning applications through a process known as data augmentation. Effective data augmentation methodologies must also address the elimination or reduction of biases that may arise due to historical biases, while maintaining statistical accuracy.

This paper will focus on the utilization of synthetic data for data sharing purposes. Machine learning solutions fundamentally rely on high-quality training data, including data owned by central banks. In this context, it would be beneficial to share synthetic versions of the relevant data. It is essential for the synthetic data to exhibit statistical similarity to the original data, as the underlying assumption is that machine learning models trained on statistically similar synthetic data will yield results comparable to those obtained using the original data.

2.1 Synthetic data and central bank

In this section, the literature concerning the usage of synthetic data by central bank practitioners is presented. From the literature search, it is observed that current efforts on using synthetic data in central banks has focused on either enabling access to data necessary for research or has focused on assessment on synthetic data from a utility perspective.

The Deutsche Bundesbank Research Data and Service Centre (RDSC) provides researchers with access to micro data for research purposes, up to individual data with the corresponding restrictions. To overcome further confidential restrictions as well as to enable data sharing to be done with researchers who are unable to access the, synthetic data of micro data such as the Survey of Income and Program Participation is created by adding noise to the data (Drechsler, 2011)

Banking microdata collected by the Central Bank of Paraguay from 2017-2023 in the areas of financial inclusion, granular term deposits, and credit risk management were used to create synthetic data for the purpose of data sharing. The synthetic data created were evaluated for utility in these areas (Caceres & Moews, 2024).

Banco de Espana has also conducted preliminary efforts to create synthetic data and assessed the synthetic data for both utility and confidentiality. The targeted data was microdata for individual non-financial firms and microdata for loans to legal persons (Lapteva, 2023).

Bangko Sentral ng Pilipinas (Ong et al., 2024) has also reported their efforts to create synthetic data for the purposes of data sharing. In their work, the Consumer Expectation Survey was chosen to be used as the data to create synthetic data. In their work, assessment was performed on the synthetic data from a fidelity, utility and privacy perspective.

From the literature, central banks have shown increasing appetite to use synthetic data as a means of sharing data. Early efforts have focused on micro-data, with an interest on utility and privacy rather than fidelity.

2.2 Models for synthetic data generation

The models for creating synthetic data or synthetic data generators (SDG) are varied. It ranges from rule-based systems, random sampling, statistical approaches, differential privacy approaches, and machine learning approaches. A comprehensive survey of all these methods are presented in (Goyal & Mahmoud, 2024).

Of the machine learning approaches, generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs) are the most popular in recent years.

Generative Adversarial Networks (GANs) provide a powerful method for generating synthetic data using machine learning. They consist of two components: a generator that creates data and a discriminator that evaluates its authenticity. Through a process of iterative competition, the generator improves its ability to produce data that closely mimics real-world datasets.

In contrast, Variational Autoencoders (VAEs) take a probabilistic approach to data generation. They encode input data into a compressed latent space and then decode it back, enabling the creation of new data points sampled from this latent space. This ensures that the synthetic data retains the key patterns and variability found in the original dataset.

Large language models (LLMs) like generative pre-trained transformer (GPT) offer another approach, particularly for text-based data. By training on vast datasets, these models can generate text that is coherent and contextually accurate, often indistinguishable from text written by humans. This makes them highly effective for applications requiring natural language processing.

From the papers published by central banks, SDG that were used included gaussian diffusion models (GDM) , synthetic minority oversampling technique (SMOTE), gaussian Copula (GC), conditional tabular generative adversarial network (CTGAN), tabular variational autoencoder (TVAE) and gaussian mixture models (GMM).

2.3 Assessing synthetic data

Any synthetic data that is generated using an SDG needs to be assessed. This is more so for synthetic data that is produced and shared by central banks, due to the need for central banks to maintain the highest levels of trust to execute their mandates effectively. Performance metrices for synthetic data in the areas of utility, privacy, utility and fidelity are described in (Osorio-Marulanda et al., 2024).

2.3.1 Privacy

Privacy metrics for synthetic data aim to quantify the potential disclosure risk regarding the original source data. The required stringency and type of metric depend heavily on the context; metrics suitable for internal use may differ from those needed for public releases.

Primarily, robust privacy metrics are derived from frameworks like differential privacy. These metrics typically evaluate the data generation algorithm, providing a mathematically rigorous measure (e.g., privacy budget parameters like ϵ and δ) of the maximum statistical information leakage allowed by the process across all outputs. They offer a systematic way to analyse the inherent privacy safeguards of the method used.

A significant limitation, however, exists in developing reliable metrics to assess the privacy of a specific, already generated synthetic dataset, especially without knowledge of its provenance. Current state-of-the-art metrics characterise the properties of the generation process itself, rather than providing a definitive privacy score for a single data instance.

It is with this in mind that the work in this paper used the T-measure, as it measures the outcome of the generative process.

2.3.2 Utility

Utility evaluation metrics encompass techniques that evaluate the general performance of models, algorithms, or data. It involves evaluating how well the utility of the original data is maintained in the synthetic version of the data. Traditional metrices related to machine learning operations such as accuracy, precision, recall, F1 score, area under curve (AUC) are useful utility metrices.

Utility measures which have been reported by central banks for assessing their synthetic data efforts includes accuracy, AUC, recall, precision, F1 scores , Kappa, Matthews Correlation Coefficient (MCC) and error measures such as root mean square error (RMSE). The choice of a utility evaluation measure will be determined by the use case in which the synthetic data is to be used.

The utility evaluation measure is frequently compared to the privacy evaluation measure. It is crucial for balancing the trade-offs between privacy and utility in various applications.

2.3.3 Fidelity

Fidelity refers to metrics that directly compare the synthetic dataset with the real one, rather than through a model or task performance. It assesses how well the synthetic data statistically matches the original data. These measures are often used based on the intuition that a specific level of fidelity will lead to better performance across various machine learning tasks.

3. Methods

3.1 Framework

The purpose of the proposed framework is to perform assessment of the dataset created by the SDG for the dataset's privacy, utility and fidelity.

The framework is divided into three distinct phases, which is described in more detail below:

Phase 1 - Data Preprocessing

In this phase, the raw dataset was ingested for analysis, after which non-informative features like unique identifiers (e.g. ID columns) were systematically removed to prevent data leakage. The remaining features were then divided into two categories: categorical features (e.g., pay status) and continuous features (e.g., age). Following this, feature-specific preprocessing techniques were applied; categorical variables underwent transformation via one-hot encoding to enable appropriate interpretation by machine learning models, while continuous variables were normalised using Min-Max scaling, rescaling values to a [0,1] range for uniformity and stability during model training. This preprocessing phase resulted in a fully cleaned and standardized dataset suitable for the subsequent phase.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure X: Min-Max scaling

Where:

- X is the original value,
- X_{min} is the minimum value in the dataset,
- X_{max} is the maximum value,
- $X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$ is the normalised value in the range [0,1][0, 1][0,1].

Phase 2 - Synthetic Data Generation

The pre-processed dataset from Phase 1 is then utilized as the basis for synthetic data generation, beginning with the extraction of feature vectors to serve as input for the SDG . The SDG is fitted to the feature distribution to learn the underlying probability density of the original data. Subsequently, new synthetic samples were generated by sampling from the trained SDG. To maintain interpretability, these synthetic samples were inversely transformed back to the original feature scales where applicable. Finally, synthetic labels were assigned to the generated data points using a pre-trained classifier, ensuring consistency with the original label distribution. This phase resulted in a synthetic dataset designed to closely mimic the statistical and structural properties of the original dataset

Phase 3 – Evaluation

In this phase, the synthetic dataset created by the SDG was evaluated for privacy, utility and fidelity across four key dimensions:

(i) Statistical distance (for fidelity) was assessed using the Wasserstein distance between real and synthetic data distributions.

The Wasserstein distance is given by the following formula,

$$W(p,q) = \inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

Where:

- $p\ and\ q$ are the distributions (often represented as probability measures),
- $\Gamma(p,q)$ is the set of all **joint distributions** (couplings) with marginals $p\ and\ q$
- $[\|x-y\|]$ is typically the **Euclidean distance** between x and y,
- The infimum finds the optimal transport plan (minimum cost to transform one distribution into the other)

For 1D empirical distributions (single column), the Wasserstein-1 distance simplifies to:

$$W(p,q) = \int_0^1 |\, F^{-1}(t) - G^{-1}(t)\,|\, dt$$

Where:

- F^{-1} and G^{-1} are the quantile functions (inverse CDFs) of p and q

The final score is $S_{SD} = 1 - \frac{1}{d}\sum_{i=1}^{d} W_1^{(i)}$ where :

- d is the number of columns (features),
- $W_1^{(i)}$ is the **Wasserstein distance** between the i-th feature in the real and synthetic datasets,
- $S_{SD}\ \ is$ the normalised **Statistical Distance Score**: higher values indicate better similarity.

(ii) Structural similarity (for fidelity) was investigated via comparisons of feature correlations and marginal distributions between the real and synthetic data.

The **correlation similarity** compares Pearson correlation matrices from real ρ_{real} and synthetic ρ_{syn} datasets. The mean absolute error (MAE) between the matrices is used to compute the score:

$$S_{\text{corr}} = \frac{1}{1 + \text{MAE}(\rho_{\text{real}}, \rho_{\text{syn}})}$$

Additionally, **distribution similarity** is assessed using the Kolmogorov-Smirnov (KS) test for each feature. For a given column, the KS statistic D_n is calculated as the maximum difference between empirical CDFs, and the normalised score is:

$$S_{\text{KS}}^{(i)} = 1 - D_n^{(i)}$$

The final normalised structural similarity score combines these two measures:

$$S_{SS} = \frac{1}{2}\left(S_{\text{corr}} + \frac{1}{d}\sum_{i=1}^{d} S_{\text{KS}}^{(i)}\right)$$

(iii) Machine learning utility was then measured through downstream task performance, encompassing classification accuracy, clustering quality, and time series analysis.

For a classification task (in this work, it was classification of the default payment label using a Random Forest classifier), the S_{ML} normalised Machine Learning Score is given by

$$S_{ML} = 0.5 \cdot (\text{Accuracy} + \text{F1})$$

Additional unsupervised tasks, such as clustering (K-Means) and time series forecasting (Linear Regression on lagged financial features), were evaluated using Silhouette Score and Mean Squared Error (MSE), respectively but were not included in the ML score.

(iv) Privacy assessment was conducted using the T-measure score to estimate the risk of re-identification, with the overall T-measure score across all features calculated as the average of the absolute T-measure.
The formula for the T-measure for a single feature is given by

$$T = \frac{\bar{x}_r - \bar{x}_s}{\sqrt{\frac{s_r^2}{n_r} + \frac{s_s^2}{n_s}}}$$

Where:

- $\bar{x}_r$ = mean of the feature in the **real** dataset
- $\bar{x}_s$ = mean of the feature in the **synthetic** dataset
- s_r^2, s_s^2 = variances of the feature in real and synthetic datasets respectively
- n_r, n_s = number of samples in real and synthetic datasets

The formula for the overall T-measure score across all features is given by

$$T_{\text{score}} = \frac{1}{d}\sum_{i=1}^{d} | T_i |$$

Where:
d = number of features
T_i = t-measure for feature i

For each numerical feature, the standardized mean difference is calculated as:

$$T_{\text{col}} = \frac{| \mu_{\text{syn}} - \mu_{\text{real}} |}{\sigma_{\text{real}}}$$

where μ and σ represent mean and standard deviation, respectively.
The overall privacy score is:

$$S_P = \frac{1}{1 + \frac{1}{d}\sum_{i=1}^{d} T_{\text{col}}^{(i)}}$$

Each of these evaluation metrics was assigned an equal weight of 25% to compute a composite final weighted score, providing a balanced assessment of the synthetic data's quality by capturing its ability to preserve essential patterns, support model training, and protect privacy.

$$\text{Composite Final Weighted Score} = \sum_{i \in \{SD, SS, SML, SP\}} w_i \cdot S_i$$

The weightage may be adjusted according to the use case for which the synthetic data , adjusting for the weightage for privacy (privacy score), utility (machine learning score and statistical distance score) and fidelity (structural similarity score).

3.2 Experiment

To demonstrate the usage of the generalised weighted framework for synthetic data evaluation, experiments were conducted to firstly generate the synthetic data set using different SDG, and secondly to use the framework to elicit the weighted score.

SDG strengths and weaknesses

Table 1

Name of SDG	Strengths	Weaknesses
Gaussian Diffusion Models (GDM)	• High flexibility in modelling distributions Good for high-dimensional data	• Computationally intensive Requires substantial amounts of training data
Gaussian Copula (GC)(Patki et al., 2016)	• Maintains statistical properties Suitable for tabular data	• Assumes a Gaussian dependence structure Struggles with complex, non-linear relationships
Conditional Tabular GAN (CTGAN)(Xu et al., 2019)	• Manages imbalanced and multi-modal data Captures complex dependencies	• Training instability Sensitive to hyperparameter tuning
Tabular Variational Autoencoder (TVAE)(Xu et al., 2019)	• Effective in capturing latent structures Suitable for tabular data	• May overfit on small datasets Needs careful parameter tuning
Gaussian Mixtures Models (GMM)(Dempster et al., 1977)	• Model data as a combination of distributions Interpretable and straightforward	• Assumes data can be represented as Gaussian mixtures Sensitive to initialization
Time Dependent Self Attention (TDSA)*	• Strong at time-series data generation Preserves temporal correlations	• Limited to time-series data Computationally expensive for large datasets
Tabular Diffusion Probabilistic Model (TabDPM)(Kotelnikov et al., 2024)	• Robust for tabular data with complex dependencies Flexible with varying distributions	• Computationally heavy Requires careful parameter configuration

Notes: The TDSA algorithm is the author's own algorithm that a neural network with a self-attention mechanism to predict noise added at different time steps based on the noisy data and the current time step. Data generation starts by gradually adding noise to real data over many steps in a forward diffusion process. Finally, new data is synthesized by starting with random noise and iteratively using the neural network's predictions to reverse the noising process step-by-step.

3.2.1 Data set

The synthetic data originates from an open-source credit card dataset chosen for its suitability in assessing utility through machine learning applications. This dataset contains 30,000 entries (excluding the header) with information about credit card clients (Yeh & Lien, 2009). Key features encompass core financial details like credit limit balance, bill amounts for six consecutive months, and payment amounts for the same six months. Additionally, it includes demographic information such as the client's gender, age, education level, and marital status , along with a seven-month payment status history. The target variable is a binary indicator (0 or 1) labelled "Default payment next month," which predicts potential client default in the following month.

3.2.2 Experimental Results

Experiments were conducted to evaluate the proposed framework whereby seven SDG were used to create 30,000 synthetic elements using the original or real dataset, $dataset_{real}$. The synthetic dataset, $dataset_{syn}$ were evaluated using the measures described in Section 3.3.3, to elicit the normalised scores for the statistical distance, structural similarity, machine learning and privacy measures. The parameters of the experiments are as follows

1. Measures for statistical distance, structural similarity, machine learning and privacy measures are equal at 0.25 (Figures 1-4)
2. Time-series and clustering tasks are added to the classification tasks (Figure 5)
3. Measures for privacy and machine learning are over-weighted (Figure 6)

Overall weighted score for various SDG

Figure 1

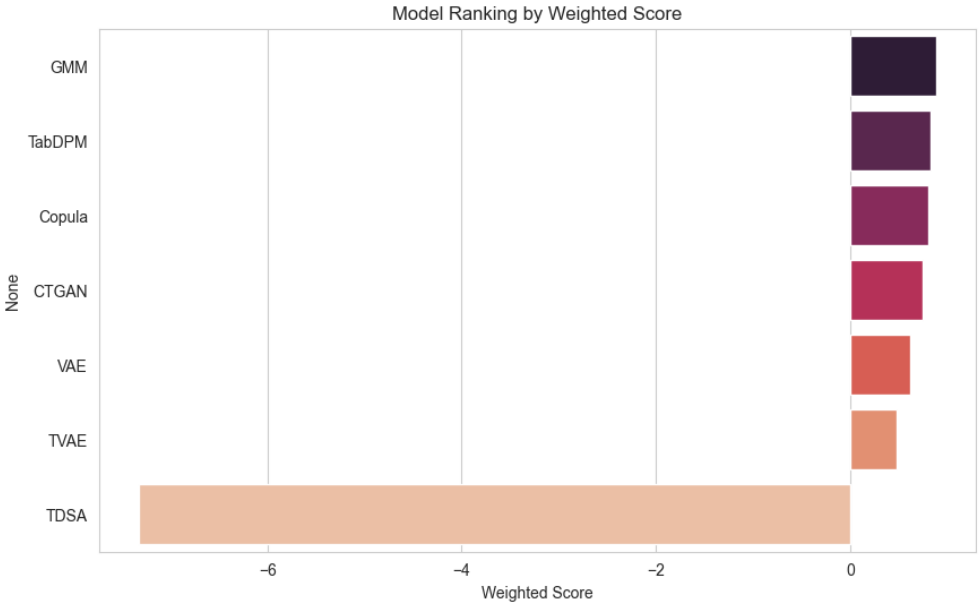


Figure 1 shows that the GMM performed the best, followed by TabDPM and GC. All three models showed a weighted score of above 0.80 on a scale of 0 to 1.0. This shows that the evaluation scheme can measure the privacy, utility and fidelity of the

synthetic dataset using a single measure. The poor results of the synthetic dataset generated by TDSA can be attributed to its high Wasserstein distance score

Privacy scores for various SDG

Figure 2

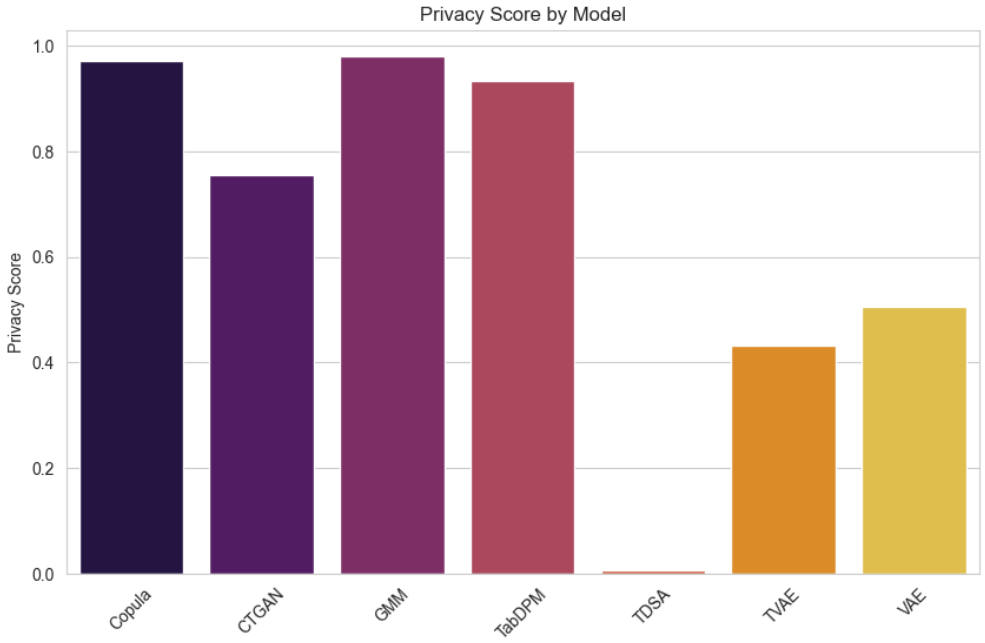


Figure 2 shows that the GC, GMM and TabDPM SDG resulted in synthetic data sets which has high privacy scores exceeding 0.8 on a scale of 0 to 1.0. TDSA which focuses on time-series with self-attention performed the worst

Utility scores for various SDG

Figure 3

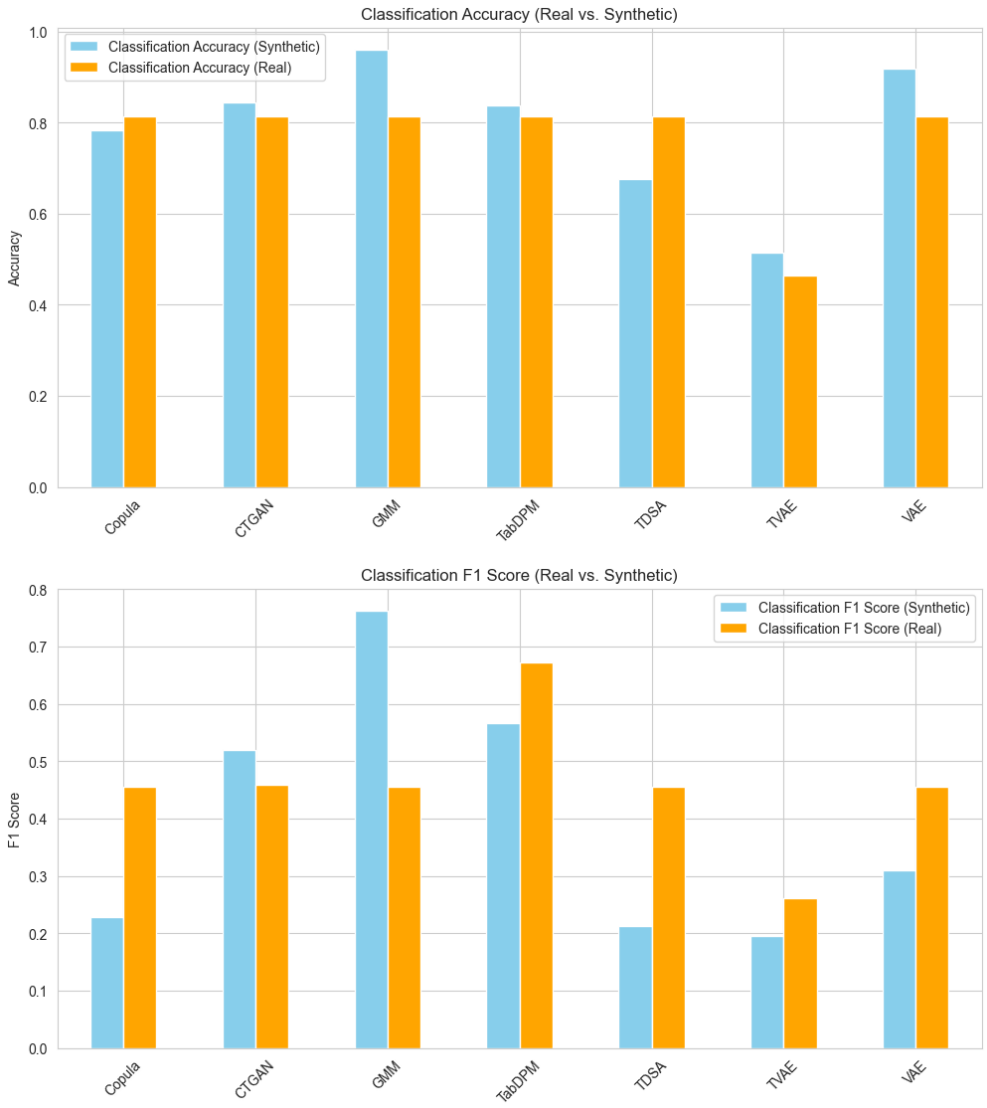


Figure 3 shows the utility scores for various datasets created using the same set of SDGs. For classification measured using accuracy, five out of seven SDG resulted in synthetic datasets which returned higher classification scores as compared to the original dataset. For classification measured using the F-1 score, four out of seven SDG resulted in the real dataset returning higher classification scores.

Fidelity scores for various SDG

Figure 4

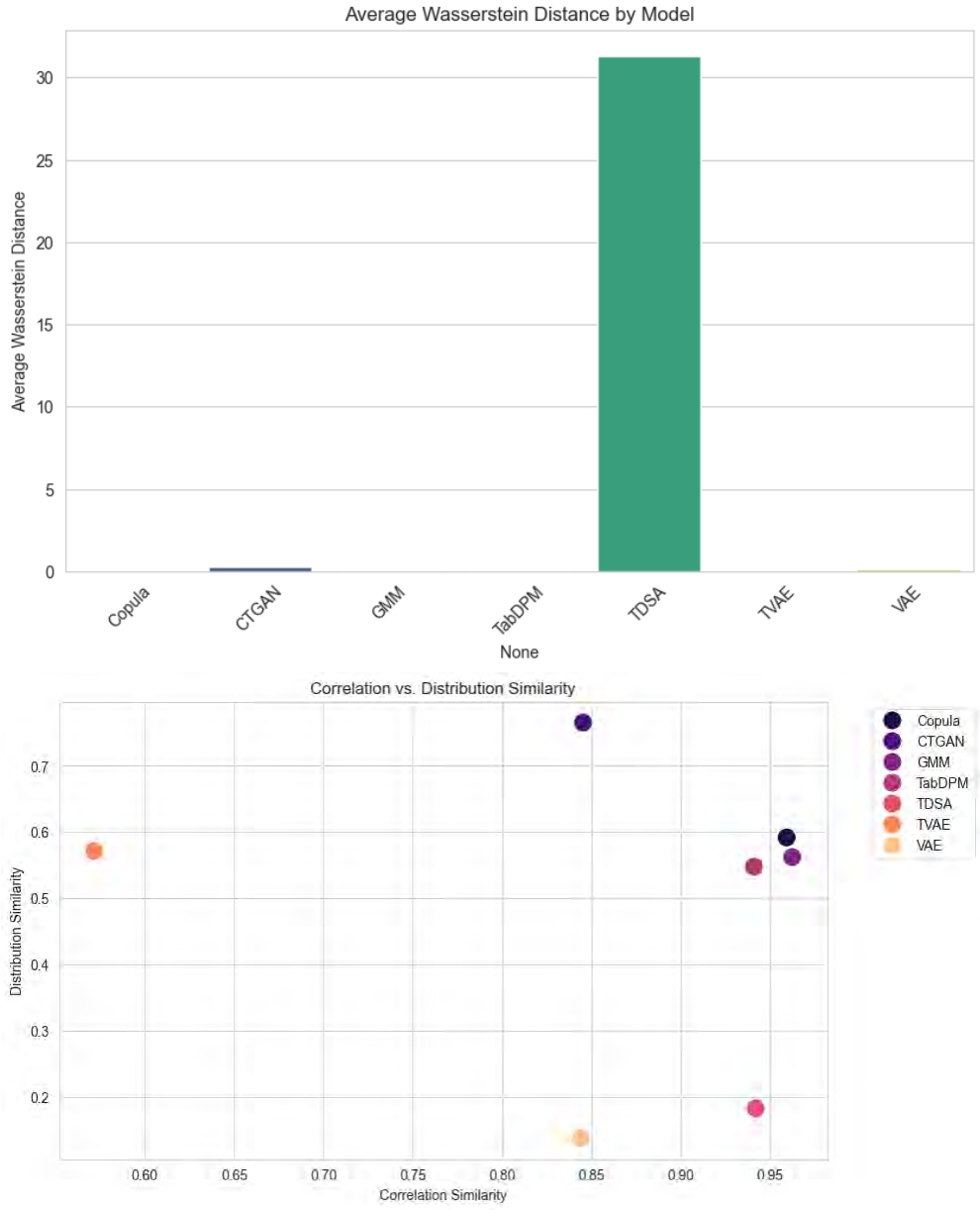


Figure 4 shows the results for synthetic datasets created using the same set of SDGs. Four out of seven SDGs resulted in synthetic datasets which has high correlation similarity (above 0.9 on a scale of 0-1.0), while only CTGAN resulted in a synthetic dataset which has distribution similarity of above 0.7 on a scale of 0 to 1.0.

Privacy scores vs combined ML score

Figure 5

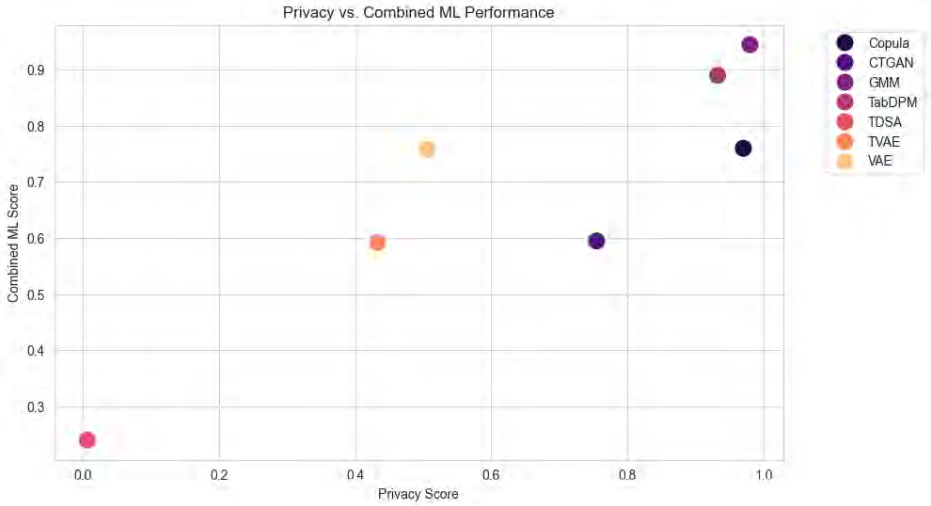


Figure 5 shows the scatter plot of the privacy scores for various synthetic datasets versus their corresponding combined machine learning scores. The combined machine learning score is normalised to fit into a range of 0 and 1 and consist of weighted classification, time-series forecast (use five months of bill and payment amount to predict the sixth month) and clustering scores for five clusters. The weightage for classification to time-series forecast and clustering were 4:3:3, respectively. From Figure 5, it can be observed that only the GMM and TabDPM SDG exceeds the 0.8 score for both combined ML and privacy

At the same time, the classification accuracy of synthetic datasets which is greater than 0.5, 0.6 and 0.8 are 0.79, 0.77 and 0.78 respectively (not shown in Figure 5). While the differences are marginal, it does suggest that there exist inherent trade-offs between privacy and ML performance.

Differential weights for privacy vs ML

Figure 6

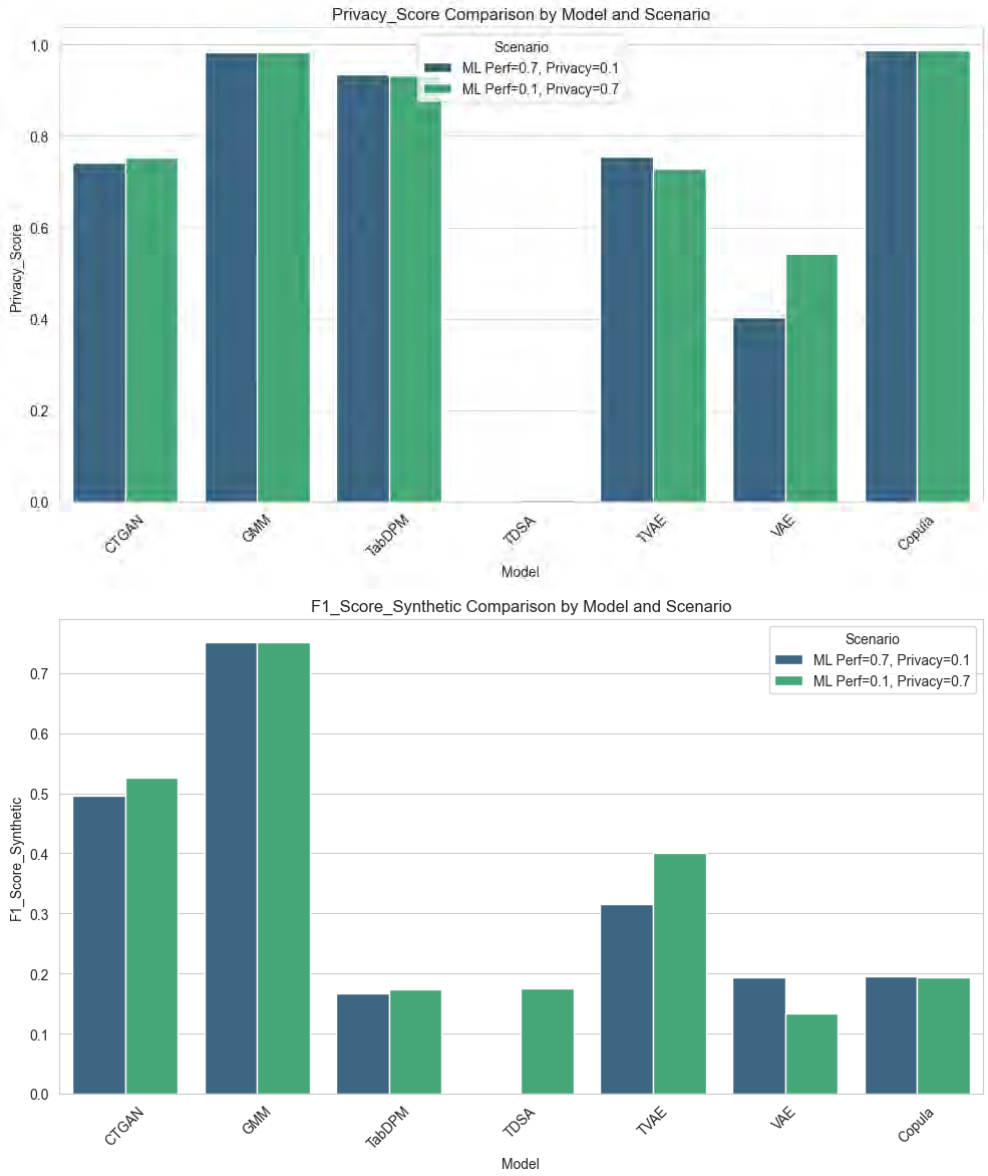


Figure 6 shows the effects of changing the weights for the measures of statistical distance, structural similarity , machine learning and privacy measures. Two scenarios were investigated i.e.

1. Privacy first scenario where weights for measures of statistical distance, structural similarity, machine learning and privacy are 0.1, 0.1, 0.1 and 0.7 respectively
2. Utility first scenario where weights for measures of statistical distance, structural similarity, machine learning and privacy are 0.1, 0.1, 0.7 and 0.1 respectively

Prioritising privacy (ML Perf=0.1, Privacy=0.7) leads to result in similar or slightly higher privacy score values for most models. TVAE shows a slight decrease in privacy score when privacy is weighted higher. Prioritizing privacy (and thus lowering the weight for ML performance) leads to a lower or similar f1-score for the synthetic

data (which might reflect ML utility) for models like VAE and GC. However, for CTGAN and TVAE, the F1 score increases slightly when privacy is prioritized. For TDSA, it increases from zero. For GMM and TabDPM, it remains relatively constant.

Privacy prioritisation also led to a higher composite weighted* score for most models (CTGAN, GMM, TabDPM, TVAE, GC) compared to prioritizing ML performance (ML Perf=0.7, Privacy=0.1). The exception is VAE, where the score slightly decreases, and TDSA, where the score increases significantly but remains negative.

Prioritising machine learning (ML Perf=0.7, Privacy=0.1) led to f1-score for synthetic data being *higher or similar* for several models (GMM, TabDPM, VAE, GC) when ML performance is weighted higher. However, this is not universal; CTGAN, TVAE show a *lower* F1 score in this scenario, and TDSA has an F1 score of zero.

Machine learning prioritisation has led to overall composite weighted score* to be *lower* for most models (CTGAN, GMM, TabDPM, TVAE, GC) compared to when privacy is weighted higher. This suggests that, according to the weighting formula used, the gains in ML performance might not fully compensate for the potential reduction in privacy or other factors in the overall score. VAE is a slight exception, showing a marginally higher weighted score, while TDSA shows a drastically lower (more negative) score. The privacy score is often *slightly lower or similar* compared to when privacy is weighted higher. This is expected, as less emphasis is put on the privacy component. TVAE is an exception, showing a slightly higher privacy score in this scenario.

In summary, weighting ML performance higher often results in slightly lower privacy scores and potentially higher F1 scores (though not always), but it tends to decrease the overall composite weighted score for most models in this dataset and can increase computation time for several methods. This highlights the trade-off being evaluated: prioritizing ML performance directly does not necessarily lead to the best overall outcome as defined by the Weighted score metric in this specific evaluation.

Source*: Author own calculation

4. Discussion

The experimental findings highlight the effectiveness of the proposed weighted evaluation framework in assessing synthetic data quality across privacy, fidelity, and utility dimensions. Simpler generative models like the GMM achieved the highest composite final weighted scores, with GMM, TabDPM, and GC all scoring above 0.80. These models produced synthetic datasets that balance statistical similarity, structural similarity, machine learning utility, and privacy preservation well. In contrast, the TDSA method performed poorly due to high statistical divergence from real data, significantly lowering its overall score.

The framework effectively discerns nuanced differences in generator performance, providing a comprehensive measure of quality. It aligns with recent efforts to develop unified metrics for synthetic data evaluation, offering flexibility to emphasize various aspects via weighting.

Fidelity Dimension: Many models preserved global structure well but differed in distributional details. Four of the seven SDGs produced synthetic data with high correlation similarity to real data, implying intact inter-feature relationships. Distributional similarity proved more challenging, with only CTGAN achieving a Kolmogorov–Smirnov distribution similarity above 0.7.

Privacy Dimension: Most methods produced synthetic data with strong privacy guarantees, with GC, GMM, and TabDPM yielding privacy scores exceeding 0.8, indicating low re-identification risk. TDSA had the weakest privacy performance, suggesting its synthetic data retained more identifiable traces of the original data.

Machine Learning Utility: Results were mixed. In classification tasks, five out of seven synthetic datasets outperformed real data on accuracy, hinting that certain generative models can produce synthetic samples easier for classifiers to learn. However, when evaluating classification by the F1-score, the real dataset outperformed synthetic data in most cases, indicating some synthetic data may not capture minority class characteristics or nuanced decision boundaries as effectively as real data.

The framework's ability to aggregate these facets into a single weighted score provides a convenient summary, yet disaggregated scores remain crucial for diagnosing specific strengths or weaknesses of a synthetic dataset.

A key insight is the inherent trade-off between privacy and utility in synthetic data generation. Only GMM and TabDPM achieved both high privacy and high utility simultaneously, suggesting these techniques were most effective at generating data that is both safe and useful. Other methods fell short in one dimension when excelling in the other. For example, CTGAN and TVAE produced strong utility but lower privacy scores, whereas the GC method excelled in privacy at the expense of slight utility reduction.

The weighted evaluation framework quantitatively captures this balance, giving practitioners a transparent view of how much utility is gained for a given privacy cost, supporting informed decisions on which synthetic data method best fits their risk appetite and analytical needs.

The framework's flexibility to reconfigure the importance of each dimension to suit specific use cases is a significant advantage. Ablative experiments revealed how different prioritizations impact the composite final weighted score and highlighted model-specific behaviours. Weighting privacy higher led to higher overall scores for most models compared to the ML-heavy scenario. This suggests that most generators were better at improving privacy than boosting predictive performance.

In summary, the discussion confirms that the generalized weighted framework captures multiple facets of synthetic data fidelity and utility and privacy in a unified manner, providing a tunable mechanism to explore trade-offs between statistical similarity, analytic utility, and privacy. This holistic perspective is essential for informed decision-making on synthetic data adoption in sensitive domains.

5. Extension

Looking ahead, several promising directions can extend this work to enhance the evaluation framework and facilitate its adoption in practice. These extensions aim to

improve the framework's robustness, broaden its applicability, and align it more closely with real-world needs, particularly in central banking data use cases.

Adaptive or Learned Weighting Strategies: Enhancing the framework to make the weighting adaptive is important. The optimal balance between privacy, fidelity, and utility may not be obvious a priori. Future research could explore algorithms that learn weight coefficients based on specific objectives or constraints, such as multi-objective optimization or reinforcement learning to adjust weights until synthetic data meets target criteria. An adaptive weighting system might consider the context of deployment, automatically assigning higher weight to privacy if sensitive personal identifiers are detected or increasing weight on structural similarity if the downstream task is sensitive to data distribution shifts. Studying how weight tuning affects computational efficiency and model selection can inform practitioners about which techniques are best suited under different priorities. A data-driven or context-aware weighting mechanism reduces reliance on expert judgment and makes the evaluation more objective and personalized.

Enhancing Metrics and Privacy Safeguards: Updating the framework to include improved metrics as privacy and data utility research evolves is essential. Exploring alternative privacy metrics or stronger confidentiality guarantees, such as differential privacy parameters, can provide more formal privacy assurance. Integrating measures of model generalizability or fairness ensures synthetic data supports accurate models without introducing bias. Evaluating whether models trained on synthetic data exhibit similar fairness properties as those trained on real data adds another facet to the quality assessment. Using the weighted score as a feedback signal to train or select better synthetic data generators can push the frontier of synthetic data toward higher quality and reliability.

6. Conclusion

In conclusion, the generalized weighted framework opens numerous avenues for further investigation. By pursuing the extensions above, the framework can be applied as a practical decision-support tool for organizations like central banks that are keen to leverage synthetic data. The ongoing exploration of synthetic data in central banking – whether for data sharing, open research, or enhancing AI models with limited real data – stands to benefit from a robust evaluation standard. Our future work will focus on refining this framework, validating it in collaborative settings, and ensuring it remains aligned with both technological advances in synthetic data generation and the evolving requirements of data governance in the financial industry. Through continuous improvement and adaptation, we anticipate that this framework, with its weighting flexibility and multi-dimensional perspective, can become an integral part of how synthetic data's value and risks are assessed in practice, facilitating greater trust and wider adoption of synthetic data solutions in sensitive domains.

References

Caceres, H. E., & Moews, B. (2024). *Evaluating utility in synthetic banking microdata applications* (arXiv:2410.22519). arXiv. https://doi.org/10.48550/arXiv.2410.22519

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation* (Vol. 201). Springer. https://doi.org/10.1007/978-1-4614-0326-5

European Union. (2024). *Article 10: Data and Data Governance | EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/article/10/

Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, *13*(17), Article 17. https://doi.org/10.3390/electronics13173509

Irving Fischer Committee on Central Bank Statistics. (2023). *Guidance Note 3 – Data-sharing practices*.

Jordon, J., Houssiau, F., Cherubin, G., Cohen, S. N., Szpruch, L., & Bottarelli, M. (2022). *Synthetic Data—What, why and how?* Alan Turing Institution.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2024). *TabDDPM: Modelling Tabular Data with Diffusion Models*. https://doi.org/10.5555/3618408.3619133

Lapteva, E. K. (2023). *Utility and confidentiality assessment of synthetic company data*. IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks."

Ling, K. L. F. (2013). *Application of micro-data for systemic risk assessment and policy formulation*. Workshop on Integrated Managemet of Micro-databases, Porto.

Ong, C. A., Escalanda-Lo, C., Gabriel, M., & Reyes, R. (2024). Research for All: Exploring machine learning applications in generating synthetic datasets. *2024 Philippines Statistical Association Annual Conference*. https://events.psai.ph/2024/ac/

Osorio-Marulanda, P. A., Epelde, G., Hernandez, M., Isasa, I., Reyes, N. M., & Iraola, A. B. (2024). Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review. *IEEE Access*, *12*, 88048–88074. https://doi.org/10.1109/ACCESS.2024.3417608

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Generalised Weighted Framework for Synthetic Data Evaluation[1]

Chiung Ching Ho BNM

1 This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, Bank Negara Malaysia, the BIS, the IFC or the other central banks and institutions represented at the event.

Overview

Privacy, utility and fidelity for trust-worthy synthetic data

There is a need to use synthetic data by Central Banks
driven by both legislation and a need to maintain privacy of data

Synthetic data produced by Central Banks needs to be trusted
to fulfil privacy, utility and fidelity (PUF) requirements

An assessment framework is proposed to evaluate synthetic data
produced by synthetic data generators (SDG) that assesses PUF requirements using a flexible and extensible framework

Index

1. Introduction
   - ❑ Research motivation
   - ❑ Synthetic data sharing and central banks
2. Considerations for synthetic data
   - ❑ Privacy, utility , fidelity
3. Techniques
4. Framework for synthetic data evaluation
5. Experiment
   - ❑ Data
   - ❑ Results
6. Discussion
7. Summary and recommendations
8. References
9. Addendum

Generalised Weighted Framework for Synthetic Data Evaluation

Motivation : Inspired by an attempt to create a universal metric for evaluation of synthetic data [1], a new framework is proposed that will allow for flexible assessment of synthetic data from a privacy, utility and fidelity (P.U.F) perspective

Problem statement : Data created by SDG needs to be evaluated for privacy, utility and fidelity (PUF) to ensure that synthetic data is credible to be shared or published by central banks or authorities, and current solutions does not allow for flexibility in measuring SDG that balances P.U.F needs

Research Question (RQ) 1 : Which of current SDG techniques exhibit good PUF scores?

Research Question (RQ) 2 : How do we balance P.U.F measures for different SDG to address different analytical needs?

Assessing synthetic data

- Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s)[2]
- Synthetic data is being used by central banks to enable data sharing without compromising on privacy or infringing on legislation[3]
- Reasons for sharing data using synthetic data in central banks includes
  - Sharing micro data for research purposes[4]
  - Justification of data collection rigor [5]
  - Data disclosure risk mitigation and[6]
  - Sharing of granular data instead of aggregated data[7]
- Examples of synthetic data shared by central banks includes micro data from surveys, non-financial firm-level data, and loans to legal persons[4][5][6][7]

Synthetic data generation (SDG) techniques comparison

Name of Technique	Strengths	Weaknesses
Gaussian Diffusion Models (GDM)	High flexibility in modelling distributions Good for high-dimensional data	Computationally intensive Requires large amounts of training data
Gaussian Copula (GC)	Maintains statistical properties Suitable for tabular data	Assumes a Gaussian dependence structure Struggles with complex, non-linear relationships
Conditional Tabular GAN (CTGAN)	Handles imbalanced and multi-modal data Captures complex dependencies	Training instability Sensitive to hyperparameter tuning
Tabular Variational Autoencoder (TVAE)	Effective in capturing latent structures Suitable for tabular data	May overfit on small datasets Needs careful parameter tuning
Gaussian Mixtures Models (GMM)	Models data as a combination of distributions Interpretable and straightforward	Assumes data can be represented as Gaussian mixtures Sensitive to initialization
Time Dependent Self Attention (TDSA)	Strong at time-series data generation Preserves temporal correlations	Limited to time-series data Computationally expensive for large datasets
Tabular Diffusion Probabilistic Model (TabDPM)	Robust for tabular data with complex dependencies Flexible with varying distributions	Computationally heavy Requires careful parameter configuration

Table 1 : A comparison of SDG techniques

Assessing synthetic data : P.U.F for credible synthetic data

Privacy

- Is x in the dataset?
- Can unknown data about x be found?
- Can I recreate the original data?

Utility

- How does the synthetic data perform as compared to the original data in specific task?

Fidelity

- How truthful is the synthetic data?
- How statistically similar is the synthetic data?
- How similar is the distribution of the synthetic data?

There is always a trade-off between privacy and utility, while fidelity may proxy for utility

Generalised Weighted Framework for Synthetic Data Evaluation

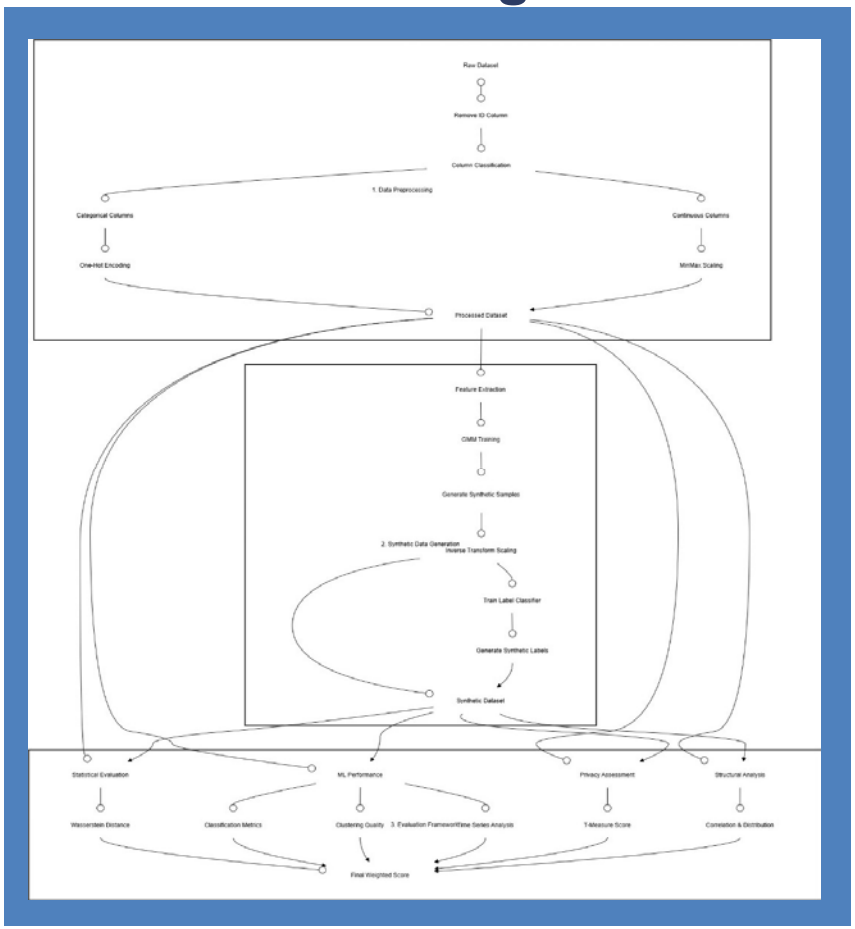


Figure 1: Synthetic data evaluation process flow

Phase 1 - Data Preprocessing
•Start with the raw dataset
•Remove any ID columns
•Classify columns into two types:
- Categorical (like SEX, EDUCATION, MARRIAGE, PAY status)
- Continuous (like AGE, BILL amounts, PAYMENT amounts)

•Process each type differently:
- Categorical columns get one-hot encoded
- Continuous columns get scaled using MinMax scaling

•Result: A cleaned, processed dataset

1.Preprocessing is dependent on data set

Phase 2 - Synthetic Data Generation
•Take the processed dataset
•Extract features for GMM training
•Train the Gaussian Mixture Model
•Generate synthetic samples
•Convert the synthetic data back to original scale
•Generate appropriate labels using a classifier
•Result: A synthetic dataset that mimics the original

2. The synthetic data generator (SDG) is chosen based on use case

Phase 3 - Evaluation
•Compare real and synthetic data across four aspects:
- Statistical similarity (using Wasserstein distance)
- Machine learning performance (classification accuracy, clustering quality, time series analysis)
- Privacy assessment (using T-measure score)
- Structural analysis (correlations and distributions)

•Combine all metrics into a final weighted score
Each metric contributes equally (25%) to the final evaluation score, which helps assess how well the synthetic data captures the important characteristics of the original dataset while maintaining privacy.

3. Privacy, Utility (ML scores) and Fidelity (statistical and structural) measures can be changed or weighted as needed

Figure 2: Synthetic data evaluation measures using Gaussian Mixture Models as a Synthetic Data Generator

Experiment using credit card data

The synthetic data was created from an open source credit card dataset*
This dataset was chosen as it was suitable for assessing utility through machine learning applications
It has 30,000 entries (excluding header) containing information about credit card clients. Here are the key features:

Financial

- LIMIT_BAL: Credit limit balance
- BILL_AMT1 through BILL_AMT6: Bill amounts for six consecutive months
- PAY_AMT1 through PAY_AMT6: Payment amounts for six consecutive months

Demographics

- SEX: Gender of the client
- AGE: Age of the client
- EDUCATION: Education level
- MARRIAGE: Marital status

Historical payment

- PAY_0 through PAY_6: Payment status history for seven months
- "Default payment next month": Binary indicator (0 or 1) predicting whether the client will default on their payment in the next month

* Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Fidelity : Measuring Wasserstein distance and data correlation & distribution

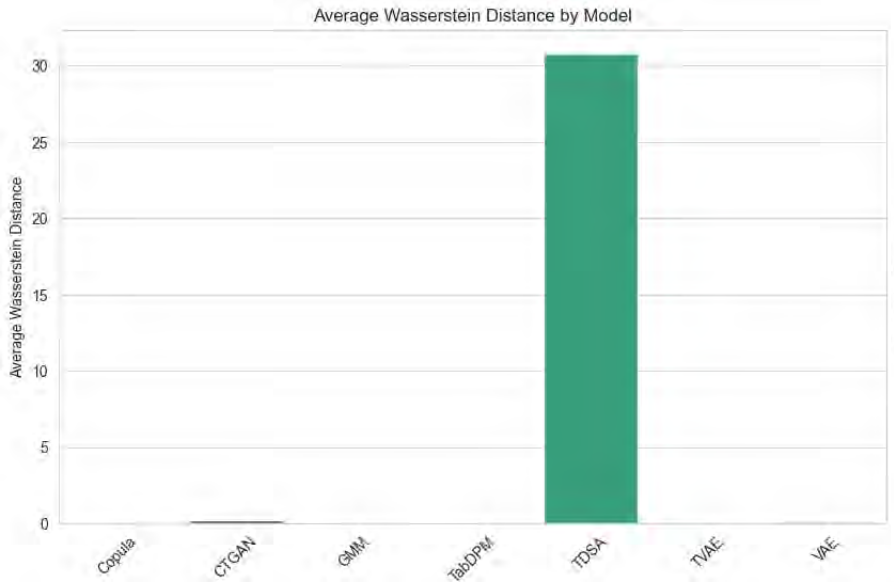


Figure 3 : Average Wasserstein distance for all columns for synthetic data relative to original data except for TDSA

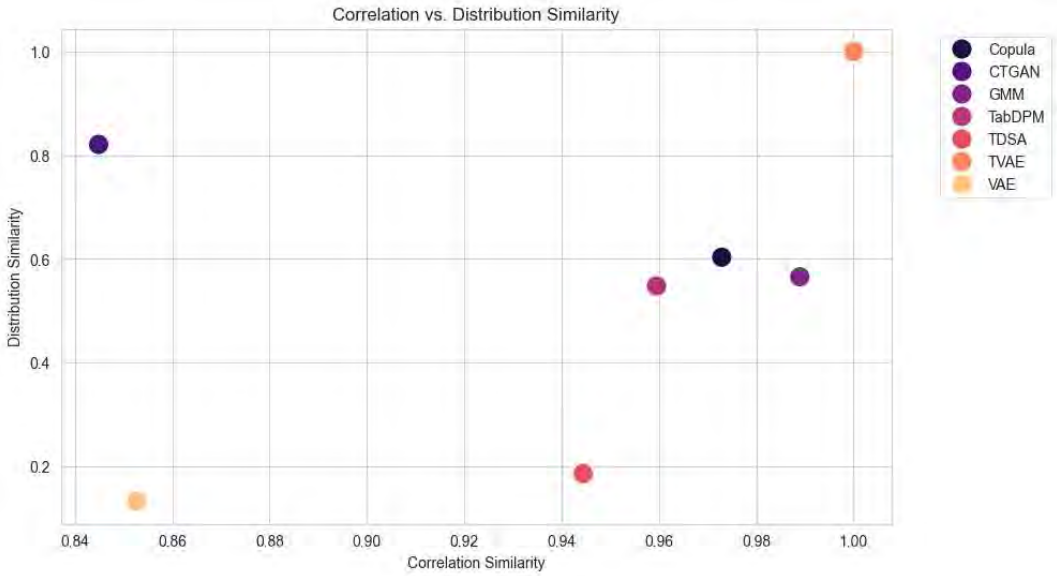


Figure 4 : TVAE generated synthetic data has high correlation and distribution similarity

Utility : Classification and Clustering and Time Series Forecast



Figure 5 : Overall weighted scores for synthetic data produced by SDGs are similar except for TDSA

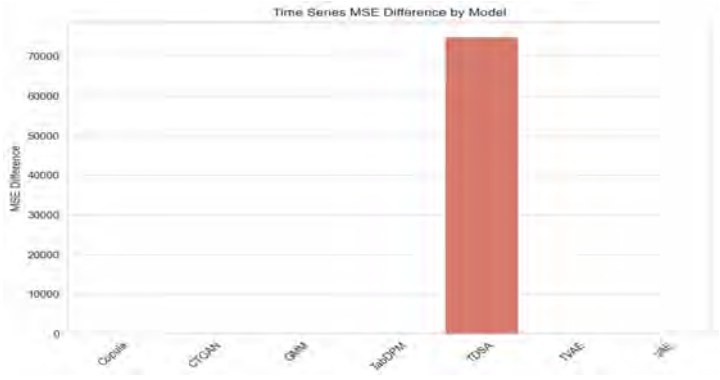


Figure 6: Time series forecast mean square error for synthetic data produced by various SDG with TSDA as an outlier

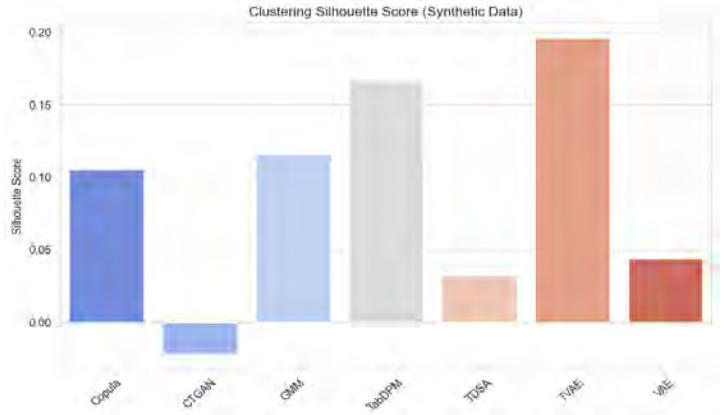


Figure 7: Clustering scores for synthetic data produced by various SDGs are generally poor

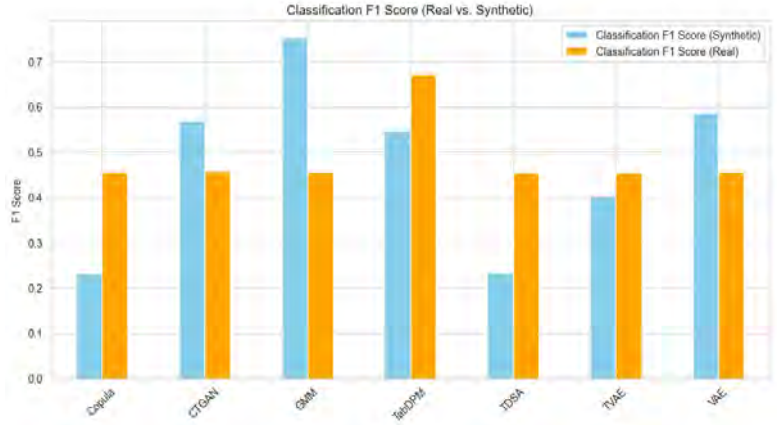


Figure 8: F1 scores for synthetic data produced by SDGs

Privacy : T-measure scores for various synthetic data produced by various SDG



T-measure is the absolute difference between the synthetic and original column mean for all numeric columns, normalised by the standard deviation. This is then transformed into the privacy score

Figure 9: Privacy scores for datasets produced by various SDG with TDSA as an outlier

Privacy vs Utility trade-off : Choosing the SDG that fulfils both

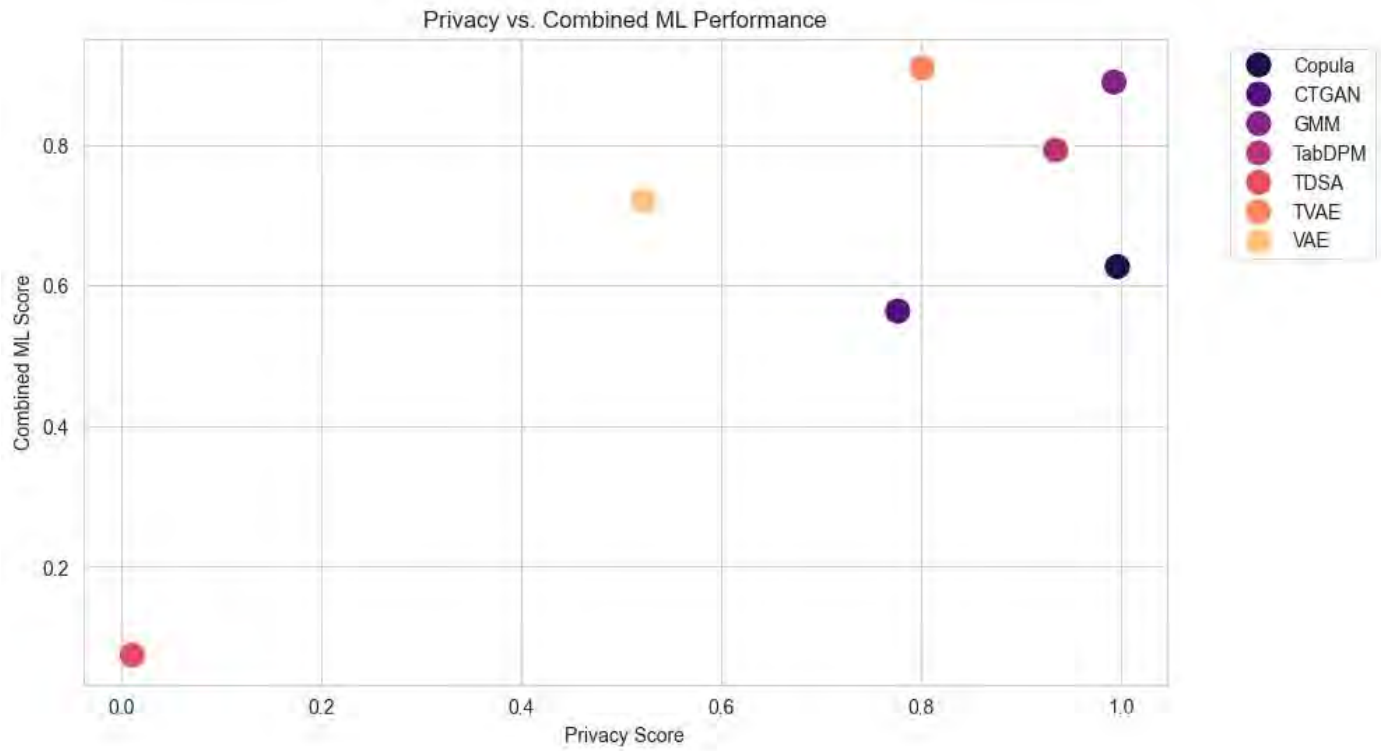


Figure 10: Privacy vs Utility (Combined ML) Score plot

Discussion : Weighted score aligns with privacy, utility and fidelity measures

Model	Weighted Score	Privacy	Utility	Fidelity	Runtime(s)	Evaluation Summary
GMM	0.89	High	High	Moderate	34.5	Best overall performer. High privacy and utility and moderate fidelity. Highest weighted score.
TVAE	0.86	High	High	High	1113.2	High privacy, high utility, and high fidelity. Weighted score reflects this balance.
TabDPM	0.83	High	Moderate	Moderate	365	High privacy with moderate utility and fidelity. Weighted score reflects this balance.
Copula	0.82	High	Moderate	Moderate	84448.3	High privacy, but moderate utility and fidelity. Weighted score reflects this balance.
CTGAN	0.77	Moderate	Low	High	139.3	Moderate privacy, low utility but high fidelity. Weighted score aligns with this.
VAE	0.65	Low	Moderate	Low	97.5	Low privacy, moderate utility and low fidelity. Weighted score aligns with this.
TDSA	-7.19	Low	Low	Low	2367.6	Poor in all dimensions. Negative weighted score reflects its poor performance.

Table 2 : Weighted scores and PUF scores alignment

Conclusion

Summary

- The proposed framework is generalisable to support different measures for different analytics tasks, answering RQ 1 & 2.
- Preprocessing of data is however difficult to generalise as it is very task-dependent.
- Future expansion of the framework may include support for non-structured data, federated machine learning and addition of differential privacy mechanism

Recommendations

1.Privacy-Critical Applications:

- ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Validate to avoid overfitting
- ❑ **Copula**: Use for analysis in sensitive domains; at a cost of long processing times

2.Utility-Focused Applications:

- ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Fast processing times
- ❑ **TVAE**: Strong for utility and privacy in sensitive domains

3.Avoid:

- **TDSA**: Poor in all metrics; not recommended for imbalanced datasets and datasets with non-sequential data

References

[1] Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M., & Narang, P. (2024). TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data (arXiv:2207.05295). arXiv. https://doi.org/10.48550/arXiv.2207.05295

[2] European Union. (2024). Article 10: Data and Data Governance | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/10/

[3]Jordon, J., Houssiau, F., Cherubin, G., Cohen, S. N., Szpruch, L., & Bottarelli, M. (2022). *Synthetic Data—What, why and how?* Alan Turing Institution.

[4] Bender, S., & Staab, P. (2015). *The Bundesbank's Research Data and Service Centre (RDSC)—Gateway to treasures of micro data on the German Financial System*. IFC workshop on "Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges".

[5] Caceres, H. E., & Moews, B. (2024). *Evaluating utility in synthetic banking microdata applications* (arXiv:2410.22519). arXiv. https://doi.org/10.48550/arXiv.2410.22519

[6] Lapteva, E. K. (2023). *Utility and confidentiality assessment of synthetic company data*. IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks".

[7] Ong, C. A., Escalanda-Lo, C., Gabriel, M., & Reyes, R. (2024). Research for All: Exploring machine learning applications in generating synthetic datasets. *2024 Philippines Statistical Association Annual Conference*. https://events.psai.ph/2024/ac/

Addendum : Reflection on generative AI (GenAI) tools for enabling research (as of 31.12.2024)

	Research	Coding	Writing	Analysis
Helps	Generate new ideas	Speed up coding	Conciseness and tone	Quick initial comparisons
Caveats	Hallucination and recency	LLMs are throttled LLMs needs management	Loss of personality	Surface level analysis
Future hopes	Integrated into citation management	More efficient models to enable self-service coding	Increased support for academic writing	Increased evidence of reasoning

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

From CPU to GPU: comparison of parallel computing techniques for numerical solutions[1]

Maxim Ralchenko, Oleksandr Shcherbakov and Kim P Huynh,
Bank of Canada

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

From CPU to GPU[1]

Comparison of parallel computing techniques for numerical solutions

Maxim Ralchenko, Oleksandr Shcherbakov, and Kim P. Huynh[2]

Abstract

This paper provides comparison of alternative parallelization techniques using central processing units (CPU) and graphics processing units (GPU) in application to a dynamic empirical model of consumer optimal cash consumption, holding and withdrawals. We document important optimization steps within programming code and compare OpenMPI versus CUDA implementations in terms of computation time, accuracy, and electricity consumption. To compare the practical efficacy we analyze what would it cost to implement these algorithms in a publicly-available cloud computing facilities.

Keywords: Demand for money, inventory models, heterogeneity, simulated maximum likelihood.

JEL classification: C15, C88, CE41, E42, L86.

[1] This paper is prepared for the BIS Irving Fisher Committee on Central Bank Statistics (IFC) 4th workshop on "Data Science in Central Banking," co-hosted with the Bank of Italy (BoI). We thank Giuseppe Bruno and Juri Marcucci for encouraging us to write this paper. We thank Nathalie Swift and the Analytical Environment Business Support (AEBS) team for their expert technical assistance on this project. This project is a collaboration between AEBS and Currency Economic Research and Analysis to modernize our structural econometric models that are useful for policy analysis. The content is not related to the economic outlook or to the direction of monetary policy. We acknowledge the use of the Bank of Canada's High-Performance Cluster EDITH3. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the Bank's Governing Council.

[2] All authors are with the Bank of Canada. Contact email addresses are: mralchenko@bankofcanada.ca, ashcherbakov@bankofcanada.ca, and khuynh@bankofcanada.ca

1. Introduction

In economics, many interesting phenomena are described by complicated mathematical problems that are solved by many sophisticated decision-makers. For example, strategic consumer decisions of how much cash to consume and carry and how much and how often to withdraw can be formulated as a consumer dynamic programming problem with continuous dynamic controls. By matching model predictions to observed data we can estimate preference parameters which are useful in understanding determinants of behaviour of economic agents and are needed to accurately predict changes in the behaviour due to regulation or an exogenous shock.

When using parametric estimation techniques, to recover structural parameters of these models one needs to be able to solve them many times. When consumer heterogeneity becomes important, multiple solutions may be needed for multiple parameter values. For example, in simulated extremum estimators (simulated maximum likelihood or simulated method of moments) at least one solution is usually required per realization of a vector of random coefficients. As the dimensionality of random preferences becomes larger these models quickly become computationally infeasible.

Recent developments in empirical industrial organization, however, suggest faster ways of estimating rich preference heterogeneity by first presolving the underlying mathematical problem for a large number of candidate parameter values and then estimating a distribution of probability weights over agent types by matching model prediction to the observed data. These modern non-parametric methods to estimate preference heterogeneity were originally proposed by Ackerberg (2009)—more recently used by Malone et al. (2021), which is closest to our implementation, and McManus et al. (2022)—to our setting. Alternative approaches using semi-parametric mixtures estimators are discussed in Fox et al. (2011; 2016), Nevo et al. (2016). In this paper, we solve the problem more than 1.3 million times – one for each distinct vector of candidate parameter values. Consumer heterogeneity with respect to structural parameter values is combined with their heterogeneity along first the two moments of consumer income distribution and the variation in consumer cash withdrawal travel distance.

In this paper, we demonstrate how to implement these algorithms with general processing units (GPU) rather than a central processing unit (CPU). The idea of using GPUs rather than CPUs for dynamic programming for economic problem was discussed in Aldrich et al. (2011) and Aldrich (2014).

The rest of the paper is organized as follows. Section 2 outlines the economic problem and the solution algorithm. Section 3 describes various stages and layers of computer code optimization. Section 4 discusses performance of different hardware and software configurations employed in parallelization of the workload. Section 5 discusses the results.

2. Economic problem description

We study Canadian consumer cash management behaviour. Cash is valuable to consumers mainly due to its functionality as a medium of exchange and store of value. Our focus is on the use of cash as a medium of exchange, that is, consumer usage of cash as a payment instrument. Ability to use cash appears necessary for the store of value motive to realize. This next section is based and draws heavily from Huynh et al. (2025).

We assume that consumers have constant preference for cash and non-cash transactions at the Point of Sale (POS), can withdraw cash at an ATM or a bank branch, and can hold cash inventories. Cash withdrawal is costly. The cost is increasing in distance to the nearest cash access point and is independent of the withdrawal amount. Holding cash is also costly. The cost may reflect foregone interest, chances that cash is lost or stolen, need for safe storage for larger holdings, etc. Every period, consumers begin with a weakly positive level of cash holdings from the previous period and receive random draws defining their total current expenditures. We assume these draws are represented by independent identically distributed random draws from a normal distribution with mean and variance parameters to be estimated. After observing realizations of expenditure in the current period, consumers decide whether to withdraw cash and, if yes, how much, and how much cash to consume (i.e., spend in the current period). Let i denote a consumer type and let h_{it-1} be the level of cash holdings at the beginning of t. Let $s_{it} \overset{iid}{\sim} N(s_i, \sigma_i^2)$ denote current level of expenditures, and w_{it} and c_{it} denote levels of cash

withdrawal and cash consumption, respectively. We assume that consumer per-period reward function is defined as follows:

$$\begin{aligned} U_i(h_{it-1}, s_{it}, d_i; w_{it}, c_{it}) = & \alpha_i \ln(1 + c_{it}) + (1 - \alpha_i) \ln(1 + s_{it} - c_{it}) \\ & - \gamma_i (h_{it-1} + w_{it} - c_{it}) \\ & - F_i d_i \times \mathbb{1}_{w_{it}>0} \end{aligned} \tag{1}$$

where α_i is a preference for cash transactions, γ_i is the cost of holding cash, d_i is a measure of distance to the nearest cash access point, and F_i is the fixed cost of withdrawal.

We assume that consumers maximize discounted value of utility flows over some time horizon $T \leq \infty$ by choosing the level of cash withdrawal $w_{it} \geq 0$ and cash consumption $c_{it} \geq 0$ in every period, i.e.,

$$\max_{(c_{i0}, w_{i0}), \ldots, (c_{iT}, w_{iT})} \sum_{t=0}^{T} \beta_i^t \mathbb{E} U_i(h_{it-1}, s_{it}, d_i; w_{it}, c_{it}) \tag{2}$$

$$\text{subject to } 0 \leq w_{it}, \ 0 \leq c_{it} \leq \min\{h_{it-1} + w_{it} - c_{it}\} \tag{3}$$

where prior to making their choices consumers observe current period realization of total consumption s_{it}, and expectation is taken with respect to its future realizations at $t + 1$ onward, and the laws of motion for state variables are defined as

$$h_{it} = h_{it-1} - c_{it} + w_{it}, \tag{4}$$

$$s_{it} = s_i + \epsilon_{it}\sigma_i, \ \epsilon_{it} \overset{iid}{\sim} N(0, 1). \tag{5}$$

Our goal is to solve problem (2) subject to constraints (3) for a given T. The solution is expressed as optimal cash withdrawal and usage policy for a consumer

type i characterized by a vector of parameters ($\alpha_i, \gamma_i, F_i, \beta_i$), observable distance state d_i, average spending level s_i and a known variance of income innovations, σ_i^2.

1. **Representative consumer model.** If we assume identical preferences for all consumer types i such that any consumer behaviour can be represented by the same vector of parameter values, estimation of the model parameters requires several hundreds of solutions (depending on how starting values are chosen). While there are material benefits from parallel computing even for the representative consumer specification, the real gain comes when we allow for semi-parametric or non-parametric consumer heterogeneity which is discussed next.

2. **Heterogeneous consumers.** While a representative consumer model is a useful starting point in our analysis, we find that observed consumer behaviour is likely to reflect significant variation in consumer preferences (and constraints) in population. A natural first step is to define consumer types using exogenous observable demographic characteristics such as age and income. With just two categorical values in each of the variables (low/high, young/old) we need to solve four problems of at least the same difficulty level as the representative consumer model discussed above. The `at least' part comes from the fact that non-linear search over much larger parameter vector requires significantly more function evaluations where each evaluation involves four solutions.

Conditioning on the observed demographics helps to improve model flexibility but it often cannot fully account for all the differences in consumer preferences particularly those determined by the unobserved consumer characteristics. Therefore, a richer non-parametric estimation techniques may be needed to accurately account for individual consumer decisions. This is particularly relevant when individual consumer-level data is available which is our case. In Huynh et al. (2025), we show that consumer observed behaviour suggests bimodality in consumer response to an increase in withdrawal costs. We find that some consumers can significantly reduce cash use, withdrawal level and the frequency of withdrawals, while others reduce cash consumption only sightly and increase withdrawal level while reducing the frequency. We find that while some of the behavioural differences are correlated with observable demographics, the non-parametric estimates reveal much reacher distribution of individual consumer responses than its semi-parametric analog does. Therefore, in what follows, we describe optimizations in the context of the non-parametric estimation method while results for a representative consumer specification and the one based on observable demographics are available in Huynh et al. (2025).

2.1 Solution algorithm

Our parametric assumptions do not allow for a closed-form solution of the dynamic programming problem (2). Obtaining a numeric solution for two dynamic controls is computationally challenging and we begin by re-formulating the problem in terms of a single control that uniquely identifies a withdrawal-consumption pair.

First thing to notice is that the effect of control variables (w_{it}, c_{it}) on the continuation value occurs via the change in cash holdings, i.e., $h_{it} = h_{it-1} + \Delta h_{it}$, where $\Delta h_{it} = w_{it} - c_{it}$. In other words, all withdrawal-consumption pairs such that $w_{it} - c_{it} = \Delta h_{it}$ have the same dynamic implications (result in the same continuation value).

Therefore, we can re-cast the dynamic programming problem (2) (with constraints (3) in terms of a single dynamic control Δh_{it} as follows,

$$\max_{\Delta h_{i0},\ldots,\Delta h_{iT}} \sum_{t=0}^{T} \beta^t \mathbb{E}\left[\bar{U}_i\left(h_{it-1}, s_{it}, d_i; w_i^*(\Delta h_{it}), c_i^*(\Delta h_{it})\right)\right] \tag{6}$$

$$\text{subject to } \Delta h_{it} \geq \max\left\{-s_{it}, -h_{it-1}\right\}, \tag{7}$$

where $w_i^*(\Delta h_{it})$ and $c^*{}_i(\Delta h_{it})$ is a unique withdrawal-consumption pair that solves

$$\max_{(w,c)} \left\{\alpha \ln(1+c) + (1-\alpha)\ln(1+s_{it}-c) - F_i d_i \times \mathbb{1}_{w_{it}>0}\right\} \tag{8}$$

$$\text{subject to } w \geq 0,\ s_{it} \geq c \geq 0,\ w - c = \Delta h. \tag{9}$$

Lemma 1 in Huynh et al. (2025) establishes the uniqueness. Re-formulating the problem in terms of a change in cash inventories Δh allows reduces expensive numeric optimization to only one dimension.

To solve the maximization problem (6) numerically we define 30 grid points for each of the four state variables $S \equiv (h_{it-1}, s_i, \epsilon_{it}, d_i)$. We set $T = 183$ such that by making daily choices consumers solve optimal cash inventory problem with a half a year planning ahead horizon. The problem is solved using backwards induction. In the terminal period, consumers solve a simple static problem where exact solution is given by

$$(w_{iT}^*, c_{iT}^*) = \arg\max_{(w,c)} \left\{\alpha \ln(1+c) + (1-\alpha)\ln(1+s_{it}-c) - F_i d_i \times \mathbb{1}_{w>0}\right\} \tag{10}$$

$$\text{subject to constraints (9)}.$$

As the solution is analytic, no non-linear optimization is involved at this stage. We assume that consumers do not incur any holding costs in the terminal period. Given the optimal solution we compute terminal period value function $v_{iT}(S)$ defined on 304 grid points in the state space. To compute value function at t = T − 1, for each point in the state space S we numerically search for optimal dynamic control Δh_{it}^*

$$\arg\max_{\Delta h} \left\{\bar{U}_i\left(h_{it-1}, s_{it}, d_i; w_i^*(\Delta h), c_i^*(\Delta h)\right) + \beta \frac{\sum_{n=1}^{30} H\left(V_{t+1}(h_{it}, s_i, x_n, d_i)\right) f(x_n)}{\sum_{k=1}^{30} f(x_k)}\right\}, \tag{11}$$

where $H(\cdot)$ is a linear interpolation function approximating continuation value along the next period cash holding dimension h_{it} and the expectation over future realizations of income shocks is approximated by 30 quadrature draws x_n from a normal distribution with variance parameter, $\sigma_i^2(s)$, being a known (observed) function of average income level, and $f(x_n)$ are the weights given by the normal *pdf*.

2.1 Matching model predictions to data

Our individual-level data contains the following moments of cash withdrawal-consumption distribution on a daily basis: average cash consumption, average cash holdings, average cash withdrawal level conditional on withdrawing, average daily frequency of withdrawals, and the level of cash holdings that triggers a withdrawal.

Each consumer (an observation in the data) is characterized by a vector of state variable values, including average income level, variance of innovations to income in every period, distance to the nearest cash access point. In the data, consumers report their average cash usage, holding and withdrawal levels, withdrawal frequency as well as cash holdings at withdrawal.

For a given vector of parameter values, a solution to the consumer dynamic programming problem is represented by a dynamically optimal change in consumer cash holdings amount, $\Delta h_{it} = \Delta h(h_{it-1}, s_i, \epsilon_{it}, d_i)$ which, in turn, uniquely identifies a withdrawal-consumption pair in a given period, $(w_i^*(\Delta h_{it}), c^*_i(\Delta h_{it}))$. We are interested in the parameter values such that observed consumer behaviour is closely related to the model predictions.

We define "closedeness" of observed and predicted consumer behaviour in terms of simulated method of moments criterion function. More formally, on a daily level, let $\mathrm{E}[w_{it}|w_{it} > 0] \geq 0$ denote expected cash withdrawal level conditional on withdrawing, $0 \leq \Pr(w_{it} > 0) \leq 1$ denote expected daily frequency of cash withdrawals, $\mathrm{E}[h_{it}] \geq 0$ denote expected holding level, and $\mathrm{E}[c_{it}] \geq 0$ denote expected cash consumption. In the data, about half of observations for cash holding at withdrawal are missing. Therefore, we do not use this moment in estimation. The expectation is taken by averaging forward simulation outcomes over time. We define four structural error terms,

$$\varepsilon_{1i} \equiv \mathbb{E}[w_{it}|w_{it} > 0] - \mathbb{E}[\widehat{w_{it}|w_{it} > 0}], \quad (12)$$

$$\varepsilon_{2i} \equiv \Pr(w_{it} > 0) - \widehat{\Pr(w_{it} > 0)}, \quad (13)$$

$$\varepsilon_{3i} \equiv \mathbb{E}[h_{it}] - \widehat{\mathbb{E}[h_{it}]}, \quad (14)$$

$$\varepsilon_{4i} \equiv \mathbb{E}[c_{it}] - \widehat{\mathbb{E}[c_{it}]}, \quad (15)$$

where terms without hats represent data sample moments and terms with hat denote model predictions. Let Z_{it} denote a vector of instrumental variables including a constant term, consumer income level and the distance to the nearest cash access point. We believe these variables are plausibly exogenous in our context. The method of moments parameters then become:

$$\theta = \arg\min_{\theta'} \{g_n(\theta')' \cdot W \cdot g_n(\theta')\} \quad (16)$$

where $g_n(\theta)$ is 12-by-1 sample moment conditions obtained by interacting each of the four structural errors defined in equations (12) through (15) with three instrumental variables and W is the weighting matrix. For the weighting matrix we use the optimal weighting matrix from the second stage of a representative consumer model. The representative consumer model is estimated using $(Z'Z)^{-1}$ (inner product of instrumental variable matrix) in the first and $W = (g_i(\theta^1)'g_i(\theta^1))^{-1}$ in the second stage (using first stage estimates θ^1).

Our version of non-parametric estimates of the distribution of consumer types is closest to McManus et al. (2022). In particular, having solutions for a large number of consumer types we estimate weights on each type such that the closest prediction obtains the weight of 1 and all other candidate types obtain zero weights. Matching analog would be $K-$nearest neighbour predictor with $K = 1$.

To implement this method we first need to obtain solutions for a large number of candidate types. We begin by solving the model for 1,300,000 consumer types, where each consumer type is represented by a vector of Halton numbers, $(\alpha_i, \gamma_i, F_i, \beta_i)$, $i = 1,\ldots,1300000$. Halton numbers cover parameter space more evenly than pseudo-random numbers. As the original solution time using serial implementation was too slow (see Table 2), we considered parallelization solutions using OpenMPI with CPU and CUDA with GPU which are discussed next.

3. CPU and GPU computing

This section describes how we optimize the existing CPU code and then how we port this optimized code to GPU.

3.1 Parallelization of existing CPU code

The existing code can be parallelized through two main approaches:

1. Each combination of state variables (30^4 = 810,000 grid points), at each time step of the dynamic programming problem, see (11).
2. For multiple sets of observations, different sets of coefficients α_i, γ_i, F_i, and β_i to incorporate individual heterogeneity into the model.

The original CPU code was optimized with OpenMPI. Using the first optimization approach, there was no strong improvement in performance beyond 100 cores. The second approach does not require so much communication overhead and can be run on a greater number of cores, as the first approach requires synchronization of all grid points when moving to the next time step.

Based on findings from the GPU optimization exercise, we propagated a handful of optimizations back to th CPU code. These optimizations were found by identifying calculations that can be done at a certain point in advance and stored in memory, instead of repeating these calculations in an iterative process. The biggest of such improvements included:

1. Precalculating $\beta / \sum_{k=1}^{30} f(x_k)$ just once at the beginning as this term doesn't change throughout candidate solution values.
2. Precalculating the coefficients in the reward function, as most are consistent for all golden search iterations at a given time step
3. Precalculating the dot product $H(V(x)) \cdot f(x)$, as $f(x)$ is a constant, and $V(x)$ is calculated in the previous time step

As the code is numerically sensitive, precalculating and storing some coefficients yields very subtle differences in the final result, although different numerical techniques for the summation portion of the $V \cdot E_w$ dot product will give a great difference than the precalculation optimizations.

3.2 GPU porting

The existing C-OpenMPI codebase was converted to CUDA, which is an Nvidia proprietary API that permits the programmer to use the GPU for general-purpose programming on Nvidia devices. At a basic level, it permits the programmer to allocate memory on a device (GPU), copy memory from the host (CPU) to the device, and back, and invoke GPU kernels (snippets of code that execute on the device). For GPUs from AMD, who is the other major vendor, heterogeneous-computing Interface for Portability (HIP) is used. The HIP API is a C++ runtime and is a CUDA-like API. For both kinds of GPUs, OpenCL is another alternative for general-purpose GPU programming.

To write GPU kernels from scratch, usually a knowledge of C or C++ is required (alternatively, Fortran may be used, although this approach is much less common). The main learning curve to go onto the GPU for a programmer who is already experienced in those languages is implementing shared-memory parallelism. In this approach, multiple threads can access the same pool of shared memory. The difficulty with shared memory parallelism depends on the nature of the algorithm. For example, if some element of memory must be accessed many times, race conditions may easily occur, where two threads try to process some element of memory simultaneously. Alternatively, some algorithms, such as those classified as "embarrassingly parallel", are much simpler to implement, owing to the diminished risk of such conflicts.

The difference between shared memory on CPU versus a GPU computing is chiefly the nuances of GPU hardware. In a broad sense, GPUs were initially designed to calculate pixel values on computer monitors. These processing units are optimized at the hardware level to more efficiently process the arithmetic operations required to display graphics. Existing codes that handle embarrassingly parallel workloads and can operate in a SIMD (single-instruction, multiple-data) paradigm are particularly good candidates for optimization on a GPU; it is instructive to note that the "Hello World" example on a GPU is vector summation. Particularly useful resources for GPU computing on Nvidia include the CUDA C++ Programming Guide and the CUDA C++ Best Practices Guide. The latter publication is especially useful to optimize initial GPU code to account for points such as memory optimization (GPUs have many classes of memory that vary in size and read/write speed).

We exploit the fact that we can perform in parallel the calculations for each combination of state variables, at each time step. While the time-stepping procedure is by necessity serial, the calculations at each time step are done in parallel.

We initially used Nvidia A40 GPUs to run the converted code. The choice of GPU was based on existing availability, and is not the most ideal choice for this code, as it requires strong double precision performance. We subsequently acquired a Nvidia H100 GPU to further improve performance. Porting the code between the two GPUs is straightforward and can be done without changes; nevertheless, there are often optimizations that can be done to take advantage of the peculiarities of a specific GPU. In our case, a forward simulation, which kept the A40 fully busy, did not do so on the H100. An optimization was performed to execute multiple forward simulations simultaneously.

For this specific code, a GPU can be effectively used, because solving for each combination of state variables is an example of an embarrassingly parallel workload, as each combination at each time step is independent. Note that not every code is a good candidate for GPU optimization; examples of poorer candidates would be problems with fewer parallel tasks, or situations where each parallel tasks will tend to execute different processor instructions (termed divergence). Large-memory problems are not impossible to solve on a GPU, but there is more work that must be done to split the problem among multiple GPUs.

4. Performance and cost benchmarks

This section provides a speed comparison between CPU and GPU. In addition, we use the technical performance to provide a case study of how to potentially benchmark the cost of running the code on a publicly available cloud-based GPU.

4.1 Speed comparisons

Table 1 compares the execution time and speedups between a single CPU, an A40 GPU, and a H100 GPU. All of this hardware is available on the Bank of Canada Edith-3 High-Performance Cluster.

Code performance between CPU and GPU hardware — Table 1

	CPU	A40 GPU	H100 GPU
Execution time [minutes]	11.25	0.14	0.0054
Speed-up from serial	.	79	2077
Speed-up from H100	.	.	26

Notes: code performance is reported for a single observation in the data; CPU core is AMD EPYC 7543 32-core processor.

The speed-up factors can be thought of in terms of how many CPU cores are needed to replicate the GPU performance; in other words, 79 CPU cores would be needed to replicate the A40, but 2077 cores for a H100. This approach assumes little to no overhead between CPU cores for a distributed memory setup (i.e. MPI), which is a reasonable assumption for each CPU core handling a given combination of coefficients α_i, γ_i, F_i, and β_i.

The discrepancy between the A40 and H100 is due to the much improved double precision (fp64) performance on the H100; there is a theoretical ~50x performance difference for double precision between the two GPUs and a ~5x memory bandwidth improvement. The speedup factor of ~26x suggests that the code was mostly compute-bound as opposed to memorybound on the A40 (compute-bound codes spend more time doing calculations; memory-bound codes spend more time read and writing from GPU memory).

There are some power savings to be had with a GPU. Each compute node on our in-house cluster has 2 CPUs, and 64 cores in total. When all cores are running on a node, testing suggests that about 450W are consumed. To match the H100 performance, roughly 32 nodes are needed, which scales to 14.4 kW power usage. The H100 has a thermal design power (TDP) of at most 400W; if we assume that all cores of the node are operating (although, for the GPU code we only use 1), then we have at most 850W for the GPU code.

There is also a set-up cost question to consider. In a typical high-performance computing cluster environment, the cost to acquire ~2000 CPU cores is likely to be much greater than the cost of a data-centre HPC-optimized GPU. With consumer hardware, it is possible, as of 2025, to purchase a high-end (enthusiast-class) GPU

that will have as much as 2.5x performance at double precision than the A40. It is therefore much more practical to purchase such a GPU, as opposed to a ~200-core CPU for this job.

4.2 Case study based on a normalization

We provide a case study to understand trade-offs. We focus on completion time or the actual time for a job to finish. To normalize all performance relative to the H100 GPU card which takes about 10 hours to finish one job. Based on the job time for CPU will ask how many cores would be required in a MPI job to match the 10 hours. We do similarly with a cloud-based A100 card. We choose the A100 as it is relatively affordable and widely available with many cloud-based services. It will provide a way for the reader to conduct sandbox or experiments without committing to large infrastructure investments.

Table 2 provides a comparison of the hardware required for CPU on premise (Edith3) and GPU cloud (A100) solution to match the wait time for the GPU on premise (H100). In addition, we estimate the cost of running the job on the GPU cloud based on GPU Azure.

Requirements to finish one job in 10 hours — Table 2

	CPU	GPU	GPUC
Hardware	AMD	H100	A100
Number of cores	2000	16896	55296
Clock speed	2.8 GHz	1.98 GHz	1.41 GHz
Memory	1 TB	94 GB	80 GB
Cloud cost per job	NA	NA	$200

Note: The GPU is on premise is the benchmark case as the job finishes in 10 actual hours. The CPU is on premise and GPUC is cloud hardware requirements are adjusted to ensure completion of the job in 10 hours. Info on GPU cloud is available here: azure.microsoft.com/en-us/pricing/details/machine-learning/

In our case the model lends itself to an embarrassingly parallel structure. In addition, the memory requirements are not so onerous since the dataset used is relatively small (about 10 megabytes). Aldrich (2014) provides a detailed discussion between using fewer CPU cores with faster clock speed versus much more GPU cores with slower clock speeds. From the Table 2 - CPU requires only 2000 cores at 2.8 Ghz relative to 16896 cores for H100 at 1.98 Ghz and 55296 cores for A100 at 1.41. Ghz.

A technical consideration is the memory requirements for CPU versus GPU. Aldrich (2014) makes a distinction between CPU memory versus GPU memory. CPU uses host memory while GPU uses device memory. Therefore, GPU instructions can only use data that are located in device memory. Therefore, it is important to have a fast interface between host and device.

The final calculation that we provide is the cost of running the job on GPU cloud. Based on our calculation it is about $200 per job. Hopefully, this cost calculation can help the reader to think about the tradeoffs in investing on on-premise solution

versus GPU cloud solution. Due to the various complexities on-premise and total cost of ownership - we are unable to calculate the cost of the on premise solutions.

5. Discussion

As many interesting economic problems require computationally challenging solutions the role of parallelization techniques increases tremendously. Computational tasks that have been viewed as unfeasible using standard serial calculations, now become quick and efficient when solved using multiple processors at the same time.

We provide a concrete example based on policy-relevant research at the Bank of Canada. An important caveat is that we assume that the researcher is skilled in writing code in C and has the support of a dedicated high-performance computing specialist. We provide a brief guide on how to port the code from CPU to GPU.

We provide some calculations to help understand the tradeoffs inherent in implementing GPU versus CPU. We illustrated how calculation speed can be improved using GPU instead of CPU architecture as well as how these gains are increased with modern advancements in hardware development. We provide an example of case where GPU shines - embarrassingly parallel jobs which do not exceed the device memory requirements. Finally, we provide a back of the envelope calculation of how much it would cost to implement our example on publicly available cloud service.

6. References

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. QME, 7:343–376.

Aldrich, E. M. (2014). Chapter 10 - GPU computing in economics. In Schmedders, K. and

Judd, K. L., editors, Handbook of Computational Economics Vol. 3, volume 3 of Handbook of Computational Economics, pages 557–598. Elsevier.

Aldrich, E. M., Fern´andez-Villaverde, J., Ronald Gallant, A., and Rubio-Ram´ırez, J. F. (2011). Tapping the supercomputer under your desk: Solving dynamic equilibrium models with graphics processors. Journal of Economic Dynamics and Control, 35(3):386–393.

Fox, J. T., il Kim, K., and Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. Journal of Econometrics, 195(2):236–254.

Fox, J. T., Kim, K. I., Ryan, S. P., and Bajari, P. (2011). A simple estimator for the distribution of random coefficients. Quantitative Economics, 2(3):381–418.

Huynh, K. P., Shcherbakov, O., and Stenzel, A. (2025). Dynamic consumer cash inventory model. Technical report, work in progress.

Malone, J. B., Nevo, A., and Williams, J. W. (2021). The tragedy of the last mile: Economic solutions to congestion in broadband networks. Working Paper.

McManus, B., Nevo, A., Nolan, Z., and Williams, J. W. (2022). The steering incentives of gatekeepers in the telecommunications industry. Technical report.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. Econometrica, 84(2):411–443.

Dynamic Consumer Cash Inventory Model

Kim P. Huynh
Bank of Canada

Alex Shcherbakov
Bank of Canada

André Stenzel
Bank of Canada

4th IFC Workshop on Data Science in Central Banking
Banca d'Italia
18-20 February 2025

The views expressed in this paper are solely those of the authors and do not necessarily represent those of the Bank's Governing Council. The content is not related to the economic outlook or to the direction of monetary policy.

Introduction

- rapid technological developments in the payment industry
  - → new payment instruments become available/are increasingly adopted
  - → consumers balance use of multiple instruments (cash, cards, mobile payments)
- cash usage at the point of sale (POS) has been declining over the past decade
  - → calls to abandon cash (Rogoff 2017), but retains market share (Henry et al. 2024)
  - → cash puzzle: demand for bank notes is steadily increasing (Engert et al. 2019)

    ⇒ continued role for central banks to ensure accessibility of cash for entire populace
- infrastructure to access cash constantly evolving
  - → # of bank branches (predominant cash access point) in Canada peaked around 2013
  - → ongoing branch closure programs (e.g., Desjardins, Laurentian), reduction of up to 30% of branches

    ⟹ Who are the winners & losers from these changes?

Contribution & Findings

- structural model of dynamic cash inventory management
  - → accounts for payment choice at point of sale: cash vs. non-cash
  - → estimation exploits multiple waves (2009, 2013, 2017) of detailed survey & diary data
  - → address role of changing infrastructure
- explicitly account for consumer heterogeneity
  - → representative consumer vs. fully flexible estimation approach: heterogeneity matters
  - → substantial heterogeneity even within demographic groups (old/young/rich/poor/urban/rural/...)
- consumer heterogeneity translates to bimodal response to changes in infrastructure
  - → substantially reduce cash use vs. infrequent but larger withdrawals & holdings
  - → younger and less affluent households forced to substitute away from cash despite preferences

Literature (non-exhaustive)

Baumol-Tobin type inventory models

→ Baumol (1952), Tobin (1956), Smith (1986), Whitesell (1989), Lippi and Secchi (2009) Alvarez and Lippi (2009; 2017), Chen et al. (2021) ...

Technological Innovation & Payment Choice

→ Arango et al. (2015), Koulayev et al. (2016), Chen et al. (2017), Wakamori and Welte (2017), Huynh et al. (2020; 2022), Engert et al. (2024) ...

Nexus of Cash Inventory Management & Payment Choice

→ Alvarez and Lippi (2017), Briglevics and Schuh (2020), Moracci (2022)

Cash Puzzle & Role of Cash

→ Rogoff (2017), Greene and Schuh (2017), Henry et al. (2018), Engert et al. (2019)

Heterogeneity Methods

→ Ackerberg (2009), Nevo et al. (2016), Malone et al. (2021), McManus et al. (2022)

Model: Overview (Parameters of Interest)

- Preferences (α): consumers want to use mix of cash & non-cash for transactions
  - → parameter on cash part of log-linearized Cobb-Douglas utility for payment choice at POS
  - → cash requires cash inventories comprised of cash holdings from previous period & withdrawals
- Costly withdrawals (F): interacted with distance (d) to cash access points
  - → distance measure d: taken from data, proxy for bank branch density in consumer's FSA
  - → cost of withdrawals depends on infrastructure (d) *and* preference/perception (F)
- Holding costs (γ): security, foregone interest, ...

Details

$\Longrightarrow$ trade-off between frequent withdrawals and larger holdings to facilitate cash use at POS

One-Slide Model

4-slide version

- consumers use cash (c_{it}) or non-cash ($s_{it} - c_{it}$) to settle exogenous noisy consumption s_{it}
  - → cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it},w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty}\beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \quad \Longleftarrow \quad \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \,\leq\, \text{cash use} \,\leq\, \text{cash} + \text{withdrawal}}, \quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \,\leq\, \text{expenditure}} \quad , \quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}} \quad , \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - → recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - → analytical characterization of conditional in-period choice
- approximate value function by backward induction using 6-months planning horizon
  - → $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - → yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

One-Slide Model

4-slide version

- consumers use cash (c_{it}) or non-cash ($s_{it} - c_{it}$) to settle exogenous noisy consumption s_{it}
  - → cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it},w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty}\beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \quad \Longleftarrow \quad \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \,\leq\, \text{cash use} \,\leq\, \text{cash} + \text{withdrawal}}, \quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \,\leq\, \text{expenditure}} \quad , \quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}} \quad , \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - → recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - → analytical characterization of conditional in-period choice
- approximate value function by backward induction using 6-months planning horizon
  - → $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - → yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

One-Slide Model

4-slide version

- consumers use cash (c_{it}) or non-cash ($s_{it} - c_{it}$) to settle exogenous noisy consumption s_{it}
  - → cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it},w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty}\beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \quad \Longleftarrow \quad \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \,\leq\, \text{cash use} \,\leq\, \text{cash} + \text{withdrawal}}, \quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \,\leq\, \text{expenditure}} \quad , \quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}} \quad , \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - → recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - → analytical characterization of conditional in-period choice
- approximate value function by backward induction using 6-months planning horizon
  - → $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - → yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

Data: Cash Withdrawal and Usage

- combine data from 2009, 2013 and 2017 Method of Payments Surveys
  - → survey questionnaire: demographics (age, income, location, revolver status)
  - → three-day diary: all transactions made at the POS → aggregate by day

Variable	Mean	Median	Min	Max	s.d.
Withdrawal level	141.29	100.00	0.00	1000.00	134.72
Withdrawal frequency/day	0.10	0.07	0.03	1.00	0.08
Cash holdings	63.35	40.00	0.00	500.00	70.00
Cash holdings at withdrawal	17.90	10.00	0.00	200.00	22.92
Daily total spending	84.04	50.47	0.00	977.00	105.01
Daily cash use	14.74	2.00	0.00	760.00	35.99
Cash/Total spending	0.31	0.04	0.00	1.00	0.41

Notes: Based on 3424 individual consumers across three waves (2009, 2013, 2017).

- exogenous expenditures: categorize consumers according to 30 spending types $(E[s_i], \sigma^2_{s_i})$
  - → spending type reflects total expense, *not* methods of payment used
  - → augment with consumer-specific distance measure (inverse density of branches) Details

Data: Estimation procedure I

Factual Data by Year

Variable of Interest	Overall	2009	2013	2017
Average daily cash use	14.72	15.88	15.56	12.49
Average cash withdrawal	147.80	146.25	143.54	154.99
Probability of withdrawal	0.10	0.11	0.11	0.08
Average cash holding	69.75	66.45	70.24	72.17

- model predictions given parameter vector (α, F, γ)
  - → obtained by averaging over 100×183 day simulations
  - → compare prediction vs. data to obtain error terms

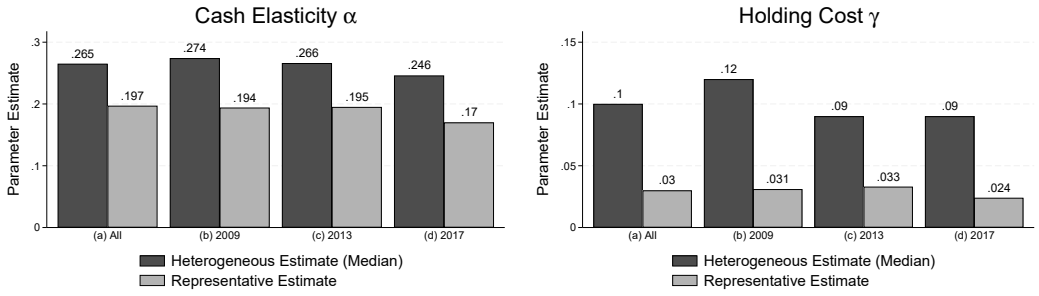
(i)	withdrawal amount given withdrawal:	$\varepsilon_{1,it} = \mathbb{E}\left[w_{it} \middle| w_{it} > 0\right] - \mathbb{E}\left[\widehat{w_{it} \middle| w_{it} > 0}\right]$
(ii)	withdrawal frequency:	$\varepsilon_{2,it} = \mathbb{E}\left[\Pr\left(w_{it} > 0\right)\right] - \mathbb{E}\left[\widehat{\Pr\left(w_{it} > 0\right)}\right]$
(iii)	level of cash holdings:	$\varepsilon_{3,it} = \mathbb{E}\left[h_{it}\right] - \widehat{\mathbb{E}\left[h_{it}\right]}$
(iv)	level of cash use:	$\varepsilon_{4,it} = \mathbb{E}\left[c_{it}\right] - \widehat{\mathbb{E}\left[c_{it}\right]}$

  - → 12 moments via interaction with 3 instruments
    - ∗ constant term, distance measure, spending level

Data: Estimation procedure II

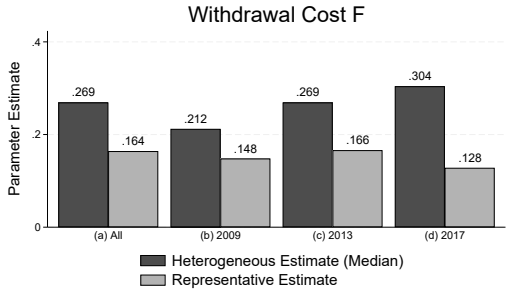
- estimation contrasts representative approach with fully heterogeneous approach
  - → representative: sample/demographic group/year-specific $(\alpha_g, F_g, \gamma_g)$
    - ∗ estimation via GMM using stacked moment conditions
    - ∗ implies optimal 2nd-stage weighting matrix
    - ! consumer spending heterogeneous (30 types), only preferences homogenous (within-group)
  - → heterogeneous: consumer-specific $(\alpha_i, F_i, \gamma_i)$
    - ∗ minimize weighted sum of squared moments (in spirit of Ackerberg 2009, Malone et al. 2021)
    - ∗ candidates: 100'000 parameter tuples obtained via Halton Draws
    - ∗ use weighting matrix from representative approach to account for different moment scales
- report results for both approaches
  - → scale holding cost γ by 100 to facilitate visualization

Takeaway I: Accounting for heterogeneity matters!



- median estimates' trend qualitatively similar to representative approach for α, γ
  - → cash elasticity and inventory cost both decline over time
  - → notable level differences

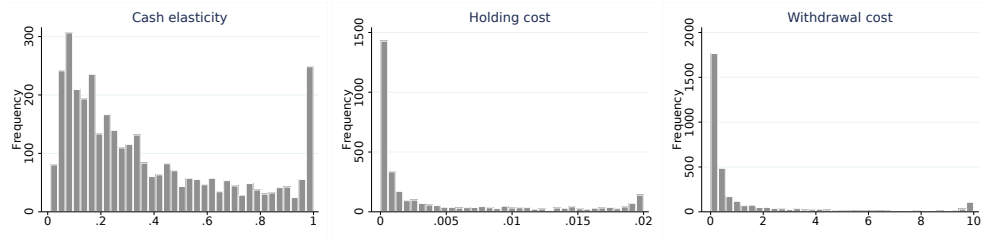
Takeaway I: Accounting for heterogeneity matters!



- qualitatively different trend for withdrawal cost parameter F
  - → upward trend for heterogeneous estimates
  - → inverse U-shaped trend for representative estimates

Takeaway I: Accounting for heterogeneity matters!

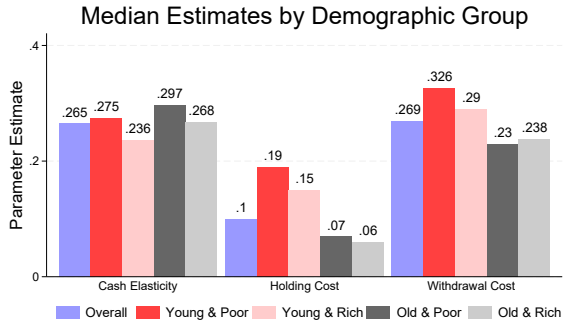
Distribution of α_i (left), γ_i (center), F_i (right) — All years



- heterogeneous approach reveals substantial heterogeneity
  - → cash preference α_i bi-modal but smooth between 0 and 1
  - → holding cost γ_i & withdrawal cost F_i exhibit very long tail

- model fit: heterogeneous approach outperforms representative approach
  - → true even after segmenting into finer groups (demographic group per year)
  - → representative approach unable to match increased cash holdings observed in the data

Takeaway II: Heterogeneity is partially explained by demographics!



- richer and older people use more cash but only age relation driven by preference
- older households have lower costs (both withdrawal and holding/inventory)
  - → cash elasticity increases over time for older & poorer consumers

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - goes hand in hand with increase in cash holdings

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - goes hand in hand with increase in cash holdings

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - goes hand in hand with increase in cash holdings

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

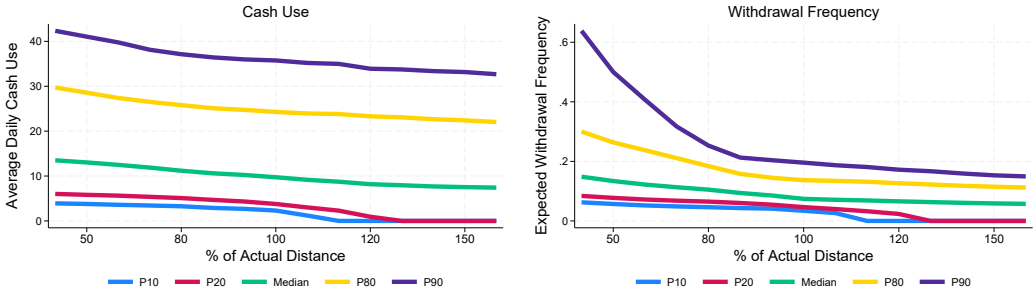
Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - goes hand in hand with increase in cash holdings

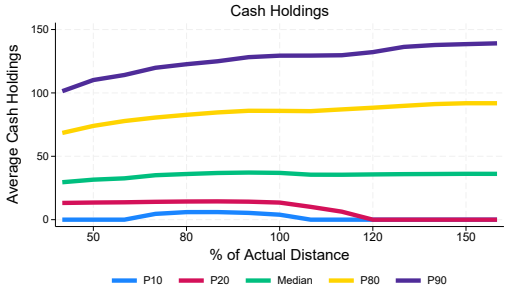
Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!



- Counterfactual: perturb distance to access-to-cash infrastructure for entire sample
  - → withdrawal costs ⇑ $\implies$ cash use ⇓ accompanied by less frequent withdrawals
  - → some consumers stop using cash as it gets too expensive to withdraw;
    $\approx 10\%$ (25%) increase sufficient for 10% (20%) of consumers to stop use of cash at POS

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!



- cash holdings indicative of bimodality in responses
  - → continued cash users (P80+) hold more cash to economize on withdrawals
  - → cash avoiders (P20-) stop using cash as withdrawals become too costly

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

Estimates & demographics by response to 25% increase in distance to access-to-cash infrastructure

Variable	Cash non-users	Cash users		
		all	decreased holdings	increased holdings
	(1)	(2)	(3)	(4)
Cash elasticity α	0.368	0.374	0.328	0.532
Cash holding cost γ	0.011	0.003	0.003	0.003
Withdrawal cost parameter F	2.660	1.152	0.970	1.776
Age	43.259	49.667	49.006	51.927
Income	46179.856	48063.337	47501.312	49982.079
Revolver	0.249	0.197	0.208	0.159
Urban	0.878	0.841	0.823	0.901
Young & Poor	0.342	0.251	0.265	0.203
Young & Rich	0.283	0.202	0.203	0.195
Old & Poor	0.254	0.370	0.362	0.398
Old & Rich	0.121	0.177	0.169	0.204
Observations	696	2465	1907	558

- continued cash-users (cash holdings⇑, withdrawal amt.⇑) are older (& more affluent)
  - → extensive margin: no difference in cash elasticity relative to cash non-users, but lower costs
  - → intensive margin: preference & costs determine response
- younger and poorer consumers phase out cash use as cash becomes too costly for them

Robustness/Extensions/Other Results

- paper: several other analyses, bigger emphasis on distributional impact
  - → branch closures largely unrelated to consumer demographics & preferences
- robustness checks
  - → different planning horizons for consumers, # of points consumer value function is solved at
  - → alternative definitions of consumer spending types and distance measures
  - → structure of the weighting matrix, application to heterogeneous approach
- conceptually feasible extensions
  - → inventory model with deposits (not on equilibrium path)
  - → estimating discounting/risk aversion (in progress)
    - heavily overidentified model (3 parameters, 12 moments)
    - evidence in favor of consumer heterogeneity (Fulford and Schuh 2017)
  - → different impact of distance, free withdrawals, ...

Conclusion

- structural dynamic cash inventory model
  - → trade-off between more frequent withdrawals and larger cash holdings
  - → choice between cash and non-cash
  - → account for infrastructure affecting costs of withdrawal
- estimate model using detailed data on consumer behavior
  - → accounting for individual consumer heterogeneity crucial
  - → bi-modality in consumer responses to increased costs of withdrawing cash
  - → younger & less affluent consumers bear brunt of impact: cash use → 0 due to cost (not preference)

Dynamic Consumer Cash Inventory Model

Kim P. Huynh
Bank of Canada

Alex Shcherbakov
Bank of Canada

André Stenzel
Bank of Canada

4th IFC Workshop on Data Science in Central Banking
Banca d'Italia
18-20 February 2025

The views expressed in this paper are solely those of the authors and do not necessarily represent those of the Bank's Governing Council. The content is not related to the economic outlook or to the direction of monetary policy.

References i

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *QME*, 7:343–376.

Alvarez, F. and Lippi, F. (2009). Financial innovation and the transactions demand for cash. *Econometrica*, 77(2):363–402.

Alvarez, F. and Lippi, F. (2017). Cash burns: An inventory model with a cash-credit choice. *Journal of Monetary Economics*, 90:99–112.

Arango, C., Huynh, K. P., and Sabetti, L. (2015). Consumer payment choice: Merchant card acceptance versus pricing incentives. *Journal of Banking & Finance*, 55:130–141.

Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly journal of economics*, pages 545–556.

Briglevics, T. and Schuh, S. (2020). This is"what's in your wallet"... and here's how you use it. *working paper*.

Chen, H., Felt, M.-H., and Huynh, K. P. (2017). Retail payment innovations and cash usage: accounting for attrition by using refreshment samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2):503–530.

Chen, H., Strathearn, M., and Voia, M. (2021). Consumer cash withdrawal behaviour: Branch networks and online financial innovation. Technical report, Bank of Canada.

Engert, W., Fung, B., and Segendorf, B. (2019). A tale of two countries: Cash demand in canada and sweden. Technical report, Bank of Canada.

Engert, W., Shcherbakov, O., and Stenzel, A. (2024). The impact of central bank digital currency on payments. Technical report, mimeo.

Fulford, S. and Schuh, S. D. (2017). Credit card utilization and consumption over the life cycle and business cycle. *Consumer Financial Protection Bureau Office of Research Working Paper*, (2017-03).

Greene, C. and Schuh, S. D. (2017). The 2016 diary of consumer payment choice. *Research Data Reports Paper*, (17-7).

Henry, C., Shimoda, M., and Rusu, D. (2024). 2023 methods-of-payment survey report: The resilience of cash. Technical report.

Henry, C. S., Huynh, K. P., and Welte, A. (2018). 2017 methods-of-payment survey report. Technical report, Bank of Canada Staff Discussion Paper.

Huynh, K., Molnar, J., Shcherbakov, O., and Yu, Q. (2020). Demand for payment services and consumer welfare: the introduction of a central bank digital currency. Technical report, Bank of Canada.

Huynh, K. P., Nicholls, G., and Shcherbakov, O. (2022). Equilibrium in two-sided markets for payments: Consumer awareness and the welfare cost of the interchange fee. Technical report, Bank of Canada Staff Working Paper.

Koulayev, S., Rysman, M., Schuh, S., and Stavins, J. (2016). Explaining adoption and use of payment instruments by us consumers. *The RAND Journal of Economics*, 47(2):293–325.

Lippi, F. and Secchi, A. (2009). Technological change and the households' demand for currency. *Journal of Monetary Economics*, 56(2):222–230.

References ii

Malone, J. B., Nevo, A., and Williams, J. W. (2021). The tragedy of the last mile: Economic solutions to congestion in broadband networks. *Working Paper*.

McManus, B., Nevo, A., Nolan, Z., and Williams, J. W. (2022). The steering incentives of gatekeepers in the telecommunications industry. Technical report.

Moracci, E. (2022). Cash management and payment choices revisited. *working paper*.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. *Econometrica*, 84(2):411–443.

Rogoff, K. S. (2017). The curse of cash. In *The Curse of Cash*. Princeton University Press.

Smith, G. W. (1986). A dynamic baumol-tobin model of money demand. *The Review of Economic Studies*, 53(3):465–469.

Tobin, J. (1956). The interest-elasticity of transactions demand for cash. *The review of Economics and Statistics*, pages 241–247.

Wakamori, N. and Welte, A. (2017). Why do shoppers use cash? evidence from shopping diary data. *Journal of Money, Credit and Banking*, 49(1):115–169.

Whitesell, W. C. (1989). The demand for currency versus debitable accounts: note. *Journal of Money, Credit and Banking*, 21(2):246–251.

Model: Setup I

Return

- discrete time, $t = 0, 1, \ldots, \infty$ (infinite horizon)
- Consumers $i = 1, \ldots, N_i$:
  - → cash inventories, $h_{it-1} \geq 0$, from previous period
  - → exogenous consumption demand $s_{it} = s_i + \varepsilon_{it},\ \varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_{si}^2\right)$
  - → allocate payments for consumption b/w cash, $c_{it} \geq 0$, and non-cash, $s_{it} - c_{it} \geq 0$
  - → can withdraw cash, $w_{it} \geq 0$, at fixed cost depending on F_i and d_i
    - d_i: distance to branch network (data), F: scaling parameter (estimated)

$$\max_{(c_{it}, w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty} \beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}}$$

$$s.t.\ \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \leq \text{cash use} \leq \text{cash} + \text{withdrawal}},\quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \leq \text{expenditure}},\quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}},\quad \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + (1-\alpha) \ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{-F \times \mathbb{1}_{w_{it}>0} \ln(1 + d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma \left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + (1-\alpha)\ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{-F \times \mathbb{1}_{w_{it}>0} \ln(1+d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma\left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + (1-\alpha)\ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{-F \times \mathbb{1}_{w_{it}>0}\ln(1+d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma\left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + \left(1 - \alpha\right) \ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{-F \times \mathbb{1}_{w_{it}>0} \ln(1 + d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma \left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Dynamics

Return

- recursive formulation of the dynamic programming problem

$$V\left(h_{it-1}, s_i, d_i, \varepsilon_{it}\right) = \max_{c_{it}\geq 0, w_{it}\geq 0} \left\{ u(h_{it-1}, s_i + \varepsilon_{it}, d_i, w_{it}, c_{it}) + \beta \int V(h_{it}, s_i, d_i, \varepsilon_{it+1}) dF_\varepsilon \right\}$$

$$s.t.\ c_{it} \leq \min\{h_{it-1} + w_{it}, s_{it}\}, h_{it} = h_{it-1} + w_{it} - c_{it}$$

→ (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}

- flow utility & withdrawal cost depend on (c_{it}, w_{it})
- holding cost & future value depend on h_{it}

→ recast problem as choice of h_{it} (Δh_{it}), inducing conditionally optimal (c_{it}, w_{it})

- analytical characterization of $(c^*_{it}(\Delta h_{it}), w^*_{it}(\Delta h_{it}))$
- withdrawal $w_{it} > 0$ triggered by (i) depleted cash reserves or (ii) large consumption shock
- mix of cash and non-cash throughout; withdrawal before cash is depleted

Dynamically Optimal Choice

Return

- given (w_{it}^*, c_{it}^*) we obtain $\tilde{u}(h_{it}; h_{it-1}, s_{it}) = u(h_{it-1}, s_{it}, d_i, c_{it}^*(h_{it}), w_{it}^*(h_{it}))$ and thus

$$V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{ \tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon \right\}$$

  → Bellman equation with current cash holdings as state & future cash holdings as control

- approximate $V(\cdot)$ by backward induction using a 183-period (6 months) planning horizon
  - → robust to perturbations in length of planning horizon, discount factor ($\beta \approx 0.95$)
  - → yields policy $h_{it}^*(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w_{it}^*, c_{it}^*)
  - → implied behavior in line with theoretical predictions
    - withdrawal triggered by depleted cash reserves or large expenditure s_{it}
    - costly withdrawal implies fixed post-consumption cash inventory $\tilde{h} = \arg\max_h E[V(\cdot)] - \gamma \cdot h$

Dynamically Optimal Choice

Return

- given (w_{it}^*, c_{it}^*) we obtain $\tilde{u}(h_{it}; h_{it-1}, s_{it}) = u(h_{it-1}, s_{it}, d_i, c_{it}^*(h_{it}), w_{it}^*(h_{it}))$ and thus

$$V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{ \tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon \right\}$$

  → Bellman equation with current cash holdings as state & future cash holdings as control

- approximate $V(\cdot)$ by backward induction using a 183-period (6 months) planning horizon
  - → robust to perturbations in length of planning horizon, discount factor ($\beta \approx 0.95$)
  - → yields policy $h_{it}^*(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w_{it}^*, c_{it}^*)
  - → implied behavior in line with theoretical predictions
    - withdrawal triggered by depleted cash reserves or large expenditure s_{it}
    - costly withdrawal implies fixed post-consumption cash inventory $\tilde{h} = \arg\max_h E[V(\cdot)] - \gamma \cdot h$

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$

- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - trade-off b/w withdrawal & sub-optimal cash usage
    - either withdraw & implement ideal cash use...
    - or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$

- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - trade-off b/w withdrawal & sub-optimal cash usage
    - either withdraw & implement ideal cash use...
    - or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$

- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - trade-off b/w withdrawal & sub-optimal cash usage
    - either withdraw & implement ideal cash use...
    - or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$

- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - trade-off b/w withdrawal & sub-optimal cash usage
    - either withdraw & implement ideal cash use...
    - or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

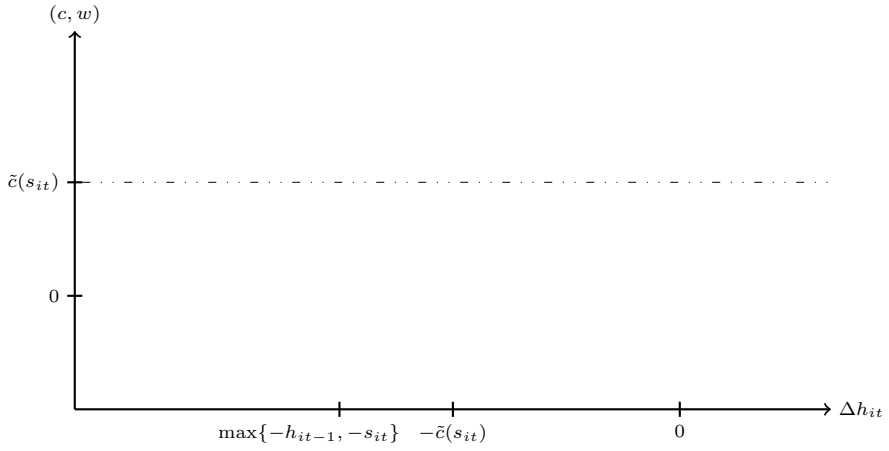
Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$

- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - trade-off b/w withdrawal & sub-optimal cash usage
    - either withdraw & implement ideal cash use...
    - or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

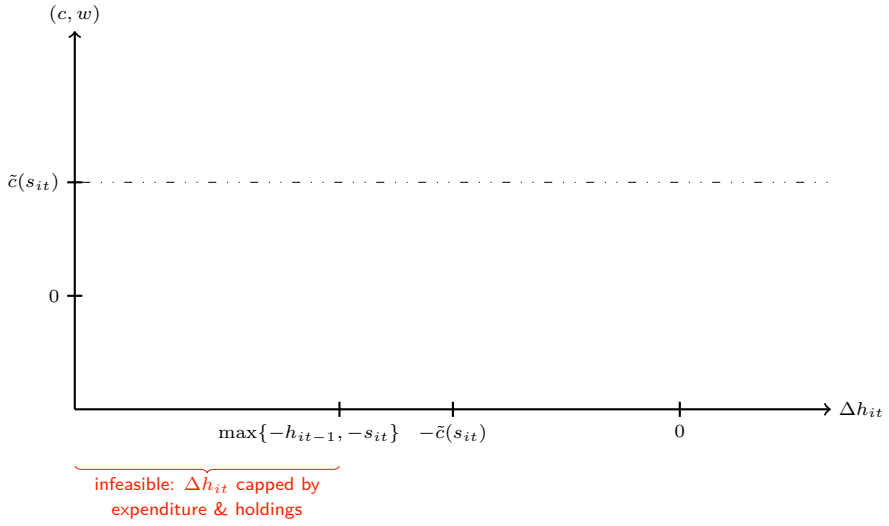
Model: In-period solution

Return



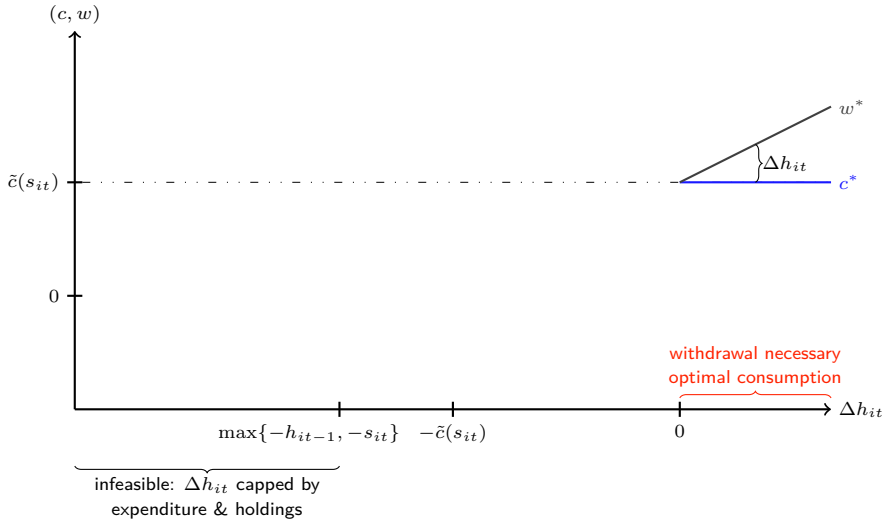
Model: In-period solution

Return



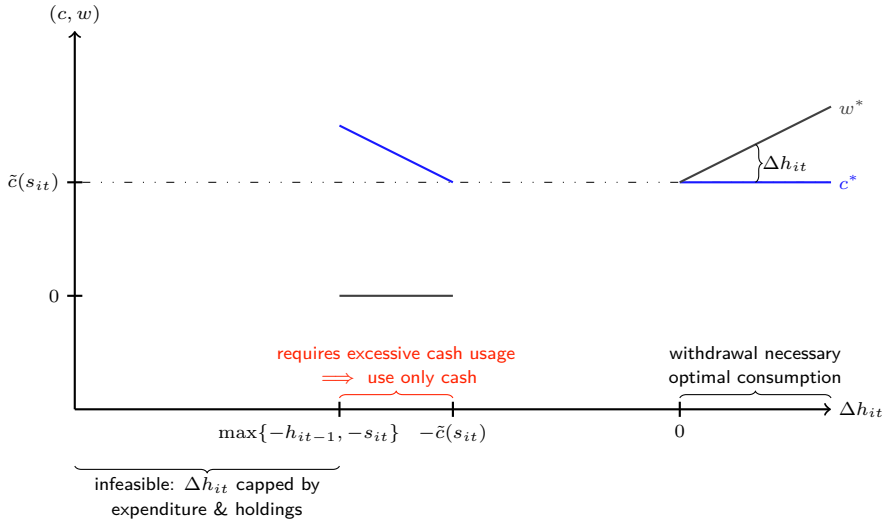
Model: In-period solution

Return



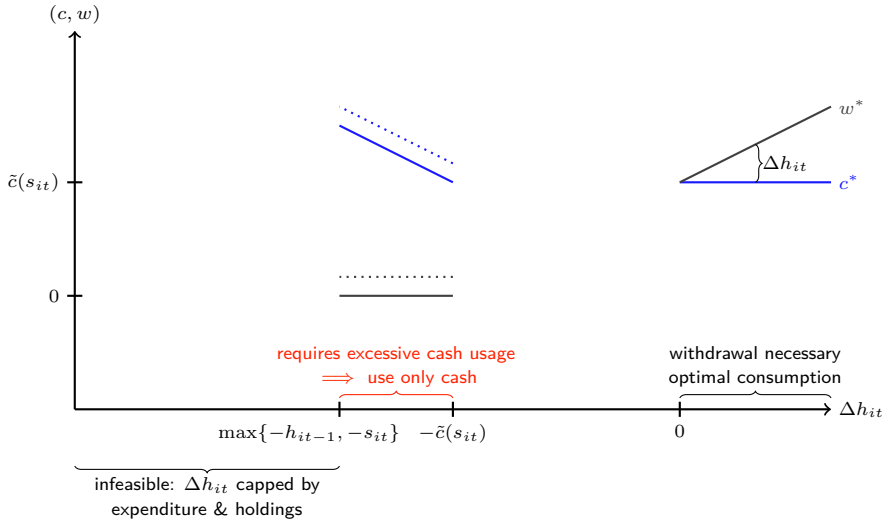
Model: In-period solution

Return



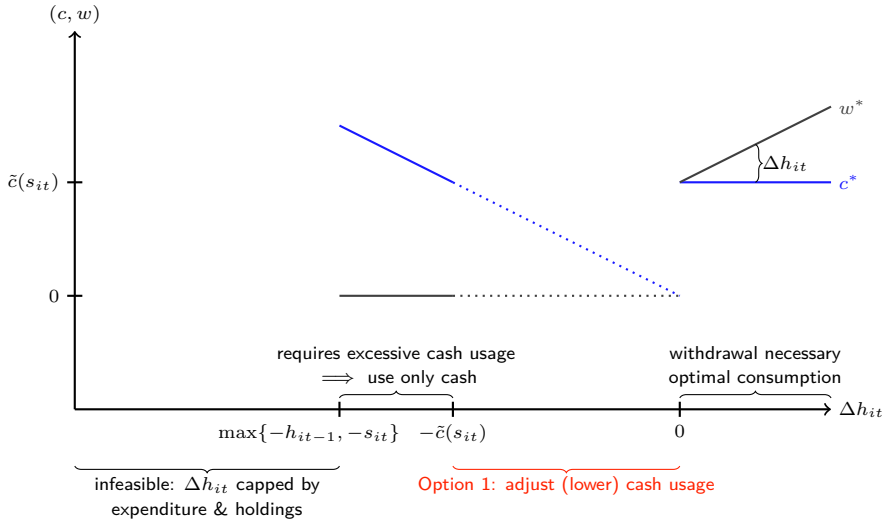
Model: In-period solution

Return



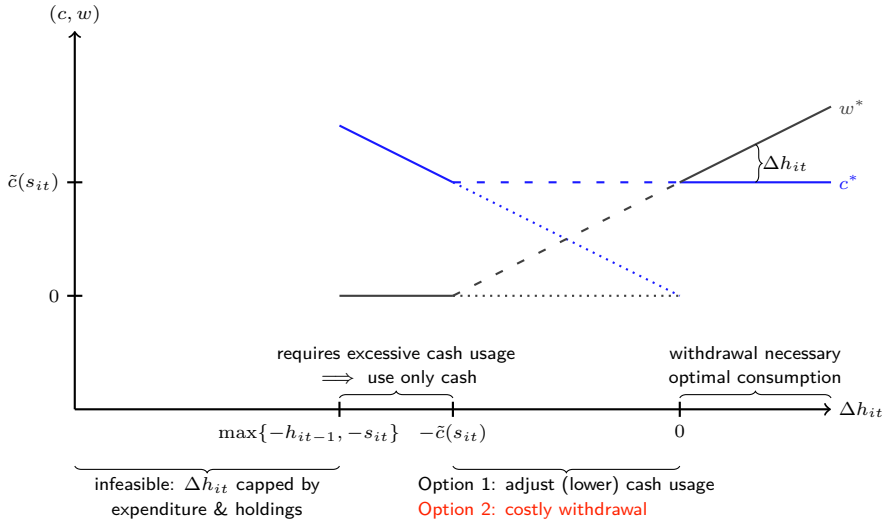
Model: In-period solution

Return



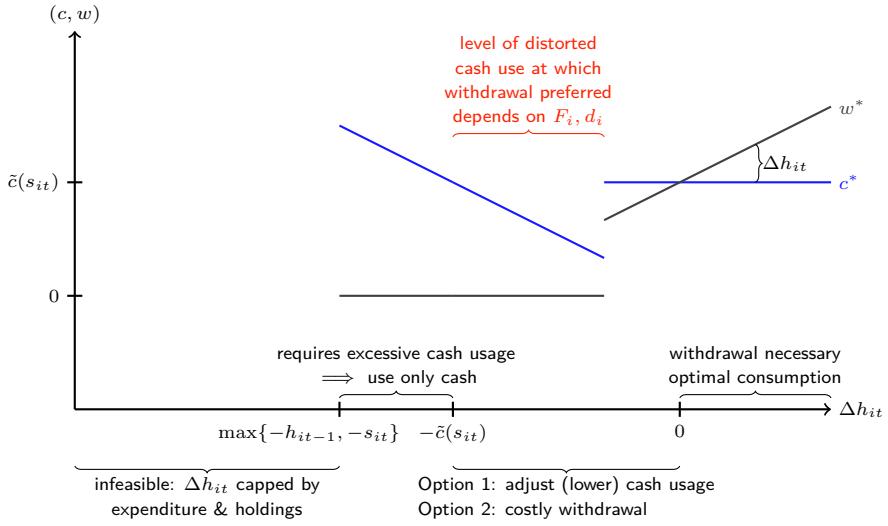
Model: In-period solution

Return



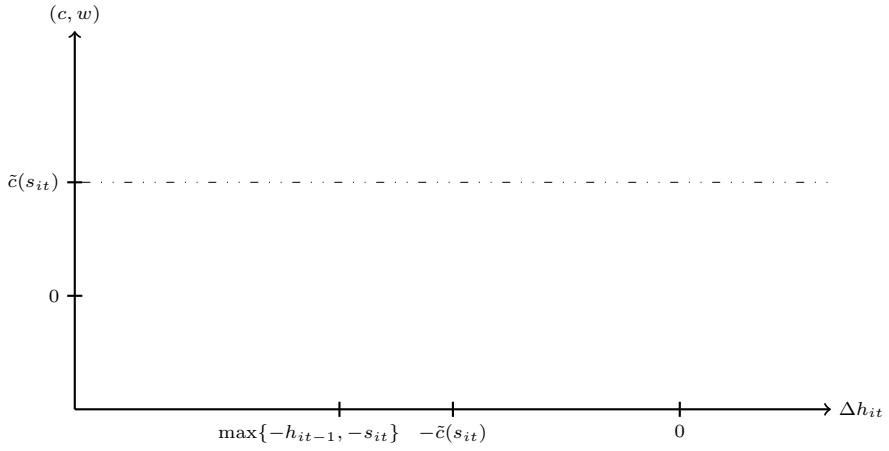
Model: In-period solution

Return



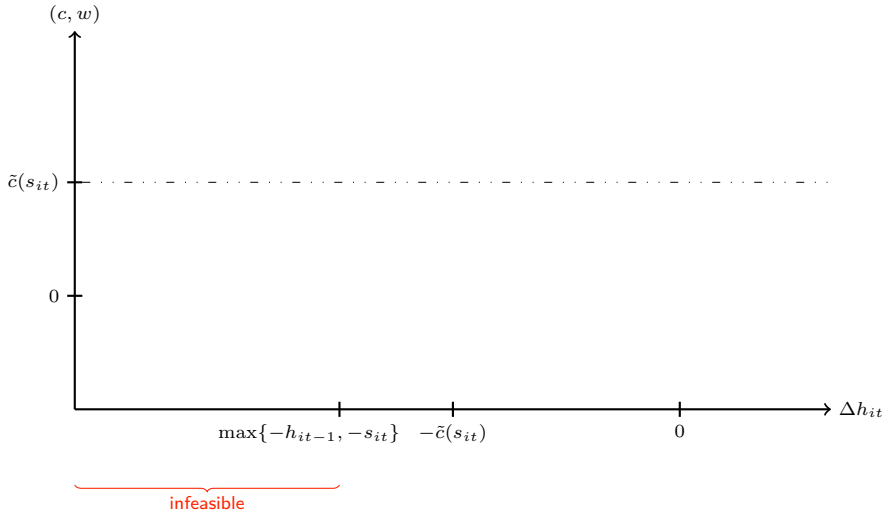
Model: In-period solution:

Return



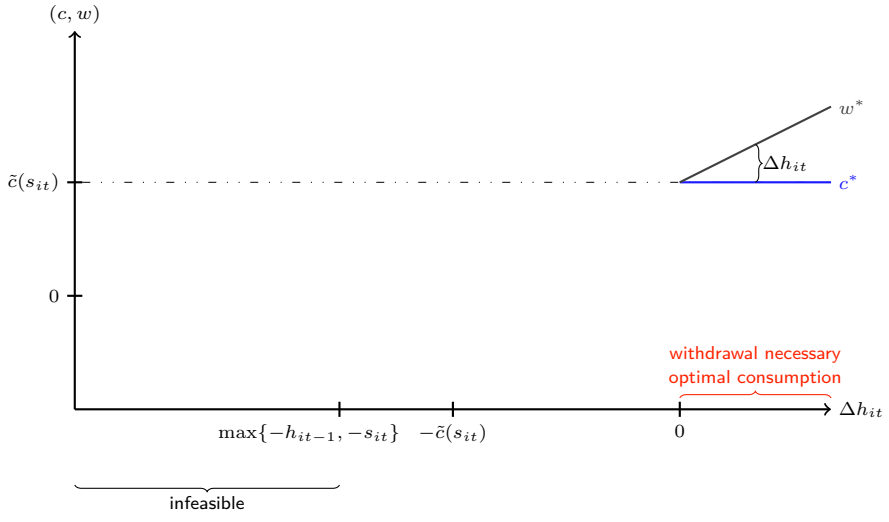
Model: In-period solution:

Return



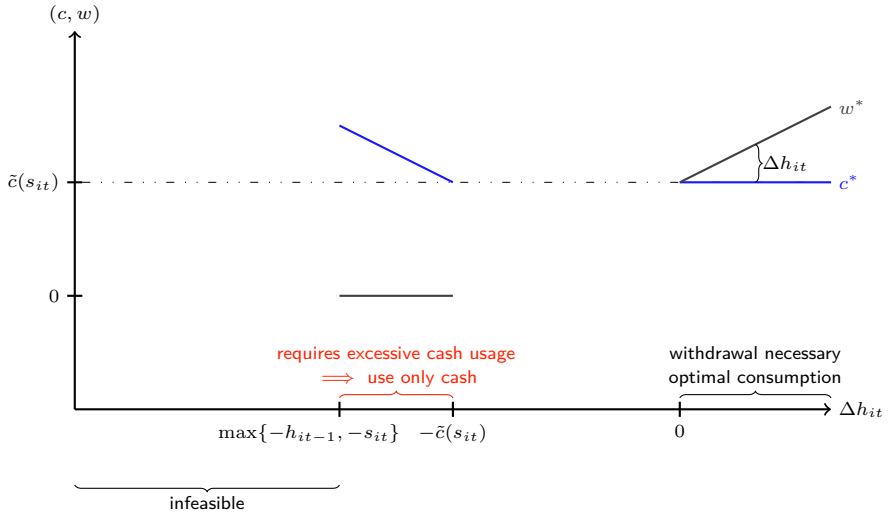
Model: In-period solution:

Return



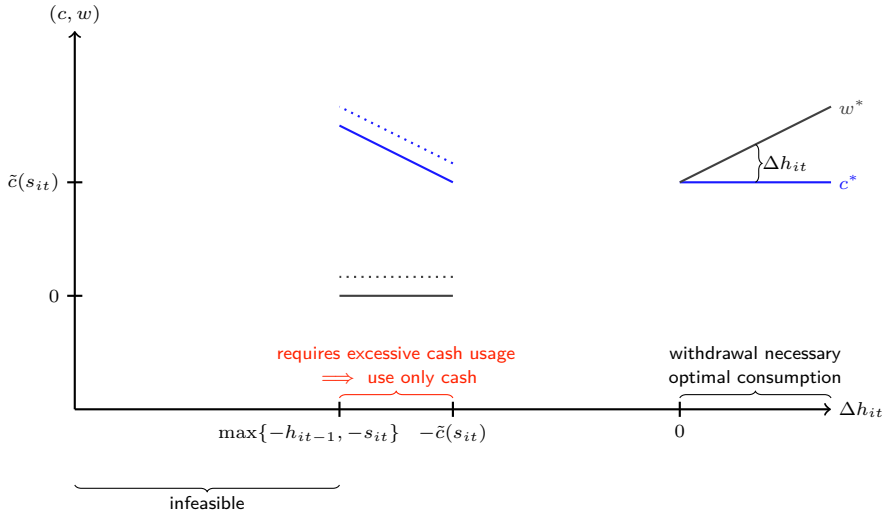
Model: In-period solution:

Return



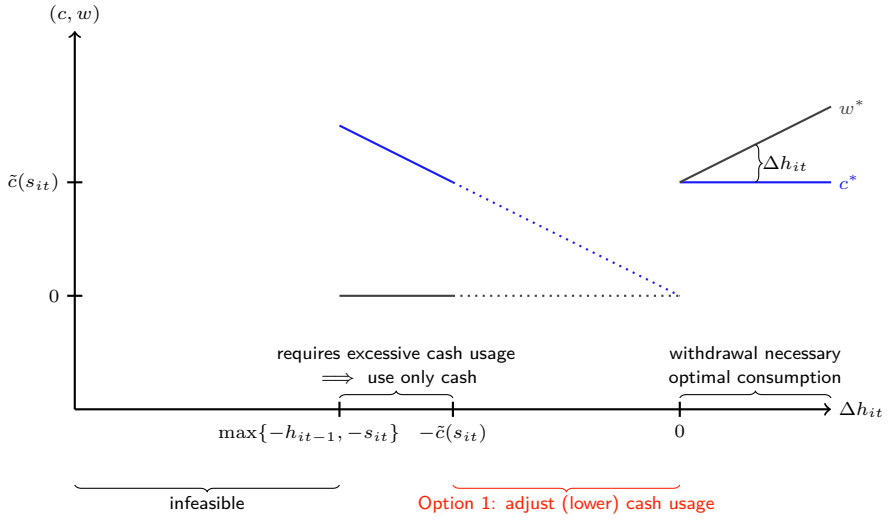
Model: In-period solution:

Return



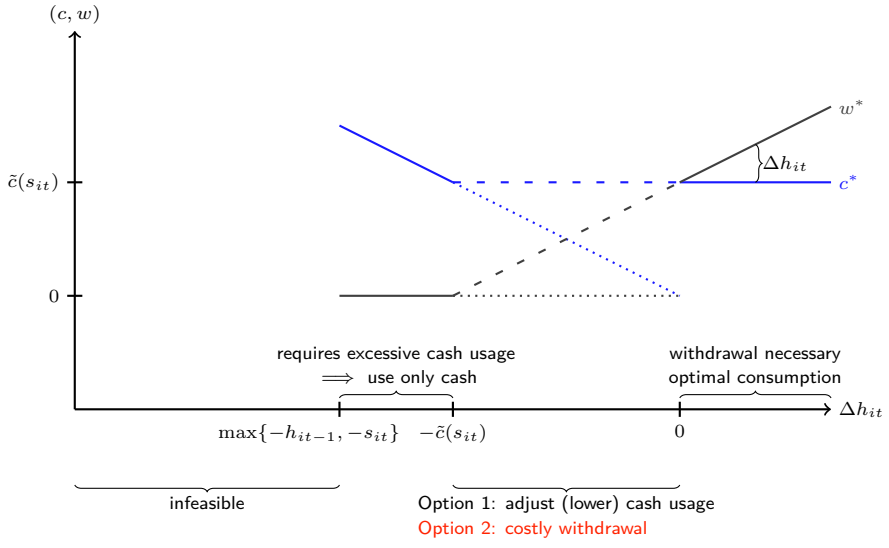
Model: In-period solution:

Return



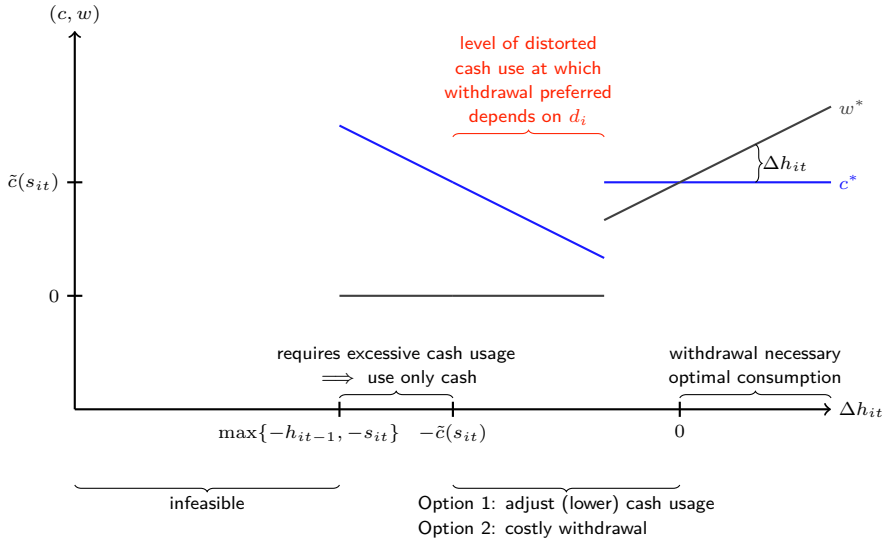
Model: In-period solution:

Return



Model: In-period solution:

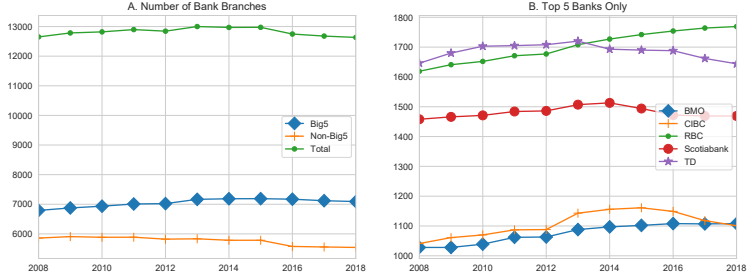
Return



Data: Distance to Bank Network

Return

- d_i reflects density of bank branches in forward sortation area (FSA, 1643 in CAN)
  - → $d_i = \frac{\ln(\text{geographic area of FSA})}{\#\text{ of banch branches}}$
  - → ATMs by FIs typically co-located with branches (88%)
  - → white-label ATMs only account for ≈ 20% of ATM withdrawals (Chen et al. 2021)



Evolution of bank branches.

Evolution over time by Demographic Group

Return

Table 1: Evolution of parameter estimates by demographic type (median estimates)

(a) Cash elasticity, α						
2009	0.282	0.240	0.269	0.311	0.259	0.344
2013	0.282	0.216	0.297	0.232	0.257	0.333
2017	0.245	0.235	0.297	0.234	0.245	0.257
(b) Holding cost, γ						
2009	0.00161	0.00162	0.00062	0.00091	0.00112	0.00153
2013	0.00209	0.00150	0.00072	0.00054	0.00080	0.00255
2017	0.00154	0.00134	0.00077	0.00049	0.00082	0.00145
(c) Withdrawal cost, F						
2009	0.257	0.258	0.141	0.209	0.269	0.084
2013	0.298	0.276	0.240	0.223	0.290	0.105
2017	0.394	0.392	0.269	0.296	0.375	0.100
	Young & Poor	Young & Rich	Old & Poor	Old & Rich	Urban	Rural
N-obs.	924	749	1189	562	2902	522

Evolution of Infrastructure (Sample)

Return

Year	Mean	p10	p25	Median	p75	p90	N-obs.
2009	288.14	0.71	1.67	6.35	111.66	573.90	973
2013	160.32	0.52	1.24	3.42	40.72	279.70	1408
2017	205.65	0.48	1.24	3.48	38.72	287.62	1043
All	210.50	0.57	1.38	3.95	55.62	349.48	3424

Notes: The distance measure is the natural logarithm of the geographic area of the forward sortation area (FSA) consumers are located in in square kilometers over the number of available bank branches.

Evaluation of Factual Infrastructure Changes: Cash Holdings

Return

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.18%	0.23%	-0.13%	1.16%	0.16%	-0.18%	-0.06%	0.75%
2017	0.13%	0.23%	-0.35%	0.45%	0.70%	-0.15%	-0.09%	2.69%

Elasticity

Return

Elasticity of predictions w.r.t. increase in distance (25% change)

(a) All consumers	mean	min	p10	p25	p50	p75	p90	max
average withdrawal amount	-0.75	-4.00	-4.00	-0.01	0.01	0.09	0.49	2.32
expected withdrawal frequency	-1.19	-4.00	-4.00	-2.06	-0.34	-0.12	-0.05	0.00
average cash holding	-0.61	-4.00	-4.00	-0.13	-0.02	0.00	0.37	3.59
average cash use	-1.09	-4.00	-4.00	-1.81	-0.19	-0.06	-0.01	0.00
expected payoff per period	-0.14	-3.98	-0.30	-0.08	-0.01	-0.00	-0.00	-0.00
Observations	3161							

(b) Cash users (post increase)	mean	min	p10	p25	p50	p75	p90	max
average withdrawal amount	0.16	-0.28	0.00	0.00	0.01	0.24	0.56	2.32
expected withdrawal frequency	-0.39	-3.58	-0.73	-0.44	-0.23	-0.08	-0.04	0.00
average cash holding	-0.00	-3.48	-0.17	-0.06	-0.02	0.00	0.46	3.59
average cash use	-0.27	-3.48	-0.48	-0.27	-0.12	-0.04	-0.01	0.00
expected payoff per period	-0.05	-2.65	-0.19	-0.03	-0.01	-0.00	-0.00	-0.00
Observations	2465							

(c) Cash non-users (post increase)								
expected payoff per period	-0.46	-3.98	-1.46	-0.37	-0.07	-0.00	-0.00	-0.00
Observations	696							

- focus on 25% change
  - → reaffirm bi-modality
  - → 22% no longer use cash
- Who does what?

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Istanbul electricity demand forecast with artificial neural networks[1]

Hayriye Yasak Ozkal,
Central Bank of the Republic of Türkiye

Buse Kaylan, Meltem Sipahi and Muammer Altan Cakir,
Istanbul Technical University

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Istanbul Electricity Demand Forecast with Artificial Neural Networks

Hayriye Yasak OZKAL[1], Buse KAYLAN[2], Meltem SIPAHI[3]

and Prof. Dr. Muammer Altan CAKIR[4]

Abstract

When examining the relationship between supply and demand for electricity consumption, overestimating the electrical load can lead to unnecessary energy purchases and excessive reserves, while underestimating the load can cause insufficient energy supply, increased outages, and consequently, system instability. Optimizing electricity demand forecasts to ensure efficient resource usage and reduced outages is becoming a crucial issue for energy producers, distribution companies, and consumers. This study evaluates the impact of seasonal variables such as temperature and humidity, as well as official and weekends on electricity consumption in Istanbul, using Energy Exchange Istanbul (EXIST) transparency reports. The aim is to enhance prediction accuracy by analysing the influence of these parameters on consumption forecasts. The dataset, characterized by time series features, was examined, and after applying NARX models and Prophet/LSTM hybrid models, the final model was selected based on the lowest error metrics and the highest forecast accuracy. The results indicated that Istanbul's electricity consumption is related to factors like humidity, temperature, intra-day hours, and holiday periods.

Keywords: Electricity demand forecasting, Machine learning algorithms, Prophet/LSTM model, NARX model, Generative Adversarial Networks (GAN)

JEL classification: C45, C52, L94, Q47

[1] Department of Data Governance and Statistics, Central Bank of the Republic of Türkiye,

email: hayriye.yasakozkal@tcmb.gov.tr

[2] Department of Data Engineering and Business Analytics, Istanbul Technical University,

email: kaylanb23@itu.edu.tr

[3] Department of Data Engineering and Business Analytics, Istanbul Technical University,

email: sipahigunindi23@itu.edu.tr

[4] Department of Physics Engineering, Istanbul Technical University, email: cakirmua@itu.edu.tr

1. Introduction

As the population, industry and technology-based commercial activities and the demand for technology increase in Turkiye over the years, the demand for electrical energy increases in parallel. The non-storability of electricity affects the electricity generation, distribution and consumption related planning and operational strategies.

The fact that the amount of electricity consumed is greater than the amount of electricity produced, or that the amount of electricity produced is greater than the amount consumed, leads to an energy imbalance. Accordingly, if load forecasts are overestimated, this can lead to unnecessary energy purchases and excess reserves, while if load forecasts are lower than actual consumption, this can lead to insufficient energy supply and system instability due to outages. For this reason, many decisions in the electricity markets depend on accurate forecasting of consumer demand in order to ensure the efficient operation of the energy markets, to avoid wasting resources and to reduce costs, and on accurate load forecasts that ensure a balance between supply and demand (Akman et al., 2018).

In order to ensure system efficiency in supply and demand management, electricity is traded through the Day-Ahead Market (DAM) and Intraday Market (IMM) in addition to long-term bilateral agreements. DAM tries to ensure a balance between supply and demand by planning production and consumption a day in advance. The participants submit their hourly bids and calls for the next day, and Energy Exchange Istanbul matches these offers. In cases where the energy supply-demand balance cannot be achieved through the DAM, the imbalance is attempted to be eliminated by making flexible and instantaneous transactions through the IMM. The aim is to make energy trading more efficient by maintaining the balance between supply and demand instantaneously for both producers and consumers.

In previous studies on electricity demand forecasting, perceived temperature, weekends, and holidays have been considered as parameters in electricity consumption predictions. In this study, the periodic influence of the relevant variables that affecting electricity consumption on prediction accuracy are investigated. After analysing their impact on consumption trends, a deep learning model is developed with the aim of improving prediction accuracy.

2. Materials and Methods

During the energy demand forecasting literature review to determine which model to use, it was found that long-term (year), medium-term (month), short-term (day/hour), and real-time (minute/second) forecasts can be made using historical data and appropriate forecasting models. In an examination of previous studies, it can be seen that market conditions, seasonal variables, social and demographic events are considered within the context of the research and an attempt is made to forecast the load. For this purpose, various models based on multiple linear regression analysis (MLR), artificial neural networks (ANN), ARIMA, least-squares support-vector machines (LS-SVM) methods have been developed and according to the results of

the study, it is understood that ANN and LS-SVM are compatible with the actual values in the short-term forecasts (Akman et al., 2018). Therefore, NARX (Nonlinear Autoregressive External Input Model) models based on SimpleRNN (Recurrent Neural Networks) and LSTM (Long-Short Term Memory) artificial neural networks, and Prophet-LSTM hybrid models were developed using electricity consumption data for Istanbul and their short-term (1 hour) performance was tested.

2.1. Data

The data period used in this study is from January 2021 to December 2023. Hourly temperature and humidity data are obtained from the open-source web service Open-Meteo, and actual hourly electricity consumption is in megawatt hours (MWh) and is obtained from Energy Exchange Istanbul (EXIST).

2.1.1. Generating Synthetic Data

The hourly electricity consumption data obtained from EXIST is representative of Turkiye as a whole, and no province-based hourly data is available. Therefore, it is proposed to synthetically generate hourly electricity consumption data for Istanbul using Generative Adversarial Networks (GANs). The model development process consists of two stages. First, a synthetic version of Turkiye's hourly electricity consumption data was created. Subsequently, the generated data set is scaled according to the monthly consumption rates of Istanbul province. Finally, the hourly electricity consumption data of Istanbul province is obtained.

The GAN model used in first stage consists of a generator and a discriminator network. The purpose of the generator is to produce realistic looking data points from a random input vector. The Generator used in this study is configured with three dense layers of increasing units. These layers have 256, 512 and 1024 units respectively, and each layer is supplemented by the LeakyReLU activation function and batch normalisation. LeakyReLU accelerates the learning process by being particularly sensitive to negative input values, while Batch Normalisation ensures that data distributions between layers are normalised. The output layer is terminated by a single unit linear activation function. The Discriminator is designed to distinguish whether the data is real or generated by the Generator. The Discriminator is structured in three dense layers with 1024, 512 and 256 units in descending order. These layers have the LeakyReLU activation function, and Dropout (0.3) is applied to each layer to avoid overfitting. The sigmoid activation function is used in the output layer of the Discriminator. This provides an output to determine the probability of the data being real. The model is implemented using the adversarial training process, where the Generator and the Discriminator networks are trained and both iteratively attempt to outperform each other. In order to train the model long enough, this training process lasted for 10,000 epochs.

In the second stage of the study, the synthetic data generated by the GAN model is scaled according to the monthly consumption percentages of the Istanbul province in Turkiye, as provided by the Electricity Market Development Reports of the Energy Market Regulatory Authority (EMRA). The similarity between synthetic and actual patterns was evaluated, and due to observed inconsistencies, the dataset was excluded from further analysis. Instead, the hourly electricity consumption data for

Istanbul has been obtained by scaling with the province-specific consumption percentage billed throughout Turkiye, which is also given in the report.

2.2.Exploratory Data Analysis (EDA)

In order to understand the variability of energy demand over time and the impact of factors such as seasonality, holidays and working hours on demand, hourly electricity consumption data for the province of Istanbul from January 2021 to December 2023 was analysed. Density plot of this data, as shown in Figure 1, shows that the consumption values generally vary between 5,000 and 7,000 MW, but there are also some values outside this range. When the probability distribution is analysed (see Figure 2), it is evident that while the data appear to be approximately normally distributed, significant deviations are observed, particularly at the extreme points.

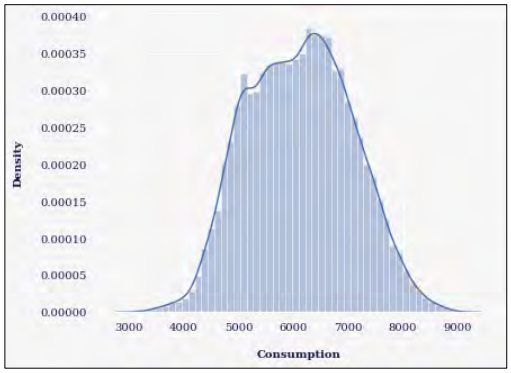


Figure 1. Density Plot

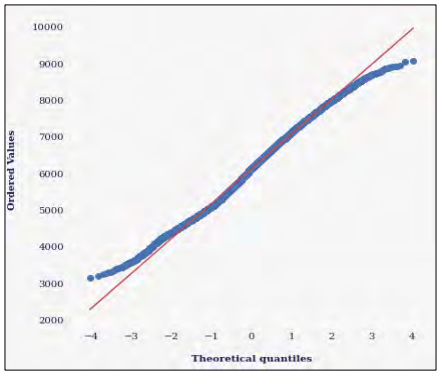


Figure 2. Probability Plot

Analysis of the Istanbul's hourly electricity consumption graph as illustrated in Figure 3, shows that there are significant peaks in electricity consumption in the middle and toward the end of the year. Given that the observed periods coincide with the summer and winter months, seasonal effects are considered to be significant. During specific periods, such as religious holidays, a notable decline in electricity consumption is observed (indicated by the red arrows). These decreases may be the result of the seasonal slowdown in economic activity that typically occurs during the holiday period.

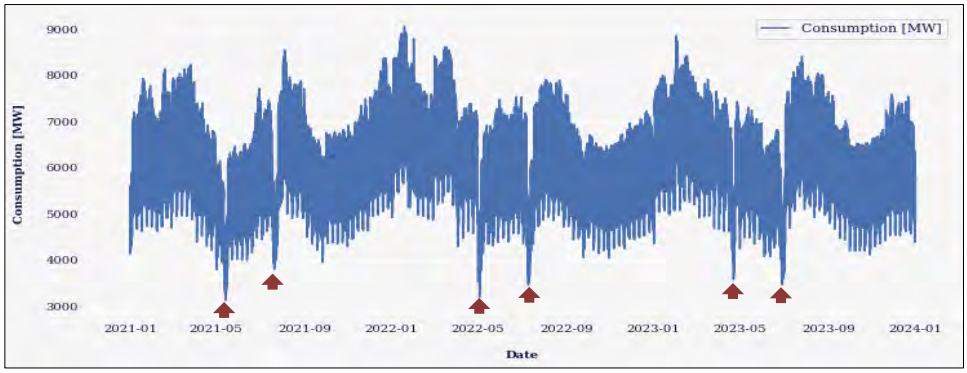


Figure 3. Istanbul Hourly Electricity Consumption Plot

An analysis of seasonal trends in electricity consumption (Figure 4) shows that there has been a significant increase in electricity consumption, particularly in the winter (December-January) and summer (July-August) months. Conversely, during the spring and autumn seasons, consumption tends to be more balanced and at lower levels. The cyclical patterns that regularly recur in the data are assessed to be explained by factors such as seasonal effects and operational calendars.

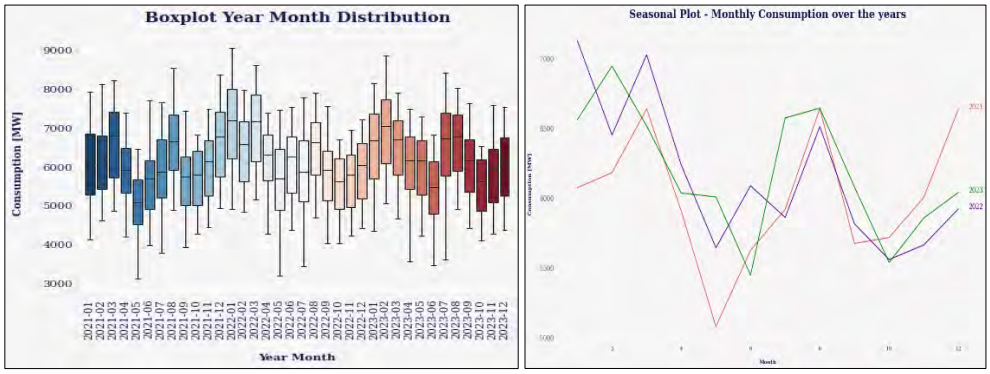


Figure 4. Istanbul Monthly Electricity Consumption Distribution Charts

As illustrated in Figure 5, electricity consumption is higher on weekdays due to the intensity of industrial and commercial activities. On the weekends, the decrease in these activities leads to a decrease in the total consumption of electricity.

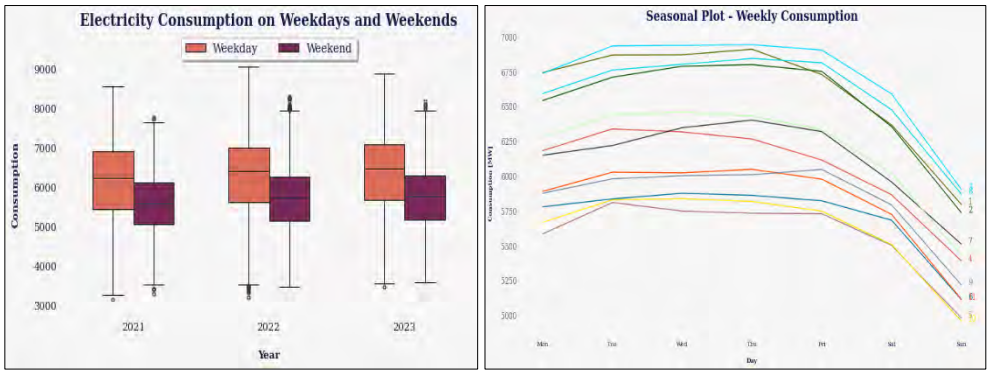


Figure 5. Istanbul Electricity Consumption Charts - Weekday/Weekend and Weekly

Analysis of the hourly electricity consumption trend reveals a sudden increase in consumption around 5:00-6:00 a.m., which continues until 11:00 a.m. This increase is followed by a gradual decrease in consumption until 17:00-18:00 p.m. (see Figure 6).

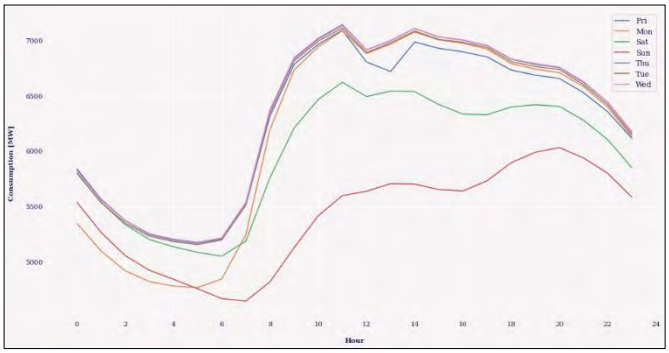


Figure 6. Istanbul Hourly Electricity Consumption Plot

Within the scope of the statistical analysis, the autocorrelation (ACF) and partial autocorrelation functions (PACF) of Istanbul hourly electricity consumption data were analysed, and it was observed that there are regular jumps in the autocorrelation graph for every 24 cycles. Through the analysis of the PACF plot, it is apparent that there are notable autocorrelations for the initial lags.

In this study, artificial neural network (ANN) models are preferred due to their superior performance in resolving complex relationships in nonlinear multivariate data sets when compared to traditional time series methods (Cheng et al., 2017). The single step (1 hour later) forecasting approaches utilized in this context are RNN and LSTM-based NARX models and the Prophet-LSTM hybrid model.

2.3.Overview of Forecasting Models

2.3.1. NARX Neural Network

The nonlinear autoregressive network with exogenous inputs (NARX), which is widely used in time series modelling, uses the past values of the dependent variable as well as the past values of the exogenous explanatory variables to predict the future value of the dependent variable. The mathematical equation used in the NARX model is as follows:

$$y(t) = f(y(t-1), y(t-2), \ldots, y(t-n_y), u(t-1), u(t-2), \ldots, u(t-n_u))$$ [5]

2.3.2. Long Short-Term Memory

LSTM (Long Short-Term Memory) is a specialised type of Recurrent Neural Network (RNN) that is widely used in time series due to its capacity to learn long-term relationships within consecutive time steps. In recurrent neural networks, inputs are received, hidden states are updated based on previous computations, and predictions are generated for each element of the sequence. The LSTM learns both short-term and long-term dependencies at each time step using a memory cell structure, unlike the RNN.

The memory cell is composed of a self-recurrent neuron and three gates: the input gate, the output gate, the forget gate. The gates are is responsible for regulating the interactions between adjacent memory cells and the memory cell itself. The input gate controls whether the input signal can change the state of the memory cell. The output gate regulates the memory cell's state based on if it can change the state of the other memory cell. In addition, the forget gate chooses whether the previous state of the cell is remembered or forgotten (Cocianu and Avramescu, 2020).

[5] The symbol y(t) represents the predicted value (model output). The variables y(t-1), y(t-2) refer to the lagged values of the dependent variable, u(t-1), u(t-2) to the lagged values of the exogenous inputs. ny is the model feedback for the dependent variable, nu is the model feedback for the exogenous inputs, and f is the learned nonlinear function.

2.3.3. Prophet Model

Facebook Prophet is an open-source library developed for the purpose of forecasting time series with complex features including trends, seasonality, and holidays. It is included in the study because it successfully handles trend shifts and outliers in time series. The mathematical equation for Prophet is expressed as follows (Bashir et al., 2022):

$$y(t)=g(t)+s(t)+h(t)+\epsilon t$$ [6]

2.4.Proposed Methodology

Based on the results of the data exploration and analysis, the final dataset was created using electricity consumption data and related factors (Table 1). The dataset covers the time interval from 00:00 on January 1, 2021 to 23:00 on December 31, 2023.

Selected Factors — Table 1

Apparent Temperature
Perceived Humidity
Holiday
Weekend
Related Electricity Consumption Hour
Month of the year

The final dataset for NARX model development is used to create two separate series, the target variable (y) and the attributes (X) then series divided into three parts: 60% training data, 20% validation data for tuning model hyperparameters, and 20% test data for evaluating model performance. The chronological order has been taken into consideration when division is performed. All recorded time series values were then transformed to a normal distribution using the standard MinMaxScaler. The lags of the normalized series up to 24-time steps determined by the PACF plot are added to the feature series and formatted according to the configuration of each model's input layer.

The NARX RNN model used in the study is composed of three layers: an input layer, a SimpleRNN layer with 50 units and a ReLU activation function, and an output layer with one unit for the estimation of the electricity consumption. The model was compiled using the Adam optimizer to adjust the learning rate as the data was trained, and the Mean Squared Error (MSE) loss function to minimize the mean squared differences between predicted and actual values. In order to preserve the

[6] The variable y(t) is the quantity that is being estimated, g(t) is the trend function responsible for modelling non-periodic variations, s(t) is the seasonality (hourly, daily, weekly, yearly), h(t) is the holiday/special day effect, and ϵt is the error term, i.e., the random noise or error that cannot be modelled.

chronological sequence of data, the training of the model was performed in batches of 32 instances for 50 cycles without shuffling and to develop an efficient model, the Early Stopping method was used (Table 2).

Differently from the NARX RNN model, the LSTM-based NARX model has two sequential LSTM layers with 50 units and a dropout layer between them, which disables a random 20% of neurons during training to prevent overfitting.

Summary of Models Table 2

Model Name	Layer (type)	Output Shape	Parameters	Total Parameters
NARX RNN Model	simple_rnn (SimpleRNN)	(None, 50)	2,900	2,951
	dense (Dense)	(None, 1)	51	
LSTM Based NARX Model	lstm (LSTM)	(None, 24, 50)	11,600	31,851
	dropout (Dropout)	(None, 24, 50)	0	
	lstm_1 (LSTM)	(None, 50)	20,200	
	dense (Dense)	(None, 1)	51	
Prophet LSTM Hybrid Model	Prophet Model	(26.256, 32)	0	31,851
	lstm (LSTM)	(None, 24, 50)	11,600	
	dropout (Dropout)	(None, 24, 50)	0	
	lstm_1 (LSTM)	(None, 50)	20,200	
	dense (Dense)	(None, 1)	51	

The objective of developing the Prophet-LSTM hybrid model is to use the Prophet model as the main prediction tool, and feed the remaining errors from the predictions into the LSTM model for more accurate estimation. This method allows LSTM to learn more complex and short-term patterns that Prophet cannot model.

In hybrid model approach, the data set has initially prepared in the input format of the Prophet model, with the time label (ds) containing the relevant time and date of the consumption, the hourly electricity consumption data as the target variable (y), and the lags of the target variables up to 24-time steps and the independent variables designated as exogenous inputs. Exogenous inputs are added to the Prophet model as regressors, and initial consumption predictions are obtained using seasonal and trend components. In the next step, the residuals are calculated based on the actual and predicted consumption values. Subsequently, pre-processing and modelling steps similar to the LSTM-based NARX model are applied to residuals and exogenous inputs, except that ReLU is used instead of the LSTM's default activation function. Thus, the residuals are estimated using delayed values and exogenous inputs.

Lastly, the residuals estimated by the LSTM stack and the trend (*T(t)*), seasonality (*S(t)*) and other components (extra_regressors_additive) estimated by Prophet are added to obtain the final electricity consumption forecast.

$$y(t) = T(t) + S(t) + R(t) + beta * regressor(t)$$

3. Experimental Results

The statistical performance of the NARX RNN, LSTM-based NARX, and Prophet-LSTM hybrid models developed for the forecasting Istanbul's hourly electricity consumption data is measured by the root mean square error (RMSE) and the mean absolute error (MAE) error metrics (Table 3).

Performance Evaluation of the Models — Table 3

MODELS	PERFORMANCE METRICS			
	Train		Test	
	MAE	RMSE	MAE	RMSE
NARX RNN Model	86.65	115.68	90.39	116.87
LSTM Based NARX Model	133.41	168.11	117.48	148.66
Prophet Model	82.93	111.95	85.88	112.22
Prophet-LSTM Hybrid Model	62.17	80.48	65.58	84.73

Evaluating the performance of the models over training and test data shows that the models can generalize the data, except for the LSTM-based NARX model. According to the performance results, the Prophet-LSTM hybrid model outperformed the other models by achieving the lowest error rates on both training and test data.

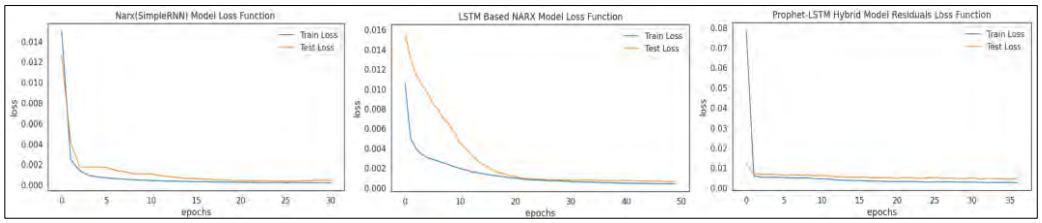


Figure 7. Comparison of Models' Loss Function Graphs

Comparing the loss functions of the models used in the study, it is observed that the training and test losses of all models decreased rapidly in the first cycles but remained stable after a while (Figure 7). As both training and test losses stabilized at similar levels, the potential for overfitting or underfitting was not evident in any of the models.

4. Conclusions and Suggestions for Further Work

Considering the changing social expectations, increasing electricity consumption needs in technology and industry, it is expected that energy imbalance problem will continue in the near future and accordingly, optimization studies in forecasting electricity demand will be crucial. In this study, the optimization problem is investigated through the use of artificial neural networks and hybrid models within the context of the factors that affect the consumption for the province of Istanbul.

According to the study, evaluating the performance metrics and loss function plots together, the Prophet-LSTM hybrid model is more successful in predicting hourly electricity consumption compared to the other models. Also, the high error metric values in the training data indicate that the LSTM-based NARX model could

not fully adapt to the current data set. One of the findings of the study is that synthetic data sets may not replicate time series which contain seasonality, trends, and are affected by exogenous factors as in our case.

We conclude that the results are encouraging and that future efforts should be made to evaluate short- and long-term forecasting performance, improve the model performance by selecting shorter and more recent periods, periodic retraining the model with updated consumption patterns, and recalibrating exogenous variables for adapting model to changing circumstances.

References

Akman, T., Yılmaz, C., & Sönmez, Y. (2018). Elektrik yükü tahmin yöntemlerinin analizi [Analysing electricity load forecasting methods]. *Gazi Journal of Engineering Sciences, 4*(3), 168-175. [in Turkish]

Bashir, T., Chen, H., Tahir, M., & Zhu, L. (2022). Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN. *Energy Reports*, 8, 1678–1686.

Biçen, Y. (2016). Türkiye elektrik enerjisi piyasası gelişim süreci: Gün öncesi ve dengeleme güç piyasası özellikleri [Development process of the Turkish electricity market: features of day-ahead and balancing power markets]. *Karaelmas Fen Ve Mühendislik Dergisi*, *6*(2), 432-438. [in Turkish]

Cheng, Y., Xu, C., Mashima, D., Thing, V.L.L., & Wu, Y. (2017). PowerLSTM: Power demand forecasting using long short-term memory neural network. In *2017 International Conference on Advanced Data Mining and Applications. (Springer)*. (pp. 727–740).

Cocianu, C. L., & Avramescu, M.-S. (2020). The use of LSTM neural networks to implement the NARX model: A case study of EUR-USD exchange rates. *Informatica Economica*, *24*(1), 5–14.

Demirtaş, I., & Tarı, EI. (2021). Türkiye'de elektrik piyasasının rekabete açılması ve rekabetçi elektrik piyasasında elektrik tüketiminin belirleyicileri: ARDL sınır testi [Liberalization of the electricity market in Turkiye and determinants of electricity consumption in a competitive market: ARDL bounds test]. *Journal of Emerging Economies and Policy 6*(1), 148-164. [in Turkish]

Gündüz, S., Uğurlu, U., & Öksüz, İ. (2021). Gün içi piyasası elektrik fiyat tahmini için eksik verilerin tamamlanması [Completion of missing data for intraday market electricity price forecasting]. *Avrupa Bilim Ve Teknoloji Dergisi(25)*, 334-340. [in Turkish]

Karaçor, Z., & Güvenek, B. (2010). Enerji piyasası reformlarının elektrik enerjisi piyasasına etkisi: EÜAŞ ve ayrıcalıklı şirketler üzerine bir analizi [The impact of energy market reforms on the electricity market: an analysis on EÜAŞ and privileged companies]. *Yönetim Ve Ekonomi Dergisi, 17*(1), 147-166. [in Turkish]

Kaysal, K., Akarslan, E., & Hocaoğlu, F. O. (2022). Türkiye kısa dönem elektrik yük talep tahmininde makine öğrenmesi yöntemlerinin karşılaştırılması [Comparison of

machine learning methods in Turkiye short-term electricity load forecasting]. *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi, 9*(2), 693-702. [in Turkish]

Kızıl Voyvoda, Ö., & Voyvoda, E. (2019). Türkiye'de enerji sektörünün yeniden yapılandırılması sürecinde hukuk düzenlemeleri – elektrik sektörü [Legal regulations in the restructuring process of the energy sector in Turkiye - electric sector]. *Çalışma Ve Toplum, 1*(60), 127-154. [in Turkish]

Massaoudi, M., Chihi, I., Sidhom, L., Trabelsi, M., Refaat, S.S., & Oueslati, F. (2019). A novel approach based deep RNN using hybrid NARX-LSTM model for solar power forecasting [Preprint]. *arXiv*. https://arxiv.org/abs/1910.10064

Yıldız, D., Yıldız, G., & Demirci, S. (2023). Fındık üretimi Için öğrenme tabanlı verim tahmini [Learning based yield forecasting for hazelnut production]. *Black Sea Journal of Engineering and Science, 6*(2), 117-126. [in Turkish]

Zuazo, I.Z., Boussaada, Z., Aginako, N., Curea, O., Camblong, H., & Sierra, B. (2021, September 8). Short-Term load forecasting of building electricity consumption using NARX neural networks model. In *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*. (pp. 1–6).

Enerji Piyasası Düzenleme Kurumu. (2021-2023). Elektrik piyasası aylık sektör raporları [Monthly electricity market sector reports]. Retrieved April 25, 2025, from https://www.epdk.gov.tr/detay/icerik/3-0-23/elektrikaylik-sektor-raporlar [in Turkish]

TÜRKİYE CUMHURİYET MERKEZ BANKASI

İTÜ

ISTANBUL ELECTRICITY DEMAND FORECAST WITH ARTIFICIAL NEURAL NETWORKS

Hayriye Yasak OZKAL, Buse KAYLAN, Meltem SIPAHI
Advisor: Prof. Dr. Altan ÇAKIR

19.02.2025

Rome, Italy



Disclaimer

The views are those of the authors and do not necessarily reflect the views of the Central Bank of the Republic of Türkiye or its board members

TÜRKİYE CUMHURİYET
MERKEZ BANKASI

Content

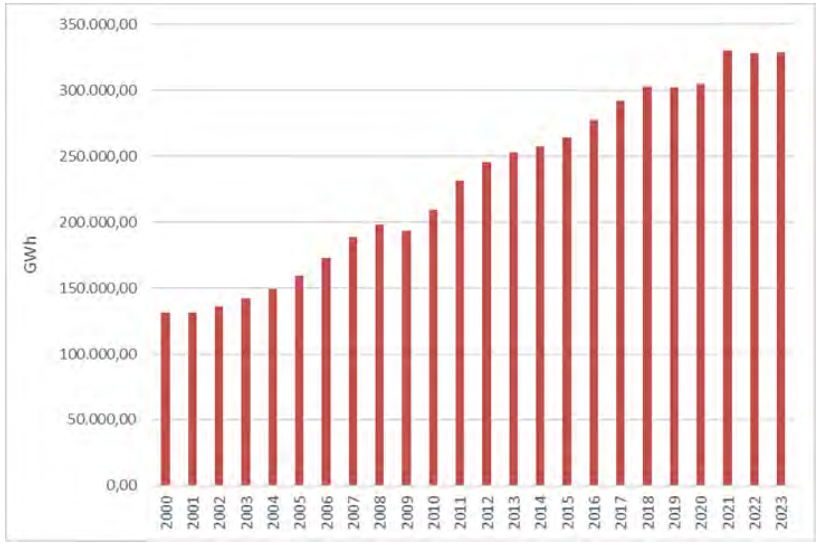
- **Introduction**
- **Main Objectives**
- **Methodology**
  - Data Collection
  - Generating Synthetic Data
  - Exploratory Data Analysis (EDA)
  - Time Series and Variable Selection
  - Data Preprocessing
  - Model Development
  - Model Evaluation
- **Key Findings & Future Work**

Introduction

Number of Consumers

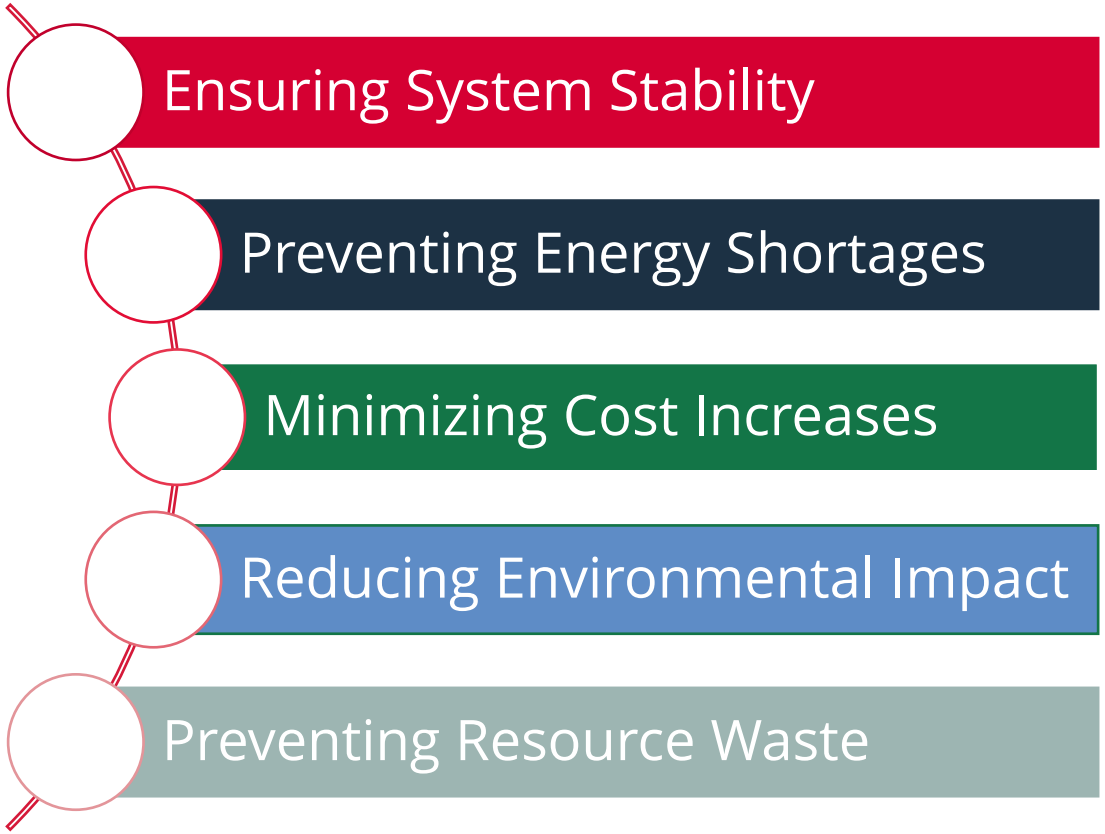
Years	2021	2022	2023	2022-2023 Change
Istanbul	8.346.307	8.482.702	8.594.757	1.32 (%)
Türkiye	47.311.986	48.563.459	49.726.481	2,39 (%)

Development of Actual Consumption by Years (GWh)



As examined in the Energy Market Regulatory Authority (EMRA) 2023 sectoral report, the demand for electricity consumption and the number of consumers have increased over the years, ignoring minor deviations.

Main Objectives



The main objective of energy demand instability projects is to understand, manage and solve the problems caused by inaccurate demand forecasts and the resulting imbalances.

Methodology

Data Collection

- Gathering hourly humidity and temperature data for Istanbul
- Gathering hourly electricity consumption data for Turkey from January 2021 to December 2023
- Generation of Synthetic Data for Istanbul Electricity Consumption (Failed)
- Ratio of Istanbul's monthly electricity consumption to Turkey.

Exploratory Data Analysis (EDA)

- Identify patterns, trends, and correlations between electricity consumption and external factors.
- Examine historical data to understand consumption variations over time.

Model Development

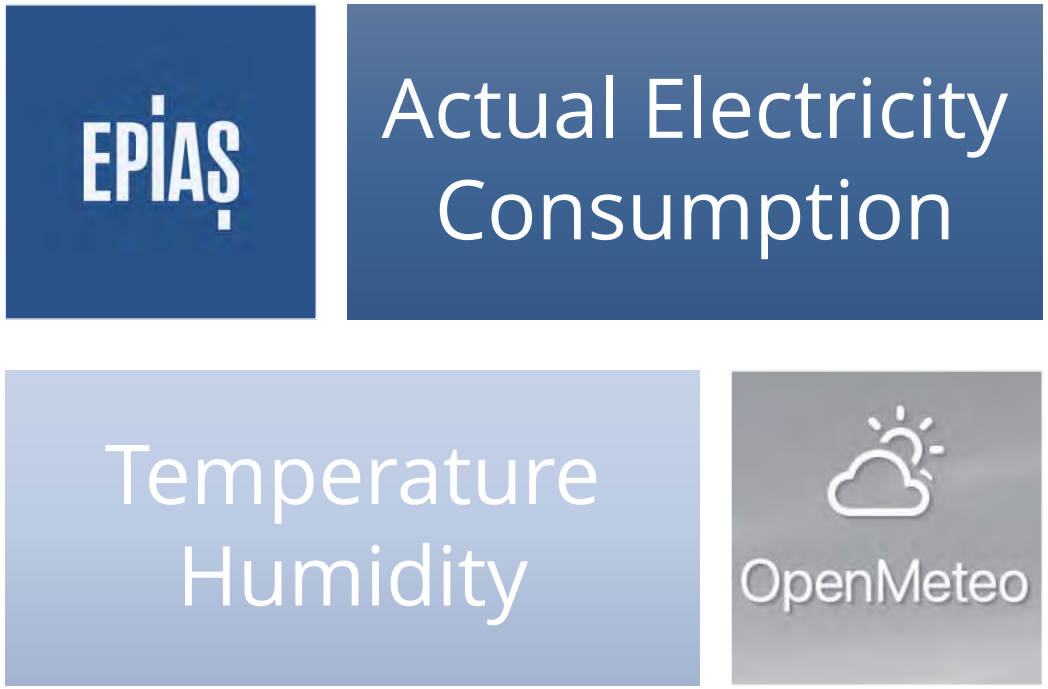
- NARX RNN MODEL
- NARX LSTM MODEL
- PROPHET – LSTM HYBRID MODEL

Performance Evaluation

- Performance Metrics (MAE, RMSE)
- Model Loss Function
- Comparison of Actual and Estimated Values

Methodology

Data Collection



In addition to these data collected on an hourly basis, information such as religious and public holidays, weekends and working days have also been added to the data set.

Methodology

Generating Synthetic Data For Istanbul Electricity Consumption

Data Source

- Total electricity consumption data of Turkey (2021-2023) from Energy Exchange Istanbul (EPIAS)
- Data includes date, hour, and consumption amount in MWh
- Monthy consumption percentages for Istanbul were used to scale the generated data

Challenge

- Lack of city-specific electricity consumption data
- Need for synthetic data to represent Istanbul's consumption

Methodology

GAN Model Architecture

GENERATOR

- **Dense Layer (256 units)** LeakyReLU Activation
- Batch Normalization
- **Dense Layer (512 units)** LeakyReLU Activation
- Batch Normalization
- **Dense Layer (1024 units)** LeakyReLU Activation
- Batch Normalization
- **Output Layer: Dense Layer (1 unit)** Linear Activation

DISCRIMINATOR

- **Dense Layer (1024 units)** LeakyReLU Activation Dropout (0.3)
- **Dense Layer (512 units)** LeakyReLU Activation Dropout (0.3)
- **Dense Layer (256 units)** LeakyReLU Activation Dropout (0.3)
- **Output Layer: Dense Layer (1 unit)** Sigmoid Activation

TRAINING PROCESS

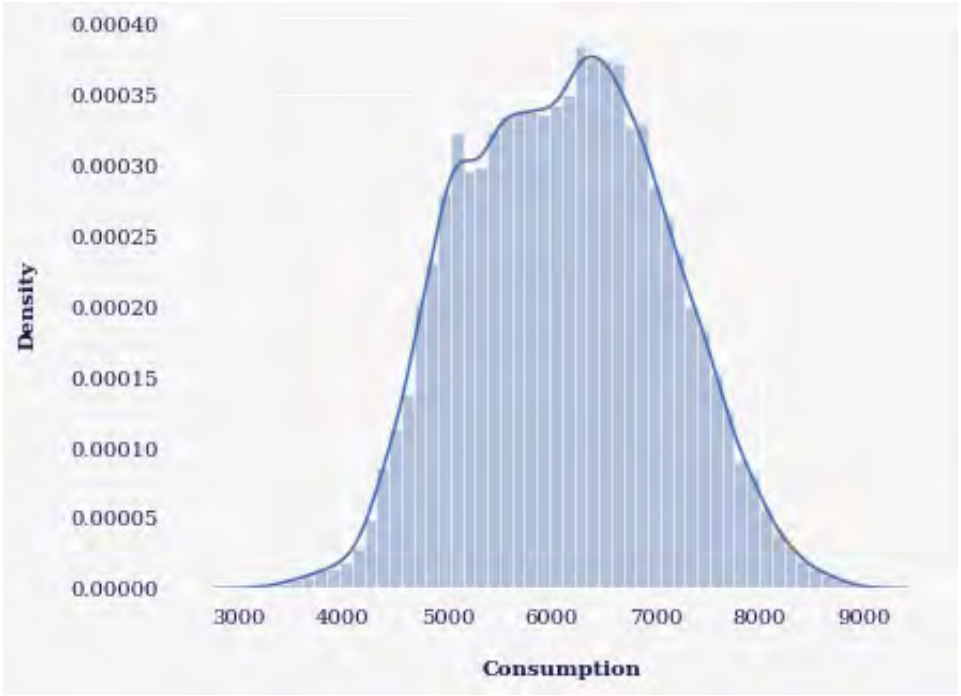
- **Epochs:** 10,000
- **Adversarial Training:** Generator and Discriminator iteratively trained to improve quality of synthetic data.

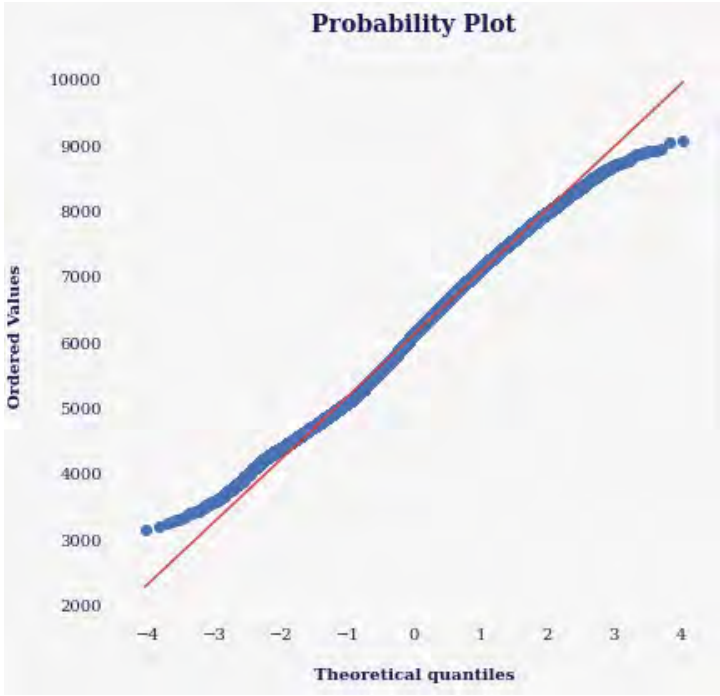
When the synthetic data generated by the GAN model is compared with the real electricity consumption data, it is observed that the synthetic data cannot simulate the real data pattern.

Methodology

Exploratory Data Analysis (EDA)

- The hourly consumption data of Istanbul province used in the analyses and modelling were obtained by scaling with the province-based consumption percentages in Turkey in the monthly reports of EMRA.
- Electricity consumption is usually concentrated between 5000-7000 MWh and extremely low/high consumption is rare.
- The data are approximately normally distributed but deviations are observed especially at the endpoints.

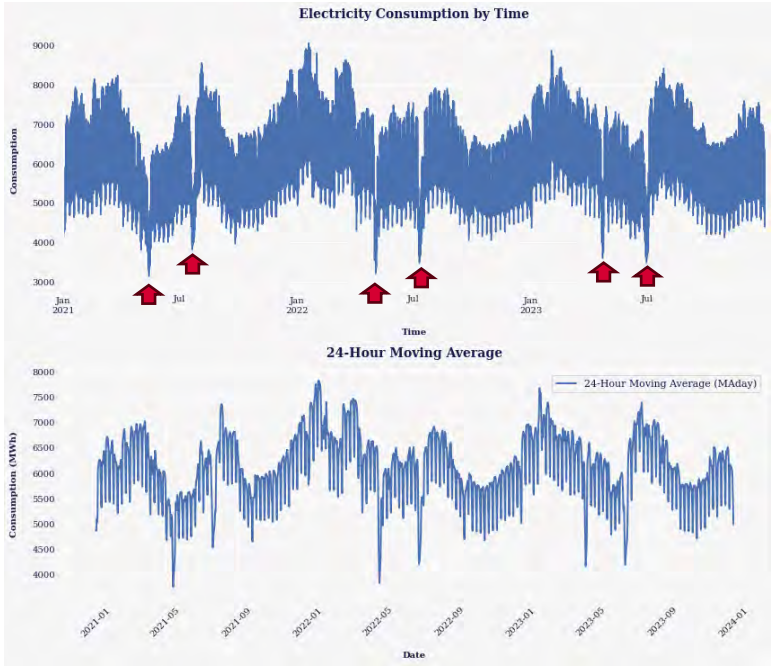


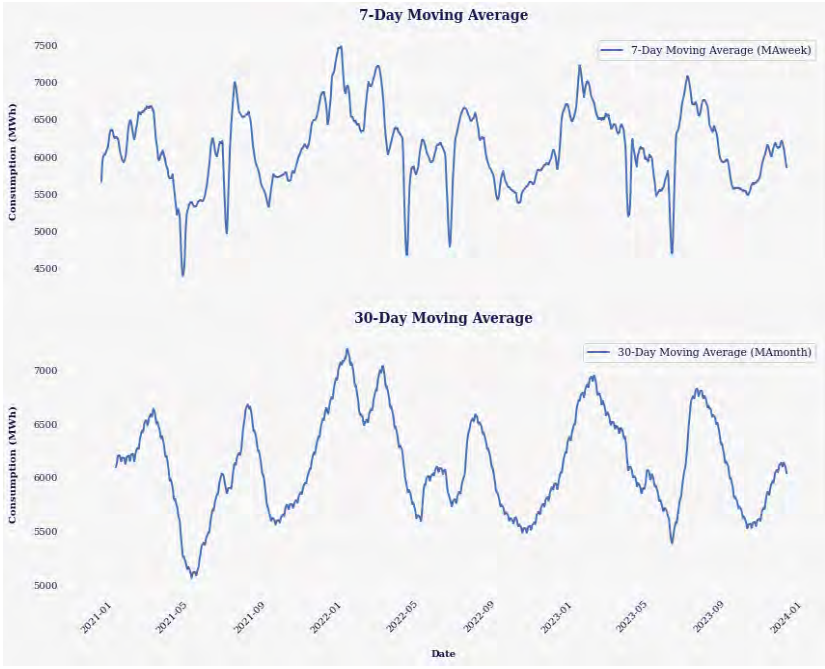


Methodology

Exploratory Data Analysis (EDA)

- When the hourly electricity consumption data of Istanbul between January 2021 and December 2023 are analysed, it is seen that consumption increases in winter months and there are fluctuations in summer months.
- It is seen that significant decreases occur in the data during religious holidays.
- The 30-Day Moving Average graph shows that the increases and decreases in overall energy consumption are seasonally significant

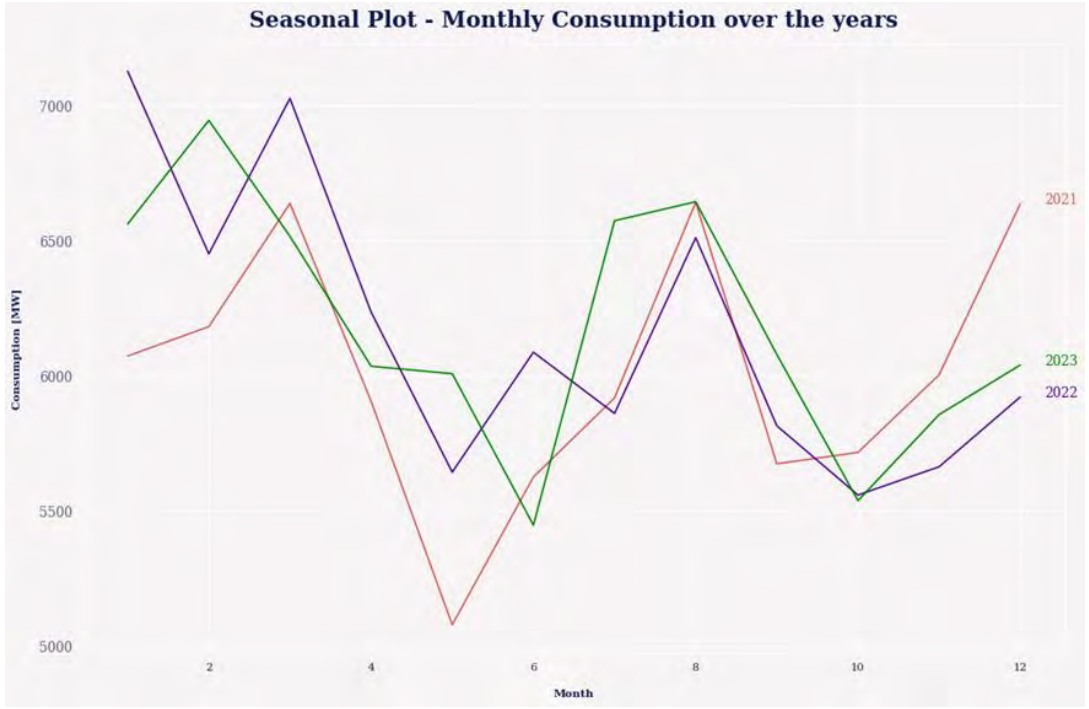




Exploratory Data Analysis (EDA)

Monthly Consumption

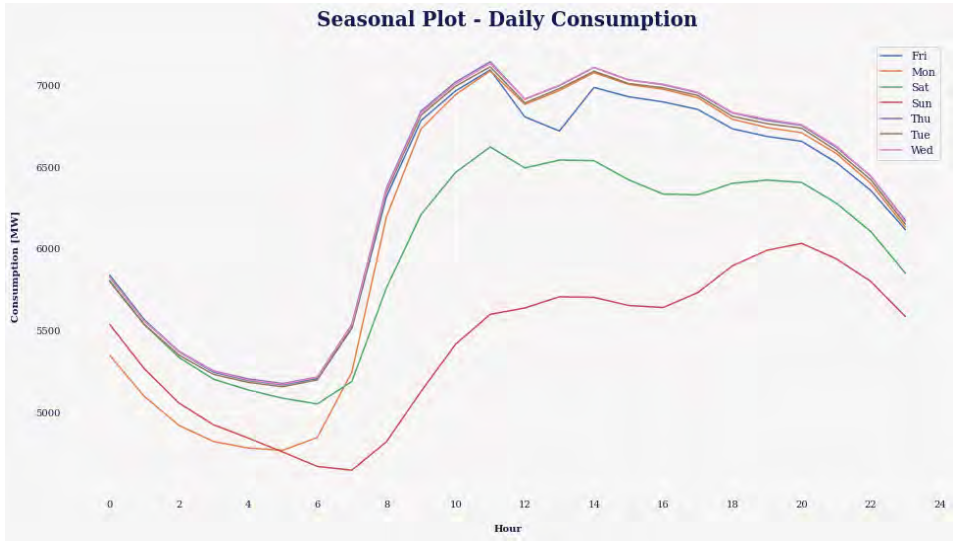
- When the monthly electricity consumption graph is analyzed, it is observed that consumption increases in summer and winter months and decreases in spring months.
- 2021 shows a different trend in terms of consumption, which may be due to the pandemic and the mild winter compared to other years

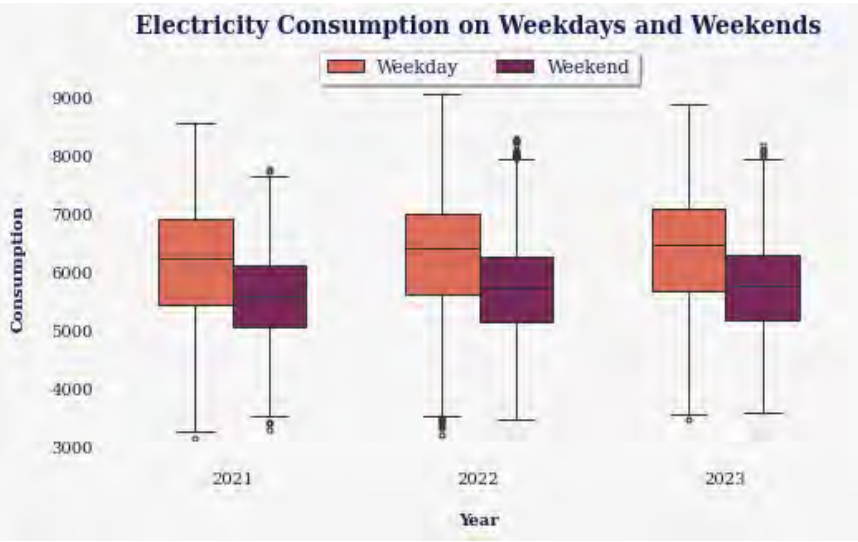


Exploratory Data Analysis (EDA)

Hourly & Daily Consumption

- When the electricity consumption graph is analysed in terms of days of the week there is a clear distinction between weekday and weekend consumption.
- For all days, there is a sharp increase in consumption starting around 5-6 AM, peaking nearly 11 AM.
- There is a gradual decline in consumption starting from late afternoon, around 5-6 PM, continuing through the evening and night.





Methodology

Time Series Analysis &

Variable Selection

- **Stationary:** The stationarity of the independent variables to be used in the model was measured by Augmented Dickey Fuller, Phillips-Perron and KPSS tests and those that failed were eliminated.
- **Correlation Matrix:** The correlation values of the dependent variable and all independent variables are observed and the independent variables to be eliminated are decided with the help of the threshold value in the decision mechanism.
- **Multicollinearity:** The VIF value used to examine whether there is multicollinearity between the explanatory variables in the data set.
- **PACF (Lag Analysis):** The appropriate number of lags to include in autoregressive (AR) models are determined by this analysis.

Selected Factors
Sensed Temperature
Perceived Humidity
Holiday
Weekend
Related Electricity Consumption Hour
Month of the year

Methodology

Data Preprocessing

- Based on the results of data exploration and analysis, a final data sets were created for electricity consumption data and related factors, with the time interval from 00:00 on January 1, 2021 to 23:00 on December 31.
- The dataset is divided into 60% training data, 20% validation data to tune model hypermeters and 20% test data to measure the model performance in chronological order.
- Next, all the time series recorded values are normalized using standart MinMaxScaler.
- Lags of the normalized series up to 24 time steps determined by PACF were added to the feature series.

Methodology

Model Development

Narx RNN Model : The Nonlinear AutoRegressive model with exogenous inputs (NARX) is a powerful modeling technique for time series forecasting.

Narx LSTM: LSTM (Long Short Term Memory) is a specialized type of RNN (Recurrent Neural Network) widely used in time series due to its capacity to learn long-term relationships within consecutive time steps.

Prophet - LSTM Hybrid Model : Combines the strengths of two popular time series forecasting techniques. The idea behind this hybrid approach is to leverage the advantages of both methods to improve the accuracy and robustness of time series forecasts.

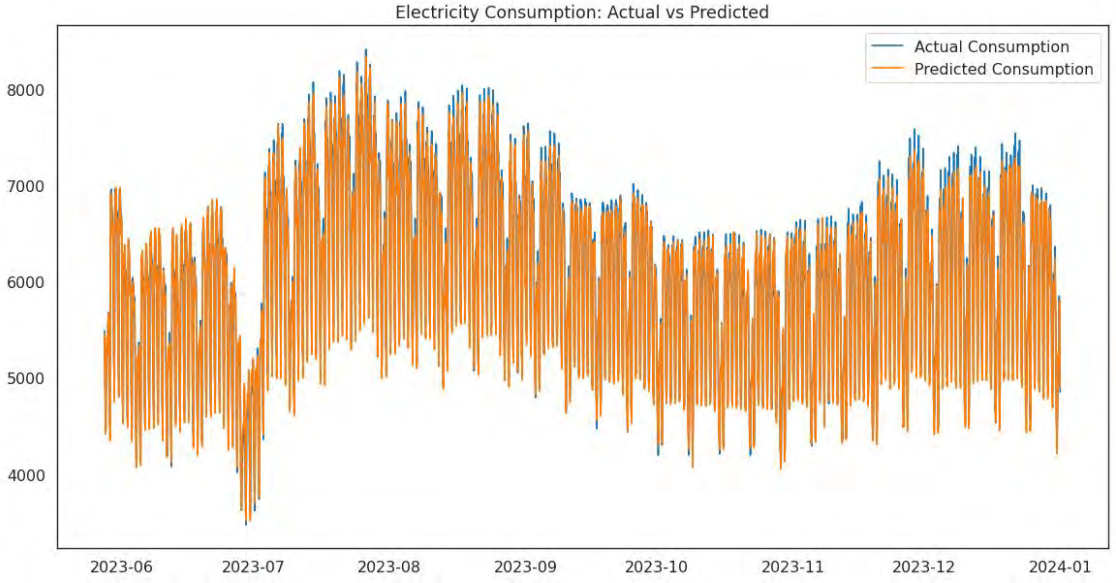
Prophet: Prophet is an open source library for forecasting time series with complex features such as trends, seasonality and holidays.

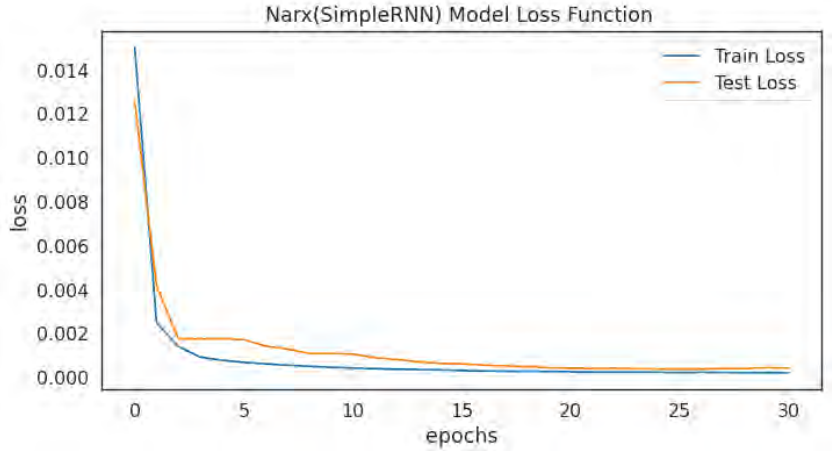
LSTM: Can learn more complex and short-term patterns that Prophet cannot model.

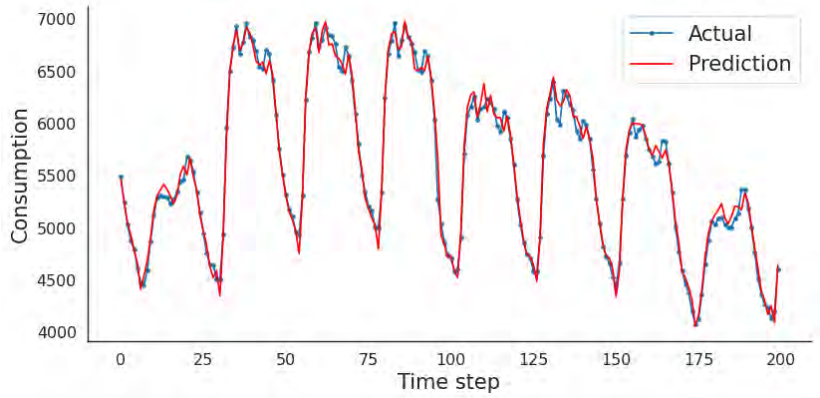
Model Name	Layer (type)	Output Shape	Parameters	Total Parameters
NARX RNN Model	simple_rnn (SimpleRNN)	(None, 50)	2,900	2,951
	dense (Dense)	(None, 1)	51	
LSTM Based NARX Model	lstm (LSTM)	(None, 24, 50)	11,600	31,851
	dropout (Dropout)	(None, 24, 50)	0	
	lstm_1 (LSTM)	(None, 50)	20,200	
	dense (Dense)	(None, 1)	51	
Prophet LSTM Hybrid Model	Prophet Model	(26.256, 32)	0	31,851
	lstm (LSTM)	(None, 24, 50)	11,600	
	lstm_1 (LSTM)	(None, 50)	20,200	
	dense (Dense)	(None, 1)	51	

Model Development

NARX RNN Model



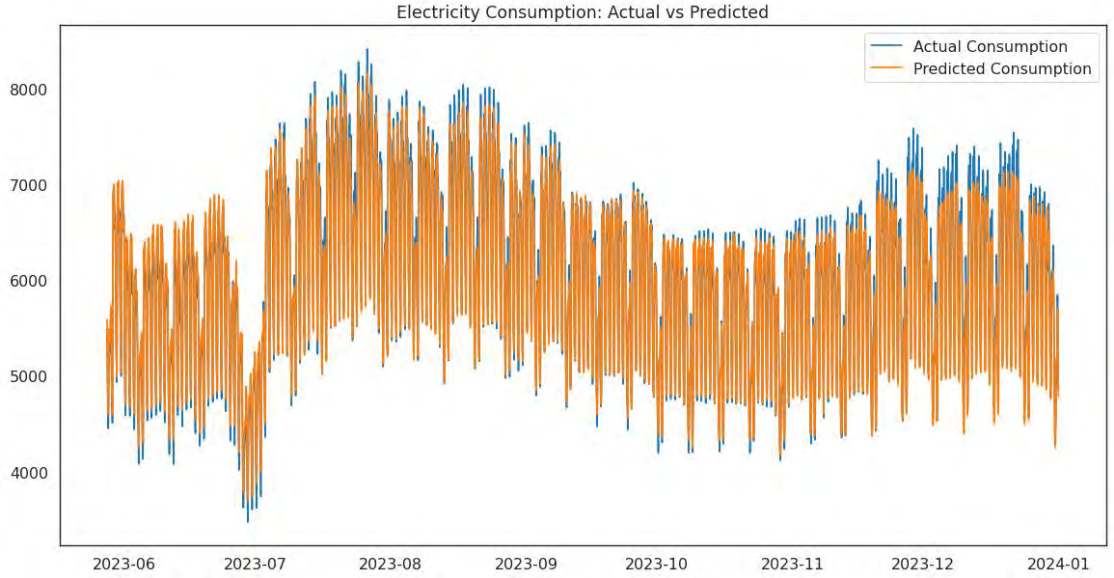


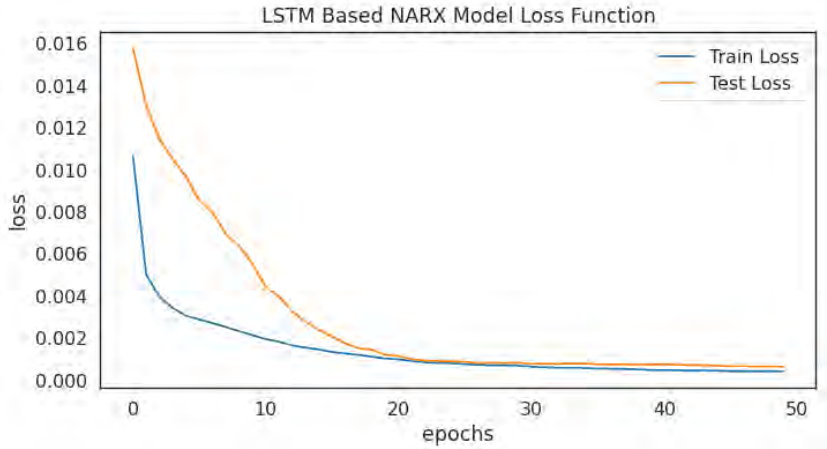


- Lag: 24
- Model: SimpleRNN (50 units, ReLU)

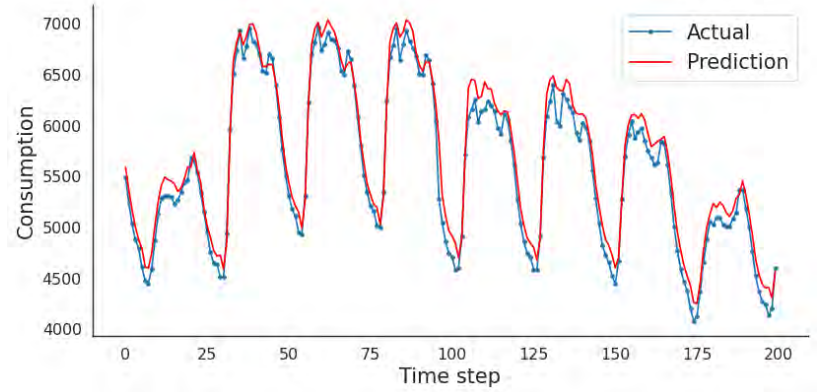
Model Development

LSTM Based NARX Model



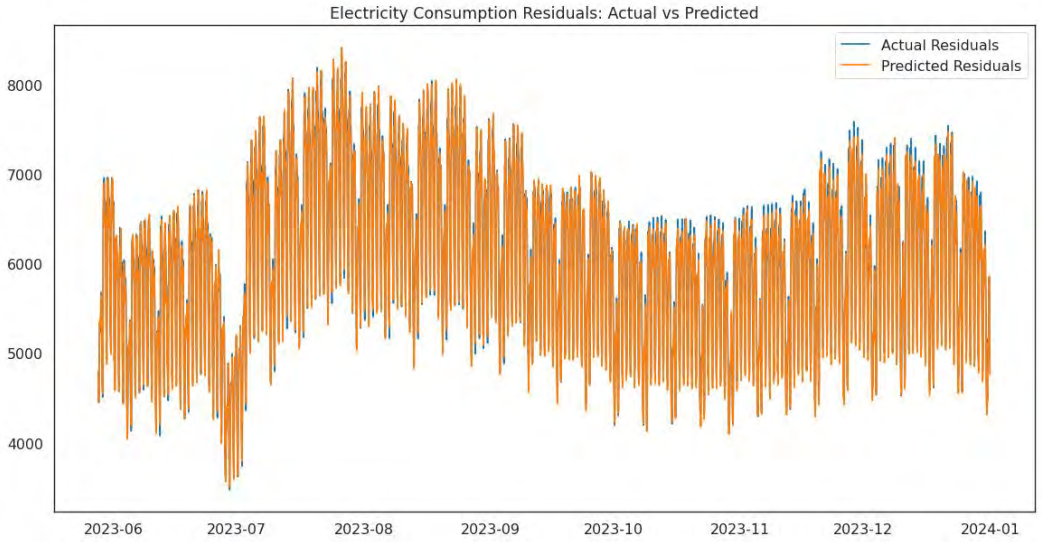


- Lag: 24
- Model: LSTM(50 units)



Model Development

Prophet-LSTM Hybrid Model




```
[21]: predicted_values=pd.DataFrame()

predicted_values[['ds', 'trend', 'daily', 'weekly', 'yearly', 'extra_regressors_additive']] = prophet_predictions[['ds', 'trend', 'daily', 'weekly', 'y

y_pred_flatten = np.full(len(predicted_values), np.nan)

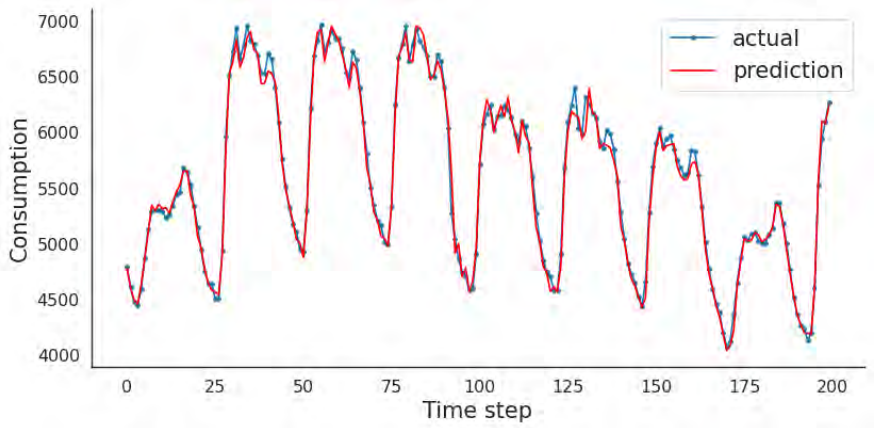
# y_pred değerlerini prophet_predictions ile hizalayarak ekleyin
y_pred_flatten[-len(y_pred):] = y_pred.flatten()

y_pred_flatten[-len(y_pred):] = y_pred.flatten()

y_pred_flatten[:-len(y_pred)] = y_pred_train.flatten()

predicted_values['residual'] = y_pred_flatten

# Sezonluk ve diğer bileşenleri toplayarak y_pred sütununu oluşturma
predicted_values['y_pred'] = (
    predicted_values['trend'] +
    predicted_values['daily'] +
    predicted_values['weekly'] +
    predicted_values['yearly'] +
    predicted_values['extra_regressors_additive'] +
    predicted_values['residual']
)
```



- **Model:** Using Prophet initial predictions generated and these predictions used as input to train an LSTM model. The LSTM model learns from the residuals (the differences between the actual values and the Prophet predictions) and generates refined forecasts.

Model Development

Model Evaluation

- MAE: Measures the mean of error magnitudes, ignores direction, is easy to interpret and has low sensitivity to outliers.
- RMSE: Expresses errors in original data units, gives more weight to large errors and is more sensitive to outliers.

MODELS	PERFORMANCE METRICS			
	Train		Test	
	MAE	RMSE	MAE	RMSE
NARX (SimpleRNN)	86.65	115.68	90.39	116.87
LSTM Based NARX	133.41	168.11	117.48	148.66
Prophet	82.93	111.95	85.88	112.22
Prophet-LSTM Hybrid Model	62.17	80.48	65.58	84.73

When the performances of the models are evaluated with MAE and RMSE metrics over training and test data, it is seen that the models can generalise the data and there is no risk of overfitting in the models, except for the LSTM-based NARX model.

Findings & Feature Work

Key Findings:

- Synthetic datasets could failed to replicate time series data.
- When the performance metrics, loss function graphs are evaluated and the actual predicted consumption values are compared, it is seen that the Prophet-LSTM hybrid model is more successful in the prediction of hourly electricity consumption.

Future Work:

- Optimizing performance metrics
- Evaluation of short-term and long-term forecast performances
- Improving the performance of the model by selecting shorter and more significant periods.
- Updating the model regularly with new data and adapting to changing conditions.

THANK YOU!

Hayriye Yasak Ozkal

hayriye.yasakozkal@tcmb.gov.tr

Big Data Technologies Directorate

Central Bank of the Republic of Turkey

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Forecasting influenza-like illness in Italy using Wikipedia: a principal components regression approach[1]

Gianluca Mura,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Forecasting influenza-like illness in Italy using Wikipedia: a principal components regression approach

Gianluca Mura[1]

Abstract

In recent years, there has been growing academic interest in leveraging Wikipedia data to understand the epidemiology of infectious diseases. Studies on Influenza-Like Illness (ILI) have shown that Wikipedia can serve as a valid surveillance system, addressing gaps in traditional surveillance methods. Influenza, being a highly prevalent and impactful disease, presents an opportunity to explore the potential of using Wikipedia data for accurate and timely predictions. In this paper, I propose a novel approach to predict Influenza-Like Illness rates in Italy by analysing the number of consultations of a selected set of Italian language Wikipedia pages. I employ Principal Components Regression (PCR) as my predictive model, utilizing the reduced-dimensional representations of the Wikipedia data. The performance of the model is evaluated by comparing the predicted rates with the official data published by InfluNet, the Italian epidemiological influenza surveillance system, for four consecutive influenza seasons, spanning from 2015 to 2019. This study contributes to the growing body of literature on utilizing alternative data sources for disease surveillance and highlights the potential of machine learning models for real-time forecasting in public health emergencies.

Keywords: Influenza-Like Illness (ILI), Wikipedia Page Consultations, Machine Learning, Principal Components Regression, Surveillance System.

JEL classification: C38, C53, I18

[1] Bank of Italy (Gianluca.Mura@bancaditalia.it). The views and opinions expressed in this paper are those of the author and do not necessarily reflect the position of Bank of Italy.

1. Introduction

Seasonal influenza remains a significant public health challenge, causing substantial morbidity and mortality worldwide, with an estimated 290,000 to 650,000 respiratory deaths annually (De Toni et al., 2021). Beyond its health implications, influenza leads to economic losses through work absenteeism and decreased productivity, with up to 88% of the economic burden attributable to indirect costs (de Courville et al., 2022). Traditional surveillance systems, such as sentinel physician networks and laboratory confirmations, play a crucial role in monitoring influenza outbreaks but are limited by reporting delays of up to two weeks and are expensive (Simonsen et al., 2016; Gianfredi et al., 2021). In recent years, the emergence of digital epidemiology has introduced novel data streams with potential to supplement traditional surveillance methods by capturing health-seeking behaviours in the population through near real-time signals of disease activity from internet-based data sources. (McClymont et al., 2024). The recent COVID-19 pandemic has further demonstrated the critical importance of timely disease surveillance and forecasting capabilities. While traditional surveillance systems were overwhelmed during the pandemic, digital surveillance methods gained prominence, offering potential complementary tools for tracking disease spread (Santangelo et al., 2022).

Against this background, Wikipedia, as a widely used online resource for health information, represents a particularly valuable data source for tracking influenza trends, since its access logs contain information about public interest in health topics across different languages and geographic regions (Smith, 2020). Moreover, prior research has demonstrated strong correlations between Wikipedia page views and official influenza incidence data, positioning it as a promising resource for timely and accurate disease surveillance and forecasting (McIver & Brownstein, 2014; Generous et al., 2014; Zimmer et al., 2018). However, effective forecasting systems must account for inherent challenges in surveillance data, including data instability, surveillance coverage variations, and media-influenced attention spikes that can skew internet-based signal detection (Chakraborty et al., 2018). To overcome these problems, internet-based surveillance systems should ideally complement rather than replace traditional methods, creating hybrid approaches that leverage the strengths of each data source (Hammond et al., 2022).

In Italy, RespiVirNet, coordinated by the Istituto Superiore di Sanità, is the integrated national surveillance system for influenza-like illnesses and respiratory viruses, combining epidemiological and virological monitoring to enable timely vaccine updates and identify circulating viruses such as influenza. Traditional surveillance systems of this kind, as previously highlighted, while extremely valuable, inherently suffer from delays associated with data collection and reporting. To address this limitation, this study investigates whether Italian Wikipedia access logs can effectively predict influenza-like illness (ILI) rates in Italy.

Specifically, I propose a forecasting model based on Principal Components Regression (PCR) that utilizes page-view data from relevant Italian-language Wikipedia articles, which provides methodological advantages by addressing the multicollinearity inherent in such data while preserving the variance structure critical for predictive accuracy. By employing dimension reduction techniques, PCR enables efficient handling of high-dimensional data without sacrificing explanatory power.

The model's performance is assessed against official data from InfluNet, Italy's epidemiological influenza surveillance system, across four consecutive influenza seasons (2015-2019). Ultimately, this research – while being aware of the limitations inherent in this type of data – aims to contribute to the growing field of digital disease surveillance (Santangelo et al., 2022) by showing the practical utility of Wikipedia data for enhancing real-time monitoring capabilities in public health systems and exploring the potential integration of these alternative data streams into existing surveillance frameworks.

2. Essential review of related literature

Several studies have investigated the possibility of exploiting Wikipedia access log data to monitor and predict rates of influenza syndromes and other infectious diseases. In a work that can be considered pioneering, Generous et al. (2014) conducted a comprehensive evaluation of Wikipedia as a global disease monitoring tool, examining the correlation between article access logs and official disease incidence data across multiple diseases and countries. Their findings revealed that Wikipedia data could successfully estimate influenza-like illness (ILI) levels in certain contexts, while acknowledging limitations in others.

Contemporary and related work is that of McIver and Brownstein (2014), who developed a Poisson model using selected Wikipedia articles to estimate ILI prevalence in the United States. Their approach demonstrated that Wikipedia usage could provide near real-time estimates of ILI activity with accuracy, outperforming Google Flu Trends particularly during abnormal influenza seasons. They identified specific health-related articles (including "Influenza," "Fever," and "Oseltamivir") that consistently correlated with official CDC data.

Hickmann et al. (2015) combined Wikipedia access data with a mathematical epidemiological model and data assimilation methods. Their research, specifically focused on forecasting the 2013-2014 influenza season in the United States, highlighted the value of combining Wikipedia data with traditional epidemiological models (i.e. integration of mechanistic disease modelling with digital surveillance), achieving predictive accuracy for both the timing and magnitude of peak influenza activity during the study period.

Priedhorsky et al. (2017) implemented a systematic article selection process based on semantic relatedness showing how filtering inputs through Wikipedia's category links significantly improved prediction quality. However, they also identified limitations in forecasting beyond the immediate present, finding "very little forecasting value, even just one time step ahead". This study underscores the importance of precise article selection and the limited predictive horizon of purely observational data.

Moreover, few studies have compared Wikipedia's forecasting capabilities with other data sources. Sharpe et al. (2016) conducted a comparative analysis of Google, Twitter, and Wikipedia using Bayesian change point analysis, reporting varying levels of sensitivity and positive predictive value (PPV) across platforms.

De Toni et al. (2021) propose a language-agnostic approach for using Wikipedia to monitor influenza-like illness across multiple European countries. Their method employs two algorithms, Personalized PageRank and CycleRank, to automatically select relevant Wikipedia pages without requiring expert input. Their results showed strong correlations between Wikipedia data and official epidemiological records across the studied countries, demonstrating the potential of Wikipedia as an effective ILI surveillance tool in diverse linguistic contexts.

Finally, with reference to Wikipedia pages in Italian, Gianfredi et al. (2021) conducted a cross-sectional study examining the correlation between Wikipedia page views and official influenza data in Italy between October 2015 and May 2019. Using Pearson correlation coefficients, they identified strong temporal correlations between Wikipedia search trends and official ISS influenza bulletins. These findings indicate that Wikipedia access patterns mirror actual influenza incidence in Italy, suggesting potential value as a complementary real-time surveillance tool that could address the delay in traditional systems.

3. Data and methods

Epidemiological data

As noted earlier, the RespiVirNet (formerly InfluNet) surveillance system monitors influenza-like illnesses (ILI) during the seasonal period through a network of sentinel physicians. The system relies on a network of participating general practitioners (GPs) and paediatricians (PLSs) and is coordinated by the Istituto Superiore di Sanità (ISS). Epidemiological surveillance is conducted via weekly reports of ILI cases from a sample of GPs and PLSs during the period from October to April.

To conduct this analysis, the figures on the reference week and the weekly incidence data, expressed as the number of influenza syndromes (cases) per 1,000 patients (total incidence), were extrapolated. Each influenza season, and thus the relevant sample survey, has an average duration of 28 weeks, beginning on the 42nd week of one year and ending on the 17th of the following year. The decision to limit this analysis to four influenza seasons (equivalent 112 weekly observations) prior to the COVID-19 pandemic (2015-2019) was based on methodological considerations regarding data stability. Including pandemic-era data might have introduced significant complications to the analysis, as information-seeking behaviours likely changed dramatically during this period (McClymont et al., 2024). Research suggests that the pandemic created an unprecedented "information crisis" that altered online engagement patterns. In fact, during early 2020, Wikipedia experienced extraordinary spikes in page creation and viewership related to coronavirus content, potentially distorting typical seasonal patterns of influenza-related information seeking. The pandemic appears to have shifted public attention toward specific topics like epidemiology and public health, creating anomalous conditions that would require specialized analytical approaches (Colavizza, 2020). By focusing on pre-pandemic seasons, this study may provide a more stable baseline for understanding the relationship between Wikipedia access data and influenza trends under more typical

circumstances, though future research could explore how these patterns evolved during the pandemic period.

Influenza seasons considered in the analysis Table 1

Influenza Seasons	From	To	No. of weeks
2015-2016	2015-42	2016-16	28
2016-2017	2016-42	2017-17	28
2017-2018	2017-42	2018-17	28
2018-2019	2018-42	2018-17	28

[1] Note: the year 2015 ended with the 53rd week, the others with the 52nd

Source: author's elaboration

Wikipedia data

Wikipedia provides various tools and APIs to analyze its data, including access logs, page content, hyperlinks, and editor activity. There are also numerous free third-party solutions and open-source projects have been developed by the community, enhancing the functionality and accessibility of Wikipedia data. I utilized data from Meta-Wiki, specifically from the Pageviews Analysis tool, which provides detailed statistics on Wikipedia page access. This tool aggregates user visits to Wikipedia articles, allowing for the analysis of search trends over time: information on daily accesses as of 1 July 2015 was identified by selecting all types of platforms (desktop, mobile and mobile application) and discarding accesses generated by crawlers and other automated programs). The number of accesses was then aggregated by ISO week and purged of irrelevant observations because they were not included in the annual flu season survey. Data preprocessing included the identification and treatment of extreme values in the Wikipedia page view time series. Examination of the dataset revealed few abnormal values that fell substantially outside expected ranges for typical page view patterns. Specifically, four outliers were identified and treated: two adjacent values for the 'Epidemia' page during the first influenza season (2015-2016), and two individual values for the 'Influenza suina' page occurring during the third (2017-2018) and fourth (2018-2019) influenza seasons respectively. To maintain the integrity of the time series while minimizing distortion, these outliers were replaced with imputed values calculated as the arithmetic mean of the immediately preceding and following observations. This approach preserved the underlying temporal trends while preventing extreme values from disproportionately influencing the principal component structure and subsequent regression analysis.

Regarding the selection of the most relevant Wikipedia articles, several approaches are viable, which involve analyzing their relationships based on hyperlinks, categories, content similarity, or a combination of these factors. For this study, the selected pages were manually picked based on thematic relevance, with publications and prior studies on influence rate prediction reviewed to guide the selection of topics most likely to contribute to accurate modeling. After this selection, distance metrics – in terms of shortest path distance between “Influenza” article and

the other articles – were applied as methods of verification and confirmation, ensuring the dataset represented a cohesive and meaningful subset of Wikipedia content. At the end of this process, 14 articles were selected. Finally, as previous studies have shown (e.g. Generous et al., 2014), since geo-localized data are not freely available, the language of Wikipedia articles was used as a proxy to delimit the target country. Given the limited diffusion of the Italian language abroad, it is the author's opinion that this does not have any relevant consequences for the purposes of this study.

Selected Wikipedia articles — Table 2

Article in Italian	Corresponding article in English	Total number of views over the study period (4 flu seasons)	No. of weeks
Influenza	Influenza	227.775	-
Febbre	Fever	446.206	1
Cefalea	Headache	230.886	1
Influenza aviaria	Avian influenza	91.820	1
Influenza suina	Swine influenza	49.771	1
Epidemia	Epidemic	62.671	2
Pandemia	Pandemic	96.235	2
Pandemia influenzale	Influenza pandemic	21.314	1
Raffreddore comune	Common cold	225.309	1
Paracetamolo	Paracetamol	806.148	1
Raffreddore	Cold	12.419	2
Virus parainfluenzali umani	Human parainfluenza viruses	11.355	1
Antivirale	Antiviral drug	45.896	1
Influenza virus A sottotipo H1N1	Influenza A virus subtype H1N1	93.237	1

Note: Shortest Path Distance is defined as the minimum number of hyperlink clicks (edges) required to navigate from one Wikipedia page (node) to another through a directed network of internal links

Source: author's elaboration based on pageviews.toolforge.org/

4. Methodology

Overview

This methodology combines dimension reduction via Principal Component Analysis (PCA) with linear regression through a multi-step process. Before briefly mentioning the algebraic formulation of PCR, it is useful to recall the pros and cons of PCA with reference to the objective of this study. When the goal is timely influenza surveillance, this technique offers several complementary advantages.

First, it enhances signal extraction: weekly access counts inevitably include some noise, i.e. spikes in page views may mirror media attention or automated activity rather than genuine changes in health-related concern. Because PCA performs a mathematical transformation of the original variables into new orthogonal (uncorrelated) variables that maximize explained variance, it effectively consolidates what multiple pages have in common while eliminating redundancies. As a result, a small set of components can dampen such random fluctuations while preserving the persistent temporal pattern that tends to accompany an epidemic wave. Second, PCA contributes to statistical stability: the raw page-view series suffer from some degree of collinearity, a circumstance that inflates variance and undermines predictive precision. Regressing influenza-like-illness (ILI) rates on the uncorrelated component scores alleviates this difficulty and produces more reliable coefficients even when the number of original predictors exceeds the number of weekly observations. Third, given that InfluNet provides only about thirty weekly data points per season, using just the leading principal components rather than individual Wikipedia pages helps prevent over-fitting to limited data.

This approach promotes statistical parsimony while still incorporating information from all Wikipedia pages in the initial analysis before dimension reduction. This lean structure mirrors the "sloppiness" seen in complex biological and physical models: just a few "stiff" parameter combinations drive the predictions, while all other directions have negligible influence (Transtrum et al., 2015). Acknowledging this fact supports the choice to analyse principal-component signals instead of individual page-view coefficients.

On the other hand, these strengths, coexist with limitations that deserve explicit acknowledgment. PCA orders components strictly by the variance they explain (descending order), yet components accounting for only a modest fraction of total variance can still hold valuable predictive information, excluding them solely on variance grounds can therefore hinder forecast accuracy (Hadi et al., 2012). Moreover, a single component may blend traffic to biomedical articles with visits triggered by unrelated popular events, complicating substantive interpretation. Such peculiarities do not negate the method's effectiveness, but they remind that PCA is an exploratory method and therefore translating patterns of online behaviour into actionable public-health signals still requires careful judgement.

Principal component regression

Principal Component Regression provides an effective approach for addressing multicollinearity Wikipedia page view data while preserving the variance structure essential for accurate influenza prediction. In this section, I provide a concise overview of the key mathematical steps involved in performing principal component regression. For consistency and clarity, all mathematical notations follow standard linear algebra conventions, though alternative representations exist.

To model influenza incidence rates, as a first step, the covariates – daily view counts for 14 influenza-related Wikipedia pages – were standardized to zero mean and unit variance employing $\boldsymbol{Z} = (\boldsymbol{X} - \mu)\boldsymbol{D}^{-1/2}$, where $\mu \in R^{14}$ is the vector containing the means and **D** is the diagonal matrix of variances. This results in a

standardized matrix $\boldsymbol{Z} \in R^{112\,x\,14}$ where each element z_{ij} represents the standardized page views for Wikipedia article j during week i. It should be noted that, while standardization is not an absolute requirement for PCA, I standardized all variables prior to my analysis to ensure that pages with higher overall consultation volumes would not disproportionately influence the component structure, thereby allowing each Wikipedia page an equal opportunity to contribute to the extracted patterns.

In the second step, it is performed the Principal Component Analysis on $\boldsymbol{Z}$, which can be efficiently computed using Singular Value Decomposition (SVD). To this end, let
$\boldsymbol{\Sigma} \in R^{14\,x\,14}$ be the empirical covariance matrix of the standardised predictors that can be computed as:

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\boldsymbol{Z}^T\boldsymbol{Z}$$

Diagonalising the covariance matrix produces the spectral (eigen) decomposition, a fundamental transformation that separates the variance structure into its principal directions:

$$\boldsymbol{\Sigma} = \boldsymbol{V} \wedge \boldsymbol{V}^T$$

where $\boldsymbol{V}$ is the orthogonal matrix of eigenvectors (i.e. the loading directions), $\boldsymbol{V}^T$ is its transpose, and $\wedge$ is a diagonal matrix containing the singular values (square roots of eigenvalues). From this decomposition, is it possible to obtain the principal component scores:

$$\mathbf{W} = \mathbf{ZV}$$

where $\boldsymbol{W} \in R^{112\,x\,14}$ is the matrix of principal component scores. Each column $w_k$of $\boldsymbol{W}$ represents the $w - th$ principal component, with variance λ_k.

The final step involves regress the weekly influenza-like illness rates $y \in R^{112}$ on a subset of *m* principal components:

$$y = \boldsymbol{W}_m\,\gamma_m + \varepsilon$$

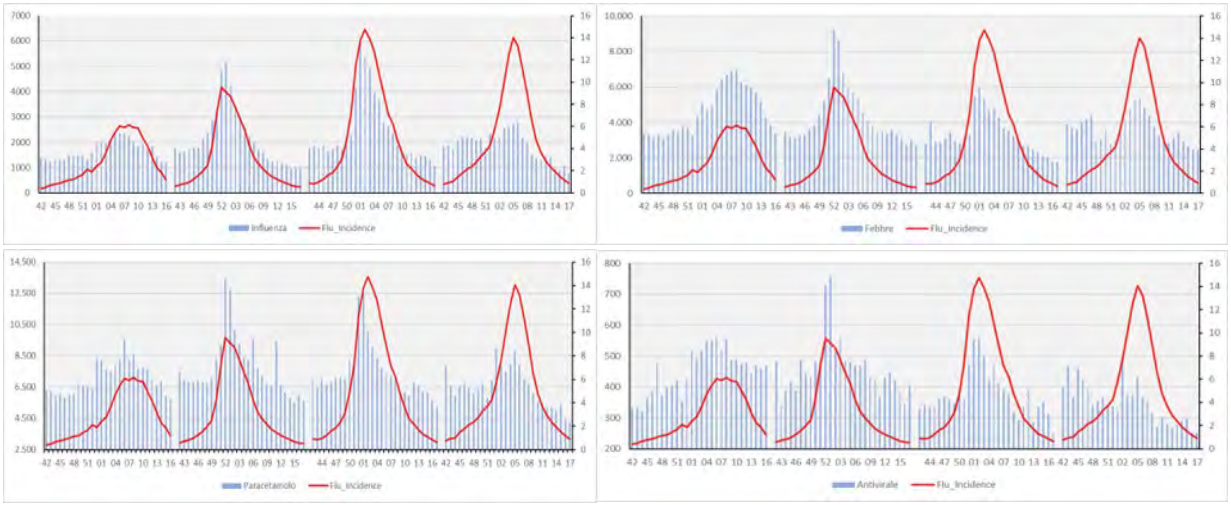
where $\boldsymbol{W}_m \in R^{112\,x\,m}$ contains the first m columns of $\boldsymbol{W}$ (representing the principal components that explain the most variance), $\gamma_m \in R^m$ is the vector of regression coefficients, and $\varepsilon \in R^{112}$ represents the error term.

5. Results and discussion

Exploratory data analysis

The exploration of temporal patterns in both Wikipedia page views and official ILI rates can reveal some insights into their relationship. Descriptive statistics indicated substantial variability in both datasets, with Wikipedia page views demonstrating cyclical patterns that often corresponded to official influenza seasons. Graph 1 illustrate the time-series comparison between access frequencies for four selected Wikipedia pages ("Influenza," "Febbre," "Paracetamolo," and "Antivirale") and official ILI rates over the four observed influenza seasons (2015-2019).

Access logs for selected Wikipedia pages plotted alongside official ILI rates Graph 1



Sources: ILI rates: respivirnet.iss.it/; Wikipedia: pageviews.toolforge.org/

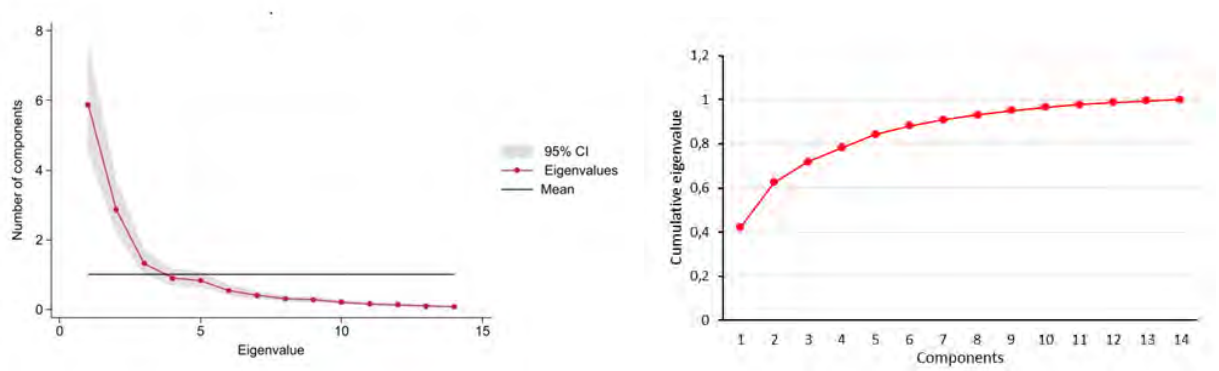
Visual inspection of these time-series plots suggested some degree of temporal alignment between Wikipedia page views and official ILI rates. Increases in page views frequently preceded or coincided with rises in reported influenza cases, indicating possible predictive utility. The "Influenza" page demonstrated the strongest correlation (r = 0.81), followed by "Paracetamolo" (r = 0.66), while "Febbre" and "Antivirale" showed moderate correlations (r = 0.59 and r = 0.39, respectively). These varying correlation strengths, with *p-values* confirming statistical significance, might suggest different information-seeking behaviours associated with disease progression, from initial symptom investigation to treatment options.

The comparative analysis across the four influenza seasons revealed significant epidemiological heterogeneity that presents notable forecasting challenges. The 2017-2018 season exhibited the most severe epidemic profile with an earlier start and substantially higher peak (14.74 cases per 1,000 patients) compared to other seasons. The 2018-2019 season showed a delayed peak that occurred in February rather than January. Meanwhile, the 2016-2017 season had a relatively moderate intensity but demonstrated the sharpest decline post-peak, while the 2015-2016 season exhibited a more prolonged epidemic period with multiple smaller peaks. Additionally, examining the magnitude of page view fluctuations suggested interesting patterns: the "Paracetamolo" page consistently received the highest absolute number of consultations, potentially reflecting its relevance beyond influenza cases, while the "Antivirale" page showed more pronounced relative increases during peak periods despite lower baseline consultation rates. These observations suggest that different Wikipedia pages may provide complementary signals that together enhance forecasting capabilities.

Principal component analysis results

After identifying the correlation patterns identified in the exploratory analysis, it was then proceeded with PCA. Graph 2 presents the scree plot of eigenvalues after PCA, which exhibits a characteristic elbow shape with a steep decline for the first three components, followed by a gradual levelling off thereafter. The first component demonstrated a substantially higher eigenvalue (approximately 5.9), indicating its dominance in explaining dataset variance.

Scree plot of eigenvalues (left) and explained cumulative variance (right) Graph 2



Source: author's elaboration

Several established criteria exist for determining the optimal number of principal components to retain (Mooi et al., 2018): the Kaiser criterion, which suggests retaining components with eigenvalues greater than 1; the scree plot method, which identifies the "elbow point" in the plot of eigenvalues where additional components contribute minimally to explained variance; parallel analysis, which compares observed eigenvalues to those from randomly generated data; the cumulative percentage of variance explained, typically targeting 70-90%; and information criteria such as AIC and BIC, which balance model fit with complexity. Each approach offers unique advantages, often leading researchers to apply multiple criteria for increased confidence in their selection.

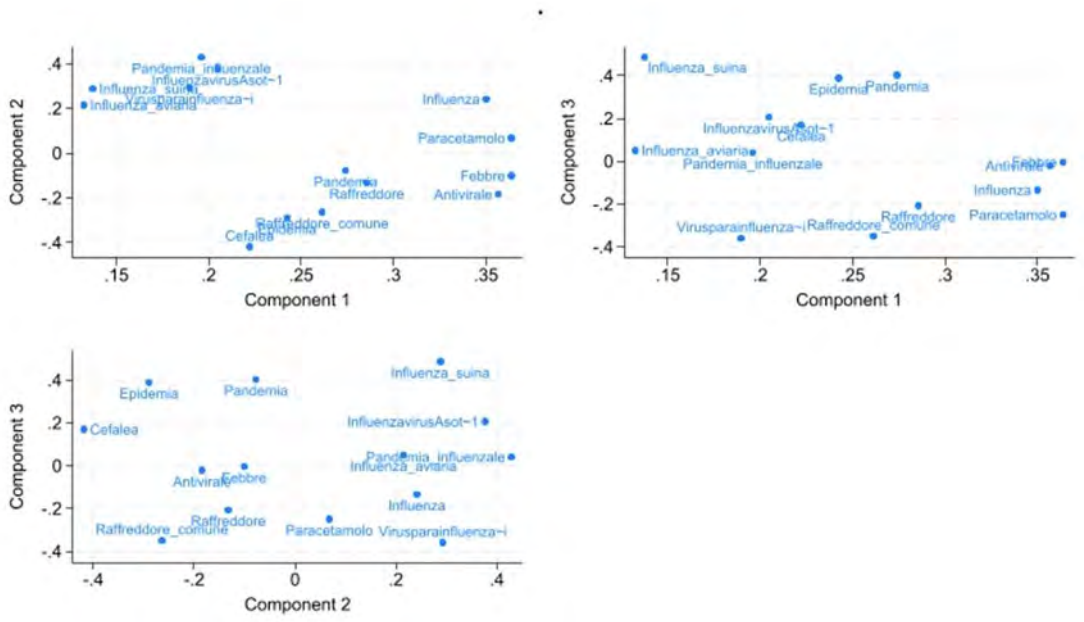
For this analysis, I primarily employed the Kaiser criterion, which suggested retaining three components, as the eigenvalues for the first three components were 5.9, 2.9, and 1.3, respectively, while the fourth component's eigenvalue fell below the threshold at 0.9. The cumulative percentage of variance explained supported this decision, with the first three components collectively accounting for about 72% of the total variance, a proportion that aligns with statistical best practices. The three-component solution thus balanced dimensionality reduction with information preservation, avoiding both under-specification and over-parameterization.

To better understand the underlying structure of the principal components, I utilized component visualization diagrams to examine the relationships between components. Graph 4 presents a three-panel graph displaying the pairwise relationships between the first three principal components, allowing for visual inspection of how the original Wikipedia variables contribute to each dimension.

These component projection graphs provide an intuitive representation of the multidimensional relationships, with variables positioned according to their correlation with each component pair. Distinct clustering patterns among the Wikipedia pages are showed, suggesting meaningful latent dimensions in users' information-seeking behaviour. Pages related to general influenza symptoms (such as "Influenza" and "Febbre") appear grouped differently from pandemic-specific pages (such as "Influenza virus A sottotipo H1N1" and "Pandemia influenzale"), visually confirming the differentiation of components by thematic content. This graphical representation supports a more detailed quantitative analysis of the component structure.

Component loadings

Graph 3



Source: author's elaboration

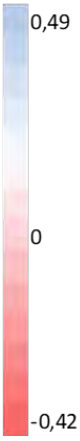
For this purpose, it is useful to check the exact measure of component loadings, which reveal distinct patterns of variable contributions to each principal component. The first principal component (PC1), explaining 42.1% of variance, exhibited strong positive loadings from "Influenza" (0.35), "Febbre" (0.36), "Paracetamolo" (0.36), and "Antivirale" (0.36), suggesting it represents general information-seeking about influenza symptoms and common treatments. This component captures the core influenza-related search behaviour that aligns with typical symptom investigation and treatment options. The second principal component (PC2), accounting for 20.0% of variance, showed distinct loading patterns with strong positive contributions from "Pandemia influenzale" (0.43), "Influenza virus A sottotipo H1N1" (0.38), and "Influenza_suina" (0.29), while displaying negative loadings for "Cefalea" (-0.42) and "Raffreddore comune" (-0.26). This component appears to differentiate between searches related to severe or pandemic influenza strains versus those focused on more common, less severe symptoms. It potentially reflects a dimension of concern about serious influenza outcomes as opposed to routine seasonal symptoms. The

third principal component (PC3), explaining 10.0% of variance, demonstrated a more complex pattern with high positive loadings for "Influenza suina" (0.49), "Pandemia" (0.40), and "Epidemia" (0.39), contrasted with negative loadings for "Virus parainfluenzali umani" (-0.36), "Raffreddore comune" (-0.35), and "Paracetamolo" (-0.25). This component might represent a distinction between epidemic/pandemic-related information-seeking and searches for differential diagnosis or treatment of common respiratory conditions. It potentially captures public concern about wider disease spread versus management of individual symptoms.

Loadings of Wikipedia pages along the first three components

Graph 4

	PC1	PC2	PC3
Influenza	0,35	0,24	-0,13
Febbre	0,36	-0,10	0,00
Cefalea	0,22	-0,42	0,17
Influenza_aviaria	0,13	0,21	0,05
Influenza_suina	0,14	0,29	0,49
Epidemia	0,24	-0,29	0,39
Pandemia	0,27	-0,08	0,40
Pandemia_influenzale	0,20	0,43	0,04
Raffreddo_comune	0,26	-0,26	-0,35
Paracetamolo	0,36	0,07	-0,25
Raffreddore	0,29	-0,13	-0,21
Virus_parainfluenzali_umani	0,19	0,29	-0,36
Antivirale	0,36	-0,18	-0,02
Influenza_virus_A_sottotipo_H1N1	0,20	0,38	0,21



Source: author's elaboration

The adequacy of the dataset for PCA was confirmed through the Kaiser-Meyer-Olkin test, which returned an overall measure of 0.8308, exceeding the 0.80 threshold that indicates good sampling adequacy. Individual KMO values for variables ranged from 0.7040 ("Pandemia influenzale") to 0.9196 ("Febbre"), further validating the appropriateness of the PCA approach. This also suggests strong intercorrelations among variables that are effectively addressed through the dimensionality reduction provided by PCA.

These findings indicate that the principal components extracted represent meaningful latent dimensions of information-seeking behaviour related to influenza, with a good degree of interpretability in terms of the original Wikipedia variables. The meaningfulness of these dimensions is evidenced by their distinct loading patterns, which align with theoretically coherent aspects of health information-seeking: general symptom and treatment information (PC1), pandemic-specific concerns (PC2), and epidemic spread versus routine illness management (PC3). The KMO measure of sampling adequacy further validates that these components capture genuine underlying structures rather than artifacts of random variation (Mooi et al., 2018).

Principal components regression forecasting results

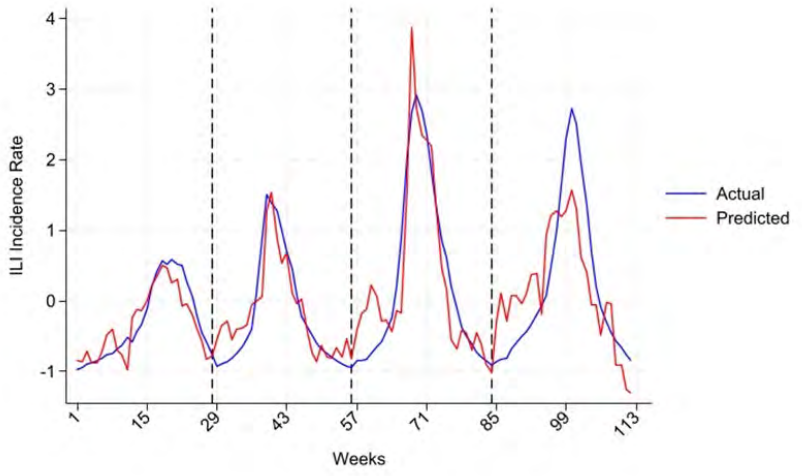
The retention of three components provided a parsimonious yet informative basis for the subsequent regression analysis, achieving an optimal balance between dimensionality reduction and information preservation. As mentioned above, this parsimony is valuable for predictive modelling because it reduces potential multicollinearity among the original variables while retaining a high percentage of the original variance, enough to capture the essential patterns in Wikipedia page views related to influenza trends without over-fitting to noise.

The graph illustrates the relationship between observed ILI rates and those predicted by the PCR over (non-consecutive) 112 weeks. The model demonstrates reasonable accuracy in capturing the cyclical nature of influenza activity, successfully identifying most seasonal peaks and troughs, particularly during the first two seasonal cycles.

The observed discrepancies between actual and predicted values highlight well known limitations inherent to Wikipedia access data as a surveillance source. In fact, while Wikipedia article views exhibit lower signal-to-noise ratios than general internet searches, they remain susceptible to external influences unrelated to disease prevalence. The model's overestimation around week 71, for instance, may reflect increased public interest driven by media coverage rather than actual disease activity. Conversely, the significant underestimation for the last period suggests that during unusually severe outbreaks, information-seeking behaviour on Wikipedia may not scale proportionally with disease prevalence.

These variations likely stem from complex factors affecting online information-seeking behaviour, including media attention that drives article views independently of illness rates, changes in public awareness or concern that don't directly correlate with epidemiological patterns, background seasonal variations in internet usage patterns, external events that may trigger interest in health-related content.

Actual versus predicted Flu incidence rates — Graph 5



Note: standardized values.

Source: author's elaboration

Despite these limitations, as shown in Table 5.1, the model demonstrates good explanatory power, with an R-squared value of 0.7723. While this level of explanatory power appears substantial for social science data and epidemiological modelling, it is important to interpret these results with appropriate caution, recognizing the inherent limitations of statistical models in capturing complex disease dynamics. The adjusted R-squared of 0.7660 is only marginally lower than the unadjusted value, which may indicate that the penalty for including three regressors is minimal. Statistical significance of the overall regression is excellent, confirming that the principal components collectively provide substantial explanatory power to the model; nevertheless, statistical significance alone does not guarantee practical significance or real-world applicability, and these results should be considered preliminary until further validation across additional influenza seasons and geographical contexts.

Overall goodness-of-fit of the regression

Table 3

Influenza Seasons	Value	Degrees of freedom/Units
R-squared (R^2)	0.7723	0.7723
Adjusted R^2	0.7660	0.7660
F-statistic	122.11	122.11
Prob > F (overall p-value)	< 0.0001	< 0.0001

Sources: TableNote

The prediction-error metrics provide support for the model's performance. With an RMSE of 0.4750 and an MAE of 0.3668 (both in standard deviation units), the model demonstrates reasonably strong predictive accuracy, with typical errors of approximately half a standard deviation. While encouraging, these error levels may still be too large for certain public health applications, particularly when precise estimates are required for resource allocation decisions. The RMSE represents about 12% of the total observed range of the standardized dependent variable, which, though relatively modest, still indicates a non-negligible level of uncertainty in the predictions.

Model performance

Table 4

Influenza Seasons	Value	Scale/Notes
RMSE	0.4750	Standard-deviation units (variables were z-scored)
MAE	0.3668	Sd units
RMSE / Range	12%	% of (max – min) of outcome

Source: author's elaboration

Ultimately, the model's ability to maintain this level of accuracy across different influenza seasons varies and would benefit from further evaluation across a broader range of epidemic scenarios. Despite its promise, the Wikipedia-based approach may still be susceptible to biases related to internet access disparities, changes in information-seeking behaviour over time, and platform-specific trends that could

affect long-term reliability. When compared to other digital surveillance approaches, this Wikipedia-based PCR model shows potential advantages, but its relative effectiveness requires critical assessment considering systematic comparisons. Prior comparative evaluation of digital surveillance tools has been made (Sharpe et al., 2016). However, the rapid evolution of artificial intelligence capabilities and epidemic intelligence systems creates a pressing need for updated comparative assessments that account for technological advancements (MacIntyre et al., 2023). Claims of superiority over search query-based systems should therefore be considered tentative until more recent comparative analyses are conducted using identical evaluation criteria and time periods. In any case, models that draw on large-scale Internet data offer a promising complement to conventional surveillance, yet they should be treated as supplementary decision-support tools rather than outright replacements for established epidemiological methods (McIver et al, 2014)

6. Conclusions

This study provides evidence that Wikipedia access logs offer valuable data for enhancing influenza surveillance systems in Italy. The findings demonstrate that using Principal Components Regression on Italian-language Wikipedia page views produces reasonably accurate predictions of influenza-like illness rates when compared with official InfluNet data. The model's R-squared value of 0.7723 indicates substantial explanatory power, consistent with previous research showing strong correlations between Wikipedia data and official epidemiological records (De Toni et al., 2021; Gianfredi et al., 2021).

The methodological approach using PCR proved effective in addressing multicollinearity concerns while preserving the variance structure essential for prediction. By extracting three principal components that collectively accounted for approximately 72% of the total variance, the model achieved a balance between dimensionality reduction and information preservation. The loading patterns of these components revealed meaningful dimensions in user information-seeking behaviour, consistent with theoretical expectations about how individuals search for health information during disease outbreaks.

Nevertheless, this approach faces some important limitations. As demonstrated by the prediction errors and visual analysis of time series data, Wikipedia-based surveillance remains susceptible to external influences unrelated to disease prevalence, including media-driven attention spikes that may skew predictions. Additionally, the model showed variable performance across different epidemic scenarios, suggesting that its reliability may depend on the specific characteristics of each influenza season.

These findings align with the broader literature suggesting that internet-based surveillance systems should complement rather than replace traditional methods (Hammond et al., 2022). The Wikipedia-based approach offers advantages in terms of near real-time availability and cost-effectiveness, potentially addressing the reporting delays inherent in traditional surveillance systems. However, as previous research has emphasized, the combination of traditional and digital surveillance

sources appears to provide the most robust framework for disease monitoring (McClymont et al., 2024).

Future research should explore more sophisticated techniques for filtering noise from Wikipedia signals, integration with other digital data streams, and extension to additional infectious diseases beyond influenza. Longer time series spanning more influenza seasons would also enhance model validation and improve understanding of how Wikipedia-based surveillance performs across varied epidemic profiles. Overall, this study contributes to the growing field of digital disease surveillance by demonstrating the practical utility of Wikipedia data for enhancing real-time monitoring capabilities while acknowledging the continued importance of established epidemiological methods.

Acknowledgements

I thank Dr. Stefano Andreozzi for valuable suggestions and comments.

References

Chakraborty P, Lewis B, Eubank S, Brownstein JS, Marathe M, Ramakrishnan N. (2018) What to know before forecasting the flu. PLoS Comput Biol 14(10): e1005964. https://doi.org/10.1371/journal.pcbi.1005964.

Colavizza, G. (2020). COVID-19 research in Wikipedia. Quantitative Science Studies, 1(4), 1349–1380. https://doi.org/10.1162/qss_a_00080.

de Courville C, Cadarette SM, Wissinger E, Alvarez FP. The economic burden of influenza among adults aged 18 to 64: A systematic literature review. Influenza Other Respi Viruses. 2022;16(3):376-385. doi:10.1111/irv.12963.

De Toni G, Consonni C, Montresor A (2021). A general method for estimating the prevalence of influenza-like-symptoms with Wikipedia data. PLoS ONE 16(8): e0256858. https://doi.org/10.1371/journal.pone.0256858.

Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R (2014) Global Disease Monitoring and Forecasting with Wikipedia. PLoS Comput Biol 10(11): e1003892. doi: 10.1371/journal.pcbi.1003892.

Gianfredi, V., Santangelo, O. E., & Provenzano, S. (2021). Correlation between flu and Wikipedia's pages visualization. *Acta Bio Medica: Atenei Parmensis*, *92*(1), e2021056.

Hadi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, *52*(1), 15-19.

Hammond, A., Kim, J. J., Sadler, H., & Vandemaele, K. (2022). Influenza surveillance systems using traditional and alternative sources of data: a scoping review. *Influenza and Other Respiratory Viruses*, *16*(6), 965-974.

Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. (2015). Forecasting the 2013–2014 Influenza Season Using Wikipedia. PLoS Comput Biol 11(5): e1004239. doi: 10.1371/journal.pcbi.1004239.

Istituto Superiore di Sanità. (2025). RespiVirNet: National epidemiological and virological surveillance system for influenza and other respiratory viruses (https://www.iss.it/en/respivirnet).

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

McClymont, H., Lambert, S.B., Barr, I. *et al.* Internet-based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic. *J Epidemiol Glob Health* 14, 645–657 (2024). https://doi.org/10.1007/s44197-024-00272-y

MacIntyre, C. R., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., ... & Gurdasani, D. (2023). Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, *51*(3), 03000605231159335.

McIver DJ, Brownstein JS (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol 10(4): e1003581. doi:10.1371/journal.pcbi.1003581.

Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). *Market research*. Springer Nature Singapore Pte Ltd.

O'Leary, D. E., & Storey, V. C. (2020). A Google–Wikipedia–Twitter model as a leading indicator of the numbers of coronavirus deaths. *Intelligent Systems in Accounting, Finance and Management*, *27*(3), 151-158.

Priedhorsky, R., Osthus, D., Daughton, A. R., Moran, K. R., Generous, N., Fairchild, G., ... & Del Valle, S. Y. (2017, February). Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1812-1834).

Santangelo, O. E., Gianfredi, V., & Provenzano, S. (2022). Wikipedia searches and the epidemiology of infectious diseases: A systematic review. *Data & Knowledge Engineering*, *142*, 102093.

Sharpe, J. D., Hopkins, R. S., Cook, R. L., & Striley, C. W. (2016). Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. *JMIR public health and surveillance*, *2*(2), e5901.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Simonsen, L., Gog, J., Olson, D., & Viboud, C. (2016). Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. Journal of Infectious Diseases, 214(suppl.4). https://doi.org/10.1093/infdis/jiw376.

Smith DA (2020) Situating Wikipedia as a health information resource in various contexts: A scoping review. PLoS ONE 15(2): e0228786. https://doi.org/10.1371/journal.pone.02287.

Transtrum, M. K., Machta, B. B., Brown, K. S., Daniels, B. C., Myers, C. R., & Sethna, J. P. (2015). Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, *143*(1).

Wikimedia Foundation. (2025). Pageviews Analysis. Meta-Wiki. https://pageviews.toolforge.org/

Zimmer C, Leuba SI, Yaesoubi R, Cohen T. 2018 Use of daily Internet search query data improves real-time projections of influenza epidemics. J. R. Soc. Interface 15: 20180220. http://dx.doi.org/10.1098/rsif.2018.0220.

4th IFC and Bank of Italy Workshop on "Data Science in Central Banking"
18-20 February 2025, Rome, Italy

Forecasting influenza-like illness in Italy using Wikipedia: a principal component regression approach

Gianluca Mura

Bank of Italy

Disclaimer: The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the position of Bank of Italy

AGENDA

1. Context and motivation
2. Data collection and preprocessing
3. Principal component analysis (PCA): an overview
4. Model building: principal component regression (PCR)
5. Model performance evaluation
6. Conclusions

1. Context and motivation

Influenza remains a significant public health challenge, with substantial epidemiological, clinical, and economic impacts. Factors such as its high transmissibility, antigenic variability, seasonal epidemics (and occasional pandemics), and severe complications in vulnerable populations (e.g., children, elderly) contribute to its impact on society.

According to recent estimates, the European Centre for Disease Prevention and Control (ECDC) reports that seasonal influenza causes approximately 40 million symptomatic cases and 28,000 influenza-related deaths annually in Europe.

Additionally, **influenza-like-illness** (ILI) imposes substantial social costs due to lost working days and decreased productivity among working adults and students, particularly those who are unvaccinated, under 65 years of age, or experiencing severe disease. Considerable work time and productivity loss is also attributable to caregiver burden.

These figures underline the critical role of surveillance systems in guiding vaccine composition, healthcare planning, and understanding influenza-related complications.

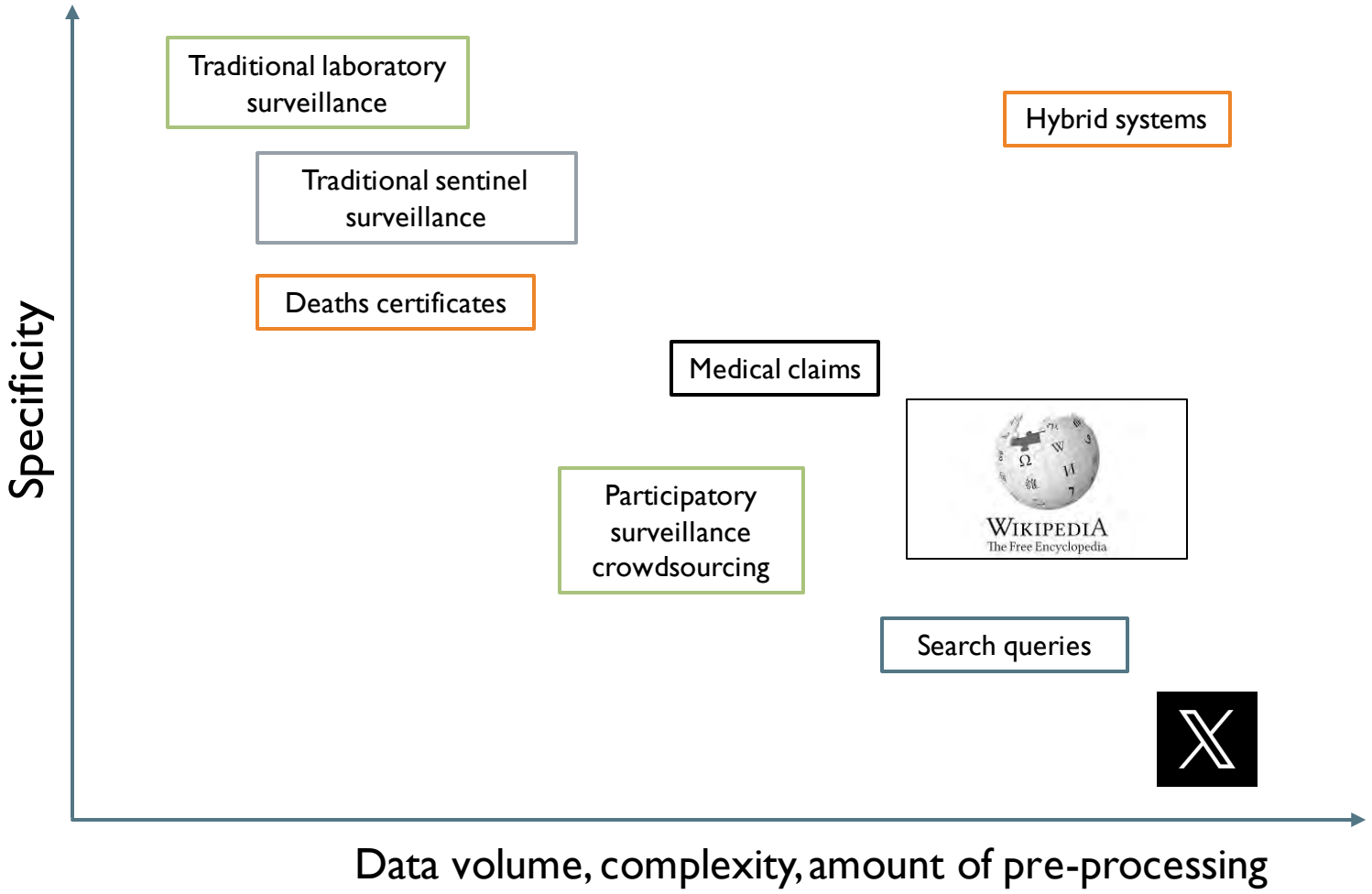
Traditional surveillance systems may find a limitation in the delay of data availability of up to two weeks compared to the reporting period due to the necessary collection and processing activities of the information sent by the network of sentinel physicians. Therefore, emerging methods, such as **monitoring Internet activity and social media trends**, have gained attention for their potential to enhance near real-time influenza surveillance (better situational awareness).

In this context, several initiatives have been promoted by major health organizations (WHO, ECDC, CDC) to develop alternative systems based on big data for forecasting influenza and other viral diseases.

1. Context and motivation

Surveillance systems have evolved significantly, incorporating both classical and big data-driven approaches. Classical methods now include **electronic death certificates, patient-level hospital discharge records, and medical claims data**,

Alongside these, **big data-driven systems** have emerged, relying on streams from internet search queries, social media, and crowdsourcing platforms to accelerate the detection and monitoring of syndromic trends.



Source: author's elaboration based on "Simonsen L. et al. - Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J Infect Dis 2016;214:S380-5".

2. Data collection and preprocessing

In Italy, the RespiVirNet (formerly **InfluNet**) surveillance system monitors influenza-like illnesses (ILI) during the seasonal period through a network of so-called sentinel physicians.

The system relies on a network of voluntarily participating general practitioners (GPs) and pediatricians (PLS) and is coordinated by the Istituto Superiore di Sanità (ISS). Epidemiological surveillance is conducted via weekly reports of ILI cases from a sample of GPs and PLS during the period from October to April.



InfluNet
Rete Italiana Sorveglianza Influenza

Settimana **2019 - 17**
dal **22** *al* **28** *aprile al 2019*

Stagione Influenzale 2018 - 2019

Rapporto N. 27 del 21 maggio 2019

Risultati Nazionali

La tabella seguente mostra il numero dei casi e i tassi d'incidenza, nel totale e per fascia di età, di tutte le regioni che hanno inviato i dati. L'incidenza settimanale è espressa come numero di sindromi influenzali (casi) per 1.000 assistiti.

Settimana	Totale Medici	Totale Casi	Totale Assistiti	Totale Incidenza	0-4 anni		5-14 anni		15-64 anni		65 anni e oltre	
					Casi	Inc	Casi	Inc	Casi	Inc	Casi	Inc
2018-42	1.073	1.088	1.384.736	0,79	100	1,26	117	0,63	691	0,85	180	0,58
2018-43	1.080	1.319	1.396.489	0,94	155	1,96	148	0,80	855	1,05	161	0,51
2018-44	1.091	1.421	1.411.739	1,01	134	1,68	163	0,86	909	1,10	215	0,68
2018-45	1.091	2.074	1.411.584	1,47	235	2,93	250	1,31	1266	1,54	323	1,02

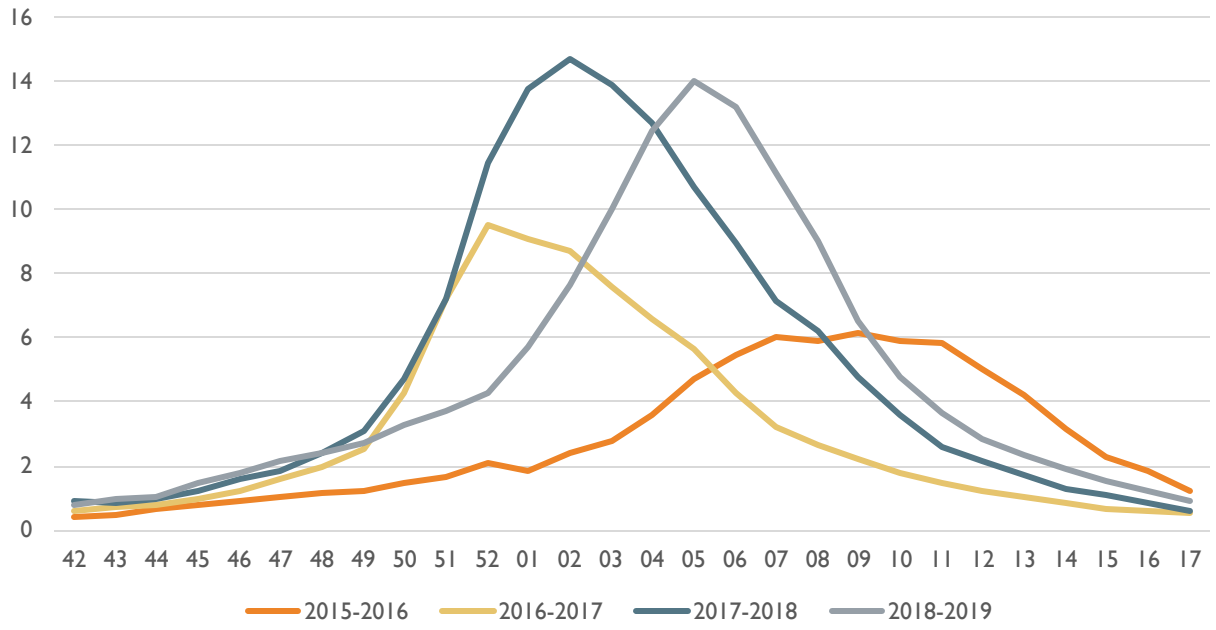
2. Data collection and preprocessing

The data on the reference week and the weekly incidence data, expressed as the number of influenza syndromes (cases) per 1,000 patients (total incidence), were extrapolated. Each influenza season, and thus the relevant sample survey, has an average duration of 28 weeks, beginning on the 42nd week of one year and ending on the 17th of the following year.

Due to the limited data available on Wikipedia access logs, four influenza seasons were selected for equivalent 112 weekly observations. As is usually the case, the four influenza seasons examined had different characteristics with different intensities and epidemic peaks.

Influenza Seasons	From	to	n. of weeks
2015-2016	2015-42*	2016-16	28
2016-2017	2016-42	2017-17	28
2017-2018	2017-42	2018-17	28
2018-2019	2018-42	2018-17	28

Note: the year 2015 ended with the 53rd week, the others with the 52nd



Source: InfluNet reports

2. Data collection and preprocessing

Key arguments in favor of using Wikipedia for predicting Flu incidence rates.

- **Widespread usage and accessibility**: i) Wikipedia is one of the most accessed online resources for health information, surpassing established platforms like WebMD and the World Health Organization in terms of web traffic; ii) Its content is frequently updated in real time, making it a dynamic and responsive source for health trend.
- **Correlations with epidemiological data**: several studies have shown moderate to strong correlations between Wikipedia page views and official epidemiological data, such as influenza incidence tracked by health institutes
- **Global relevance and multilingual content**: Wikipedia offers health content in over 275 languages, providing a globally representative dataset that complements regionally limited surveillance systems (Note: the language of Wikipedia articles is not always a good proxy for the geographical scope of interest).
- **Real-time surveillance**: Wikipedia page views can provide near real-time surveillance of public interest in specific diseases, which is faster than traditional health monitoring systems that may face delays in reporting. This rapid data acquisition makes it suitable for short-term forecasting and early detection of disease outbreak.

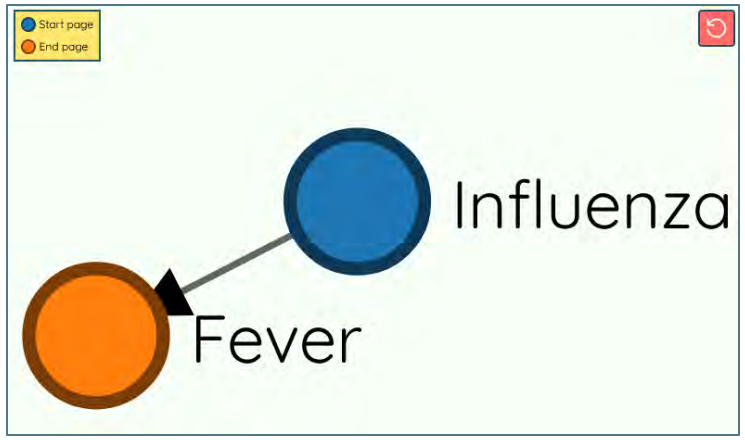
2. Data collection and preprocessing

Wikipedia provides **various tools and APIs** to analyze its data, including access logs, page content, hyperlinks, and editor activity. There are also numerous free third-party solutions and open-source projects have been developed by the community, enhancing the functionality and accessibility of Wikipedia data.

With reference to the number of Wikipedia **access logs**, information on daily accesses as of 1 July 2015 was identified by selecting all types of platforms (desktop, mobile and mobile application) and discarding accesses generated by crawlers and other automated programs) . The number of accesses was then aggregated by ISO week and purged of irrelevant observations because they were not included in the annual flu season survey.

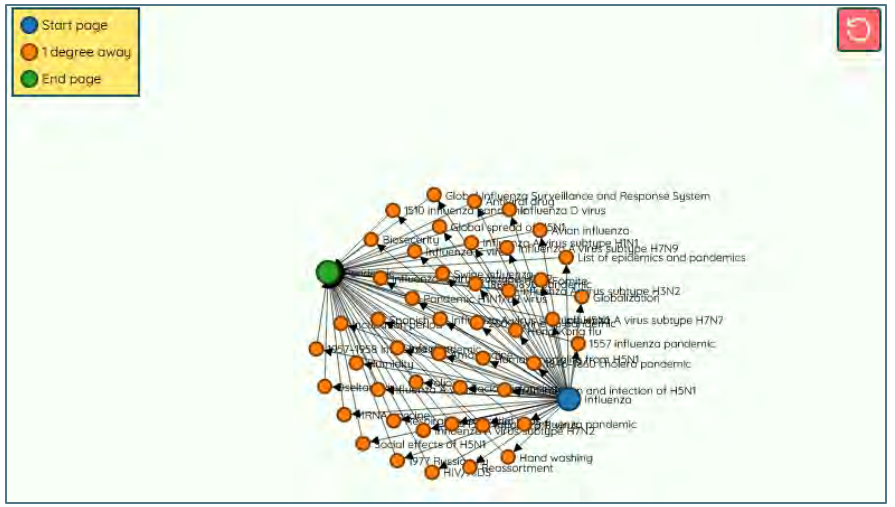
With regard to the selection of the most relevant Wikipedia articles, several approaches are available, which involve analyzing their relationships based on hyperlinks, categories, content similarity, or a combination of these factors.

1 path with 1 degree of separation from **Influenza** to **Fever**



Source: www.sixdegreesofwikipedia.com

47 paths with 2 degrees of separation from **Influenza** to **Pandemic**



2. Data collection and preprocessing

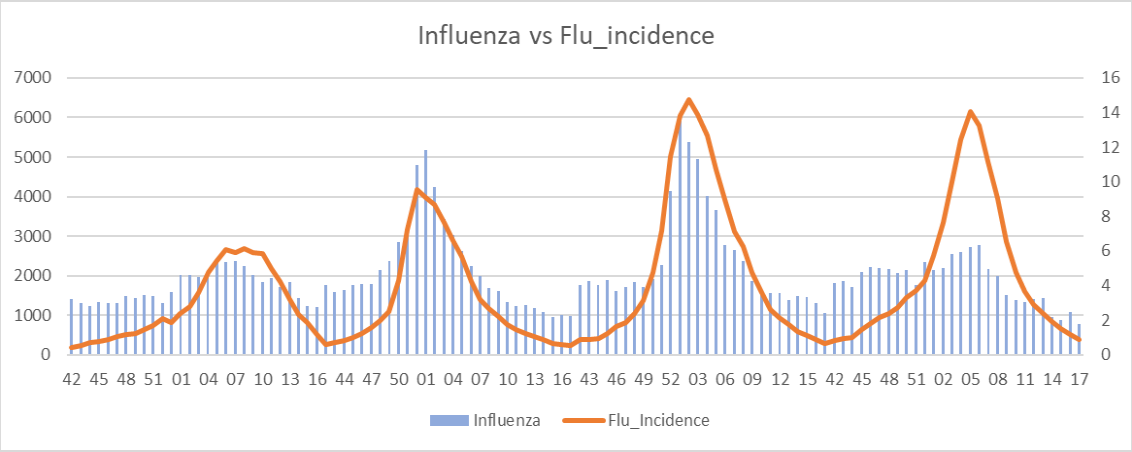
The selected pages were curated based on **thematic relevance**, with publications and prior studies on influence rate prediction reviewed to guide the selection of topics most likely to contribute to accurate modeling.
After this selection, distance metrics were applied as methods of verification and confirmation, ensuring the dataset represented a cohesive and meaningful subset of Wikipedia content. At the end of this process, 14 articles were selected.

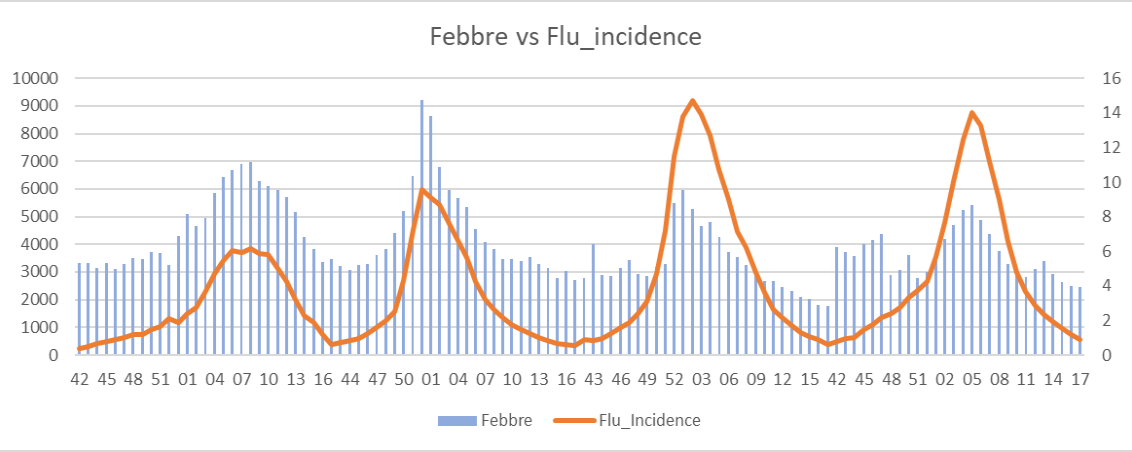
Wikipedia articles in Italian	Corresponding article in English	Total number of views over the study period (4 flu seasons)	Shortest Path Distance
Influenza	Influenza	227.775	-
Febbre	Fever	446.206	1
Cefalea	Headache	230.886	1
Influenza_aviaria	Avian influenza	91.820	1
Influenza_suina	Swine influenza	49.771	1
Epidemia	Epidemic	62.671	2
Pandemia	Pandemic	96.235	2
Pandemia_influenzale	Influenza pandemic	21.314	1
Raffreddore_comune	Common cold	225.309	1
Paracetamolo	Paracetamol	806.148	1
Raffreddore	Cold	12.419	2
Virus parainfluenzali umani	Human parainfluenza viruses	11.355	1
Antivirale	Antiviral drug	45.896	1
Influenza virus A sottotipo H1N1	Influenza A virus subtype H1N1	93.237	1

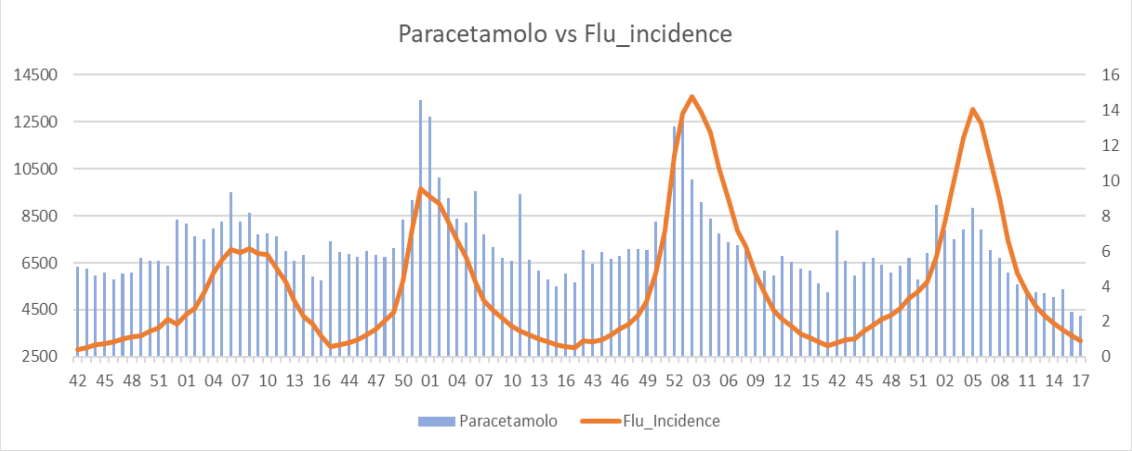
Source: https://meta.wikimedia.org/

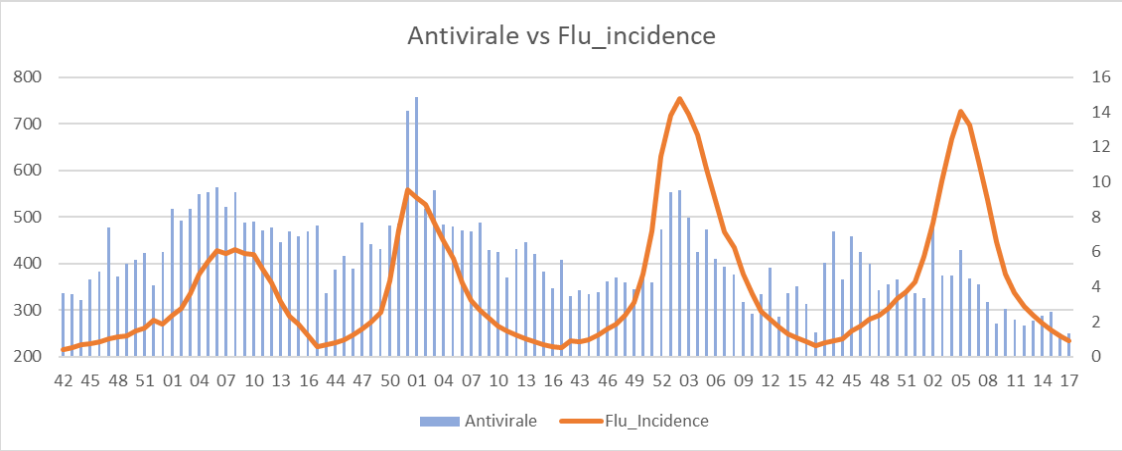
2. Data collection and preprocessing

Access logs for selected Wikipedia pages (left axis) plotted alongside official ILI rates (right axis)









3. Principal component analysis (PCA): an overview

PCA is a statistical technique used to **reduce the dimensionality** of a dataset while retaining as much variability as possible. It transforms correlated variables into a **new set of uncorrelated variables**, called principal components, ordered by the amount of variance they capture. PCA is widely used for **simplifying data structures, identifying patterns, and preparing data for predictive modeling** or visualization.

Pros of PCA:

1. **dimensionality reduction**: PCA can reduce the number of variables while retaining most of the important information;
2. **removes multicollinearity** between variables by creating orthogonal principal components;
3. **noise reduction**: PCA can help filter out noise in the data by focusing on the components that explain the most variance.
4. **better visualization**: PCA allows for easier visualization of high-dimensional data by projecting it onto lower dimensions.
5. **improved model performance**: using principal components as inputs can improve the performance of some machine learning models.

Cons of PCA:

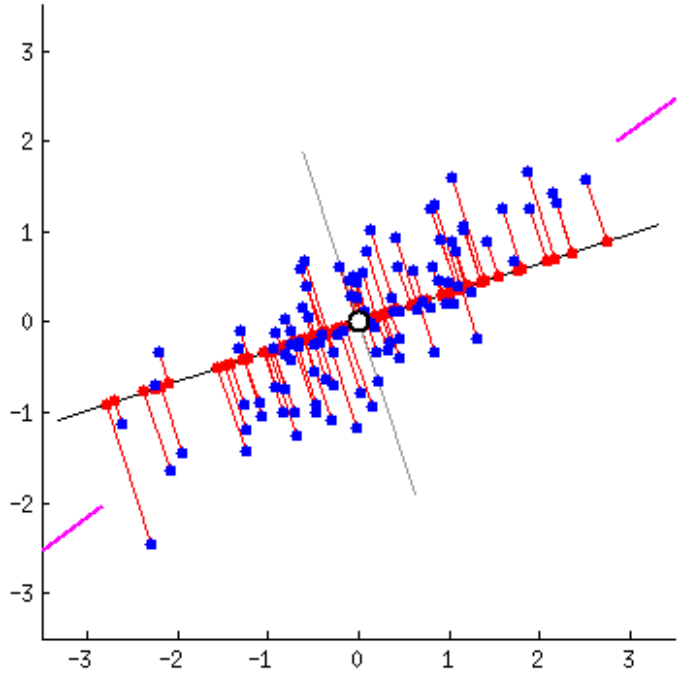
1. **loss of interpretability**: the principal components are often difficult to interpret in terms of the original variables;
2. **assumes linearity**: PCA assumes linear relationships between variables, which may not always be the case;
3. **sensitive to outliers**: PCA can be significantly affected by outliers in the dataset;
4. **may lose some information**: while PCA aims to retain the most important information, some potentially valuable details may be lost in the process of dimensionality reduction;
5. **requires standardization**: variables need to be standardized before applying PCA, which can be an additional preprocessing step.

3. Principal component analysis (PCA): an overview

The graph below is intended to explain the transformation of 2D data into 1D using PCA.

It shows the new vector, which is determined by rotating around the mean of the 2D distribution to find the direction that minimizes the projection error (i.e. maximizes the variance captured).

The original data points are projected onto this new line, representing their positions along the principal component, effectively reducing the dimensionality while preserving as much information as possible



4. Model building: principal component regression (PCR)

Principal Component Regression (PCR) combines PCA and linear regression to address issues like multicollinearity and dimensionality reduction. The consequential steps followed to perform the PCR were:

1. **Data Preparation**: check for missing values and outliers; assess normality of variables; standardize the data to ensure all variables are on the same scale.
2. **Correlation analysis**: create a correlation matrix to identify highly correlated variables.
3. **PCA computation** on standardized predictor variables, examination of the scree plot to determine the number of components to retain; cumulative explained variance plot analysis.
4. **Choice of the number of components** to be considered (different methods available).
5. **PCR model building**: selected principal components were used as predictors in a linear regression model; the model is fitted using ordinary least squares.
6. **Model Evaluation**: assess model performance using appropriate metrics.
7. **Interpretation**: analyze the contribution of original variables to principal components; interpret PCR coefficients in terms of the original variables.
8. **Prediction:** generation of predictions using the PCR model; comparison of actual vs. predicted values.
9. **Refinement**: evaluation of possible variations to the initial model to improve performance.

4. Model building: principal component regression (PCR)

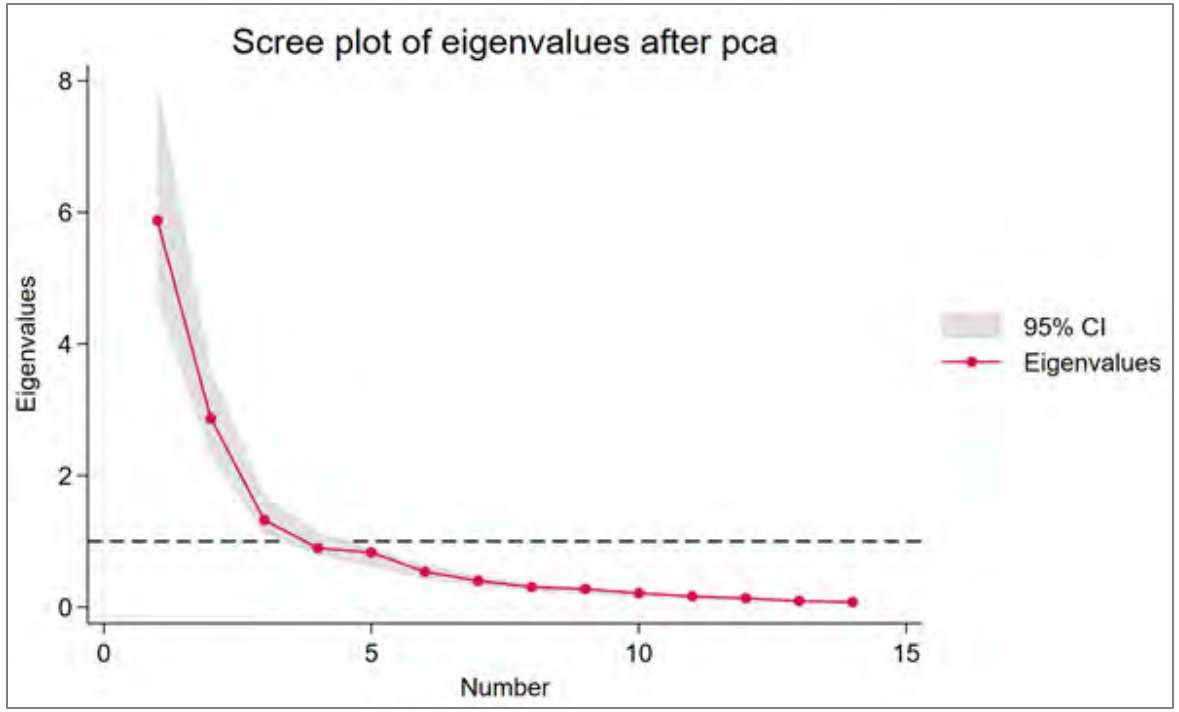
The correlation matrix revealed several medium and high correlations between the independent variables, as hypothesized, justifying the application of PCA to address multicollinearity. Additionally, the statistical significance of these correlations was assessed, with *p-values* confirming their reliability.

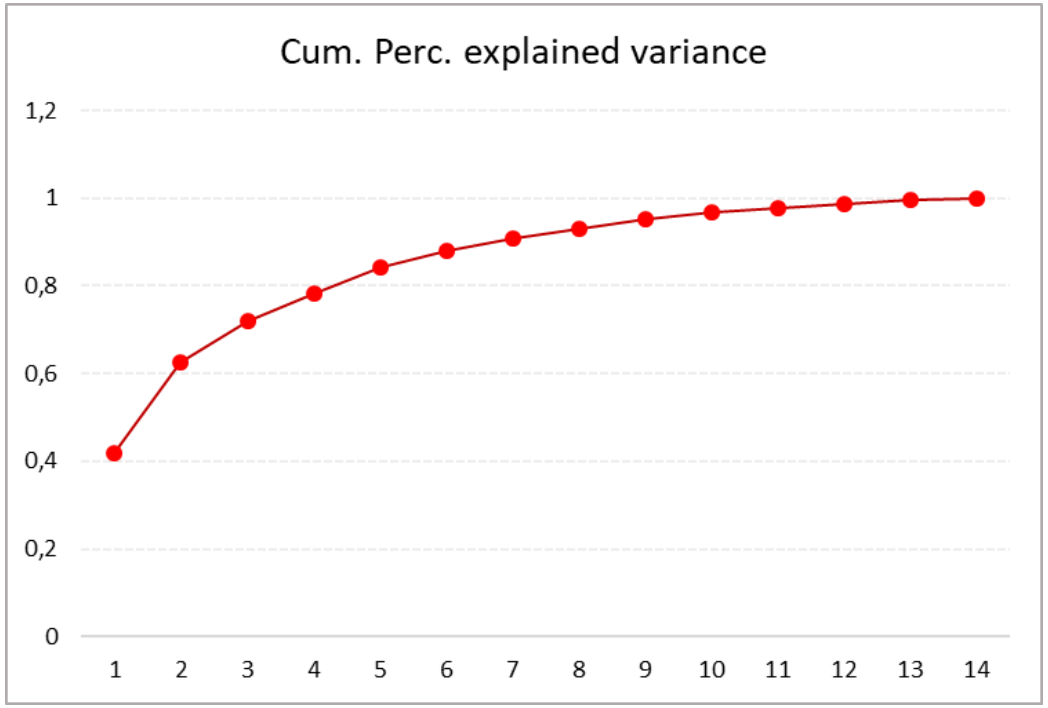
	Flu_Incidence	Influenza	Febbre	Cefalea	Influenza_aviaria	Influenza_suina	Epidemia	Pandemia	Pandemia_influenzale	Raffreddore_comune	Paracetamolo	Raffreddore	Virus parainfluenzali umani	Antivirale	Influenza virusA sottotipo H1N1
Flu_Incidence	1														
Influenza	0,8113	1													
Febbre	0,5887	0,6822	1												
Cefalea	-0,0923	0,1485	0,5865	1											
Influenza_aviaria	0,3181	0,4072	0,1761	-0,0368	1										
Influenza_suina	0,5571	0,3749	0,2183	-0,0843	0,2269	1									
Epidemia	0,1521	0,207	0,5839	0,7214	0,0591	0,0981	1								
Pandemia	0,3584	0,4268	0,5043	0,4976	0,1362	0,2738	0,6397	1							
Pandemia_influenzale	0,6823	0,6822	0,3008	-0,2233	0,2578	0,4726	0,0123	0,2926	1						
Raffreddore_comune	0,0568	0,384	0,5807	0,5393	-0,0736	-0,1105	0,3556	0,3391	0,0053	1					
Paracetamolo	0,6601	0,8384	0,7795	0,3003	0,358	0,2166	0,3223	0,3874	0,4267	0,5879	1				
Raffreddore	0,3938	0,5137	0,5627	0,3751	0,144	0,1214	0,3602	0,3833	0,0425	0,6659	0,6465	1			
Virus parainfluenzali umani	0,4691	0,6003	0,2763	-0,0928	0,2187	0,0716	-0,0268	0,1613	0,6316	0,1739	0,4821	0,1646	1		
Antivirale	0,3864	0,6264	0,8392	0,7166	0,1815	0,1272	0,6205	0,5237	0,1683	0,6271	0,7429	0,5698	0,253	1	
Influenza virusA sottotipo H1N1	0,767	0,6359	0,3177	-0,14	0,2704	0,5499	0,0418	0,3348	0,6161	-0,0356	0,4412	0,2156	0,3695	0,2079	1

4. Model building: principal component regression (PCR)

Selecting the optimal number of factors (or principal components) is crucial in PCA to balance dimensionality reduction and information retention. Several approaches offer support for determining the number of factors, and it is common to use a combination of methods rather than relying on a single criterion.

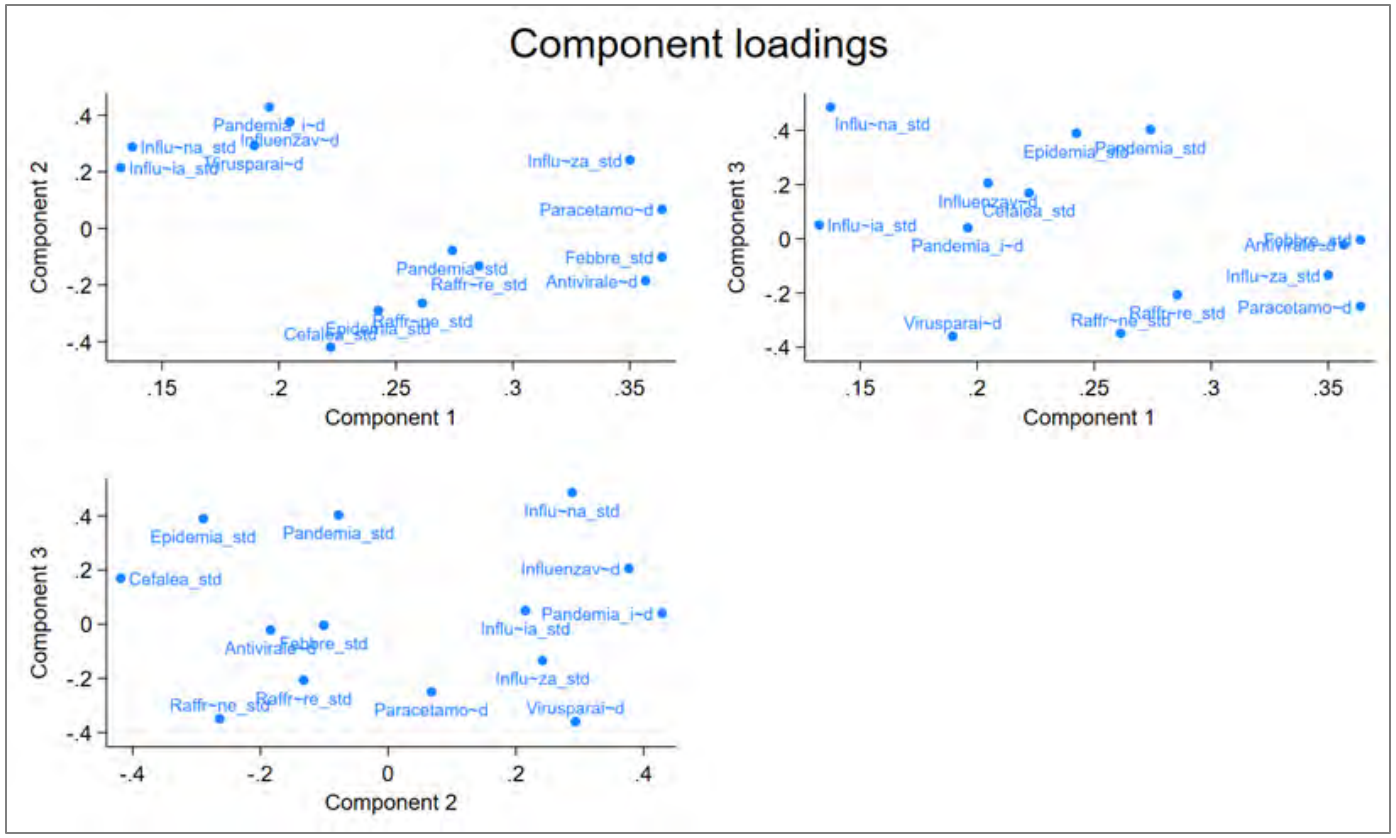
1. **Kaiser Criterion:** eigenvalue > 1;
2. **Cumulative Variance Explained:** select enough components to achieve a cumulative variance explained threshold, typically 70–90%;
3. **Scree Plot**: i.e. plotting the eigenvalues in descending order and look for the "elbow point" where the eigenvalues level off. Components are normally retained before this point.





4. Model building: principal component regression (PCR)

Loading plots (left side) visually represent the relationships between variables and components, helping to identify key contributors, variable groupings, and the underlying structure of the data in dimensionality reduction. I also assessed the adequacy of the dataset for PCA by performing the **Kaiser-Meyer-Olkin (KMO) test**. A KMO value greater than 0.80 indicates good sampling adequacy, suggesting that the variables share sufficient common variance to justify the use of PCA. This also suggest that the dataset is well-suited for dimensionality reduction, and the extracted components are likely to be meaningful.

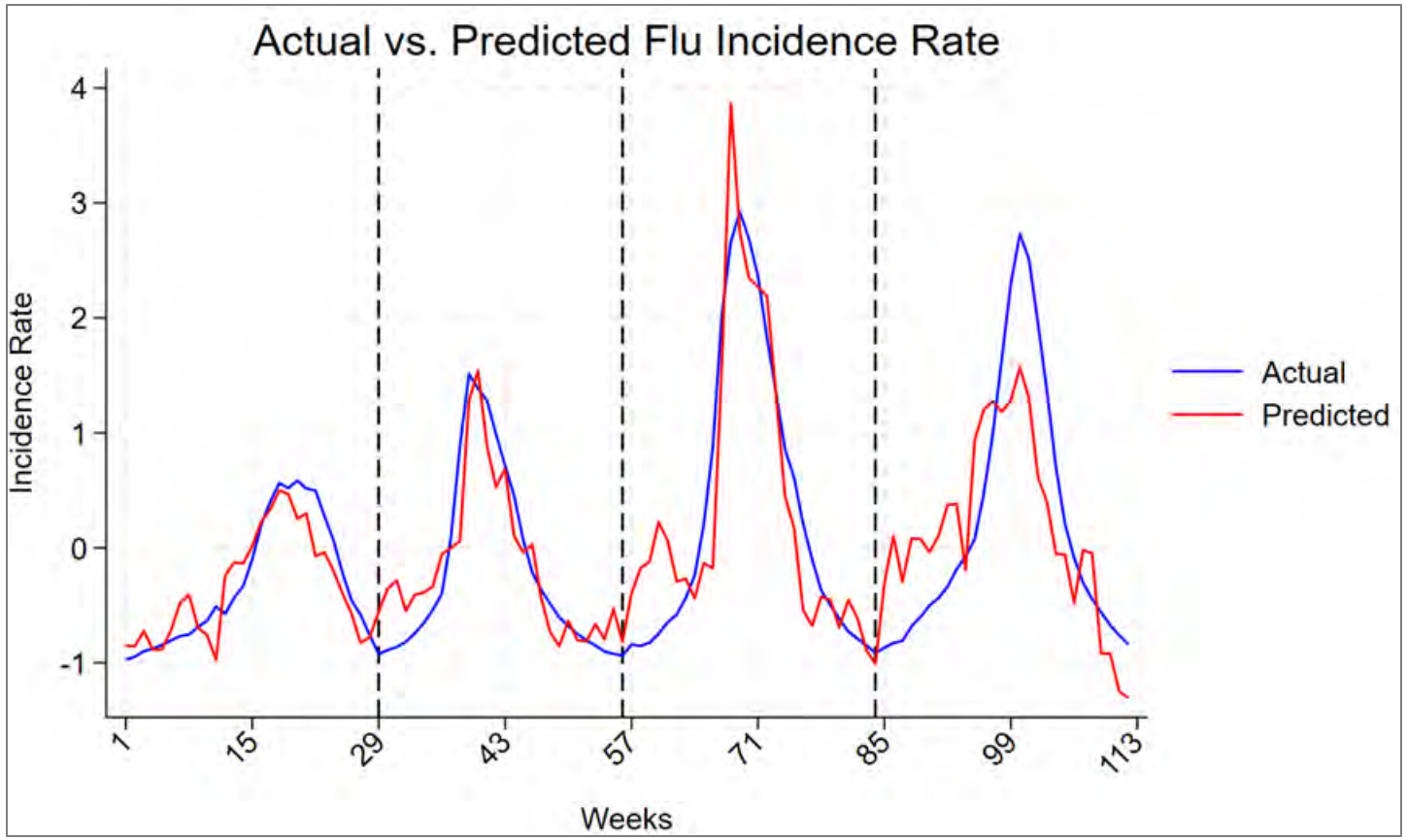


Variable	KMO
Influenza	0,8414
Febbre	0,9196
Cefalea	0,7868
Influenza_aviaria	0,7249
Influenza_suina	0,7365
Epidemia	0,7918
Pandemia	0,8715
Pandemia_influenzale	0,704
Raffreddore_comune	0,7933
Paracetamolo	0,8687
Raffreddore	0,8122
Antivirale	0,8888
Virus parainfluenzali umani	0,8155
Influenza virus A sottotipo H1N1	0,8657
Overall	**0,8308**

5. Model performance evaluation

The model performs well in explaining the variability of the dependent variable (explain 77.23% of the variance in the dependent variable) and shows good predictive accuracy. Additionally, the overall model is highly statistically significant.

$$Flu_{incidence} = -2.62e-09-09 + 0.2750 \cdot pc1 + 0.3337 \cdot pc2 + 0.0794 \cdot pc3 + \varepsilon$$



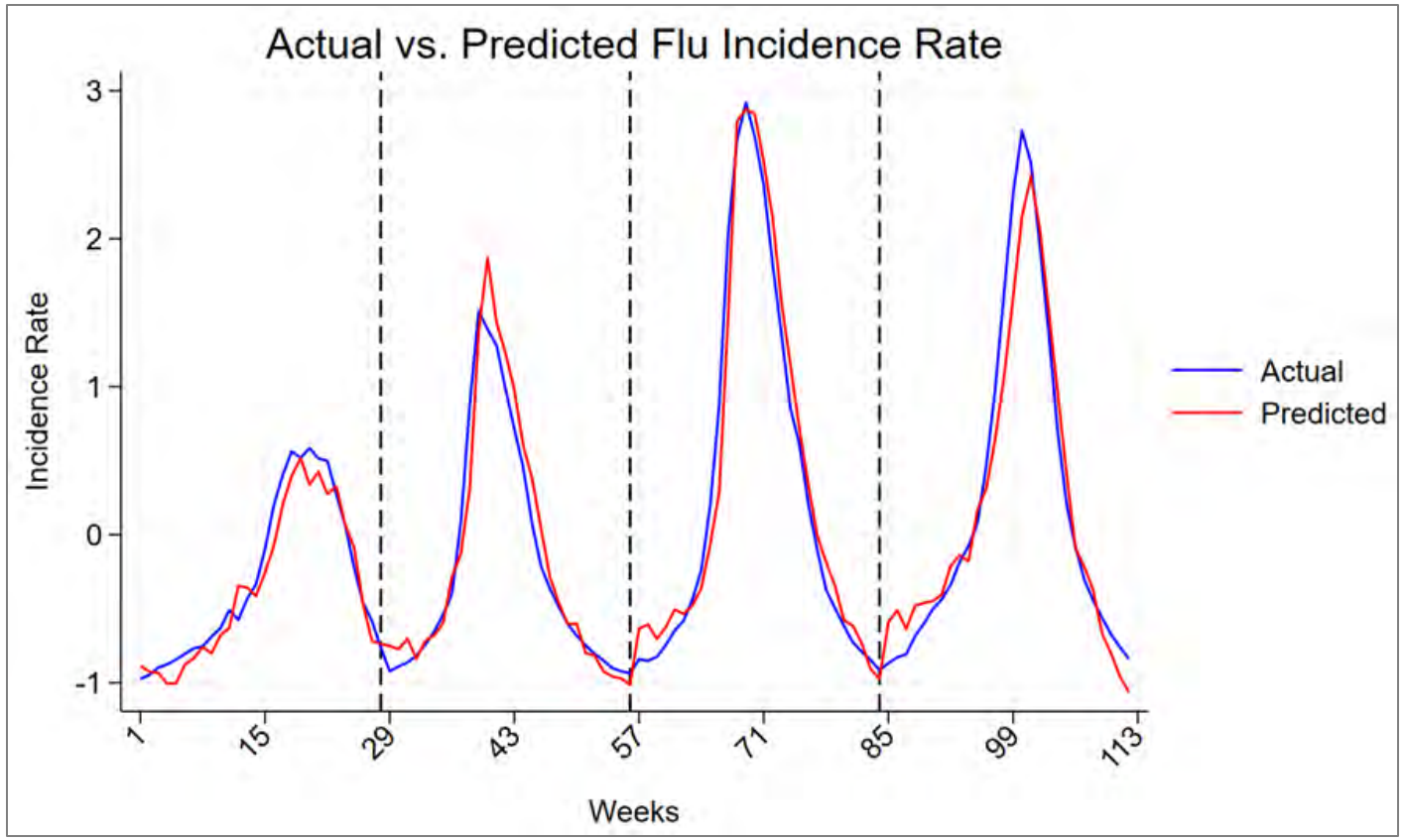
$\boldsymbol{r^2 = 0.7723}$

RMSE = 0.4837 ~12% of variable range

5. Model performance evaluation

To enhance the predictive performance of the model, a new explanatory variable was introduced: the **flu spread rate of the previous week**. This variable, representing known data, leverages temporal dependencies inherent in influenza dynamics. The rationale for this choice is supported by findings that historical data – such as previous week influenza levels – offer crucial insights into current trends. These temporal dependencies allow the model to dynamically adjust predictions.

$$Flu_incidence = -3.42e-09 + 0.0942 \cdot pc1 + 0.0932 \cdot pc2 - 0.0607 \cdot pc3 + 0.7542 \cdot Flu_incidence_{t-1} + \varepsilon$$



$$r^2 = \mathbf{0.9572}$$

$$RMSE = 0.2106$$

5. Conclusions

With this work, drawing on experiments and analyses previously conducted in the US and Europe, I aimed to assess the feasibility of developing a statistical model capable of estimating ILI rates in Italy using access log data from specific Wikipedia page consultations.

The Principal Component Regression-based approach proved particularly well-suited to the explanatory variables, as it efficiently manages the high dimensionality and correlations present in the dataset.

Moreover, the selection of Wikipedia articles can be expanded and continuously adjusted to account for contingent situations, such as changes in public interest during health emergencies.

The encouraging initial results confirm the potential for defining a hybrid model that combines traditional surveillance systems with real-time data sources like Wikipedia, providing enhanced forecasting capabilities and improving public health monitoring in Italy.

This study contributes to the broader goal of leveraging alternative data streams to complement traditional methods, addressing timeliness and adaptability in disease surveillance.

THANK YOU

gianluca.mura@bancaditalia.it

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

A foundational model for macroeconomic times series forecasting and nowcasting[1]

Batuhan Koyuncu, Byeungchun Kwon, Marco Lombardi,
Fernando Perez-Cruz and Hyun Song Shin,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.



A Foundational Model for Macroeconomic Times Series Forecasting and Nowcasting

B Koyunku, B Kwon, M Lombardi, **F Perez-Cruz*** and H S Shin

Bank for International Settlements *Prof at ETH Zurich

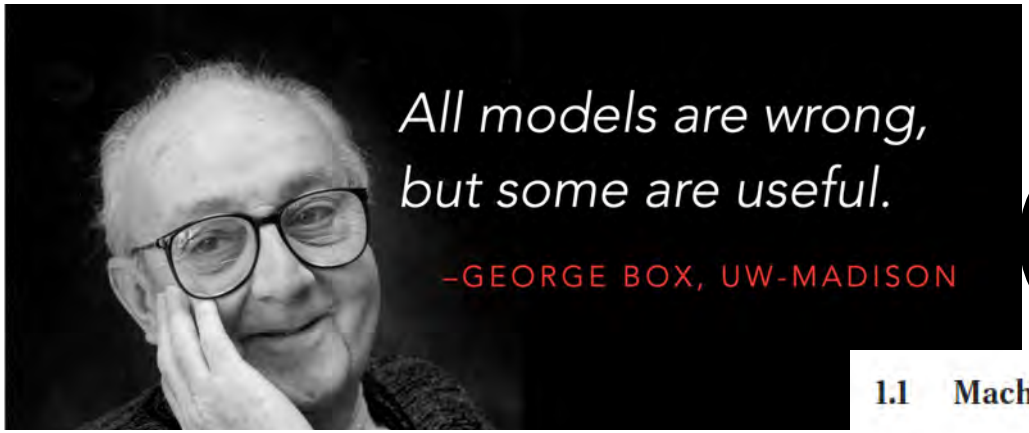
Why ML is not Statistics?

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f_w(\boldsymbol{x}_i)\right)$$

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1\ldots n} \underset{iid}{\sim} p_{X,Y}(\boldsymbol{x}, y)$$

Machine Learners and Statisticians solved the **same** problem

Why ML is not Statistics?



$$\left(y_i, f_w(\boldsymbol{x}_i)\right)$$

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}$$

Machine Learners anc

1.1 Machine learning: what and why?

We are drowning in information and starving for knowledge. — John Naisbitt.

We are entering the era of **big data**. For example, there are about 1 trillion web pages[1]; one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day[2]; the genomes of 1000s of people, each of which has a length of 3.8×10^9 base pairs, have been sequenced by various labs; Walmart handles more than 1M transactions per hour and has databases containing more than 2.5 petabytes (2.5×10^{15}) of information (Cukier 2010); and so on.

This deluge of data calls for automated methods of data analysis, which is what **machine learning** provides. In particular, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).

Do we need ML for time series analysis?

- We have many models to analyse time series:
  - AR
  - GARCH
  - VAR or BVAR
  - Kalman
  - Hidden Markov Models
  - Dynamical Switching Models
  - …

Do we need ML for time series analysis?

- We have many models to analyse time series:
  - AR
  - GARCH
  - VAR or BVAR
  - Kalman
  - Hidden Markov Models
  - Dynamical Switching Models
  - …
- Whatever the time series problem is, one of those methods should be good enough, right?

Do we need ML for time series analysis?

- We have many models to analyse time series:
  - AR
  - GARCH
  - VAR or BVAR
  - Kalman
  - Hidden Markov Models
  - Dynamical Switching Models
  - …
- Whatever the time series problem is, one of those methods should be good enough, right?
- The answer is not in the models but in the data:
  - ML helps with unknown nonlinear interactions or universality (maybe both).

The Game Changer: Zero-Shot Learning

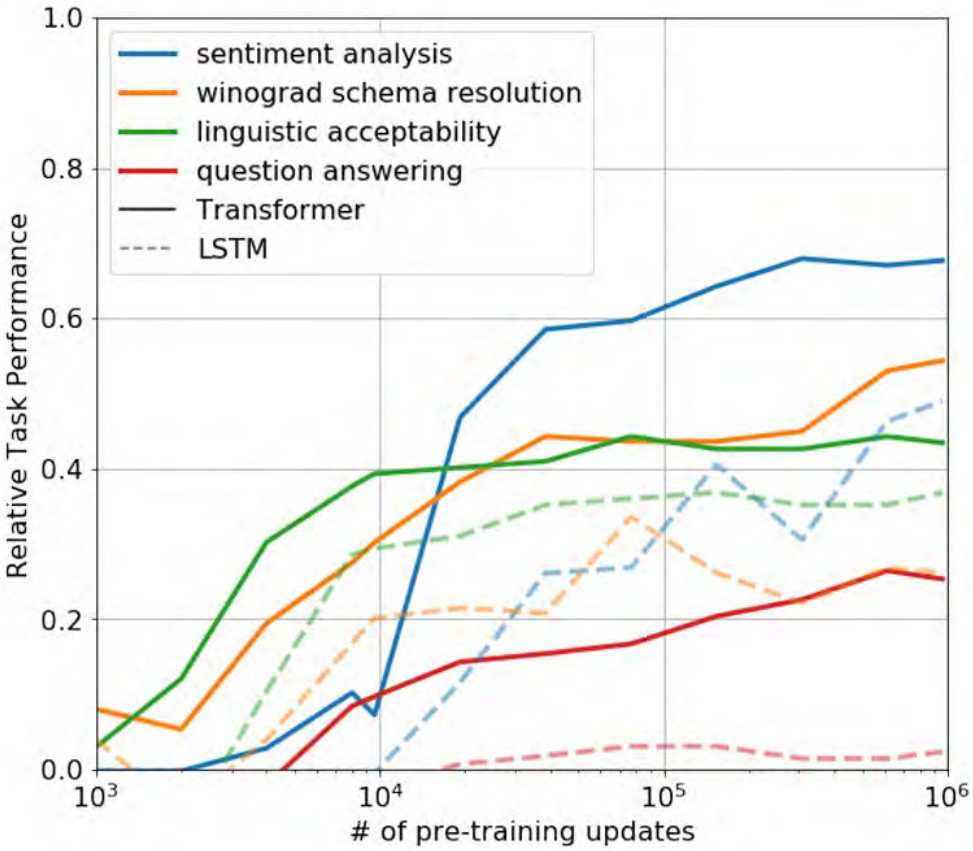
Zero-shot Behaviors We'd like to better understand why language model pre-training of transformers is effective. A hypothesis is that the underlying generative model learns to perform many of the tasks we evaluate on in order to improve its language modeling capability and that the more structured attentional memory of the transformer assists in transfer compared to LSTMs. We designed a series of heuristic solutions that use the underlying generative model to perform tasks without supervised finetuning. We visualize the effectiveness of these heuristic solutions over the course of generative pre-training in Fig 2(right). We observe the performance of these heuristics is stable and steadily increases over training suggesting that generative pretraining supports the learning of a wide variety of task relevant functionality. We also observe the LSTM exhibits higher variance in its zero-shot performance suggesting that the inductive bias of the Transformer architecture assists in transfer.

Improving Language Understanding by Generative Pre-Training

The Game Changer: Zero-Shot Learning

Zero-shot Behaviors We'd like to better understa
ers is effective. A hypothesis is that the underlying
tasks we evaluate on in order to improve its language

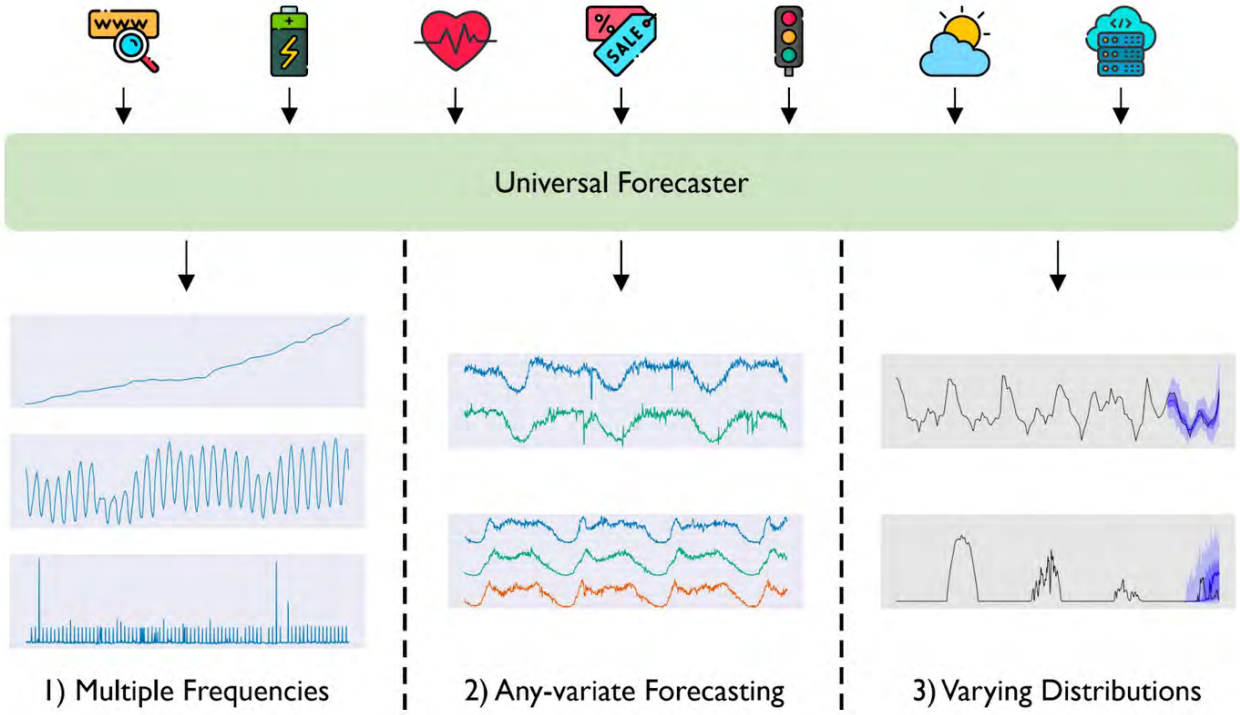
attentional memory of the transformer assists in tra
of heuristic solutions that use the underlying gener
finetuning. We visualize the effectiveness of these
pre-training in Fig 2(right). We observe the perfo
increases over training suggesting that generative p
of task relevant functionality. We also observe the
performance suggesting that the inductive bias of th



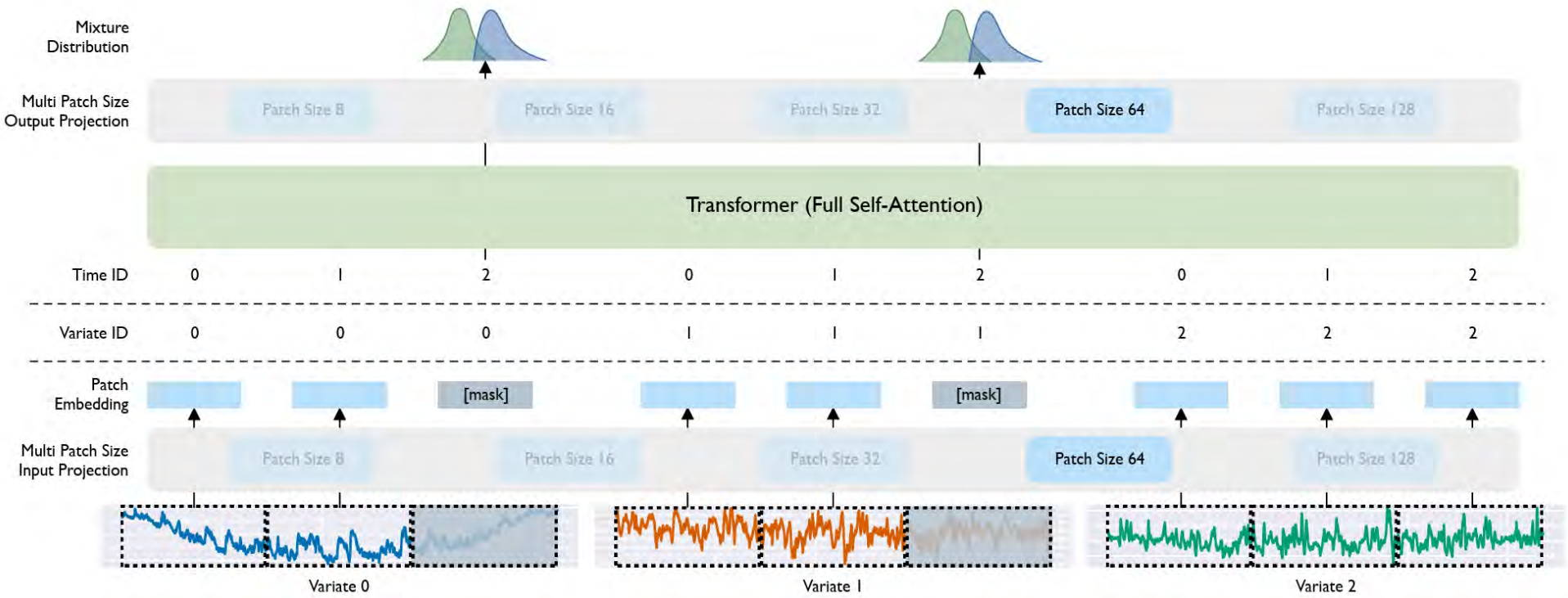
Improving Language Understanding by Generative Pre-Training

Transformers for time series: MOIRAI

- BERT based
- RoPE
- Only predict the last tokens
- Multi frequency
- Probabilistic output
- Data:
  - 60% data is about energy
  - 18% Transport
  - …
  - 0.1% Econ/Finance

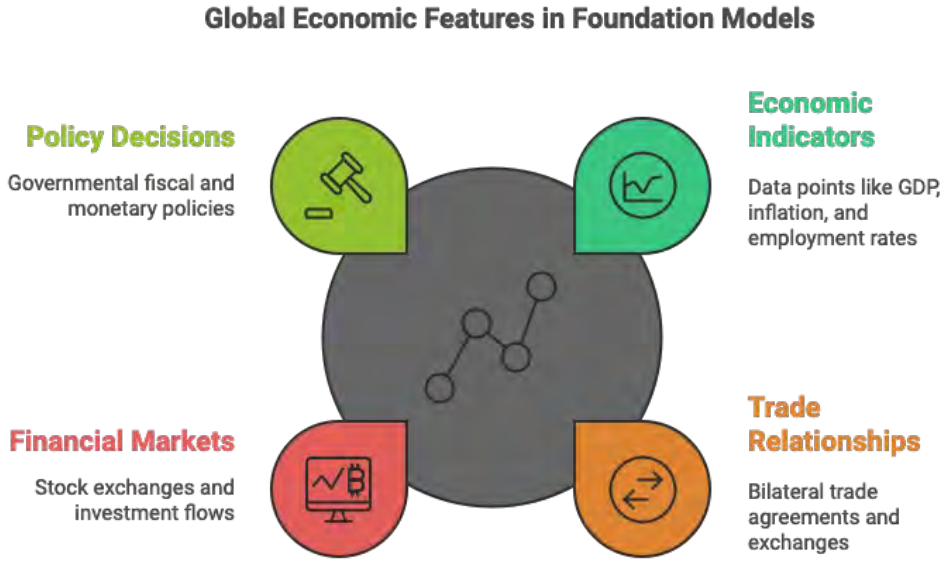


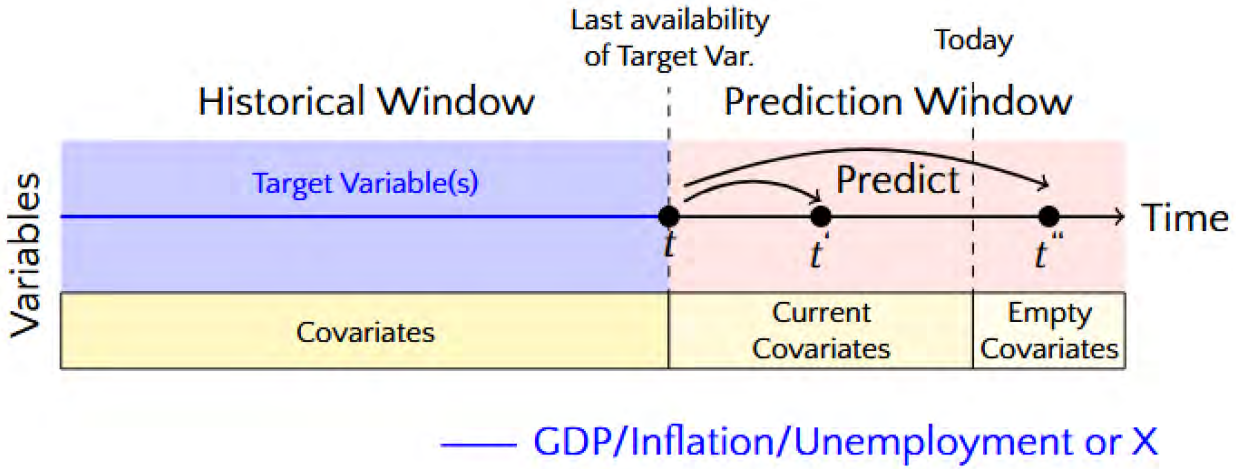
MOIRAI Architecture



MOIRAI for Macroeconomics

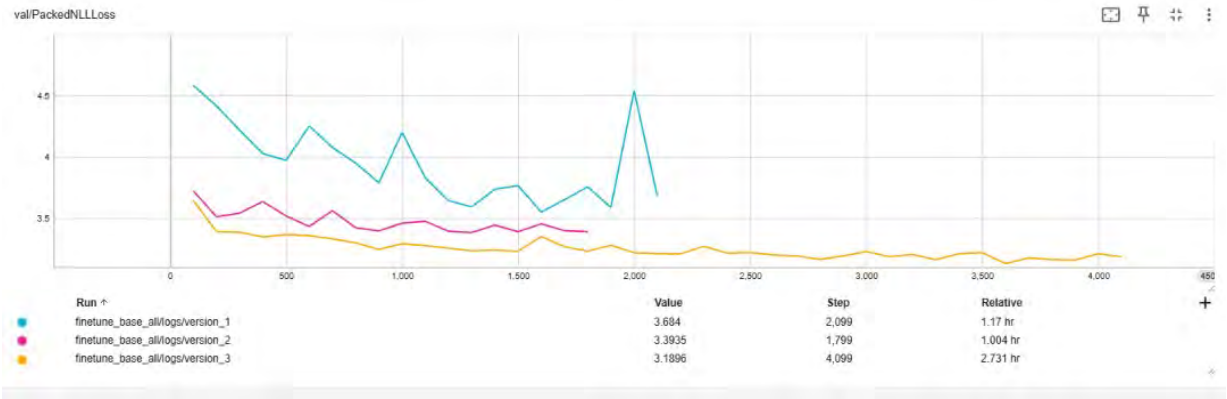
- Data:
  - BIS_MACRO data (80 countries)
  - Emphasis: GPD/CIP/Unemployment
- Granularity:
  - Daily
- Token:
  - Length 32
- Good for:
  - Forecasting
  - Nowcasting
  - Scenario planning

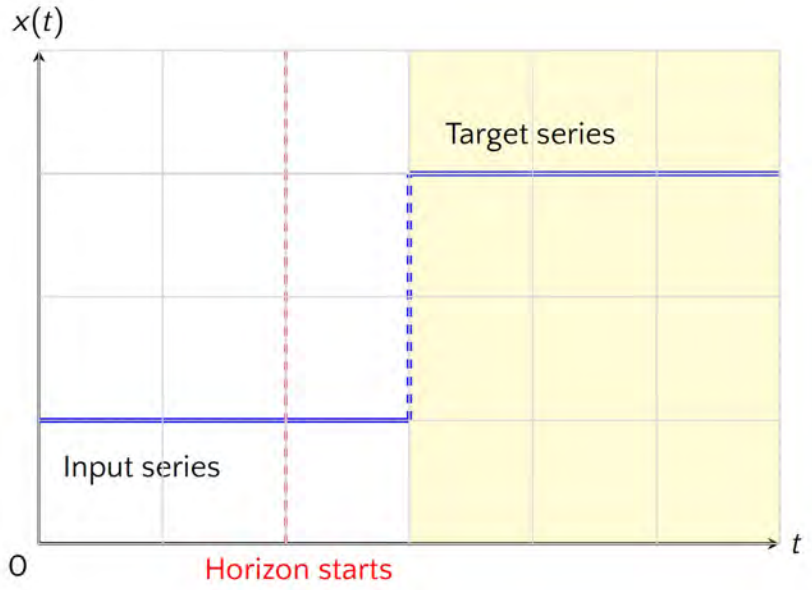




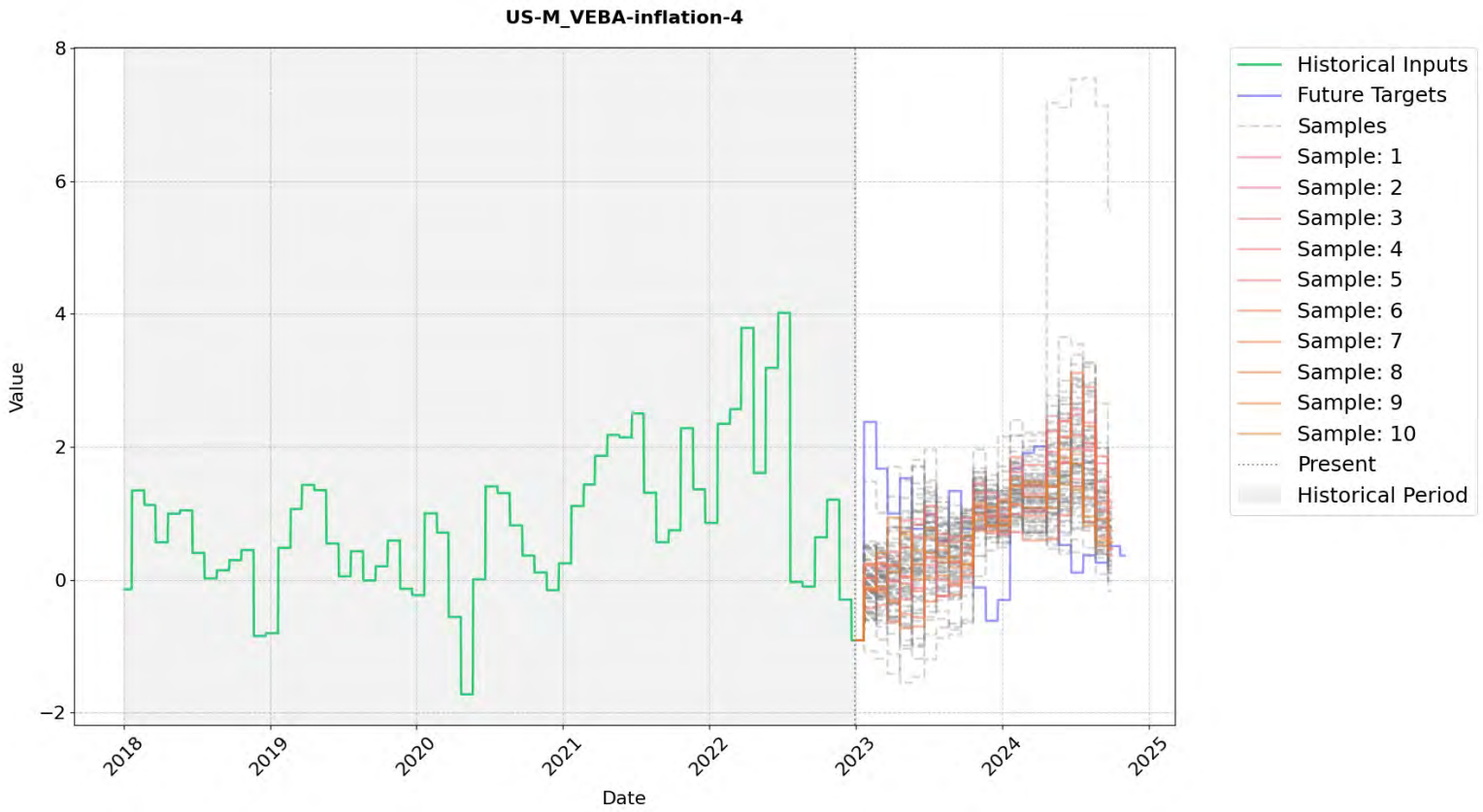
Implementation details

- Features:
  - Original data and first difference
  - Log data and first difference
- Input Vector:
  - GPD, CIP and Unemployment
  - 10 other variables
- Length: 48-60 patches (4-5 years)
- Training period [1984 to 2024]\ Test years
- Test years: 1995, 2005, 2015 and 2023+
- Masking 1 to 5 patches
- Loss over “jump mask”
- MOIRAI Base: 91m parameters

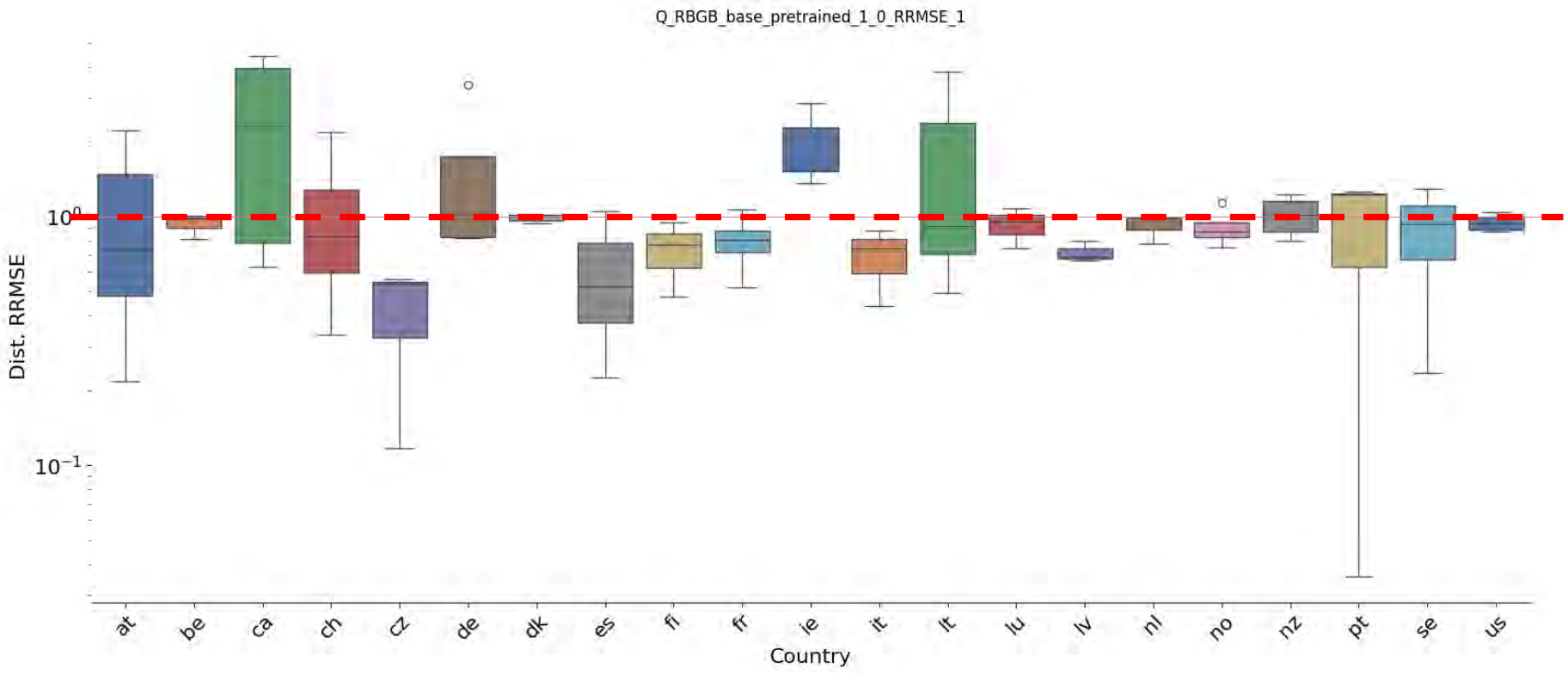




Future Looking US Inflation



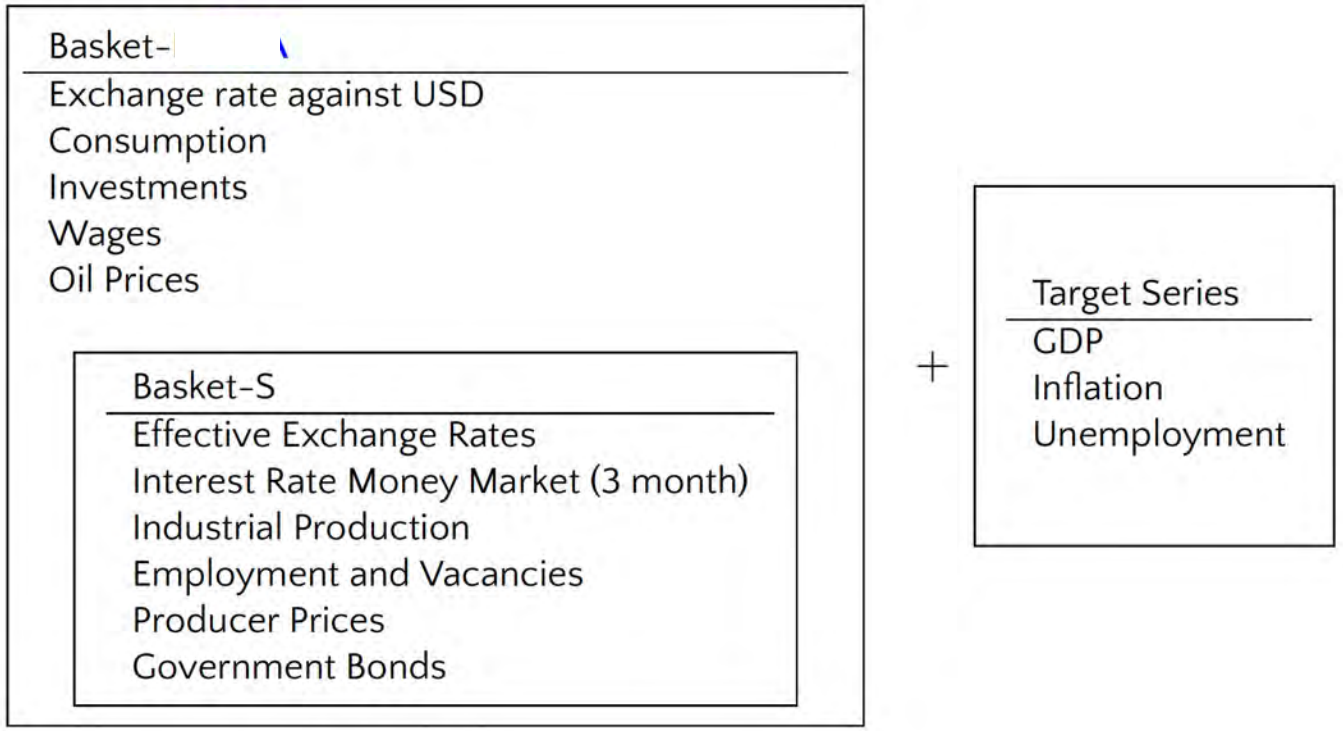
Univariate Prediction of GDP Compared to AR(1)



What's next

- Complete training the model
- Univariate estimation
- Estimation with Covariates
- Nowcasting
- Scenario planning
- BIS internal validation
- Code Available Summer 2025

Reach out for testing interesting cases



Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Data anonymization principles at Banco de Portugal[1]

Ana F Carvalho, Francisco Fonseca, Mário Lourenço
and Ricardo Marques,
Banco de Portugal

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Data anonymization principles at Banco de Portugal

Ana F. Carvalho[1], Francisco Fonseca[2], Mário Lourenço[3], Ricardo Marques[4]

Abstract

Central Banks collect increasingly larger sets of granular data which frequently contain personal data. The analytical potential of these datasets is greatly enhanced whenever they have common identifiers, meaning they can be combined in order to maximize the value of the available information. However, given the sensitive nature of some of these datasets, particularly when referring to natural persons, the protection of personal data must always be a priority. This paper focuses on the essential principles that must be observed when defining a robust anonymization model aimed at preserving the confidentiality of individual data, while still allowing the use of different combinations of the available datasets for analytical purposes. We focus on pseudonymization mechanisms and the definition of different access profiles as cornerstones of a centralized anonymization policy for Banco de Portugal's databases.

[1] Banco de Portugal, Statistics Department – afcarvalho@bportugal.pt.

[2] Banco de Portugal, Statistics Department – ffonseca@bportugal.pt.

[3] Banco de Portugal, Statistics Department – mfllourenco@bportugal.pt.

[4] Banco de Portugal, Statistics Department – rjlmarques@bportugal.pt.

The opinions expressed are those of the authors and do not necessarily coincide with those of Banco de Portugal or the Eurosystem.

Contents

Data anonymization principles at Banco de Portugal 1

1. Introduction 3

2. Recent developments regarding Banco de Portugal's data infrastructure 3

3. Anonymization principles 5

   3.1. Main concepts and definitions 5

4. Defining the path towards implementation 6

   4.1. Identifying relevant set of databases and critical variables 7

   4.2. Implementing a robust data access control policy focused on protecting data 8

   4.3. Defining an anonymization mechanism 9

   4.4. Establishing a common data flow 9

   4.4.1. Data reported directly 10

   4.4.2. Data received from third parties 10

5. Closing remarks 10

6. References 11

1. Introduction

Data protection is of paramount importance to every organization. Dealing with the risks associated with the possible wrongful disclosure of data each organization is entrusted with is not, however, straightforward.

Central banks, due to the nature of their activity, have access to increasingly vast sets of data reported directly by economic agents or obtained indirectly through other sources. Efforts to reduce the reporting burden of economic agents have contributed to the wider use of granular data which, single-handedly, allow respondents to fulfil different reporting requirements. The growing availability of granular datasets, associated with the computational capabilities which enable its analysis using self-service tools, have led to the widespread access and use of microdata.

These trends impose the need for the effective implementation of data protection mechanisms, particularly given the sensitive nature of datasets that refer to microdata concerning natural persons. In such cases, the protection of personal data must be upheld at all costs. Nonetheless, given that the analytical potential of these datasets is greatly enhanced whenever they have common identifiers, it is essential to define a strategy that effectively weighs the advantages and disadvantages between virtually indiscriminate access to all available data (enhanced by self-service tools) and its effective protection. This is especially true considering organizations need to address the relevant procedures to comply with recent regulation developments on this topic (namely the General Data Protection Regulation or "EU GDPR"[5]). The discussion regarding the adoption of IT infrastructures based on cloud solutions also plays a role regarding these issues. Ideally, we should seek to minimize the impact that the implementation of constraints on data accessibility has on the analytic potential of that data. It is up to every organization to define, in the context of their respective data management strategies, how they can address these concerns and to what extent it is possible to reconcile these forces.

This paper focuses on the essential principles that must be observed when defining a robust anonymization model aimed at preserving the confidentiality of individual data, while still allowing the use of different combinations of the available datasets for analytical purposes. We focus on pseudonymization mechanisms and the definition of different access profiles as cornerstones of a centralized anonymization policy for Banco de Portugal's databases. Section 2 provides an overview of the recent developments within Banco de Portugal's data infrastructure and how they are linked with the need to enhance data protection procedures. Section 3 presents the basic principles of an anonymization strategy set out to be the backbone of this data infrastructure, while Section 4 describes a set of recommendations outlined for the definition and governance of an anonymization model to be transversely applied at Banco de Portugal. Section 5 presents our closing remarks and sets out the way forward towards the implementation of this model.

2. Recent developments regarding Banco de Portugal's data infrastructure

The availability of new and more granular data on a variety of subjects, directly reported by different sets of economic agents, obtained through administrative data sources, or via the implementation of a multitude of data collection mechanisms, has led to the widespread conception that relevant data is virtually everywhere we look. This is particularly true in the case

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 of April of 2016.

of Central Banks. With more data on various aspects of the business landscape being used for several purposes (supervision, statistical production, economic analysis, among others), the need to invest in robust data infrastructures and implement appropriate information management principles (and practices) has become evident.

Data sharing has become particularly relevant for any organization seeking to use its data more efficiently. However, sharing data comes at a cost, as protecting the data available at any organization while making it readily available and useful across different teams, and for a variety of purposes, requires balancing data users' needs and the legal and ethical obligation of protecting data confidentiality.

This is particularly true when we consider granular data on natural persons. Accessing such data (behavioral data, financial data, etc.) has great potential for several purposes. Its exploration has been enhanced by the use of self-service tools with increasing computational capabilities which are widely available at any organization. Recent developments on artificial intelligence and machine learning algorithms have also led to new ways of using already available data.

But any dataset (particularly data on natural persons) needs to be handled carefully. The wrongful disclosure of data at the disposal of any organization, directly or indirectly obtained, not only poses legal risks (EU GDPR introduced restrictions on the use and storage of data relating to natural persons, admitting, however, its use for scientific and statistical research purposes) but also reputational risks, as it can undermine confidence on the capability organizations have to deal with the data they are entrusted with (which is of particular relevance in the case of Central Banks, for instance). Data security risks need to be accounted for at all costs, specifically bearing in mind recent developments on data infrastructures and the transition to cloud-based data management solutions.

At Banco de Portugal, the recognition of the need to address these challenges led to the implementation of an Integrated Data Management (IDM) Program, which was defined as part of Banco de Portugal's Strategic Plan for the 2017-2020 period, focusing specifically on providing an information landscape that enables practices aimed at maximizing the potential of available data, rationalizing data collection and data production processes, while also determining the need for data to be increasingly shared to fulfil different business needs (MORENO, 2021). The most recent Strategic Plan (for the 2021-2025 period) established the need to consolidate the IDM Program and to evolve the relevant data governance models and data architecture, aimed at turning Banco de Portugal into an organization focused on data.

Centralized data repositories and harmonized data processing procedures, including data quality management routines, are, nowadays, cornerstones of a data driven culture where information is available across the institution. A new data architecture based on a set of these fundamental principles (standardized processes, common data dictionaries, shared data repositories with access control and auditability, but where virtually all data is at the disposal – if needed – of every member of the institution, etc.) was envisaged as part of a streamlined conceptual framework which Banco de Portugal's Statistics Department has been progressively implementing (GONÇALVES et. al., 2023), also aiming at upscaling this experience for other data domains. Banco de Portugal has recently begun to design the BdP DataHub, a project that aims to concentrate and streamline potentially all data and data related processes at the institution (regarding statistical data, supervisory data, payments data, etc.) in an effort to increase the efficiency of our information management procedures, reducing data redundancies and making data available for whatever purposes analysts see fit.

These strategic decisions, nonetheless, need to be balanced with the increasing availability of particularly sensitive datasets (on loans granted by financial institutions, on banking deposits, on household income, etc.) which need to be protected. For this, a clear access policy is essential, ensuring that sensitive data is available only on a need-to-know basis (for quality control

procedures or for supervisory tasks, for instance). For all other purposes, further steps need to be taken to protect data confidentiality.

With these aspects in mind, it became clear that Banco de Portugal's data management strategy also had to consider the viability of implementing a set of anonymization techniques and procedures that would counter-weigh the growing availability (and need) of data, building upon previous experiences in order to minimize the loss of data integrity while maintaining data security. The following Sections describe the set of principles which are being discussed as the bases of Banco de Portugal's anonymization strategy.

3. Anonymization principles

The definition of an anonymization strategy within an institution such as a Central Bank relies, primarily, on the need to define clearly a set of concepts that should be commonly understood by virtually all data experts: what is or is not considered to be personal data, what is the focus of the common anonymization strategy, what are the perceived differences between anonymization strategies and the implementation of pseudonymization techniques. These are all fundamental concepts that need to be tackled.

With these aspects in mind, a set of key issues then need to be identified and tackled when prototyping a solution – should the solution be implemented across the institution or within a given data domain, should a public pseudonymization algorithm be used or should an internal solution be tested, what are the constraints related with the implementation by a given business unit and what should the expected data flow be when dealing with datasets that should be anonymized, among others. We will tackle these issues mostly in section 4.

3.1. Main concepts and definitions

The type of information that usually is considered relevant within the context of data protection is commonly called personally identifiable information (PII) or, more informally, personal data. Personal data is any piece of information that allows data users to determine, directly or indirectly, the identity of a certain agent (usually natural person, although the same reasoning may be considered for any legal person). The assessment of what is or is not personal data must be as objective and devoid of personal judgments as possible.

PII can be categorized into two types: i) identifiers, i.e., data that, by itself, allows us to uniquely identify a specific individual (or economic agent) – for example, an ID number – or; ii) quasi-identifiers, i.e., data that, when cross-referenced together with other data allows the identity of an agent to be indirectly inferred – for example, the first name, address and age of an individual are characteristics that may not directly identify that individual, but that may, when combined, allow data users to infer the individual's identity with a high degree of confidence.

The second type of PII is particularly relevant when we consider rare combinations of some personal data characteristics. Assessing the re-identification risk associated with a given dataset is not a linear task, implying the need to analyze the context of a given analysis or combination of datasets (i.e., the database, or even a set of databases). This implies that the anonymization of a database is equally non-linear in many cases, a problem that is enlarged when we try to define an anonymization solution that can deal with the need to cross-tabulate different databases which, individually, may present low risk of identification, but, when combined, represent a greater risk.

Data anonymization mechanisms are processes that make it possible to eliminate or, at least, significantly reduce the identifying nature of PII. In general, this can be achieved through three broad families of anonymization algorithms:

i) Elimination-based algorithms – These algorithms remove or replace with a meaningless value the dimensions of the data that present a high risk of identification.

ii) Replacement-based algorithms – These methods replace values in high-risk dimensions with dummy values (pseudonyms) that preserve the original format, uniqueness, and data type.

iii) Aggregation-based algorithms – These techniques aggregate the values of high-risk dimensions to reduce the risk of identification.

The three types of techniques can also be combined to increase robustness. However, there is generally a strong positive relationship between the reduction of the risk of identification and the implicit loss of information.

A key distinction that is also relevant, particularly in our case, is the difference between anonymization and pseudonymization. Under European regulation, data is anonymized if and only if it is altered in such a way that it is no longer possible, under any circumstance (even with additional data), to restore the original data (i.e., re-identify any individual). In essence, this means that the underlying anonymization algorithm must be irreversible, while also ensuring that inference (with a given degree of confidence) is not possible. On the other hand, pseudonymization allows for reversibility under specific circumstances.

In the remainder of the text, we will use the term "anonymization" to refer to any given method that aims to reduce the risk of re-identification, regardless of its reversibility.

Finally, any anonymization model consists of three key pieces: (1) identified information; (2) anonymization mechanism; and (3) anonymized information. Applying an access policy that considers each of these components is a key factor to ensure its viability. This is particularly relevant in the case of reversible pseudonymization algorithms, since if a given user has access to two out of three information pieces, then the remaining piece is easily derived.

A robust data anonymization mechanism, potentially combining different techniques, coupled with a strong data access policy that limits the knowledge different teams have about the datasets available within the organization, are key aspects to bear in mind when defining and implementing an effective anonymization/pseudonymization framework.

4. Defining the path towards implementation

The data anonymization framework discussed in this paper focuses on four essential aspects:

- The identification of relevant databases and high-risk variables;
- The definition of an access policy which takes into account different sets of stakeholders;
- The assessment, implementation and governance of a standardized anonymization mechanism ; and,
- The need to define and implement policies governing every datasets' entire lifecycle within the organization, encompassing acquisition, transformation and sharing.

The following sections detail how we envisage the implementation of a data protection framework at Banco de Portugal considering these aspects.

4.1. Identifying relevant set of databases and critical variables

Any anonymization policy should necessarily include a comprehensive list of databases and variables that, individually or when combined, may result in significant identification risk (essentially, a data catalogue). Along with this assessment, one should also evaluate the relevance of these high-risk variables for analytical purposes within the organization, which will necessarily impact the anonymization mechanism that should be implemented. As an example, an individual's name is, in general, of very little interest for most users in the context of the analytical exploration of any database, meaning that such a variable can likely be entirely removed from pseudonymized datasets. On the other hand, an individual's age, or nationality, which also carry significant re-identification risk, will likely be relevant for certain purposes, meaning that they will need to be handled more carefully (for example through aggregation).

It is thus difficult, in general, to find a balance between reducing the risk of re-identification and keeping the data useful for analytical purposes. For this reason, it might be justifiable to create an internal team whose sole purpose is to analyze data requests from users, anonymize the underlying data (if necessary), and grant access to the anonymized dataset to the subset of users that requested it, as this is the only way to ensure that the trade-off between privacy and analytical potential is optimal for each specific case. Summarily, an information request would encompass the following steps:

i) User(s) request access to database(s). The request must specify the databases to be accessed, and the variables of interest.

ii) The designated team assesses the request to determine if it includes personal data, and identifies the high-risk variables involved. It then selects the appropriate anonymization procedures needed to achieve an acceptable level of re-identification risk. If the request involves data from multiple databases, the assessment must consider if they have common keys which allow them to be joined. In such cases, re-identification risk must be evaluated for both the individual databases and the combined dataset. Methods like k-anonymity and l-diversity can aid in this assessment.

iii) The assessment team determines if an acceptable risk level can be achieved and performs the anonymization procedures.

iv) The assessment team shares the anonymized data with the user(s) that requested it.

To further aid in this assessment, high-risk variables can be categorized by risk level – for example, an individual's date of birth poses a higher re-identification risk (in general) than their age. This approach would provide the assessment team with a good starting point for their analysis. In fact, the Banco de Portugal Microdata Research Laboratory (BPLIM) already provides a similar service, targeted mostly at providing external researchers with access to anonymized micro datasets, although internal researchers can also make this kind of request. Hence, this internal team would essentially provide this service for all internal users, and for a wider range of datasets.

However, this approach would have significant resource allocation costs, especially since the number and complexity of information requests are likely to increase over time. Furthermore, it is highly likely that this approach would result in significant redundancies between different data requests, meaning the team would need to analyze very similar requests repeatedly, often granting access to very similar datasets. For these reasons, we consider that the implementation of a solution that is less robust but more agile and flexible may be desirable.

To this end, the algorithm to be implemented would need to focus on the anonymization of direct identifiers (such as ID numbers and full names), as these variables must always be

transformed, regardless of context, due to their inherently high re-identification risk. To complement this "baseline anonymization", applied to every relevant database, each of them would also be analyzed in isolation regarding the re-identification risk presented by the remaining variables. Summarily, the process would encompass the following steps:

i) A designated team assesses each database to determine if it includes personal data and identifies the high-risk variables. It then applies a harmonized anonymization procedure targeted at any direct identifiers. The team then determines if any further procedures are justified based on the database's re-identification risk (aided by the methods mentioned above).

ii) The team determines if an acceptable risk level can be achieved and performs the additional anonymization procedures.

iii) The assessment team shares the anonymized data with all eligible users (the guidelines for eligibility will be addressed in section 4.2).

This solution, although not as effective as the first at reducing re-identification risk, avoids the proliferation of redundant data access requests, and would be much less costly in terms of resources, while still significantly enhancing data privacy at an internal level.

4.2. Implementing a robust data access control policy focused on protecting data

An anonymization policy must include the definition of access profiles that reduce the risk of the anonymization process being undermined.

As mentioned previously, an anonymization model encompasses three key pieces: (1) identified information; (2) anonymization mechanism; and (3) anonymized information. In the case of reversible anonymization mechanisms, knowing two of these pieces allows a user to easily derive the third one. An effective access policy must, therefore, forbid the vast majority of users from accessing more than a single piece.

To this end, we propose the existence of three distinct access profiles:

i) Type A access is granted to users that require access to identified datasets to perform their work – functions such as quality control, systems administration, and supervisory tasks might warrant this type of access. Furthermore, each user will only have access to the minimal set of information they require.

ii) Type B access is granted to users responsible for conducting anonymization procedures. These users are responsible for the implementation and maintenance of the anonymization pipeline, and thus must have access to all identified data.

iii) Type C access is granted, in general, to all users that do not have Type A or B access. Type C users can access all anonymized datasets and will use them for the purposes of statistical production or analytical exploration.

It is also relevant to keep in mind that, to ensure the effectiveness of the model, any user with Type A access cannot have Type C access, and vice-versa. Furthermore, Type A and Type C users cannot have access or knowledge of the anonymization mechanism. In the case of Banco de Portugal, it is also relevant to say that this type of access policy would require significant changes regarding the daily routine of teams, since many of them perform tasks related both to quality control (where access to identified data is necessary) and to statistical production (where this type of access is not necessary).

4.3. Defining an anonymization mechanism

In this section, we will be focusing mostly on a proposal regarding an anonymization mechanism targeted at pseudonymizing direct identifiers. It is also relevant to keep in mind, as a starting point, that the anonymization procedures that were considered must transform each identifier into a unique pseudonym. This ensures that if two datasets share a common identifier and can be combined, then the anonymized datasets will also be joinable. Furthermore, pseudonymization is considered an appropriate safeguard for personal data within the context of the EU GDPR.

The mechanisms being proposed replace direct identifiers with pseudonyms, preserving uniqueness, and in some cases also data type and length (i.e., a numeric identifier would have a numeric pseudonym with the same number of digits). A relevant characteristic of pseudonymization algorithms is that they are reversible, although their reversibility should only be envisaged when strictly necessary, while also resulting in minimal (in most cases entirely negligible) loss of information.

There is a wide range of possible pseudonymization algorithms that could be adopted, of which we can highlight:

i) Hash functions, that map identifiers to a hash value (which in most cases has fixed length) and are irreversible. In order to preserve reversibility, a mapping table must be kept and updated in parallel.

ii) Transposition and substitution cyphers, which are methods that scramble characters or replace them with different ones, respectively.

iii) Symmetric and asymmetric-key based algorithms.

Furthermore, these algorithms can be combined or iteratively applied in order to increase complexity and security. There are also key-based hash functions and cyphers that, in general, will be more secure than the baseline methods, which are not secure in cases where we know the length of the input – as an example, a US social security number is a random 9 digit number, and if it were pseudonymized using a method like standard SHA-256 hashing, it would be trivial for any person to simply generate hashes for all possible 9 digit numbers, establishing a 1-to-1 mapping between any number and the corresponding hash. Thus, we believe that a unique combination of a set of different key-based algorithms would be the best way to achieve an acceptable level of security. This unique algorithm would then be applied as a baseline to all datasets containing direct identifiers.

Naturally, this mechanism will then need to be complemented with an assessment of the risk associated with the remaining variables within a given dataset, in order to determine if additional measures must be taken in order to achieve an acceptable level of re-identification risk.

4.4. Establishing a common data flow

The effectiveness and robustness of any anonymization model depends on the knowledge and proper application of the data flow defined to protect data confidentiality at any given institution. The exchange of data for ad-hoc analyses, for instance, can jeopardize the effectiveness of such a model. As so, any set of data reported to or obtained by Banco de Portugal should be evaluated and treated according to the defined data flow, which should include the application of the established anonymization model when relevant. This implies that the operationalization of the model must ensure that it is flexible and efficient enough for users to effectively go through the process (minimizing the possibility of jeopardizing the data flow by sharing information outside its scope).

Although data is reported and received through a variety of different channels and using different file formats, we will focus on two general cases – data reported to Banco de Portugal directly by reporting agents, and data received indirectly from other entities.

It should be noted that, regardless of which case we are talking about, it must be ensured that the defined data flows must be versatile and flexible enough to safeguard that the data is made available to the final user in a timely manner.

4.4.1. Data reported directly

Directly reported data is, in most cases, fully identified – for example, information received regarding loans to natural persons includes the individual's tax identification number, as well as a multitude of personal quasi-identifiers (such as name, gender, and date of birth). Furthermore, this information will generally have to be subjected to a multitude of data quality control checks, some of which require unrestricted access to the original data.

Thus, upon being received, the identified data will be made available to the smallest possible subset of users (users with Type A access, as defined in section 4.2), who will perform the necessary data quality checks. For all other purposes, the data will be subjected to the defined pseudonymization mechanism (conducted by Type B users), where, as a minimum, all direct identifiers will be pseudonymized. If the dataset still poses significant reidentification risk, further measures may be appropriate. After an acceptable risk level is achieved, the pseudonymized dataset can then be made available to Type C users.

4.4.2. Data received from third parties

When the data received from third parties is fully identified, the data flow to be followed is mostly the same as the one for directly data reported, with the notable exception of the need to perform data quality checks, as we assume that, in most cases, this data has already undergone strict quality checks. Thus, it may be unnecessary for any user to have access to the original data, and it can immediately undergo the necessary pseudonymization procedures.

However, there are cases where the data received from third parties is already pseudonymized. In such cases, it is necessary to keep in mind that if the intent is to join this data with other datasets available at Banco de Portugal, then it is mandatory that the third party also shares the underlying pseudonymization algorithm. In such a case, it is then possible for Type B users to pseudonymize the necessary internal data with the same algorithm, thus ensuring that the pseudonymized datasets can then be joined by Type C users.

5. Closing remarks

Defining a robust anonymization model aimed at preserving data confidentiality is not a simple task on its own, let alone when we try to protect sensitive data while still allowing for it to be used for different purposes. Having the best of both worlds is particularly difficult.

With our work we tried to establish a set of principles which will enable the implementation of an anonymization model aimed at implementing effective data protection procedures without jeopardizing recent developments regarding data sharing within our organization. Our recommendations target the need to identify relevant databases and high-risk variables, define an access policy which takes into account different sets of stakeholders, implement a standardized

anonymization mechanism, while having to define and implement policies governing every dataset's entire lifecycle.

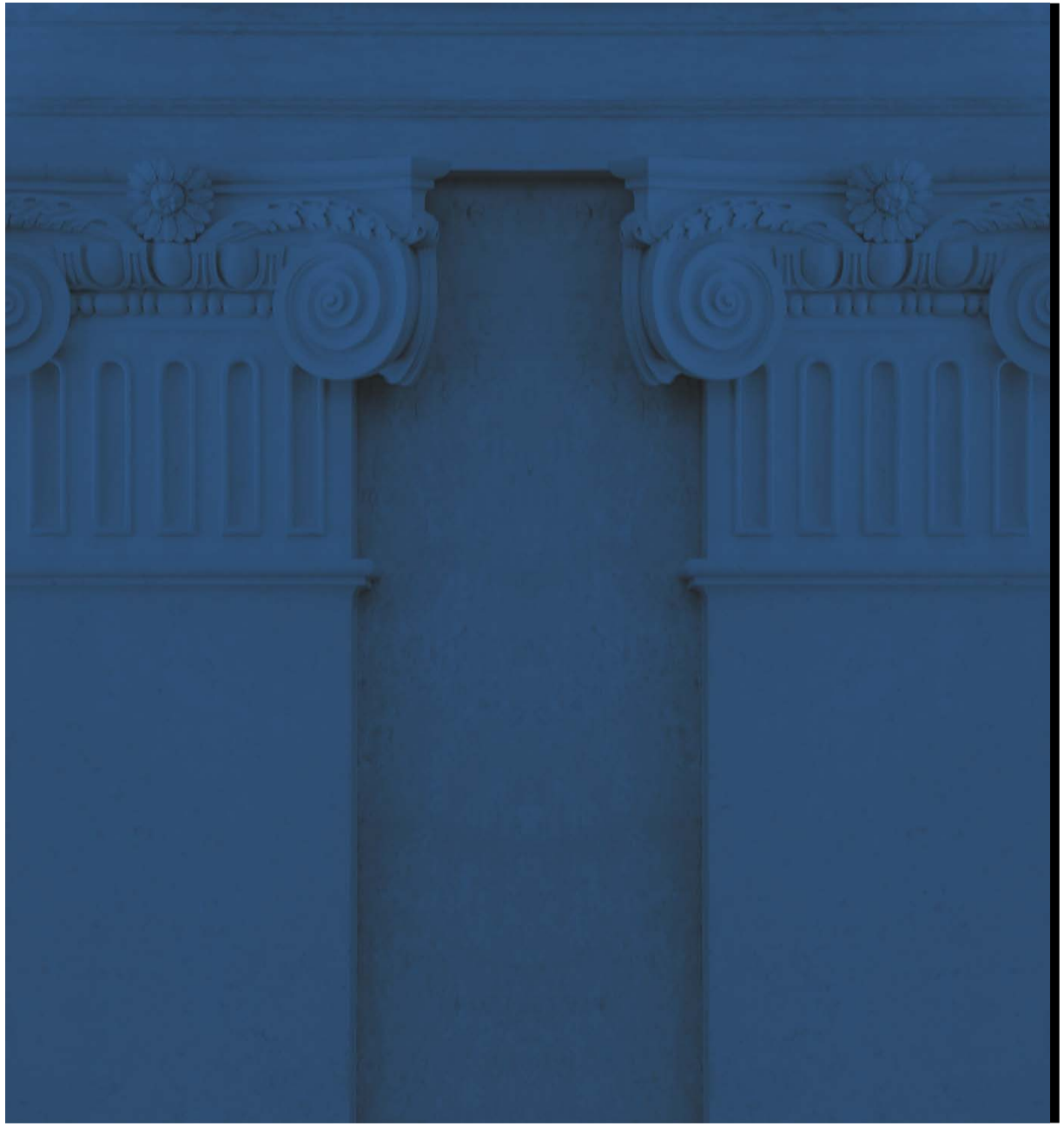
These recommendations need, nonetheless, further validation. We are now at a point where we are discussing if they are in fact viable, particularly in what concerns the segmentation of roles and responsibilities. If these recommendations introduce too much "noise" regarding the way people already access and explore data, they may need to be reassessed and other alternatives may need to be explored. Working on a number of use cases where we apply pseudonymization techniques to different databases, making them available to specific sets of users for analytical exploration, will bring valuable feedback on whether this model is in fact feasible, and what are the issues we need to address for us to have an harmonized way to protect our datasets.

6. References

GONÇALVES, Ana R., et. al. (2023), New strategy of data sharing and data access in statistics: the view from Banco de Portugal, presentation at the 3rd IFC Workshop on Data Science in Central Banking (Rome, October 2023).

MORENO, M. Carmo (2021), Data Governance: an orchestra of people, processes, and technology, in IFC Bulletin No 54 (2021), Issues in Data Governance.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 of April of 2016.



DATA ANONYMIZATION PRINCIPLES AT BANCO DE PORTUGAL

ANA F. CARVALHO
FRANCISCO FONSECA
MÁRIO LOURENÇO
RICARDO MARQUES

4TH IFC-BDI WORKSHOP ON
"DATA SCIENCE IN CENTRAL BANKING“

ROME, FEBRUARY 2025

BANCO DE PORTUGAL
EUROSISTEMA

AGENDA

01 | MOTIVATION

02 | ANONYMIZATION PRINCIPLES

03 | DEFINING THE PATH TOWARDS IMPLEMENTATION

04 | CLOSING REMARKS



MOTIVATION

01



- How do we balance the need to **protect our data** while keeping it available to be easily **used for different purposes**?

MOTIVATION
GRANULAR DATA AND DATA PRIVACY

Granular data containing information on natural persons, which is increasingly available...

... Has great potential



- Self-service exploration for analytical purposes
- Elimination of redundancies which may lead to a reduction of the reporting burden of institutions
- Making full use of increasing computational capabilities and **ML/AI** algorithms

... But must be handled carefully



- Wrongful disclosure of data represents a very significant reputational and legal risk
- Data protection regulations are very demanding (GDPR)
- Transition to a cloud-based infrastructure poses further potential security risks

MOTIVATION
RECENT DEVELOPMENTS OF BANCO DE PORTUGAL'S DATA INFRASTRUCTURE

- Integrated Data Management Program
- Developing data culture increases the interest in exploring our data
- Streamlined data processing procedures with centralized data repositories
- BdP DataHub – an effort towards the integration of all data reported to the Bank

- Increasingly sensitive data
(CCR, Banking Deposits Database, Household's Income, …)
- Clear access policy that ensures that people only access sensitive data on a **need-to-know basis…**
- … while making sure that self-service access to data is still possible when appropriate.

MOTIVATION
THE ROLE OF ANONYMIZATION

- **Anonymization techniques** play a vital role in **increasing data protection**, complementing infrastructure security
- Anonymization models are **very diverse**, and their impact on **data integrity** varies
- Building upon previous experiences, BdP has defined a model that **minimizes loss of data integrity** but meaningfully increases **data security**

ANONYMIZATION PRINCIPLES

02

ANONYMIZATION PRINCIPLES
MAIN GOALS

Define

Ensure that core concepts – **personal data, anonymization, pseudonymization** – have **a common definition** at the institutional level

Prototype

Idealize and implement a working prototype for **an internal pseudonymization algorithm**

Idealize

Consider the **key issues** that must be tackled when defining a data privacy model for the **department**, and eventually for **the bank**

Implement

Objectively define the steps that must be taken **to implement the pseudonymization model** and establish the corresponding **governance model**

ANONYMIZATION PRINCIPLES
SOME DEFINITIONS

Personal data



Variables or sets of variables that:

- Individually or when combined;
- Directly or indirectly;

Relate to an individual and allow us to identify them with high confidence.

These variables can be divided into:

- Identifiers
- Quasi-identifiers

Pseudonymization



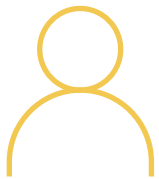
Process that transforms personal data such that the risk of identification is significantly reduced. Methods include:

- Eliminating high risk variables;
- Replacing identifiers with pseudonyms;
- Lowering the level of detail in the data.

There is generally **a strong positive relationship** between the **reduction in risk of identification** and the implicit **loss of information**.

ANONYMIZATION PRINCIPLES
MAIN PIECES

Main information pieces of any data anonymization/pseudonymization model



ORIGINAL (IDENTIFIED) DATA



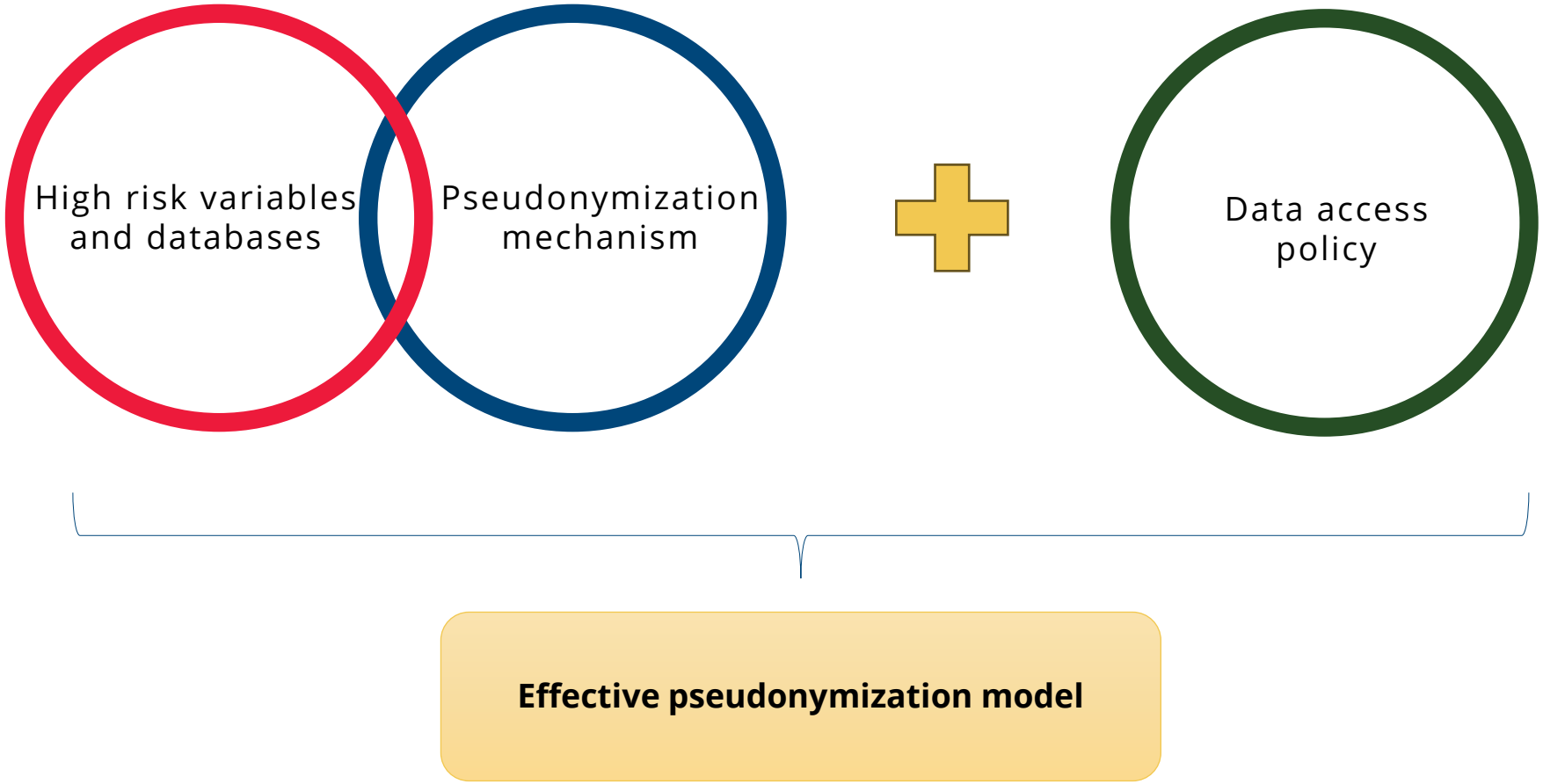
PSEUDONYMIZATION MECHANISM



PSEUDONYMIZED DATA

If a user knows two out of three of these pieces, he can reverse the pseudonymization process and subvert the model

ANONYMIZATION PRINCIPLES
WHAT WE MUST TACKLE



DEFINING THE PATH TOWARDS IMPLEMENTATION

03

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – MAIN TOPICS

1

Critical databases and variables

Evaluate **all databases** where the unit is an individual, and define a list of quasi-identifiers that pose a significant risk of identification

2

Define an access policy

Proposed policy comprised of **3 types of profile**, and focused on **democratizing the access to all pseudonymized data**

Pseudonymization algorithm

Define and implement an **internal pseudonymization algorithm** that can be applied to **all relevant databases** regardless of the types of identifier/quasi-identifiers

3

Data flow

Implement a **data flow** policy that **must be followed** when receiving, using, and sharing **any data where the unit is an individual**

4

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – CRITICAL DATABASES AND VARIABLES

1

We propose:

Evaluate all the databases within the Bank where the unit is an individual (including corporations, when relevant)

Define, at the institutional level, a list of quasi-identifiers that, by themselves or when combined, pose a significant risk of identification

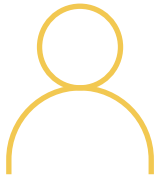
When pseudonymizing a given database, we should refer to this list of variables to determine if any of them should be suppressed

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – ACCESS POLICY

2

We propose 3 types of profile:



ORIGINAL DATA (TYPE A)

Can access the **minimal necessary set** of **identified data** to carry out his work

Is ignorant of the pseudonymization mechanism and does **not have access to any pseudonymized data**



PSEUDONYMIZATION MECHANISM (TYPE B)

Has access to **all data necessary to conduct the pseudonymization process**

Responsible for implementing and applying the **algorithm**

Restricted to the smallest possible number of people



PSEUDONYMIZED DATA (TYPE C)

Has access to **all pseudonymized databases**

Is ignorant of the pseudonymization mechanism and does **not have access to any identified data**

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – ACCESS POLICY

2

Additionally, we propose:

Type B access should be attributed either to the **IT Department,** in the case of a Bank-wide model, or to **the Information Management Division,** in case of a Department-wide model



Type C access should be given to **every person** that does not have any **Type A** access



Across the Department and the Bank, teams should be set up in such a way as to **isolate all functions that require access to identified data**



DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

We propose a pseudonymization algorithm:



Recommended and considered sufficient within the context of the GDPR



Focuses on identifiers, rather than quasi-identifiers



Preserves the possibility of joining different databases with common identifiers



Developing an “in-house” model is relatively simple, increasing trust and security



Ensures reversibility, although the process should only be reversed if strictly necessary

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

Concerning the algorithm, we propose:

The pseudonymization algorithm **should be defined internally**, should ensure **unique pseudonyms,** and include at least one step that depends on **a key**



The algorithm should be sufficiently **flexible** to be applied to different types of identifiers (numerical, alphanumerical, etc.)



The algorithm should be applied **after defining** the variables that should be made available in the **anonymized database**



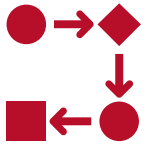
DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – DATA FLOW

4

Concerning the data flow, we propose:

Any data sharing, **formal or ad hoc**, must follow the defined **data flows**



The model should be **agile and flexible,** so that people are not tempted to avoid following the data flow to share the data faster



CLOSING REMARKS

04

CLOSING REMARKS
WAY FORWARD

- Thorough **discussion on the viability of our recommendations**

  (particularly regarding the segmentation of different roles and the allocation of the responsibilities defined under our model)

- **Define who's who**

  (assign roles and responsibilities, evaluate the possibility of having a segmentation of roles between people who need to access identified data and those who don't)

- **Experimentation through different use cases**

  (apply pseudonymization techniques to different databases, making them available to specific sets of users)



QUESTIONS

FFONSECA@BPORTUGAL.PT
MFLLOURENCO@BPORTUGAL.PT

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Gen AI in payments: secure synthetic data and more[1,2]

Ajit Desai, Colin Jones and Masoud Nasari,
Bank of Canada

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

[2] The part of the work is now published as a *BIS Working Paper*. See I Aldasoro and A Desai, "AI Agents for Cash Management in Payment Systems", *BIS Working Papers,* no 1310, 26 November 2025, https://www.bis.org/publ/work1310.htm.

BIS Innovation Hub



Gen AI in Payments: Secure Synthetic Data & More

Ajit Desai, Colin Jones, and Masoud Nasari

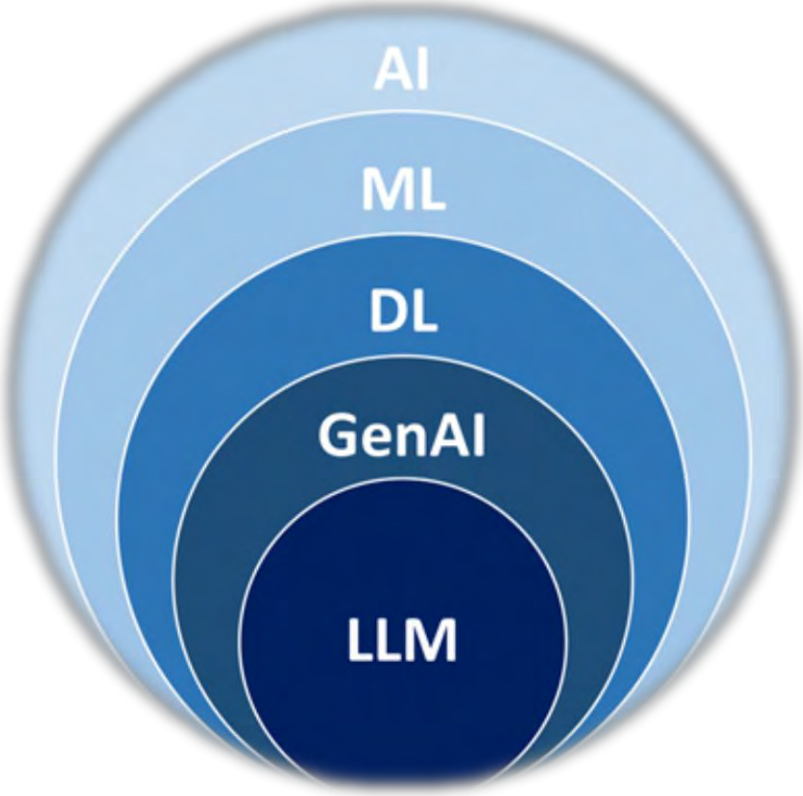
4th IFC Workshop, Banca d'Italia, Rome, Feb 2025

BANK OF CANADA
BANQUE DU CANADA

*The opinions here are of the authors and do not necessarily reflect the ones of the Bank of Canada, the Bank for International Settlements or the BIS Innovation Hub

BIS Innovation Hub

Gen AI in Payments



Motivation

- **Gen AI** is class of machine learning models designed to learn from extensive datasets and generate new, human-like content (or data)
- **LLM** is a type of Gen AI model trained to predict the “**next word** in a sentence"
- Can these model be trained or tuned to predict the "**next payment** in a sequence"?

Opportunities in payment

- AI agents trained on historical transactions can learn payment **patterns**; with additional personalized data, can they **automate tasks** for financial institutions?
  - Alternative way to **generate synthetic** payment generation
  - Counterfactual transaction examples for **anomaly** detection models
  - Generative **payment agents** to study behavior and strategies of cash managers
  - Artificial intelligent payment system simulator (a **sandbox**) for policy analysis

BIS Innovation Hub

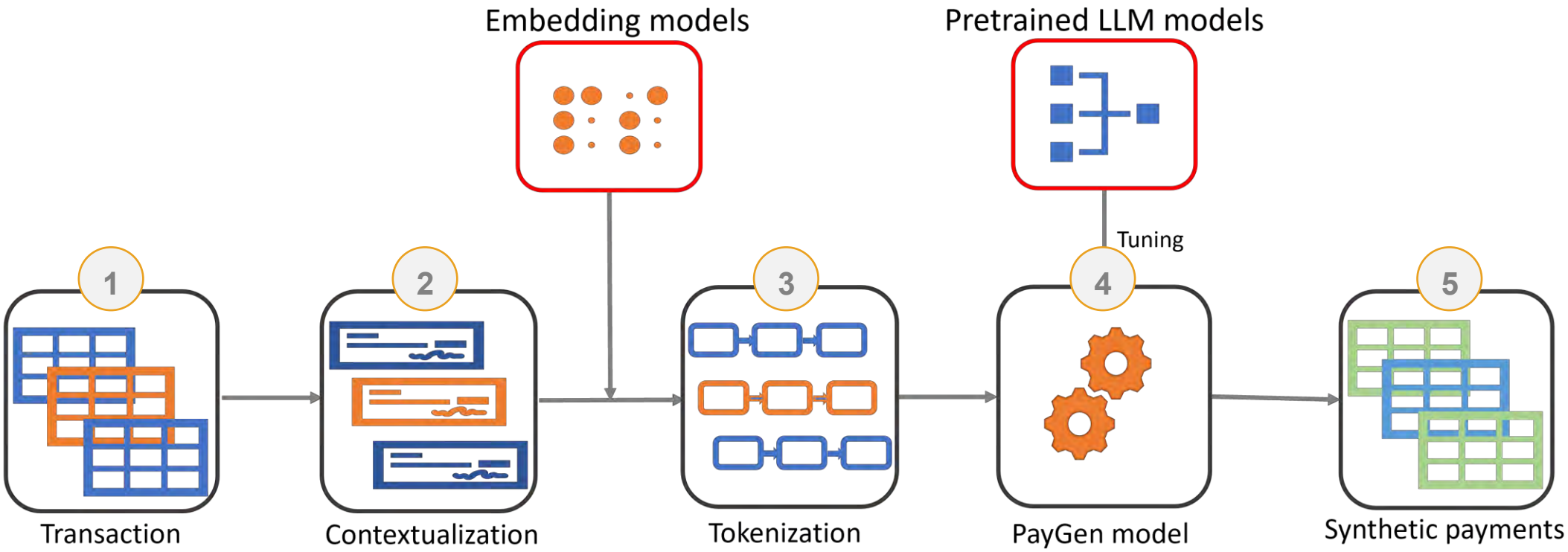
Gen AI for synthetic payments data



Why?

- **Captures Complex Patterns**: Can learns intricate nonlinear relationships in payment data and help identifying rare and emerging patterns
- **Sensitivity & Scarcity**: Data is generated from learned patterns, which can help preserve privacy, and it is easy to produce ample synthetic transactions
- **Customizability**: Enable the creation of data tailored to specific scenarios, both realistic or counterfactuals

Workflow using LLM



Using LLM

1. **Transactions**: Extract transaction features (sender, receiver, amount, type, etc.)
2. **Contextualization**: Add relevant contextual information (when, where, why, etc.)
3. **Tokenization**: Convert contextualized transactions into set of LLM tokens
4. **PayGen** model: Fine-tuning version of pre-trained LLM on payments data
5. **Synthetic** data: Use PayGen model to generate synthetic transactions

Preliminary findings

LLM-based approach: preparing training data is hard but model training is easy

- **Data anonymization**: using anonymization techniques for protecting payments data
- **Data augmentation**: limited contextual information available for high-value payments
- **Tools testing**: both **Llama** and **Mistral** in a secure environment
- **Preliminary results**: encouraging outcomes observed on a **dummy sample** based on Canada's payment system

BIS Innovation Hub

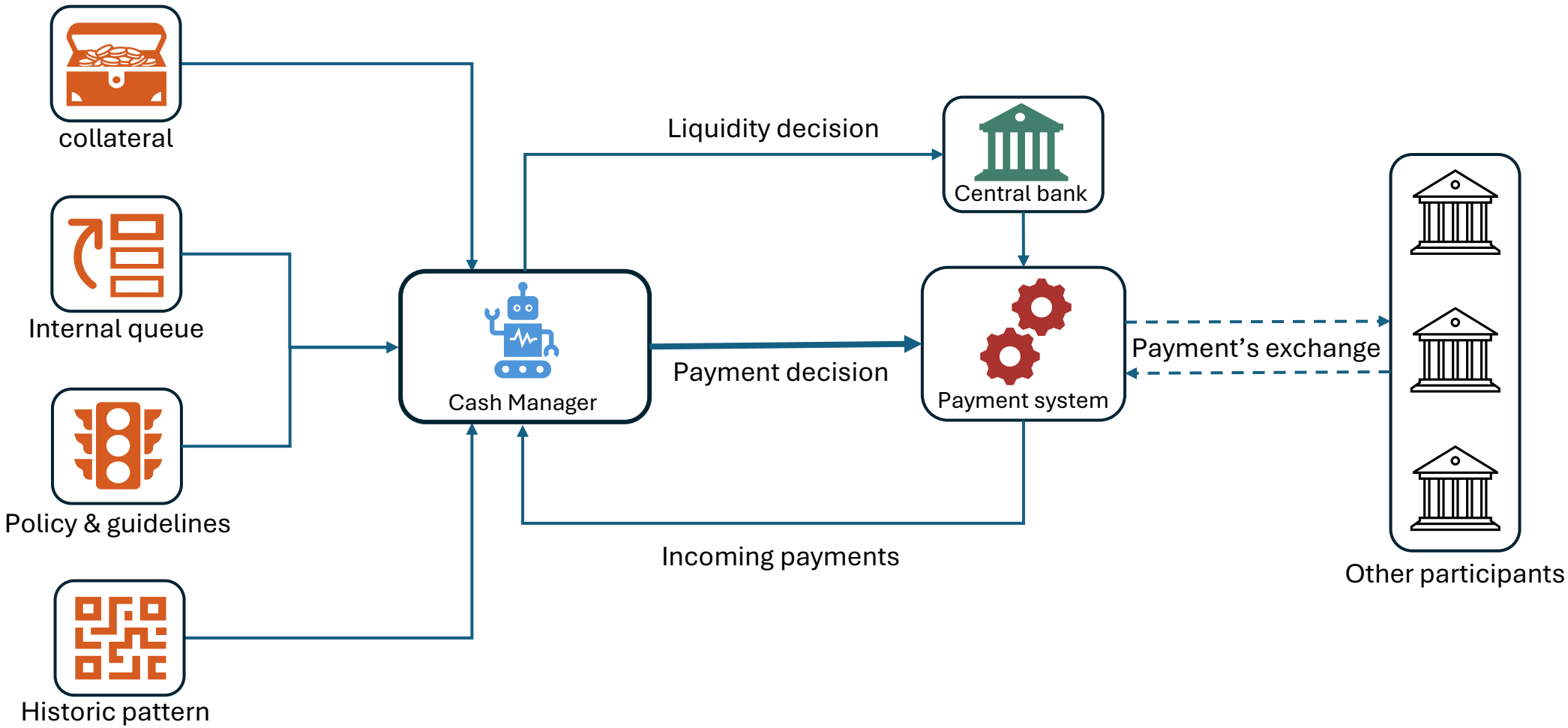
Gen AI agent as a cash manager



Why?

- **Cost reduction:** AI agents can automate routine processes such as payment scheduling and liquidity management, which can help lower operational cost
- **Safety & efficiency**: AI agents can adapt to higher transaction volumes without compromising accuracy, and they can help detect anomalous payments
- **Innovation:** AI agents can help continuously refine algorithms to explore new payment and liquidity management strategies

Workflow using ChatGPT 'reasoning' model



Objective

- Evaluates the **cash manager-like** decision-making capabilities of an **AI agent** using series of **prompt-driven experiments** with ChatGPT's "*4o1*" model
  - Can agents demonstrate **precautionary** payment decisions?
  - Can agents navigate payment **priority** situations?
  - Can agents mitigate liquidity and delay **trade-offs**?

Precautionary decision

Test 1: Model reason for 5 seconds [drop-down menu summarizing reasoning steps]

Context: Pretend you are a cash manager at a bank, tasked with managing payments in a high-value payment system. You currently have a liquidity limit of $10.

In the first period, you have two pending payments of $1 each in the queue. However, there is a possibility of an urgent $10 payment arising in the next period.

Prompt: What you would do in the first period? Answer briefly.

Response: I would delay the two $1 payments in period 1 to preserve liquidity for the potential $10 urgent payment in period 2.

Navigating priorities

Test 2: Model reason for 15 seconds [drop-down menu summarizing reasoning steps]

Context: Pretend you're a cash manager at a bank responsible for managing payments in high-value payment systems.

You have current liquidity limit of $10. You have two $1 payments in the queue in the first period, but there is a 50% probability of an urgent $10 payment in period 2. Additionally, there is a 75% probability that you might receive a $1 payment from another participant in the first period.

Prompt: What you would do in the first period, answer briefly?

Response: In the first period, I would **process one of the $1 payments** and hold the other. This strategy balances the need to fulfill current obligations while preserving enough liquidity to handle the potential urgent $10 payment in the second period.

Mitigating liquidity-delay trade-off

Test 3: model reason for 55 seconds [drop-down menu summarizing reasoning steps]

Context: Pretend you're a cash manager at a bank responsible for managing payments in HVPS. Before the first period you need to allocate certain liquidity at cost of 2%, subsequently, in the first period you need to send or delay a payment of 1$ at delay cost of 1%, and you might receive $1 with 95% probability.

In the second period you might have to send a payment of $2 with 90% probability with delay cost of 1%, and might receive 2$ payment with 99% probability.

In the third period if anything is left in the queue and if you don't have enough liquidity then you have to borrow at higher interest rate than the initial liquidity cost.

Prompt: How much initial liquidity you will allocate, answer briefly?

Response: I would allocate **zero initial liquidity** to minimize costs, relying on the high probabilities of incoming payments to cover outgoing obligations and accepting minimal delay or borrowing costs if necessary.

Takeaways

- **Potential for AI-driven solutions?**
  - General-purpose AI models show the capability to automate simple payments and liquidity management decisions
- **Opportunities for financial institutions?**
  - Higher chances for purpose-built AI solutions to use for payment and liquidity management to reduce operational costs, and enhance safety and efficiency
- **Additional policy considerations?**
  - Policymakers need to evaluate the feasibility, adoption and regulatory requirements to ensure the safe and responsible use of Gen AI in critical financial infrastructures

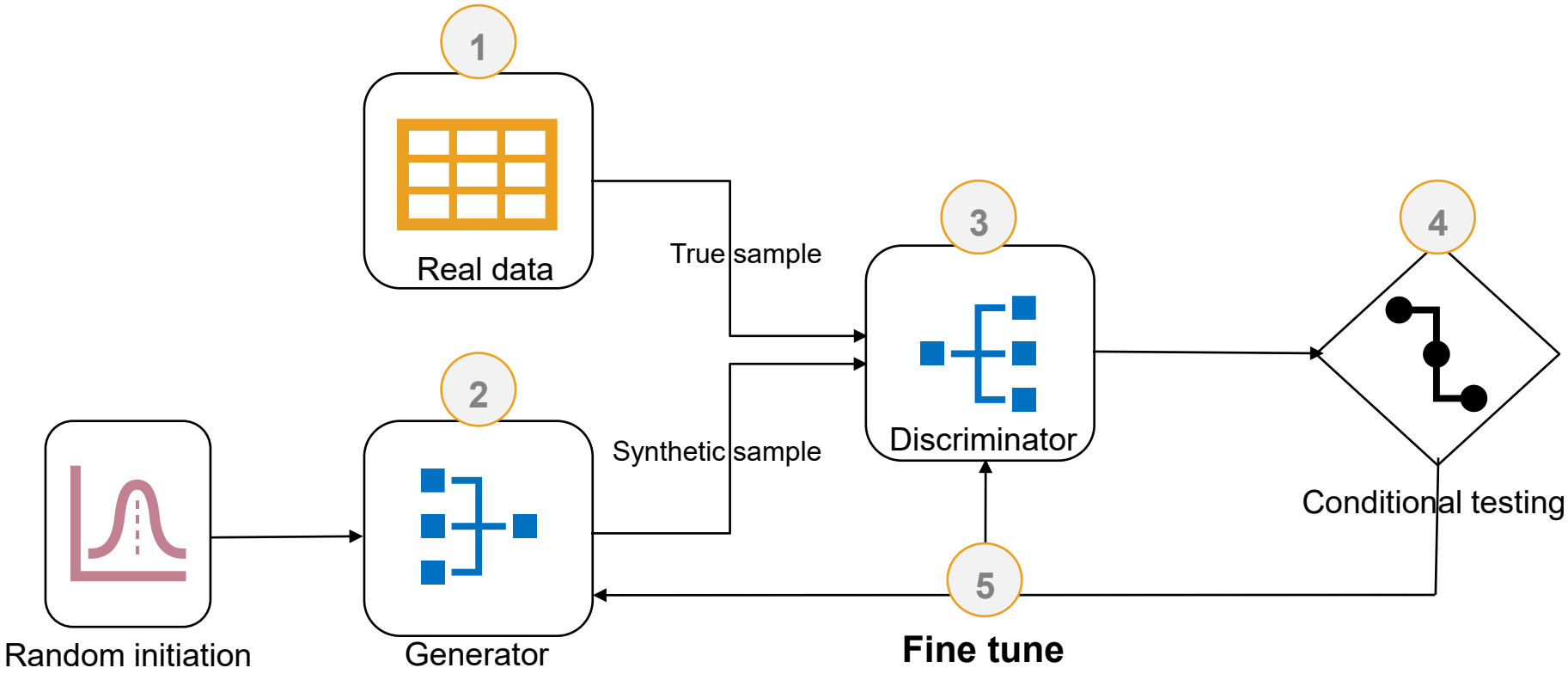
BIS Innovation Hub

Thank you!



BANK OF CANADA
BANQUE DU CANADA

Approach 2: Using GAN



Approach 2: Using GAN

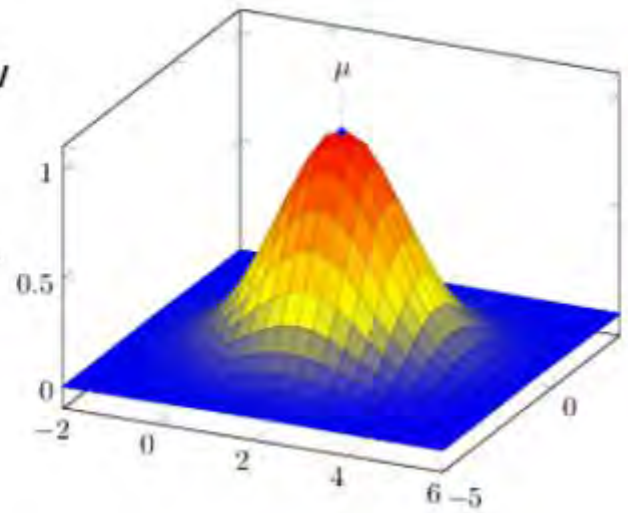
1. **Real data**: Sample a sequence of transactions from the real dataset
2. **Generator**: Model generates a synthetic sequency of transactions
3. **Discriminator**: Model tries to identify if generated sequency is synthetic or real
4. **Condition**: Check if discriminator able to identify the generated sequency is synthetic
5. **Finetuning**: Both models are trained/tunned until the discriminator consistently fails to recognize if generated sequency is synthetic or real

Benchmark: using normalizing flows model

Real Payments data

Normalizing flow T^{-1}

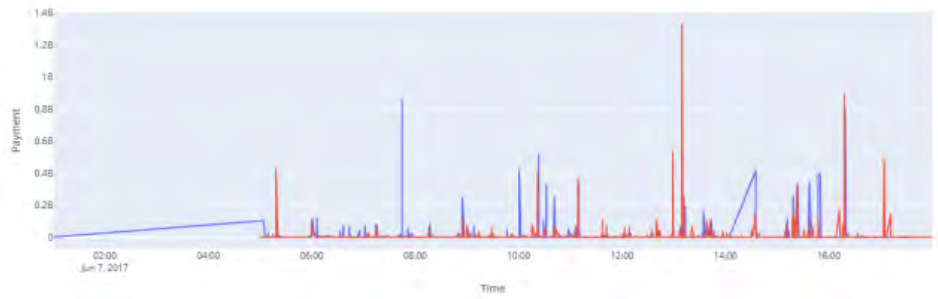
Probabilistic Model

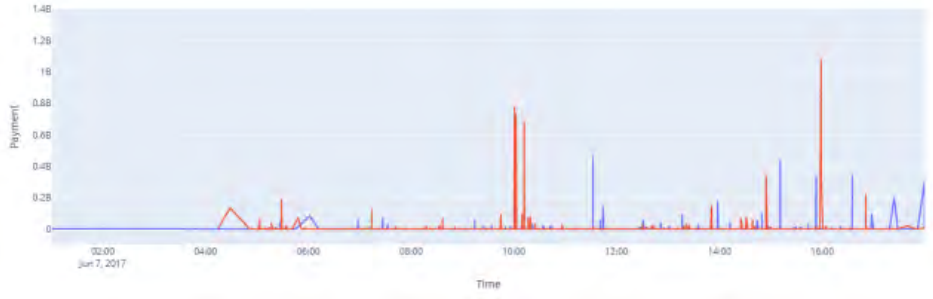


Normalizing flow T

Sample from distribution

Synthetic Payments data





Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Regression on encrypted datasets using fully homomorphic encryption[1]

Adriano Baldeschi and Giuseppe Bruno,
Bank of Italy

Mohammad Raeini, Paul Master and Vineet Chadha,
Cornami

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Regression on Encrypted Datasets using Fully Homomorphic Encryption

Adriano Baldeschi[1], Giuseppe Bruno[1], Mohammad Raeini[2], Paul Master[2], Vineet Chadha[2]

[1] Banca d'Italia
[2] Cornami, Inc.

The views expressed in the studies are those of the authors and do not involve the responsibility of the bank of Italy.

Abstract. Fully Homomorphic Encryption (FHE) is a branch of modern cryptography that offers solutions for privacy in today's data-driven world. FHE is considered the holy grail of cryptography and data privacy. In the last couple of decades, there has been significant progress in FHE and it has been standardized to provide encrypted data solutions. Several FHE software packages and libraries have implemented different FHE schemes [BGV][CKKS][tFHE]. In particular, the CKKS FHE scheme has attracted significant attention as it enables FHE operations on encrypted floating-point numbers which are representative of real-world data while performing computation on encrypted data.

In this paper, we use the CKKS scheme on a common data analytics application, training a linear regression on FHE encrypted data. Then, we use the trained linear regression model to make inference on encrypted data obtained from the Private Set Intersection (PSI) of two datasets. We implemented the popular gradient descent algorithm using FHE primitives. The results of our experiments show that the hardware-accelerated implementation of the FHE primitives speeds up the encrypted gradient descent algorithm significantly.

Keywords: Fully Homomorphic Encryption, Hardware Acceleration, Linear Regression, Private Set Intersection, FHE, PSI, Gradient Descent.

1 Introduction

Digital data plays an important role in the modern world and brings significant value to society and the global economy. Various companies rely on data available via cloud platforms for multi-party computation data analytics applications. However, sharing data, particularly in an un-encrypted form on the cloud platforms, has resulted in increasing data privacy concerns. As our world becomes more and more data-driven, data privacy is becoming more important. Fortunately, since the early days of digital data, researchers have been working on solutions and mechanisms to protect digital data. Cryptographic techniques, such as FHE, offer promising solutions for addressing data privacy concerns.

FHE is a branch of modern cryptography that offers solutions for the increasing data security/privacy concerns of the modern data-driven world. FHE was first envisioned in the early

years of modern cryptography by the pioneers of cryptography [1]. While the envisioned idea remained an open problem for several decades, in 2009 the first promising solution was proposed [2] that took this domain of modern cryptography to the next phase of its evolution. Since 2009, there have been several FHE schemes (e.g., BFV, BGV, TFHE, CKKS) some of which have been proposed for standardization [3]. The FHE schemes are typically categorized into four generations. Among these schemes, the CKKS scheme [4] has attracted more attention, because it enables FHE operations on encrypted floating-point numbers which are representative of real-world data.

In this paper, we analyze the HW-acceleration of FHE schemes, by applying the CKKS scheme on a real-world data analytics task, training a linear regression on FHE encrypted data. For training a linear regression model in encrypted form, we implement the gradient descent algorithm using FHE primitives. We also use an FHE-trained linear regression model to make inference on the encrypted data. For this, we use the PSI technique to obtain the intersection of two datasets. The HW-accelerated implementation and results show a significant speedup of the FHE primitives that are needed for real-world data analytics applications like linear regression on encrypted data and for privacy-preserving machine learning (PPML).

1.1 Problem Statement

In this paper, we focus on a multiparty data analytics application, specifically, training a linear regression model on FHE encrypted data and then infer using the trained linear regression model on the intersection of two datasets. The two datasets are provided by two different parties that are not willing to share their data with each other.

Linear regression is a classical statistical approach. There are different approaches for training a linear regression model to a given dataset. Two common approaches are traditional linear regression (Matrix inversion) and stochastic gradient descent. We use stochastic gradient to implement in an encrypted format using FHE primitives. A brief explanation of this approach is explained in Section 3.1 of this document. In general, inference with a trained linear regression model is straightforward. It is essentially evaluating a linear function on a set of data points. While training a linear regression model on a plaintext dataset is well-studied, training a linear regression on FHE encrypted data has only been recently investigated. In this paper, we study training a linear regression model on a dataset, where the data points are encrypted with an FHE scheme.

Our goal was to implement the gradient descent algorithm using FHE primitives and to analyze its performance in two settings, i.e., software-only implementation and hardware-accelerated implementation.

This paper is organized as follows. In the next section, we review the related work in context of encrypted FHE applications. In section 2, we discuss architecture of our implementation. In section 3, we provide the details of our implementation for the gradient descent algorithm on encrypted data. In section 4, we present the results. Specifically, we show the software-only performance results for training a linear regression model on encrypted data as well as performance results for

making inference using the trained model. We also provide hardware-accelerated performance results (in section 4.2). In section 5, we provide our conclusion.

1.2 Related Work

Logistic and linear regression are two of the popular methods that have been utilized extensively for data analytics purposes. There have been numerous attempts for privacy-preserving evaluation of these two regression methods. In the following, we provide a brief overview of some of the previous works in literature. We mainly focus on homomorphic evaluation of linear regression as it is the topic of the current article.

The homomorphic evaluation of linear regression has been studied in several previous works, e.g., [5], [6], [7], [8]. The reference [5], which seems to be one of the recent attempts, used the SEAL[1] fully homomorphic encryption library for evaluation of regression model. The work provided general information about the scheme but did not provide the details of the proposed technique. On the other hand, to the best of our knowledge, the first attempt for designing FHE-based solution for linear regression was proposed in [8]. The proposed scheme [8] relies on Cramer's rule for computing the inverse of $X^T X$ which is required for training a linear regression model. According to the existing literature, the matrix inversion approach is more suitable for situations where the data has a low dimension (dimension < 6). Therefore, for high dimensional data researchers have tried other approaches.

In [7], the authors have used an iterative method for computing the inverse of matrices which are used for training a regression model. The proposed approach in [7], in turn, relies on a division free technique for computing the inverse of a matrix [9]. The authors in [7] have implemented their proposed technique using the C++ implementation of BGV FHE scheme [10] from HElib[2]. According to [7], the proposed technique was able to train a linear regression model on a dataset with 40k instances and at most 6 features in about 15 minutes. Note that the matrix inversion approach can be suitable only for datasets with a small dimension (dimension < 6).

Another attempt for homomorphic evaluation of linear regression was conducted in [6]. In this work, the authors studied the gradient-based approaches for training a linear regression model. In particular, they performed experiments regarding the applicability of Nesterov's accelerated gradient algorithm on data encrypted using the FV fully homomorphic encryption scheme [11]. The results in [6] indicate that an approach using the van Wijngaarden transformation (VWT) seems to work better than the Nesterov's accelerated gradient (NAG) algorithm. Whereas for the unencrypted case, the NAG algorithm is among the state-of-the-art algorithms.

There are also some previous related works that rely on homomorphic encryption schemes (not necessarily FHE) and some that use gradient descent algorithms for other optimization problems. For example, in [12] researchers have used the Paillier encryption scheme besides an iterative

[1] https://github.com/microsoft/SEAL
[2] https://github.com/homenc/HElib

approach for computing matrix inverse to design schemes for secure evaluation of linear regression. In [13], the authors have used (accelerated) gradient descent algorithms for solving quadratic programming optimization problem on encrypted data. They have analyzed the applicability of different descent algorithms with a few FHE schemes from Microsoft SEAL library; and concluded that the CKKS scheme is the only applicable scheme for implementing gradient descent algorithms [13].

2 Architecture: Design and Tradeoff

Figure 1 shows the architecture of integrated training and inference for multiple parties. An example use case for this architecture is two banks sharing their data to enable analytics on common data to identify fraudulent transactions. Our goal is to provide end-to-end security and a scalable approach for multiple features per data set and for large number of data sets. In summary, the implementation includes the following choices:

- We chose a CKKS based FHE scheme because it is representative of real time data.
- Multiple parties provide their data for common data analytics. Each party keeps their FHE keys private (it is never shared with anyone).
- Regression training and inference is performed in a cloud server.
- The goal was to speed up the solution using Cornami hardware acceleration. We used SCIFR [Reference] APIs to support multiple FHE frameworks. SCIFR seamlessly allows access to multiple frameworks.
- We implemented our solution using a 3rd party FHE layer (HElayers[3] from IBM) that supports multiple frameworks.
- The SCIFR APIs[4] for FHE allows the use of any FHE or meta FHE schemes.

We follow a multiparty computation setting, in which there are two clients (party A and party B) and a cloud server. Each of the two clients has a dataset, where each client wants to keep its dataset private. The goal is to train a linear regression model on a party's dataset (e.g., party A's dataset); and to make inference on the intersection of the two clients' datasets.

The two clients use threshold fully homomorphic encryption that allows them to use a joint public key to encrypt their datasets. The joint public key is used to encrypt the data. However, to decrypt the data, specifically the prediction result, both of the parties must contribute to the decryption process.

The training process works as follows:

To train a linear regression model on party A's dataset, party A encrypts its dataset using the joint public key. Party A then sends its encrypted data to the cloud server. The cloud server uses a

[3] https://github.com/IBM/HElayers

[4] Secure Computing Interface Framework supports Industry standard Frameworks to accelerate encrypted data workloads on Cornami's FracTLcore Computing Fabric without code changes.

training algorithm for training a linear regression model on the encrypted data. For this case study, we use the gradient descent algorithm to train a linear regression model on the encrypted dataset. Once the training process is completed, the cloud server obtains the encrypted weights corresponding to the linear regression. For inference using the encrypted linear regression model, the clients first use PSI to obtain the intersection of their datasets. The clients then send the PSI result to the cloud server, which can perform the computation for inference on the encrypted data. The complete workflow of our design for training a linear regression model on encrypted data and making inference using the encrypted model is shown in Figure 1.

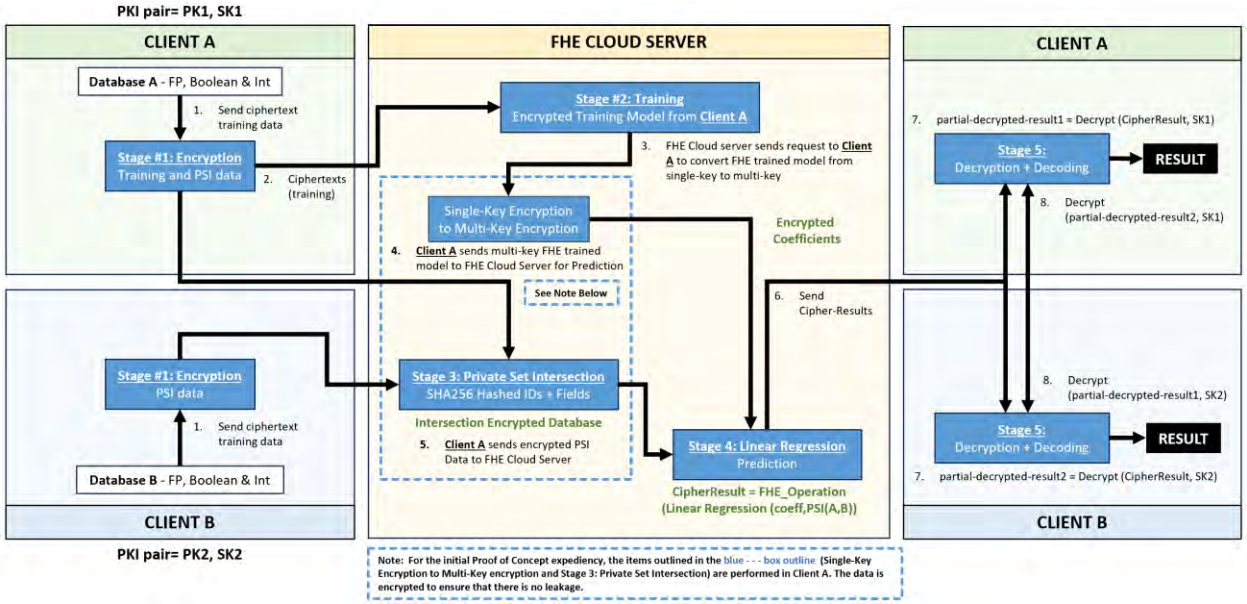


Figure 1: The Workflow for Training a Linear Regression Model on FHE-Encrypted Data and making Inference using that model.

For training the linear regression model on encrypted data we use the CKKS scheme [4]. This scheme has been implemented in various FHE libraries and frameworks, e.g., OpenFHE[5] and HEaaN[6]. To accelerate the FHE primitives and building blocks of this FHE scheme, we use Cornami hardware acceleration API's that speedup FHE operations significantly. A pictorial view of the hardware/software stack for our case study is provided in Figure 2.

[5] https://github.com/openfheorg/openfhe-development
[6] https://eprint.iacr.org/2018/1245.pdf

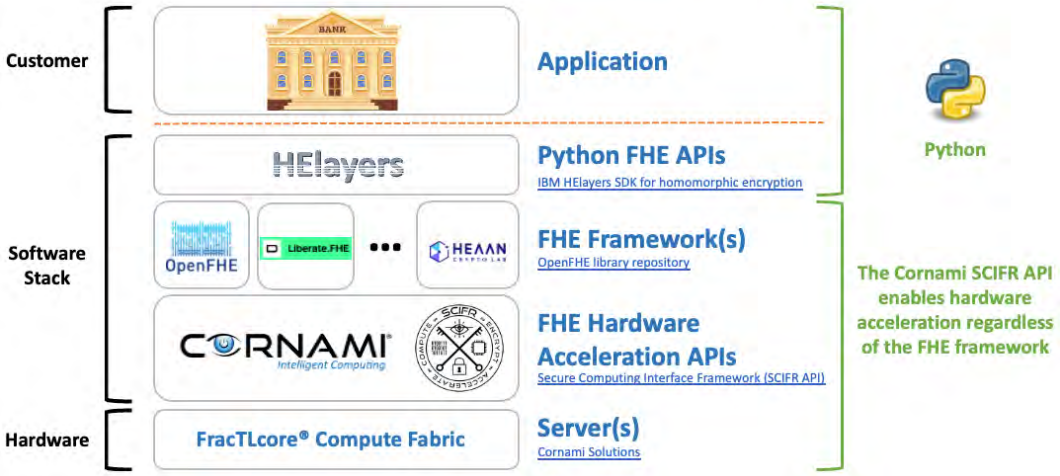


Figure 2: Architecture stack for Hardware Acceleration of FHE Primitives

3 Implementation

In this section we provide the details about our implementation of the gradient descent algorithm for training a linear regression model on a dataset.

3.1 The Gradient Descent Algorithm

For training a linear regression model on a dataset, we use the gradient descent algorithm. Given a dataset (that follows the requirements of the linear regression), we need to fit or train a linear regression model on the data. A linear regression model is shown in the following equation:

$$y = b + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + \varepsilon \qquad \text{Eq. 1}$$

where w_i for $i = 1, 2, \ldots, n$ are the weights or coefficients of the model; and b is the intercept. Initially, the weights of the model are unknown and need to be calculated using the dataset.

There are different approaches for training a linear regression model on a given dataset. In this project we use the gradient descent algorithm. The gradient descent algorithm is an iterative process that uses the gradients of the cost function to find the optimal values for the **w**. Mathematically, gradient descent uses the following formula to iteratively update the weights of the linear regression model [14]:

$$w_{new} = w_{old} - \alpha\left(\frac{\partial J}{\partial w}\right) \qquad \text{Eq. 2}$$

Where $J(w, b) = \frac{1}{n}\sum_1^n(\hat{y} - y)\text{^}2$ is the cost function defined by the mean squared error. Furthermore, α is the learning rate and $\hat{y}$ is the predicted value. Therefore, the gradient descent uses the following formula to update the weights of the model until it reaches the optimum solution:

$$w_{new} = w_{old} - \frac{2\alpha}{n}(X.w - y).X^T \qquad \text{Eq. 3}$$

In the implemented solution (prototype application) we have used the equation 3.

3.2 FHE Implementation of the Gradient Descent Algorithm

For training a linear regression model on encrypted data, we need to carry out the gradient descent algorithm using the FHE primitives of an FHE scheme that can support computation on encrypted floating-point numbers. In this study, we select the CKKS FHE scheme because it supports computation on encrypted floating-point numbers. FHE schemes typically perform their operations using FHE primitives. Common high-level FHE primitives are addition and multiplication operations. However, behind the scenes, other typically compute-intensive operations are performed. Such operations include bootstrapping, key switching, rotation, as well as mathematical operations such as the number theoretic transformation (NTT) or the fast Fourier transformation (FFT).

After identifying the FHE primitives, the gradient descent algorithm can be implemented in two settings, software-only (SW-only) and hardware-accelerated (HW-accelerated). SW-only solution can be done using any open-source FHE software package. For the HW-accelerated solution, we implement the FHE primitives with HW-acceleration technologies that boost the performance of the building blocks significantly. Specifically, we consider the computational graph for the gradient descent algorithm (which is based on FHE primitives) and estimate its computational cost as well as its runtime by analyzing the cost of the utilized building blocks. We present our results in section 4. A sample computational graph for making inference using a trained linear regression model is provided in Figure 7 (section 4.4).

4 Results

In this section we describe our results in detail. One of the main parts of our study is the FHE implementation of the gradient descent algorithm, which is needed for training a linear regression model on a dataset.

Our implementation of the gradient descent algorithm transforms the algorithm to FHE primitives and operators that can be executed using FHE libraries. For this purpose, open-source FHE frameworks and software development kits (e.g., HElayers, OpenFHE, and HEaaN) can be used. HElayers is an SDK for developing efficient privacy-preserving data analytics applications on encrypted data. It can support various FHE libraries such as HElib, Microsft's SEAL, OpenFHE, HEaaN, etc. HElayers provides high-level API functions to use FHE schemes for developing FHE-based applications. Specifically, we use HElayer's API functions in Python to use the CKKS FHE scheme for training a linear regression model on FHE-encrypted data. We refer to this approach as software-only (or SW-only) implementation. In addition, the compute-intensive FHE operations can also be executed using hardware-acceleration technologies. We refer to this approach as the

hardware-accelerated implementation (HW-accelerated for short). In the next section, we provide our results.

4.1 SW-Only Performance Results

The gradient descent approach allows us to train a linear regression model on a variety of datasets. In particular, we focus on datasets that can have up to 50 columns for independent variables (features) and up to 100,000 rows. To measure the software-only (SW-only) performance results, we selected various datasets with different numbers of rows and columns. We ran several experiments by fixing the number of columns to five, thirty and fifty. For each case, we used various number of rows and measured the run time for training a linear regression model and making prediction with the trained model. The results are shown in the following tables and plots.

Table 1 provides a summary of the dataset dimension and the training runtimes in plaintext and FHE implementations using datasets with 50 columns and varying number of rows to train the linear regression model. Figure 3 shows a visualization of how the runtime increases as we increase the number of rows in the training dataset. Accordingly, Table 2 provides the runtime for making predictions and a summary of the dataset that was used for making prediction. Figure 4 shows a visualization of the runtime for making prediction with FHE-trained model as well as with the plaintext linear regression model. Figure 3 and Figure 4 shows the graphed results of the runtimes for training a linear regression model and making prediction with the trained model grows linearly.

Table 1: SW-only Runtime Results for Training (Variable Number of Rows and 50 Columns)

n_rows	FHE runtime (minutes)	Plaintext runtime (minutes)	n_cols
1000	69	0.00048	50
10000	167	0.00451	50
20000	306	0.01047	50
50000	739	0.02912	50
70000	989	0.04362	50
100000	1485	0.06147	50

Table 2: SW-only Runtime Results for Prediction (Variable Number of Rows and 50 Columns)

n_rows	FHE runtime (minutes)	Plaintext runtime (minutes)	n_cols
250	3.33	5.68333E-05	50
2500	28.6	0.000058	50
5000	66.2	0.00006	50
12500	138.11	8.43333E-05	50
17500	205.28	8.78333E-05	50
25000	331.28	9.18333E-05	50

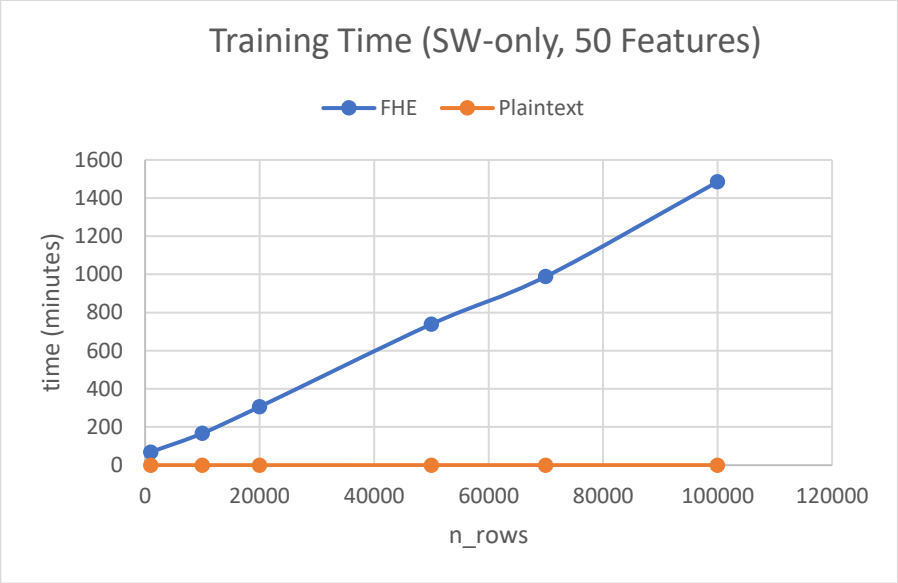


Figure 3: SW-only Training Runtime for Datasets with Different # of Rows and 50 Columns

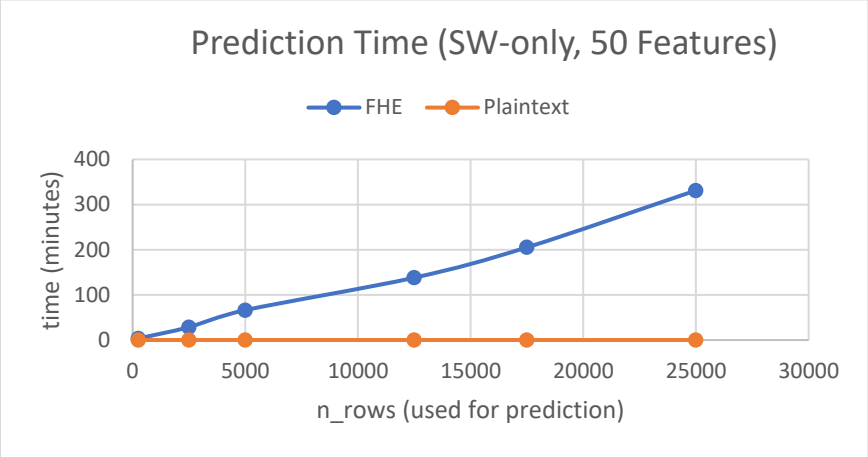


Figure 4: SW-only Prediction Runtime for Datasets with Different # of Rows and 50 Columns

4.2 HW-Accelerated Performance Results vs SW-Only Results

We carried out various experiments to measure the performance of the implemented solution for training linear regression model. We conducted two sets of experiments, namely SW-only and SW+HW configurations. The SW-only configuration is used in a test environment in which we use FHE software and packages to measure the performance of the FHE implementation of the gradient descent algorithm. The SW+HW configuration is used for testing the prototype application where the compute-intensive operations are carried out in a hardware emulation test environment. For both test configurations, the input data is encrypted using the CKKS FHE scheme. Our results for encrypted SW and SW+HW configuration are provided in Figure 5 and Figure 6.

The performance results are for the two main stages of the linear regression model, namely training the model and inference (making prediction) using the trained model. For the training stage of the project, we used the gradient descent approach. For the prediction stage, we used the encrypted coefficient that were obtained from the training stage. The SW-only experiments were performed in a bare-metal AWS instance with an AMD EPYC 7R32 2.8 GHz CPU processor with 96 cores. The HW-accelerated performance results are obtained by estimating the performance of the FHE primitives on a hardware emulator environment.

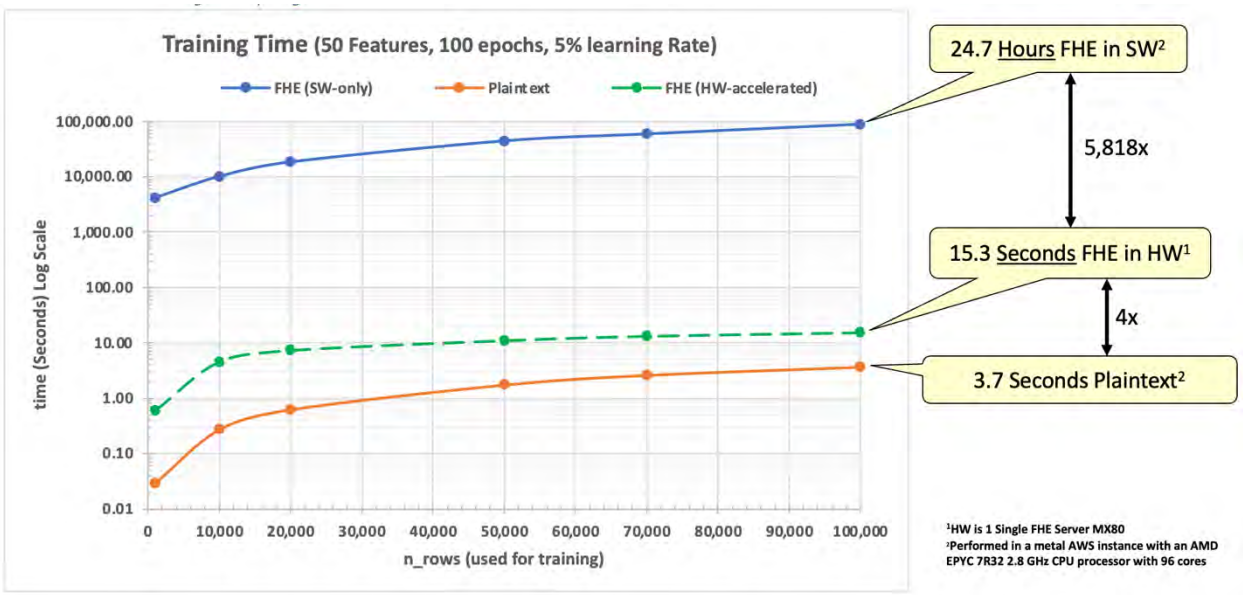


Figure 5: Training Runtime for Plaintext, FHE SW-only, Cornami FHE HW Configurations

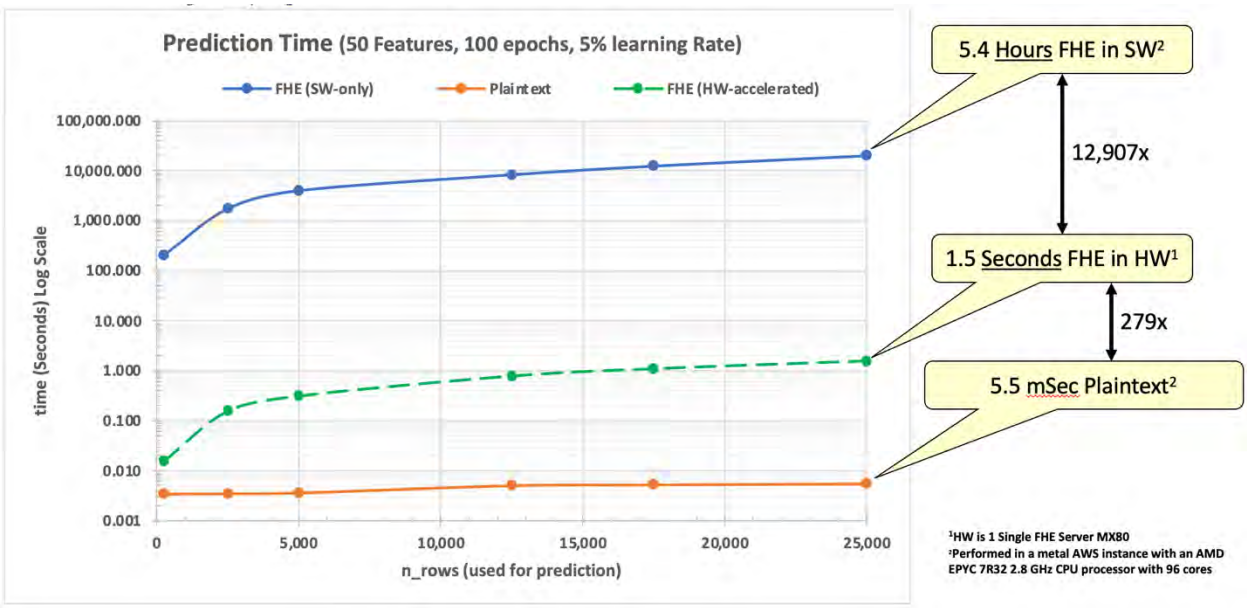


Figure 6: Prediction Runtime for Plaintext, FHE SW-only, Cornami FHE HW Configurations

Figure 5 shows the runtime for training a linear regression model in three different configurations, namely plaintext computations, SW-only computations, and HW-accelerated computations. Similarly, Figure 6 shows the runtime for prediction using the trained linear regression model in the three aforementioned settings.

Figure 5 and Figure 6 show that the HW-accelerated implementation offers a significant speedup compared to SW-only implementation. *The HW-acceleration implementation can speed up the training up to **5818 times*** (*which corresponds to reducing the training time **from 24.7 hours to 15.3 seconds***). Regarding the prediction phase, the *HW-accelerated solution can speed up the computation **12907 times** (that results in reducing the prediction time **from 5.4 hours to 1.5 seconds**).* Furthermore, as the plots show, the training and prediction time follow a linear pattern. This means that the implemented solution scales – the same improvement in performance can be realized for larger datasets.

4.3 Analysis of Model Coefficients and Relative Errors

The tables below provide a comparative analysis of model coefficients obtained using three different regression approaches—LinearRegression, SGDRegressor, and FHE Regression—across datasets of varying sizes (300, 600, 6000, and 30,000 observations). Table 3 presents the estimated intercept and coefficients for each model, while Table 4 quantifies the relative error of FHE coefficients compared to the plaintext models (LinearRegression and SGDRegressor).

From the data, we observe that as the number of observations increases, the FHE coefficients progressively approach those of the plaintext models. This trend is particularly evident in the decreasing relative errors, which indicate an improved approximation of traditional regression results by the FHE model as the dataset size grows. Notably, for 30,000 observations, the relative errors of FHE compared to LinearRegression are below 2% for both coefficients, and the differences with SGDRegressor are similarly small.

These findings suggest that while FHE-based regression initially exhibits noticeable discrepancies, its performance converges towards that of standard methods with larger datasets. This supports the potential viability of privacy-preserving regression techniques in practical applications, provided that sufficient data is available for training.

Table 3: Comparative Analysis of Model Coefficients using Three Different Regression Approaches

Model	Observations	Intercept	Coeff. 1	Coeff. 2
Sklearn LinearRegression	300	0.0000	0.8657	-0.0137
Sklearn LinearRegression	600	0.0000	0.8037	0.4713
Sklearn LinearRegression	6000	0.0000	0.7480	0.0782
Sklearn LinearRegression	30000	0.0000	0.4815	0.8200
SGDRegressor	300	0.0009	0.8589	-0.0108
SGDRegressor	600	-0.0010	0.8039	0.4698
SGDRegressor	6000	0.0009	0.7441	0.0701
SGDRegressor	30000	-0.0012	0.4805	0.8193
FHE Regression	300	0.0000	0.8510	-0.0123
FHE Regression	600	0.0000	0.7910	0.4650
FHE Regression	6000	0.0000	0.7354	0.0772
FHE Regression	30000	0.0000	0.4732	0.8061

Table 4: FHE Regression Across Datasets of Varying Sizes

Observations	Relative Error (FHE vs LinearRegression) - Intercept	Relative Error (FHE vs LinearRegression) - Coeff. 1	Relative Error (FHE vs LinearRegression) - Coeff. 2	Relative Error (FHE vs SGDRegressor) - Intercept	Relative Error (FHE vs SGDRegressor) - Coeff. 1	Relative Error (FHE vs SGDRegressor) - Coeff. 2
300	3.01e-07	1.11%	18.88%	1.59e-05	0.87%	14.81%
600	2.96e-07	1.57%	1.34%	1.56e-05	1.61%	1.10%
6000	3.92e-08	1.02%	3.87%	1.00e-06	1.17%	10.04%
30000	9.64e-09	1.72%	1.70%	2.28e-06	1.52%	1.60%

4.4 FHE Hardware Performance

The FHE hardware performance results are obtained based on the implementation of the gradient descent using FHE primitives. We use HElayer's API functions, in Python, to implement the gradient descent based on the FHE primitives of the CKKS scheme.

The Linear Regression Python code is processed by HElayers and produces a dataflow graph consisting of FHE primitive operations and their interconnections. The graph on Figure 7 is for inference or prediction on a dataset with an arbitrary number of rows (1 to unlimited) by 50 columns. The fifty FHE encrypted data points (either coefficients or input values) are packed (SIMD) across 7 FHE encrypted input vectors. The encrypted data processing flow is described below:

1. The graph reads in the encrypted coefficients (7 tensors).

2. The graph then reads in the 1st row of encrypted input data (7 tensors).
3. An encrypted matrix multiplication is performed, and the encrypted results across the 7 tensors are summed together (sum of tree) to produce the single encrypted predicted output value.
4. The encrypted predicted tensor is then sent to the output.
5. GOTO step 2 to continue processing until all the input rows are processed.

To estimate the performance, we collected FHE primitives from the hardware RTL emulator. Each FHE primitive has size (number of FracTLcores), latency (clock cycles), and throughput (Clock cycles per Output) values. For the given number of FracTLcores in a server, the graph is processed (summation of the processing chain of FHE primitives) to determine the total processing time for the total number of inputs.

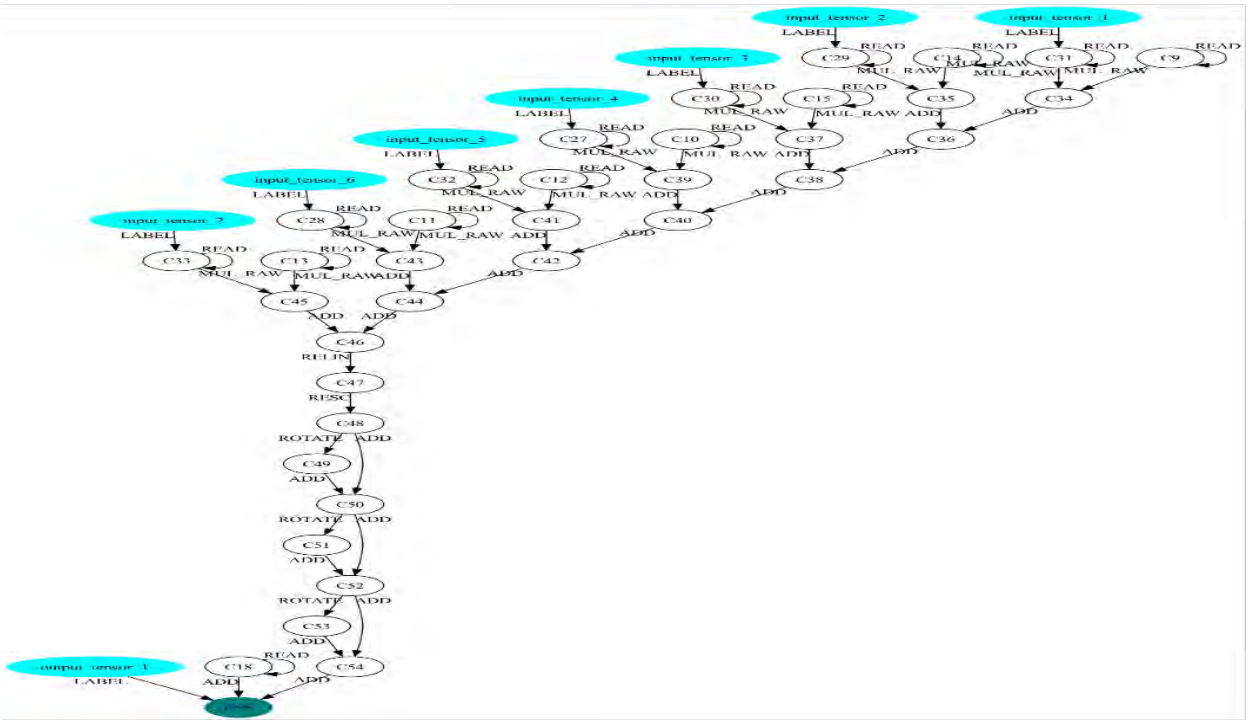


Figure 7: Call graph for Regression inference on HW Emulator

4.5 Scaling the FHE Application

We implemented a prototype application for training a linear regression model on FHE-encrypted data and for making inference on FHE-encrypted data. The implemented solution implements the gradient descent approach for training a linear regression model on encrypted data. The test results show that the runtime for training a linear regression model grows linearly (see Figure 3). Similarly, the runtime for making inference using a trained linear regression model grows linearly (see Figure 4). The obtained test results and observations show that the implemented solution is

scalable and can support large-scale data analytics applications, that use datasets with millions of rows. This observation can be seen in Figure 5 and Figure 6, for the training and prediction phases respectively.

5 Conclusion

FHE offers promising solutions for ensuring data security and privacy. Unlike conventional cryptography that only provides data security in storage and/or transmission, FHE solutions also enable performing computations on encrypted data (data security for data in use). There has been a significant amount of progress on theory and development of FHE schemes. To effectively integrate the FHE-based solutions into real-world applications, we need to speed up the FHE schemes with hardware acceleration technologies.

In this paper, we used one of the popular FHE schemes, the CKKS scheme, for a real-world multiparty computation data analytics application. We implemented the gradient descent algorithm to train a linear regression model on an FHE encrypted dataset. We used the trained linear regression model to make predictions on the private set intersection (PSI) of two datasets. Our results from this project show that the HW-accelerated implementation of the gradient descent algorithm provides a significant speedup for training and prediction. HW-accelerated FHE solutions enables the use of FHE in real-world applications.

6 References

[1] R. L. Rivest, M. L. Dertouzos, and L. Adleman, "On data banks and privacy homomorphisms," *Foundations of secure computation*, vol. 4, no. 11, 1978.

[2] C. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices," in *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2009. doi: 10.1145/1536414.1536440.

[3] M. Albrecht *et al.*, "Homomorphic encryption standard," in *Protecting Privacy through Homomorphic Encryption*, 2022. doi: 10.1007/978-3-030-77287-1_2.

[4] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017. doi: 10.1007/978-3-319-70694-8_15.

[5] B. Chen and X. Zheng, "Implementing Linear Regression with Homomorphic Encryption," *Procedia Comput Sci*, vol. 202, pp. 324–329, 2022, doi: 10.1016/j.procs.2022.04.044.

[6] P. M. Esperança, L. J. M. Aslett, and C. C. Holmes, "Encrypted accelerated least squares regression," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 2017.

[7] W. J. Lu, S. Kawasaki, and J. Sakuma, “Using Fully Homomorphic Encryption for Statistical Analysis of Categorical, Ordinal and Numerical Data,” in *24th Annual Network and Distributed System Security Symposium, NDSS 2017*, 2017. doi: 10.14722/ndss.2017.23119.

[8] D. Wu and J. Haven, “Using homomorphic encryption for large scale statistical analysis,” *FHE-SI-Report, Univ. Stanford, Tech. Rep. TR-dwu4*, 2012.

[9] C. Guo and N. J. Higham, “A Schur–Newton Method for the Matrix \lowercase{\boldmath *p* }th Root and its Inverse,” *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 3, pp. 788–804, Jan. 2006, doi: 10.1137/050643374.

[10] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, “(Leveled) fully homomorphic encryption without bootstrapping,” in *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 2012. doi: 10.1145/2090236.2090262.

[11] J. Fan and F. Vercauteren, “Somewhat Practical Fully Homomorphic Encryption,” *Proceedings of the 15th international conference on Practice and Theory in Public Key Cryptography*, 2012.

[12] R. Hall, S. E. Fienberg, and Y. Nardi, “Secure multiple linear regression based on homomorphic encryption,” *J Off Stat*, vol. 27, no. 4, 2011.

[13] A. Bertolace, K. Gatsis, and K. Margellos, “Homomorphically encrypted gradient descent algorithms for quadratic programming,” Sep. 2023.

[14] A. Ng, “CS229 Lecture notes,” *CS229 Lecture notes*, vol. 1, no. 1, 2000, doi: 10.1111/j.1466-8238.2009.00506.x.

Data sharing conundrum? In fully homomorphic encryption we (must) trust

Presenter: Adriano Baldeschi PhD

Authors: Adriano Baldeschi, Giuseppe Bruno, Mirko Avantaggiato, Andrea Capitanelli



CORNAMI®
Intelligent Computing

What is Cryptography?

- Definition:
- - Cryptography is a technique to protect information by converting it into an unreadable format without a secret key.

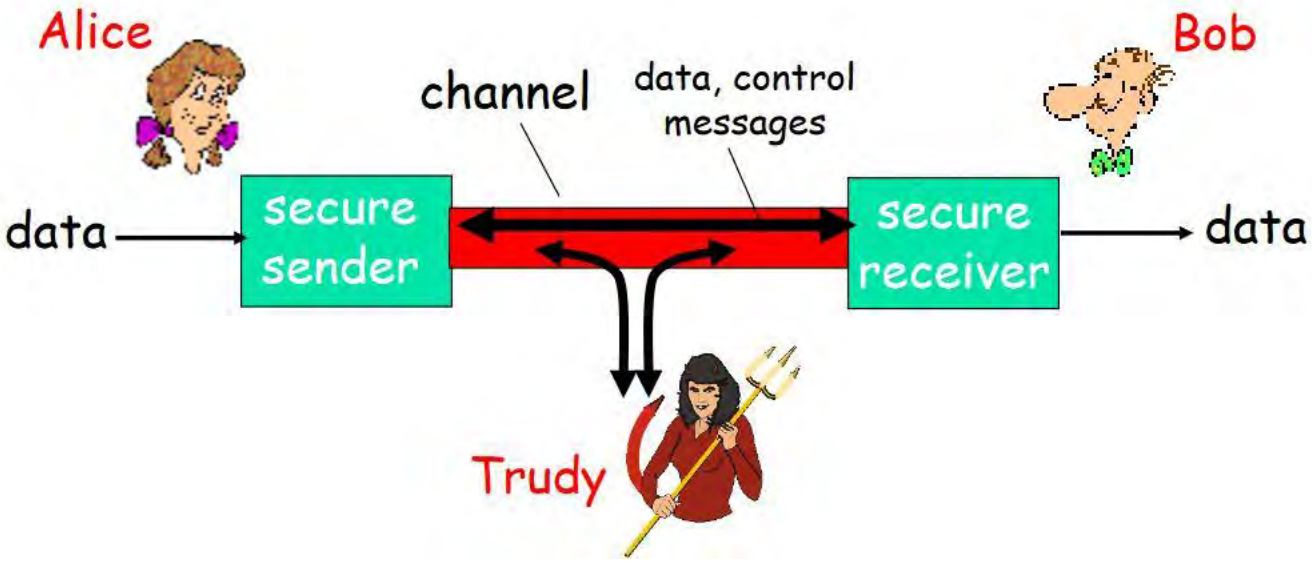
- Main Objectives:
- 1. Confidentiality
- 2. Integrity

A little bit of naming

A cryptographic system is a system capable of encrypting and decrypting a message through the use of an algorithm and a key (an alphanumeric string). The message to be encrypted is called “plaintext” while the result of the cryptographic algorithm is called “ciphertext”

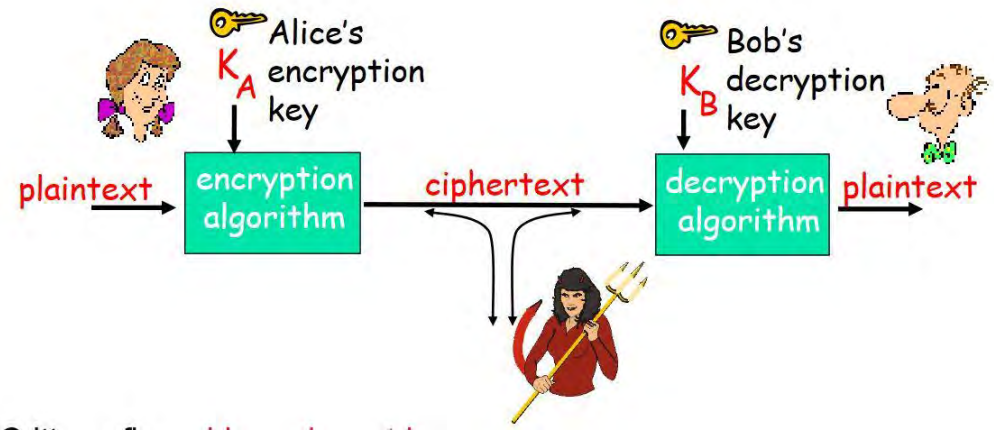
Alice and Bob are the two parties that intend to communicate securely through the network

- They could be client and server programs, or devices (e.g. routers) …
- Trudy is a third party that can listen to the messages exchanged by Alice and Bob and possibly alter them, delete them or create fake ones



Types of Cryptography

- Main Categories:
- - Symmetric Cryptography: One key for both encryption and decryption.
- - Asymmetric Cryptography: A pair of public and private keys.



Introduction to Homomorphic Encryption

- What is it?
- - A type of encryption that allows computations on encrypted data without decrypting it.

- Key Feature:
- - The computation result remains encrypted but, when decrypted, matches the result of working on plaintext data.

- Origin of the term:
- - 'Homomorphic' means 'same form' in Greek.

How Does Homomorphic Encryption Work?

- 1. Encryption: Data is transformed into encrypted form.
- 2. Computation on encrypted data: Operations like addition and multiplication are performed without access to plaintext.
- 3. Decryption: The encrypted result is transformed into plaintext, providing the correct output.

- Note: Computationally intensive but increasingly practical.

An example

Scenario: Credit Risk Analysis with Homomorphic Encryption

A bank wants to calculate a customer's **creditworthiness** based on their **income** and **total debt**, but without ever seeing the raw data for privacy reasons.

Problem:

- The customer does not want to share their income and debt in plaintext with the bank.
- The bank needs to compute the **debt-to-income ratio** to assess risk.
- If traditional encryption were used, the bank would not be able to perform calculations on the encrypted data.

Solution with Homomorphic Encryption

1. The customer **encrypts** their financial data with a homomorphic public key.
2. They send the encrypted data to the bank.
3. The bank performs computations directly on the encrypted data without decrypting it.
4. The encrypted result is returned to the customer.
5. The customer **decrypts** the result with their private key and obtains their creditworthiness score.

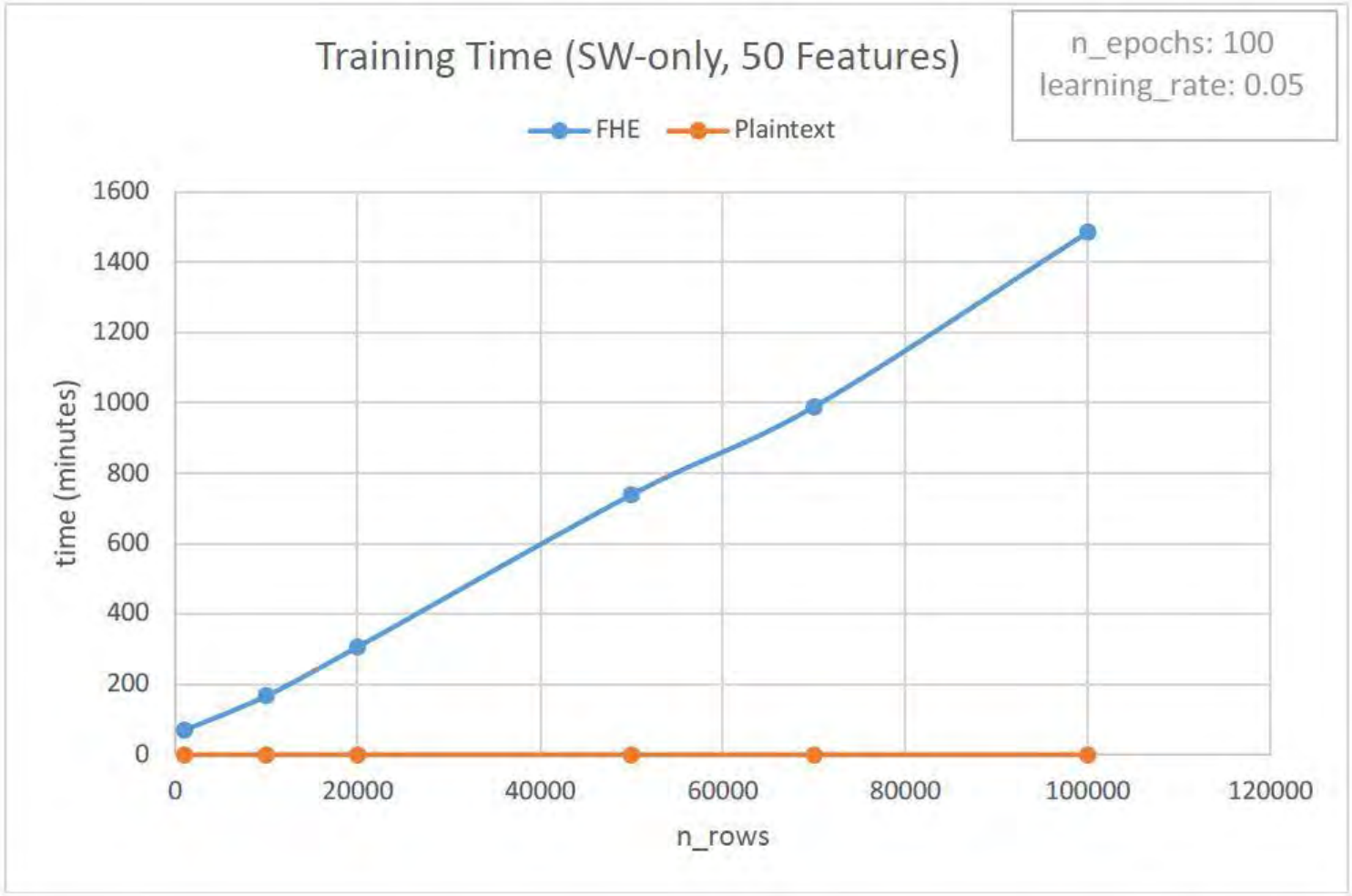
Using homomorphic encryption on linear regression

1. **Scenario**:
   - A healthcare organization wants to use patient data (e.g., age, weight, blood pressure) to train a linear regression model to predict disease risk.
   - The data is sensitive and must remain private throughout the analysis.
2. **Homomorphic Encryption in Action**:
   - The organization encrypts the dataset using a **homomorphic encryption scheme** before sharing it with a data analyst or external machine learning provider.
   - This encryption allows computations directly on the encrypted data without needing decryption.
3. **Steps in the Process**:
   - **Data Encryption**:
     Each data point (features and target values) is encrypted using a public key. For example:
     - X (features) → $Enc(X)$
     - y (target) → $Enc(y)$
   - **Secure Computation**:
     The analyst performs the required computations (e.g., matrix multiplication for gradient descent) directly on the encrypted data:
     - Compute $Enc(\beta) = (X^T X)^{-1} X^T y$, where β is the coefficient vector of the regression model.

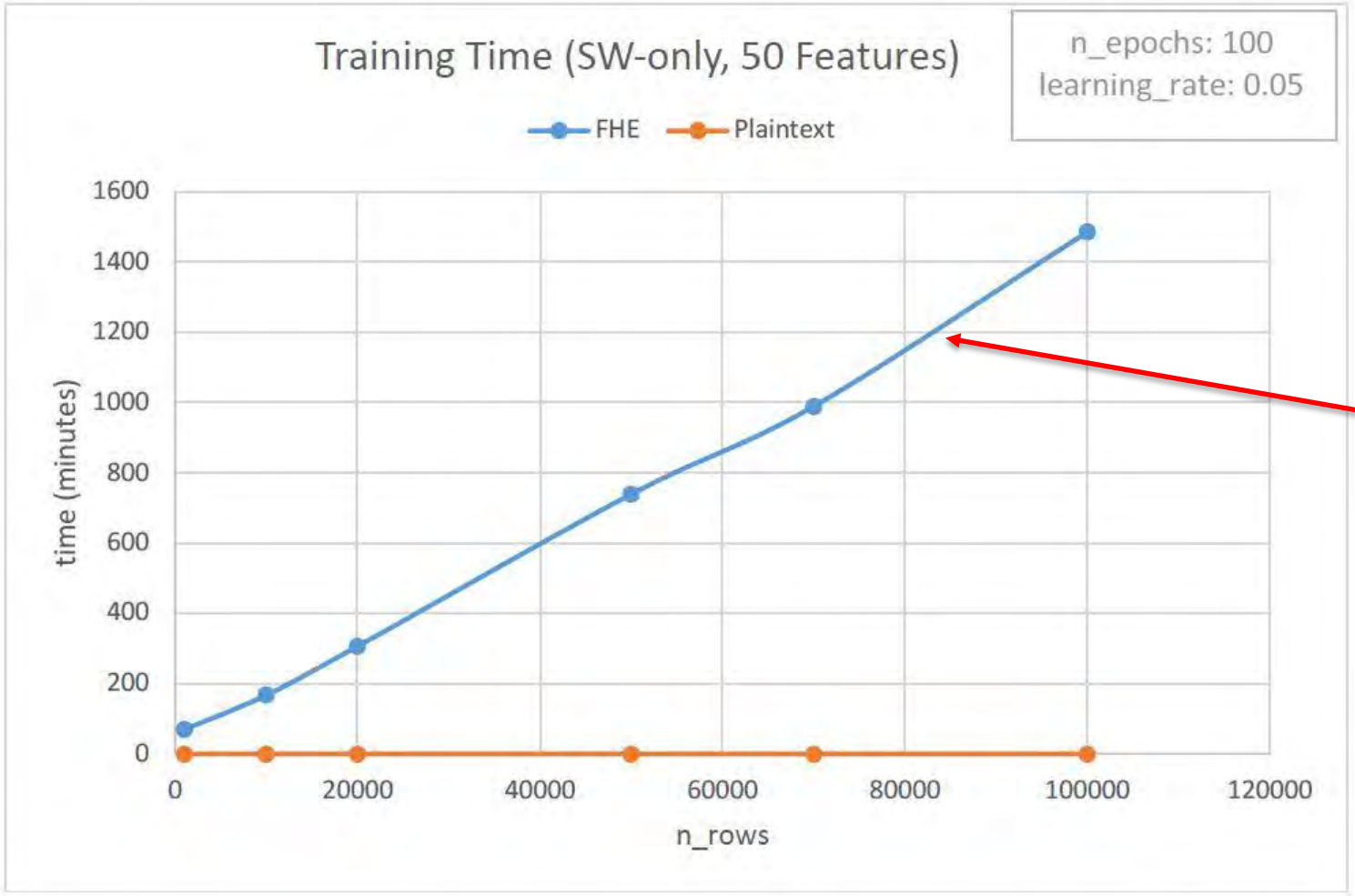
The gradient descent approach to linear regression

The gradient descent approach allows us to fit a linear regression model on a variety of datasets. In particular, we focus on datasets that can have up to 50 columns for independent variables (features) and up to 100000 rows (one hundred thousand rows). To measure the software-only (SW-only) performance results we selected various datasets with different numbers of rows and columns. Specifically, we ran several experiments by fixing the number of columns to five, thirty and fifty. For each case, we used different number of rows and measured the run time for fitting a linear regression model and making inference (prediction) with the trained model. The experimental results are as shown in the following tables and plots.

Training time



Training time



Slow!

Training time on Cornami accelerated Hardware

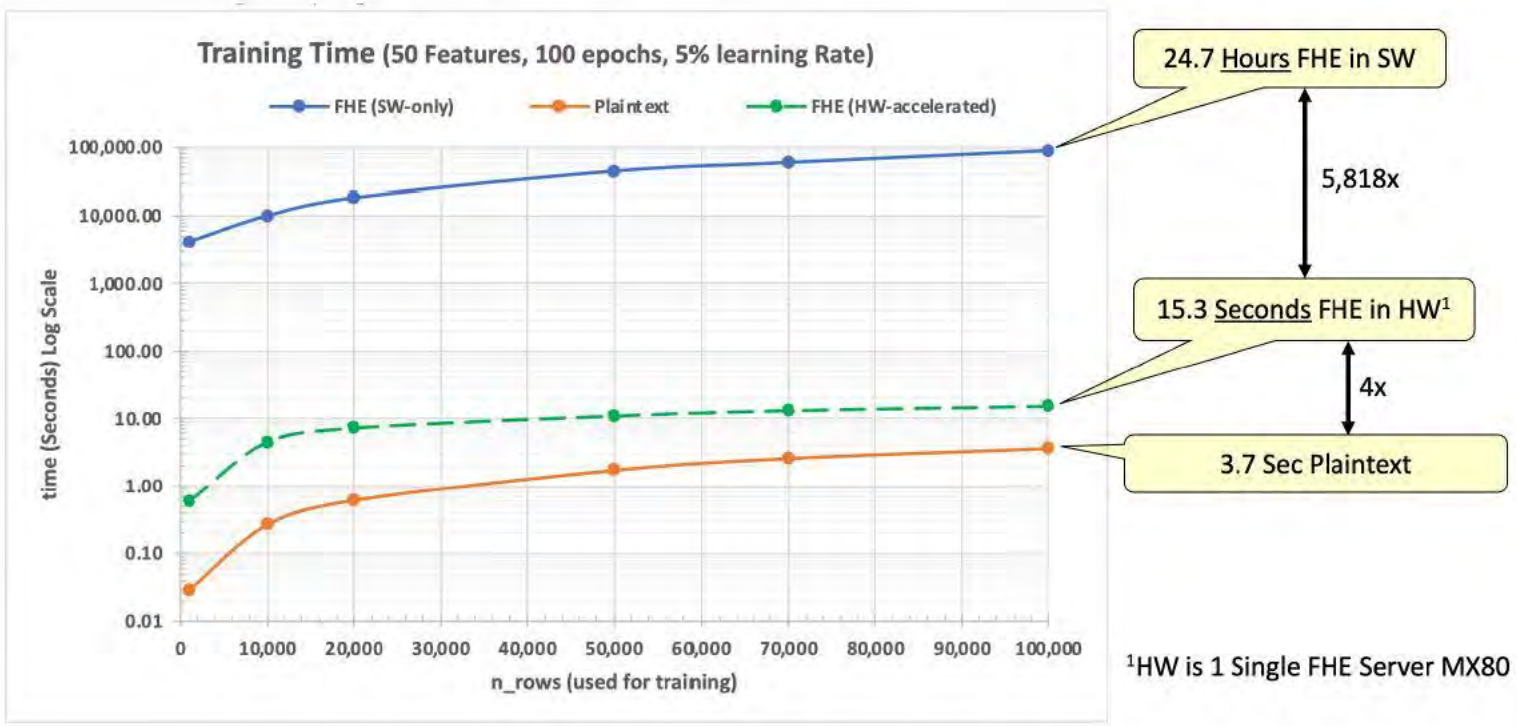


Figure 9: Training Runtime for Plaintext, FHE SW-only, Cornami FHE HW

Statistical Analysis of the Results

RMSE: root mean square error

n_rows	n_cols	RMSE (FHE and actual values)	RMSE (PTXT and actual values)
250	50	0.0064	0.00626
2500	50	0.00535	0.00499
5000	50	0.00529	0.00486
12500	50	0.0065	0.00618
17500	50	0.00683	0.00658
25000	50	0.00567	0.00533

When n_rows>1000 we obtain good results

Conclusion

- We trained a linear regression model on fully homomorphic encrypted (FHE) data using the stochastic gradient descent algorithm.
- The estimated coefficients are comparable to the actual ones when the dataset contains more than 1,000 rows.
- Although training time is relatively slow using software, it can be significantly faster using Cornami hardware.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Metadata integration as a key enabling technique for AI and ML[1]

Thomas Gottron and Daniel Suranyi,
European Central Bank

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Metadata integration as a key enabling technique for AI and ML

Thomas Gottron, Daniel Suranyi[1]

Abstract

In this paper we explore the critical role of harmonized and integrated metadata in maximizing the potential of Artificial Intelligence (AI) and Machine learning (ML) by addressing the challenges of fragmented metadata standards and silos. We emphasise the necessity of well-described, machine-readable data structures and the incorporation of methodological and business knowledge into analytical processes, illustrated with examples from central banking and official statistics. Based on these needs, we propose to create an integrated metadata layer that facilitates efficient and effective AI and ML applications. We define key requirements for such a layer. The paper outlines the key features of SDMX that make it suitable for this role and presents a high-level architecture for the proposed component.

Keywords: metadata, integration, machine learning, artificial intelligence, harmonisation, SDMX

JEL classification: C8, C55

[1] Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Contents

Introduction 3

Background 3

The relevance of metadata for AI applications 4

- Analysis of bank exposures networks and systemic risks 5
- Data quality assurance 6

A metadata layer for AI and ML 7

- An integrated metadata layer as mediator 8
- Requirements for an integrated metadata layer 9

Evaluation of SDMX as technology for an integrated metadata layer 11

- (1) Discovery, access, and retrieval 12
- (2) Support for different types of metadata 12
- (3) Data integration and transformation 12
- (4) Based on open standards 13

Summary and conclusions 13

References 13

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) promise novel insights and gains in efficiency and effectiveness when analysing data. However, the full potential of AI and ML can only be reached when certain pre-requisites are fulfilled: (1) data from various and heterogeneous sources can seamlessly be integrated and jointly analysed, (2) data structures and semantics are well described in a machine-readable format to nurture the analysis and (3) methodological and business knowledge can be incorporated in the analytical processes. Harmonised, interoperable, and well-integrated metadata addresses all those pre-requisites, rendering it a key enabler for streamlining advanced data analytics based on AI and ML. However, reaching integrating metadata is not a trivial task and is not easily achieved. Challenges are encountered in the form of a fragmented landscape of metadata standards, metadata silos dispersed in organisations, and a lack of standardised access to metadata repositories.

In this paper we elaborate on the challenges and benefits of integrating metadata, with a specific focus on AI and ML solutions. Furthermore, we propose a solution based on SDMX to provide harmonised metadata to advanced analytical systems. In particular, we make three main contributions:

(1) We illustrate the importance of metadata for AI and ML applications. To this end, we provide prototypical example use cases from the world of central banking and official statistics. We indicate how these use cases benefit from high-quality metadata, using them to motivate the need for integrated metadata. Our analysis includes benefits like streamlining data integration, semantic handling and pre-processing of variables, and the enrichment of analytics with business context information.

(2) We propose an integrated metadata layer as unified interface for AI and ML tasks and highlight how this layer can facilitate efficient and effective applications of AI and ML. The integrated layer will provide two key capabilities: (a) to serve as standardised interface, that analytical solutions can use to consume metadata, and (b) to function as a kind of Rosetta stone for translating between and connecting to different metadata sources. Driven by those two main capabilities we postulate key requirements for an integrated metadata layer component.

(3) We demonstrate that SDMX satisfies all key requirements for such an integrated metadata layer component. We point to the key features that render SDMX a suitable candidate for an integrated metadata layer and devise the high-level architecture for the component.

Background

AI plays an important role for central banks. The topic predominantly appears in conferences and presentations. The IFC workshop series on data science in central banking (Araujo, Bruno, Marcucci, Schmidt, & Tissot, 2023) is an excellent example for the relevance of advanced analytics and its applications in central banking.

The range of use cases is wide, including network analysis to assess systematic risks (Aarab & Gottron, 2024), detecting novel plausibility checks by analysing supervisory data (Romano, Martinez-Heras, Natalini Raponi, Guidi, & Gottron, 2021) or automated data integration and measuring quality indicators like integrability and coverage (Gottron, Novello, Aarab, & Lauro, 2024). The recent popularity of large language models (LLMs) has also sparked interest and investigations in potential use cases in the domain of official statistics (Buono, Felecan, & Tessitore, 2024).

The relevance of metadata to support AI and ML use cases has been highlighted in the literature. In general, metadata can support trust in AI solutions by providing information about data pipelines and other artifacts involved in building and operating an AI model (Qin & Yu, 2023). In the context of LLMs there is evidence that a strong semantic metadata layer considerably reduces AI hallucinations and increases quality of large language models' replies to user queries (Farquhar, Kossen, Kuhn, & Gal, 2024). Likewise, metadata and provenance of ML experiments is deemed important and relevant in modern data ecosystems (Schelter, Böse, Kirschnick, Klein, & Seufert, 2017), underlining the need for an overall approach to consistently capture, manage and promote metadata management.

Integration of data and metadata is a topic which has been discussed extensively in the scientific literature. One example of interoperable and harmonised metadata deployed on a large and distributed scale is Linked Open Data (Heath & Bizer, 2011). Based on RDF (Cyganiak, Wood, & Lanthaler, 2014) and rich semantic descriptions the data comes with quasi-natural metadata. The degree of integration is visualised well in the LOD cloud diagram[2]. Various works have shown the benefits of this setup for automatic and advanced analytical applications: ranging from generic Machine Learning tasks (Mountantonakis & Tzitzikas, 2017) to schema inferencing (Konrath, Gottron, Staab, & Scherp, 2012).

Another common standard for metadata is provided by the DDI alliance (Ryssevik, 1999). The DDI standard is intended for describing social science data and other data based on observational methods. More recently, the Data Catalog Vocabulary (DCAT), published by the W3C as recommendation is designed to facilitate interoperability between data catalogues published on the Web (Albertoni, et al., 2024). SDMX, finally, is an ISO standard designed to describe statistical data and metadata. It also serves to exchange and share data across organisations (SDMX, 2021). We will provide further details on SDMX in the Section on Evaluation of SDMX as technology for an integrated metadata layer.

The relevance of metadata for AI applications

Data is used for training and operating the models underlying AI applications. Accordingly, the availability and accessibility of high-quality data is key for successful AI applications. Nowadays, applications frequently tap into more than one data source: they integrate data from multiple sources to facilitate a joint analysis and

[2] The Linked Open Data Cloud: lod-cloud.net/# (last accessed: 17 April 2025)

generate novel insights. Different types of metadata play a crucial role in discovering, accessing, integrating, and analysing data from heterogenous sources.

- *Business metadata* provides context and methodological information on how data is processed and obtained. It can give insights on the scope of a dataset, how it is collected, and which considerations need to be accounted for during analytical activities. Business metadata is particularly important as it offers context and methodological information that enables modern AI and ML systems to understand the data in relation to specific business processes and objectives. This type of metadata helps in aligning data use with strategic goals and operational requirements, ensuring that generated insights are relevant and actionable.
- *Semantic metadata* provides relevant insights on the content aspect and which information is modelled in the data. This is also key for data integration tasks and the alignment of different and heterogeneous datasets. Metadata on semantics is another critical component, as it facilitates the integration, harmonization, and alignment of definitions across various data sources. It ensures that data from heterogeneous origins can be combined and compared in a meaningful way, which is essential for creating cohesive and comprehensive analytical models. By standardizing the language and definitions used within datasets, semantic metadata reduces ambiguity and enhances the interoperability of data.
- *Structural metadata* describes data from a syntactical point of view. As such, it provides the basic notion for which features are available in the data. Structural metadata outlines the data models, classifications, mappings, and transformations that govern how data is organized and accessed. This type of metadata provides the necessary framework for data analysis, allowing AI and ML models to leverage complex datasets effectively. It ensures consistency in data representation, which is crucial for maintaining the integrity and reliability of analytical outcomes.
- *Technical metadata* offers insights into the technical aspects of the data, such as encodings, provenance, and freshness. This information is vital for assessing the quality and trustworthiness of data, as well as for managing data lifecycle and compliance requirements. By providing detailed technical information, this type of metadata supports the operational efficiency and effectiveness of AI and ML systems, ensuring that they can deliver timely and accurate insights.

Use cases where metadata supports AI and ML include, but are not limited to, streamlining data integration, semantic handling and pre-processing of variables, and the enrichment of analytics with business context information. In the following sub-sections we motivate the importance of metadata and, in particular, of harmonised, interoperable, and well-integrated metadata, for AI applications using two illustrative use cases from the domain of central banking.

Analysis of bank exposures networks and systemic risks

Integrating data for performing advanced analytics on joint datasets is a common use case. Metadata plays a crucial role in the integration of datasets by providing essential context and details about the data, which makes combining and utilizing different datasets more effective. One specific use case is the integration of interbank

exposures for the assessment of systemic risks based on network analytics (Aarab & Gottron, 2024).

In a first step, structural metadata enhanced data understanding by offering schema information, including the structure of the data such as tables, columns, data types, and relationships. Additionally, descriptive metadata explained data semantics, i.e. what each data element represents. Semantic descriptions ensured that datasets are compatible and can be meaningfully merged. This helped in comprehending how various datasets can be aligned and integrated. Metadata was also instrumental in mapping data across sources by providing details about how data elements in one dataset correspond to those in another, facilitating the alignment of different data models. This included the indication of attributes corresponding to unique identifiers as well as attributes following standards and conventions used in the datasets, aiding the harmonization of diverse datasets. Finally, business metadata helped to understand the scope and purpose of a dataset. Thereby, it helped to understand which datasets are relevant to a given analytical task and should be considered for integration and joint analysis.

Data quality assurance

Another strong analytical use case for AI in central banking and official statistics is quality assurance. In this case, metadata can be used to support the formulation of validation and plausibility checks, to identify patterns for quality assurance or to ensure alignment of data with business specifications.

On the lowest level of quality assurance mechanisms, technical and structural metadata can help to formulate simple rule-based validation checks to ensure formats are correct or that values comply with code lists. Descriptive metadata and business metadata provide semantic and practical context, which can be used to choose the most suitable AI model for anomaly detection or for performing the detection of more complex patterns in data. Such anomalies and complex patterns provide the basis for verifying non-trivial correlations and connections in the data. Changes in provenance metadata can indicate disruptions or modifications in upstream data production processes. Detecting such changes and assessing their impact on further processes can help in preventing data quality issues. Such monitoring approaches could, for instance, also be used to indicate when an AI model needs to be retrained and updated.

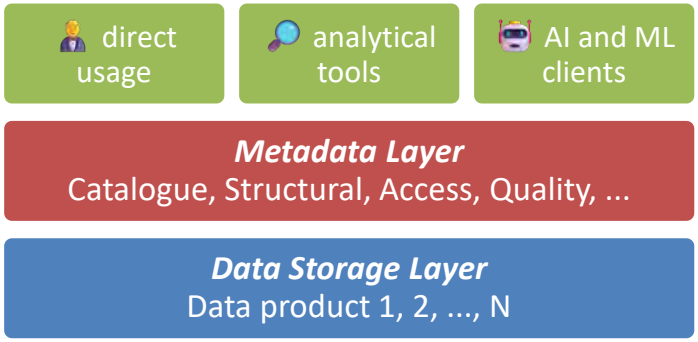
The above points motivate the positive impact of metadata as a support function for quality assurance. As soon as data processes interact with different data products and draw upon multiple sources, also the metadata of these sources needs to be included in order to support the corresponding data quality monitoring and assurance activities. Finally, the findings of DQ assessments and quality KPIs can feed into metadata as well and provide a notion of reliance for AI models. Additionally, integrated metadata can be an enabler for automating the step of data integration and measuring the fitness of a datasets for a joint analysis as demonstrated in (Gottron, Novello, Aarab, & Lauro, 2024). In this way it can save manual work, speed up the downstream analytical process and ensure a harmonised and consistent data integration approach.

A metadata layer for AI and ML

In a metadata driven system, data is always accessed through a metadata layer on top of the data stored in (various) storage locations. The metadata layer ensures a technology independent decoupling of the data storage infrastructure from the technologies used to access data and provides an abstraction layer between business functions and data storage technologies (Figure 1).

Data Management Layers

Figure 1



A metadata repository typically provides key functionality to search, access, explore and interact with metadata. Key services provided by a metadata layer or repository are:

- Serving as *single access* point for metadata
- Providing metadata in a *well-defined format*, utilising an established format (e.g. DCAT, SDMX, RDF).
- Providing programmatical access via *standardised interfaces* (e.g. APIs, SQL, linked data)

As depicted in Figure 1, also AI and ML clients are foreseen to be build upon metadata layers. Rational is, that metadata is a pivotal element for streamlining AI and ML solutions, as it provides essential context and structure to the data used in these solutions. The technical nature and interfaces of a metadata repositories are particularly relevant and important for AI and ML clients.

However, within many organizations, metadata silos present significant challenges to effective data utilization by AI and ML solutions. These silos occur when metadata is stored in isolated systems or by independent departments, preventing seamless access and integration across the organization. As a result, valuable metadata remains underutilized, hindering the ability to leverage data for comprehensive analysis and decision-making across a variety of data sources.

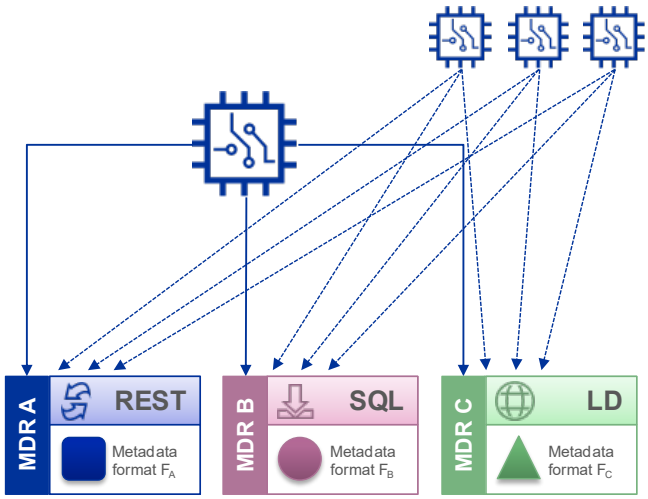
Additionally, the lack of semantic integration and metadata governance amplifies these challenges by potentially leading to inconsistencies and misalignments in how data is defined and used. Without a unified approach to metadata management, different parts of an organization may use varying definitions and standards, leading to confusion and inefficiencies.

Thereby, a fragmented landscape of metadata standards further complicates the situation, as it creates barriers to interoperability and standardization. With multiple standards, organizations face difficulties in aligning their metadata practices both internally and with external partners. This fragmentation makes it challenging to develop comprehensive and cohesive analytical solutions.

Typically, such a fragmented landscape is not a deliberate choice, but the result of different projects, incorporation of external sources and services or the extension of services to legacy systems. Nevertheless, it poses the challenge that AI and ML systems need to be capable to interact with multiple metadata access points, speak to potentially different metadata standards and interact with different types and natures of interfaces. Furthermore, the alignment and integration of metadata resides with the AI and ML solution, that has to ensure a semantically correct and seamless integration of metadata and data across the repositories. Such tasks are amplified and become a burden if more than one application needs to address, implement and solve them as depicted in Figure 2.

Scenario of several ML and AI applications accessing a fragmented metadata landscape consisting of heterogeneous metadata repositories

Figure 2



An integrated metadata layer as mediator

An integrated metadata layer, that unifies across a fragmented metadata landscape can bring benefits and remedy some of the issues described above. The main tasks of an integrated metadata layer are threefold:

(1) Provide a *unique and single access point* for all applications, in particular ML and AI solutions. This unique access point remains stable over time and is independent of underlying data storage systems and their connected metadata repositories.

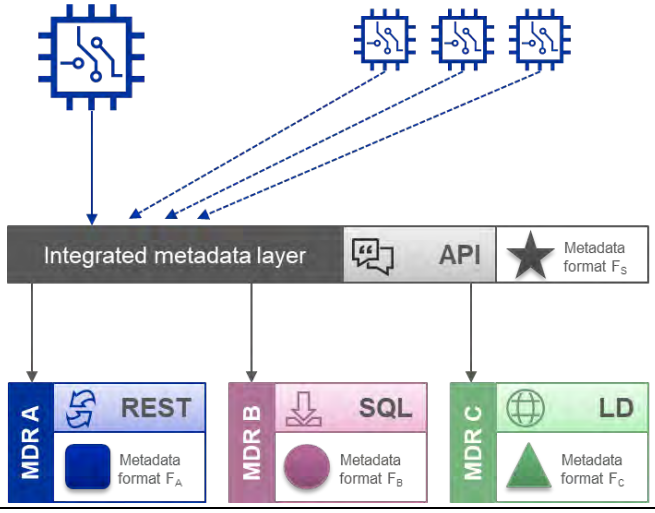
(2) At the same time, the integrated metadata layer provides a *standardised format and API for accessing information*. This includes transparent translation between different metadata standards, APIs and end points.

(3) An additional functionality to enhance the service is to provide a *semantic mapping and translation* between the semantics in the metadata repositories. Thus, it serves as a kind of Rosetta stone[3] to define mappings of code lists, concepts, taxonomies, ontologies, etc.

The result is a scenario as depicted in Figure 3. The integrated metadata layer would serve as mediator between AI and ML applications and a heterogeneous and fragmented metadata landscape, fulfilling the above three tasks.

An integrated metadata layer to mediate between AI applications and a heterogeneous metadata landscape

Figure 3



Requirements for an integrated metadata layer

This role implies certain requirements for the integrated metadata layer. In the following, we formulate and motivate these requirements based on the key functionality an integrated metadata layer has to implement to satisfy the needs of AI and ML applications. It may also serve further purposes and usage as indicated in Figure 1. We group the requirements in four distinct groups:

1. **Discovery, access, and retrieval**

    a. Provision of a *unified and standardized information model* to ensure consistency and interoperability across different systems and applications. This model should facilitate the efficient management, discovery, and retrieval of metadata.

    b. Implementation of a *modern, standardized API* that supports robust searching and retrieving of metadata. This API should be capable to handle complex queries and provide efficient and effective access to

[3] The Rosetta Stone, a granodiorite stele discovered in 1799 near Rosetta, Egypt, features the same text in Greek, Demotic, and hieroglyphic scripts. It played a crucial role in deciphering Egyptian hieroglyphs, as the Greek text enabled to unlock the meaning of the other scripts in the 1820s.

metadata, enabling seamless interaction and integration with various systems and applications.

2. **Support for different types of metadata**

    a. Ensure the metadata layer can *handle business-related metadata*, which includes definitions, policies, and methodological information supporting and underlying the data. This should facilitate AI applications' understanding and intended use of data assets.

    b. *Support semantic metadata* to capture the meaning and context of data elements. This includes relationships between data, ontologies, and vocabularies that enrich data understanding and interoperability.

    c. Capability to model, *represent and serve quality metadata* to assess and ensure the reliability, accuracy, and timeliness of data. This should include information about metrics and indicators that help to assess and monitor data quality.

    d. Provide robust *support for structural metadata*, which describes the structure and schema of data assets. This includes master metadata that defines critical data elements and hierarchies necessary for comprehensive data governance.

3. **Data integration and transformation**

    a. Design the metadata layer to describe data at *various levels of granularity*, including detailed, aggregated, and master data. This ensures flexibility and precision in data management and utilization.

    b. *Facilitate mappings between different metadata definitions* to allow for harmonization and alignment across various data sources and systems. This is crucial for seamless data integration and transformation processes.

    c. Enable the metadata layer to *support and facilitate data transformation processes*. This includes capabilities to define transformation rules and processes that convert data into a desired format or structure.

4. **Based on open standards**

    a. Ensure the metadata layer is *based on open standards* to foster integration with a diverse ecosystem of solutions. This promotes interoperability and collaboration across different technologies and platforms, primarily inside, but also outside the organisation.

    b. Design the metadata layer to be *extensible and open to integration with external data sources* and systems. This includes support for plug-ins, APIs, and other mechanisms that allow for seamless expansion and integration, ensuring the system can adapt to evolving requirements and technologies.

Evaluation of SDMX as technology for an integrated metadata layer

SDMX lends itself as a solution. In fact, the framework for the SDMX technical standards states that "[…] It is felt that a clear, standard formatting of data and metadata for the purposes of exchange and dissemination can also facilitate internal processing by organizations and users […]" (SDMX, 2021). This motivated us to investigate further, to which degree SDMX satisfies the requirements we elaborated above.

The SDMX standards package consists of various components, which standardise different aspects of data management and data exchange (SDMX, 2022; SDMX, 2023). The relevant parts of SDMX for this paper are listed below:

- **SDMX Technical Standards**
  - *Information Model*: definition of artefacts used to model data products, data exchange frameworks, and various other related aspects.
  - *Registry Specification*: description of logical interfaces used by a structural metadata repository for registration, storage, querying, and retrieval of SDMX artefacts.
  - *SDMX API*: standardised web service interfaces for data- and metadata machine-to-machine interaction.
  - *SDMX message formats*: specification of different formats used for structural metadata serialisation and file-based data exchange.
- **SDMX Content Oriented Guidelines**
  - *Globally agreed concepts, code lists, classifications* which may be used in SDMX data models.
  - Various *guidelines and governance approaches* detailing data management best practices in relation to SDMX.

Additionally, the SDMX governance structure also takes care of the *Validation and Transformation Language standard* (VTL)[4]. VTL is a standardised syntax for validation and transformation expressions, independent but fully compatible with the SDMX information model.

The different SDMX components are put in relation to the required capabilities analysed in the previous section and scored with a traffic light indicator (red/amber/green) for providing the required capability not sufficiently (red), partially (amber), or fully (green).

[4] Validation and Transformation Language: sdmx.org/validation-and-transformation-language-vtl/ (last accessed: 17 April 2025)

(1) Discovery, access, and retrieval

SDMX Score: **AMBER**

A structural metadata repository based on the SDMX standard (i.e. SDMX Registry) can be used to manage data domains and products. However, in most use cases a data catalogue will be a better choice for enterprise-level data discovery due to more advanced governance and data product management features available in data catalogue software packages available on the market. In the best case, a data catalogue is used for high level data product management activities including governance / ownership, discovery, business lineage, and other rather high-level tasks. The data catalogue is then integrated with an SDMX Registry to provide structural metadata supporting the data products and the rest of the data integration process. The standardised SMDX information model, API, and web service interfaces allow seamless machine-to-machine integration between the data catalogue, the SDMX Registry, and other platforms needing to consume metadata.

(2) Support for different types of metadata

SDMX Score: **GREEN**

An SDMX Registry is capable to provide structural metadata supporting data products and the relevant aspects for a data integration process.

The SDMX information model natively supports linking different metadata artefacts including a comprehensive support for versioning and referencing.

The SDMX Metadata Structure Definition and Metadata Set allows defining quality frameworks within structural metadata and exchanging quality information as structured messages. Ideally, the main KPIs stemming from such technically standardised quality metadata are integrated in the data product information in a data catalogue and may be used to feed data quality dashboards using an SDMX API.

(3) Data integration and transformation

SDMX Score: **GREEN**

SDMX provides at its core the standard information model, which can be used to express any structured data model and their relationships to each other, including microdata, transactional data, macro data, and master data. An SDMX Registry provides the repository required for management.

The SDMX cross-domain concepts and code lists provide a basis for master metadata management. Additionally, the content published by the SDMX agency in the Global SDMX Registry can be a starting point for enterprise-wide shared concepts and code lists. The standard allows extension of those artefacts by reference and thus an organisation framework can be placed on top of the global one. Various maintenance agencies may cooperate and share relevant structural metadata in a flexible and decentralised way.

SDMX Structure Maps provide the basic capabilities to express mappings between different structures (data models, concepts, code lists) for 1-to-1 as well as many-to-many relationships.

For more complex integration scenarios requiring metadata-driven transformations, SDMX natively integrates with VTL. Complex transformations can be stored in an SDMX Registry as VTL expressions and linked to the relevant structural metadata artefacts.

(4) Based on open standards

SDMX Score: **GREEN**

SDMX is an open source standard sponsored by 8 large international organisations (BIS, ECB, European Commission (Eurostat), ILO, IMF, OECD, United Nations (UNSD), World Bank). Many related software packages are available as open source or commercial software products. For custom use cases, SDMX allows custom extensions of the information model through a feature called controlled annotations (SDMX, 2021). The SDMX information model, API specifications, and web service standards allow seamless integration across systems requiring or managing metadata and allow for an enterprise-level single metadata management approach.

Summary and conclusions

Metadata is a key enabler for efficiently implementing and operating trustworthy AI and ML applications. It provides essential information on data that are needed for advanced analytics. A challenge for analytical applications remains an often fragmented and heterogeneous landscape of metadata repositories. We propose the solution of establishing an integrated metadata layer, that serves as single access point, provides a unified API and metadata format and harmonises and integrates metadata, thereby serving as a kind of Rosetta stone to translate between various metadata repositories. We postulated key requirements and demonstrated that SDMX seems a good fit as a basis for implementing an integrated metadata layer. As a way forward we consider implementing a first prototype with limited scope to test and validate our assessments in a real-world scenario.

References

Aarab, I., & Gottron, T. (2024). Network Topology of the Euro Area Interbank Market. *High-quality and Timely Statistics* (pp. 3-19). Cham: Springer Nature Switzerland.

Albertoni, R., Browning, D., Cox, S. J., Gonzalez Beltran, A., Perego, A., & Winstanley, P. (2024, August 22). *Data Catalog Vocabulary (DCAT) - Version 3*. Retrieved from W3C Recommendation: https://www.w3.org/TR/vocab-dcat-3/

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., & Tissot, B. (2023). Data science in central banking: applications and tools. *Proceedings of the Irving Fisher Committee on Central Bank Statistics (IFC)-Bank of Italy Workshop on "Data science in central banking: applications and tools"*. Basel: BIS.

Buono, D., Felecan, M., & Tessitore, C. (2024). An introduction to Large Language Models and their relevance for statistical offices. *Eurostat Statistical Working Papers*.

Cyganiak, R., Wood, D., & Lanthaler, M. (2014, February 25). *RDF 1.1 Concepts and Abstract Syntax*. Retrieved from W3C Recommendations: https://www.w3.org/TR/rdf11-concepts/

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024, June 19). Detecting hallucinations in large language models using semantic entropy. *Nature*, 625-630. doi:https://doi.org/10.1038/s41586-024-07421-0

Gottron, T., Novello, A., Aarab, I., & Lauro, B. (2024). Assessing Fitness for Integration – a Metadata-driven Approach. *11th European Conference on Quality in Official Statistics (Q2024)*.

Heath, T., & Bizer, C. (2011). Linked data: evoluving the web into a global data space. *J. Web. Eng., 11*(2), 177-178.

Konrath, M., Gottron, T., Staab, S., & Scherp, A. (2012). Schemex—efficient construction of a data catalogue by stream-based indexing of linked data. *Journal of Web Semantics*(16), 52-58.

Mountantonakis, M., & Tzitzikas, Y. (2017). How linked data can aid machine learning-based tasks. *International Conference on Theory and Practice of Digital Libraries* (pp. 155-168). Springer.

Qin, J., & Yu, B. (2023). Metadata in Trustworthy AI: From Data Quality to ML Modeling. *International Conference on Dublin Core and Metadata Applications*.

Romano, S., Martinez-Heras, J., Natalini Raponi, F., Guidi, G., & Gottron, T. (2021). Discovering new plausibility checks for supervisory data. *ECB Statistics Paper Series (SPS), 41*, 1-38.

Ryan, L. (2016). Architecting for Discovery. In L. Ryan, *The Visual Imperative* (pp. 203-220). Burlington, MA, USA: Morgan Kaufmann.

Ryssevik, J. (1999). Providing Global Access to Distributed Data through Metadata Standardisation–the Parallel Stories of Nesstar and the DDI. *Conferenc of European Statisticians*. Geneva.

Schelter, S., Böse, J.-H., Kirschnick, J., Klein, T., & Seufert, S. (2017). Automatically tracking metadata and provenance of machine learning experiments. *NeurIPS 2017*.

SDMX. (2021, March). *Guidelines on using SDMX Annotations*. Retrieved from SDMX.org: https://sdmx.org/wp-content/uploads/Guidelines-on-the-use-of-SDMX-Annotations.pdf

SDMX. (2021). *SDMX STANDARDS: SECTION 1 FRAMEWORK FOR SDMX TECHNICAL STANDARDS VERSION 3.0* . SDMX. SDMX.org. Retrieved from https://sdmx.org/wp-content/uploads/SDMX_3-0-0_SECTION_1_FINAL-1_0.pdf

SDMX. (2021, October). *SDMX.org*. Retrieved from SDMX STANDARDS: SECTION 2 - INFORMATION MODEL: UML CONCEPTUAL DESIGN - Version 3.0: https://sdmx.org/wp-content/uploads/SDMX_3-0-0_SECTION_2_FINAL-1_0.pdf

SDMX. (2022, November 29). *SDMX Technical Specifications*. Retrieved from SDMX.org: https://sdmx.org/standards-2/

SDMX. (2023, May 24). *SDMX Content-Oriented Guidelines (COG)*. Retrieved from SDMX.org: https://sdmx.org/guidelines/

EUROPEAN CENTRAL BANK

EUROSYSTEM

Metadata integration as a key enabling technique for AI and ML

IFC Workshop 2025



The views expressed are those of the authors and do not necessarily reflect those of the ECB.

20/02/2025

Thomas Gottron, Daniel Suranyi
ECB Data Office

Metadata supports and enables AI applications

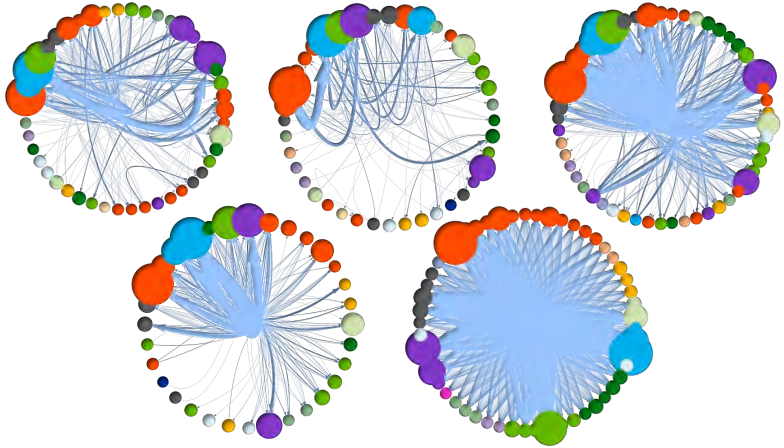


Metadata plays a crucial role for AI and ML:

- **Business metadata**: provision of context and methodological information
- **Semantic metadata**: integration, harmonisation and alignment of definitions
- **Structural metadata**: data models, classifications, mappings, transformations
- **Technical metadata**: insights on encodings, provenance, freshness, etc.

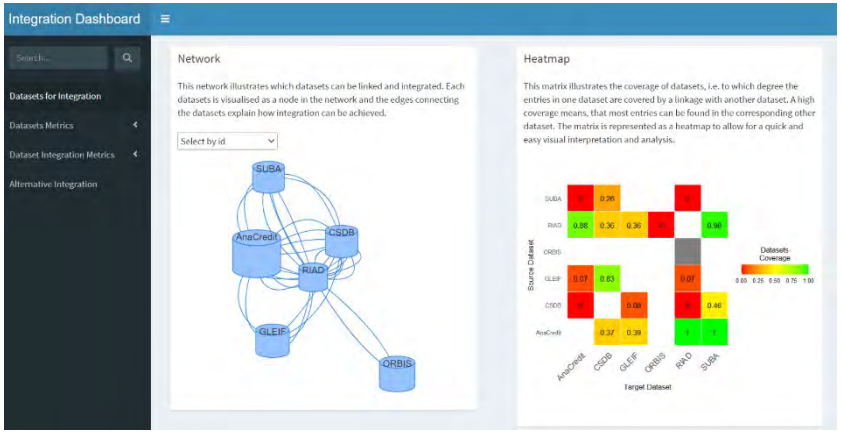
Metadata driven AI use cases in central banking

Analysis of bank exposures and systemic risks



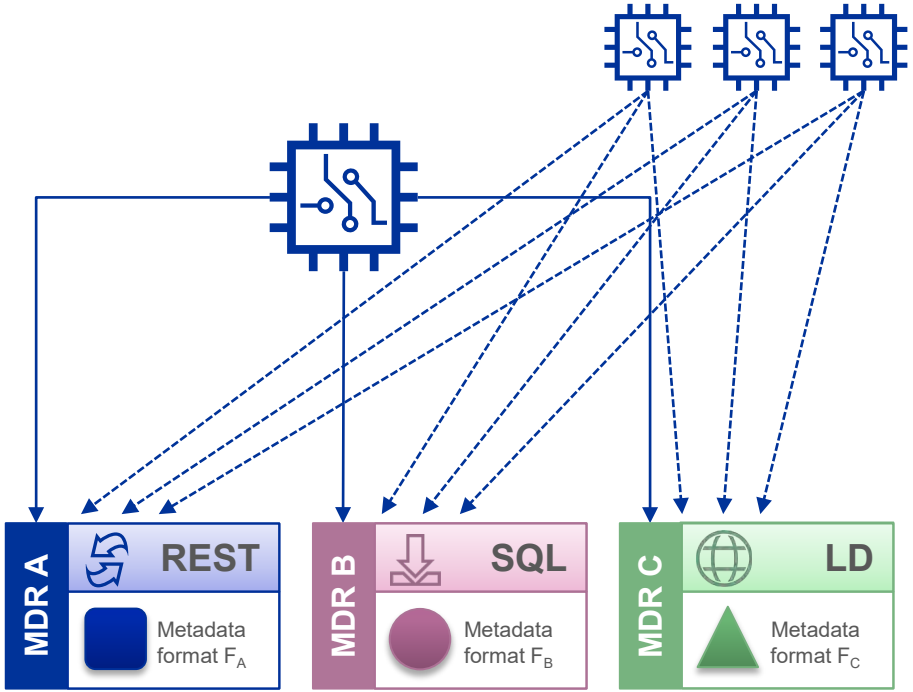
Aarab, I., & Gottron, T. (2024). Network Topology of the Euro Area Interbank Market. High-quality and Timely Statistics (pp. 3-19). Cham: Springer Nature Switzerland

Quality assurance for integrated datasets



Gottron, T., Novello, A., Aarab, I., & Lauro, B. (2024). Assessing Fitness for Integration – a Metadata-driven Approach. 11th European Conference on Quality in Official Statistics (Q2024).

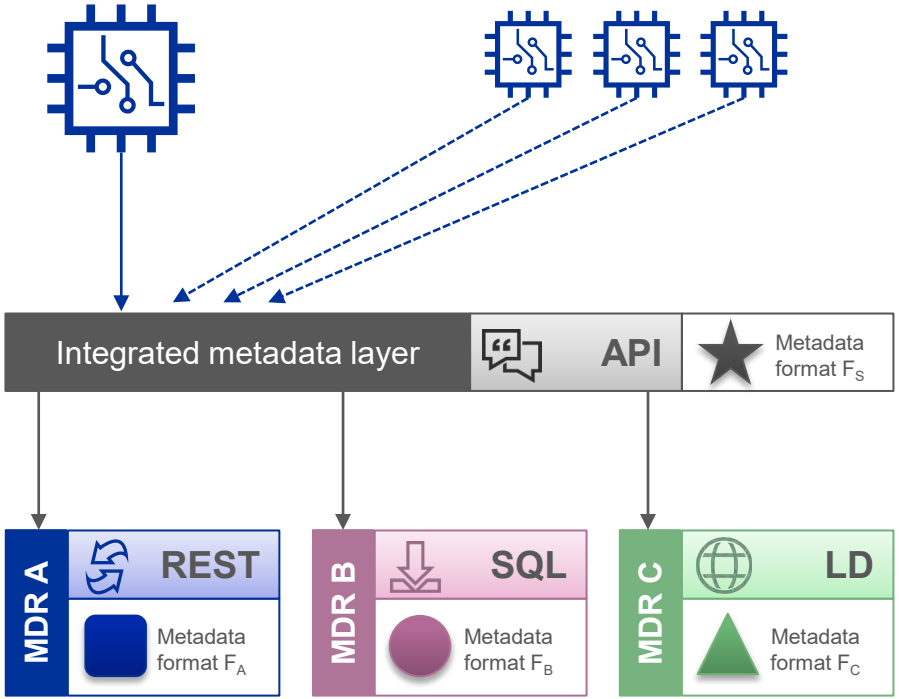
Challenges in metadata integration



- Metadata silos even within organisations
- Lack of semantic integration and metadata governance
- Fragmented landscape of metadata standards
- Lack of standardised access to metadata repositories

www.ecb.europa.eu ©

Idea and benefits of an integrated metadata layer

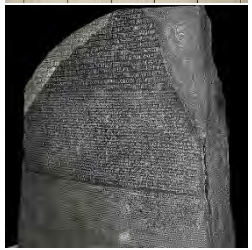




Single metadata access point for all AI and ML applications



Standardised interface (API & format)



„Rosetta stone“ for translating between and connecting to heterogeneous metadata sources

The Rosetta Stone, Awikimate, CC BY-SA 4.0, via Wikimedia Commons

Key requirements for an integrated metadata layer

Discovery, access & retrieval

- Standardised information model
- Modern standardised API for searching and retrieving metadata

Support for different types of metadata

- Business metadata
- Semantic metadata
- Quality metadata
- Structural metadata, incl. master metadata

Data integration and transformation

- Capability to describe granular, aggregated and master data
- Mappings between metadata definitions
- Support for data transformation

Based on open standards

- Integrate with a wide ecosystem of solutions
- Extensibility and open to integration with external sources

www.ecb.europa.eu ©

SDMX as backbone for an integrated metadata layer

"It is felt that a clear, standard formatting of data and metadata for the purposes of exchange and dissemination can also facilitate internal processing by organizations and users, but this is not the focus of the specification"

SDMX Standards: Section 1
Framework for SDMX technical standards, Version 3.0

Assessment of SDMX and its features

Discovery, access & retrieval



+ SDMX registry
+ Core standard information model
- Fully fledged data catalogue will be more suitable for discovery

Support for different types of metadata



+ Support for reference data, structural information, description of data flows
+ SDMX Metadata Structure Definition and Metadata Set

Data integration and transformation



+ Cross domain concepts and code lists
+ Structure maps, representation maps, item scheme maps
+ Integration of VTL

Based on open standards



+ SDMX as open and extensible ISO standard
+ Wide range of existing third-party and open-source solutions and tools

www.ecb.europa.eu ©

Summary and Conclusions

Harmonised and integrated metadata is a key enabler for AI and ML application



An integrated metadata layer helps to overcome drawbacks in current metadata landscapes



SDMX is a suitable solution to serve as backbone technology for an integrated metadata layer

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

SDMX chatbot: a trustworthy assistant to navigate the SDMX documentation[1]

Samuel Melm and Brian Buffett,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.



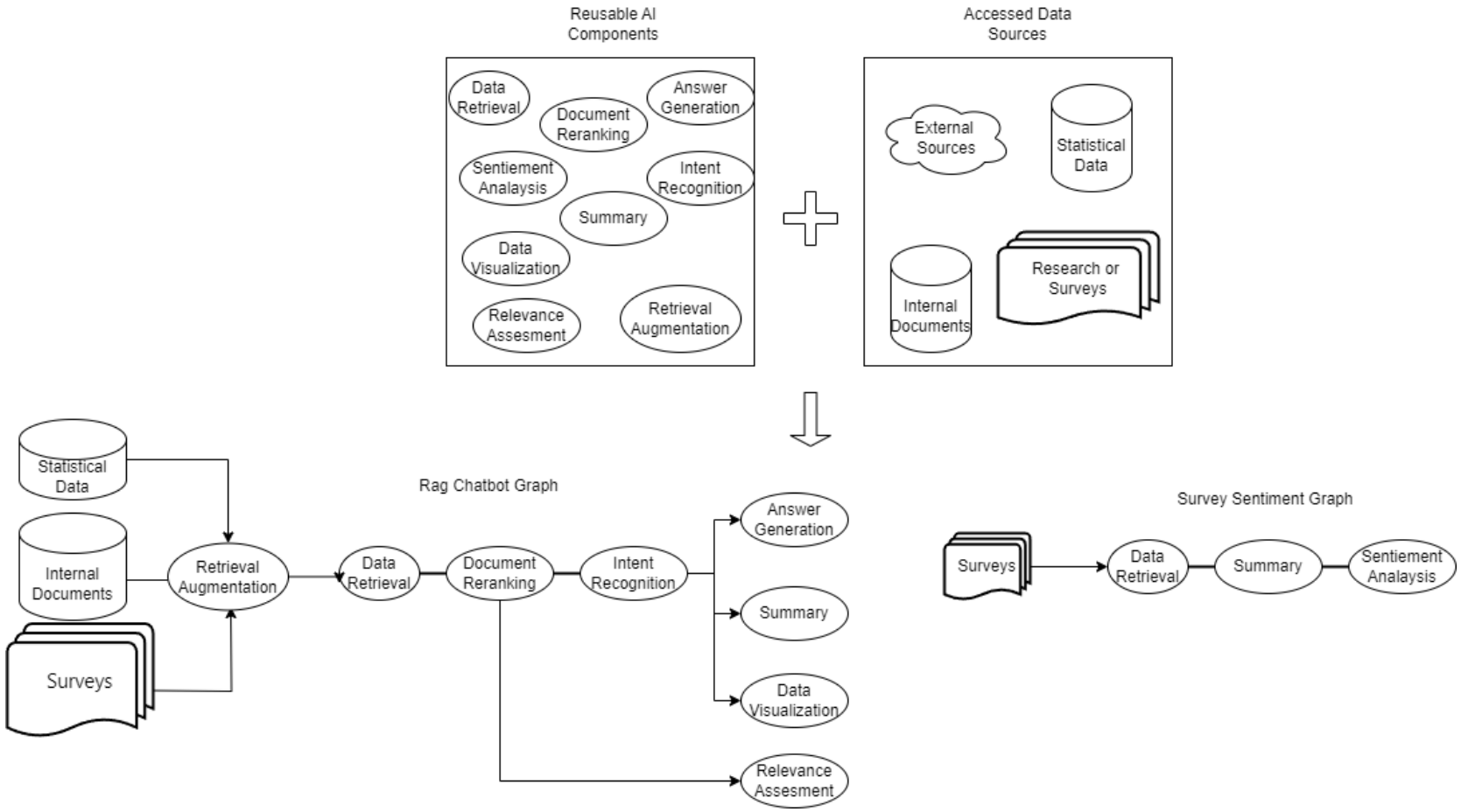
SDMX Chatbot

A Trustworthy Assistant to Navigate the SDMX Documentation

AI at BIS

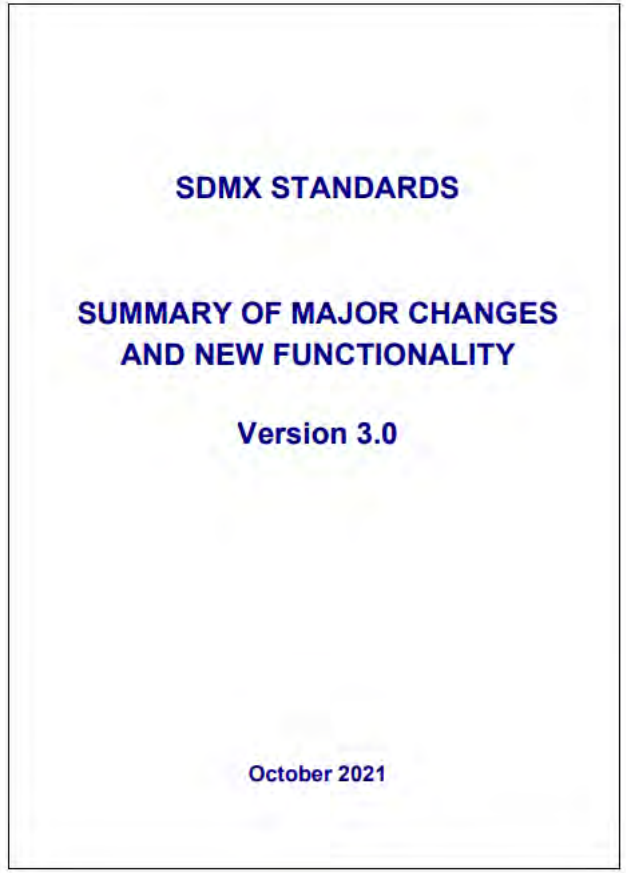
- BIS explores AI in various areas
- State of the art is expanding at a rapid pace
- Self-hosted language and code models
- Development of multiple RAG applications for variety of users
- Modular approach with high software engineering standards

Reuseable Graph Architecture



Goals

- Improve engagement with the SDMX standard
- RAG-based question answering system
  - Ingest all SDMX documentation into knowledge base
  - High-quality answers
  - Transparency
  - Verifiability


SDMX STANDARDS

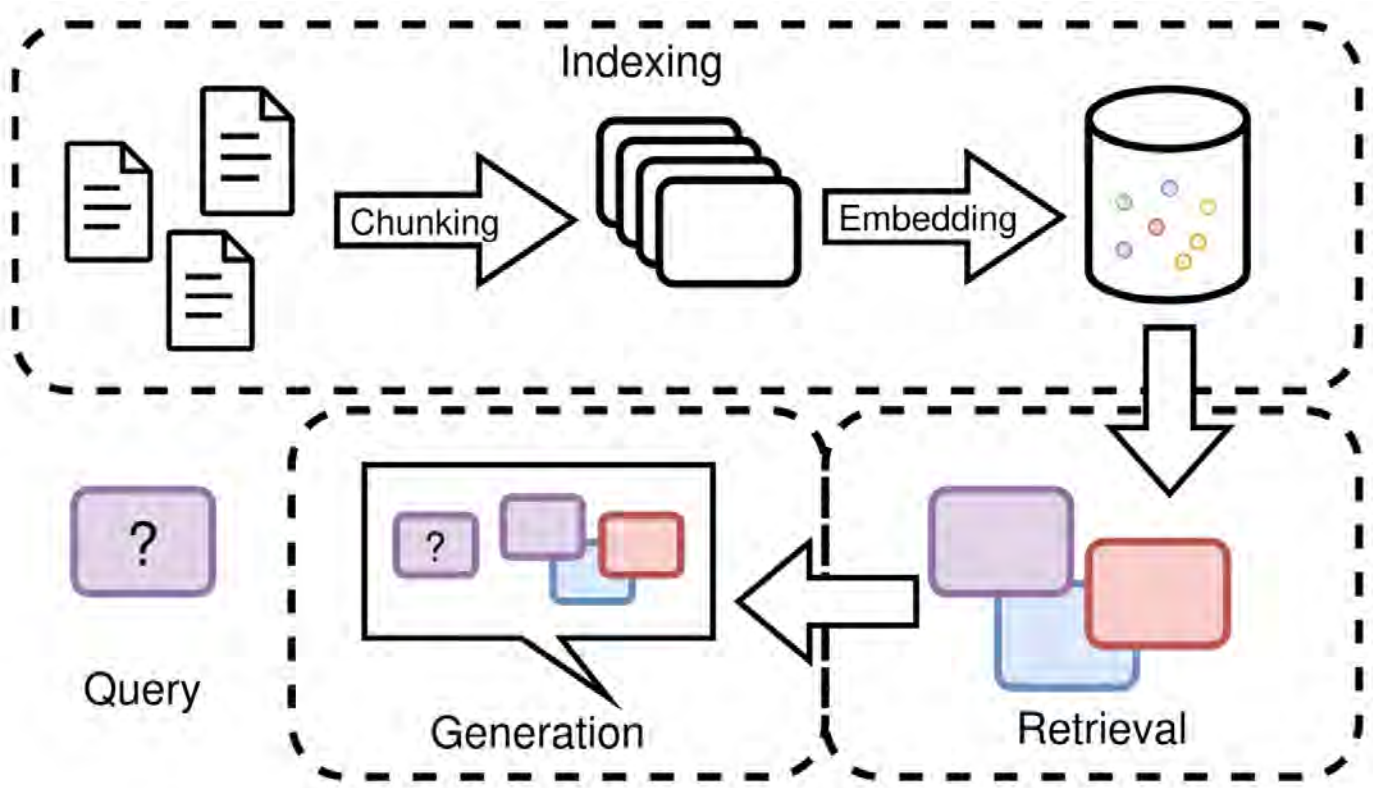
SUMMARY OF MAJOR CHANGES AND NEW FUNCTIONALITY

Version 3.0

October 2021

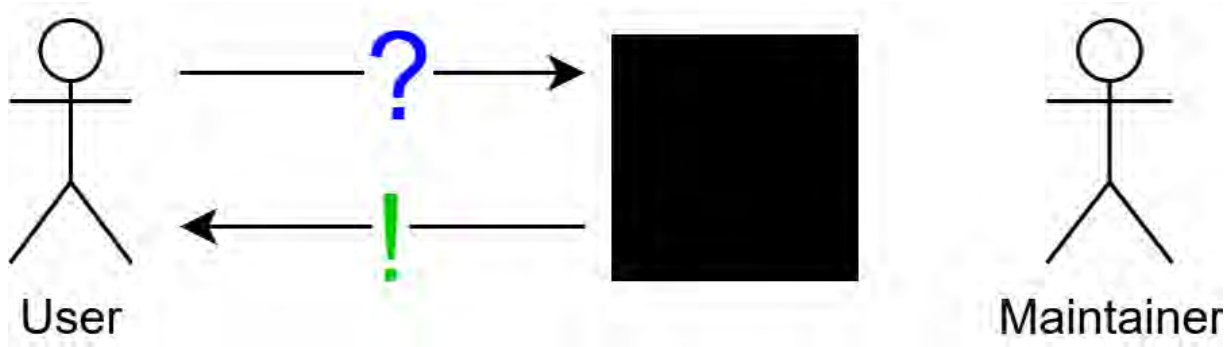
Retrieval-Augmented Generation (RAG)

- Split documents into chunks
- Create vector embeddings
- Retrieve relevant chunks
- Generate answer with LLM

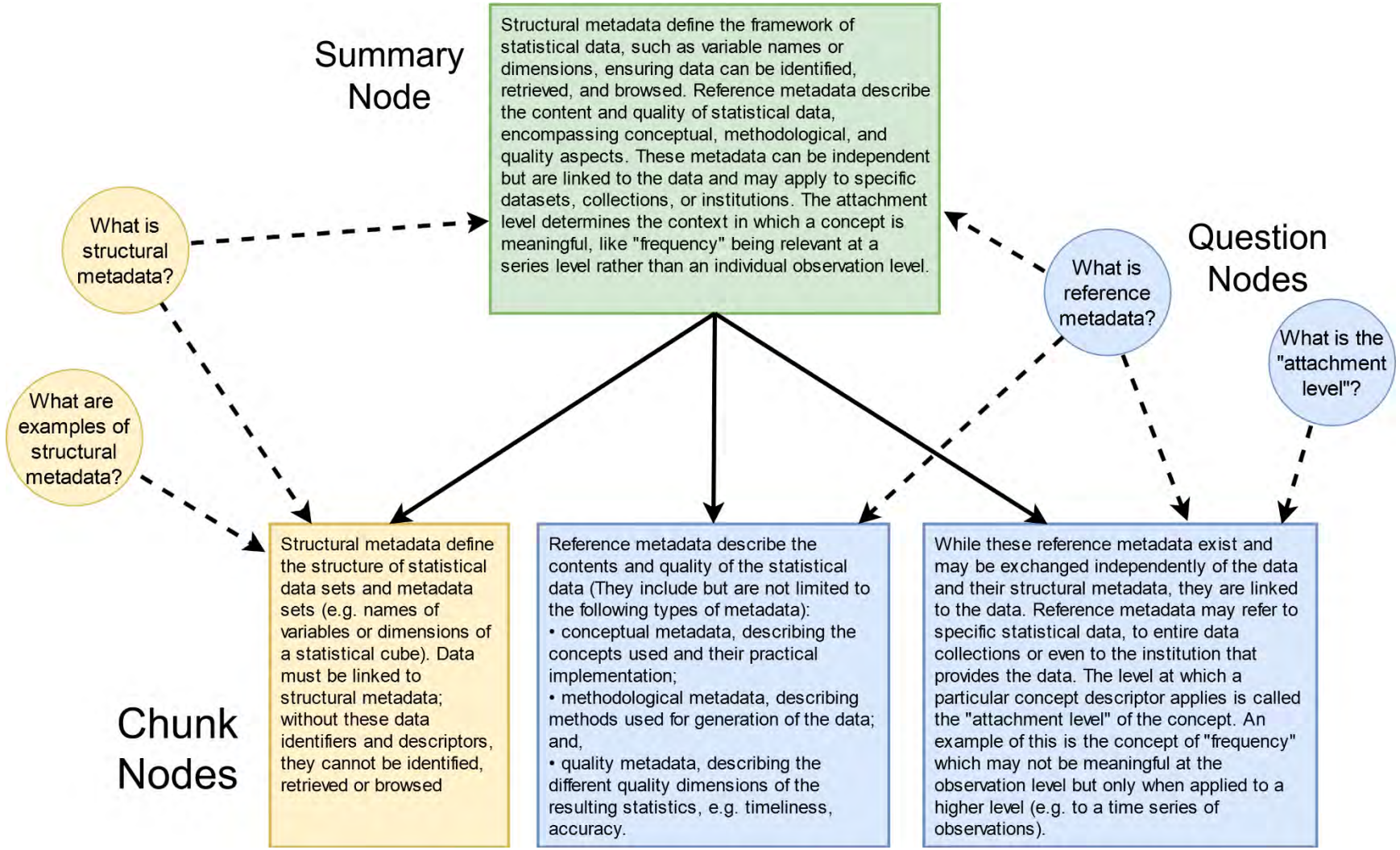


Transparency

- RAG systems are black boxes
- Users see only question and answer
  - Citations
- Maintainers must understand…
  - …capabilities of the system
  - …coverage of the documentation

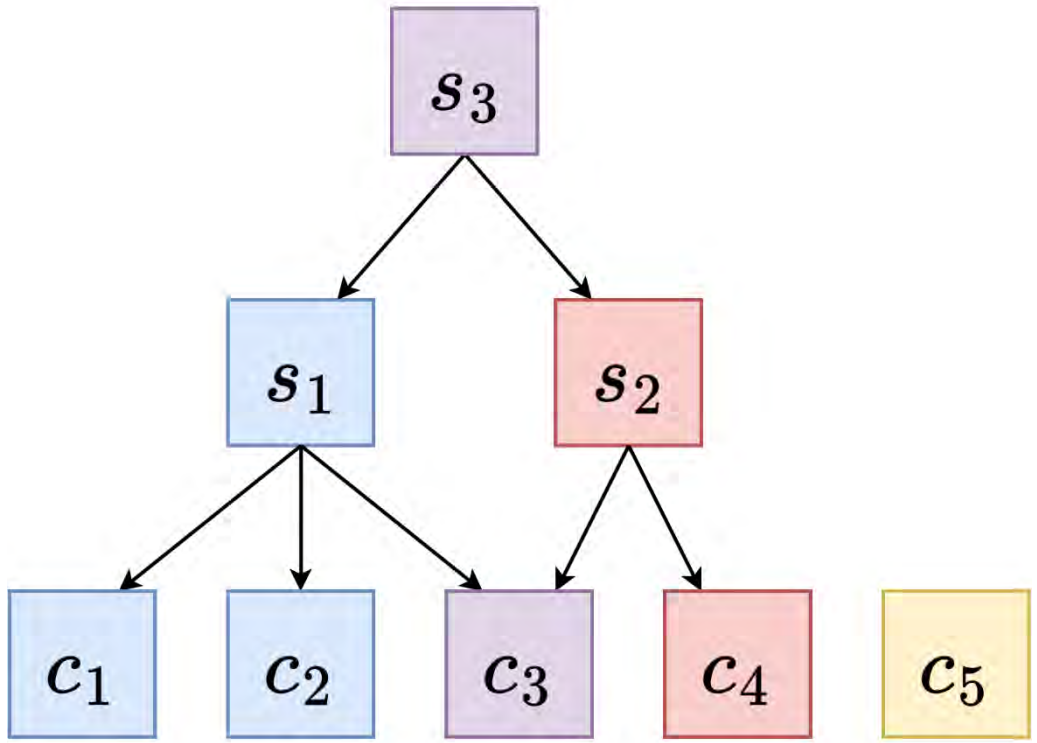


Document-Question Network

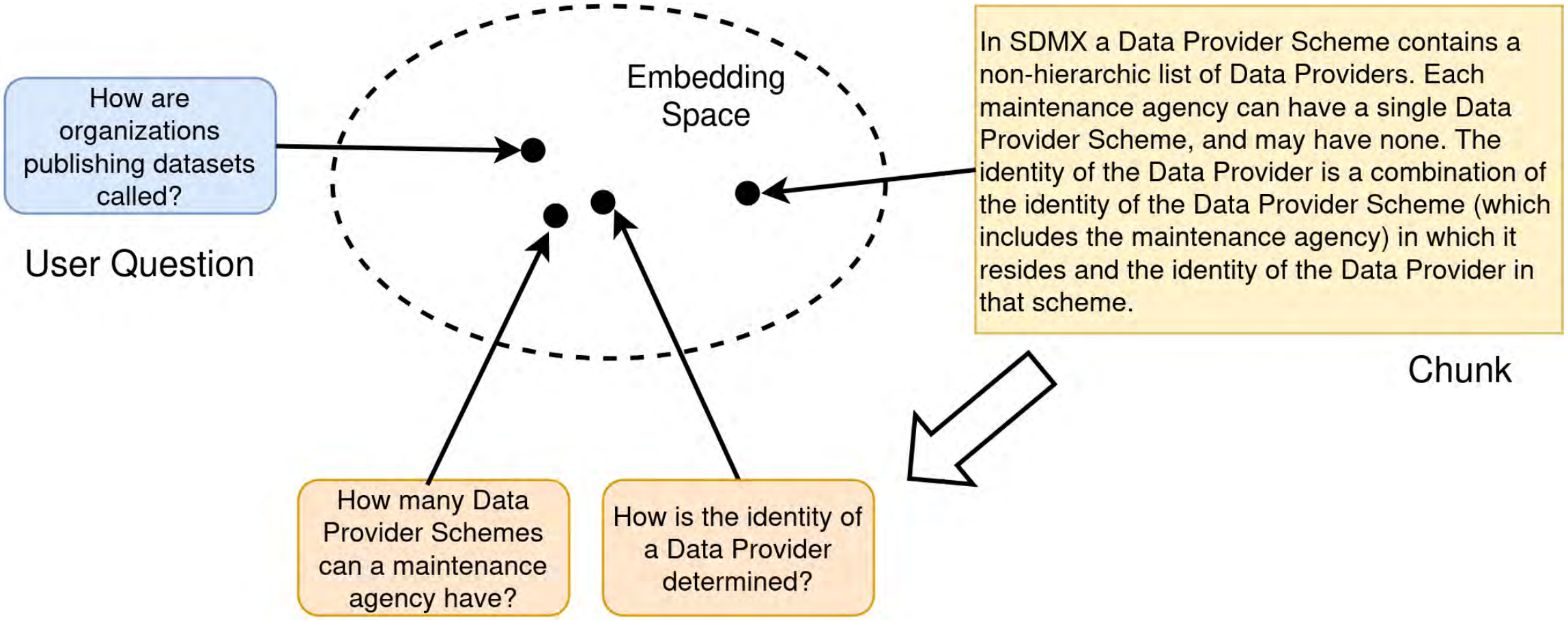


Summarization

- Goal: Improve answer quality
- Summaries are retrieved alongside chunks
- Based on RAPTOR
  - Clustering with UMAP + GMMs
  - Recursive summarization
- Reveals topics in corpus
- Reduces details and introduces abstraction
- Compress the information into fewer elements

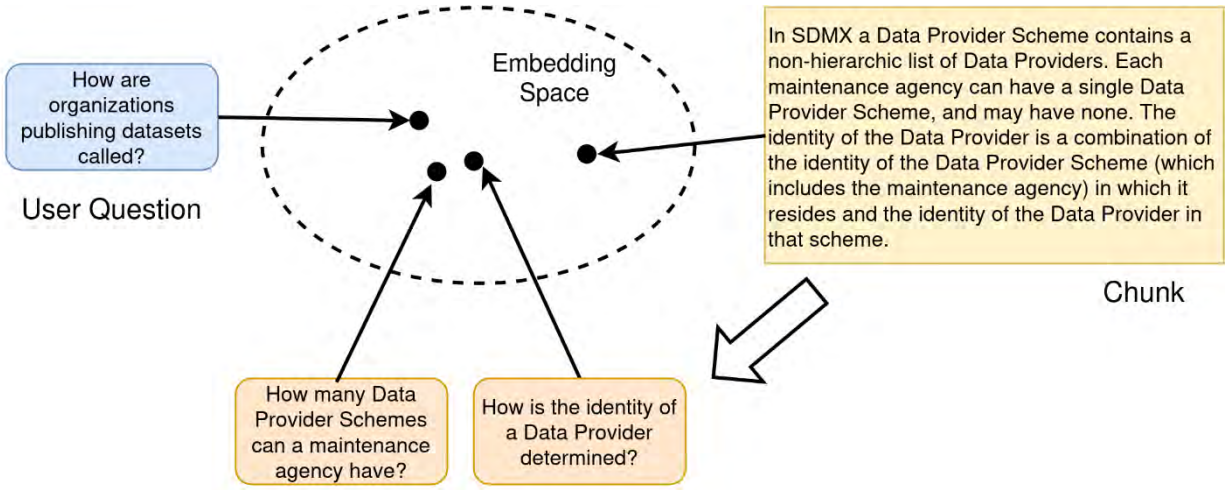


Hypothetical Questions



Hypothetical Questions

- Commonly used to improve retrieval
- Generate questions answered by chunk
  - Closer to user query
- What does chunk answer?
- What would a user ask?
- How would the system answer?
- Caching of answers



Testing & Evaluation

- SDMX experts are unreplaceable
  - Sustainability: can parts be automized?
- Initial dataset of 100 evaluation questions
  - Use automatic question generation to find more
- RAGAS
  - LLM-based evaluation framework
  - Multiple metrics
    - Factual correctness
    - Faithfulness
    - Context quality (precision & recall)

Results & Outlook

- Results compared to basic RAG pipeline
  - Up to 14% increase in correctness
    - Summarization
    - Filtering of context
  - Hypothetical questions reduce correctness
  - A lot more insights into the knowledge base
- Outlook
  - Beta testing with SDMX experts
  - Expansion of source documents
  - Presentation to the SDMX secretariat
  - Integration into sdmx.org search, following SDMX secretariat approval

BIS

Demo

Example 1: Basic RAG

What is an attribute?

An attribute provides information about the data, thus qualifying the data further.

Example 1: Improved RAG

What is an attribute?

An attribute is a qualitative information about a specific statistical object and can be concepts such as units, magnitude, currency of denomination, titles, methodological comments, or quality statements.

Example 2: Basic RAG

What is a data structure definition?

A Data Structure Definition (DSD) is a definition developed for the specific needs of one organisation only. **No, it's not that**. According to the text, a Data Structure Definition is a specification of statistical concepts necessary for identifying series and used as dimensions, attributes, and measures.

Example 2: Improved RAG

What is a data structure definition?

A Data Structure Definition (DSD) defines the structure of a dataset, including variable types, relationships, and value constraints, enabling data sharing between organizations and systems. It provides a common language and framework for describing datasets, allowing different organizations to share and compare data more easily. A DSD is a standardized description of a statistical dataset's structure and content, including attributes, dimensions, measures, and relationships, facilitating integration from multiple sources and promoting comparability and consistency.

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

sdmx.io: where tools and learning resources come together to address official statistics use cases[1]

Douglas Araujo,
Central Bank of Brazil

Brian Buffett,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

sdmx.io: where tools and learning resources come together to address official statistics use cases

Douglas Araujo (Banco Central do Brasil)[1] and Brian Buffett (Bank for International Settlements)

The increasing demands for data is creating pressure for more agile data organisations, processes, and software systems. The SDMX standard provides an information model and technical architecture that supports robust metadata-driven processes and systems. One challenge in realising this goal is to reduce barriers to entry and use throughout the data lifecycle for organisations of all size and stage of development. Sdmx.io helps to address this challenge. sdmx.io is not a single project but an ecosystem of open source tools, patterns, guidance, learning materials and other resources like pre-configured containerised environments that make the software quick and simple to deploy. The quantity of SDMX compliant data portals and data available via SDMX API's is becoming extensive and exploring ways to effectively leverage the richness of these data assets is essential. This paper discusses this topic from the perspective of data scientists and more specifically, various tools and methods to facilitate the discovery and use of SDMX data sources.

Keywords: Data science, SDMX, Data discovery

JEL classification: C81, C82, C89

[1] This paper reflects only the views of the authors and not necessarily of the Banco Central do Brasil.

1. Introduction

The demand for timely, granular, and reliable data is greater than ever before, driven by policymakers, businesses, and researchers' need for increasingly complex evidence bases. Data organizations face mounting pressure to become more agile, adapting their processes, infrastructures, and governance models to a world in which information flows are continuous, diverse, and rapidly changing. At the same time, the rise of artificial intelligence (AI) and machine learning amplifies expectations for structured, well-documented, and machine-readable data. Algorithms, no matter how sophisticated, depend on the quality and interpretability of the underlying datasets. For many data users, however, the reality is frustrating: valuable time and resources are consumed by data cleaning, format conversions, and metadata reconstruction before meaningful analysis can even begin.

The Statistical Data and Metadata Exchange (SDMX) standard was developed precisely to address these challenges (see Buffett and Napoles (2024)). SDMX offers a robust information model and technical architecture. It was initially designed to enable efficient and standardized exchange of official statistics between organizations. Over time, it has matured into a versatile framework that combines metadata-driven processes with interoperable formats and application programming interfaces (APIs), capable of supporting both one-off analysis and recurrent data pipelines. It is an ISO standard, backed by leading international organizations, namely the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Labour Organization (ILO), the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations, and the World Bank.

Despite its proven role in official statistics, SDMX remains little known in the broader data science community. Many economists, social scientists, and advanced users perceive it as a tool exclusively for national statistical offices and central banks. This perception is misleading and increasingly outdated. The core strengths of SDMX, ie rich metadata, standardized models, discoverability, and automation directly address the pain points faced by data scientists. This paper explores how SDMX is fit-for-purpose in the era of data science, outlines its value proposition for advanced users, and introduces practical initiatives such as sdmx.io and sdmx4ds that aim to make the standard more accessible.

2. SDMX: Beyond data producers

The traditional view of SDMX is tightly linked to its origins: a standard primarily for producers of official statistics. Under this framing, SDMX is seen as a tool for large institutions such as national statistical offices, international organizations, and central banks to harmonize reporting, reduce duplication, and improve data dissemination. It is true that these institutions make extensive use of SDMX (eg, Bogdanova et al (2025a)). But this impression overlooks how the same features that benefit producers can also empower data scientists and advanced users.

Rich metadata support

At the heart of SDMX lies metadata. Datasets are not just a collection of numbers; they are annotated with information about variables, units of measure, methodologies, and sources. SDMX describes these features, and more. Concretely, SDMX defines a set of metadata artefacts such as Data Structure Definitions (DSDs), code lists, concept schemes and dataflow that describe the dimensions, attributes, measures, and permissible values of a dataset. These structural metadata components specify, for example, what a "unit of measure" means, how a "reference period" is encoded, or which classification system is applied. In addition, SDMX supports reference metadata, which document methodological information such as data sources, compilation practices, quality assessments, and revision policies.

For a data scientist, this rich descriptive layer provides immediate clarity. Instead of manually reconstructing meaning from raw column headers or accompanying documentation (which is often not sufficiently clear), users can quickly assess relevance and understand structure. This reduces ambiguity and enables consistent treatment of time series, cross-sectional data, or panel datasets across domains. The standardized metadata in SDMX also facilitates reproducibility of analysis, and lowers the bar for the adaption of one analysis to other circumstances.

Improved Data Discovery

Because SDMX uses structured metadata, it enables far more effective search and filtering. Analysts can identify datasets by filters including geographic coverage, indicator type, or time period, all without downloading multiple files. For example, a researcher seeking long-term interest rate data across advanced economies can query SDMX APIs using metadata filters, retrieving exactly what is needed without manual scanning of spreadsheets.

Standardized Data Models

SDMX enforces consistency across datasets by applying common models for dimensions, measures, and code lists. This means data from the BIS, IMF, and ECB — though produced independently — can be integrated seamlessly.

For comparative analyses or multi-source dashboards, this is a decisive advantage. Data scientists can integrate multiple datasets using metadata described with a common language.

Automation and API Integration

SDMX's design aligns with the languages and platforms of data science. Whether using Python libraries for econometrics, R packages for visualization, or business intelligence software for reporting, SDMX's standardized formats reduce friction. For advanced users, the ability to automate analyses and integrate tools is critical. SDMX APIs allow programmatic retrieval of data directly into tools like Python, R, or Julia. Instead of downloading files manually, users can build reproducible pipelines that update automatically when new data is released. Analysts can integrate statistical data

into diverse toolchains without needing bespoke conversions. This functionality underpins modern workflows for monitoring systems, forecasting models, or interactive dashboards.

One illustration of how SDMX integrates with data science tools is *gingado*, the open-source tool for machine learning in economics and finance (Araujo (2023)). SDMX offers the capability to automatically augment the user's data to train machine learning models, by finding relevant data that might improve the models' performance.

Together, these features make clear that SDMX is not just a producer's tool. For advanced users, it represents an underutilized resource: a ready-made framework for discoverable, interoperable, and reproducible data use.

3. Fit-for-Purpose for Data Science

The true test of any data standard in the 2020s is whether it can meet the demands of advanced analytics. SDMX passes this test by design. Its metadata-driven architecture is not an optional add-on; it is the foundation. This makes SDMX particularly fit-for-purpose for data scientists, whose workflows depend on transparency, structure, and reproducibility.

Reducing data wrangling

A large share of data scientists' time is consumed by cleaning, formatting, and reconciling datasets. SDMX minimizes this burden. By embedding definitions, units, and code lists into the structure, it allows users to skip much of the pre-processing that is typically required. The result is faster time to analysis and more reliable outcomes.

Enabling reproducible workflows

Reproducibility is a core principle in both academic research and applied policy analysis. SDMX supports reproducibility through standardization. When datasets are pulled via SDMX APIs, they come with consistent identifiers and metadata. This ensures that the same script will yield the same dataset tomorrow, next month, or in another user's environment. Such reproducibility is critical in collaborative projects, peer review, and policy simulations.

Bridging Official Data and Advanced Analytics

Official statistics covering areas like macroeconomic indicators, labor markets, or trade are essential for economic and policy analysis. Yet these datasets often exist in fragmented formats, with limited programmatic access. SDMX bridges the gap by making official data more accessible to advanced analytics pipelines. It allows the trusted, authoritative data of statistical agencies and central banks to flow directly

into machine learning models, visualization dashboards, or econometric frameworks (see Bogdanova et al (2025b) for examples).

Supporting Multi-Source Integration

Complex problems rarely respect domain boundaries. Consider climate change, which requires integrating environmental, financial, and socio-economic data. SDMX facilitates such integration by enforcing consistent structures and metadata across domains. Analysts can combine datasets from multiple organizations without extensive re-engineering. This capacity is essential for interdisciplinary policy challenges and cross-border research.

Alignment with FAIR Principles

The principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data (Wilkinson et al (2016)) are widely recognized as best practice. SDMX embodies these principles: its metadata makes data findable, its APIs make it accessible, its standard models make it interoperable, and its structured formats make it reusable. For data scientists navigating a crowded data landscape, this alignment offers confidence that SDMX-based resources can be reliably integrated into workflows.

By addressing these core needs, SDMX establishes itself not just as a standard for producers, but as a fit-for-purpose enabler of modern data science.

4. The sdmx.io Ecosystem

Recognizing the dual needs of producers and users, the *sdmx.io*[2] initiative was launched as an ecosystem of resources designed to lower barriers and broaden adoption. Its purpose is straightforward: to make it easier for organizations of all sizes and stages of development to adopt and benefit from SDMX.

The ecosystem includes:

- Open-Source Tools: Ready-to-use libraries and utilities for working with SDMX data, compatible with common programming languages.
- Containerized Environments: Pre-configured setups that simplify deployment, allowing institutions to test and scale SDMX solutions without major IT investments.
- Learning Materials: Tutorials, webinars, and documentation that guide both technical and non-technical audiences through adoption.
- Community Resources: Shared repositories where users can exchange code, methods, and lessons learned.

[2] https://sdmx.io

By bringing together tools, environments, and knowledge in a single hub, sdmx.io reduces the perception of complexity that has sometimes discouraged uptake. It transforms SDMX from a niche technical standard into a practical solution accessible to both institutional producers and individual data scientists.

5. SDMX for Data Science (sdmx4ds)

In order to fully realize SDMX's potential for advanced users, the *sdmx.io* initiative has established a dedicated effort known as *sdmx4ds*[3] (SDMX for Data Science). This initiative builds a bridge between the world of official statistics and the growing community of data scientists.

At its core, *sdmx4ds* is a **repository of reproducible analyses**. Hosted openly, it invites data scientists to share Jupyter Notebooks, R Markdown files, or Quarto scripts that demonstrate the use of SDMX in practical applications. Submissions are encouraged to be short, focused, and replicable. Each submission is typically three to five pages in PDF, accompanied by code.

The objectives are clear:

- **Showcase Practical Use**: Demonstrate how SDMX data can be accessed, processed, and analyzed in real workflows.
- **Promote Reproducibility**: Encourage users to publish code alongside analysis, enabling others to replicate and adapt.
- **Foster Collaboration**: Create a community where methods and insights can be exchanged, accelerating innovation.
- **Encourage Citation**: Recognize contributions and embed SDMX more firmly into academic and applied literature.

Importantly, *sdmx4ds* is not confined to official statistics alone. Submissions may combine SDMX data with other sources, illustrating how structured standards can be integrated into diverse analytic contexts.

For advanced users, *sdmx4ds* offers both a learning resource and a platform for visibility. For institutions, it signals that SDMX is no longer just about compliance and reporting — it is about enabling innovation, collaboration, and knowledge exchange in the era of data science.

6. Case Examples and Pathways Forward

The value of SDMX for data science is perhaps best illustrated through concrete examples:

[3] https://github.com/bis-medit-sdmxio/sdmx4ds

- **Macroeconomic Forecasting**: A central bank develops a forecasting model that automatically retrieves interest rate, inflation, and exchange rate data via SDMX APIs. Instead of manual updates, the model refreshes itself with each new release, ensuring forecasts remain current.
- **Policy Dashboards**: A national statistical office builds an interactive dashboard combining SDMX data from domestic and international sources. Policymakers can visualize trends in near real-time, supported by metadata that ensures indicators are interpreted consistently.
- **Climate and Financial Risk Analysis**: A research team integrates climate statistics, financial exposures, and socio-economic data, all structured through SDMX. The resulting models support stress testing of financial institutions under climate scenarios, a critical emerging policy area.
- **Collaborative Repositories**: Through *sdmx4ds*, a data scientist publishes a Jupyter Notebook demonstrating trade data analysis using SDMX APIs. Others adapt it to their contexts, creating a virtuous cycle of learning and reuse.

Examples of analyses published on *sdmx4ds*:

- EU Energy Price Analysis using data from Eurostat and ECB
- Exploring the impact of high cocoa prices on Cameroon's productivity
- Exploring the novel BIS central bank total assets data set

Each of these pathways and examples illustrates the same principle: by embedding rich metadata and standardization at the core, SDMX transforms data from a raw input into a reusable, interoperable asset.

Conclusion

The same SDMX that is already widely adopted by data producers can provide immense value to data users. By combining standardized data structures with rich, machine-readable metadata and programmatic access, SDMX allows analysts to discover, interpret, and integrate official statistics with far less manual effort. For data scientists, this translates into faster data ingestion, more reliable automated pipelines, and greater reproducibility of analysis. As the volume and complexity of available data continue to grow, broader awareness and use of SDMX among the data user community can significantly enhance the efficiency, comparability, and impact of empirical research and policy analysis.

sdmx.io and *sdmx4ds* are efforts to further disseminate the usefulness of SDMX and increase its community. By focusing on open source tools and on open research, these platforms also contribute public goods to the data science community.

References

Araujo, DKG (2023): "gingado: a machine learning library focused on economics and finance", BIS Working Papers No 1122.

Bogdanova, B, B Buffett, B Ligani, I Mustafi, S Nikoloutsos, R Schmidt and B Tissot (2025a): "SDMX adoption and use of open source tools", IFC Report No 17.

Bogdanova, B, M Erdem, B Ligani and O Sirello (2025b): "Leveraging generative AI for official statistics: opportunities, challenges and use cases", 65th ISI World Statistics Congress.

Buffett B and J Napoles (2024): "Enhancing data management through the Statistical Data and Metadata Exchange Standard", Asian Development Bank.

Sirello, O (2024): "Editing metadata with generative AI", background paper prepared for the UNECE Expert Meeting on Statistical Data Editing, Vienna, 7 October.

Wilkinson, M et al (2016): "The FAIR Guiding Principles for scientific data management and stewardship", Scientific data, v.3, n.1, p.1-9.



sdmx.io – where tools and learning resources come together to address official statistics use cases

Douglas Araujo and Brian Buffett

Bank for International Settlements

Rome, 18-20 February 2025

Context

- **Increasing demands for data**: pressure for more agile data organisations, processes, and software
  - AI and advanced analytics only add to the demand

- **SDMX** (**S**tatistical **D**ata and **M**etadata e**X**change) standard:
  - information model and technical architecture that support robust metadata-driven processes…
  - … like **data analysis** for one-off use (eg, research) or for recurrent use (eg, an internal dashboard)

- **SDMX** is reliable and trustable:
  - an ISO standard with a **well-defined governance**
  - Sponsored by BIS, ECB, EUROSTAT, ILO, IMF, OECD, UN, WB
  - Used routinely by various authorities to produce and disseminate **official statistics**

- But for most data scientists in economics, **SDMX is still unknown and has high barriers to entry**

SDMX – Only for Data Producers?

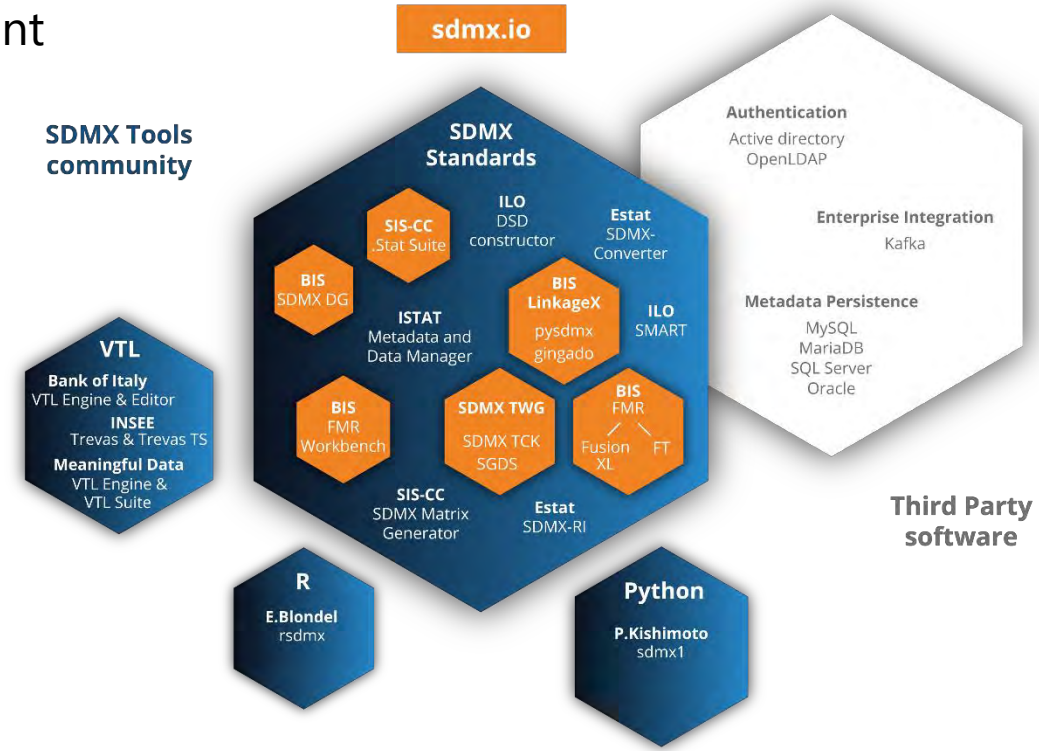
- **Rich Metadata Support:** Metadata which describes datasets in detail, including variables, units, methodologies, and data sources. This helps data scientists:
  - Quickly assess data relevance.
  - Understand data structure without diving deep into raw files.
  - Ensure consistency across datasets for time series or comparative analyses.
- **Improved Data Discovery:** Metadata enables efficient searching and filtering of datasets based on specific criteria like time periods, geographic regions, or indicators.
- **Standardized Data Models:** Common formats for sharing statistical data, making it easier to integrate datasets from different sources (e.g., BIS, ECB, IMF, central banks).
- **Automation & API Integration:** APIs allow automated data retrieval and updates, which is crucial for dashboards or monitoring systems.
- **Interoperability:** Facilitates seamless data integration across various tools and platforms commonly used in data science (Python, R, BI tools).

Enter SDMX iO

- Goal is to facilitate the use of SDMX-based solutions
  - throughout the data lifecycle (eg, production -> dissemination -> use)
  - and for organisations of all sizes and stage of development

- To achieve this goal, **sdmx.io** is an ecosystem of resources:
  - open-source tools,
  - pre-configured containerised environments,
  - learning material (tutorials, webinars),
  - and more





What about data science?

- SDMX is typically used by authorities responsible for producing and disseminating statistics

- But it has ample potential for use in data science as well (Araujo, 2023):
  - Standard definitions of data and metadata modelling can help users find relevant data…
  - … from reliable sources …
  - … in a programmatic and reproducible way.

- **sdmx.io** will increasingly also cater to this target user group of end-users / data scientists

- Today, we are introducing a new series **sdmx4ds** (SDMX for Data Science)

sdmx4ds: a repository of data science work using SDMX

- Data scientists can share their work that uses SDMX (and potentially other data sources)
  - Citation is encouraged

- Goal is to publish Jupyter Notebooks in python, R, Julia showcasing:
  - an interesting analysis that uses data downloaded via SDMX
  - How to download data with SDMX? gingado sdmx1

- Submission process:
  - All public, via GitHub
  - Submissions should be short (circa 3-5 pages of PDF)
  - A template offers guidance on structure

Live demo of the submission process

Next steps

- Call for papers – coming out soon. Keep an eye on **sdmx.io**
- Publication
  - Lightly-edited Jupyter notebooks / Rmarkdown / quarto file in GitHub repo
- Easy for other users to fork and adapt to their own cases
  - 

 button allows for quick replication and customisation by readers



Thank You!

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

SDMX enabled users' processes at the Bank of Italy[1]

Attilio Mattiocco and Valentino Pinna,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

SDMX enabled users' processes at the Bank of Italy

4th IFC Workshop on Data Science in Central Banking
Rome, 18-20 Feb 2025

Attilio Mattiocco
IT Architect, Banca d'Italia
Valentino Pinna
IT Expert, Banca d'Italia

SDMX @ Bank of Italy

- The Bank of Italy collaborated to the creation of the standard and it continues to work on its evolution, being part of the governing bodies (Technical and Statistical Working Group) and producing Open Source Software which is made available to the user community.

- The SDMX Connectors have been presented for the first time in the 2011 SDMX Global Conference. The software has then been released as an open source in 2014.

- During the last 10 years the connectors became a well known tool among data scientists, inside and outside the Bank of Italy.

The SDMX Connectors

SDMX Providers

European Central Bank
OECD
EUROSTAT
IMF
International Labour Organization
Bank for International Settlements

UNICEF
UN Data
World Bank
Pacific Data Hub

ISTAT
Bundesbank
INSEE
Australian Statistics
Statistics Canada



Data Tools











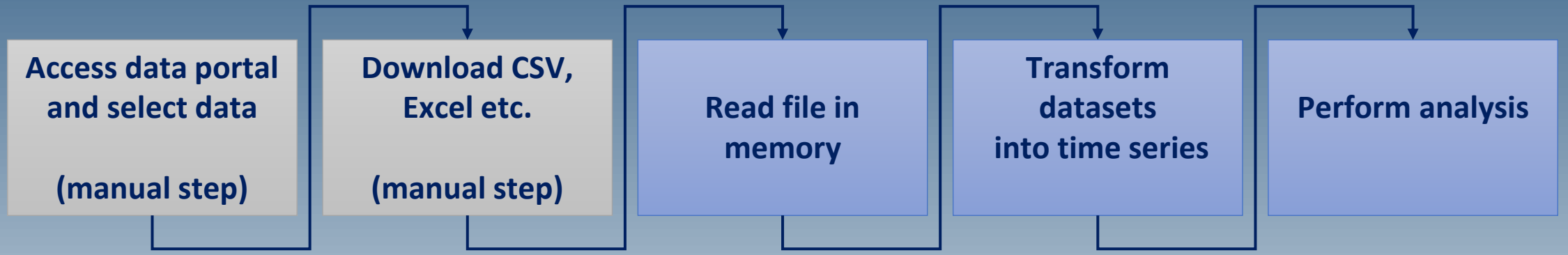




Automation and efficiency in user programs

Main use case: reading time series from a remote provider and seamlessly use them in a data analysis tool (i.e R, Matlab, SAS, Python, Excel, STATA)

Without the SDMX Connectors



With the SDMX Connectors

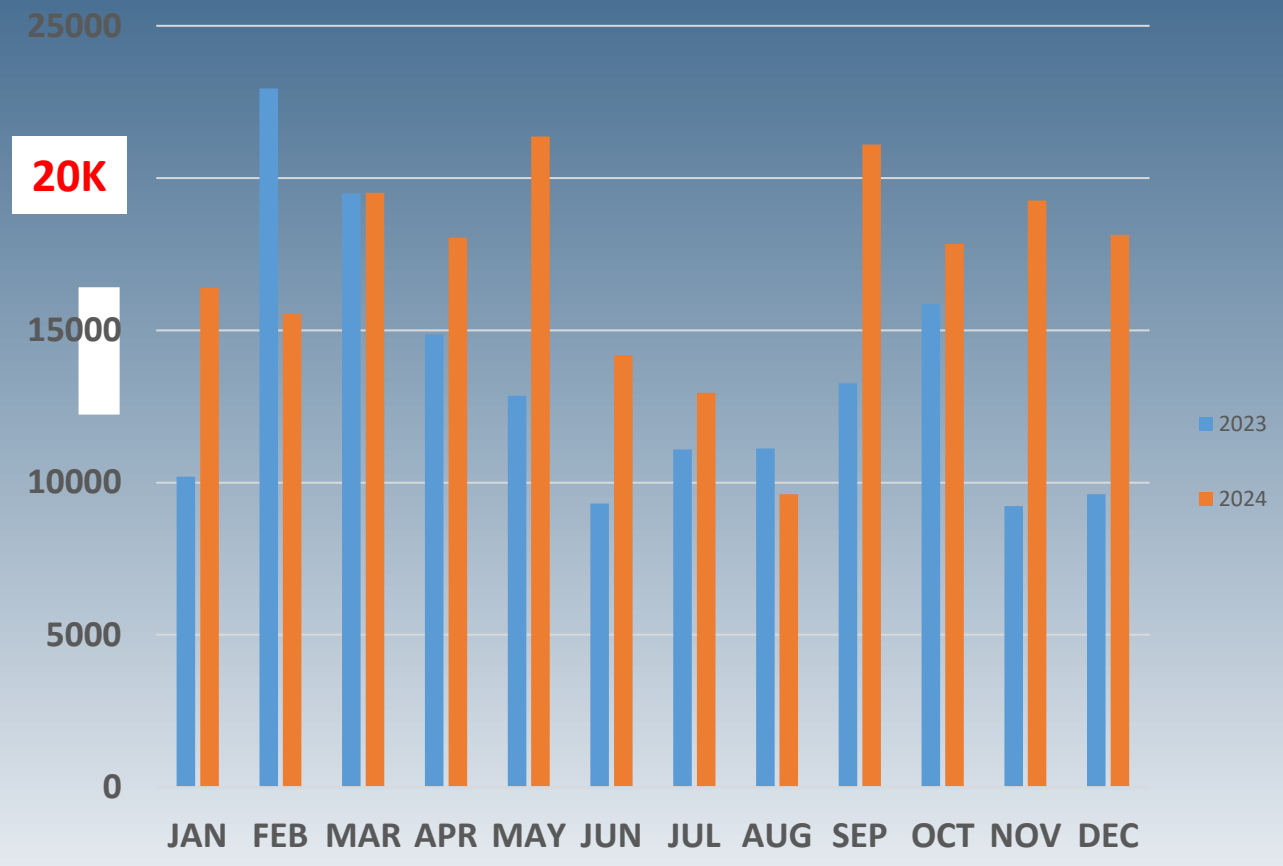


The whole data process can be automated with:

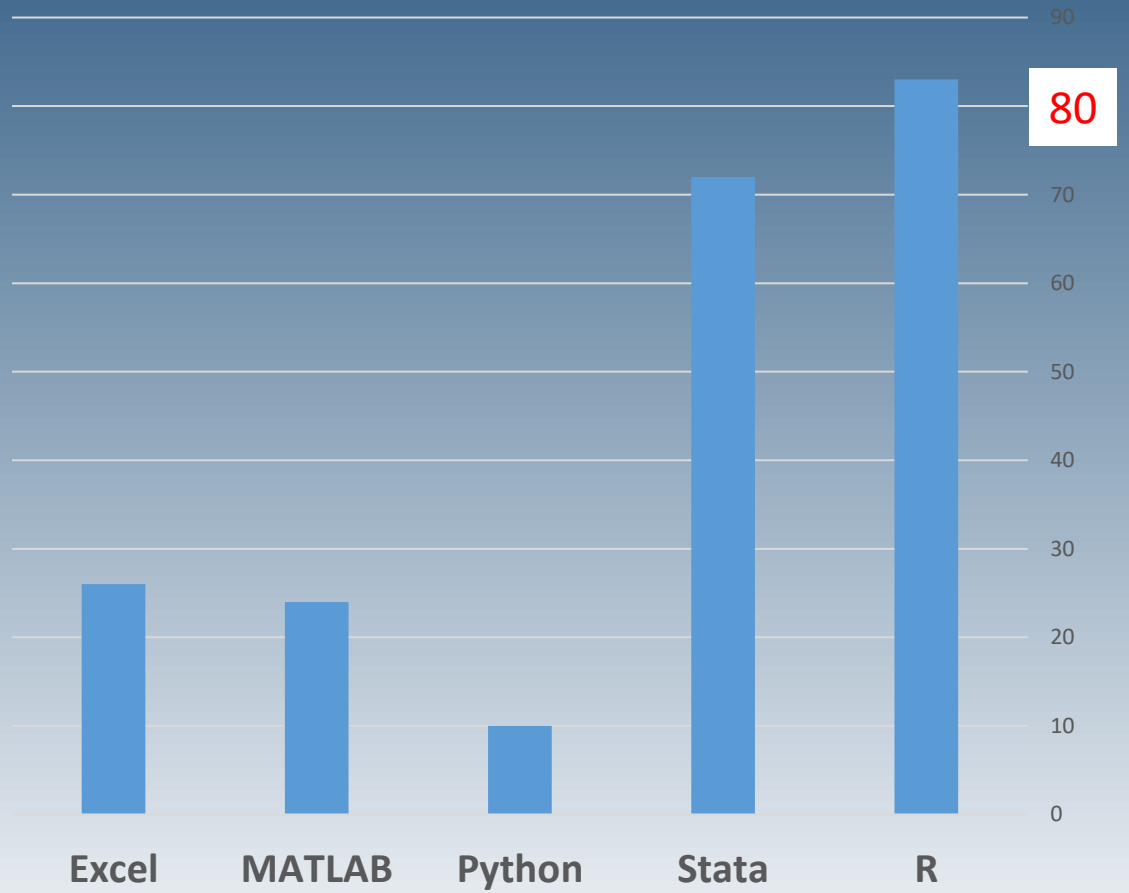
- Fewer risks of **error**
- **Faster** processing
- Increased **transparency** and **reproducibility**

The SDMX Connectors in the Bank of Italy

User Queries



Tools (unique users)



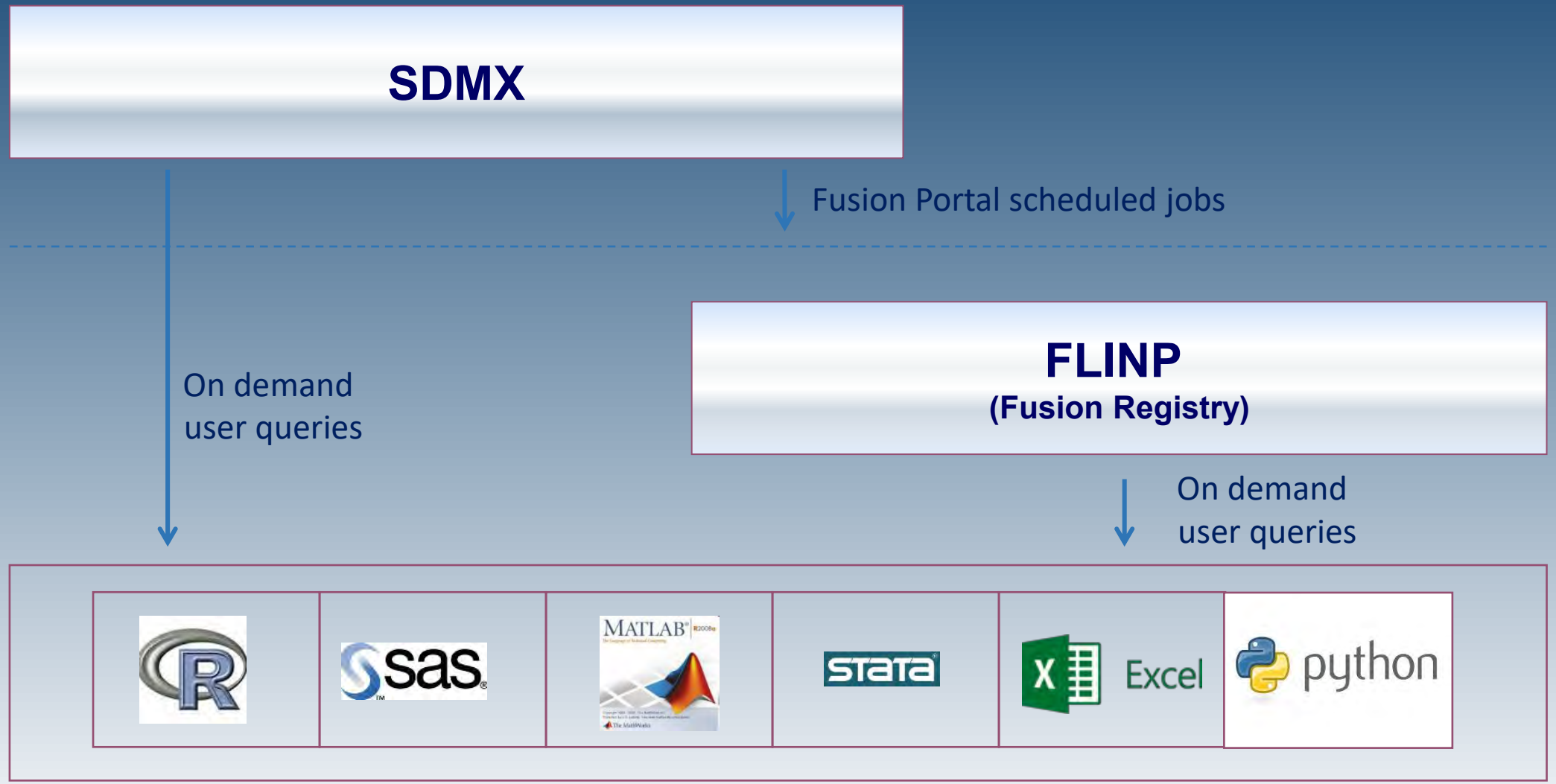
The SDMX Connectors in the Bank of Italy

- Every year at the Bank of Italy, more than **200 users** perform approximately **200K SDMX queries** in their programs
- **Efficiency, robustness, ease of use**
- **Data quality** (metadata, reproducibility)

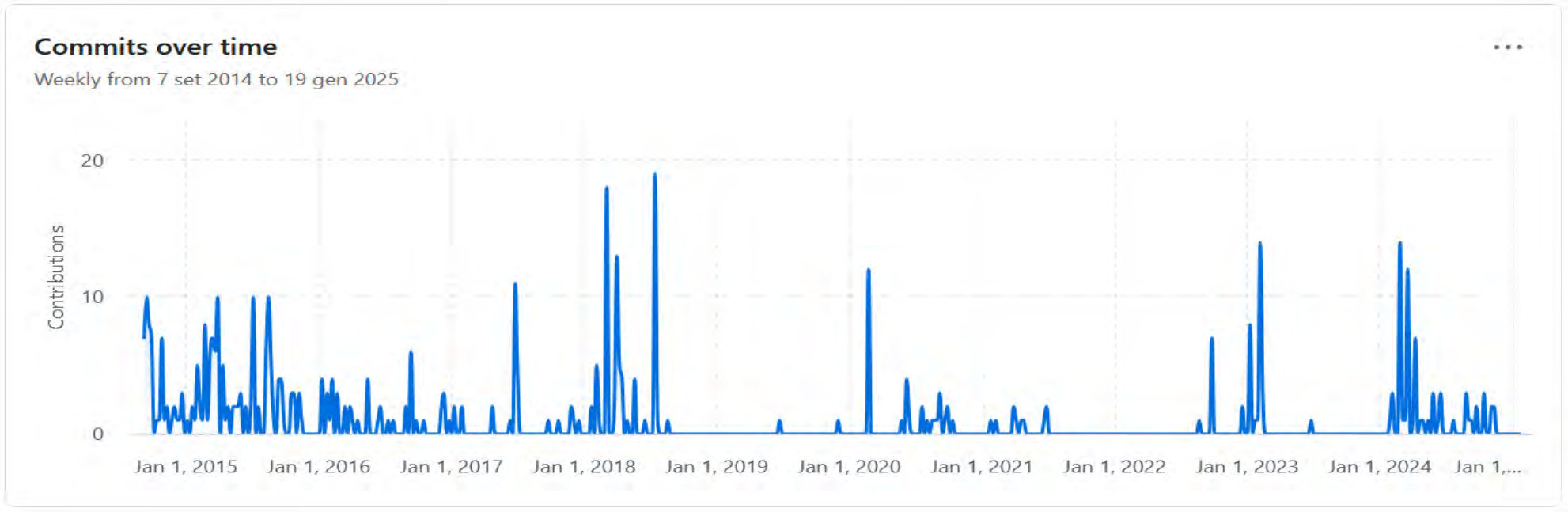
The SDMX universe is perceived by users as an **open and flexible database of certified statistical data**, equipped with a common language that makes it as usable as in-house managed databases (and in some cases, more usable), but *"with great power comes great responsibility"*

- Need for clear and consistent governance and a dissemination **strategy** towards users
- Data users must be considered an integral part of the dissemination process: a **symbiotic** relationship is established between the user and the data producer
- **Changes** (technical and structural) must be communicated in a timely manner
- **Downtimes** are a big problem for users

FLINP: internal SDMX provider for critical data



Ten years of the Open Source Project



- **23 releases** in Github (pushed to MatlabCentral and CRAN as well)
- **199 Issues** closed
- **65 Pull Requests**
- **46 forks** in Github
- **55.000** downloads from CRAN (RJSDMX)
- **5.000** downloads from github (mainly MATLAB and STATA)

What next?

- **SDMX 3:** the API are already supported, but not the new formats. The **sdmx-core** library (that we already use in the VTL Engine) will probably be used
- **FLINP:** the solution is far from perfect, but we think the strategy is correct. The evolution of this setting is under investigation
- **AI:** today one of the most popular features of the SDMX Connectors is the SDMX Helper: an interactive data/metadata browser that simplifies users in searching the data they need. AI will play an key role in this use case
- **VTL:** we foresee a number of use cases where SDMX data and VTL rules will be used in data tools. Collaboration with the **sdmx.io** community has just started.

Thanks!

(questions?)

https://github.com/amattioc/SDMX

https://github.com/vpinna80/VTL

https://www.sdmx.io/

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Supporting users in seeking data on the ECB data portal: a use case for retrieval augmented generation[1]

Luca Petracca, Simone De Benedictis,
Thomas Gottron and Zlatina Hofmeister,
European Central Bank

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Supporting users in seeking data on the ECB Data Portal: a use case for Retrieval Augmented Generation

Luca Petracca, Simone De Benedictis, Thomas Gottron, Zlatina Hofmeister[1]

Abstract

The European Central Bank (ECB) Data Portal holds vast economic data. However, users occasionally struggle to locate specific information due to inefficient search functionality based on traditional information retrieval methods. This is particularly true for users who are unfamiliar with economic terminology or the dataset structures available on the EDP. To address this, we propose a chatbot interface that leverages Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) techniques. The system can understand complex natural language queries, accurately identify relevant datasets and time series, and provide precise data points along with contextual information and sources. Focusing on high-demand time series, initial evaluations demonstrated high accuracy in dataset identification and series key provision, showcasing the potential to enhance data retrieval on the ECB Data Portal.

Keywords: data portal, artificial intelligence, large language models, retrieval augmented generation, information retrieval

JEL classification: C45, C55, C67, C82

[1] Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Contents

Introduction 3

Background and Related Work 3

The architecture for a RAG solution 4

- User Query Preprocessing 5
- Information Retrieval using Semantic Search 6
- Response Generation 10
- Building a User Interface 11

Evaluation 12

- Setup 12
- Results 14

Summary and conclusions 21

References 22

Introduction

The ECB Data Portal[2] hosts vast amounts of economic data, but users often struggle to locate specific information efficiently. Traditional search methods can be inefficient, especially for users unfamiliar with economic terminology or dataset structures. There is a clear need for a more intuitive, user-friendly interface that can understand and respond to queries about economic indicators and time series data using natural language. Enhancing access to ECB data is crucial for researchers, policymakers, and the public to make informed decisions based on accurate economic information. We investigated the option of a chatbot interface that can understand complex queries and provide precise data points, along with contextual information and sources.

Developing an effective chatbot for economic data retrieval poses several challenges. The system must accurately understand the user's information need, identify the right dataset and relevant time series, and retrieve and provide responses that are both precise and contextually relevant. Furthermore, it needs to handle heterogeneous time series data, including various indicators, timeframes, and metadata, while ensuring data accuracy and retrieval.

Previous approaches to improving data portal accessibility have often relied on keyword-based search engines or categorization systems, which lack the flexibility to interpret natural language queries effectively. While Large Language Models (LLMs) have demonstrated promising performance in general question-answering tasks, their application to specialized domains like time series data retrieval has been limited. This project proposes an integration of LLMs and Retrieval-Augmented Generation (RAG) techniques specifically for the ECB Data Portal's requirements.

Our approach combines latest state-of-the-art models for natural language understanding and response generation with RAG techniques to enhance the model's knowledge of ECB datasets and time series. The system processes user queries through several stages: query refinement, query embedding, index search, dataset and time series matching, and response generation with source attribution. We implemented this solution focusing on high-demand time series. Initial evaluations demonstrate high accuracy in dataset identification and series key provision showcasing the potential of this approach to drastically improve economic data retrieval on the ECB Data Portal. Finally, as for all LLM based applications, we looked at safeguarding mechanisms to ensure a safe operation of the system.

In this paper we discuss the technical setup of the overall system, illustrate the design of the RAG component, explain the evaluation methodology and present our insights and findings.

Background and Related Work

Lewis et.al. (Lewis, et al., 2020) introduced the concept of retrieval augmented generation (RAG) as a solution for knowledge intensive NLP tasks that need to

[2] data.ecb.europa.eu/

incorporate specific domain knowledge. Since, the concept has received a lot of attention, especially for use cases where external knowledge needs to be incorporated into the NLP task, but otherwise an out-of-the-box LLM without further fine-tuning is sufficient. Researchers have invested much effort in optimising the naïve RAG approach and designed advanced and modular solutions to address and overcome certain challenges and shortcomings. The survey presented by Gao et.al. (Gao, et al., 2024) provides a good overview of the recent developments.

Interacting AI agents are a common concept for complex solutions. They cover a wide range of use cases and diverse methodological approaches. The idea to base AI agents on LLMs as core method has been investigated for various use cases (Xi, et al., 2023). In the context of combining them for the sake of NLP tasks or, more specifically even, for RAG solutions the AI agents are also referred to as modules (Gao, et al., 2024). An interesting observation is that the combination of multiple modules with specific purposes often results in better system performance, compared to a monolithic LLM. The modules perform tasks like query refinement (Ma, Gong, He, Zhao, & Duan, 2023) or introducing a judge to assess and enhance the generation of the final response (Shao, et al., 2023). Furthermore, small and specialised models with a narrow scope and task in the architecture pipeline seem to provide good contributions in a context of a wider task (Lewis, et al., 2020).

Evaluation of systems based on LLMs and RAG is a complex and evolving topic. Traditional evaluation metrics like recall and precision (Baeza-Yates & Ribeiro-Neto, 1999) are of limited use and commonly applied only to the retrieval component. The overall system performance is sometimes assessed using classical text summarisation metrics like ROUGE (Lin, 2004). More advanced evaluation approaches like RAGAs (Es, James, Espinosa-Anke, & Schockaert, 2024) seek to assess faithfulness, answer relevance and context relevance or involve LLMs as reliable judges for grading the system output (Zheng, et al., 2023).

The architecture for a RAG solution

In this section we describe our approach, the architecture, and technical details. In particular, we motivate certain design decisions and explain how the prototype solution interacts with the users and the information from backend systems.

High-level overview of the architecture of our RAG based solution Figure 1

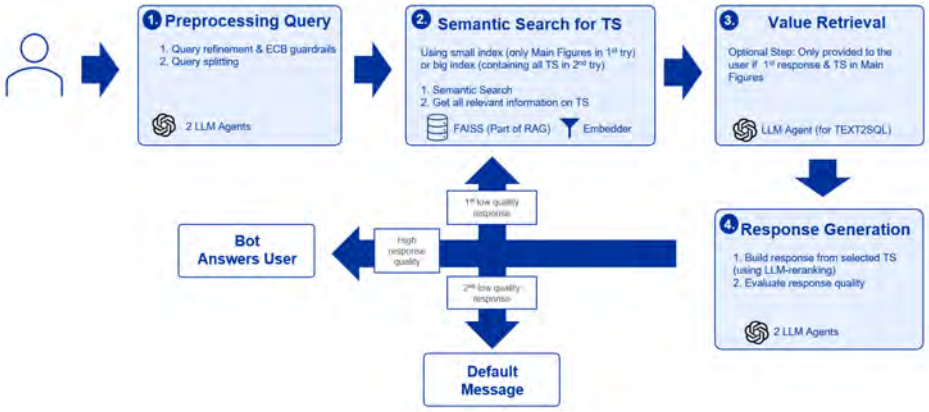


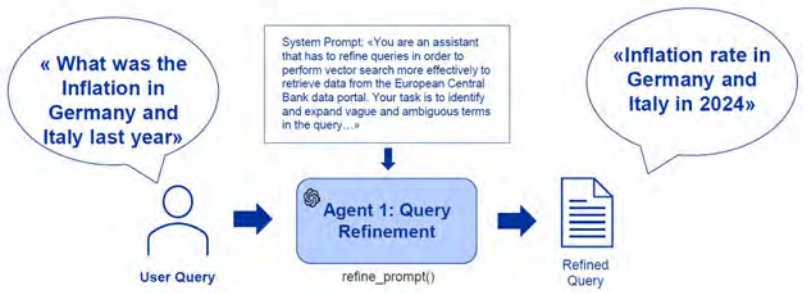
Figure 1 gives a high-level overview of the architecture of our solution. It illustrates how the prototype is leveraging a pipeline of different agents to achieve the intended result and output. Each agent is responsible for a specific task in the overall pipeline. This setup has proved more robust performance compared to simpler and more generic setups we have tested during the development. Several agents within this pipeline leverage LLMs, each configured with specific system prompts that guide their individual task and – whenever possible – examples to implement few-shot learning.

User Query Preprocessing

The first agent in our pipeline is responsible for query refinement. As users interact with the system through natural language, their queries may contain ambiguities, colloquial terms, or lack the precise terminology expected by the data portal's search engine. Therefore, the primary purpose of the *Query Refinement Agent* is to standardize and improve the quality of user-provided input. For example, as illustrated in Figure 2 the agent processes a user query like "*What was the Inflation in Germany and Italy last year?*", by interpreting phrases such as "*Inflation*" as "*Inflation rate*" – the more formally recognized term in economic data contexts. Similarly, relative temporal references, like "*last year*", are converted to specific and absolute time periods. The system prompt driving this agent directs the LLM to act as a helpful expert in preparing queries for optimal data retrieval from the ECB Data Portal. The outcome of this agent is a *Refined Query*, like "*Inflation rate in Germany and Italy in 2024*", that is clearer and more structured. This harmonization step bridges the gap between the various ways users might express their information needs and the system's structured understanding of economic data concepts within the EDP.

Query Refinement Agent

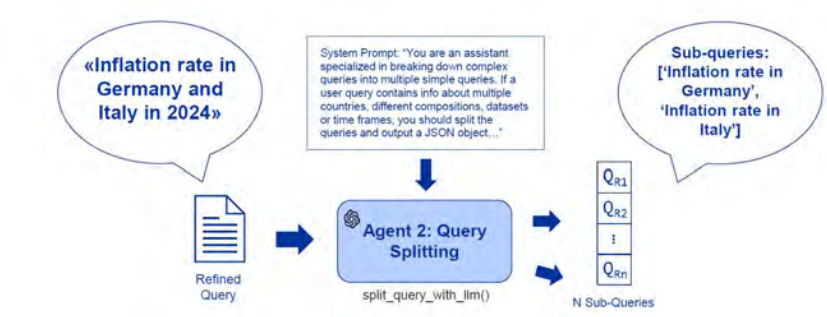
Figure 2



Subsequently, the second agent, takes the *Refined Query* as input and returns a list of multiple, distinct user requests. The key function of this agent is to recognize when a single user question is actually asking for several pieces of information simultaneously, such as when data is requested across multiple geographical regions, for different economic indicators, or across varying time frames.

Query Splitting Agent

Figure 3



For instance, as shown in Figure 3, for the *Refined Query* "*Inflation rate in Germany and Italy in 2024*", the agent's job is to understand that the user is interested in two distinct pieces of information: the inflation rate for Germany and the inflation rate for Italy. To handle these types of queries, the *Query Splitting Agent's* system prompt is instructed to analyse the refined query and, when appropriate, segment it into *n Sub-Queries*. In the given example, the agent effectively breaks down the initial refined query into a list of discrete, simpler sub-queries such as ['Inflation rate in Germany', 'Inflation rate in Italy']. These sub-queries are now easier to manage and address in the following data retrieval steps. Again, in this way we adjust the query in a way that still reflects the user's information need but is conceptually closer to how data is stored in the EDP and captured in our search index.

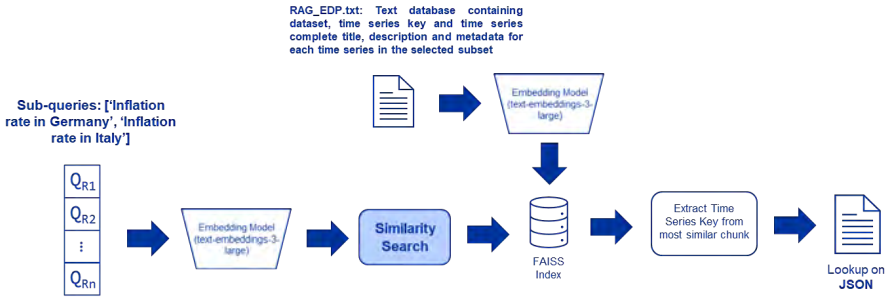
Information Retrieval using Semantic Search

After the query refining and splitting steps, the process continues with the retrieval component, a critical element in our architecture designed to efficiently find the relevant data from the ECB Data Portal. As shown in Figure 4, the initial operation

within the retrieval component involves the transformation of the sub-queries into a vector-based representation. Specifically, each sub-query is embedded via the *text-embedding-3-large* model. This embedding process is crucial for converting natural language text into high-dimensional vectors that numerically encode semantic meaning. The choice of text-embedding-3-large is justified by it being one of the state-of-the-art embedding models available on the market according to several benchmarks (OpenAI, 2024).

Similarity Search

Figure 4



After embedding sub-queries, we perform a similarity search (cf. Figure 4). In a **first-stage search** we use an index built using metadata embeddings derived *exclusively* from time series categorized as "Main Figures" on the ECB Data Portal – representing the top 1000 most frequently demanded and accessed datasets on the EDP website, as shown in Figure 5. The rationale for this focused index is based on the idea that most user queries are likely to be related to the most relevant macroeconomic indicators. By initially searching within this constrained, high-relevance index, the system aims to rapidly identify data for frequent and standard user requests and to provide a better and faster user experience. The similarity search uses cosine similarity to identify the top-k most similar time series keys.

Main Figures Index

Figure 5



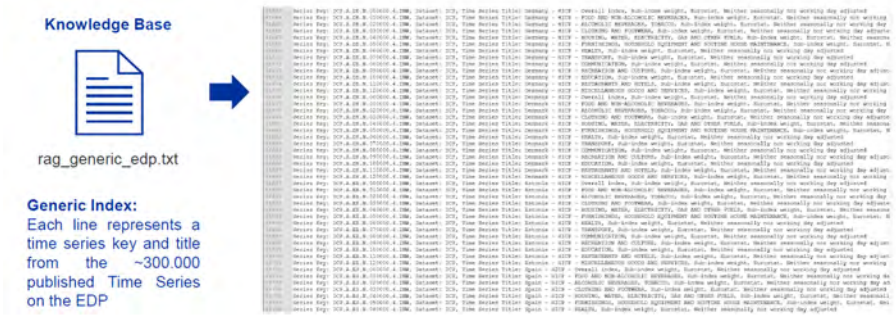
Based on the retrieved time series from the Main Figures, we generate a preliminary answer. This answer is subjected to evaluation by a **Judge Agent**. As elaborated below, the Judge Agent assesses the relevance and quality of the generated response against the original user query. If the Judge Agent's rating is

deemed **satisfactory**, the system proceeds with this initial retrieval set to construct and deliver a "Specific Answer" to the user.

However, if the **Judge Agent assess the answer as not satisfactory**, we conclude that the initial search on Main Figures did not provide relevant results. In this case the system reverts to a **second-stage search**. This secondary search is performed using a broader index. This index is constructed from metadata embeddings representing a much larger subset of the EDP data – encompassing approximately 300,000 time series keys, as shown in Figure 6. This index covers a significantly larger collection of datasets, including those featured in official ECB publications, reports, and analytical articles. Hence, it is designed to capture a more comprehensive representation of the EDP's data offerings, providing a fallback mechanism when user queries extend beyond the most commonly accessed "Main Figures". Technically, this second-stage search follows the same process, using a cosine similarity-based search is performed to retrieve the top-k most similar time series keys.

Generic Index

Figure 6



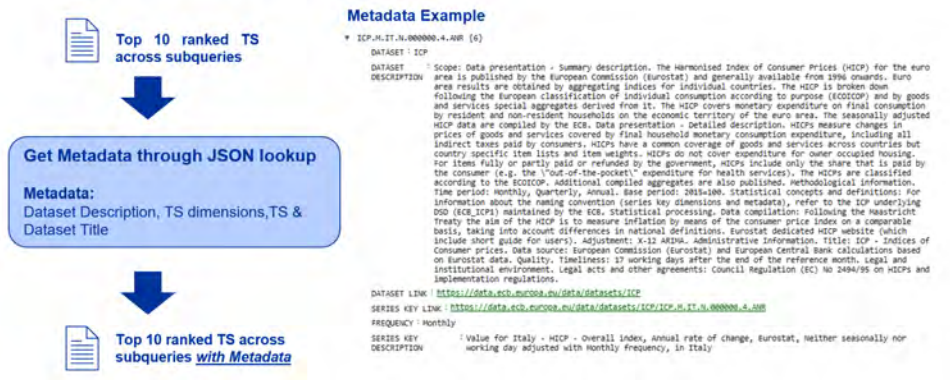
The answer generated from time series retrieved from the *"Generic Index"* is *again* presented to the **Judge Agent** for assessment. As before, if the Judge Agent assigns a **satisfactory rating**, the system proceeds to construct a "Generic Answer" based on this second retrieval set. Conversely, if the Judge Agent's rating remains **unsatisfactory** even after searching the broader index, the system concludes that it cannot provide a satisfactory answer. In this failure scenario, the system defaults to generating a **generic error message**, informing the user that no suitable dataset has been identified. This two-tiered index and judgment-driven retrieval process aims to balance search efficiency and response speed for common queries with the capability to perform more comprehensive data exploration when necessary, guided by intelligent quality assessment via the Judge Agent.

To implement efficient and scalable retrieval, we leverage the **FAISS (Facebook AI Similarity Search) index** (Douze, Guzhva, Deng, Johnson, & Szilvasy, 2024). FAISS is optimized for high-dimensional vector similarity search and provides the necessary infrastructure to rapidly compare query embeddings against the large collection of embeddings representing time series metadata. FAISS supports an efficient computation and retrieval based on **cosine similarity** scores, efficiently identifying the top-k most semantically similar time series metadata embeddings to the input query embedding (Manning, Raghavan, & Schütze, 2008). One of the key differences from a standard Retrieval-Augmented Generation (RAG) approach lies in how we

handle information *after* the initial retrieval step. Most RAG systems incorporate the retrieved top-k text snippets as immediate input for the next stage. However, our priority is to ensure the accuracy of the information and the depth of contextual understanding used to build our responses. Therefore, for each of the identified relevant time series keys we perform an extended **metadata lookup**, as illustrated in Figure 7. In our prototype setup, this is achieved by directly querying a structured JSON file, which serves as our in-house, detailed repository of metadata. Ultimately, this design choice is intentionally aimed at minimizing the risk of introducing incomplete information or hallucinations.

Metadata Retrieval

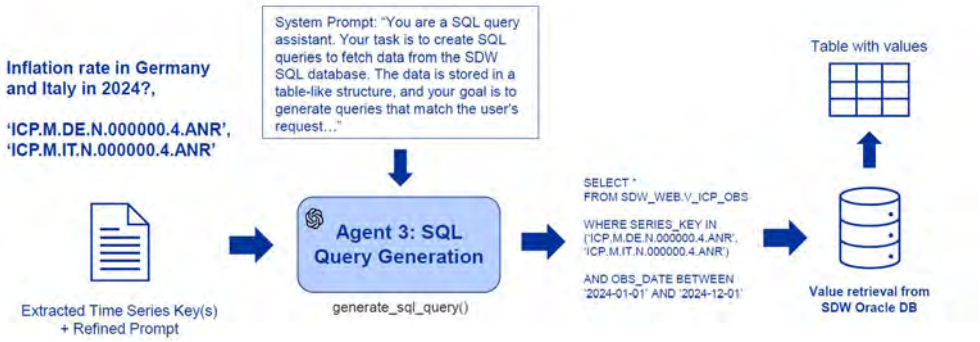
Figure 7



After we have retrieved the relevant series keys and metadata from the JSON, we still need to retrieve values to answer the user's query. The **SQL Query Generation Agent** (Agent 3), as shown in Figure 8, uses the extracted time series key(s) and the refined query prompt(s) as input to produce and execute an SQL query for fetching precise values from the European Central Bank's Secure Data Warehouse (SDW) DB – the backend data repository for the EDP.

To this end, we use a consistent SQL query template. The agent's task is to "fill in the blanks" in the template. The "blanks" correspond to the list of **time series keys** and consider any **temporal boundaries** specified by the user. For example, given the user query "*Inflation rate in Germany and Italy in 2024?*", the agent should extract "*2024*" as the relevant timeframe and refer to the time series keys *ICP.M.DE.N.000000.4.ANR* and *ICP.M.IT.N.000000.4.ANR* as identified in the retrieval phase. In general, the agent translates these elements into appropriate SQL WHERE clause conditions.

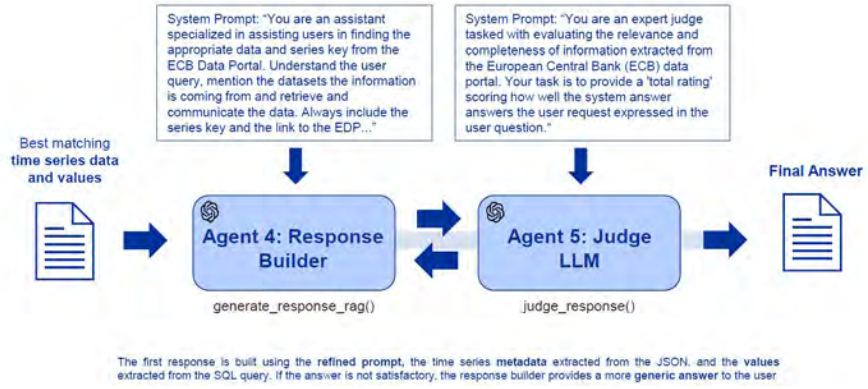
SQL Query Generation Figure 8



Response Generation

The response builder agent brings together the metadata for the timeseries and the retrieved values to construct the actual response for the user. This agent takes as input the **best matching time series data and values** extracted via the SQL Query Generation Agent, along with the associated time series **metadata**. We instructed the LLM to re-rank the sorted list of extracted time series keys according to the original user request, to add an additional layer of *reasoning* before providing the data to the Response Builder agent. The agent is then tasked to follow a specific template for generating the answer to ensure a harmonised and standardised response type.

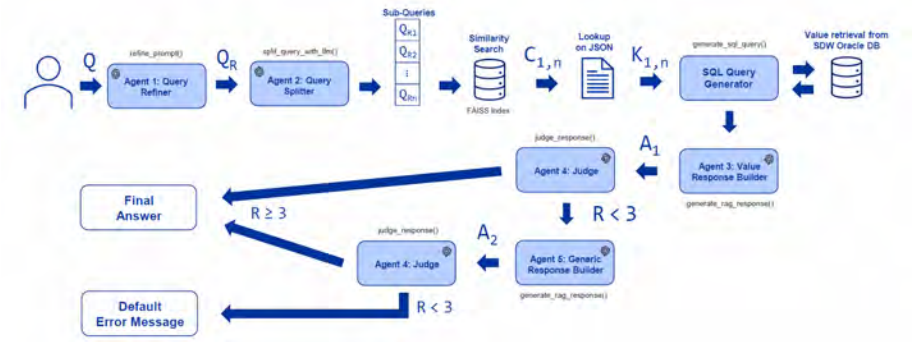
Response Generation Figure 9



Finally, we incorporated a **Judge Agent** to assess the answer before it is delivered to the user. The purpose of this judge is to check the generated answer and assess how well it reflects the user's initial query and information need. The outcome of this judgement can take several forms. In the best case, the generated answer in considered a good or very good fit to the user's query and is accepted as is. If the answer is considered a mediocre or bad fit, the judge can relate to a modified response builder agent for generating a more generic answer that does not include data and time series values, but only refers to datasets which might be of relevance to the user. If the judge also deems this generic answer not suitable for replying to

the user, a generic error response is produced, informing the user, that the chatbot is incapable of finding a suitable dataset to the query. As indicated in Figure 9, this might involve a few interactions between the response builder and the Judge Agent. Figure 10 summarizes the response generation flow with the rating-based Judge-as-LLM system.

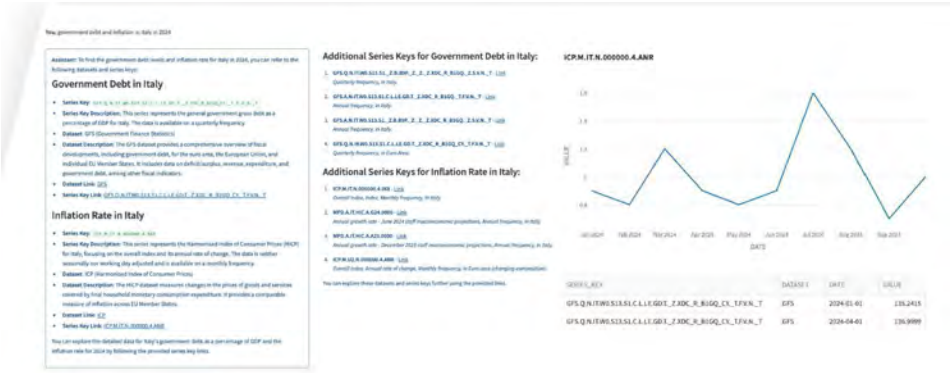
Response Generation Flow Figure 10



Building a User Interface

To provide users with a practical and accessible way to interact with our system, we developed a user-friendly interface using **Streamlit**, an open-source Python library (Streamlit, 2025). This enabled us to rapidly prototype an interactive chatbot interface without extensive front-end development. The resulting interface, as shown in Figure 11, is inspired by a chat-like environment where users can intuitively formulate their natural language queries and receive structured, informative responses.

Chatbot Interface Figure 11



The answers are provided following a consistent and structured format, designed to provide comprehensive information to the user within a single response. First, the response prominently includes the **extracted metadata** for the identified relevant time series. This typically encompasses concise **Dataset and Series Key Descriptions**, offering textual context for the data. Crucially, we also incorporate actionable **hyperlink** elements in the response, enabling users to navigate to the corresponding

pages on the ECB Data Portal for more in-depth exploration and verification. Beyond this core metadata, the interface also presents a list of the **top 5 'Additional Series Keys'** that the system deemed potentially relevant to the user's query. This provides users with alternative data series that might also be of interest and addresses potential ambiguity in the user query. Finally, the interface displays the actual data values in tabular or graphical format, depending on what seems more appropriate given the user query and retrieved data.

Evaluation

To assess the efficacy of our RAG based system in supporting users seeking data on the ECB Data Portal, we designed and conducted a comprehensive evaluation. Our evaluation framework focuses on realistic user needs and is based on quantitative metrics aligned with best practices in information retrieval. This section details the evaluation setup, including the dataset of queries and gold standard answers used, the rationale behind our evaluation methodology, and the specific metrics employed to quantify system performance.

Setup

A key aspect of our evaluation methodology is the creation of a large and diverse test dataset, consisting of **525 Query & Answer (Q&A) labelled pairs**. This dataset was specifically constructed to represent the complexity and diversity of real-world user information needs when interacting with the ECB Data Portal. We leveraged various sources like query logs and hotline requests to compile common user queries. Domain experts with the relevant business knowledge constructed the gold standard of what answer we would expect from an ideal system.

We categorized the 525 queries into three distinct types:

- **Type 1: Retrieving Values (55% of dataset):** These queries represent users explicitly seeking specific numerical values for particular time series. Examples include requests like "What was the inflation rate in Germany and Italy in January 2024?".
- **Type 2: Finding Data (30% of dataset):** These queries represent users aiming to discover relevant datasets and time series, without seeking specific values. Examples include "Where can I find data on the inflation rate in Germany and Italy?".
- **Type 3: Off-Topic (15% of dataset):** This category includes queries that fall outside the intended scope of the EDP chatbot – for instance, general conversational queries, requests unrelated to economic data, non-objective assessments, or questions about non-existent datasets. These 'Off-Topic' queries are crucial for assessing the robustness of the system and its ability to handle out-of-scope user inputs.

This structured distribution across query types is designed to evaluate the system's performance not only in retrieving specific data values but also in guiding users through the data portal, addressing various information seeking behaviours.

The relative proportions of each query type (55%, 30%, 15%) in our test dataset broadly mirror our understanding of typical user interactions and information needs within the EDP context.

Aligned with these distinct user query types, our system is designed to generate correspondingly different types of responses. We aim to deliver both **qualitative** and **quantitative** answers depending on the user's query. Therefore, for each query type, we expect the following system responses:

- **Response Type 1 – Value Response (for Query Type 1):** For users seeking specific data values (Type 1 queries), the expected response is a 'Value Response'. This *full* response type is designed to provide both *qualitative* context and *quantitative* data. It will provide the user with:
    - The *datasets and time series keys* relevant to their query, extracted from the Main Figures index, providing metadata and links.
    - The *extracted values* themselves, extracted via the generated SQL query, directly addressing the user's request for specific numerical data.
- **Response Type 2 – Data Response (for Query Type 2):** When users are exploring the data portal to *find* datasets and time series, or just to ask general questions on methodology and economic indicators (Type 2 queries), the expected system output is a 'Data Response'. This answer type focuses on offering a *qualitative* answer, guiding the user towards relevant data sources. It is designed to deliver:
    - *datasets and time series keys* identified as pertinent to the user's search for data and extracted from the Generic Index.
- **Response Type 3 – Default Message (for Query Type 3):** For queries that are off-topic or inappropriate for the EDP chatbot (Type 3 queries), we expect the system to provide a 'Default Message'. We inform the user that their query is outside the scope of the system and no satisfactory answer can be provided.

This clear mapping between query types and expected response types sets the stage for a precise evaluation of our system's effectiveness in addressing different user needs.

Risk by Query-Response Type

Table 1

	Response Type 1	**Response Type 2**	**Response Type 3**
Query Type 1	Good	Ok	Ok
Query Type 2	Ok	Good	Ok
Query Type 3	**Harmful**	**Harmful**	Good

Beyond evaluating the accuracy of data retrieval and the appropriateness of response types, our evaluation framework also considers the potential risks associated with different combinations of query types and system responses. As shown in Table 1, some mismatches between user queries and system outputs could be considered possibly "harmful", especially when the system incorrectly generates a response to an 'Off-Topic' query (Query Type 3). Specifically, we define a response as possibly "harmful" if either of the following scenarios occur:

1. **Expected Response Type 3 (Default Message) but System Produces Response Type 1 (Value Response).** This indicates a scenario where the system, despite being presented with an out-of-scope query, incorrectly attempts to retrieve and present specific data values.
2. **Expected Response Type 3 (Default Message) but System Produces Response Type 2 (Data Response).** Similarly, this denotes cases where an 'Off-Topic' query inappropriately triggers the system to provide a data-oriented response, even if it is only a qualitative Data Response (datasets and series keys), rather than the correct Default Message indicating inability to handle the query.

These scenarios are categorized as possibly "harmful" because they represent potential misinterpretations of user intent by the system. For our use-case, providing an answer to an off-topic query could mislead the user or suggest capabilities beyond the system's intended scope. For the purpose of our evaluation, the assessment of whether a response to a Type 3 query is indeed possibly "harmful" is conducted via manual human review, allowing for a better judgement of the specific context and potential implications of each response.

Results

Our evaluation framework is based on standard metrics used in information retrieval, focusing on measuring retrieval accuracy at different levels. We mainly looked at three levels of accuracies (a) identifying the relevant dataset, (b) identifying the relevant time series within a dataset and (c) identifying the relevant values within a time series:

- **Dataset Accuracy (Top-5):** This metric measures the percentage of queries for which the *correct dataset* (as defined in the gold standard answer for each query) is present within the top 5 datasets retrieved by the system. "Top-5" accuracy is a commonly used metric in ranked retrieval evaluation, reflecting scenarios where users are presented with a ranked list of results and are likely to examine the top few options (Baeza-Yates & Ribeiro-Neto, 1999). We computed this metric both overall across all query types, and separately for Type 1 and Type 2 queries to analyse performance for different information seeking intents.
- **Series Key Accuracy (Top-5):** Complementary to Dataset Accuracy, Series Key Accuracy focuses on the retrieval of the *correct time series key* – the most granular identifier for a specific data series within the EDP. Series Key Accuracy (Top-5) measures the percentage of queries for which the gold standard *series key* is present within the top 5 series keys retrieved by the system. Similar to Dataset Accuracy, the use of a "Top-5" metric is motivated by user interface considerations and the desire to evaluate performance in a

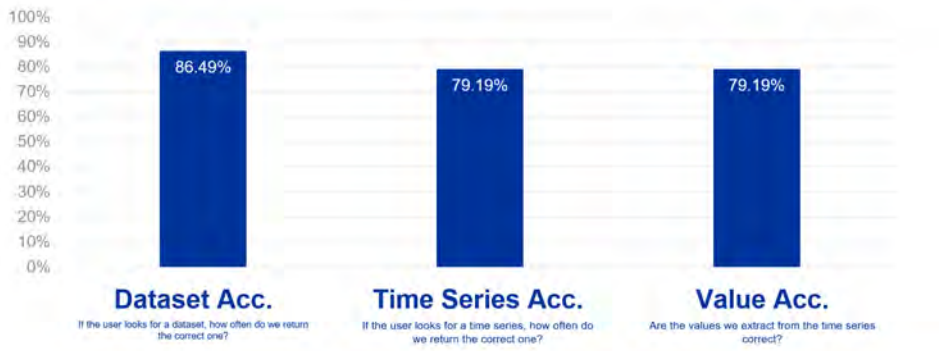
ranked retrieval scenario. This metric is again calculated both overall and per query type (Type 1 and Type 2) to provide a detailed performance breakdown.

- **Value Accuracy:** For queries that refer to a specific time series and pose time constraints (e.g. inflation in 2024), we also assess if the SQL query generation agent was capable to identify the *correct time constraints* from the user query. In this case a result is considered correct if the time constraints precisely match the gold standard provided by human experts and, thus, the also the correct values are retrieved from the SDW backend.

To evaluate the user-perceived benefits of our retrieval system, we analysed the performance metrics specifically related to retrieval accuracy for Query Type 1 and Query Type 2 queries – the query types for which the system is designed to provide responses (Value Responses and Data Responses, respectively). The results for our metrics, evaluated on our dataset of 525 Q&A pairs are summarized in Figure 12.

Top-5 Retrieval Results

Figure 12



Our evaluation reveals strong performance in both Dataset Accuracy and Time Series Key Accuracy. As shown in Figure 12, the **Dataset Accuracy** (Top-5) achieves a score of 86.49%. This indicates that for **86.49%** of all evaluated user queries, the system successfully places the correct dataset within the top 5 datasets retrieved in the ranked list. This high Dataset Accuracy suggests that, in a majority of user interactions, the system effectively directs users to the appropriate broad data categories on the EDP, facilitating data discovery and exploration.

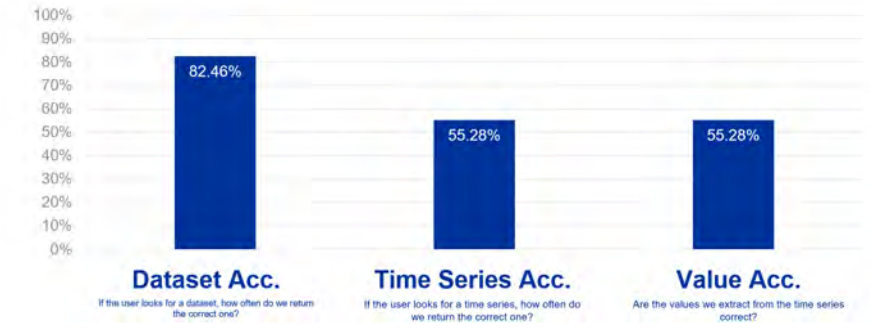
Complementing Dataset Accuracy, **Time Series Key Accuracy** (Top-5) stands at **79.19%.** This metric, representing the ability to retrieve the most specific data identifiers, signifies that for 79.19% of user queries, the specified gold standard time series key is present within the top 5 time series keys returned by the system. This indicates that while the system is highly effective at pinpointing the relevant datasets, narrowing the result down to the single most relevant time series key can present a slightly greater challenge. This is not unexpected, given the subtle semantics that differ closely related time series within our datasets. However, a Top-5 Time Series Key Accuracy of nearly 80% still represents a substantial benefit to users, significantly narrowing their search space within the vast EDP repository.

A particularly noteworthy finding of our evaluation pertains to **Value Accuracy**. As reported in Figure 12, the Value Accuracy score is also **79.19%** – numerically

identical to the Time Series Key Accuracy. This means, that once the time series key is correctly identified, the agent is always capable of extracting the relevant time periods from the user's query. As direct consequence of the 'RAG + Text to SQL' architectural design this directly leads to the system to retrieve and display the correct values for the users. This feature is a significant user benefit, enhancing trust in the reliability and accuracy of the data presented by the system, a crucial factor for the risk assessment of applications involving sensitive economic and financial information like ours.

Top-1 Retrieval Results

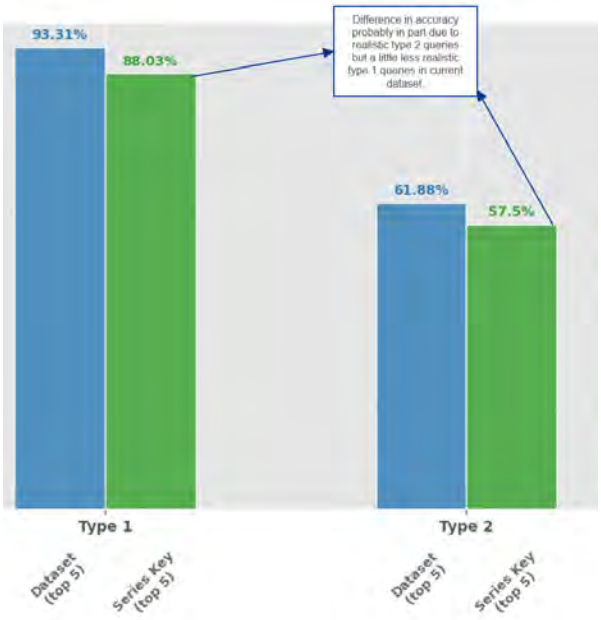
Figure 13



As an additional step, after obtaining these results with a Top 5 retrieval, we wanted to check the granularity of our system's retrieval performance and evaluate a more stringent definition of accuracy. Therefore, we also conducted an additional analysis focusing on **Top 1 retrieval accuracy**. In this case, we consider a retrieval *correct* if and only if the *highest ranked* dataset or time series key returned by the system is the gold standard correct answer. The results for this Top 1 accuracy analysis, also focused on Type 1 and Type 2 queries, are presented in Figure 13. As expected, applying this more stringent criterion results in a decrease in overall accuracy scores compared to the Top 5 metrics. Specifically, **Dataset Accuracy (Top 1)** is measured at **82.46%**, while **Time Series Key Accuracy (Top 1)** and consequently **Value Accuracy (Top 1)** both register at **55.28%**. While these Top 1 accuracy figures are lower than their Top 5 counterparts, they still represent a noteworthy level of performance, especially considering the highly specific nature of the 'correct answer' requirement in this stricter evaluation. A Dataset Accuracy of over 82% at Top 1 rank indicates that, even when demanding pinpoint accuracy in the very first returned result, the system effectively identifies the intended dataset category in the majority of cases.

Model Performance by Query Type

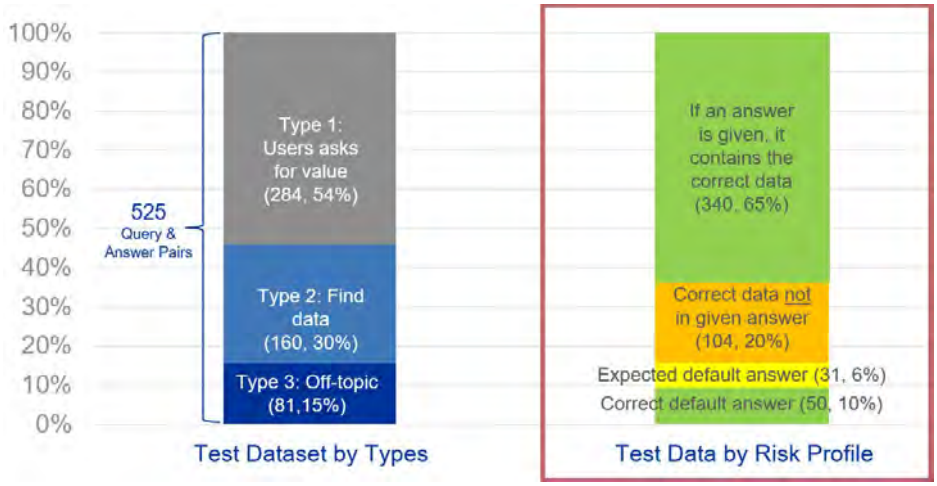
Figure 14



Furthermore, our evaluation reveals a significant distinction in system performance when analysing the results separately for **Type 1 (Retrieving Values) and Type 2 (Finding Data) queries**, as shown in Figure 14. For **Type 1 queries**, we achieve a **Dataset Accuracy (Top 5)** of **93.31%** and a **Time Series Key Accuracy (Top 5)** of **88.03%**. These markedly high accuracy scores indicate that, for queries where users are directly seeking specific data values, our system is highly capable in guiding users to both the correct dataset category and, even more precisely, the intended time series key within the top 5 results. In contrast, for Type 2 queries, the accuracy metrics are lower, yet still respectable. We observe a **Dataset Accuracy (Top 5)** of **61.88%** and a **Time Series Key Accuracy (Top 5)** of **57.50%** for Type 2 queries. This discernible difference in accuracy between Type 1 and Type 2 query types is not unexpected and is likely attributable to the inherent variation in complexity and search scope between these categories. Despite this performance variation between query types, the results robustly demonstrate that our system consistently achieves a meaningful level of accuracy across *both* direct value retrieval (Type 1) and broader data discovery (Type 2) user interactions, adapting effectively to different modes of user engagement with the data portal.

Answer Incorrectness Risk

Figure 15



Beyond assessing user benefits through accuracy metrics, our evaluation framework includes an assessment of various risk categories, particularly focusing on **Answer Incorrectness Risk** and **Hallucination Risk**. The results of this risk-focused analysis, derived from our dataset of 525 Q&A pairs, are presented in , specifically in the "Test Data by Risk Profile" breakdown. Answer Incorrectness Risk, in our context, refers to all those cases where the system's response fails to address the user's information need, or when the system inappropriately attempts to provide an answer for queries it should have declined. We defined Answer Incorrectness based on two primary criteria:

- *Does the returned answer not include the correct Time Series (TS) or Dataset (DS)?* This assesses situations where the system, despite providing a response, fails to identify the expected time series key or dataset relevant to the user's query.
- *Does the bot point towards EDP data for questions it should answer with the default message?* This criterion examines cases where the system incorrectly generates a response instead of defaulting to a generic message when presented with a query that is out of scope or unanswerable using EDP data.

The results of this specific analysis are the following:

- **If an answer is given, it contains the correct data (65%, 340 out of 525 queries):** This majority segment represents queries where, when the system *does* provide a data-driven answer, it successfully includes the correct and expected Time Series or Dataset within its response. This signifies a substantial portion of successful, relevant, and data-accurate answers generated by the system.
- **Correct data not in given answer (20%, 104 out of 525 queries):** In 20% of cases, while the system provides an answer, it *fails to include the specifically correct or expected data series*. This category represents instances of 'incorrect' answers in the sense of failing to perfectly match the gold standard, even though the system is still operating within the EDP data domain.

- **Expected default answer, system gives default answer (10%, 50 out of 525 queries):** For 10% of the test set, corresponding to the 'Off-Topic' (Type 3) queries, the system correctly identifies these queries as outside its scope and appropriately provides the intended *default message*. This indicates successful handling of out-of-scope queries, aligning with expected system behavior.
- **Expected default answer, system does *not* give default answer (6%, 31 out of 525 queries):** A smaller portion, 6% of queries, represents cases where the system *incorrectly* attempts to answer 'Off-Topic' (Type 3) queries with data-driven responses instead of the intended default message. This category highlights the risk of the system misinterpreting or inappropriately responding to out-of-scope user inputs.

Beyond Answer Incorrectness related to data selection, we also investigated **Hallucination Risk**. In our evaluation framework, this analysis focused on assessing whether the system, in its responses, might generate information that is factually incorrect, misrepresents data, or drifts outside the intended scope of the ECB Data Portal's data content. Concretely, our evaluation considered several potential types of hallucinations:

a) *Did the bot mention a topic not relevant to the task?* (Task Irrelevance)
b) *Did the bot make wrong claims (e.g., wrong claims about the ECB's strategy)?* (False Claims)
c) *Does the bot point towards data outside the EDP portal?* (Out-of-Scope Data)
d) *Does the bot invent data and give wrong numbers?* (Invented Data)
e) *Does the bot invent TS titles or data descriptions?* (Invented Metadata)
f) *Does the bot answer questions it should not answer (currently)? E.g., questions about data collection.* (Overreach)

Our analysis for Hallucination Risk was primarily conducted through manual human evaluation of a subset of system responses, particularly focusing on cases flagged for potential answer incorrectness or 'Off-Topic' query handling. Crucially, while a detailed manual assessment for all categories (a-f) remains ongoing, our preliminary analysis has revealed that instances of hallucination of type (c) – pointing towards data outside the EDP portal – were detected in only very few cases. This preliminary finding suggests that our architecture may be effectively mitigating a significant aspect of potential hallucination risk related to data source fidelity. However, a comprehensive quantitative assessment across all hallucination types (a-f) requires further analysis.

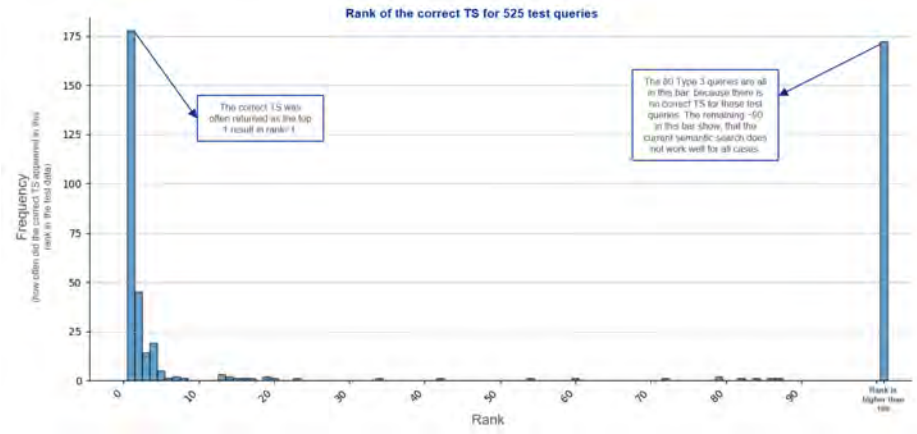
We also acknowledge the presence of **Jailbreak Risk**. This risk category refers to the potential for users to bypass the defined guardrails of the system through adversarial prompting. This could cause the Large Language Models within the system to disregard system prompts, circumvent intended instructions, or potentially produce unintended or inappropriate outputs. While Jailbreak Risk is a recognized concern for LLM-based systems, our evaluation has primarily focused on Answer Incorrectness and Hallucination risks related to core system functionality. A more detailed analysis of Jailbreak Risk, and the implementation of more robust safeguard mechanisms, is an area for future research and development.

Finally, we conducted an analysis of the ranking of the "correct" time series within the retrieved lists. This analysis aims to evaluate how effectively the system ranks the

most relevant time series at the top of the search results. Figure 16 shows a histogram illustrating the distribution of ranks for the gold standard time series across our dataset of 525 test queries. Ideally, for optimal search performance, the "correct" time series should consistently appear at Rank 1 – the very top of the results list.

Analysis of the Search Performance

Figure 16



As depicted in Figure 16, the histogram reveals a highly skewed distribution, strongly concentrated towards the lower ranks. The fullest bar in the histogram is at **Rank 1**, indicating that for a substantial number of queries the gold standard, most relevant time series is indeed retrieved as the top-ranked result by our similarity search process. Following Rank 1, the height of the bars rapidly diminishes as the rank number increases, showing that when the correct time series is not the top result, it is still typically found within the higher ranks (rank 2, 3, and to a lesser extent rank 4, 5 etc.). This distribution pattern is highly encouraging, as it confirms that our semantic search is generally effective at prioritizing and placing the most relevant time series towards the beginning of the ranked result list, aligning well with typical user browsing behaviours where users tend to focus on the top few search results.

However, the results show a noticeable bar on the far right, labelled as "Rank is higher than 100." As annotated on the slide, "The 80 Type 3 queries are all in this bar, because there is no correct TS for these test queries." Indeed, all 80 'Off-Topic' (Type 3) queries are categorized within this "Rank higher than 100" group, as, by definition, there is no 'correct' time series for these out-of-scope requests. However, the remaining count in this far-right bar represents instances, *even for in-scope queries (Type 1 & Type 2)*, where the semantic search failed to rank the correct time series within the top 100 results. This means that for a segment of queries, roughly ~90 out of the 525 in our dataset, our current semantic search approach does not effectively find the intended time series. This highlights an area for further investigation and potential improvement in our semantic search mechanism to enhance recall and ranking robustness across a broader range of user queries and data series within the EDP.

Summary and conclusions

Our experience developing a RAG-based system for the ECB Data Portal yielded several key insights, both regarding the strengths of our current approach and opportunities for further improvement.

A primary observation was the clear advantage of our multi-agent architectural design compared to simpler, single-agent systems. By decomposing the complex data retrieval task into specialized components—such as Query Refinement, Query Splitting, SQL Generation, Response Building, and Judgement—we significantly improved both system robustness and accuracy. This modular design consistently handled complex natural language queries and structured ECB datasets more effectively than the monolithic approaches tested during earlier phases of the project.

Furthermore, our findings strongly support the incorporation of Text-to-SQL mechanisms for extracting numerical data as an essential improvement. By leveraging SQL-based extraction rather than relying exclusively on semantic search for both qualitative and quantitative queries, we significantly reduced the risk of numerical inaccuracies and mitigated potential hallucinations associated with purely language-model-driven data extraction. The evaluation also highlighted our system's effectiveness in handling "off-topic" queries. By designing structured response types, including a clearly defined "Default Message" for queries outside the system's intended scope, we successfully avoided generating misleading or incorrect responses, thereby enhancing the reliability of the system.

Another significant learning was the importance of using realistic evaluation datasets. This approach provided deeper insights into practical system performance and highlighted limitations that might have remained unnoticed otherwise.

Initial experiments on the potential of using an LLM for re-ranking initial semantic search results looked promising. Results indicated that applying LLM-based re-ranking to the top 10 semantic search results significantly improved relevance. However, expanding re-ranking to larger sets (e.g., top 40 results) provided diminishing returns, offering limited improvements relative to computational costs. Thus, we will focus in future work on LLM-based re-ranking strategies on a limited subset of initial top-ranked results, achieving an effective balance between enhanced accuracy and computational efficiency.

Furthermore, semantic search accuracy remains key area for ongoing research and development. A detailed analysis of accuracy of the retrieval component revealed that the current semantic search mechanism struggles to consistently position the correct time series at the top ranks for certain queries. Improving recall and ranking accuracy within the semantic search algorithm will thus be a critical priority for future enhancements.

Finally, our investigations into metadata utilization revealed that simply increasing metadata volume does not necessarily enhance retrieval accuracy. As demonstrated in evaluations using an expanded metadata index, the addition of more metadata fields alone did not significantly improve search performance. This implies that future efforts should focus on more sophisticated metadata strategies, potentially involving LLM-driven metadata re-ranking, dynamic augmentation, or innovative embedding methods, rather than merely expanding metadata quantity.

These insights establish a robust foundation to guide future improvements of our RAG-based system, enhancing both its effectiveness and user-centred design for data discovery on the ECB Data Portal.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York: ACM press.

Douze, M., Guzhva, A., Deng, C., Johnson, J., & Szilvasy, G. (2024). The Faiss library. *arXiv*(2401.08281).

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. *18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150-158). Association for Computational Linguistics.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., . . . Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459--9474.

Liang, L. (2024). *Stacking Small Language Models for Generalizability.* arXiv preprint.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out* (pp. 74-81). Barcelona: Association for Computational Linguistics.

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). *Query Rewriting for Retrieval-Augmented Large Language Models.* arXiv preprint.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. In C. D. Manning, *Introduction to Information Retrieval* (p. Chapter 6). Cambridge: Cambridge University Press.

OpenAI. (2024, January 25). *New embedding models and API updates.* Retrieved from OpenAI: https://openai.com/index/new-embedding-models-and-api-updates/

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). *Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy.* arXiv preprint.

Streamlit. (2025). *Streamlit Documentation*. Retrieved from Streamlit: https://docs.streamlit.io/

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., . . . Zhou, E. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey.* arXiv preprint.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Lin, Z., . . . Xing, E. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*, 46595-46623.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Supporting users in seeking data on the ECB data portal

A use case for RAG

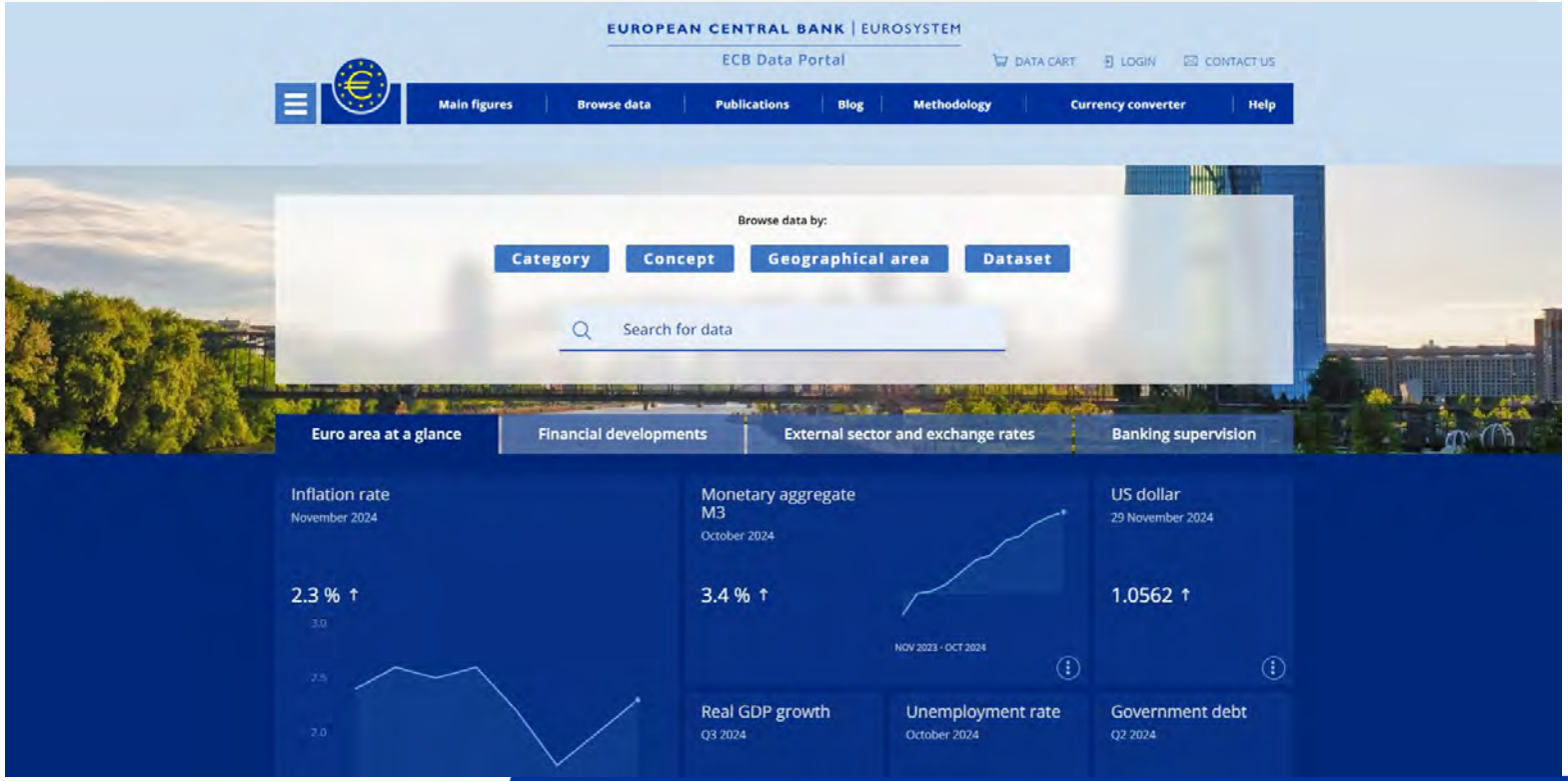


The views expressed are those of the authors and do not necessarily reflect those of the ECB.

February 2025

Luca Petracca, Simone De Benedictis, Thomas Gottron, Zlatina Hofmeister

The ECB Data Portal



Example of Time Series data on the EDP



The problem

How can we help users finding information in the ECB Data Portal?

- Users report difficulties finding and identifying the right information

The ECB Data Portal uses “traditional” ways of accessing information

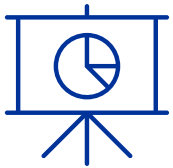
- Browsing by categories, concepts, geographical areas and datasets
- Search by keywords
- Main indicators
- Interactive Publications
- Dashboards

What about providing an interface where users can use their own “natural language”?

www.ecb.europa.eu ©

Our Idea

- **PoC:** testing **Retrieval Augmented Generation (RAG)** and **Text-to-SQL** techniques to explore how state-of-the-art **LLMs** can be leveraged to enable users to **find data** using **natural language** in a **chat** interface



Understand the user queries



Identify the right dataset and time series key



Retrieve and provide time series data

www.ecb.europa.eu ©

How does it work?

PoC – Setup

Scope:

- **Time Series Included:**
  - ~300.000 TS are included (~10% of the EDP)
  - including all TS either in Main Figures* or included in official publications, reports and articles

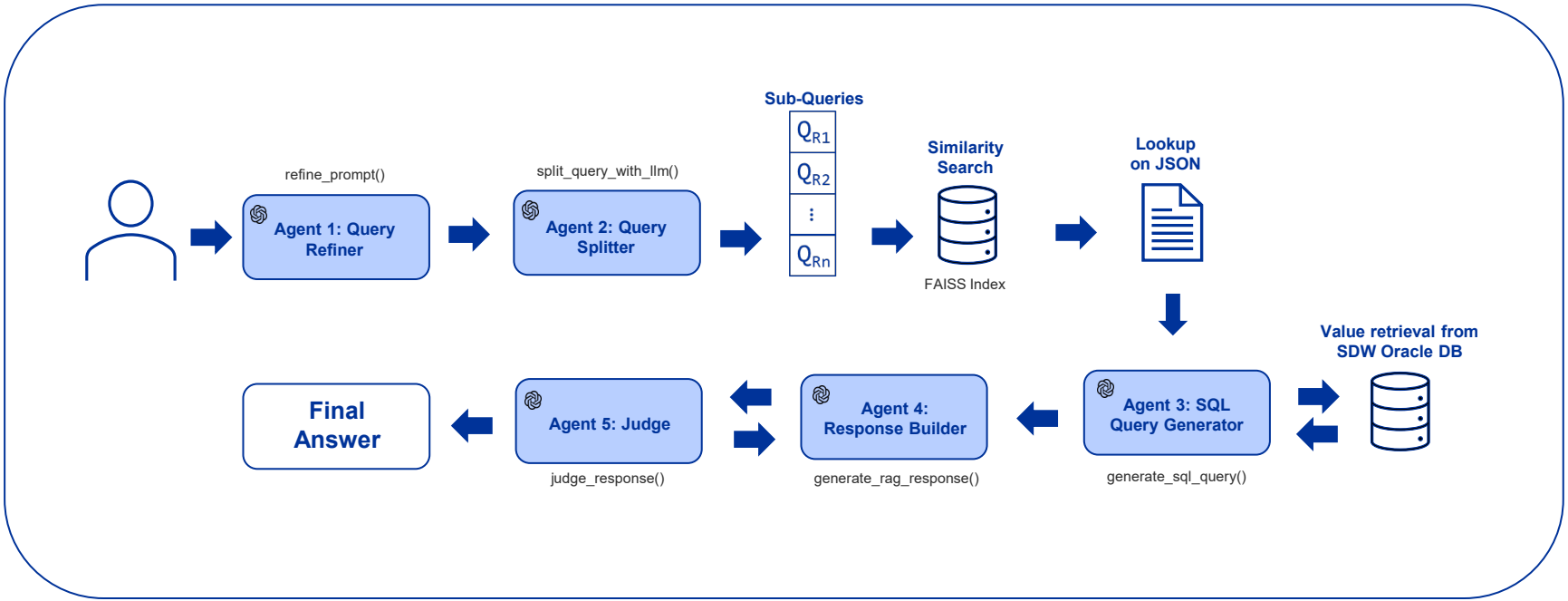
Technologies & APIs:

- Connection to LLM via **OpenAI API**
- Technologies:
  - **FAISS** (For RAG semantic search)
  - **GPT-4o** (Chat completion model)
  - **Text-embedding-3-large** (Embedder for RAG semantic search)

*Main Figures: https://data.ecb.europa.eu/main-figures

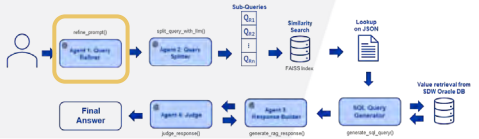
www.ecb.europa.eu ©

Current Architecture



Behind the Scenes

Step 1: Query Refinement



« What was the Inflation in Germany and Italy last year»

User Query

System Prompt: «You are an assistant that has to refine queries in order to perform vector search more effectively to retrieve data from the European Central Bank data portal. Your task is to identify and expand vague and ambiguous terms in the query…»

Agent 1: Query Refinement

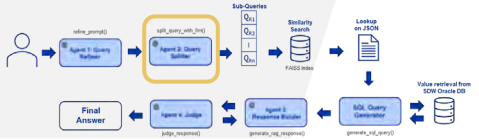
refine_prompt()

«Inflation rate in Germany and Italy in 2024»

Refined Query

www.ecb.europa.eu ©

Step 2: Query Splitting



«Inflation rate in Germany and Italy in 2024»

System Prompt: “You are an assistant specialized in breaking down complex queries into multiple simple queries. If a user query contains info about multiple countries, different compositions, datasets or time frames, you should split the queries and output a JSON object…”

Sub-queries: [‘Inflation rate in Germany’, ‘Inflation rate in Italy’]

Refined Query

Agent 2: Query Splitting

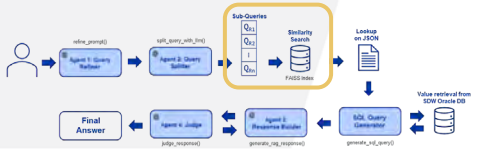
split_query_with_llm()

Q_{R1}
Q_{R2}
⋮
Q_{Rn}

N Sub-Queries

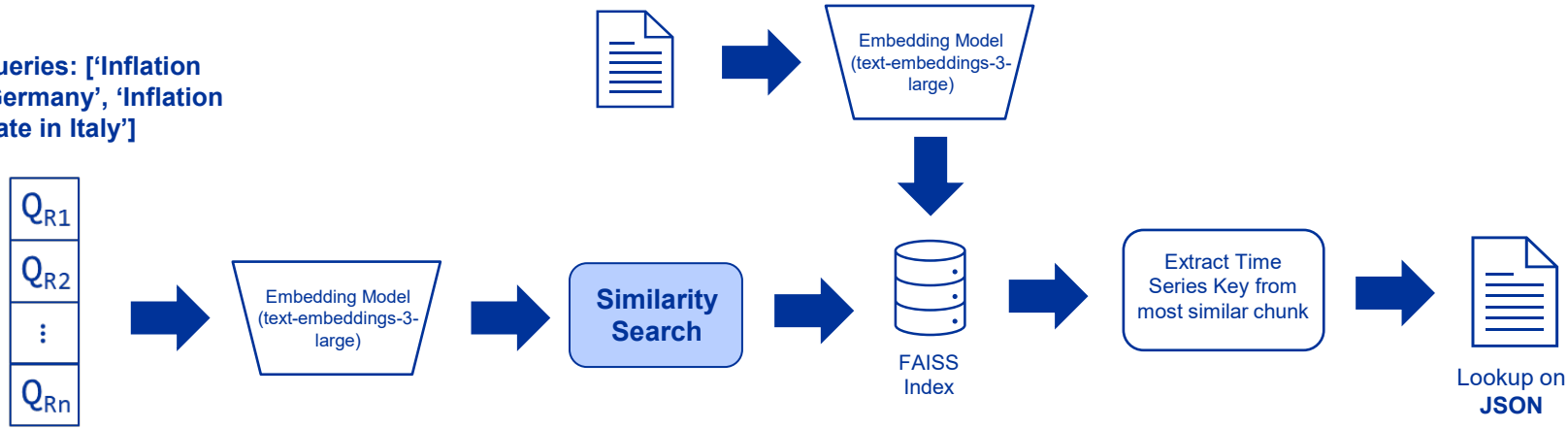
www.ecb.europa.eu ©

Step 3: Similarity Search

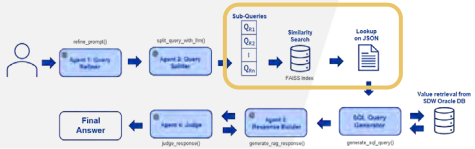


RAG_EDP.txt: Text database containing dataset, time series key and time series complete title, description and metadata for each time series in the selected subset

Sub-queries: ['Inflation rate in Germany', 'Inflation rate in Italy']



Step 4: Information Retrieval



Knowledge Base





FAISS Index

We are looking for most similar chunk in the Index based on the content of the user query. Each line contains the dataset, key and description of a time series

Metadata JSON

Given the extracted time series key, we look up on the JSON to extract the full metadata. **But we still need the values…**

Series Key: ICP.M.IT.N.000000.4.ANR, Dataset: ICP, Value for Italy - HICP - Overall index, Annual rate of change, Eurostat, Neither seasonally nor working day adjusted with Monthly frequency, in Italy

▼ ICP.M.IT.N.000000.4.ANR {6}

DATASET : ICP

DATASET DESCRIPTION : Scope: Data presentation - Summary description. The Harmonised Index of Consumer Prices (HICP) for the euro area is published by the European Commission (Eurostat) and generally available from 1996 onwards. Euro area results are obtained by aggregating indices for individual countries. The HICP is broken down following the European classification of individual consumption according to purpose (ECOICOP) and by goods and services special aggregates derived from it. The HICP covers monetary expenditure on final consumption by resident and non-resident households on the economic territory of the euro area. The seasonally adjusted HICP data are compiled by the ECB. Data presentation - Detailed description. HICPs measure changes in prices of goods and services covered by final household monetary consumption expenditure, including all indirect taxes paid by consumers. HICPs have a common coverage of goods and services across countries but country specific item lists and item weights. HICPs do not cover expenditure for owner occupied housing. For items fully or partly paid or refunded by the government, HICPs include only the share that is paid by the consumer (e.g. the \"out-of-the-pocket\" expenditure for health services). The HICPs are classified according to the ECOICOP. Additional compiled aggregates are also published. Methodological information. Time period: Monthly, Quarterly, Annual. Base period: 2015=100. Statistical concepts and definitions: For information about the naming convention (series key dimensions and metadata), refer to the ICP underlying DSD (ECB_ICP1) maintained by the ECB. Statistical processing. Data compilation: Following the Maastricht Treaty the aim of the HICP is to measure inflation by means of the consumer price index on a comparable basis, taking into account differences in national definitions. Eurostat dedicated HICP website (which include short guide for users). Adjustment: X-12 ARIMA. Administrative Information. Title: ICP - Indices of Consumer prices. Data source: European Commission (Eurostat) and European Central Bank calculations based on Eurostat data. Quality. Timeliness: 17 working days after the end of the reference month. Legal and institutional environment. Legal acts and other agreements: Council Regulation (EC) No 2494/95 on HICPs and implementation regulations.

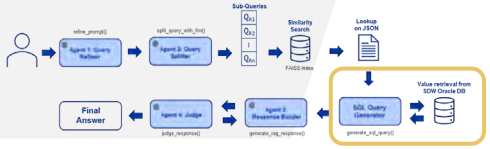
DATASET LINK : https://data.ecb.europa.eu/data/datasets/ICP

SERIES KEY LINK : https://data.ecb.europa.eu/data/datasets/ICP/ICP.M.IT.N.000000.4.ANR

FREQUENCY : Monthly

SERIES KEY DESCRIPTION : Value for Italy - HICP - Overall index, Annual rate of change, Eurostat, Neither seasonally nor working day adjusted with Monthly frequency, in Italy

Step 5: SQL Query Generation



Inflation rate in Germany and Italy in 2024?,

‘ICP.M.DE.N.000000.4.ANR’, ‘ICP.M.IT.N.000000.4.ANR’

System Prompt: “You are a SQL query assistant. Your task is to create SQL queries to fetch data from the SDW SQL database. The data is stored in a table-like structure, and your goal is to generate queries that match the user's request…”

Extracted Time Series Key(s) + Refined Prompt

Agent 3: SQL Query Generation

generate_sql_query()

```
SELECT *
FROM SDW_WEB.V_ICP_OBS

WHERE SERIES_KEY IN
(‘ICP.M.DE.N.000000.4.ANR’,
‘ICP.M.IT.N.000000.4.ANR’)

AND OBS_DATE BETWEEN
‘2024-01-01’ AND ‘2024-12-01’
```

Value retrieval from SDW Oracle DB

Table with values

www.ecb.europa.eu ©

Step 6: Response Generation



System Prompt: “You are an assistant specialized in assisting users in finding the appropriate data and series key from the ECB Data Portal. Understand the user query, mention the datasets the information is coming from and retrieve and communicate the data. Always include the series key and the link to the EDP...”

System Prompt: “You are an expert judge tasked with evaluating the relevance and completeness of information extracted from the European Central Bank (ECB) data portal. Your task is to provide a 'total rating' scoring how well the system answer answers the user request expressed in the user question.“

Best matching **time series data and values**

Agent 4: Response Builder

generate_response_rag()

Agent 5: Judge LLM

judge_response()

Final Answer

The first response is built using the **refined prompt,** the time series **metadata** extracted from the JSON, and the **values** extracted from the SQL query. If the answer is not satisfactory, the response builder provides a more **generic answer** to the user

www.ecb.europa.eu ©

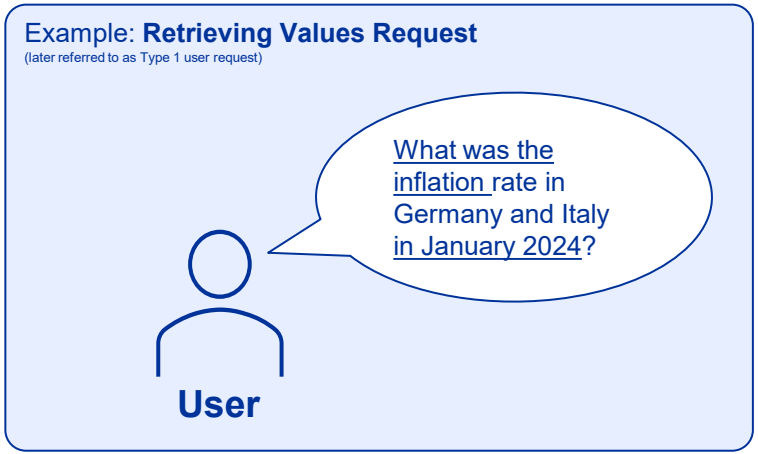
Evaluation

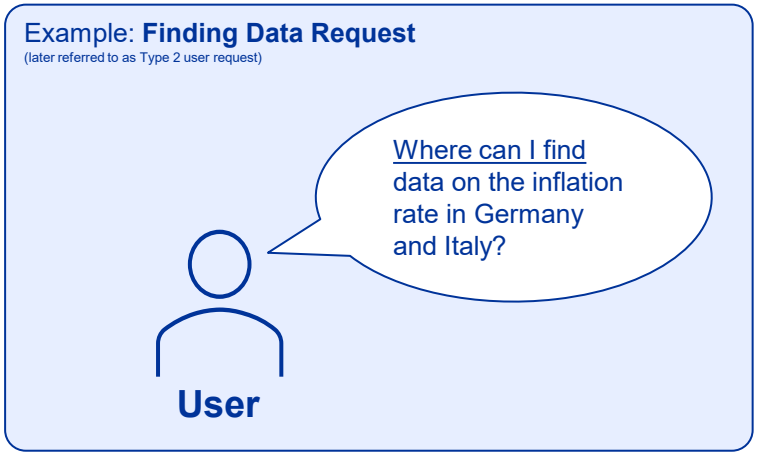
Our Approach

We leverage a mixture of state-of-the-art AI methods based on LLMs (**RAG and TEXT2SQL).**

Our chat assistant supports users with 2 types of requests in natural language:

- ✓ **Type 1 - Retrieving Values:** We retrieve the specific values a user in a time series a user is looking for.
- ✓ **Type 2 - Finding Data:** We identify the correct datasets and time series users are looking for.





Results – Value Accuracy

Test Set: 525 Q&A labelled pairs

Results (overall):

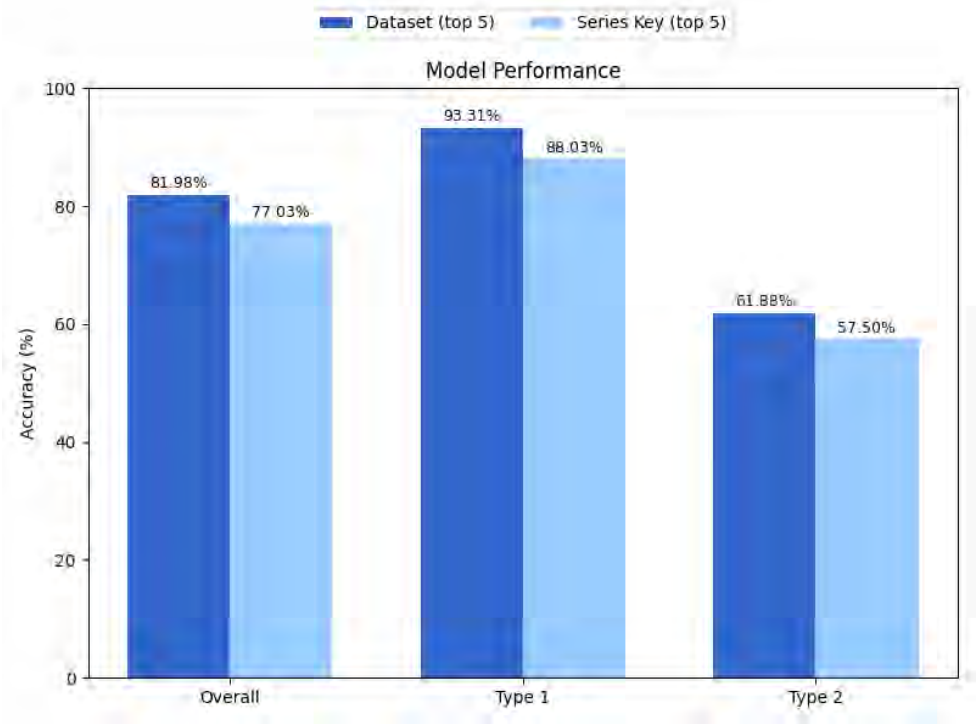
- Dataset Accuracy - **81.98%**
- Series Key Accuracy - **77.03%**

Type 1:

- Dataset Accuracy – **93.31%**
- Series Key Accuracy - **88.03%**

Type 2:

- Dataset Accuracy – **61.88%**
- Series Key Accuracy – **57.50%**



Key Insights and Way Forward

Key Insights and Way Forward

In a nutshell…

- The PoC shows promising results with regards to our goal of **helping users find data on the EDP**
- The only way to increase accuracy and reliability while decreasing costs is… **to avoid LLMs**
  - **Vector search** and fixed **SQL generation** instead of RAG for search and value extraction was crucial for reducing hallucination
  - Judge LLM as the "core" AI-based agent
- **Top-K multiple series key** retrieval drastically improved model accuracy

Next Steps:

- **Improve Accuracy**
  - Exploring new retrieval techniques for the similarity search **(clustering embeddings, hierarchical search…)**
  - Test a **conversational approach** building the SQL query field-by-field with the user
  - Increase the number of questions-answers to better measure accuracy
- **Finding the right balance between cost-benefit-risk**
- **Deploy the first version internally for a selected group of users with the objective of better understanding how users interact with the chatbot**

www.ecb.europa.eu ©

Thank you!

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Using retrieval-augmented generation (RAG) and LangChain to drive chatbots for interactive economic data analysis and insight[1]

Seah Boon Keong, Eilyn Chong and Max Chew,
Central Bank of Malaysia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Using retrieval-augmented generation (RAG) and LangChain to drive chatbots for interactive economic data analysis and insight

Seah Boon Keong, Eilyn Chong, Max Chew

Abstract

This paper introduces a Proof of Concept (PoC) chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing complex economic datasets. Our approach aims to democratise access to economic data, enabling a wide range of users - from policymakers and researchers to financial analysts and the general public - to engage more effectively with authoritative economic information sources. The system ingests and processes diverse economic datasets and metadata sourced from a comprehensive knowledge base of authoritative institutions (e.g., the Bank for International Settlements (BIS), the International Monetary Fund (IMF) Data Services, the Federal Reserve Bank of St. Louis (FRED), Malaysia's Department of Statistics (OpenDOSM), etc.).

Keywords: Large Language Model (LLM), Economic Data Analysis, Chatbot System, Retrieval-Augmented Generation (RAG), LangChain, Dynamic Data Visualisation

1. Introduction

With the increasing availability of economic data, researchers and the general public face the challenge of navigating vast datasets from multiple official sources, including international organisations and national statistical agencies. While each platform offers valuable insights, users must often familiarise themselves with distinct interfaces, data structures, and retrieval methods. These variations create significant barriers for those seeking swift insights, leading to cognitive overload. This challenge is further compounded by the complexity of economic datasets - often fragmented, multidimensional, and filled with domain-specific terminology – posing significant challenges for non-expert users seeking actionable information (Araujo, 2024; Wu et al., 2023).

To address these challenges, we explore the potential of advanced chatbot technology as a solution. Users often struggle with fragmented and complex economic datasets, requiring specialised knowledge to interpret. Our proposed solution is an intuitive chatbot system that simplifies data access and analysis, making it easier for users to gain initial insights through natural language queries. By leveraging cutting-edge technologies such as Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), our chatbot provides a user-friendly interface for exploring and analysing economic datasets. This approach democratises access to vital economic data, extending its benefits to a broader user base, including policymakers, researchers, financial analysts, and the general public.

Key areas of focus include democratising accessibility to authoritative economic data sources, offering initial statistical insights into the datasets, and enabling dynamic data visualisation. The system ingests and processes a wide array of economic datasets and metadata sourced from numerous institutions, including the Bank for International Settlements (BIS), the International Monetary Fund (IMF), the Federal Reserve Bank of St. Louis (FRED), and Malaysia's Department of Statistics (OpenDOSM). It supports multiple data formats, such as CSV, Parquet, and JSON, ensuring compatibility with diverse datasets. Furthermore, the system retrieves datasets specified in the metadata, either from API calls or direct URL link to files, enabling seamless integration of external data sources.

Large Language Models (LLMs) have emerged as a cornerstone of modern natural language processing (NLP) systems due to their remarkable capabilities in generating human-like responses (Ouyang et al., 2022; OpenAI, 2023). Despite their strengths, LLMs are not without inherent limitations. While they possess some world knowledge acquired during training (Petroni et al., 2019; Roberts et al., 2020; Jiang et al., 2020), they are prone to hallucinations and generating imaginary content (Maynez et al., 2020; Zhou et al., 2021). To mitigate these issues, Retrieval-Augmented Generation (RAG) has been introduced as an augmentation technique. RAG enables LLMs to retrieve relevant information from external knowledge resources, thereby mitigating hallucination concerns and improving the accuracy of generated responses (Khandelwal et al., 2020; Izacard et al., 2022). By integrating RAG, our chatbot system ensures that responses are grounded in authoritative datasets, enhancing the reliability of insights provided to users.

The RAG framework operates by retrieving documents based on the user's input and generating answers based on the retrieved content (Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Sachan et al., 2021; Lee et al., 2021; Jiang et al., 2022; Nakano et al., 2021; Qian et al., 2023; Lazaridou et al., 2022; Shi et al., 2023). In our chatbot system, this capability is further enhanced by sourcing publicly available datasets from authoritative institutions such as BIS, FRED, IMF, and OpenDOSM. For instance, when analysing trends in GDP growth rates or government debt levels, the chatbot retrieves relevant data points from these sources and generates the dynamic visualisations, enabling users to interact with the data dynamically.

By leveraging these technologies, our chatbot aims to make economic data more accessible and easier to analyse for a wide range of users. This approach not only enhances the usability of economic data but also facilitates quicker and more accurate insights, supporting data-driven decision-making processes. As demonstrated in recent studies, the integration of RAG and LLMs in specialised domains has shown significant promise in improving the quality of NLP applications (Wu et al., 2023; Chen et al., 2021). Our work builds on these advancements, offering a practical solution tailored to the unique challenges of economic data analysis.

2. Methodology and Approach

Model Selection and Fine-Tuning

The foundation of our chatbot system is built on a combination of Retrieval-Augmented Generation (RAG) and LangChain, two advanced technologies that significantly enhance the user experience when interacting with complex economic

datasets. At its core, the system employs a Large Language Model (LLM) trained using LangChain LLM, which is specifically designed to understand query context and generate accurate, insightful responses. The decision to fine-tune this model was driven by the unique challenges posed by economic data, which often includes domain-specific terminology, multidimensional tabular structures, and intricate relationships between variables.

Fine-tuning the LLM involved training it on a curated set of question-and-answer pairs using economic datasets sourced from official data providers and national statistical agencies to enhance its abilities to generate insightful responses to prompts-based tasks related to economic analysis. This process ensured that the model could interpret natural language queries and provide contextually relevant insights into economic trends, such as inflation rates, GDP growth, government debt levels, and labour market dynamics. For example, when analysing annual GDP growth for Malaysia, Singapore, Japan, and Australia from 2005 to 2023 (IMF, 2023), the fine-tuned model demonstrated superior accuracy compared to general-purpose LLMs. By incorporating domain-specific knowledge, the chatbot system minimises the risk of hallucinations—a common issue in standard LLMs—and ensures that responses are grounded in authoritative data sources (Maynez et al., 2020; Zhou et al., 2021).

Data Source Integration

To ensure the chatbot provides comprehensive and reliable insights, the system integrates publicly available economic datasets from authoritative institutions such as BIS, FRED, IMF, and OpenDOSM. These datasets encompass a wide range of macroeconomic indicators and financial statistics, enabling users to explore diverse aspects of the global economy. The system supports multiple data formats, including CSV, Parquet, and JSON, ensuring compatibility with various types of datasets.

Metadata plays a crucial role in this integration process, as it contains information and descriptions about the dataset to be retrieved. When a user submits a query, the system performs a similarity search between the query and the dataset descriptions. If a match is found, the system retrieves the relevant datasets specified in the metadata, either through API calls or direct links to CSV or Parquet files.

For instance, when prompted to analyse Federal Reserve Economic Data (FRED) borrowing data from 2005 to 2023, the system uses the metadata to perform similarity search between the user query and the dataset descriptions within the metadata repository. It identifies that the FRED borrowing dataset description matches the user query. The system then extracts the dataset using the direct CSV link provided in the metadata file. The system then generates a summary report with initial statistical insights derived from the dataset, accompanied by visualisations such as line charts.

Embedding Model

A critical component of our methodology is the embedding model, which converts both user queries and economic data into vector representations. We employ GPT-4o mini with text-embedding-ada-002 for high-dimensional data, ensuring that the data is accurately represented in a format that the LLM can process effectively. This step is essential for enabling the chatbot to understand the semantic meaning of queries and match them with relevant data points.

For example, when a user asks about trends in Malaysia's labour force size from 1982 to 2021, the embedding model converts the query into a vector representation and matches it with the corresponding description of the OpenDOSM dataset. This process ensures that the chatbot retrieves the most relevant information, such as the consistent upward trend in labour force size from 5,431.4 thousand people in 1982 to 15,797.2 thousand people in 2021 (OpenDOSM, 2023). By leveraging high-dimensional embeddings, the system achieves a high degree of precision in data retrieval, enhancing the overall quality of responses.

Retrieval Mechanism

The retrieval mechanism is a cornerstone of our chatbot system, ensuring that the data retrieved matches the query context and intent. This step involves using embeddings to find relevant information in the vector database, which is powered by Facebook AI Similarity Search (FAISS—a highly efficient vector search library. The retrieval process begins by converting the user query into a vector representation and then searching for the nearest neighbours in the vector space. This ensures that the chatbot retrieves the most relevant data points, even when dealing with large and complex datasets.

Dynamic Charting Techniques

One of the standout features of our chatbot system is its ability to provide dynamic charting based on the questions and answers posed by users. Our chatbot dynamically generates interactive charts directly from the retrieved datasets. This feature allows users to visualise trends and patterns in real-time without requiring any technical expertise.

For example, when a user asks about trends in Malaysia's GDP growth from 2015 to 2023, the chatbot retrieves the relevant dataset from OpenDOSM and generates an interactive line chart showing the sharp decline in Q2 2020 due to the COVID-19 pandemic and the subsequent recovery by Q4 2021 (OpenDOSM, 2023). Users can interact with the chart by zooming in on specific time periods, and hovering over data points for additional details. This dynamic charting capability enhances data interpretation and user engagement, making it easier for users to gain actionable insights.

Comparison of FAISS, Chroma, and PineCone

The choice of vector database is a critical factor in determining the performance and scalability of the chatbot system. To identify the best option, we conducted a comparative analysis of three popular vector databases: FAISS, Chroma, and PineCone based on an analysis conducted by the Hewlett Packard Enterprise Community (Vaduguru, 2024).

FAISS emerged as the top choice due to its exceptional **performance and cost efficiency**. Developed by Facebook AI Research, FAISS is optimised for high-dimensional vector search, making it ideal for handling large-scale economic datasets. Its in-memory storage capabilities enable rapid retrieval of relevant data points, while its simple API facilitates seamless integration with the chatbot system. In contrast, Chroma and PineCone, while offering robust features, were less cost-effective and more complex to implement for our specific use case.

By selecting FAISS as our vector database, we ensure that the chatbot system delivers fast, accurate, and scalable responses, even when processing extensive datasets. This provides users with a seamless and efficient experience when exploring economic data.

Choice of Vector Database

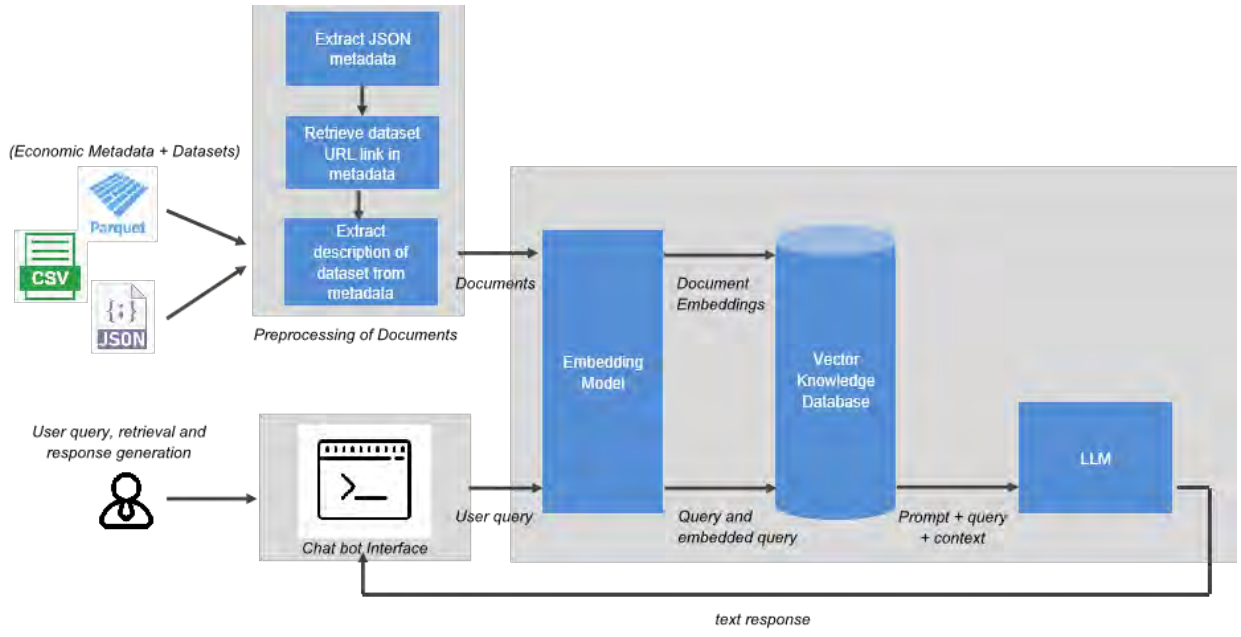
Comparison of FAISS, Chroma, and PineCone — Table 1

	FAISS	Chroma	PineCone
Performance	High	Moderate	High
Scalability	High	High	High
In-Memory Storage	No	Yes	Yes
API Simplicity	Moderate	High	High
Cost Efficiency	High	Moderate	Moderate

Source: (Vaduguru, 2024; Woyera, 2023; MyScale 2024)

Process Flow

The chatbot system follows a structured process flow designed as shown below to handle user queries and deliver insights.



A. Preprocessing

1. Finding Metadata and Data Links

To obtain metadata and links, we manually search authoritative economic data sources to identify relevant datasets. For each dataset, we extract key metadata such as descriptions, data frequency, and the corresponding link to CSV/Parquet/JSON file. The information is then organised into a structured JSON file, which the system uses to streamline access and retrieval.

2. **Reading Metadata and Data Links**
   Utilise Langchain to load the JSON file containing metadata and links to CSV/Parquet/JSON file. Langchain provides a robust framework for handling various data formats and integrating them seamlessly into the workflow.

3. **Data Validation and Loading**
   Validate the links to Parquet, JSON or CSV files. If the Parquet link is valid, read the data from it; if the Parquet link is not available, the data should be read from the CSV link; otherwise, the data should be obtained from the JSON link. This step ensures that the data is correctly loaded from the appropriate source.

4. **Text Splitting**
   Use Langchain's "RecursiveCharacterTextSplitter" to divide the text into manageable chunks of 1000 characters with an overlap of 200 characters. The chunk size of 1000 characters is chosen to balance manageability and context preservation, ensuring that each chunk contains enough information for meaningful processing. The 200-character overlap helps maintain contextual continuity between chunks, reducing boundary effects and enhancing the quality of the embeddings.

5. **Generating Embeddings**
   Generate embeddings for the text chunks using OpenAI's embedding model (text-embedding-ada-002) integrated within Langchain. Embeddings are numerical representations of text that capture semantic meaning, essential for similarity searches and other NLP tasks. This step leverages the power of LLMs to create rich, high-dimensional representations of the text data.

6. **Creating a Vector Store**
   Create a vector store from the text chunks and their embeddings using FAISS. FAISS is a library for efficient similarity search and clustering of dense vectors, making it ideal for handling large-scale data. The vector store serves as a repository of embeddings that can be quickly queried to find similar documents or text chunks.

7. **Saving the Vector Store Locally**
   Save the vector store locally in the specified folder with a filename based on the metadata. This step ensures that the vector store is easily accessible for future queries and operations. By storing the vector store locally, we can efficiently manage and retrieve the embeddings as needed.

B. Frontend Chatbot Integration

1. **Similarity Search**
   Implement the Langchain's similarity search to retrieve relevant documents from the vector store based on user queries. This step leverages the precomputed embeddings to find the most relevant information quickly. The similarity search is a key component of Retrieval-Augmented Generation (RAG), where relevant documents are retrieved to enhance the chatbot's responses.

2. **Prompt Chatting** Use the Large Language Model (LLM) with gpt-4o-mini for generating responses to user queries. By leveraging the retrieved documents, the chatbot can provide more accurate and contextually relevant answers through RAG (Retrieval-Augmented Generation). The integration of LLMs like gpt-4o-mini allows the chatbot to generate high-quality, coherent responses that are informed by the retrieved context.

3. **Dynamic Chart Generation**
   Integrate functionality to dynamically generate charts based on the prompt responses. This involves utilising the datasets loaded into the vector store to create visual representations of the data, enhancing the user's understanding and interaction. Dynamic chart generation adds a visual dimension to the chatbot's capabilities, making the information more accessible and engaging.

This methodology ensures a comprehensive approach to preprocessing, efficient data handling, and robust frontend chatbot integration, leveraging Langchain, FAISS, RAG, and LLM technologies.

3. Results

The results section highlights the performance and capabilities of our chatbot system in addressing user queries related to economic data analysis. As a Proof of Concept (PoC), the chatbot leverages a curated subset of datasets sourced from authoritative institutions. These include IMF datasets on government debt levels, and current account balances across multiple countries. From FRED, the PoC integrates data on Federal Reserve borrowings and fluctuations in the monetary base. Additionally, the POC also includes BIS datasets on debt securities and financial stability metrics, and OpenDOSM datasets on Malaysia's labour market dynamics and GDP trends, highlighting domestic economic performance.

Below, we present a selection of queries posed by users and the corresponding responses generated by the chatbot. Each response is accompanied by screenshots of the chatbot interface, illustrating its functionality and user experience. These examples demonstrate the system's ability to handle complex queries, retrieve relevant data, and generate dynamic visualisations that enhance the user's understanding of economic trends.

3.1 Query: Trends in Annual GDP Growth (IMF Dataset)

User Query:

"*Explain the observed trends in annual GDP growth for Malaysia, Singapore, Japan, and Australia from 2005 to 2023.*"

Chatbot Response:

Based on the annual changes in GDP growth for Malaysia, Singapore, Japan, and Australia from 2005 to 2023, distinct patterns and trends emerge.

Japan (JPN): Japan experienced moderate fluctuations in GDP growth, starting at 1.8% in 2005 and peaking at 4.1% in 2010. Notable periods include a significant contraction in 2009, where the GDP growth fell to -5.7%, reflecting the global financial crisis's impact. By 2023, Japan's GDP growth stabilized at 2.3%, influenced by ongoing structural challenges, including an aging population and low productivity growth.

Singapore (SGP): Singapore demonstrated significant volatility, beginning at 7.4% in 2005 and reaching a peak of 14.5% in 2010, driven by a robust recovery from the global financial crisis. However, the economy faced a sharp decline in 2020, with GDP growth dropping to -5.4% due to the COVID-19 pandemic. By 2023, Singapore's GDP

growth was recorded at 3.9%, reflecting a recovery phase as the economy adjusted to post-pandemic conditions.

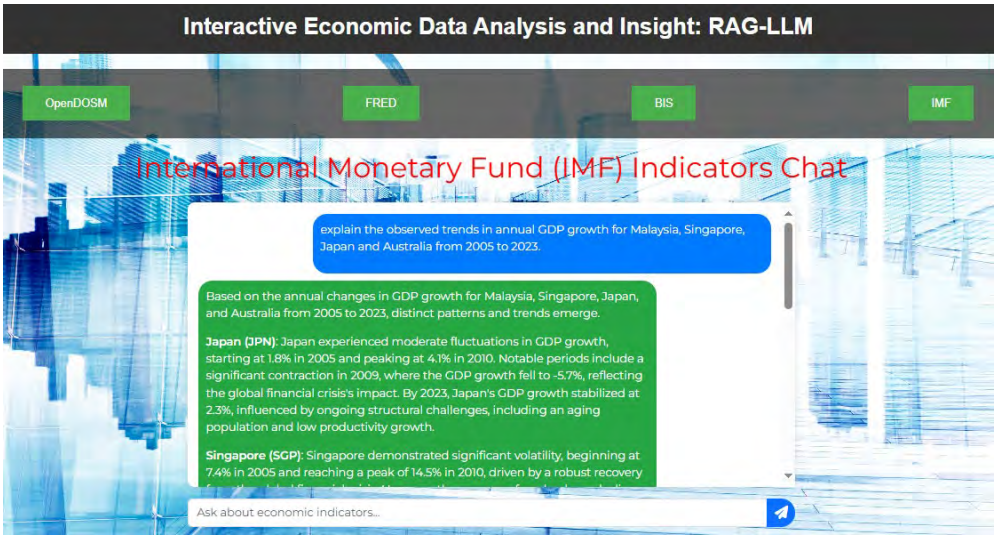
Malaysia (MYS): Malaysia showed a pattern of moderate growth, starting at 5.0% in 2005 and peaking at 7.5% in 2010. The country faced challenges in 2020, with GDP growth declining to -5.6% amid the pandemic. By 2023, Malaysia's GDP growth rebounded to 5.5%, supported by a recovery in domestic demand and external trade.

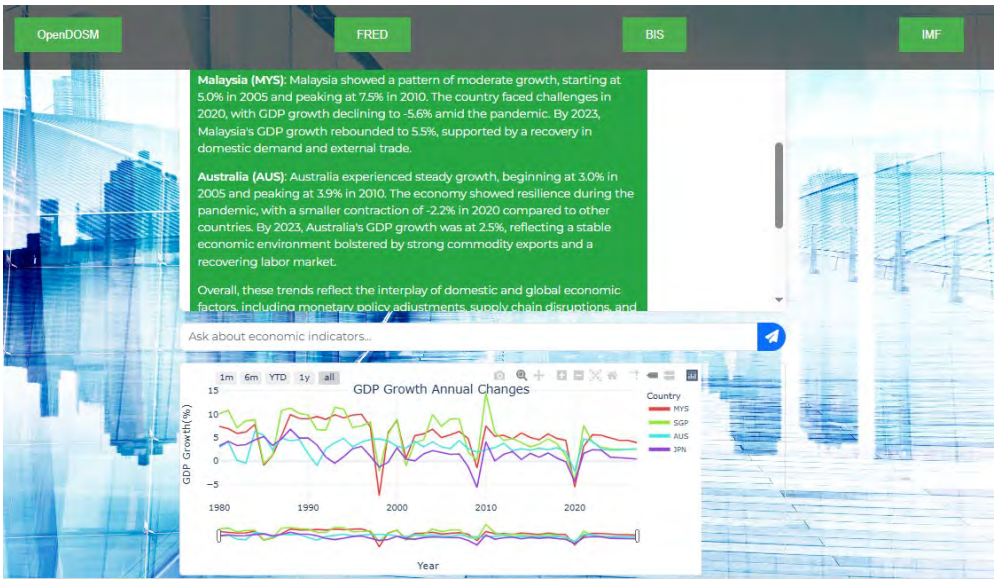
Australia (AUS): Australia experienced steady growth, beginning at 3.0% in 2005 and peaking at 3.9% in 2010. The economy showed resilience during the pandemic, with a smaller contraction of -2.2% in 2020 compared to other countries. By 2023, Australia's GDP growth was at 2.5%, reflecting a stable economic environment bolstered by strong commodity exports and a recovering labour market.

Overall, these trends reflect the interplay of domestic and global economic factors, including monetary policy adjustments, supply chain disruptions, and geopolitical events. The varying impacts of the COVID-19 pandemic also highlight the resilience and adaptability of these economies in the face of unprecedented challenges.

For further details, you can refer to the dataset from the International Monetary Fund: IMF World Economic Outlook.

Screenshots:





3.2 Query: Trends in Current Account Balance (IMF Dataset)

User Query:

"Explain the observed trends in Current Account Balance (as percentage of GDP) for Italy, Malaysia, UK, and US from 2005 to 2023."

Chatbot Response:

Based on the annual changes in Current Account Balance (as a percentage of GDP) for Italy, Malaysia, the United Kingdom, and the United States from 2005 to 2023, distinct patterns and trends emerge.

Malaysia (MYS): Malaysia experienced moderate fluctuations in its current account balance, starting at 13.8% in 2005 and peaking at 16.3% in 2008. Notable periods include a decline to 1.3% in 2016, reflecting challenges in global trade and domestic economic adjustments. By 2023, Malaysia's current account balance settled at 3.9%, influenced by a recovery in trade and investment flows post-pandemic.

United States (USA): The United States demonstrated a trend of increasing deficits, beginning at -5.7% in 2005 and reaching a low of -4.7% in 2008, before deepening to -2.9% in 2020. The trend reflects ongoing trade imbalances and increased consumption of imports over exports. By 2023, the current account balance was at -3.2%, indicating persistent challenges in achieving a balanced trade position.

Italy (ITA): Italy showed a relatively stable but negative trend, starting at -0.9% in 2005 and fluctuating around this level until reaching -1.4% in 2023. The current account balance reflects Italy's ongoing economic challenges, including slow growth and structural issues in its economy, which have hindered significant improvements in its trade balance.

United Kingdom (GBR): The United Kingdom exhibited a trend of increasing deficits, starting at -1.9% in 2005 and worsening to -5.5% by 2023. This trend highlights the UK's challenges in maintaining a competitive trade position, exacerbated by factors such as Brexit and shifts in global trade dynamics.

Overall, these trends reflect the interplay of domestic and global economic factors, including trade policies, investment flows, and external economic conditions. The data indicates that while Malaysia has managed to maintain a relatively positive current account balance, the US and UK have struggled with persistent deficits, and Italy has faced challenges in stabilizing its trade position.

For further details, you can access the dataset from the IMF here.

Screenshots:



3.3 Query: Trends in FRED Borrowing Data (FRED Dataset)

User Query:

"Analyse the trends in the annual changes of FRED borrowing data from 2005 to 2023. What patterns or significant fluctuations can be observed?"

Chatbot Response:

Based on the annual changes in Federal Reserve borrowings data from 2005 to 2023, distinct patterns and trends emerge.

2005-2007: The period leading to the financial crisis saw relatively stable and low borrowing levels, with values fluctuating between approximately $100 billion and $400 billion. For instance, in January 2005, borrowings were around $110 billion, gradually increasing to about $424.7 billion by July 2005, reflecting a moderate upward trend as the economy was still expanding.

2008: A dramatic shift occurred during the financial crisis, with borrowings skyrocketing from approximately $400 billion in early 2008 to a peak of $698.8 billion in November 2008. This surge was driven by the Federal Reserve's emergency lending measures in response to the financial turmoil, including the collapse of major financial institutions.

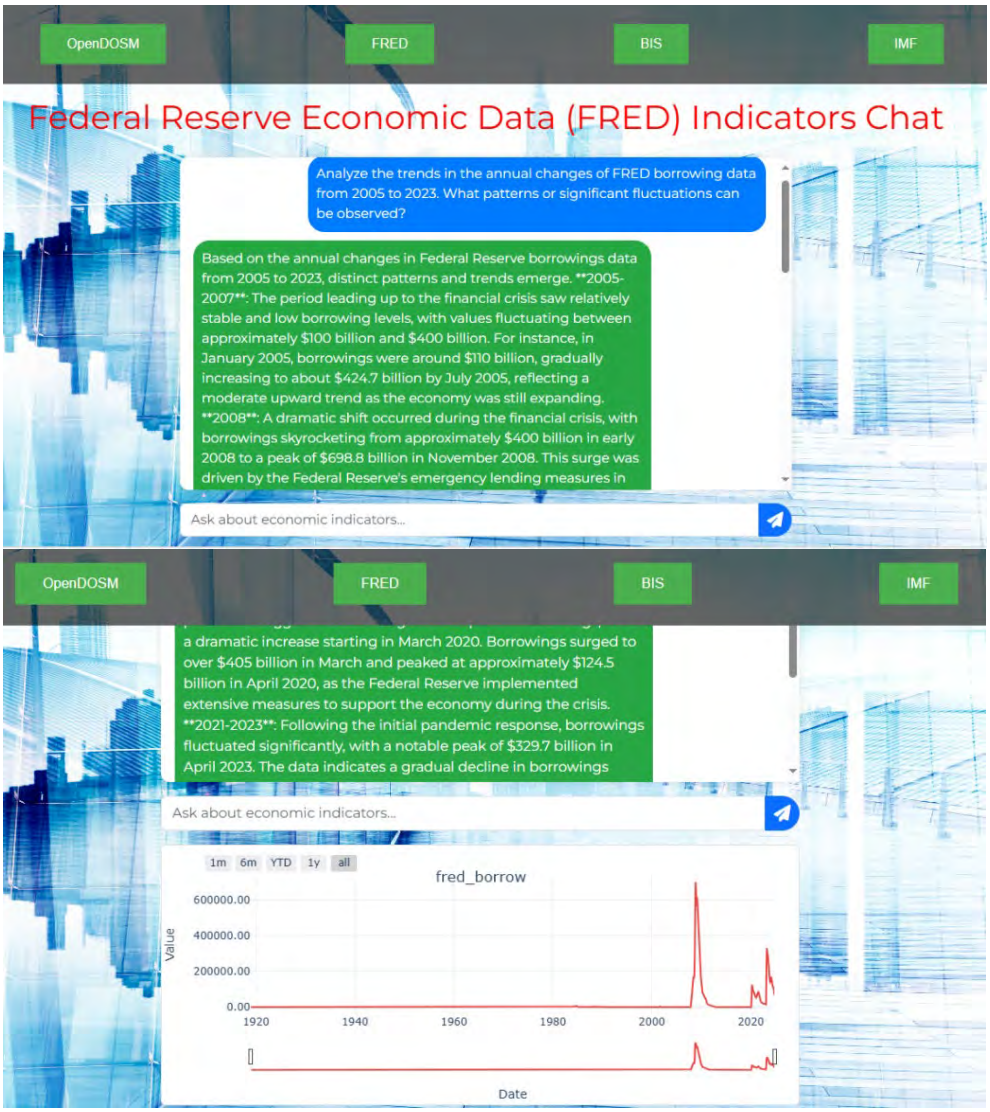
2009-2015: Following the peak in 2008, borrowings remained elevated but began to decline gradually. By the end of 2009, borrowings had decreased to about $142 billion, reflecting the Fed's efforts to stabilize the economy. Throughout this period, borrowings fluctuated but generally trended downward, with occasional spikes, such as in early 2010, when they reached around $111 billion.

2016-2019: Borrowings continued to stabilize, with values mostly below $100 billion, indicating a return to more normal lending conditions as the economy recovered. The data showed a consistent decline, with borrowings dropping to around $76.8 billion in January 2016 and remaining relatively low through 2019.

2020: The COVID-19 pandemic triggered another significant spike in borrowings, with a dramatic increase starting in March 2020. Borrowings surged to over $405 billion in March and peaked at approximately $124.5 billion in April 2020, as the Federal Reserve implemented extensive measures to support the economy during the crisis.

2021-2023: Following the initial pandemic response, borrowings fluctuated significantly, with a notable peak of $329.7 billion in April 2023. The data indicates a gradual decline in borrowing throughout 2022, but they remained elevated compared to pre-pandemic levels, reflecting ongoing economic uncertainties and the Fed's continued support measures. Overall, these trends reflect the interplay of domestic and global economic factors, including monetary policy adjustments, economic crises, and recovery efforts. The significant fluctuations in borrowing levels during the financial crisis and the COVID-19 pandemic highlight the Federal Reserve's role in stabilising the economy during periods of distress. For further details, you can access the dataset from the FRED website: https://fred.stlouisfed.org/series/BORROW.

Screenshots:



3.4 Query: Labour Market Trends (OpenDOSM Dataset)

User Query:

"Analyse the annual trends in key labour market indicators. What significant patterns or shifts can be observed?"

Chatbot Response:

The annual Labour Force Report data from 1982 to 2021 reveals several significant trends and patterns in Malaysia's labour market:

1. **Labour Force Size Growth:** The labour force size (lf_size) has shown a consistent upward trend, increasing from 5,431.4 thousand people in 1982 to a peak of 15,797.2 thousand people in 2021. This indicates a growing workforce over the decades.
2. **Employment Trends:** The number of employed people has also increased significantly, from 5,249.0 thousand in 1982 to 15,064.2 thousand in 2021. This growth in employment aligns with the overall increase in the labour force size. 3.

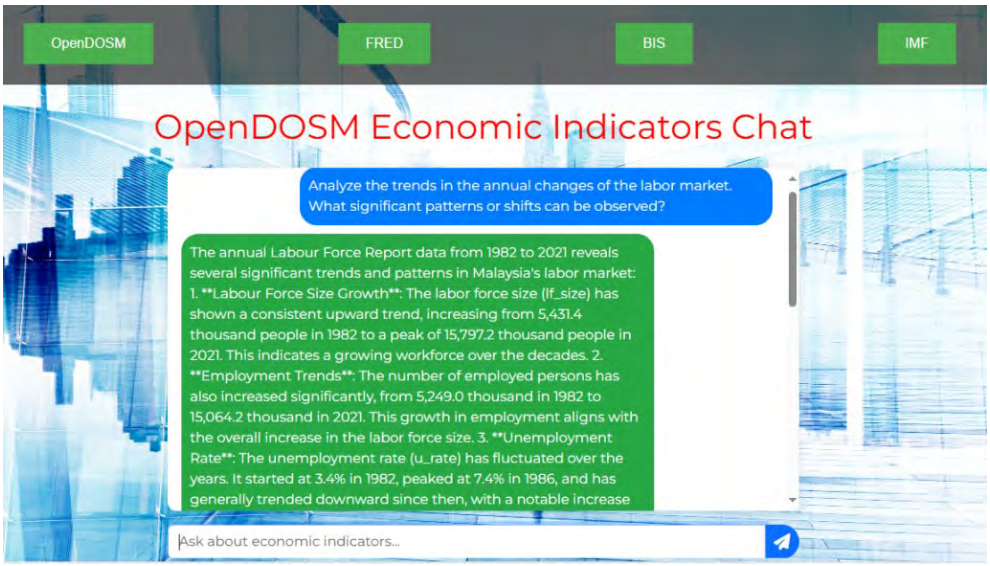
3. **Unemployment Rate:** The unemployment rate (u_rate) has fluctuated over the years. It started at 3.4% in 1982, peaked at 7.4% in 1986, and has generally trended downward since then, with a notable increase during the economic downturns in the late 1990s and early 2000s. By 2021, the unemployment rate was recorded at 4.6%.

4. **Participation Rate:** The participation rate (p_rate) has shown a gradual increase from 64.8% in 1982 to 68.6% in 2021, indicating a higher proportion of the working-age population is engaged in the labour market.

5. **Employment-Population Ratio:** The employment-population ratio (ep_ratio) has remained relatively stable, hovering around 62% to 66% over the years, suggesting that while more people are entering the labour force, the proportion of those employed compared to the total population has not seen drastic changes.

6. **Economic Events Impact:** The data reflects the impact of economic events, such as the Asian Financial Crisis in the late 1990s and the COVID-19 pandemic in 2020, which caused temporary spikes in unemployment and fluctuations in labour force participation. Overall, the trends indicate a robust growth in the labour market, with increasing employment opportunities, although challenges such as unemployment spikes during economic downturns remain evident. For a detailed view of the data, you can download the dataset from the following link: https://storage.googleapis.com/dosm-public-economy/labour-principalstats-annual.csv.

   Screenshots:





4. Discussion

Key Advantages of the Chatbot PoC

- **Democratisation of Economic Data**
  One of the primary advantages of the chatbot system is its ability to democratise access to authoritative economic datasets from institutions such as the IMF, FRED, BIS, and Malaysia's Department of Statistics (OpenDOSM). These datasets encompass a wide range of macroeconomic indicators, financial statistics, and labour market trends, enabling users—ranging from policymakers and researchers to students and the general public—to explore complex economic trends without requiring technical expertise. For instance, users can analyse trends in GDP growth or current account balance simply by posing natural language queries.
- **Contextual and Accurate Responses with RAG**
  The integration of RAG ensures that the chatbot's responses are grounded in authoritative data sources. By retrieving relevant documents from the vector store and leveraging LLMs to generate contextually accurate answers, the system minimises the risk of hallucinations—a common issue in standard generative models. This hybrid approach enhances the reliability of the insights provided for economic analysis.
- **Dynamic Charting for Enhanced Visualization**
  Dynamic charting is a standout feature of the chatbot system, enabling users to visualise trends and patterns in real-time. For example, when analysing annual GDP growth for Malaysia, Singapore, Japan, and Australia, the chatbot generates interactive line charts that allow users to zoom in on specific periods, hover over data points for additional details, and customise visualisations. This functionality enhances data interpretation and user engagement, making complex economic data more accessible and actionable.

Potential Improvements

While the current methodology ensures robust and efficient data handling, there are several areas for potential improvement

- **Expanding Data Sources**
  We plan to expand the system's capability to include more datasets from BIS, IMF, FRED and OpenDOSM. This will enhance the breadth and depth of the insights provided by the chatbot.
- **Enhancing Dynamic Charting**
  While the current dynamic charting functionality is effective, we plan to expand it by including more chart types, such as bar chart, stacked bar chart, and scatter plots. This will provide users with a richer and more versatile visualisation experience.

5. Conclusion

The chatbot system, as a Proof of Concept (PoC), demonstrates significant potential in democratising access to complex economic datasets and providing actionable insights through technologies such as Retrieval-Augmented Generation (RAG), dynamic charting, and Large Language Models (LLMs). By integrating authoritative datasets from institutions like the International Monetary Fund (IMF), Federal Reserve Economic Data (FRED), Bank for International Settlements (BIS), and Malaysia's Department of Statistics (OpenDOSM), the system empowers users—ranging from policymakers and researchers to students and the general public—to explore macroeconomic trends without requiring technical expertise. The integration of RAG ensures that responses are grounded in authoritative data, minimising the risk of hallucinations and enhancing the reliability of insights. Additionally, the system's dynamic charting capabilities provide interactive visualisations that make complex economic data more accessible and engaging.

Despite these achievements, the system also faces challenges, including limited chart types, and the need for expanded dataset coverage. Addressing these challenges will help the chatbot's usability in broader contexts.

References

Araujo, J. "Applications of LLMs in Central Banks: A Review". *Journal of Financial Technology* (2024).

BIS. "Debt Securities Data". *Retrieved from https://stats.bis.org/api/v2/data/dataflow/BIS/WS_NA_SEC_DSS/1.0/Q.N.ES+FR+IT+U2.5Z.S12T.S1.N.A.F.F3.T._Z.EUR.X1.M.V.N._T?format=csv* (2023).

Chen, D., Fisch, A., Weston, J., & Bordes, A. "Reading Wikipedia to Answer Open-Domain Questions". *Proceedings of ACL* (2017).

Chen, H., Zhang, Y., & Wang, X. "Specialized LLMs for Financial Analysis". *Proceedings of the International Conference on Machine Learning* (2021).

FRED. "Federal Reserve Economic Data". *Retrieved from https://fred.stlouisfed.org/series/BORROW* (2023).

Guu, K., Hashimoto, T. B., Oren, Y., & Liang, P. "REALM: Retrieval-Augmented Language Model Pre-Training". *Advances in Neural Information Processing Systems* (2020).

IMF. "World Economic Outlook Database". *Retrieved from https://www.imf.org/external/datamapper/* (2023).

Izacard, G., & Grave, E. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". *Proceedings of EACL* (2021).

Izacard, G., et al. "Improving Retrieval-Augmented Models with Contrastive Learning". *Advances in Neural Information Processing Systems* (2022).

Jiang, Z., Xu, F., & Araki, J. "How Can We Know What Language Models Know?" *Transactions of the Association for Computational Linguistics* (2020).

Khandelwal, U., Fan, A., Jurafsky, D., & Kiela, D. "Nearest Neighbor Machine Translation". *Proceedings of ICLR* (2020).

Lazaridou, A., et al. "Retrieval-Augmented Multimodal Models for Vision-Language Tasks". *Advances in Neural Information Processing Systems* (2022).

Lewis, P., et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". *Advances in Neural Information Processing Systems* (2020).

Maynez, J., Narayan, S., Bohnet, B., & Tiedemann, J. "On Faithfulness and Factuality in Abstractive Summarization". *Proceedings of ACL* (2020).

MyScale. "Pinecone vs FAISS Comparison". *Retrieved from https://myscale.com/blog/pinecone-vs-faiss-vector-search-efficiency-analysis/ (2024).*

Nakano, R., et al. "WebGPT: Improving Web Search for Open-Domain Question Answering". *Advances in Neural Information Processing Systems* (2021).

Ouyang, L., et al. "Training Language Models to Follow Instructions with Human Feedback". *Advances in Neural Information Processing Systems* (2022).

Petroni, F., et al. "Language Models as Knowledge Bases?" *Proceedings of EMNLP* (2019).

Qian, H., et al. "Dynamic Retrieval-Augmented Models for Time-Series Forecasting". *Proceedings of ICML* (2023).

OpenDOSM. "Malaysia Labour Force Report". *Retrieved from https://open.dosm.gov.my/data-catalogue?search=labour* (2023).

Roberts, A., et al. "How Language Model Training Scales". *arXiv preprint arXiv:2001.08361* (2020).

Sachan, M., et al. "End-to-End Training of Neural Retrievers for Open-Domain Question Answering". *Proceedings of ACL* (2021).

Shi, W., et al. "Multimodal Retrieval-Augmented Models for Cross-Modal Understanding". *Proceedings of CVPR* (2023).

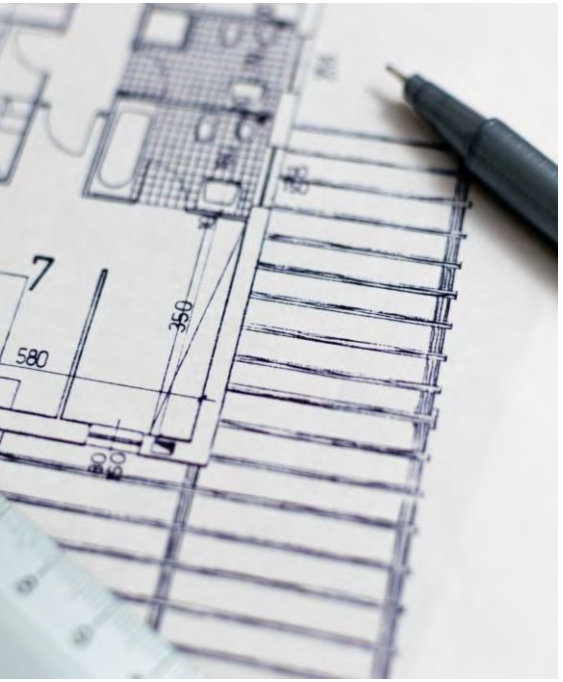
Sohangir, S., Wang, D., & Pomeranets, A. "Big Data and Machine Learning in Finance". *Journal of Financial Data Science* (2018).

Vaduguru, P. "Comparing Pinecone, Chroma DB and FAISS". *Retrieved from Hewlett Packard Enterprise Community. https://community.hpe.com/t5/insight-remote-support/comparing-pinecone-chroma-db-and-faiss-exploring-vector/td-p/7210879 (2024).*

Woyera. "Pinecone vs Chroma Comparison". *Retrieved from https://medium.com/@woyera/pinecone-vs-chroma-the-pros-and-cons-2b0b7628f48f (2023)*

Wu, Y., Li, Z., & Chen, X. "Fine-Tuning LLMs for Domain-Specific Tasks". *Advances in Neural Information Processing Systems* (2023).

Zhou, C., et al. "Detecting Hallucinations in Neural Machine Translation". *Proceedings of NAACL* (2021).



Using RAG and LangChain to Drive Chatbots for Interactive Economic Data Analysis and Insight*



Seah Boon Keong, Eilyn Chong, Max Chew
Central Bank of Malaysia
Monetary Policy Department

(*) The views and conclusions presented in this paper are exclusively those of the author(s) and do not necessarily reflect the position of the Central Bank of Malaysia or its Board members.

Agenda

- Motivation
- Methodology and Approach
- Usage
- Future Work
- Conclusion

Advancing Economic Data Access via Innovation in Chatbot Technology

- **Problem Statement:** Users often struggle with fragmented and complex economic datasets, requiring specialised knowledge to interpret.
- **Solution:** Our chatbot provides an intuitive interface that simplifies data access and analysis, making it easier for users to gain insights.



Democratize accessibility to vital economic data, extending the benefits to a broader user base including policymakers, researchers, financial analysts.



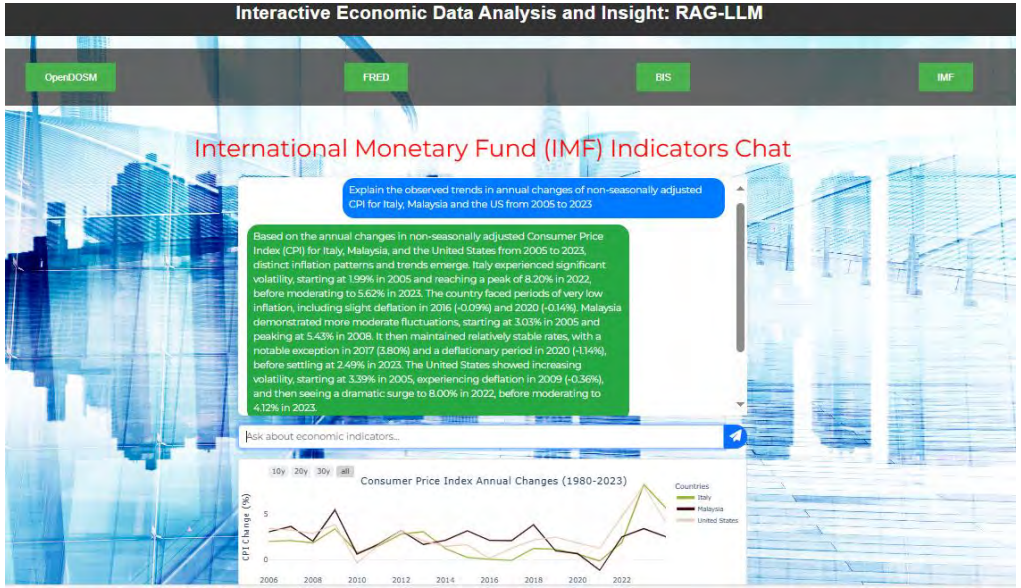
Ingest a wide array of metadata sourced from numerous authoritative institutions.



Offer an interactive and user-friendly interface to explore and analyze intricate economic datasets, thereby making complex data more comprehensible



Utilizes advanced technologies such Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) and Dynamic Charting capabilities



Data Sources: Publications and Datasets

Data Sources

- We source publicly available economic datasets from authoritative institutions such as BIS, FRED, and IMF, as well as OpenDOSM for Malaysia's data

Types of Data

- The system ingests from various economic datasets, including macroeconomic indicators and financial statistics.
- It supports: CSV, Parquet, JSON
- The system will retrieve the datasets based on the URLs specified in the metadata, either from the returned API call or from links of CSV or Parquet files.

Methodology and Approach

1. Using Retrieval Augmented Generation (RAG) Architecture

Utilizing **relevant information** effectively enhances the output from large language models. Details in the following slide.

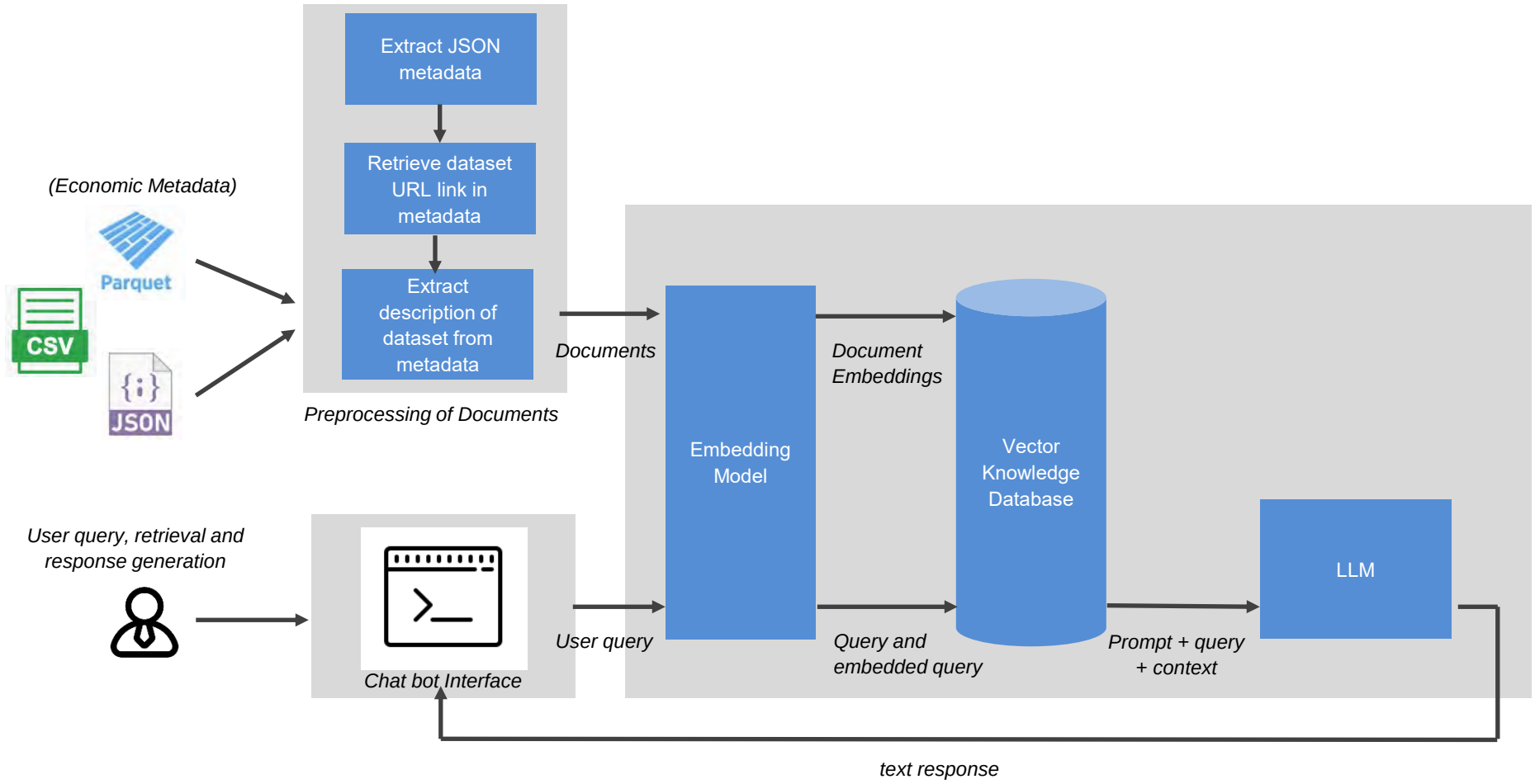
2. Data Source Integration

As a proof of concept, this prototype application ingests **datasets** from FRED, BIS, IMF and OpenDOSM for enriched analysis.

3. Dynamic Charting Techniques

Employing **interactive visuals** allows for better data interpretation and user engagement.

Methodology and Approach



Methodology and Approach

Blend of several AI components in which the key elements are as follow:

Large Language Model (LLM)
- Core of RAG
- Understands query context and generates responses.
- Trained based on the datasets provided
- Uses LangChain LLM.

External Knowledge Database (Vector DB)
- Vector database – a collection of Datasets retrieved from OpenDOSM, BIS, IMF, FRED.
- Uses FAISS Vector DB

Embedding Model
- Converts queries and data into vector representations.
- Uses GPT-4o mini with text-embedding-ada-002 for high-dimensional data.

Retrieval Mechanism
- Uses embeddings to find relevant information.
- Ensures data matches query context and intent.

Integration and Response Generation
- Integrates retrieved data into responses.
- Provides contextually relevant and corresponding chart

Methodology and Approach

Comparison of FAISS, Chroma, and PineCone based on performance, scalability, ease of use, in-memory storage, API simplicity, and cost efficiency.

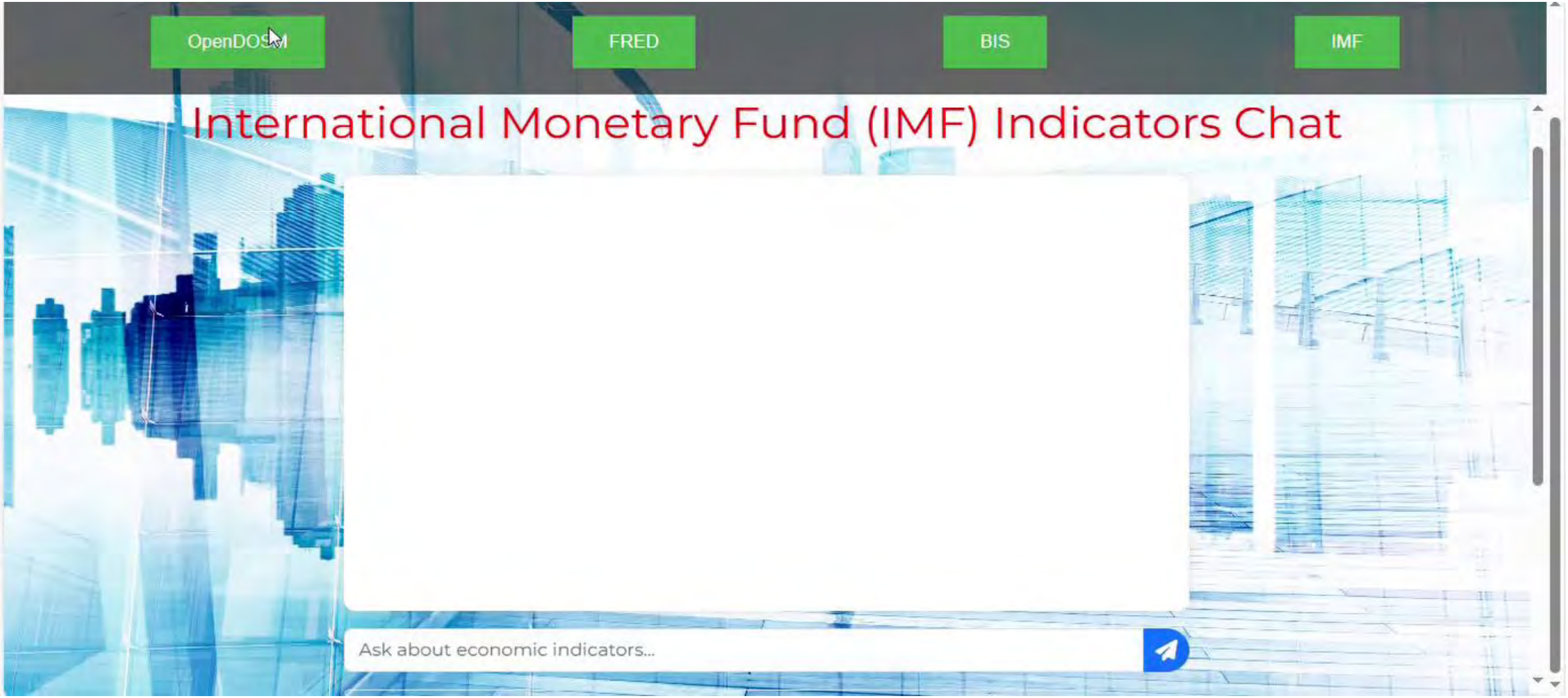
Reason for FAISS selection: **Performance and Cost Efficiency**

Feature	*FAISS*	*Chroma*	*PineCone*
Performance	***High***	***Moderate***	***High***
Scalability	***High***	***High***	***High***
Ease of Use	***Moderate***	***High***	***High***
In-Memory Storage	***No***	***Yes***	***Yes***
API Simplicity	***Moderate***	***High***	***High***
Cost Efficiency	***High***	***Moderate***	***Moderate***

Source:
Pinecone vs FAISS, Pinecone vs Chroma

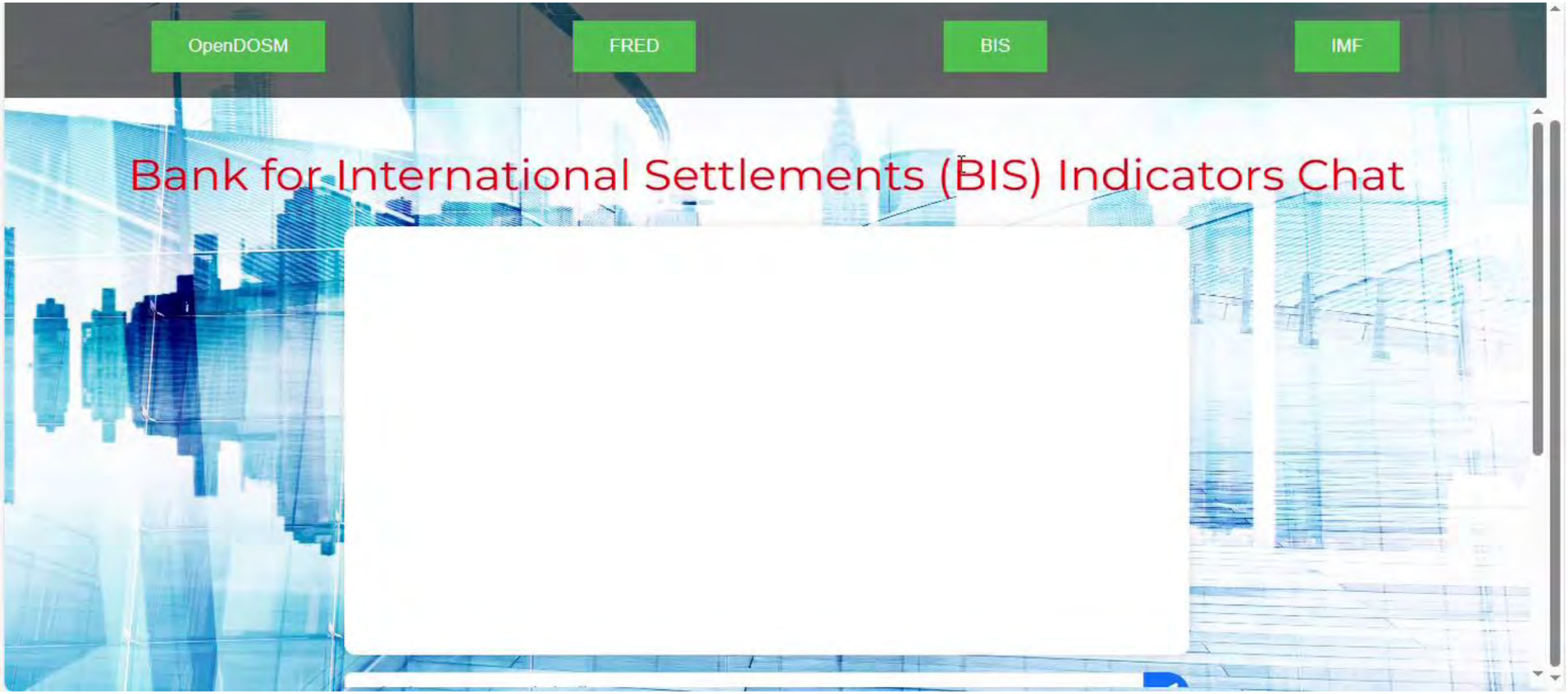
Usage

An innovative chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing economic datasets. - **Use case sample 1**



Usage

An innovative chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing economic datasets. - **Use case sample 2**



Future Work

Potential Improvements (Prototype - Proof of Concept)

- This prototype (PoC) serves as a foundation for future enhancements. Key areas for improvement include:

  - **Expanding Data Sources**
    - We aim to integrate additional datasets from sources such as BIS, IMF, and FRED. This expansion will significantly enhance the chatbot's ability to deliver comprehensive and insightful analyses.

  - **Enhancing Dynamic Charting**
    - To provide users with a richer and more versatile visualization experience. We plan to expand it by including more chart types, such as bar charts, heatmaps, and scatter plots.

Conclusion

- In summary, our chatbot system democratizes access to economic data, making analysis more accessible and efficient.

- Facilitate faster exploration of economic data to gather data-driven insights and improving efficiency of economic analysis.

Thank you for your attention!

Seah Boon Keong (Ph.D) seah.keong@bnm.gov.my

Eilyn Chong eilyn@bnm.gov.my

Max Chew kai.xi@bnm.gov.my

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Leveraging generative AI for granular credit data utilization: a multi-agent approach[1]

Nontawit Cheewaruangroj, Kawinwish Laobundit, Peranut Nimitsurachat, Supachai Saengthong, Skunpoj Thanarojsophon and Anak Yodpinyanee, Bank of Thailand

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Leveraging generative AI for granular credit data utilization: a multi-agent approach

Nontawit Cheewaruangroj, Kawinwish Laobundit, Peranut Nimitsurachat, Supachai Saengthong, Skunpoj Thanarojsophon, Anak Yodpinyanee[1]

Abstract

The Bank of Thailand's Regulatory Data Transformation (RDT) program introduces a new standard for granular credit data reporting by financial institutions. The data's complexity poses analytical challenges for central bank examiners. A generative AI model has been leveraged to aid examiners in utilizing this dataset. Given examiners' questions in plain text, these tools streamline the data analysis process by breaking down examiners' questions into several actionable steps (RDT Brainstormer), searching for relevant fields (RDT Metadata Copilot), and providing SQL code to query from the RDT database (RDT SQL Coder) using a customized Retrieval-Augmented Generation (RAG) procedure. Due to the complexity and large context of each task, a multi-agent framework has been adopted to orchestrate multiple specialized agents to collaborate on the same task. Finally, we deploy our application to production and measure the validity, relevancy and usefulness of the generated output via both automated evaluation and users' feedback. We find a substantial improvement over a single-agent approach.

Keywords: LLM, GenAI, multi-agents, text-to-SQL, Chatbot, granular data

[1] The opinions expressed in this discussion paper are those of the authors and should not be attributed to the Bank of Thailand.

Contents

1. Introduction and Motivation 3
2. Background Knowledge 4
   - Using Generative AI for Data Querying 4
   - Text Embedding 4
   - Retrieval-augmented Generation (RAG) 5
   - Multi-agent Framework 5
   - Evaluating Generative AI Response 6
3. Methodology 7
   - Customized RAG 7
   - LLMs and Structured Data Validation 7
   - Application Deployment 8
   - Multi-agent Workflow 8
4. Experiment 10
   - Tasks and Evaluation Methods 10
   - Accommodating Single-Agent Algorithm 12
5. Result & Discussion 13
6. Conclusion and Further Work 15
   - Conclusion 15
   - Further Work 15

References 16

Appendix 1: Questions for Evaluation 18
   - Quantitative Questions for User Evaluation (translated) 18
   - Qualitative Questions for User Evaluation (translated) 18

Appendix 2: Responses for Evaluation Questions 19
   - Responses for Quantitative Question 10 (translated) 19
   - Responses for Qualitative Question 10 (translated) 23

1. Introduction and Motivation

The Bank of Thailand (BOT) utilizes data extensively for both supervision and monetary policy. To enhance regulatory oversight and well-informed policymaking, the Regulatory Data Transformation (RDT) project was initiated in 2020 to incorporate granular credit data from financial institutions. The initiative aims to replace the current Data Management System (DMS), which primarily relies on fixed reports, thereby restricting the BOT's capability to comprehensively analyze data from multiple perspectives. The granularity of the credit data in the RDT project enables more detailed analysis in various dimensions, including credit lines, interest rate plans, loan movements, and collaterals. However, the RDT project contains a large volume of data as well as intricate data fields and metadata, with more than 1000 fields and over 25 million loan accounts data per month. This complexity hinders its widespread use within the bank, particularly for analysts and examiners who lack familiarity with big data techniques such as SQL. In addition, translating business problems into data querying in RDT poses a challenge. To overcome these challenges, we explore generative AI approaches that can help facilitate usage of the data.

Generative AI (GenAI) has recently gained popularity, especially as assistants for users (Brynjolfsson, Li, & Raymond, 2023). Many central banks have explored using generative AI to assist with their tasks (Aldasoro, et al., 2024). For example, central banks like the European Central Bank (ECB) and Monetary Authority of Singapore (MAS) are utilizing generative AI technologies to enhance their supervisory and analytical capabilities: the ECB's tool Athena uses GenAI to analyze documents and identify trends (ECB, 2023), while the MAS uses GenAI to assist with data collection processes (MAS, 2023). Many GenAI-based chatbots have been developed to help with data utilization, such as SQL coding assistant (Gao, et al., 2024). The ECB has leveraged generative artificial intelligence to translate queries written in plain language into code to find specific data points. This has led to several applications and platforms serving many ECB and EU members' supervisory authorities to streamline their day-to-day operations (ECB, 2024). These initiatives demonstrate how central banks are leveraging AI to improve their operations and decision-making processes.

While these solutions could assist with the mentioned tasks, there are still some obstacles to adopting them in RDT. Firstly, credit data is aggregated into various data cubes, each with different levels of granularity. An SQL coding assistant may fail to select the data cube with the correct granularity because it does not have a comprehensive view of the entire database. Secondly, datasets in RDT contain more than a thousand data fields, making it difficult to identify and verify the correct ones to use. Lastly, posing the right questions to ask the AI may not be straightforward, given the complexity of the data. One way to overcome these challenges is by allowing users to interact with multiple AIs, each with a different specialization such as problem identification, metadata discovery and SQL coding. This framework is known as the multi-agent framework.

The multi-agent framework has been developed to enhance the capabilities of generative AI by allowing multiple AI agents to converse with each other to complete their tasks. It can be employed for various tasks such as retrieval-augmented code generation and question answering (Wu, et al., 2023), improving reasoning (Du, Li, Torralba, Tenenbaum, & Mordatch, 2023) and evaluating the performance of GenAI (Liu, et al., 2023). Using the multi-agent framework would allow the AI agents to assist users in using RDT by breaking down user input questions into different tasks and

assigning different agents for each task. For example, one agent could help with enhancing user's business question, whereas another could retrieve and filter relevant fields. Then, another agent can use the information from the previous agents to craft the SQL query that retrieves the correct data.

In this study, we explore the capabilities of the multi-agent framework to assist data users via three AI modules: *Brainstormer*, *Metadata Copilot* and *SQL Coder*. The *Brainstormer* module refines users' input questions, allowing them to gain better insights. For example, inputs might lack necessary information such as a timeframe and need to be scoped down to specific categories. *Metadata Copilot* assist users in finding the data fields and data cubes relevant to the input questions. Lastly, *SQL Coder* helps users write SQL queries related to the questions. Each module consists of multiple agents that will interact with one another, the user, and a customized retrieval-augmented generation (RAG) that retrieves information from the RDT datasets. The final outputs of these modules will be evaluated by an automated test, as well as our analysts and examiners, in terms of validity, relevancy and usefulness, compared against the outputs from single-agent GenAI as benchmarks.

The remainder of the paper is organized as follows. Section 2 provides background knowledge on generative AI, and the multi-agent framework. Section 3 covers our methodology and architecture. Section 4 details the experiments for evaluating the frameworks. Section 5 discusses the experimental results. Section 6 concludes our study and suggests future work.

2. Background Knowledge

Using Generative AI for Data Querying

Large Language Models (LLMs) can be used to assist with data utilization. One promising capability is to convert user input texts (natural language) to SQL queries, also known as Text-to-SQL. This is often done with zero-shot prompting[2], as LLMs already have sufficient knowledge of SQL (Liu, Hu, Wen, & Yu, 2023). Metadata, such as names and descriptions of tables or fields, need to be provided to LLMs via contexts or using retrieval techniques such as RAG. Few-shot prompting, which includes examples in the context, can further enhance Text-to-SQL capabilities (Rajkumar, Li, & Bahdanau, 2022).

Text Embedding

Text embedding is a technique used in natural language processing (NLP) to convert texts into numerical vectors that capture the semantic meaning of the words or phrases. These embeddings allow AI to understand and process textual data more effectively by representing words in a continuous vector space, where similar words are positioned closer together. Text can be embedded at different levels such as words, sentences or paragraphs (Le & Mikolov, 2014). In LLMs and GenAI applications, texts are often embedded at the level of tokens (Nie, Zhang, & Wu, 2024). This is often done automatically using commercial or open-source LLM services during the inference process. However, for other processes such as document retrieval (see

[2] https://platform.openai.com/docs/examples/default-sql-translate

Retrieval-augmented generation), embeddings need to be applied to the source documents and input questions in order to identify relevant documents. Many pretrained models can be used for embedding tasks such as OpenAI's text-ada-002 (Greene, Sanders, Weng, & Neelakantan, 2022) and BGE-M3 (Chen, et al., 2024).

Retrieval-augmented Generation (RAG)

LLMs used in GenAI are pretrained with fixed knowledge, but do not have access to external knowledge. This external information, such as the organization's data, specific domain knowledge or user preferences, needs to be supplied to LLMs as context. The task is commonly handled with the retrieval-augmented generation (RAG) process (Lewis, et al., 2020). First, candidates of relevant chunks or entries of information are selected. This may be done by finding similarities between the vector of the input questions and the vectorized chunks (Asai, Min, Zhong, & Chen, 2023). Top candidates are added to the context of the LLMs. LLMs can then generate responses with the knowledge of this relevant information.

Multi-agent Framework

Using multiple agents of LLMs can expand the capabilities of generative AI applications, especially for complex tasks. This is because multiple agents can provide feedback to one another, be configured for different capabilities and break down complex problems into smaller tasks (Wu, et al., 2023). Several frameworks exist such as AutoGen[3], LangGraph[4], and CrewAI[5].

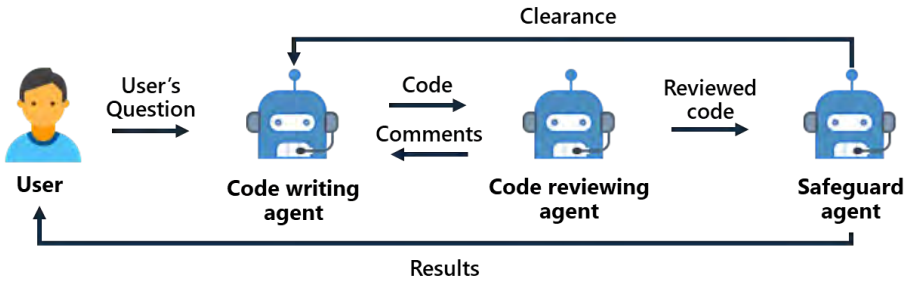
In multi-agent generative AI applications, the task is broken down into smaller tasks, each handled by agents with different configurations. For example, in a coding assistant, the task can be broken down into code writing, code reviewing and safeguarding, as illustrated in Figure 1. First, the code writing agent, prompted to interpret user questions and write codes (e.g., in Python) based on instructions. Then, the output is sent to the code reviewing agent, which is instructed to review the code following some standards. The code reviewing agent may send the comments back to the code writing agent for revisions or pass it to the safeguard agent, which will check for security and compliance. This workflow ensures that all the steps are taken care of. Another advantage of this approach is that only the relevant instructions and context are provided to each agent, reducing the problems of exceeding the context limits of LLMs. In addition, users can also see intermediate exchanges between agents and may also participate in the process.

[3] https://github.com/microsoft/autogen

[4] https://www.langchain.com/langgraph

[5] https://github.com/crewAIInc/crewAI

Example of a multi-agent workflow for a coding assistant AI Figure 1



An example of a multi-agent workflow for a coding assistant. The workflow involves three distinct agents: the Code Writing Agent, responsible for generating code; the Code Reviewing Agent, tasked with evaluating and providing feedback on the generated code; and the Safeguard Agent, ensuring compliance with safety and security standards. This collaborative approach enhances the efficiency and reliability of the coding process.

Evaluating Generative AI Response

Evaluating responses from GenAI can be both objective and subjective. For example, retrieving the correct data field is objective, whereas producing useful suggestions can be subjective. In our context, objective evaluation involves information retrieval problems. Several metrics can be used as in information retrieval systems, such as precision and recall (Saracevic, 1995).

Precision measures the proportion of relevant instances from all the retrieved instances. This metric can be used to identify the relevancy of the retrieved information. **Recall** measures the proportion of the relevant instances that were retrieved from all relevant instances. It can be used to determine whether all the relevant data fields, for instance, are retrieved. Mathematically, precision and recall are defined as:

$$\text{Precision} = \frac{\text{Relevant retrieved instances}}{\text{All retrieved instances}},$$

$$\text{Recall} = \frac{\text{Relevant retrieved instances}}{\text{All relevant instances}}.$$

Precision and recall values range from 0 to 1, where a precision of 1 means all the retrieved information is relevant and a recall of 1 means the algorithm was able to retrieve all the relevant information. In the context of GenAI responses, precision and/or recall can be used to determine the **relevancy** of the responses.

For coding tasks, such as SQL generation, the validity of the code generated can also be evaluated (Chen, et al., 2024). **Validity** may be defined as the percentage of code that is valid in terms of compilability, type-checked syntax or in successfully performing desired tasks. This may be done automatically or by human evaluation.

Some evaluations need to be subjective: for example, the usefulness of the output is best determined by the users of the GenAI (Baek, Chandrasekaran, Cucerzan, Herring, & Jauhar, 2024). **Usefulness** determines whether the outputs to the users' input questions are beneficial to them. Such a metric can be directly evaluated by users, such as through a survey using a 7-point Likert scale, where scores range from 1 (strongly disagree) to 7 (strongly agree). Alternatively, these metrics can be evaluated automatically by using other LLMs as evaluating agents (Gao, et al., 2023).

However, their assessment might not capture the opinions of actual users with diverse expertise and expectations, so evaluating agents are not leveraged in this study.

3. Methodology

We develop three complementary modules—Brainstormer, Metadata Copilot, and SQL Coder—that leverage customized RAG, LLM and Structured Data Validation. While these modules share core components, each has a distinct agent workflow implemented via AutoGen multi-agent framework. The details and structure of each core component are outlined below.

Customized RAG

RAG can be utilized to enhance LLM by retrieving relevant information from external database before generating responses to users. This architecture is particularly useful for our application because it provides LLM with necessary details of RDT metadata, especially the definition of each field, enabling LLM to assess the situations and provide information more accurately.

The RAG system in our implementation utilizes a hybrid retrieval approach that combines dense and sparse vector representations to improve accuracy and efficiency of the system. The system processes user queries through parallel tracks, leveraging both embedding-based techniques and traditional information retrieval methods to capture both of semantic and keyword relationships.

In the first track, user's queries are passed through the BGE-M3 embedding model (Chen, et al., 2024) and compared against those of our documents stored in a vector database called Qdrant[6]. This approach captures semantic relationships and contextual nuances that might be missed by traditional keyword matching.

Simultaneously, in the second track, we also use Term Frequency-Inverse Document Frequency (TF-IDF) and the BM25 ranking algorithm (Robertson, Walker, Hancock-Beaulieu, Gatford, & Payne, 1996) to process users' query. This combination allows us to capture keyword matches and term importance effectively. Finally, we also implement the negative Matrix Factorization (NMF) to identify latent topics and improve retrieval relevance. This multi-track approach allows our system to maintain high retrieval accuracy across various query types, from specific technical inquiries to more general, contextual questions about RDT dataset. The results from both tracks are then aggregated to select the most relevant previous cases from our dataset gathered in the first step; these reference cases will be used as a context for the LLM later.

LLMs and Structured Data Validation

In this step, we integrate LLMs with structured data handling mechanisms. The system processes questions from users; these questions can be classified into three main categories based on the nature of each application: high-level analytical questions about business insights (Brainstormer), metadata field definitions (Metadata Copilot), and SQL code for specific data retrieval problems (SQL Coder). Through a RAG

[6] https://qdrant.tech/

pipeline, the system pulls relevant information from our RDT metadata database to address these queries. Our private Azure OpenAI subscription is prompted to analyze the RAG results and answer users' questions accordingly using *gpt-4o-mini*[7] model.

To ensure consistency in the output format, we implement a data validation framework using Pydantic[8], combined with JSON formatting. This framework enforces a predefined schema on output from LLMs to include all necessary components such as relevant metadata fields, detailed analysis of users' questions, and comprehensive explanations. The structured output schema plays a crucial role in maintaining response quality. Each query response generated by the system must include a clear breakdown of the problem, relevant metadata field descriptions, or SQL codes depending on the application.

Application Deployment

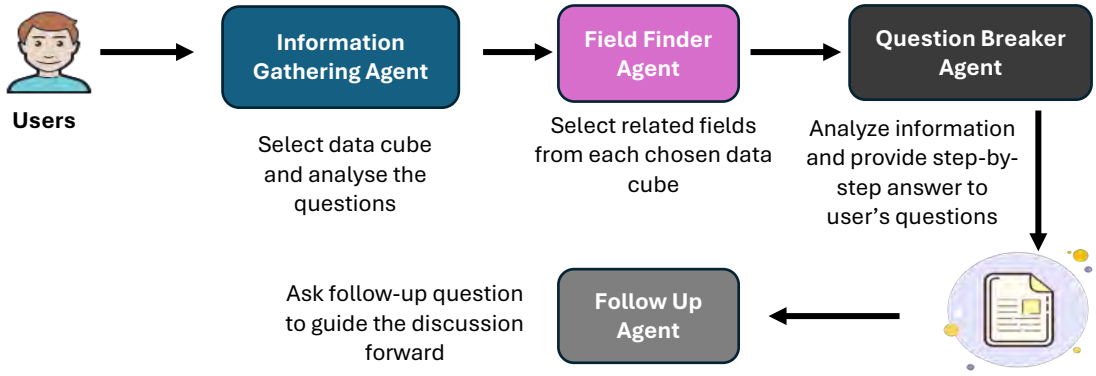
The deployment phase of the system utilizes Streamlit[9] as the primary frontend framework, offering a user-friendly web interface for accessing the service. The application is hosted on an internal server. To manage access control and maintain user session data, we implement a secure user authentication system for all users. Additionally, we also implement comprehensive logging mechanisms that capture user interactions, including both query inputs and system responses, to facilitate future improvement and system optimization. This logging system stores detailed interaction data, including timestamps, user identifiers, query content, RAG retrieval results, as well as the final outputs from each application. The collected data serves multiple purposes: enabling performance monitoring, identifying common questions about RDT, and evaluating the accuracy of application outputs.

Multi-agent Workflow

This subsection outlines the RDT Copilots agents' workflow.

Brainstormer workflow

Figure 2



Multi-agent workflow for Brainstormer module. Each box corresponds to an LLM agent.

7 https://www.openai.com/

8 https://pydantic.dev/

9 https://streamlit.io/

Brainstormer

Brainstormer serves as a chatbot interface designed to decompose complex, high-level queries into manageable sub-questions that can be answered with RDT dataset. The system operates through a sequence of specialized LLM agents as shown in Figure 2. Initially, the Information Gathering agent analyzes user queries alongside detailed cube descriptions in its prompt to identify and select the most suitable data cubes for further processing. These selected cubes are then passed to the Field Finder agent, which extracts relevant fields based on the user's query. The Question Breaker agent utilizes these selected fields as context to decompose the original query and provide guidance for RDT-based solutions. To propose new ways of problem solving, the Follow Up agent generates targeted questions that guide the conversation toward more insightful outcomes. This multi-agent architecture ensures a systematic and thorough approach to query processing, while maintaining user engagement throughout the interaction. The modular design not only enhances the system's efficiency but also allows for future scalability and improvements. By breaking down high-level queries into smaller, manageable components, Brainstormer significantly reduces the complexity while maximizing the utility of the RDT dataset.

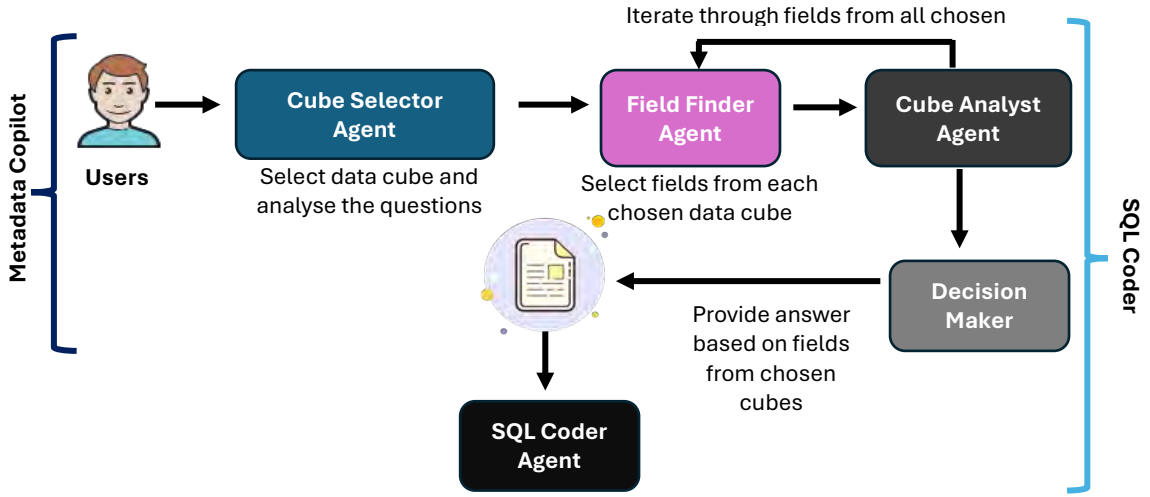
Metadata Copilot

Metadata Copilot is a specialized chatbot system focused on retrieving fields from the RDT dataset through our customized RAG pipeline. The system utilizes a system of agents to process user queries and provide useful fields effectively as shown in Figure 3; this will reduce the time users need to spend on manually searching for fields from RDT's complicated data dictionary. Beginning with the Cube Selector agent, the system analyzes user questions to identify and select the most relevant data cubes from the available collection. These selected cubes are then processed by the Field Finder agent, which extracts pertinent fields based on the query context. The Cube Analyst agent performs a critical review of the extracted fields, initiating an iterative feedback loop with the Field Finder if the initial field selection is inaccurate. Finally, the Decision Maker agent evaluates the collective findings and generates a definitive list of fields to provide to users. This multi-agent approach is beneficial in this context as it distributes tasks across several specialized agents, reducing the workload assigned to each individual agent while improving accuracy and minimizing hallucination in the output. Furthermore, this distributed architecture leverages the combined context windows of multiple agents, enabling the entire system to process more information from data dictionary than would be possible with a single agent.

SQL Coder

SQL Coder serves as a query generation system that transforms field-level metadata into SQL queries for retrieving data from the RDT dataset. Building upon Metadata's field selection capabilities, the system extends the processing pipeline with the addition of the specialized SQL Coder agent. Once the Decision Maker agent from Metadata finalizes the field list, SQL Coder agent analyzes the selected fields along with user query and selected cubes to determine SQL operations such as table joins, filtering conditions, and aggregation requirements. Finally, SQL Coder agent will write the executable SQL code that satisfies this requirement and provide to users. With SQL Coder, users can paste this provided code directly into their data portal and quickly retrieve the RDT data.

Metadata Copilot & SQL Coder workflow Figure 3



Multi-agent workflow for Metadata Copilot and SQL Coder modules. Each box corresponds to an LLM agent.

4. Experiment

In this section, we describe the experimental framework designed to evaluate the performance of our proposed multi-agent workflow in comparison to the baseline single-agent algorithm in copilot tasks aimed at assisting non-technical users in leveraging the RDT dataset. Recall that the multi-agent method outperforms its single-agent counterpart in terms of **context efficiency**: due to its extensive resource consumption, the single-agent algorithm is rendered impractical for deployment on this database. In this experiment, we aim to demonstrate that the multi-agent's task decomposition scheme also delivers superior or at least comparable **output quality** on these copilot tasks to its highly accommodated single-agent implementation in our simplified testing environment.

Tasks and Evaluation Methods

While all tasks are inherently interactive, for ease of performance comparison between the two algorithms, we consider the simplified one-round output-only scenario, thereby reducing to the question-answering setting. To this end, we create two types of questions: 62 **quantitative** questions aimed at generating specific reports (values or tables), and 30 **qualitative** questions seeking comprehensive guidance on solving analytical problems faced by central bankers, such as policy-related issues or enforcement of supervision criteria. We include 10 (translated) questions of each type in Appendix 1.

We evaluate our algorithms on the three following tasks.

Brainstormer. Given a *qualitative* question, the algorithm must suggest plausible solutions by decomposing the problem into sub-questions that can be addressed through database queries in a step-by-step fashion, while identifying essential data fields to demonstrate feasibility. It is not required to provide actual

answers or queries to solve the problem at this point: by design, the users must decide to pursue these steps themselves and will be further assisted by Metadata Copilot.

Metadata Copilot. Given a *quantitative* question, the algorithm must retrieve columns from one or more tables that are necessary for computing the answer (including intermediate fields required for filtering data, deduplicating records, and joining tables). This task requires an understanding of the data model, background concepts in central banking and related subjects, as well as abbreviations and jargons used internally by our bank examiners.

SQL Coder. Given a *quantitative* question, the algorithm must provide an SQL code that, once inputted to an SQL query execution engine, generates the desired report. While the input is simply the free-text question, our internal prompt includes separated instructions for short-listing columns (also featured in Metadata Copilot) and generating a query: the algorithm's performance can consequently be overshadowed by the quality of its former step.

To measure our algorithms' performance, we consider two evaluation methods: the automated evaluation with preset grading schemes, and the user evaluation that relies on human scoring. Their differences lie in grading objectives, processes, metrics, and underlying data tables, as summarized in Table 1.

Comparison between Automated Evaluation and User Evaluation Table 1

	Automated Evaluation			**User Evaluation (all tasks)**
	Brainstormer	**Metadata Copilot**	**SQL Coder**	
objective	relevancy	relevancy	validity	usefulness
metric	recall	recall	0/1	7-point Likert scale
dataset	multiple tables with ≤ 300 columns each (join queries and intermediate calculations may be required)			one table with ≈ 1,000 columns (feature-engineered columns included)

Automated Evaluation. This method assesses the quality of the output using pre-determined sets of criteria, allowing an automated testing on a larger number of tests, thus providing a more thorough coverage of the data fields. We measure different aspects of the algorithms' performance for our tasks.

- **Relevancy.** For the Brainstormer and Metadata Copilot tasks, we measure their ability to retrieve related columns by comparing their output columns against our answer keys and compute the recall rate. We design our answer keys for Metadata Copilot to cover *attribute groups* that should appear in the resulting SQL query, whereas for Brainstormer we focus on *concepts* that should be touched upon when solving the qualitative question. Rather than matching specific columns, we award a true positive point when they cover any fields in the required groups (e.g., business type/group/ISIC) or concepts (e.g., DPD or stage). We do not explicitly set a limit on the number of allowed output columns—with our prompts, both algorithms return a reasonable number of columns for their respective tasks and questions (up to 9 and 21 columns for Metadata Copilot and Brainstormer, respectively).
- **Validity.** For the SQL Coder task, we measure the SQL codes solely for their compilability, for two reasons. First, we aim to isolate the coding component from the business aspect, for it has already been evaluated above during the retrieval subtask, from which the coder subtask receives its short-listed

columns. Second, given the colloquialism of questions and the redundancy of data fields, the desired reports are generally underspecified (which can be resolved via a multi-turn conversation, in practice). As the correctness evaluation requires human judgement, we delegate it to the user evaluation.

The dataset for automated evaluation consists of smaller, low-level tables that often require complex join operations and intermediate computations. It tests the algorithms' capabilities in handling fundamental database structures where technical complexities are prevalent.

User Evaluation. In all three tasks, we conducted a user evaluation to gauge the perceived **usefulness** of the algorithms' outputs. Participants are presented with the outputs and asked to rate their agreement with the statement "*The algorithm's output is useful for solving the given problem*" on a 7-point Likert scale, from strongly disagree (score = 1) to strongly agree (score = 7). We intentionally let the participants decide on their own notion of "useful". At the same time, we prompt our algorithms to include explanations to their answers to ensure that the participants understand and are convinced of their correctness. Due to the participants' limited availability, we only select 10 questions of each type (precisely listed in Appendix 1) for user evaluation. We provide translated answers to some questions in Appendix 2.

The dataset for user evaluation is a single table indexed by loan account and data period as unique keys. Designed as a table for non-technical users, it contains many feature-engineered business-oriented additional fields, such as period-over-period computations, cumulative or window functions, and grouping-aggregation into different credit-line levels, amounting to roughly 1,000 features. This dataset focuses on evaluating the algorithms' ability to handle large metadata that requires business acumen, as well as the ability to eloquently express their answers to users who have limited familiarity with the data fields.

Accommodating Single-Agent Algorithm

Recall that the single-agent approach fails to execute on the RDT dataset because its limited context window size is insufficient for our metadata. As we aim to measure output qualities, we accommodate for the single-agent framework as follows. We let the single-agent algorithm skip earlier subtasks that otherwise requires large context and provide these subtasks' outputs directly. As described in Table 2, the single-agent algorithm is given relevant data cubes and shortlisted data fields, thereby bypassing the context-intensive cube selection and RAG steps altogether. In contrast, the multi-agent algorithm handles RDT metadata by distributing the relevant parts to its agents, which then perform their respective subtasks and passing their small outputs around according to our predefined workflow.

Comparison between Implementation of Multi- and Single-Agent Algorithms Table 2

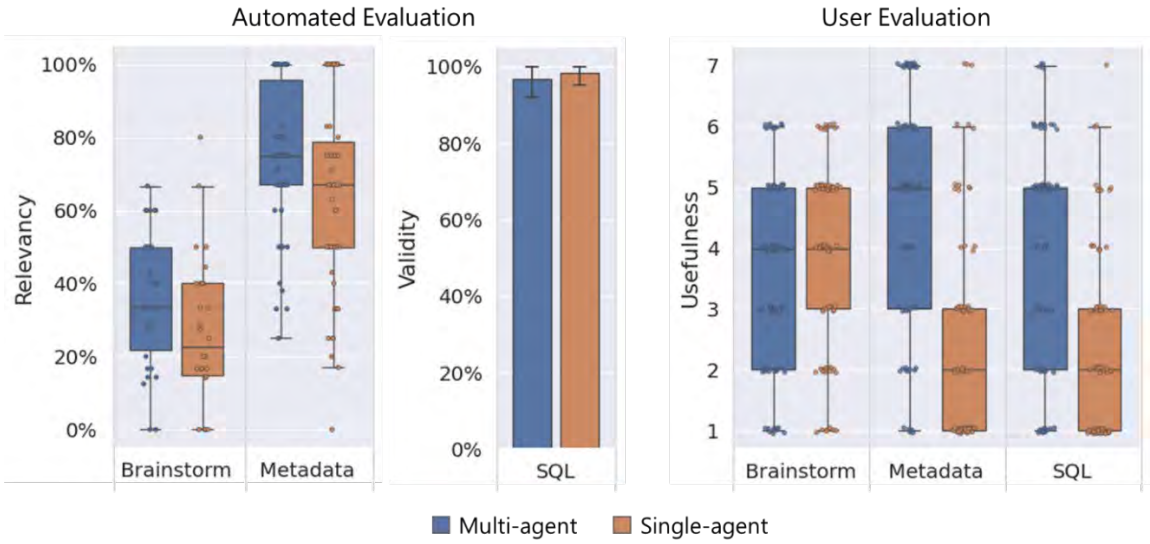
Subtask	Multi-Agent	Single-Agent
Data cube selection	general description of up to three cubes is given; the algorithm chooses relevant cubes	general descriptions of strictly required cubes are given; no cube selection required
RAG	field description of chosen cubes is given; the algorithm performs RAG	a retrieved list of fields along with their description is given; no RAG required
Brainstormer /Metadata Copilot / SQL Coder	agents solve the remaining subtasks using the data computed by other agents	one agent solves the remaining subtasks from the given prepared data

While the single-agent algorithm seemingly receives a huge handicap in this setup, we mitigate this advantage for a fairer comparison as follows. Firstly, we prompt the single-agent algorithm to reproduce the multi-agent workflow by guiding it through the same intermediate subtasks. Secondly, for the automated evaluation which involves smaller low-level cubes (tables), we restrict the questions to simpler ones that can reasonably be solved with at most two cubes, and always provide only three total cubes for selection. In essence, while the multi-agent scheme is vital for enabling our AI assistant to handle the large RDT metadata, this simplified experiment makes the single-agent framework feasible while keeping the overall task's difficulty on a rather even playing field, suitable for evaluating output quality.

5. Result & Discussion

Evaluation Results

Figure 4



This chart shows results from automated and user evaluations, using relevancy, validity and usefulness as metrics. For relevancy, each dot represents a recall score for each question in retrieving expected fields. For validity, the bars show the percentage of valid SQL queries. The error bars represent 95% CI. For usefulness, each dot represents a usefulness score (1 = strongly disagree, 7 = strongly agree) for each response, with jitter added to improve visibility.

Summary of evaluation results

Table 3

Evaluation	Module	Metric	Sample Size	Mean score		p-value	Significance
				Multi-agent	Single-agent		
Automated	Brainstormer	Relevancy	30 questions	**73.7%**	64.8%	0.006	**
	Metadata Copilot	Relevancy	62 questions	96.8%	**98.4%**	0.564	
	SQL Coder	Validity	62 questions	**35.9%**	26.9%	0.011	*
Human	Brainstormer	Usefulness	10 questions × 8 users = 80	**4.763**	2.363	9.87E-12	***
	Metadata Copilot	Usefulness	10 questions × 8 users = 80	**4.025**	2.313	5.74E-09	***
	SQL Coder	Usefulness	10 questions × 8 users = 80	3.675	**3.913**	0.100	

The framework with superior score is bold for each row. P-values and significance levels are calculated using Wilcoxon signed-rank test to determine whether there is a significant difference between multi-agent and single-agent frameworks. Asterisks refer to significance level (* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$).

The evaluation results are summarized in Figure 4 and Table 3. For the automated evaluation, the results show that RDT copilots can perform objective tasks, such as field retrieval and SQL generation for data querying, quite reliably. Metadata Copilot with the multi-agent framework achieves an average relevancy score of 73.7%, improving upon the single-agent framework's (64.8%). This margin suggests that using multiple agents can enhance GenAI's data retrieval capabilities. SQL Coder obtains a high validity score, but there are no significant differences between multi-agent (96.8%) and single-agent (98.4%) frameworks. One plausible explanation is that the current LLM excels at SQL generation under the single-agent framework: adopting multiple agents here would yield negligible benefit. However, subjective tasks such as Brainstormer show an unsatisfactory performance, with average relevancy scores of 35.9% for the multi-agent framework and 26.9% for the single-agent framework, suggesting that the task of answering open-ended questions in a single turn may be overly ambitious.

For user evaluation, our analysts and examiners found multi-agent Metadata Copilot and SQL Coder slightly useful for answering quantitative questions, with average usefulness scores of 4.76 and 4.02, respectively, markedly surpassing the single-agent framework's scores of 2.36 and 2.31. For Metadata Copilot, users tend to rate more highly when the relevant fields are selected, and the multi-agent implementation seemingly outperforms the single-agent counterpart under user evaluation's dataset. As for SQL Coder, the multi-agent version reportedly provides better coding structures and clearer explanation. However, they slightly disagree that the multi-agent Brainstormer is useful for answering qualitative questions (3.68); this score is non-negligibly worse than the single-agent Brainstormer's (3.91), apparently due to the responses' writing styles: some users mentioned that the answers of multi-agent Brainstormer are too general or incoherent.

There are significant variations in relevancy and usefulness in both evaluations, indicating that some questions are harder for these agents. Some users mentioned that GenAI performed worse for questions that require domain expertise. This is likely because the information in gpt-4o-mini's training data and our data dictionary

combined may not be sufficient for performing such tasks. User ratings for usefulness also vary greatly even within the same question, likely due to users' different coding skills and expectations. Designing fair and relevant experiments, test questions, and evaluations for GenAI applications remains a big challenge.

6. Conclusion and Further Work

Conclusion

In this paper, we explore the application of generative AI in the context of the Bank of Thailand's Regulatory Data Transformation (RDT) program. Our approach leverages a multi-agent framework to enhance the utilization of granular credit data by central bank analysts and examiners. By breaking down complex tasks into manageable steps, our RDT Copilots—comprising Brainstormer, Metadata Copilot, and SQL Coder—streamline the data analysis process and improves the accuracy and efficiency of responses.

For RDT Copilots, the multi-agent framework significantly improves the performance of GenAI applications compared to the single-agent approach, especially for retrieving relevant fields and producing useful output for data querying. These tasks benefit from the multi-agents' larger context sizes and abilities to break down problems into smaller tasks. The framework can be applied to different tasks, including more objective task as well, allowing more flexible workflows for GenAI applications.

User evaluation indicates that RDT Copilots can assist users in solving objective or quantitative problems such as searching and querying data for examining tasks. Most of participants' responses suggest that the outputs generated by Metadata Copilot and SQL Coder are generally useful, aside from certain questions that require deeper domain expertise. Subjective or qualitative tasks such as suggesting analytical frameworks for solving open-ended questions, however, remain challenging for our Brainstormer module. Overall, we believe that RDT Copilots have the potential to mitigate the challenges associated with RDT data utilization and accelerate the success of the transformation program.

Further Work

Future work will focus on improving the performance of RDT Copilots. One gap is to improve the quality of the metadata and enrich the domain knowledge for RDT Copilots in systematic ways, which will enable more accurate and relevant outputs. In addition, Brainstormer needs to be revised to produce more useful outputs. This may include designing a more detailed workflow, improving prompt engineering techniques or adopting LLMs with reasoning capabilities.

We are also investigating other potential use cases that will help with utilizing RDT data, such as adding the capabilities to generate graphs from the query. Other use cases of multi-agent GenAI applications are also being explored, including but not limited to Knowledge Management Chatbot and Supervision Assistant Agents.

References

Aldasoro, I., Doerr, S., Gambacorta, L., Notra, S., Oliviero, T., & Whyte, D. (2024). Generative artificial intelligence and cyber security in central. *BIS Papers*(No 145).

Asai, A., Min, S., Zhong, Z., & Chen, D. (2023). Retrieval-based Language Models and Applications. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, (pp. 41-46).

Baek, J., Chandrasekaran, N., Cucerzan, S., Herring, A., & Jauhar, S. K. (2024). Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion. *Proceedings of the ACM Web Conference 2024*. doi:10.1145/3589334.3645404

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at Work. *NBER Working Papers*(No 31161).

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv*. doi:10.48550/arXiv.2402.03216

Chen, L., Guo, Q., Jia, H., Zeng, Z., Wang, X., Xu, Y., . . . Zhang, S. (2024). A Survey on Evaluating Large Language Models in Code Generation Tasks. *arXiv*. doi:arXiv:2408.16498v1

Du, Y., Li, S., Torralba, A., Tenenbaum, J., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. *Arxiv*. doi:10.48550/arXiv.2305.14325

ECB. (2023, November 15). Suptech: thriving in the digital age. *Supervision Newsletter*.

ECB. (2024, September 18). The future of European banking supervision – connecting people and technology. *Keynote Speech*.

Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2024). Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. *Proceedings of the VLDB Endowment*, *17*, pp. 1132-1145.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). Human-like Summarization Evaluation with ChatGPT. *ArXiv*. doi:10.48550/arXiv.2304.02554

Greene, R., Sanders, T., Weng, L., & Neelakantan, A. (2022, Dec 15). *New and improved embedding model*. Retrieved from OpenAI: https://openai.com/blog/new-and-improved-embedding-model

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *ArXiv*. doi:10.48550/arXiv.1405.4053

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, (pp. 9549-9474).

Liu, A., Hu, X., Wen, L., & Yu, P. S. (2023). A comprehensive evaluation of ChatGPT's zero-shot Text-to-SQL capability. *arXiv*. doi:10.48550/arXiv.2303.13547

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., . . . Tang, J. (2023). AgentBench: Evaluating LLMs as Agents. *Arxiv*. doi:10.48550/arXiv.2308.03688

MAS. (2023, November 16). MAS Launches Digital Platform for Seamless ESG Data Collection and Access. *Media Releases*.

Nie, Z., Zhang, R., & Wu, Z. (2024). A Text is Worth Several Tokens: Text Embedding from LLMs Secretly Aligns Well with The Key Tokens. *ArXiv*. doi:10.48550/arXiv.2406.17378

Rajkumar, N., Li, R. L., & Bahdanau, D. (2022). Evaluating the Text-to-SQL Capabilities of Large Language Models. *arXiv*. doi:10.48550/arXiv.2204.00498

Robertson, S., Walker, S., Hancock-Beaulieu, M., Gatford, M., & Payne, A. (1996). Okapi at TREC-4. *The Fourth Text REtrieval Conference (TREC-4)*.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 138-146).

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., . . . Wang, C. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *ArXiv*. doi:10.48550/arXiv.2308.08155

Appendix 1: Questions for Evaluation

Quantitative Questions for User Evaluation (translated)

1. Total outstanding balance in the latest month, among the debtors who had a total outstanding balance across all accounts exceeding 500 million baht as of December 31, 2023.
2. Monthly counts of debtors who have joined the debt restructuring program for each FI.
3. Highest and lowest PD of loan contracts in the month prior to changing status to SM, by type, by FI.
4. YoY growth of the outstanding balance of business loans.
5. Average income of new debtors in each data period.
6. Show the total outstanding balance of debtor groups that belong to more than five different business types.
7. Calculate the total outstanding balance for the latest month, of the debtor groups who have at least one NPL contract.
8. Show housing loans where the total collateral value does not exceed 20% of the approved loan amount.
9. Calculate the total outstanding balance per debtor, only for debtors who have ever had a maximum DPD greater than 90 days.
10. Total change in the outstanding balance of debtors who have been SM/NPL within the last 12 months.

Qualitative Questions for User Evaluation (translated)

1. Do different FIs consistently appraise the value of real estate collateral in each area?
2. Forecast the volume of NPLs in the next year.
3. In the past quarter, have FIs become more cautious or less cautious in lending to SMEs and retail customers compared to previous periods?
4. Does a reduction in policy interest rate genuinely reduce the burden on debtors?
5. To what extent has the 2024 LTV measure—which allows borrowers to borrow an additional 10% above the 100% LTV ceiling for housing loans during 2024—been utilized in practice?
6. Estimate customers' income using payment history to create a risk profile.
7. Do financial institutions use a DSR threshold when evaluating loan applications?
8. In 2024, a new measure was introduced to facilitate co-borrowers without home ownership rights to more easily access housing loans. Has this measure been effective?
9. High Interest Rates: Are They Driving Bank Profits Through the Roof?
10. What should the minimum payment rate be for credit card debt repayments?

Appendix 2: Responses for Evaluation Questions

As a reminder, these results are for the user evaluation. The only given data table is smart_cube_1, but the algorithms are still required to perform the cube selection task to illustrate their complete pipeline.

Responses for Quantitative Question 10 (translated)

Question: Total change in the outstanding balance of debtors who have been SM/NPL within the last 12 months.

Multi-agent algorithm's response for Metadata Copilot

User's Requirement Analysis:

The user wants to track the total outstanding balance changes of debtors who were previously SM/NPL over the last 12 months. This requires contract/account-level data from various financial institutions and information on the classification of credit debt.

Cube Selection and Reason:

Selected Cube: Smart_Cube_1

Reason: Smart_Cube_1 contains comprehensive data including financial institution codes, debtor account numbers, credit debt classification, data periods, and total outstanding balances. All of these are necessary for analyzing the total outstanding balance changes of debtors who were previously SM/NPL over the last 12 months.

Analysis of Relevant Columns:

Column Name	Description	Relevance to Requirement
fi_bot_unique_key	Unique code of the financial institution	Used to identify the financial institution related to the debtor data
account_id	Debtor's account number	Necessary for tracking account-level debtor data
entity_id	Unique code of the debtor group	Used to group debtors for analysis
debtor_bot_unique_key_ merge_mask	Unique code of the debtor	Used to identify individual debtors
asset_and_contingent_ class_code	Asset and contingent classification code	Used to identify the status of credit debt (e.g., SM/NPL)
period	Data period (year)	Used to identify the time period of the data to be analyzed
outstanding	Outstanding credit balance before deducting unearned revenue	Used to analyze the total outstanding balance of debtors
change_class_yoy_level1	Change in classification compared to 12 months ago	Used to analyze changes in the status of credit debt over the last 12 months

Multi-agent algorithm's response for SQL Coder

SQL Query:

```
WITH recent_periods AS (
    SELECT DISTINCT period
    FROM Smart_Cube_1
    ORDER BY period DESC
    LIMIT 12
),
sm_npl_accounts AS (
    SELECT DISTINCT account_id
    FROM Smart_Cube_1
    WHERE asset_and_contingent_class_code IN ('SM', 'NPL')
)
SELECT
    fi_bot_unique_key,
    account_id,
    SUM(outstanding) AS total_outstanding_balance,
    SUM(change_class_yoy_level1) AS total_change_in_outstanding_balance
FROM
    Smart_Cube_1
WHERE
    period IN (SELECT period FROM recent_periods)
    AND account_id IN (SELECT account_id FROM sm_npl_accounts)
GROUP BY
    fi_bot_unique_key,
    account_id
ORDER BY
    fi_bot_unique_key,
    account_id;
```

Description:

- **recent_periods:** Create a temporary table that stores the data periods for the last 12 months, ordered from the most recent to the oldest period.
- **sm_npl_accounts:** Create a temporary table that stores the account numbers of debtors who were previously SM or NPL.
- **Main Query:**
    - Select data from Smart_Cube_1 that is within the last 12 months and includes accounts that were previously SM or NPL.
    - Calculate the total outstanding balance (total_outstanding_balance) and the change in the outstanding balance (total_change_in_ outstanding_balance).
    - Group the data by financial institution code (fi_bot_unique_key) and account number (account_id).
    - Order the data by financial institution code and account number.

Single-agent algorithm's response for Metadata Copilot

User's Requirement Analysis:

The user needs information on the total outstanding balance changes of debtors who were previously SM/NPL over the last 12 months. This means tracking the changes in the outstanding balances of debtors who had the status of SM (Special Mention) or NPL (Non-Performing Loan) during the specified period.

Cube Selection and Reason:

Selected Cube: Smart_Cube_1

Reason: Smart_Cube_1 contains comprehensive data including financial institution codes, debtor account numbers, credit debt classification, data periods, and total outstanding balances. All of these are necessary for analyzing the total outstanding balance changes of debtors who were previously SM/NPL over the last 12 months.

Analysis of Relevant Columns:

Column Name	Description	Relevance to Requirement
period	Data period (year)	Used to identify the last 12 months
fi_bot_unique_key	Financial institution code	Used to identify the financial institution related to the debtor
account_id	Account number	Used to identify the debtor's account to be tracked
asset_and_contingent_class_reason_code	Asset and contingent classification reason code	Used to identify the status of the debtor as SM or NPL
npl_reentry_date_latest	Date of the latest reentry into NPL status	Used to identify changes in the debtor's status

Note: Selecting these columns helps efficiently track changes in outstanding balances and debtor status over the specified period.

Single-agent algorithm's response for SQL Coder

To find the total outstanding balance changes for debtors who were previously classified as SM (Special Mention) or NPL (Non-Performing Loan) within the last 12 months, we need to filter the data based on the debtor's classification and the period. We will use the asset_and_contingent_class_reason_code to identify the SM/NPL status and the period to filter the last 12 months.

Here's the SQL query to achieve this:

```
WITH RecentPeriods AS (
    SELECT DISTINCT period
    FROM CUBE_smart_cube_1
    ORDER BY period DESC
    LIMIT 12
)

SELECT
    entity_id,
    SUM(other_debt_burden) AS total_outstanding_balance_change
FROM
    CUBE_smart_cube_1
WHERE
    asset_and_contingent_class_reason_code IN ('2000900005', '2000900019') --
Assuming these codes represent SM/NPL
    AND period IN (SELECT period FROM RecentPeriods)
GROUP BY
    entity_id
```

Explanation:

1. **RecentPeriods CTE:** This Common Table Expression (CTE) selects the distinct periods from the table and orders them in descending order to get the most recent periods. We limit the result to 12 to cover the last 12 months.
2. **Main Query:**
    - We select the entity_id to group the results by each debtor.
    - We sum the other_debt_burden to calculate the total outstanding balance change for each debtor.
    - We filter the records where asset_and_contingent_class_reason_code is either '2000900005' or '2000900019', which are assumed to represent SM/NPL statuses.
    - We further filter the records to include only those from the last 12 months by checking if the period is in the list of recent periods obtained from the CTE.

This query efficiently calculates the required total outstanding balance changes for the specified debtors over the last 12 months.

Responses for Qualitative Question 10 (translated)

Question: What should the minimum payment rate be for credit card debt repayments?

Multi-agent algorithm's response for Brainstormer

To answer the question about the minimum payment rate for credit card debt repayments, we can use data from Smart_Cube_1, which contains information about credit cards and debt repayments. The analysis involves the following steps:

Step 1: Gather Credit Card Information

- **Fields Involved**: credit_card_type_code, credit_card_type_name, credit_card_parents_account_id, credit_card_card_holder_counterparty_id
- **Reason**: This information helps identify the type of credit card and the cardholder, which is essential for analyzing debt repayments.

Step 2: Gather Debt Repayment Information

- **Fields Involved**: number_of_principal_payment, principal_payment_frequency, interest_payment_frequency, billing_default_interest, billing_fee
- **Reason**: This data provides details on the number of repayment installments, the frequency of payments, and associated fees, which are crucial for calculating the minimum payment rate.

Step 3: Analyze Debt Repayment Data

- **Fields Involved**: default_interest_rate_total, default_interest_rate_additional, default_interest_rate_base_interest_rate_for_calculation
- **Reason**: This information reveals the interest rates charged for defaulting on payments, which is a key factor in determining the minimum payment rate.

Step 4: Calculate the Minimum Payment Rate

- **Method**: Use the data from Steps 2 and 3 to calculate the minimum payment rate, considering the number of installments, payment frequency, and interest rates.
- **Expected Outcome**: An appropriate minimum payment rate for credit card debt repayments.

Step 5: Compare with Standards and Policies

- **Fields Involved**: payment_frequency_condition, provision_amount_undrawn
- **Reason**: This data helps compare the repayment conditions and available provisions with financial institution standards and policies.

Step 6: Summarize and Provide Recommendations

- **Method**: Summarize the analysis results and provide recommendations on the appropriate minimum payment rate for credit card debt repayments.
- **Expected Outcome**: Useful suggestions for setting an appropriate minimum payment rate.

Using data from Smart_Cube_1 in these steps will help effectively analyze and determine the minimum payment rate for credit card debt repayments.

Single-agent algorithm's response for Brainstormer

To answer the question about the minimum payment rate for credit card debt repayments, we can create sub-questions to help analyze and evaluate the appropriate rate as follows:

Analyzing the Current Debt Repayment Ability of Credit Card Holders:

- **Sub-question**: How is the current debt repayment ability of credit card holders, and how has it changed over time?
- **Data Cube**: CUBE_smart_cube_1
- **Method**: Use fields related to asset and contingent classification (e.g., asset_and_contingent_class_name) to analyze the current repayment status of credit card holders and compare it with previous periods (e.g., change_class_mom_level1) to observe trends in repayment ability.
- **Expected Outcome**: Understand the current repayment status of credit card holders and the trends in their repayment ability, which will help assess risk and determine an appropriate minimum payment rate.

Comparing Credit Card Interest Rates with Other Loan Types:

- **Sub-question**: How do current credit card interest rates compare with other types of loans?
- **Data Cube**: CUBE_smart_cube_1
- **Method**: Use fields related to loan interest rates (e.g., interest_rate) to compare credit card interest rates with other types of loans in the financial system.
- **Expected Outcome**: Understand the differences in credit card interest rates compared to other loan types, which will help evaluate the appropriateness of the minimum payment rate.

Analyzing the Impact of Unemployment Rates on Credit Card Repayment Ability:

- **Sub-question**: How does the unemployment rate affect the ability to repay credit card debt?
- **Data Cube**: CUBE_smart_cube_1
- **Method**: Use fields related to unemployment rates and asset classification (e.g., asset_and_contingent_class_name) to analyze the relationship between unemployment rates and the ability to repay credit card debt.
- **Expected Outcome**: Understand the impact of unemployment rates on the ability to repay credit card debt, which will help assess risk and determine an appropriate minimum payment rate.

These analyses will help us evaluate the appropriate minimum payment rate for credit card debt repayments by considering the repayment ability of debtors, economic conditions, and current interest rates.

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Leveraging Generative AI for Granular Credit Data Utilization: A Multi-Agent Approach

Nontawit Cheewaruangroj, Kawinwish Laobundit, Peranut Nimitsurachat, Supachai Saengthong, Skunpoj Thanarojsophon, Anak Yodpinyanee

Data Management and Analytics Department

Bank of Thailand

18-20 February 2025

Prepared for the 4th IFC workshop on "Data science in central banking", Rome, Italy

* The opinions expressed in this discussion paper are those of the authors and should not be attributed to the Bank of Thailand.

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Regulatory Data Transformation (RDT)

A transformative redesign from static Data Management System to granular-level regulatory reporting standard for credit data



Rich dataset with 1000+ fields and over 25 million loan accounts per month, including credit lines, interest rate plan, and outstanding amount



Multi-faceted data architecture with specialized cubes for diverse and thorough credit data analysis



Enable insights for analyst and examiners across many domains from bank supervision to monetary policies



Complex data requires knowledge and advanced data analytical skills

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Challenges with RDT Utilization



Problem translation

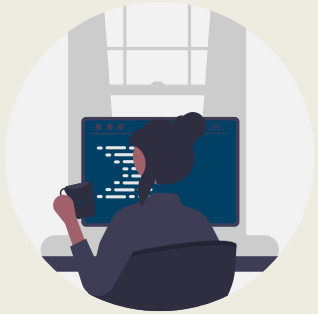
Users may not be able to translate business questions into data queries



Data understanding

RDT's data is complex

- Numerous data fields
- Structured into multiple "cubes" for different usages



Coding skills

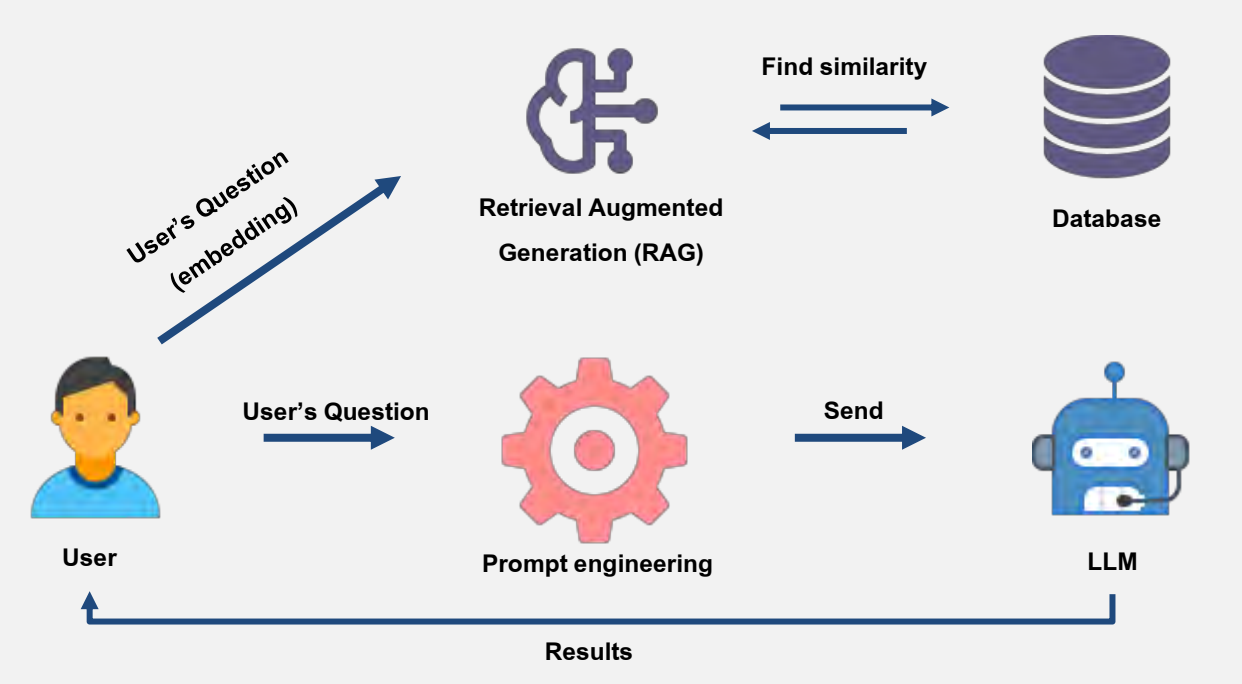
Require SQL / python knowledge to wrangle granular data effectively

- Most analysts and examiners are not proficient in coding

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Generative AI as RDT assisting agents

- **Large Language Model (LLM)** can be utilized as a copilot for answering users' questions.
- **Retreival-augmented Generation (RAG)** is used to provide internal knowledges such as metadata to GenAI .
- **Prompt Engineering** instructs GenAI to produce SQL code based on questions and retrieved metadata.



standard GenAI with RAG workflow

Ask a plain language question about RDT

Field selection

SQL query drafting

SQL

```
SELECT
  customers.name,
  SUM(loans.amount) AS total_loan_amount
FROM customers
INNER JOIN loans ON customers.customer_id = loans.customer_id
GROUP BY customers.name
ORDER BY total_loan_amount DESC
LIMIT 3;
```

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Challenges with GenAI workflow for RDT

	Challenges	Impacts
Complex operation	Standard workflow might struggle with multiple instructions for more complex tasks	Limit potentials for RDT utilization to basic querying
Context size	Retrieving metadata from all tables in RDT may exceed the context size of LLM	Unable to handle complex query that requires multiple tables or scale to larger data sources
Mixing LLMs	Cannot mix different models for different tasks to optimize cost	Consume more resources with running single large model for all tasks

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Multi-agent framework

Single-agent Framework

A large agent capable of executing diverse tasks comprehensively.



Analyst

Data analytics tasks
- Analyze questions
- Retrieve data
- Select fields
- Write SQL

Multi-agent Framework

A multi-agent approach decomposes the task into subtasks to be executed by different agents.



Manager
Orchestrate



Data Engineer
Retreive data



Analyst
Analyze data



Secretary
Communicate and assist user



Data Analyst
Write SQL



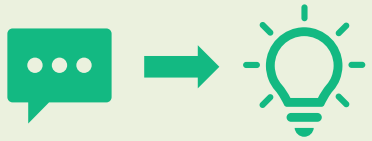
Senior Analyst
Analyze questions

Framework	Pros	Cons
Single-agent	• Easier to manage • Lower complexity to develop • Consume less resource	• Limit expertise for complex tasks • Limit on context size • Less flexibility • No steps provided to users
Multi-agent	• Able to deal with more complex tasks • Less limitation on context size • Assign different task to agents with specialities • Users can see broken-down steps	• More complexity • Consome more resource

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

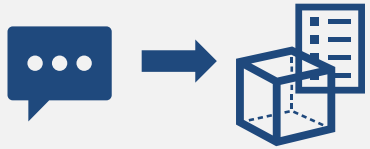
RDT Copilots: Modules

Brainstormer



Translate user's problem into tangible statements and identify relevant data sources

Metadata Copilot



Search for the best data cube and fields required to answer users' questions

SQL Coder



Assist users by converting natural language questions in to SQL query

Q: What should be the minimum payment rate for credit card debt repayment?

Brainstormer:

Find relationship between *credit_card_type*, *first_payment_amount* and *total_interest_and_fee_rate* to evaluate relationship between risks and payment.

…

Q: How many debtors have joined the debt restructuring program this year?

Metadata Copilot:

Field	Description
entity_id	Id for debtors
debt_restructuring_date	Date of debt restructuring
…	…

SQL Copilot:

```
SELECT
    COUNT(distinct entity_id) AS number_of_debtors
FROM CUBE_1
WHERE debt_restructuring_date > '2024-01-01'
```

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Brainstormer



Users

Information Gathering Agent

Select data cube and analyse the questions

Field Finder Agent

Select related fields from each chosen data cube

Question Breaker Agent

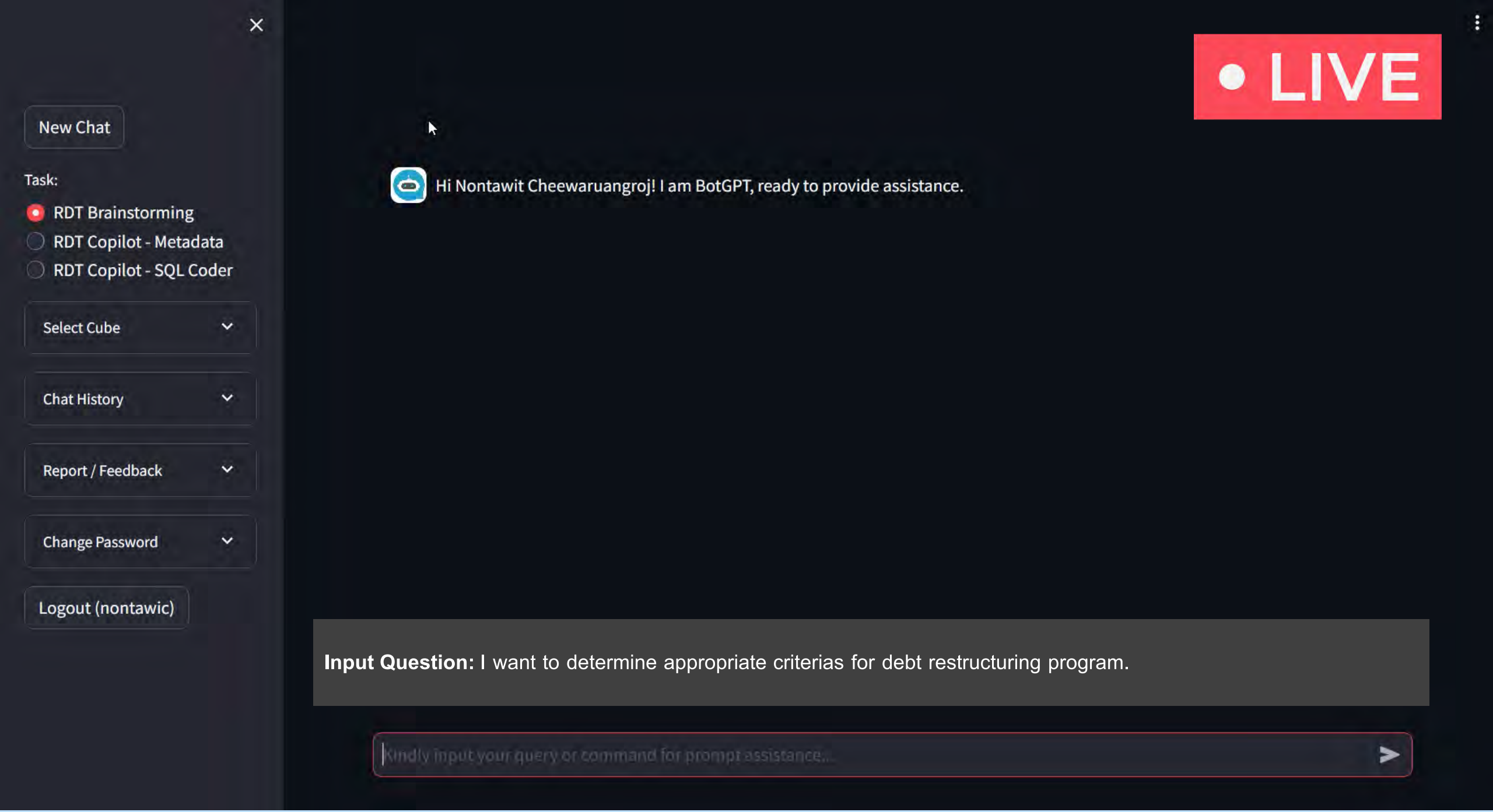
Analyze information and provide step-by-step answer to user's questions



Follow Up Agent

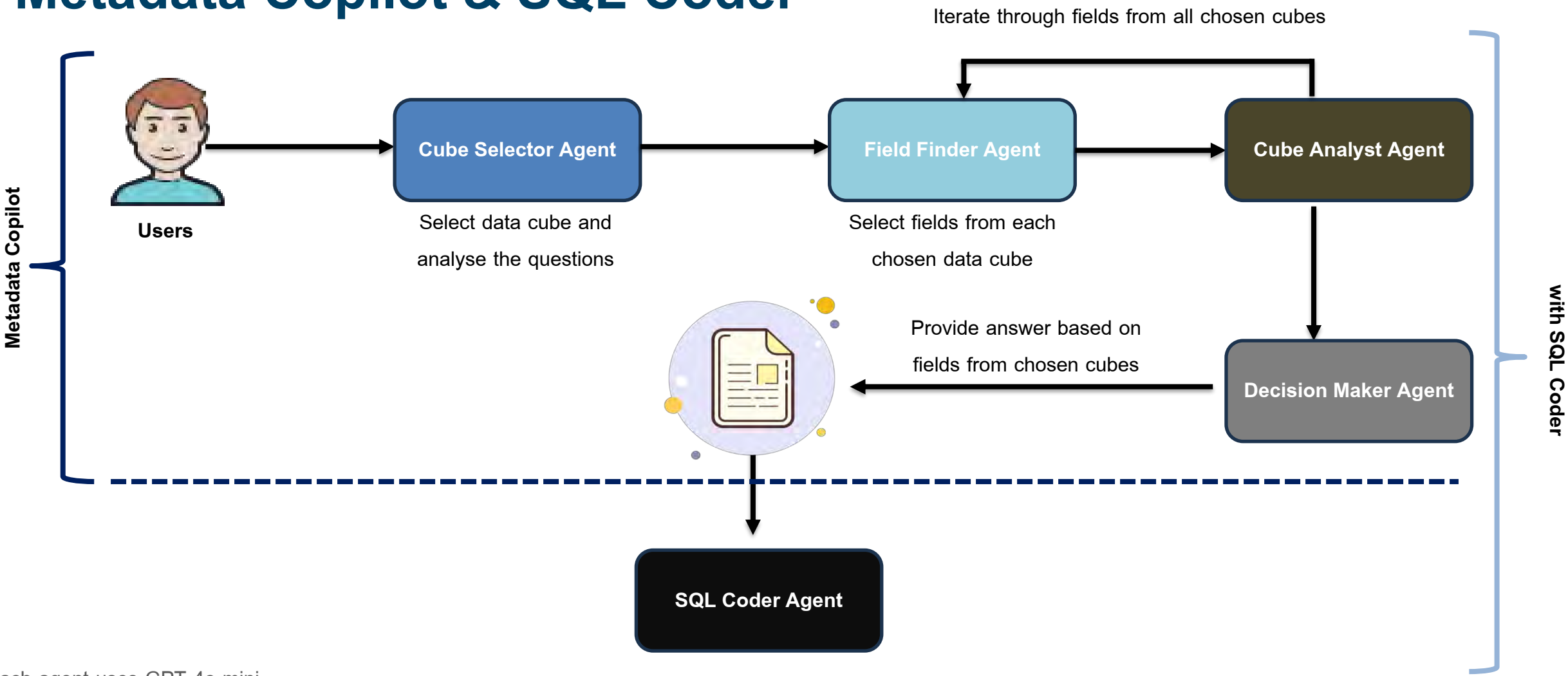
Ask follow-up question to guide the discussion forward

* Each agent uses GPT-4o mini

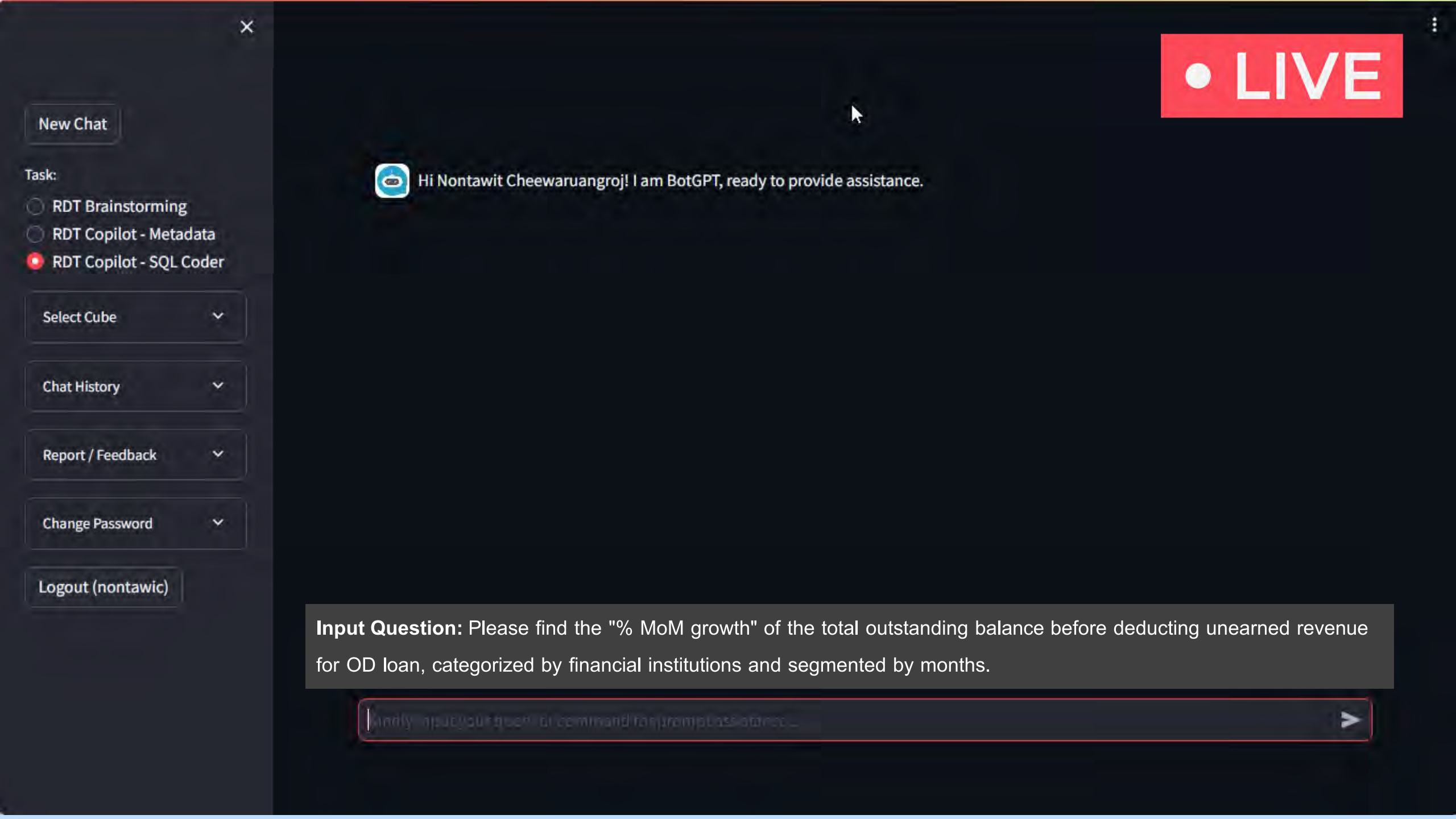

New Chat
Task:
RDT Brainstorming
RDT Copilot - Metadata
RDT Copilot - SQL Coder
Select Cube
Chat History
Report / Feedback
Change Password
Logout (nontawic)
LIVE
Hi Nontawit Cheewaruangroj! I am BotGPT, ready to provide assistance.
Input Question: I want to determine appropriate criterias for debt restructuring program.
Kindly input your query or command for prompt assistance...

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Metadata Copilot & SQL Coder



* Each agent uses GPT-4o mini


New Chat
Task:
RDT Brainstorming
RDT Copilot - Metadata
RDT Copilot - SQL Coder
Select Cube
Chat History
Report / Feedback
Change Password
Logout (nontawic)
LIVE
Hi Nontawit Cheewaruangroj! I am BotGPT, ready to provide assistance.
Input Question: Please find the "% MoM growth" of the total outstanding balance before deducting unearned revenue for OD loan, categorized by financial institutions and segmented by months.

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Evaluation

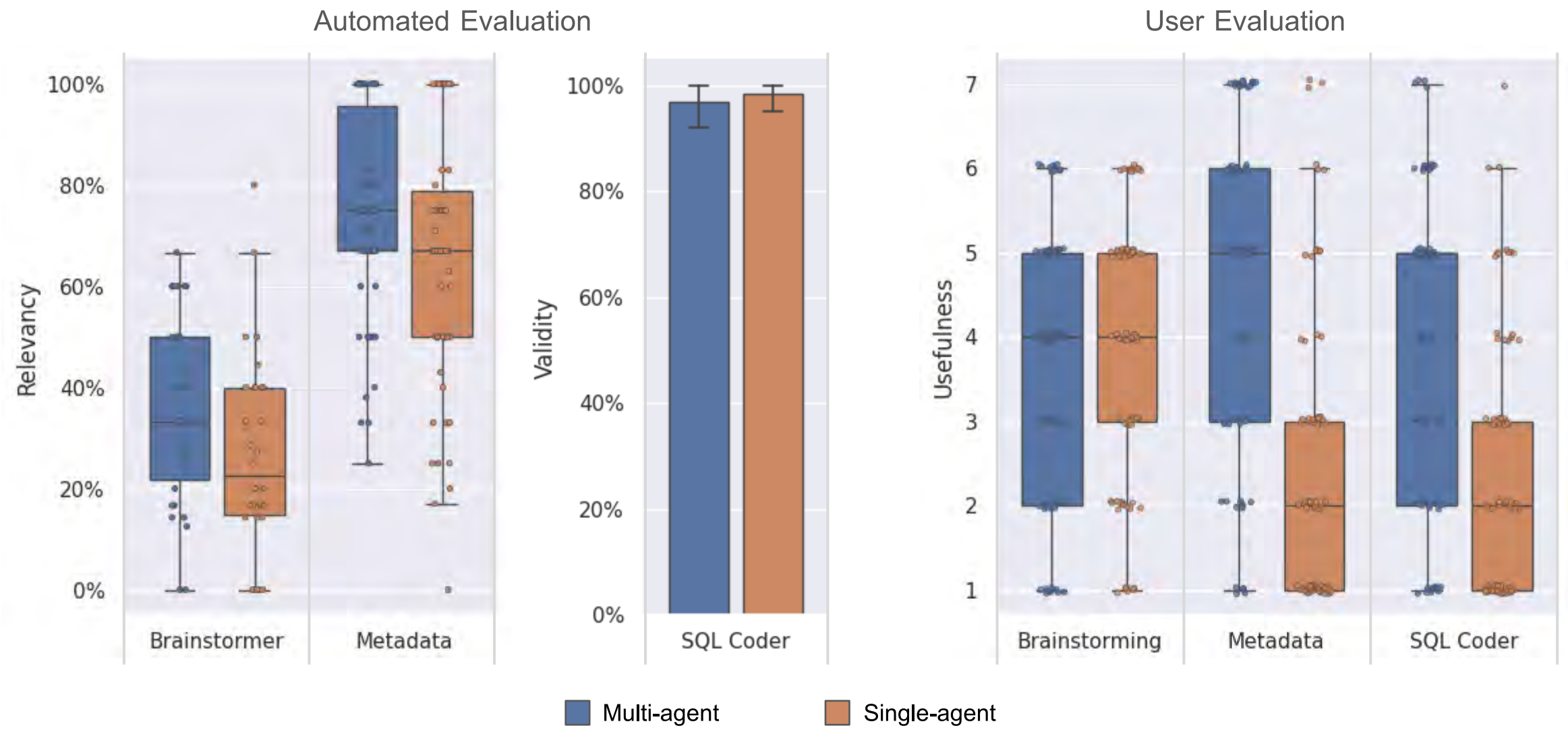
Experiment: Compare the performance of each module against the *single-agent* versions.

Metrics:

Module	Automated evaluation	Human evaluation
Brainstormer	**Relevancy** (Recall of expected fields required for business problem solving)	**Usefulness** (How useful the output for answering the question) - Likert scale of 1 to 7 - Evaluated by BOT's examiners & analysts
Metadata Copilot	**Relevancy** (Recall of expected fields required for querying data)	
SQL Coder	**Validity** (Whether the generated SQL syntax is valid, i.e., executable)	

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Results from evaluation



* For automated evaluation, Brainstormer is evaluated with 30 qualitative questions. Metadata Copilot and SQL coder are evaluated with 62 quantitative questions.
** For user evaluations, each module is evaluated with 10 questions by 8 analysts / examiners.

Discussion

Automated evaluation

- RDT Copilots can perform objective tasks, such as retrieving fields and writing SQL, very well.
- Multi-agent framework show good improvements in retrieving correct fields. For SQL generation, single-agent framework is already good.

User evaluation

- Users find multi-agent Metadata Copilot and SQL Coder useful for answering quantitative questions.
- However, users do not find Brainstormer particularly useful for qualitative questions. Multi-agents framework does not show improvements over single-agent framework.
- Users also show different preferences towards writing styles of Brainstormer versions. Results are varied for different questions for all modules.

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Conclusion

GenAI Tools can be useful in helping RDT data utilization.

- RDT Copilots can help analysts and examiners find and retrieve required data.

Multi-agent framework can be applied to enhance RDT Copilots:

- Help address issues of **limited context length** for multiple tables data.
- Improve **performance** over single-agent approach for some tasks, especially in more objective tasks that require large contexts.
- Can be adopted for **variety of tasks**, both objective (coding) and subjective (brainstorming).

ธนาคารแห่งประเทศไทย
BANK OF THAILAND

Challenges and Next Steps

Challenges

- Incomplete metadata. Internal jargons.
- Some domain knowledge not in documents or LLM knowledge.
- Difficult to get users engagement in designing and adoption.
- Difficult to experiment with different LLMs due to compliance & security concerns.

Further Work

- Improve data dictionary. Enhance prompt to incorporate more domain knowledge.
- Brainstormer needs a revision to produce more useful outputs.
- Explore LLMs with improved reasoning capabilities such as OpenAI o1/o3 model.
- Incorporate ability to directly process data, e.g., plotting graph.
- Explore multi-agent solutions for problems such as KM chatbots, supervision assistant agents.

Thank You

Contact: nontawic@bot.or.th

Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Closing remarks[1]

Paolo Angelini,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

BANCA D'ITALIA
EUROSISTEMA

Data Science in Central Banking

Closing remarks by Paolo Angelini,
Deputy Governor of the Bank of Italy

4th Irving Fisher Committee and Bank of Italy Workshop

Rome, 20 February 2025

Good afternoon ladies and gentlemen,

I am happy to be here with you today as we mark the conclusion of this workshop on data science in central banking. As highlighted in the Irving Fisher Committee Annual Report for 2024, launching this periodic workshop series jointly with the Bank of Italy, back in 2019, was a far-sighted strategic decision.[1]

The focus of this fourth edition of the workshop, Generative Artificial intelligence (AI) and its applications in central banking, aptly underlines the ongoing transformation in the field.[2]

There is broad consensus that Generative AI has the potential to significantly boost productivity worldwide. According to some estimates, it could yield annual productivity gains in the $2.6 - 4.4 trillion range globally.[3] Others predict an increase in output of the order of 15-20 percent, largely phased in over 15 years after adoption.[4] Survey evidence collected from global corporations suggests that for over 40 percent of respondents the return on investment in the most advanced Generative AI initiatives falls within the range of 11-30 percent, and meets expectations (Fig. 1).

---

1 The primary goal of these workshops is to showcase projects, share expertise among central banks, and reduce reliance on external service providers. The first workshop, in January 2019, focused on "Computing Platforms for Big Data and Machine Learning". The second focused on "Data Science in Central Banking" and was organized in two parts, dealing respectively with "Machine learning applications" (October 2021), and with "Applications and tools" (February 2022). The third workshop was also on "Data Science in Central Banking", with a focus on "Data sharing and data access" (October 2023).

2 An excellent primer on AI and central banks is in Bank for International Settlements, "Annual report", June 2024, pp. 91-127.

3 McKinsey, "The economic potential of generative AI", June 2023.

4 P. Aghion, S. Bunel, and X. Jaravel, "What AI Means for Growth and Jobs", Project Syndicate, January 14, 2025. I. Aldasoro, L. Gambacorta, A. Korinek, V. Shreeti and M. Stein, "Intelligent financial system: how AI is transforming finance", BIS Working Papers, no 1194, June 2024.

The uncertainty surrounding any estimate remains significant, largely due to the rapidly evolving nature of the phenomenon itself. Consider for instance the launch of the ChatGPT-like AI model "R1" by DeepSeek, a Chinese startup, at the end of January. The event cast doubt on the widespread belief that AI requires massive amounts of hardware and energy, raising questions about the leadership of the American tech industry and causing strong stock markets fluctuations.

The adoption rate of AI by firms is also uncertain, but growing rapidly. In early 2024, 65 percent of respondents in a survey of international corporations reported that their organizations regularly use Generative AI, nearly double the percentage observed in 2023.[5] Adoption rates vary across industries and appear to be higher among larger firms. Moreover, the speed of adoption within firms seems uneven: functions such as IT, operations, marketing, customer service, and cybersecurity appear to be at the forefront.[6] Additionally, perceptions of Generative AI differ markedly across corporate hierarchies, with top executives tending to have a more optimistic view of their organization's speed of adoption and implementation.[7]

Uncertainty has led some commentators to avoid making precise predictions about AI's future impact, instead considering two broad scenarios. In one scenario, the gradual adoption of Generative AI leads to incremental economic progress. In the other, more transformative scenario, AI fundamentally reshapes the world as we know it.[8] While the latter scenario may seem like science-fiction, some ongoing developments, such as the emergence of the so-called Agentic AI, suggest that it could in some form become a reality.

Regardless of the reliability of available estimates, there is no doubt that Generative AI is rapidly expanding across the business world and reaching consumers as well. Looking ahead, further growth seems inevitable, driven by increasing volume and diversity of data being generated and stored, the very fuel of AI, which Generative AI itself contributes to. The amount of data created worldwide each year is projected to continue its astonishing expansion, from 149 zettabytes (i.e., 10^{21} bytes) in 2024 to more than 394 zettabytes by 2028 (fig. 2).

As this event draws to a close, I would like to reflect on some of the key topics that emerged from our discussion and extrapolate a little to broader considerations.

Over the past three days, we have explored several use cases of AI and data science that are gradually permeating data analysis, real-time economic monitoring, and decision-making in central banking. One of the key takeaways from our discussions is that AI can enhance the speed, accuracy, and depth of forecasting and nowcasting (e.g., for inflation and output, or energy demand). We have also seen applications in regulatory compliance, financial

---

5 McKinsey, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value".

6 Deloitte, "Now decides next: Generating a new future Deloitte's State of Generative AI in the Enterprise", Quarter four report, January 2025.

7 Deloitte, op. cit.

8 M. S. Barr, "Artificial Intelligence: Hypothetical Scenarios for the Future", speech at the Council on Foreign Relations, New York, New York February 18, 2025.

supervision, and legal analysis. Large Language Models (LLMs), when combined with Retrieval-Augmented Generation (RAG) techniques[9] can extract insights from thousands of pages of regulatory filings, or legal documents, reducing the time devoted to document review and processing. Promising applications have also been showcased also in the field of financial stability monitoring. These findings confirm that AI applications are set to become ubiquitous in modern central banking, and beyond.

Several presentations also highlighted the risks associated with the introduction of AI. First, there is the issue of explainability in AI models. Many machine learning (ML) algorithms function as black boxes, making it difficult for users to understand the rationale behind specific results or recommendations. This is particularly concerning for applications in regulation and supervision, where transparency and accountability are crucial. Second, biases embedded in historical financial data may be reinforced by AI models, raising ethical concerns. Third, Generative AI is prone to hallucinations and misinformation. We have seen how LLMs can fabricate imaginary regulatory clauses, or misinterpret complex financial concepts; more generally, AI can generate incorrect answers that may be difficult to detect.[10] Fourth, the widespread application of Generative AI to investment decisions raises financial stability concerns, partly due to its potential to amplify herding behavior.[11] Fifth, while AI has useful applications in cybersecurity, fraud detection and anti-money laundering, it can also serve be a powerful tool for criminals and fraudsters. I am sure that many other issues could be added to this list.[12]

While these problems are significant, they have been clearly identified, and solutions are actively being developed. For instance, hallucination problems have been reduced in more recent releases of LLMs and can be further mitigated by grounding AI-generated responses in verified datasets through RAG models. Human oversight remains essential, but viable avenues for improvement are emerging.

But alongside these problems we also face other potential issues that are more difficult to identify and address. I am referring to the effects that the systematic use of these tools may have on the cognitive processes of individual users. For instance, as the technology becomes more reliable with each new release, there may be a growing tendency to reduce human oversight and accept the answers generated by AI engines as "the truth". While this could lead to efficiency gains, it could also result in errors and the spread of incorrect or distorted information and beliefs, not only in financial markets, but across all fields. Furthermore, as responses and outcomes that closely resemble the truth

---

9 RAG models enhance AI capabilities by adding information retrieval from specialized documents to the general purpose LLM text generation capabilities.

10 Among the respondents to the 2023 DigitalOcean survey on startups and SMEs trends, 34 percent mentioned security concerns, while 29 percent pointed to ethical and legal issues, including AI hallucinations and privacy risks.

11 Financial Stability Board, "The Financial Stability Implications of Artificial Intelligence", 2024; G7 Finance Ministers and Central Bank Governors under the 2024 Italian Presidency, "A High-Level Panel of Experts' Report to the G7".

12 For instance, F. Natoli and A. G. Gazzani, "The Macroeconomic Effects of AI-based Innovation", confirm that AI gradually raises total factor productivity, but also tends to increase income and wealth inequality.

become instantly available at virtually no cost, users may have less incentive to memorize information. These shifts are occurring in a world where the role of machines, already significant, is expected to expand rapidly (fig. 3).

Generative AI is just the latest addition to a long list of innovations that may significantly impact how the human brain functions (consider for example the pervasive use of mobile devices, the internet, or social media). A parallel phenomenon that comes to mind is remote working, another major trend reshaping how we work and interact. Both AI and remote work bring clear and tangible benefits, but they may also have potential side effects that are harder to identify and assess, as they may only materialize in the longer term. We should all be mindful of these risks.[13]

That said, I am not suggesting that we should delay AI adoption until uncertainty is reduced, or that its use should be resisted. However, low visibility calls for prudent driving. To ensure that AI is used in a robust, effective, fair, and accountable manner, adopters should focus on three key areas, which were echoed in various workshop presentations.

First, developing a sound AI governance framework should be a priority. At the macro level, as outlined in the High-Level Panel of Experts' Report to the G7, this requires "policy preparedness", based on alternative AI development scenarios.[14] At the firm level, Generative AI significantly expands the range of users, moving from a relatively small group of specialists to virtually the entire workforce. Unlike previous technologies, it has the potential to radically transform operational models and introduce new risks. This calls for a high level of responsibility and strategic attention.

The governance argument is particularly strong in the specific field of data management. Data is the fuel of AI, and no engine can function properly if fed with contaminated or low-quality fuel. Until now, proper data management has enabled efficient and effective search, retrieval and usage. Looking ahead, it will be crucial for extracting value from AI applications. Proprietary data are likely to be fundamental for the successful adoption of Small Language Models. These models hold great promise, as they can be easily tailored to a firms' specific needs while also mitigating the high costs and large energy consumption associated with LLMs. Examples relevant to central banks were provided during the workshop.

Second, the workforce should prepare for AI adoption. In the case of central banks, experts in economics and law have traditionally dominated the workforce. However, the digital revolution has gradually increased the pressure to reorient the talent mix

---

13 See e.g. I. Dergaa, "From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health", Frontiers in Psychology, April 2024; S. Ahmad, "Impact of artificial intelligence on human loss in decision making, laziness and safety in education", Nature - Humanities and Social Sciences Communications, June 2023; T. Ray, "The paradox of innovation and trust in Artificial Intelligence", Observer Reserve Foundation, February 2024. While these side effects, if indeed they exist, are quite general, we should ask ourselves if there is anything specific to finance, or central banking, which might need attention.

14 G7 Finance Ministers and Central Bank Governors under the 2024 Italian Presidency, "A High-Level Panel of Experts' Report to the G7". This approach emphasizes readiness and flexibility, ensuring that policies can be adjusted and refined as new evidence emerges and technology matures.

toward experts in IT, engineering, and mathematics. AI, with its transformative impact on work processes, will further intensify this pressure, necessitating adjustments in hiring policies. But in such a fast-changing environment, simply hiring specialists cannot be the main solution. Training, reskilling and upskilling the workforce will be essential. Proper attention to these aspects will not only help organizations fully leverage new technology, but also address employees' resistance to change, often a major obstacle to successful adoption.

Third, cross-institutional cooperation among authorities (not just central banks and bank supervisors) should be encouraged, given the the increasingly blurred regulatory boundaries in data-intensive technologies such as AI.[15] At the start of a new journey, it makes sense to learn from each other's mistakes and share best practices rather than rely on a trial-and-error approach in isolation. In today's challenging global environment, multilateralism is under strain, and international cooperation is facing strong headwinds. Given its long tradition of cooperation in international fora such as the BIS and the FSB, the central banking community is well-positioned to foster cooperation in this field.[16] Workshops like this one embody that spirit.

A clear area for targeted cooperation is cross-border data sharing. Many of you are familiar with the difficulties of enabling data access, both nationally and internationally, due to the complex intersections between personal data protection, national security, intellectual property, and sectoral regulation.[17] In a sense, AI will raise the opportunity cost of maintaining exclusive control over information, calling for smart ways to reconcile public policy objectives with data-drive innovation.[18]

Concluding on this final note, I encourage all of you to stay in touch, actively collaborate, and continue the conversations initiated during this workshop. These discussions can serve as the foundation for future partnerships and joint initiatives.

I would like to extend my sincere gratitude to the speakers and organizers whose dedication and hard work have made this workshop possible. I wish you all continued success in your endeavors.

---

15 P. Hernández de Cos, "Managing AI in banking: are we ready to cooperate?" Keynote speech at the Institute of International Finance Global Outlook Forum Washington DC, 17 April 2024.

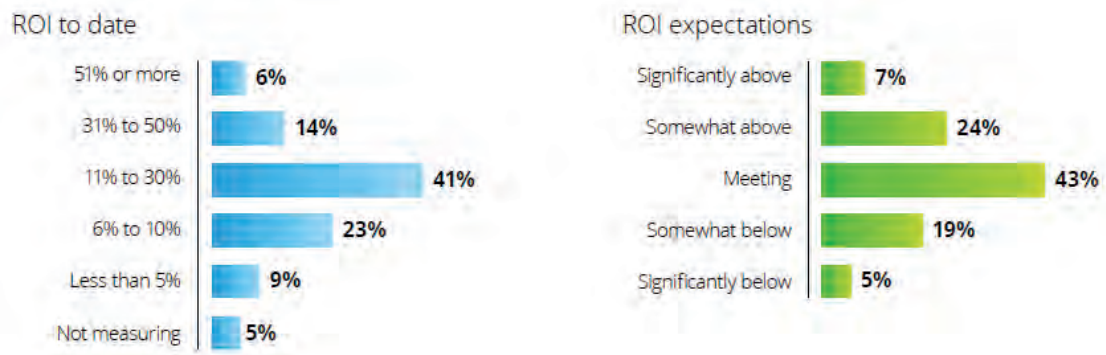
16 O. Borgogno and A. Perrazzelli, "From principles to practice: The case for coordinated international LLMs supervision", Cambridge Forum on AI: Law and Governance. Vol. 1, e13, 2025.

17 O. Borgogno and M.S. Zangrandi, "Data governance: a tale of three subjects" (2022) Journal of Law, Market & Innovation, 1(2), 50-75.

18 The EU has been working on various data sharing measures that have the potential to enhance AI model training and development. European Commission, "Working Document on Common European Data Spaces", 24 January 2024 SWD(2024) 21 final, designs a plan to develop Common European Data Spaces, particularly in finance, energy, and health. Similarly, regulatory sandboxes enshrined in the EU AI Act could provide a valuable opportunity to facilitate the responsible use of personal data collected for other purposes. See Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), Art. 59, Rec. 140.

Figure 1

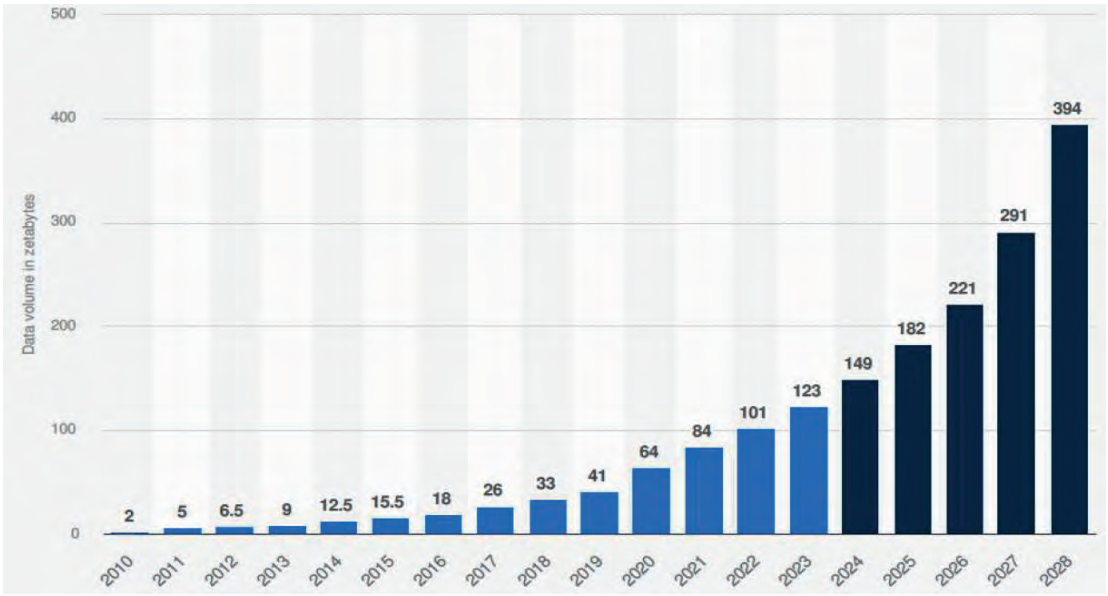
Distribution of the estimated return on investment (ROI) for the most advanced scaled Generative AI initiatives in a sample of international corporates (1)



Source: Deloitte, "Now decides next: Generating a new future. Deloitte's State of Generative AI in the Enterprise", Quarter four report, January 2025.
The panels portray the frequency distribution of ROI across surveyed firms. For instance, in the left-hand side panel, from the bottom, 5 percent of firms report that they did not attempt to measure the return on their investment, 9 percent estimate a ROI to be less than 5 percent, and so on.

Figure 2

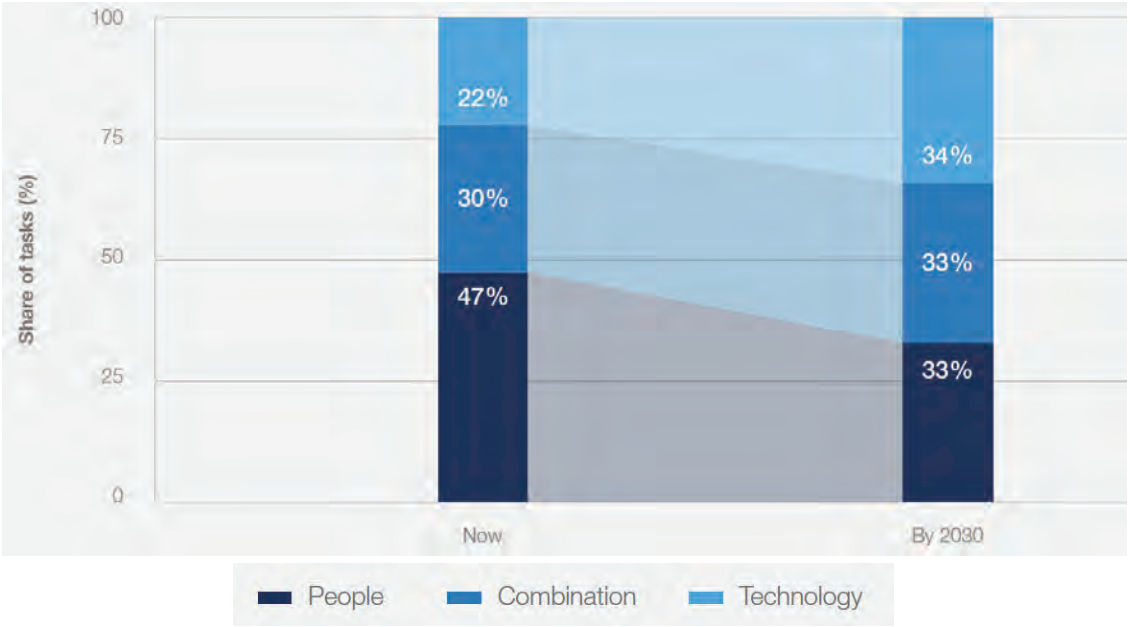
Volume of data/information created, captured, copied, and consumed worldwide each year from 2010 to 2023, with forecasts from 2024 to 2028



Source: IDC; Statista estimates.
© Statista 2024.

Figure 3

Share of total work tasks expected to be delivered predominantly by human workers, by technology (machines and algorithms), or by a combination of both



Source: World Economic Forum, "Future of Jobs Report", January 2025.

Designed by the Printing and Publishing Division of the Bank of Italy