Irving Fisher Committee on
Central Bank Statistics

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

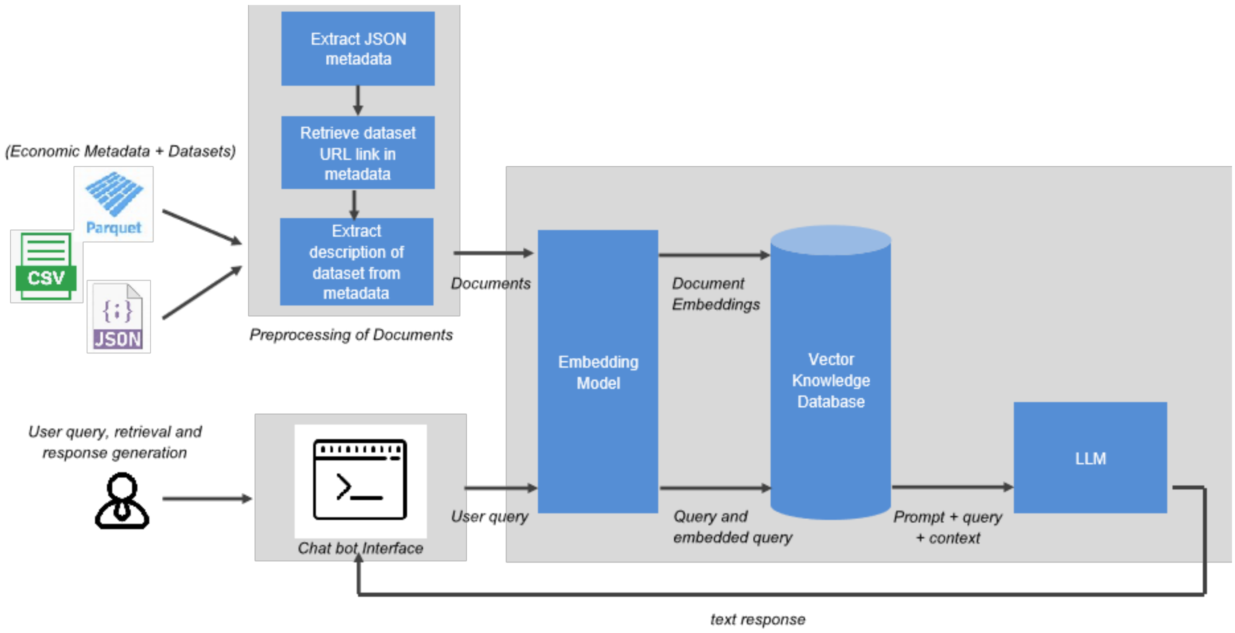
Comparison of FAISS, Chroma, and PineCone — Table 1

	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)

<u>User Query:</u>

"*Explain the observed trends in annual GDP growth for Malaysia, Singapore, Japan, and Australia from 2005 to 2023.*"

<u>Chatbot Response:</u>

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.

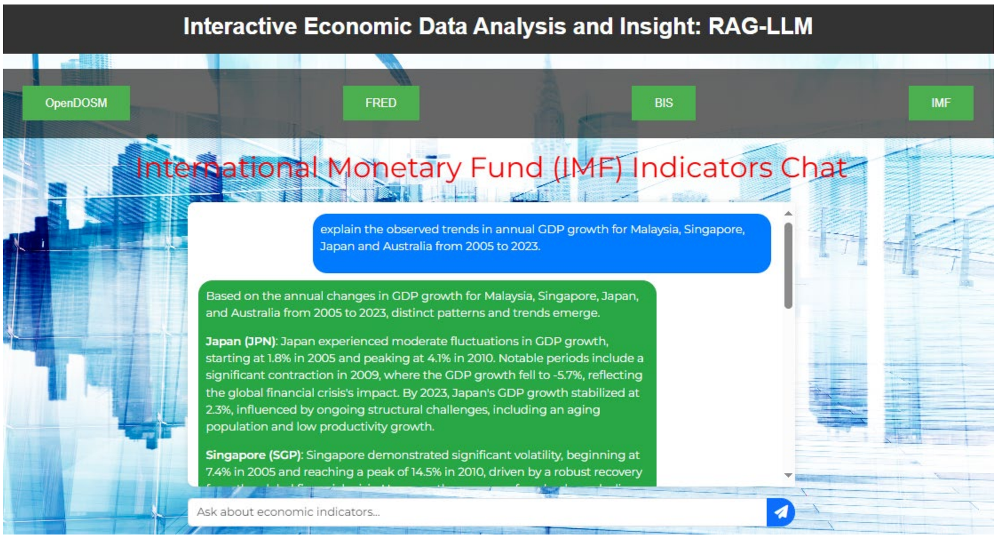
Malaysia (MYS): Malaysia showed a pattern of moderate growth, starting at 5.0% in 2005 and peaking at 7.5% in 2010. The country faced challenges in 2020, with GDP growth declining to -5.6% amid the pandemic. By 2023, Malaysia's GDP growth rebounded to 5.5%, supported by a recovery in domestic demand 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.

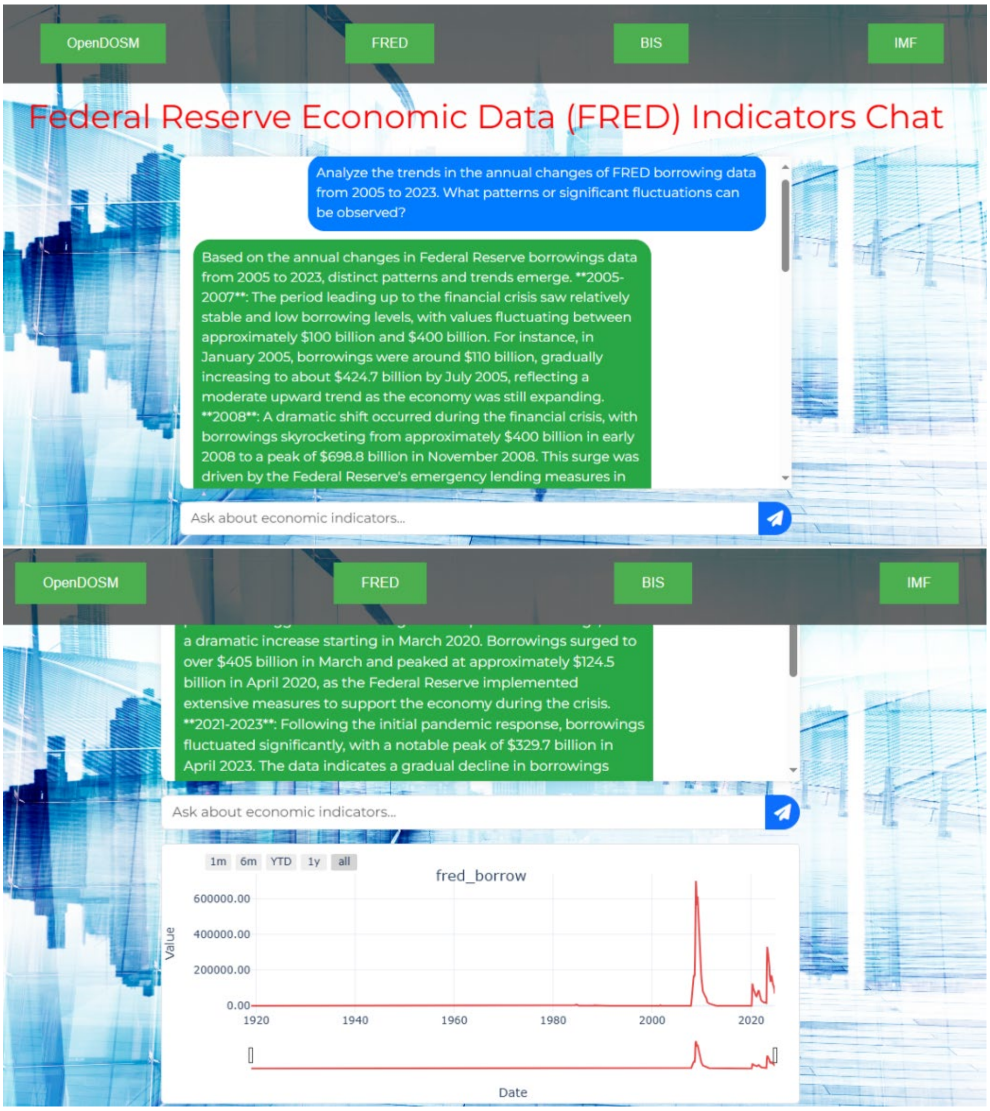
2009-2015: Following the peak in 2008, borrowings remained elevated but began to decline gradually. By the end of 2009, borrowings had decreased to about $142 billion, reflecting the Fed's efforts to stabilize the economy. Throughout this period, borrowings fluctuated but generally trended downward, 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:

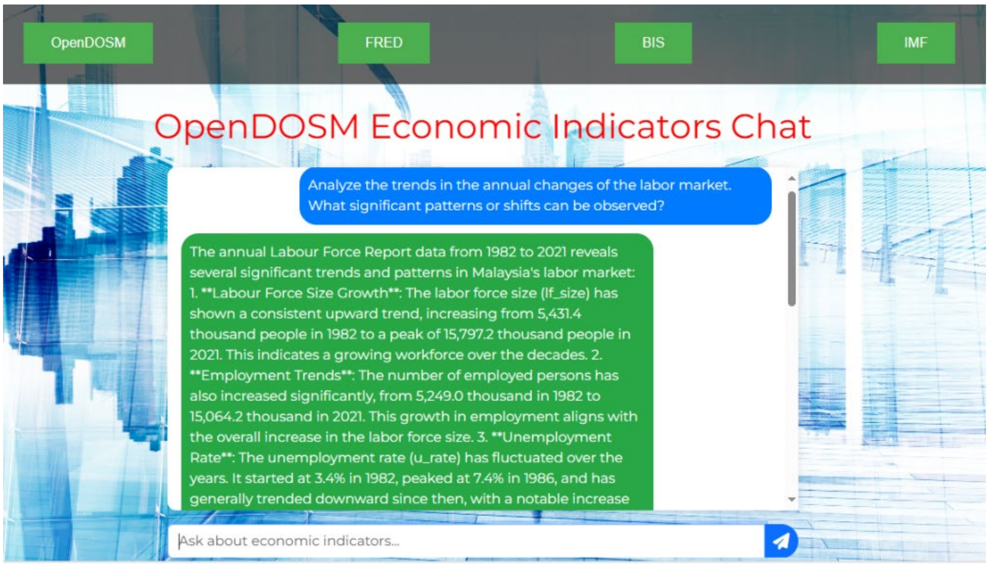
The annual Labour Force Report data from 1982 to 2021 reveals several significant trends and patterns in Malaysia's labour market:

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).

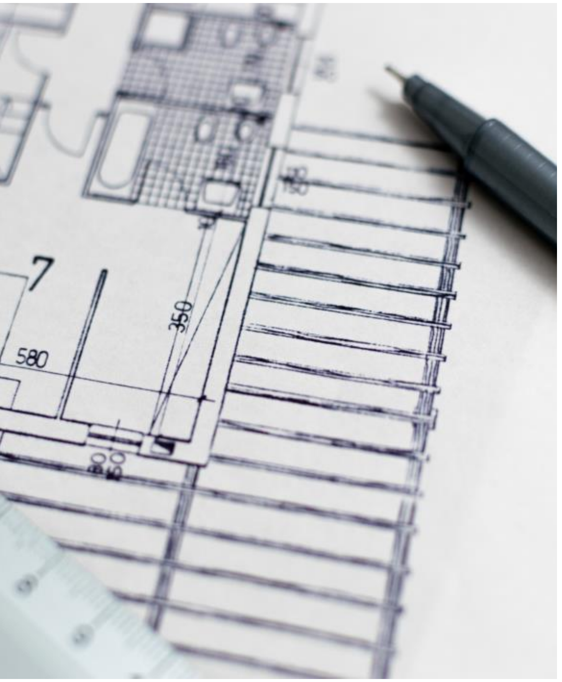
Sohangir, S., Wang, D., & Pomeranets, A. "Big Data and Machine Learning in Finance". *Journal of Financial Data Science* (2018).

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

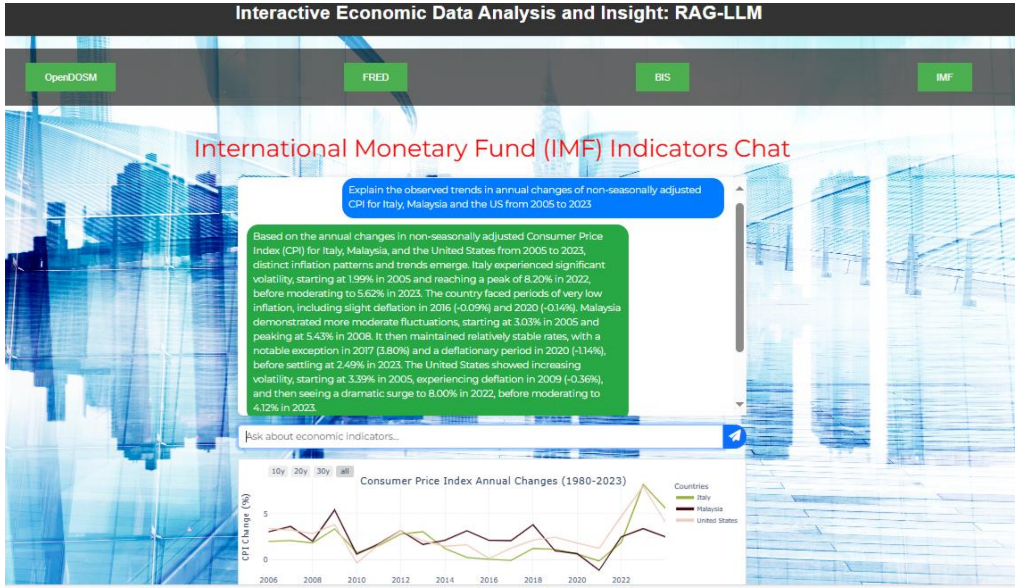
- **Problem Statement:** Users often struggle with fragmented and complex economic datasets, requiring specialised knowledge to interpret.
- **Solution:** Our chatbot provides an intuitive interface that simplifies data access and analysis, making it easier for users to gain insights.

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.

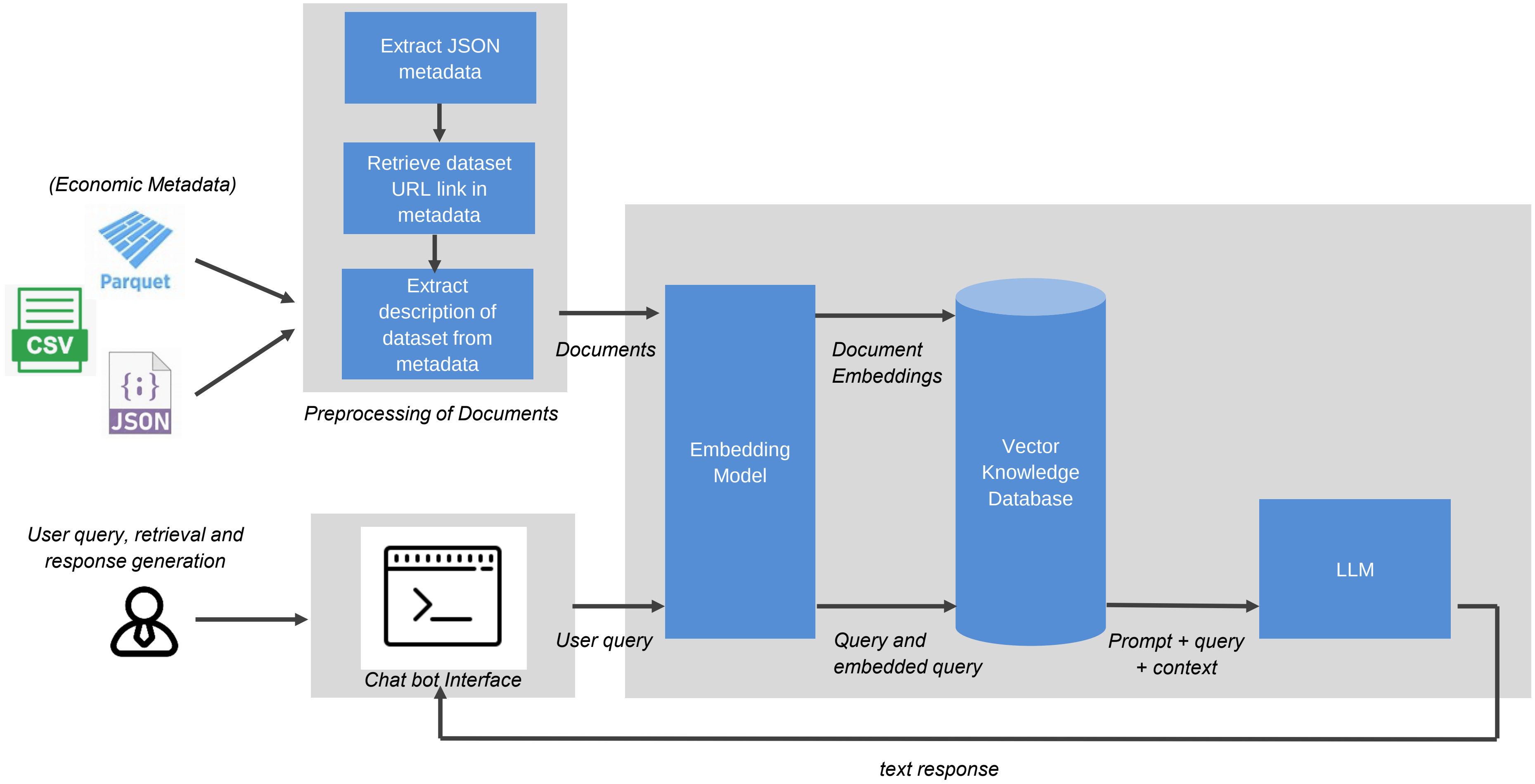
2. Data Source Integration

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.

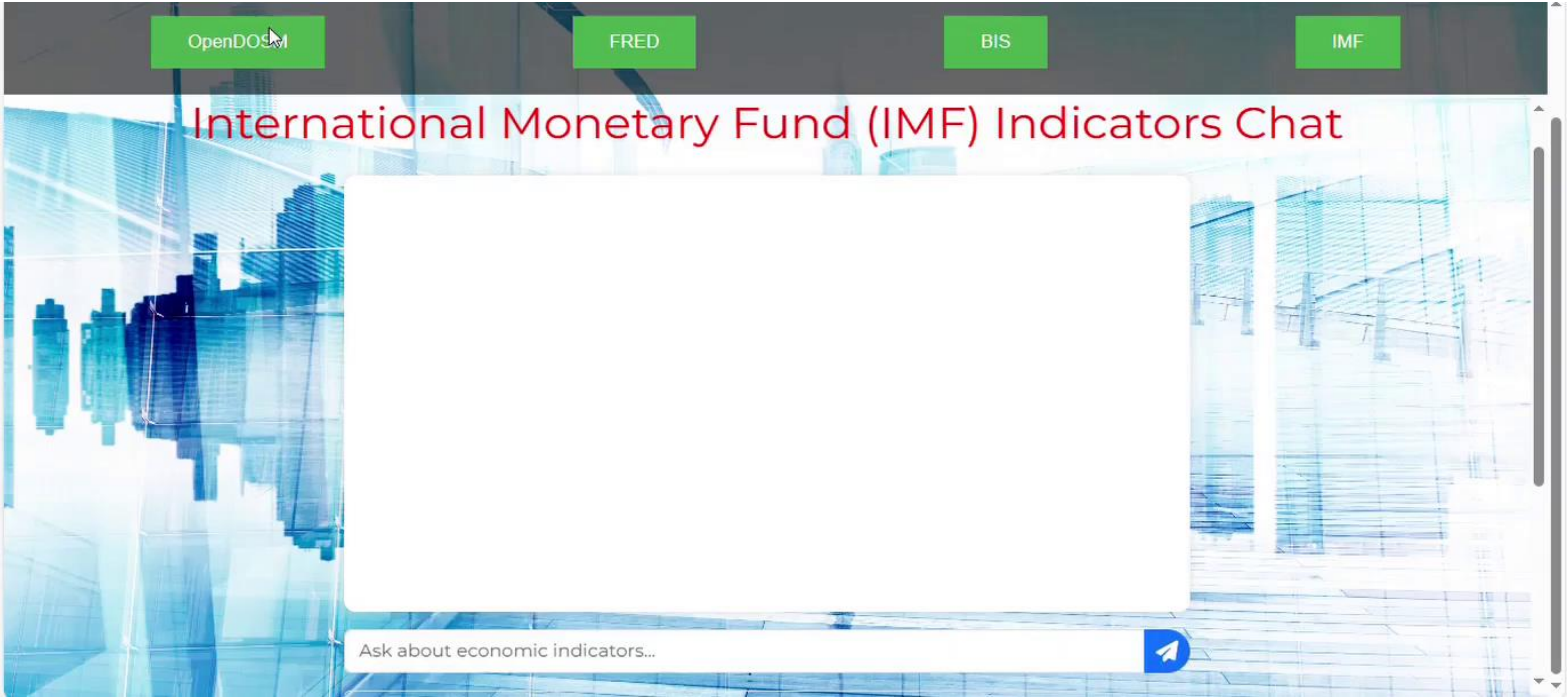
Reason for FAISS selection: **Performance and Cost Efficiency**

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

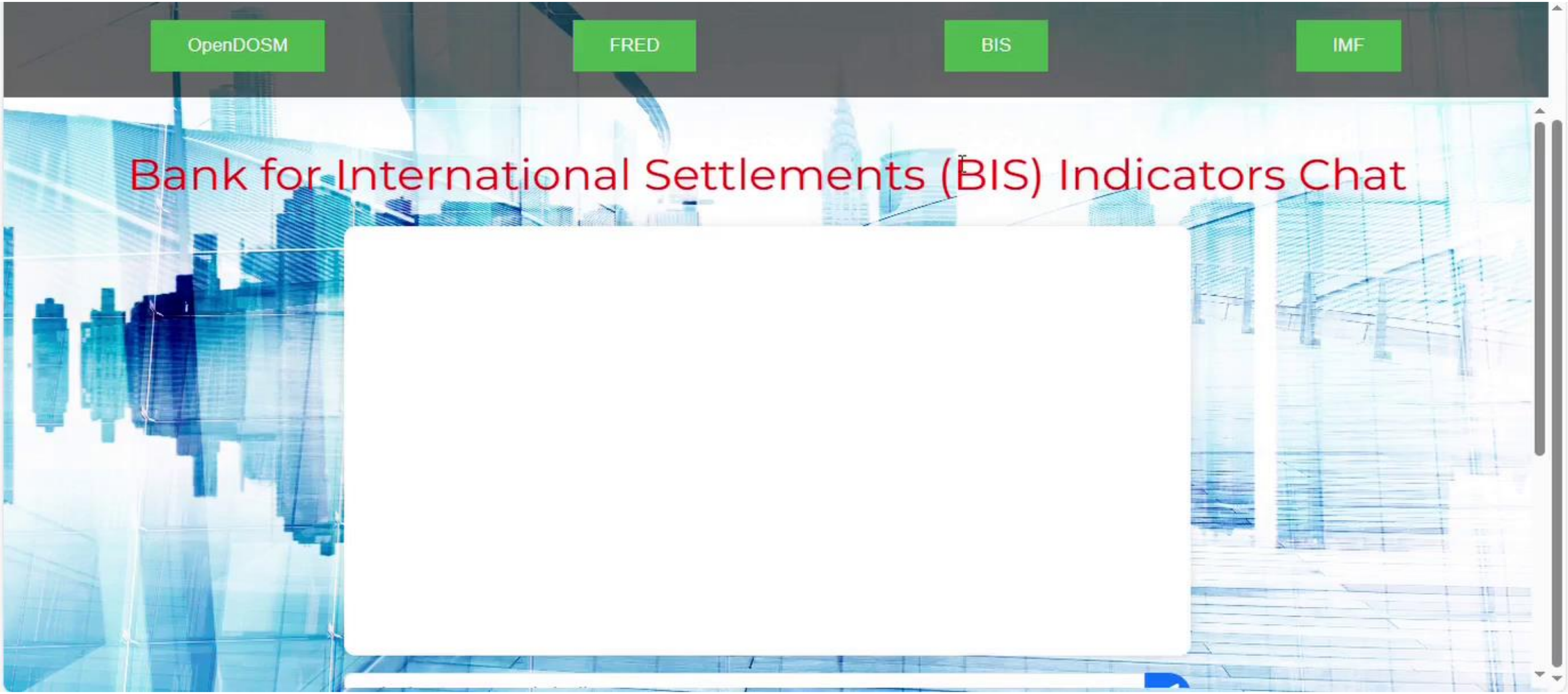
Usage

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