IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Leveraging large language models to extract data citations[1]

Sebastian Seltmann, Emily Kormanyos, Hendrik Christian Doll,
Deutsche Bundesbank

[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.

Leveraging Large Language Models to Extract Data Citations

Sebastian Seltmann[1][2], Emily Kormanyos[1], and Hendrik Christian Doll[1][3][4]

Abstract

Empirical researchers and data providing institutions currently face challenges in efficiently understanding data usage in scholarly papers. Tracing data usage in papers relies on human readers and remains time-consuming and error-prone. To address this, we explore the potential of using Large Language Models (LLMs), specifically GPT-3.5, to automate the identification and categorization of research data sources. By employing web scraping, we collect a comprehensive sample of research papers and create a human-labeled validation dataset. We analyze the accuracy of GPT-3.5 in detecting and summarizing data sources in economics and finance papers. We evaluate the detection and prediction accuracy and address the issue of false answers provided by the model. We find that LLMs can advance considerably on the status quo. Results are encouraging in terms of quality of extraction and coverage. Our paper provides a detailed description of the pipeline to implement our solution at data providing institutions, enabling the understanding of data impact and therefore efficient data provision services.

Keywords: Data citations, Automated data source classification, Large Language Models (LLMs), Natural Language Processing (NLP), Research data centers (RDCs).

JEL classification: C81, C82, C88, C87.

[1] Deutsche Bundesbank, Data Service Centre

[2] Contact: sebastian.seltmann@bundesbank.de (corresponding author)

[3] All views expressed in this report are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

[4] We wish to express our gratitude for Shir Frank and Kilian Graef for their excellent student research assistantship. Without their support, this project would not have been possible.

Contents

1. Introduction

Empirical researchers use data to conduct analyses linking theory to societal questions. To conduct such analyses, generate new research ideas, and navigate available commercial and open-source data sources, researchers need to understand which data sources academic papers analyze to arrive at their conclusions.

While scholarly citations are highly standardized and easy to use in machine-readable workflows, data citations often come in a variety of formats to date. The unstructured nature of data citations has complicated their automated extraction for the longest time. If data citations could be extracted automatically, this would give rise to a variety of valuable applications measuring data impact.

Enabled applications could include data usage networks, where extracted data citations could help researchers trace data usage across papers and domains. Extracted data citations could fuel applications, such as dataset recommender systems. Additionally, knowing where data is used enables data providing institutions to estimate of the value of dataset provision e.g. in terms of citations.

In this paper, we explore the potential of Large Language Models (LLMs) to identify mentions of data sources in research papers. Specifically, we analyze to what extent a widely adopted LLM, GPT-3.5, can accurately detect, describe, and summarize all specific mentions of data sources in research papers in the fields of economics and finance. We describe a proof-of-concept for a pipeline to extract data citations into structured lists.

To this end, we collect a large sample of research papers using web scraping techniques, thereby ensuring a sufficiently comprehensive and unbiased data coverage. We randomly sample and construct a human-labeled validation dataset to compute performance metrics of detection accuracy, i.e., to what extend the LLM is able to (i) detect all mentioned sources and (ii) whether it describes them accurately.

We find that we can advance extraction considerably compared to the status quo. Results are encouraging in terms of recall. In addition to gauging the ability of LLMs to extract data usage in research papers, we provide a guide targeted at implementing our solution at data providing institutions. By data providing institutions in this context, we refer to providers of official and commercial data, as well as research institutions.

Two recent trends enable us to take this step. First, recent technological progress in natural language processing (NLP) provides an opportunity to build a performant AI-driven reading system targeted at detecting, classifying, and connecting data source descriptions in research papers. Specifically, the increase in capabilities of large language models (LLMs), such as generative pre-trained transformer models, as popularized by ChatGPT, has been remarkable for our use-case.

Second, empirical research in economics increasingly utilizes granular data from administrative sources (e.g. Einav and Levin, 2014). Such datasets are often standardized, quality-controlled, and constitute known entities with fixed names, facilitating extraction. There is a trend towards providing this administrative data through dedicated research data centers (RDCs, Card et al., 2010).

RDCs have an interest in tracing their data, since if they can show the impact of the data they provide, this can motivate the use of public funds for data provision. Furthermore, the data providers can then target data provision efforts to impact-generating datasets. With these use-cases in mind, naturally extraction efforts have been numerous to date. In the next section, we outline the current state-of-the art in information extraction from research papers and our contribution.

2. Contribution to the status quo

2.1 Achievements to date

A large and interdisciplinary stream of prior literature has considered the task of automated extraction of dataset mentions from free texts, with or without the aid of NLP techniques. Since the widespread availability and use of LLMs is a relatively recent phenomenon, however, there is no specific prior research on the feasibility and usefulness of LLMs to this task at the time of writing and to the best of our knowledge.

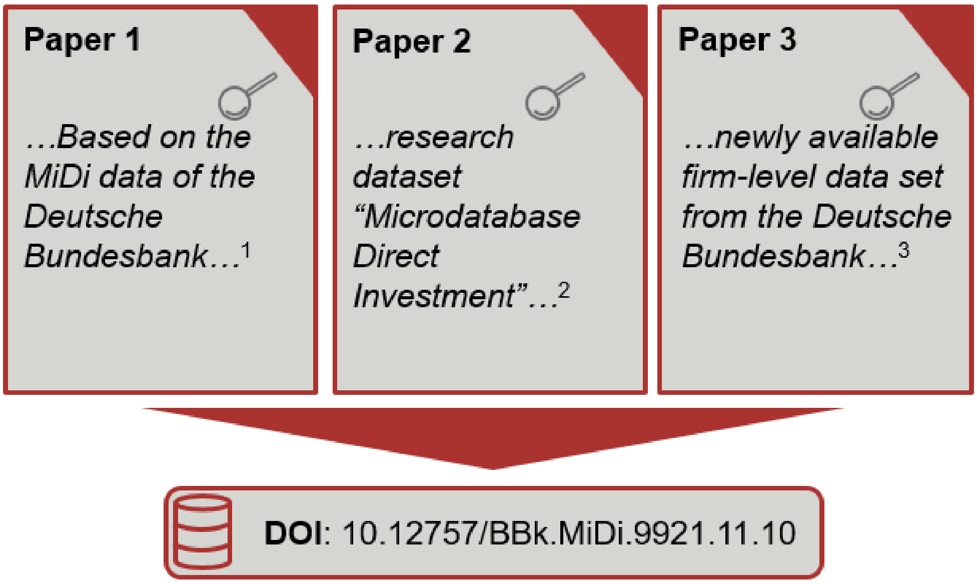

Paper 1
…Based on the MiDi data of the Deutsche Bundesbank…1
Paper 2
…research dataset "Microdatabase Direct Investment"…2
Paper 3
…newly available firm-level data set from the Deutsche Bundesbank…3
DOI: 10.12757/BBk.MiDi.9921.11.10

Figure 1: Examples for diverse citations of the same dataset. The dataset can be identified by the digital object identifier (DOI), however the DOI is often not present or cited. Sources: Authors' own depiction based on actual data citations in 1 Lipponer (2006), 2 Blank et al. (2020), and 3 Buch et al. (2005)

While one can achieve the desired outcome of automated information extraction from research papers using either (i) rule-based look-up tables, (ii) predecessor NLP models or (iii) LLMs, the use of the latter offers a hitherto unparalleled quality available to most users. Since there is a large number of a priori unknown ways to cite the same dataset, look-up tables and simple models tend to have a low recall trying to identify used datasets (Figure 1 gives an example for the complexity of the extraction task with three actual ways a dataset is cited, exemplary for numerous further possibilities).

Data citations do not necessarily have to be in textual, unstructured form. Concepts for unique and persistent identification have been introduced in the literature, for example the Digital Object Identifier (DOI). The DOI can identify any physical or digital object and is used for dataset citations as well. However, DOI usage for data citations (at least in economics and finance) remains piecemeal, therefore text extraction using Natural Language Processing (NLP) methods remain to be the method of choice for the task at hand.

LLMs offer potential of democratizing NLP applications for the wider public and academia without requiring extensive prior knowledge of computer programming and machine learning (ML). Therefore, the arrival of these models, exemplified here by the LLM GPT-3.5, warrants a new inspection of the issue of extracting dataset and methodology information from research papers. Importantly, we aim to show how providers of official statistics can leverage on LLMs with a practical application.

Prior to the popularization of LLMs, several examples from the previous literature have shown efforts for automated information extraction from research papers. The extracted entities covered by the prior literature include citations (e.g. Saier and Färber, 2020, as one recent example), papers' focal points and topic trends (e.g., Chen and Luo, 2019), or tasks and methods (e.g., Houngbo and Mercer, 2012, Kovačević et al., 2012, Eckle-Kohler et al., 2013, Luan, 2018, Jayaram and Sangeeta, 2017, or Yao et al., 2019).

Most relevant to our work are instances studying specifically the automated extraction of analyzed datasets. Such examples include the work of Boland et al. (2012) and Ghavimi et al. (2016). In contrast to the majority of earlier work on the subject, the focus of the literature has recently shifted to the use of NLP and ML techniques to identify and extract entities such as datasets, methods, or results from research papers (e.g., Kardas et al., 2020, Kumar et al., 2021, or Gemelli et al., 2023).

The sub-stream of the prior literature focusing specifically on dataset mentions has itself given rise to the construction of open-access databases for research data usage and networks.[5] Wang et al. (2022) provide a recent systematic review of the literature on extracting and evaluating scientific methods used to gain insights from the ubiquitous data sources, which characterize current empirical research.

Importantly, however, the majority of earlier work on the subject share a potential issue: The authors generally focus on research papers from the data mining, knowledge discovery, data science, computer science, ML, or bibliometric disciplines. Since papers from these disciplines might be more likely to contain particularly well written data and methodology sections compared to other disciplines, the performance gains of the devised tools and methods might conceivably be biased upward.

Closely related to our focus on dataset usages, Kumar et al. (2021) study the extraction of dataset mentions in the scientific literature for reusing and replicating used datasets and ultimately building a dataset recommender system. Such a system

[5] See Otto et al. (2020) for a discussion of several such databases.

can help researchers adequately identify the datasets best suited to answer a given research question. Ikeda et al. (2020) aim to provide a similar pathway to this goal, taking care to study a large and multidisciplinary corpus of papers, both before the widespread availability of LLMs.

Examples from the prior literature, which use LLMs to extract data citations are scarce. Compared to traditional NLP and textual analysis methods, however, LLMs pose the unique potential to facilitate applying state-of-the-art NLP algorithms by removing all barriers to entry except for the users' ability to use application programming interfaces (APIs). Therefore, such models also hold the potential of dramatically simplify complicated information extraction pipelines compared to the more complicated approach of using NLP techniques 'by hand'.

Perhaps closest to our paper is the work of Polak and Morgan (2023), who propose ChatExtract, an automated data extraction method using conversational LLMs. It requires minimal initial effort to extract data from research papers by employing engineered prompts and follow-up questions similar to the approach used in this paper. The authors demonstrate that their tool is able to provide answers with around 90% precision and recall rates on materials data, suggesting that ChatExtract is a simple and transferable tool for data extraction across various fields.

2.2 Our value proposition

Importantly, however, Polak and Morgan (2023) focus on materials data and research papers, and they demonstrate the usefulness of their tool based on known properties. By contrast, we aim to provide a solution for users who do not know the sought properties a priori, i.e., researchers or RDCs for whom reading papers front to back in order to identify data citations is not feasible. Therefore, we go beyond ChatExtract for a wider task in question (identifying unknown datasets from previously unseen research papers).

Where most of the prior literature focuses on the natural sciences or ML/NLP papers specifically, our focus is on economics and finance as two large bodies of academic literature. These academic fields tend to lean increasingly on standardized datasets either licensed from commercial providers, provided by RDCs in public institutions, or owned and disseminated by single researchers within their professional networks. Often times for any given paper, joint data from multiple sources is used to provide novel insights.

This data availability structure implies that the automated identification of datasets in these bodies of the scientific literature is crucial to understand not only the impact of single datasets, but also how they are combined with other data sources. Importantly, these applications extend beyond the work of RDCs to data providers, researchers in search for adequate datasets to answer their research questions, and the literature on academic networks, to only name a few.

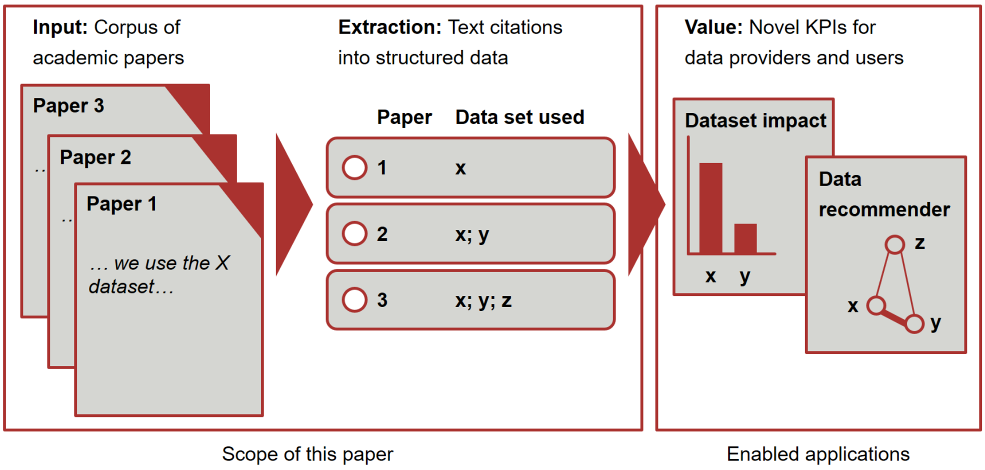


Figure 2: Scope of this paper and value proposition. Source: Authors' own depiction.

With these applications in mind, we aim to inform the informational basis. Specifically, we aim to fuel two use-cases by generating structured lists of data cited per paper. These two use-cases are (i) to be able to recommend data to researchers ("based on your interest, you might also like this data") and (ii) data impact factors ("This dataset justifies investment by generating high research or policy value"). We outline these value propositions in Figure 2**.** In the next section, we outline the data we use to extract the necessary data to inform these use-cases.

3. Data

3.1 Sample collection

In order to provide a feasibility study for Central Banks, official statistics offices and academia to obtain structured citations data from textual data mentions in papers, we use data from the Discussion Paper series of Deutsche Bundesbank. This series comprises papers written by researchers including one or more employees of the Bundesbank.

Sample

Corpus of research papers for automated extraction of data citations and evaluation sample Table 1

	Source	Sampling	Number of papers	Sample period	Labeling
Corpus to be labeled	Bundesbank discussion paper series	Full universe	1,102	02/1995 – 07/2023	Using GPT 3.5
Evaluation sample	Bundesbank discussion paper series	Random sample	103	06/1995 – 04/2023	Manually

Source: Authors

Our input sample for the dataset extraction pipeline comprises all research papers released under the Discussion Paper series of Deutsche Bundesbank between February 1995 and July 2023. The Deutsche Bundesbank discussion paper series covers a variety of different topics, such as monetary policy, banking supervision, and household economic behavior. In total, this corpus includes 1,102 research papers at the time of our data collection, published between 1995 and 2023. This constitutes the corpus to be labeled. Subsequently, it is straightforward to extend our approach to label any corpus of academic papers.

From the corpus to be labeled, we randomly select 103 discussion papers as an evaluation sample. In order to obtain a baseline to evaluate the performance of our approach, we label these papers manually. By labeling in this context we refer to reading a paper and manually noting the datasets used (if any). These manually labeled papers constitute our evaluation sample for the dataset extraction pipeline. This means that the performance evaluation metrics introduced and discussed in Sections 4 and 5 are based on estimated averages across this evaluation sample.

3.2 Manual labeling of evaluation sample

For manual labeling, we randomly assign 20 or 21 papers each to five human readers. Human readers scan the papers for specific dataset mentions. Figure 1 illustrates that dataset mentions differ widely, lacking a standardized format and standardized sections within papers. Therefore, the human readers need to read all papers thoroughly to find dataset citations.

Labeling guide

Step-by-step instructions used for human labeling of evaluation sample Table 2

Step		Instruction
1	Open the folder with the list of pdf assigned to name	
2	For each pdf	
	-	open the relevant pdf
	-	read paper for mentions of dataset used for empirical analysis
	-	if mention found, record all synonyms used for the dataset(s) that you find in the text
	-	if mention found, record the entire passage that describes the dataset(s) that you find in the text (this can range from one sentence to an entire paragraph or more)
3	Save resulting json file	

Source: Authors

For all identified datasets in the evaluation sample, we record two types of labels: short labels and long (verbose) labels. The latter corresponds simply to the full text passage where each used dataset is discussed. In this step, we record only the first mention per dataset, omitting repeated (and differing) descriptions of the same dataset.

For short labels, we record all synonyms of the dataset names for each of the manually labeled papers. If authors refer to the same dataset with different names within or across papers, recording only one strict definition as the ground truth will introduce a downward bias on the evaluation of the algorithm's performance.

If we force our algorithm to find exact matches for the identified datasets, all datasets identified with another name than that specifically mentioned in the paper would be recorded as unidentified, i.e., negatives. This is more severe when definitions are long, since the chance of *not* finding each of, e.g., ten words in the correct order is generally higher than for, say, definitions of three words length.

In addition, the ground truth will likely not be represented in any one paper, i.e., it is unlikely that several researchers use the same definition for the same dataset except for instances where dataset names are extremely short and standardized. This means that we would not be able to aggregate or compare results for the same dataset across papers if we do not undertake some sort of standardization of the dataset mentions.

When manually reading and labeling papers, we notice several ambiguities, proving the complexity of the extraction task not only for machines but also for humans (as described by Bender et al, 2020). Beyond the wide range of synonyms used to describe datasets, further particularities arise. While labeling, it is not always clear, what constitutes a dataset name. This can be present in cases, when there are descriptions in textual form instead of an actual dataset name (compare example in Figure 2). We further decide not to consider single time series or model calibration parameters as datasets.

As a result of the manual labeling process, we end up with a corpus of papers to be labeled and an evaluation sample, where we know (i) which datasets are used within, (ii) synonyms of the datasets used, if any, and (iii) the text passage, where the dataset was mentioned. With this data ready, we proceed in the next section by outlining the methodology used, specifically the extraction pipeline we built.

4. Methodology

To extract data citations automatically from the corpus of papers obtained, we construct a pipeline for assistant-style LLMs. Our goal is to build a flexible process allowing us to (i) test different variations efficiently, (ii) track the results systematically, and (iii) easily transfer code to other interested parties. In the following, we outline the components of this pipeline, including the measures we use to judge performance.

4.1 Pipeline overview

The pipeline takes as input a paper in the form of a PDF file and produces a list of datasets as output. Our pipeline can be segmented into five steps. First, we split papers into smaller segments. Second, we filter out irrelevant segments, before

– third – combining each segment with the instruction for the LLM. In the fourth step, we ask the LLM for the information contained, before finally consolidating the responses in step 5 (compare Figure 3 for a graphical depiction of the dataset citation extraction pipeline used).

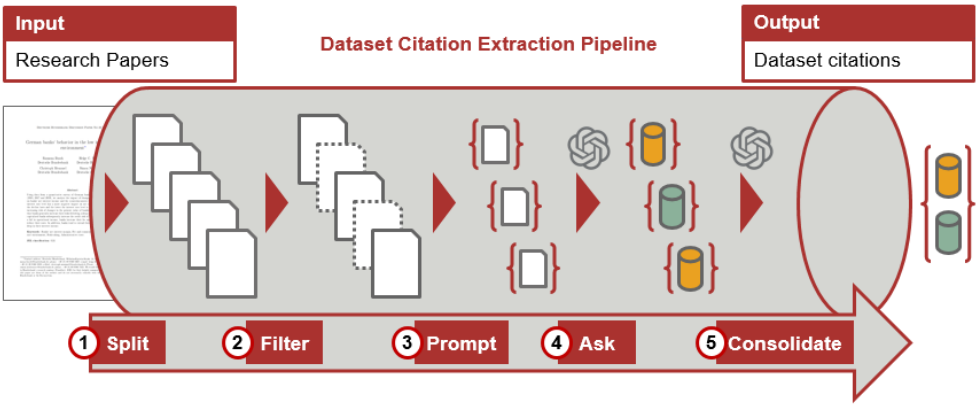


Figure 3: The Dataset Citation Extraction Pipeline used for the purpose of this study. Source: Own depiction.

In the first step, we split the document into smaller segments, or snippets. The main goal here is to ensure that each snippet will be small enough to fit into the limited context window of the LLM. A context window is the zone, where text can be analyzed coherently by the LLM, so to speak the length of the "conversation" per paper we can engage with the model. In our case, we use GPT-3.5.[6] An intuitively simple approach could be e.g. to treat each page of the document as one snippet.

While page-splitting is easy to implement, this has the drawback of splitting an ongoing paragraph into two at the page boundary. In fact, instances where contextually connected paragraphs do not end at the end of a page are most frequent, causing the LLM to consider only part of the true context of the respective page-paragraph combination at a time. A mitigation of this issue is to define an overlap between the chunks. We achieve this with `langchain`, a solution tailored to application building with LLMs.

Alternatively, we evaluate splitting documents by paragraphs by using multiple new-line commands as split separators.[7] This yields a larger number of smaller snippets

[6] More precisely, we use the `gpt-3.5-turbo` model of the OpenAI API. It is technically possible and might be sensible to compare the performance to GPT-4 with a longer context window in the future. The model used at the time of writing allowed a context window of 4,000 tokens, also referred to as 4k (approximately 5 pages of text). A token can be thought of as a piece of a word, one token is approximately 4 characters in English language. Newer versions of GPT-3.5 allow for a 16k token context window (approximately 20 pages). Recently, a 128k token context window was announced for GPT-4 (encompassing approximately 300 pages of text). In the interest of cost efficiency and technological availability at the time of writing, we relied on GPT-3.5 as a trade-off between performance and cost, but note that in the future most likely larger snippets will be used.

[7] e.g., "\n\n" in Python

compared to a page split, which leads to higher API costs but might improve performance.

The second, optional, step is to pre-filter the snippets for informative content. Independently of the splitting approach used, the text is likely to contain several snippets without relevant information for dataset extraction. Examples of such uninformative text snippets include (most) tables containing numbers, the references section or small fragments without textual content.[8] Wherever possible, we remove such snippets from consideration for further processing, prior to involvement of the LLM, based on simple heuristics.

Next, the text snippets from the paper have to be combined with instructions. This creates a prompt for a zero-shot task (compare e.g. Sun et al., 2021).[9] A prompt in this context is the plain English input we send to the model. The exact language model used determines the particulars of how to phrase the instructions. Depending on how the underlying model was designed and trained, it might allow differentiation between "system" and "user" prompts or allow fill-in-the-middle prompts.

The quality of the generated responses also differs significantly depending on the phrasing of the instructions. As it has been noted in the literature, that zero-shot prompts can outperform prompts with examples of task-completion (Reynolds & McDonell, 2021), our pipeline allows the systematic tracking of changes in output quality caused by alternative instructions. The prompts we use for this analysis can be seen in Appendix 1. In short, we ask the model identify a list of data sources, or simply write "None" if it cannot find any.

In the fourth step, we send the prompt to the LLM, consisting of instructions and text content. At this point different sampling parameters can be set. Of particular interest to us is the "temperature" parameter, which rebalances the probability distribution for the next token in the generated sequence, with high temperature corresponding to a more uniform distribution.

High temperature is useful to generate more creative output and to prevent stiff, predictable writing style. In our use-case, where we are interested in simple retrieval of factual information from a given text, we set the temperature as low as possible. Our pipeline also takes care of any connectivity issues, for example due to downtime or throttling. The result of this step is an answer to our inquiry for each individual text snippet.

In the fifth and last step, individual model replies per snippet must be consolidated into a single answer covering the entire input paper. Any single data source is likely to be mentioned multiple times across several sections of a research paper. The identification of the union of all the answer-sets is also left to the LLM. The result is supposed to be a single list of all unique data sources used in the processed research paper. We evaluate, how close the automatically-generated list is to what a human reader would expect in the next section.

[8] Tables could be informative if they contain dataset information. This can be the case, for instance, when an Appendix table includes information on further datasets used.

[9] Even though language models in our context technically are few-shot learners

4.2 Performance measures

Evaluating the consolidated results offers some interest interesting aspects. For once, LLMs like GPT-3.5 have been reported to hallucinate ("confabulation"). Further, our evaluation aims to compare two lists of free text (the consolidated model output list and the manually labeled evaluation sample), requiring fuzzy string comparisons.

While our application of fact-retrieval from a given text is not particularly susceptible to confabulation, we nonetheless need to measure the quality of the responses that our pipeline produces. This is also crucial information to test the impact of changes to our processing pipeline.

Determining the performance of the model and pipeline by comparing its output to a list labeled by a human is not trivial. In essence, we would like to gauge the similarity of the two lists. The generated lists being free text, rather than a subset of predefined labels complicate this. While our approach allows the model to identify data sources beyond our current awareness, the output then necessitates the use of fuzzy string comparisons. In Figure 4, we show the evaluation process pipeline.

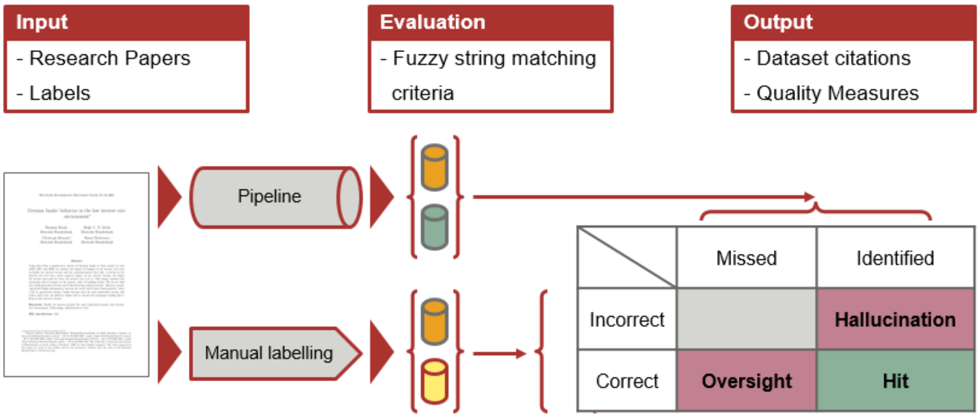


Figure 4: Assessing the quality of the model's output. Source: Own depiction

We explore different versions of this fuzzy comparison, highlighting the tradeoff between the error types based on the strictness of the match. We calculate model recall (fraction of identified "true" data sources, regardless of spurious identifications) and output similarity measured by various criteria.

For the task at hand, the model can make two kinds of errors, oversights and hallucinations. Oversights are cases, where a truly mentioned data source is not recognized as such by the model and thus not listed in the output. Hallucinations are incorrect identifications, where the models output includes some entry that does not objectively count as a data source (either because the model misunderstood or confabulated).

These two kinds of errors are not equally important in our case. For our use case, we prefer the inclusion of an inaccurate entry in the result set to the possibility of overlooking a true data source mention, as the former can more easily be detected and corrected downstream. In other words, our current focus is predominantly on recall at this step.

5. Results

When evaluating recall of our approach (i.e. the fraction of correctly identified dataset citations as a fraction of all dataset citations), we apply several criteria to compare model predictions with our evaluation sample. We (i) let the LLM judge its own results as described above, (ii) compare exact string matches, (iii) fuzzy string matches, (iv) fuzzy string matches including synonyms, and (v) the Jaccard similarity.

As these criteria are not equally strict in comparing similarity, measures of recall vary widely. In the strictest scenario, when we only count exact matches, we reach a recall of only 12% (17% conditional on the fact, that the model predicted any data to be contained at all). This result comes maybe unsurprisingly, as it signifies the variety of dataset names, i.e. complexity of the extraction task.

When we ask the LLM to evaluate its own results, we obtain a 34% recall (62% when considering only papers, where the model predicted any data). While elegant in the context of our pipeline to leverage the LLM for evaluation, too, this criterion remains a black box, as the exact evaluation metric is subjective to the model and users are unaware a priori of the direction of a potential bias.

The mildest evaluation measure is likely fuzzy synonym search, where any variations of synonyms of the dataset that are found are counted as a hit. Arguably, this may be a valid measure for our use-case, as a synonym of a dataset name is often just as correct as any other. Using this criterion, our approach finds 71% percent of dataset citations in the papers provided. Figure 5 depicts recall results for all evaluation criteria used.

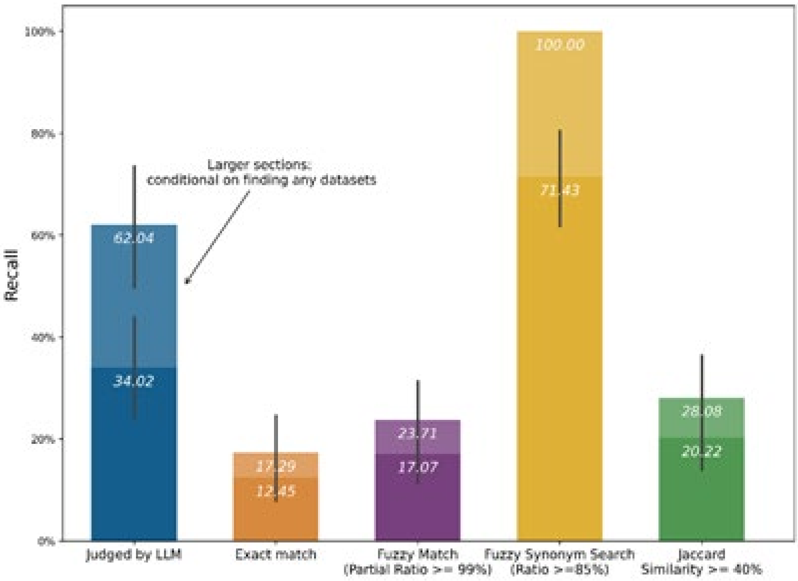


Figure 5: Performance measures of model recall based on different criteria. Source: Author's own calculations.

Note that fuzzy synonym search is not necessarily always feasible to evaluate results, if our pipeline were to be applied in another context, as it requires some domain knowledge to know synonyms of dataset names. However, providers of datasets often possess such domain knowledge. Further note that such domain knowledge is only needed for evaluation using synonyms, not to algorithm performance itself.

As the recall numbers vary considerably depending on the exact criteria used, due to the loosely defined task of comparing lists of strings, we feel compelled to comment briefly on our subjective feeling. When looking at the similarity of lists ourselves, we were surprised by the capability of the LLM to identify datasets correctly, even in the case of complicated names or infrequently used datasets.

Users of the automated dataset extraction pipeline might be interested in inspecting the performance of the LLM assistant on an individual dataset level. Here, we notice a very high ability of the algorithm to identify well-structured and well-defined datasets, in our case due to the corpus of input papers particularly from administrative data sources (datasets originating from Deutsche Bundesbank) and commercial providers. A particularity of these datasets is that they are usually clearly defined, named, and separable entities.

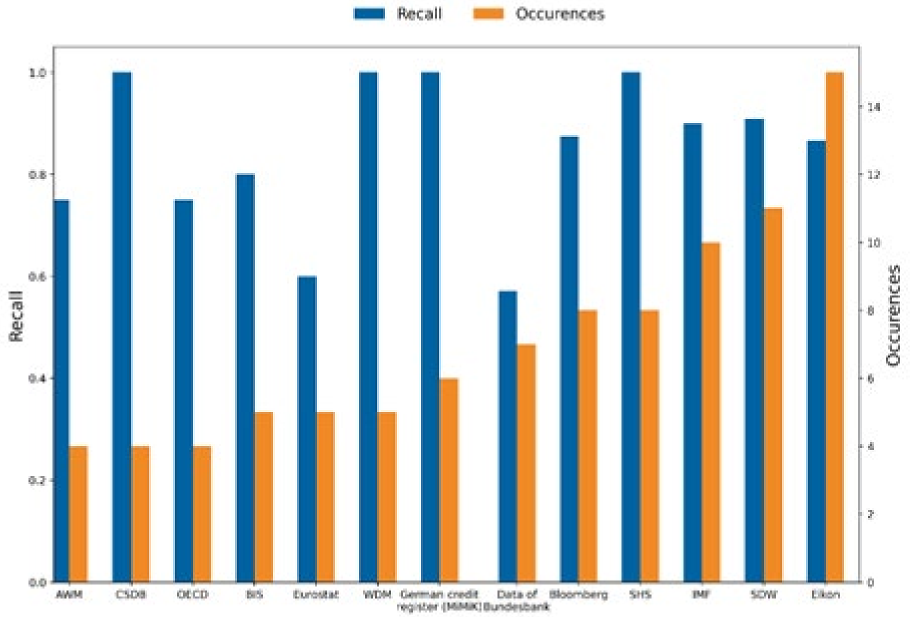


Figure 6: Recall per dataset, judged by LLM. Source: Authors own calculations.

Furthermore, the existence of relatively unique abbreviations seem to support automatic dataset extraction (e.g. "CSDB" and "SHS"[10]). Figure 6 shows recall for a number of selected administrative datasets in our sample. Before deriving ambitious conclusions from this finding, the scope of the extraction would have to be enlarged, as the number of occurrences per dataset in our sample remains limited. However, we take this as an encouraging sign for the potential of our approach.

Even considering the preliminary nature of results as part of the presented proof-of-concept, it seems already efficient to streamline the process of identifying data usage in papers by using an LLM, since the baseline to date is human labeling. This holds true even if model performance itself is deemed insufficient for certain use-cases (whereas we would argue that for our use-case it is already sufficient), as long as pre-labeling by the model increases efficiency of subsequent human labeling. Hence, using our pipeline seems to yield a fairly reliable gauge of data impact, when identifying known datasets in unknown papers.

6. Conclusions

Both, data providing institutions and data users such as empirical researchers have an interest in tracing how datasets are used. If structuredly available, information on how datasets are used in research papers would allow data providers to measure data impact and researchers to inspect data usage in a systematic way. In other words, the whole infrastructure around research paper citations (such as paper and journal

[10] "Central Securities Data Base" and „Securities Holdings Statistics", two large and well-documented European administrative datasets used for non-commercial research.

impact and citation networks) could be recreated for datasets, benefiting providers and users of data.

To date, dataset citations in research papers have not been systematically used, due to the unstructured textual nature of dataset citations. Recent advancements in NLP allow us to progress on the status quo. We present a pipeline to leverage LLMs, specifically GPT-3.5, to extract dataset citations and evaluate performance using a manually labeled evaluation sample.

We find that LLM performance is good in terms of recall. The LLM correctly identifies many of the datasets used in the papers presented to the LLM. Performance in identification is particularly good for known datasets, where domain knowledge in the form of known used synonyms of dataset names can be provided to the algorithm. However, measures of recall vary to a fair degree due to the unclear defined task of comparing two lists.

We argue that model performance is sufficiently high to measure data impact, since the proof-of-concept already offers value-generating use-cases. First, our approach considerably advances on the status quo, where manual labeling to trace data impact is often prohibitive in terms of time and effort. Here, we offer a viable pre-screening option to make tracing for humans feasible. Second, even if there are no human capacities available, our approach allows tracing data impact over time reliably, as long as dataset citation practices remain constant in a field.

Our pipeline is easily transferrable to other institutions interested in tracing their data usage in research or researchers interested in data citation networks. We use GPT 3.5 for the results described here, however the LLM can be exchanged relatively simply. Exchanging the LLM at the core of our pipeline allows interesting extensions and likely quality gains in the near future, with the quick pace of increasing LLM capabilities we currently witness.

Possible extensions include leveraging on-premise LLMs to allow tracking data impact in confidential texts, without concern of sharing confidential data outside the institution. If in a given institution, data impact is characterized predominantly by its usage in internal data-driven decision-making, it could make sense to evaluate data mentions in confidential texts, such as internal briefs, board meeting minutes, etc.

Our proof-of-concept allows important lessons learned for data providers. Data impact can now be traced automatically and efficiently. Therefore, it is now possible to provide data knowing how it is used, i.e. with a more clear user-centric focus. Therefore, data-driven decision-making and knowledge-generation can be enhanced in several key aspects.

By knowing the impact of datasets, (i) resources can be focused on value-creating data, (ii) underused data can be identified and potential usage hurdles removed, (iii) popular datasets can be promoted, (iv) combined with researcher information, tailored datasets can be recommend, and (v) frequently jointly used datasets can be provided together. With these value propositions in mind, our next step is to ingest a larger corpus of research papers to leverage on the next generation of LLMs.

References

Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K. E. & Jaradeh, M. Y. (2020). Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis, 44*, 516–529.

Bender S., Doll H. C., & Hirsch, C. (2020). Where's Waldo? Conceptual Issues when Characterizing Data in Empirical Research. In: Lane, J., Mulvany, I., & P. Nathan (Eds), "Rich search and discovery for research datasets: Building the next generation of scholarly infrastructure", Sage Publications.

Blank, S., Lipponer, A., Schild, C. J., & Scholz, D. (2020). Microdatabase Direct Investment (MiDi)–A full survey of German inward and outward investment. German Economic Review, 21(3), 273-311.

Boland, K., & Krüger, F. (2019). Distant Supervision for Silver Label Generation of Software Mentions in Social Scientific Publications. *BIRNDL@ SIGIR*, (pp. 15–27).

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying references to datasets in publications. *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings 2*, (pp. 150–161).

Buch, C. M., Kleinert, J., Lipponer, A., Toubal, F., & Baldwin, R. (2005). Determinants and effects of foreign direct investment: evidence from German firm-level data. Economic Policy, 20(41), 52-110.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *American economic association, ten years and beyond: Economists answer NSF's call for long-term research agendas*.

Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Advanced Engineering Informatics, 42*, 100959.

Eckle-Kohler, J., Nghiem, T.-D., & Gurevych, I. (2013). Automatically assigning research methods to journal articles in the domain of social sciences. *Proceedings of the American Society for Information Science and Technology, 50*, 1–8.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*, 1243089.

Färber, M., Albers, A., & Schüber, F. (2021). Identifying Used Methods and Datasets in Scientific Publications. *SDU@ AAAI*.

Gemelli, A., Vivoli, E., & Marinai, S. (2023). CTE: A Dataset for Contextualized Table Extraction. *arXiv preprint arXiv:2302.01451*.

Ghavimi, B., Mayr, P., Lange, C., Vahdati, S., & Auer, S. (2016). A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use, 36*, 171–187.

Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2019). Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. *arXiv preprint arXiv:1906.09317*.

Houngbo, H., & Mercer, R. E. (2012). Method mention extraction from scientific research papers. *Proceedings of COLING 2012*, (pp. 1211–1222).

Ikeda, D., Nagamizo, K., & Taniguchi, Y. (2020). Automatic identification of dataset names in scholarly articles of various disciplines. *International Journal of Institutional Research and Management, 4*, 17–30.

Jayaram, K., & Sangeeta, K. (2017). A review: Information extraction techniques from research papers. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, (pp. 56–59).

Kardas, M., Czapla, P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., & Stojnic, R. (2020). Axcell: Automatic extraction of results from machine learning papers. *arXiv preprint arXiv:2004.14356*.

Kovačević, A., Konjović, Z., Milosavljević, B., & Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech & Language, 26*, 105–126.

Kumar, S., Ghosal, T., & Ekbal, A. (2021). DataQuest: An Approach to Automatically Extract Dataset Mentions from Scientific Papers. Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23, (pp. 43–53).

Lipponer, A. (2006). Microdatabase direct investment-MiDi. A brief guide. Bundesbank working paper, Frankfurt.

Luan, Y. (2018). Information extraction from scientific literature for method recommendation. *arXiv preprint arXiv:1901.00401*.

Otto, W., Zielinski, A., Ghavimi, B., Dimitrov, D., Tavakolpoursaleh, N., Abdulahhad, K., . . . Dietze, S. (2020). Knowledge extraction from scholarly publications: The GESIS contribution to the rich context competition.

Polak, M. P., & Morgan, D. (2023). Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering– Example of ChatGPT. *arXiv preprint arXiv:2303.05352*.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7).

Saier, T., & Färber, M. (2020). unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics, 125*, 3085–3108.

Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. PLOS Digital Health, 2(11), e0000388.

Sun, X., Gu, J., & Sun, H. (2021). Research progress of zero-shot learning. Applied Intelligence, 51, 3600-3614.

Wang, Y., Zhang, C., & Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics, 127*, 2479–2520.

Yao, R., Hou, L., Ye, Y., Zhang, J., & Wu, J. (2019). Method and Dataset Mining in Scientific Papers. *2019 IEEE International Conference on Big Data (Big Data)*, (pp. 6260–6262).

Yoo, S., & Jeong, O. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications, 141*, 112965.

Appendix

Box 1

Prompts

The final prompt used reads as follows

A dataset is a collection of structured or unstructured data that is organized and grouped together for a specific purpose. It typically consists of multiple data points or observations related to a particular topic or subject. A dataset can include various types of information such as numerical values, text, images, audio, video, or any other form of data. It is often used in the context of data analysis, machine learning, and statistical research, where the data is utilized to extract insights, train models, or draw conclusions. Datasets can be generated through various means, including surveys, experiments, observations, or by gathering existing data from different sources.

The text after the empty lines is a scientific paper excerpt

According to the definition on the first line, search for any mentions of datasets or data-sources used in the paper's research.

If you find any, please compile a list of all mentioned datasets in this excerpt.

If there aren't any, please reply with 'None'.

Consolidation prompt

You will be provided one or more lists of datasets.

Each new list starts with '=>'.

You need to combine all the lists into a single one by removing redundant entries. Delimit the final list with simple '-' bullet points.

Recall prompt

You will be provided with text delimited by triple quotes that is supposed to be a list of datasets. Check if the following true datasets are directly contained in the answer:

...

For each of these true datasets perform the following steps:

1 - Restate the true dataset

2 - Write 'yes' if the true dataset is mentioned in the answer, otherwise write 'no'

Finally, provide a count of how many 'yes' findings there are. Provide this count as {"count":<insert count here>}.

Leveraging Large Language Models to Extract Data Citations

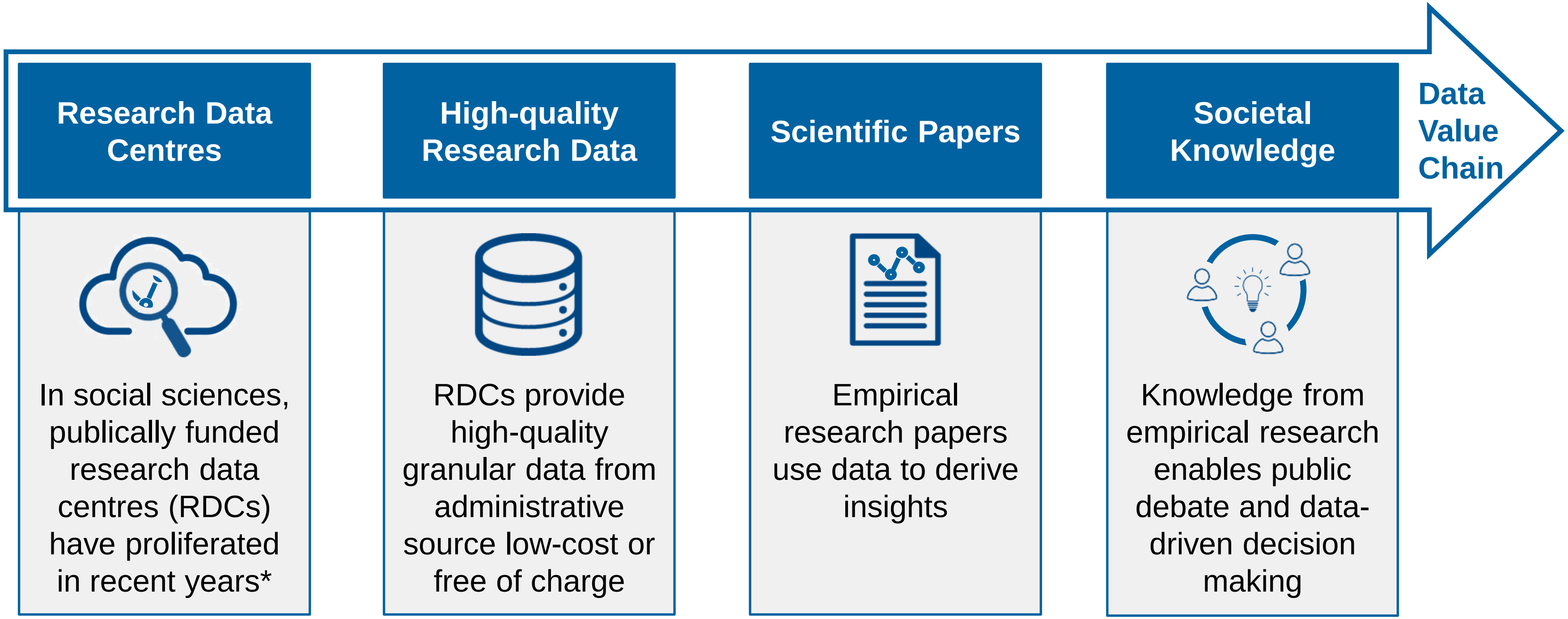
3rd IFC Workshop on Data Science in Central Banking

Sebastian Seltmann, Emily Kormanyos, Hendrik Christian Doll, Shir Frank, and Kilian Graef – Deutsche Bundesbank, Data Service Centre

The authors would like to thank colleagues in Bundesbank's Data Service Centre for valuable comments and feedback.

All views expressed in this report are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

The Data Value Chain



Data Value Chain

Research Data Centres

In social sciences, publically funded research data centres (RDCs) have proliferated in recent years*

High-quality Research Data

RDCs provide high-quality granular data from administrative source low-cost or free of charge

Scientific Papers

Empirical research papers use data to derive insights

Societal Knowledge

Knowledge from empirical research enables public debate and data-driven decision making

Since public funds are used, the public has an interest in knowing the value created by the investment

Issue: It is hard to measure data impact, as data citations in scientific papers are not standardised

* For details on how RDCs work, compare slides in appendix

Extracting Data Citations from Academic Papers

Theory

Blaschke & Hirsch (2023)[1] attempt to **measure the value of an RDC**, leveraging on manually collected info on projects and publications

Related Work

Polak & Morgan (2023)[2] propose **ChatExtract to extract materials data in research papers**, leading to accurate results when identifying known properties

Gap in Literature

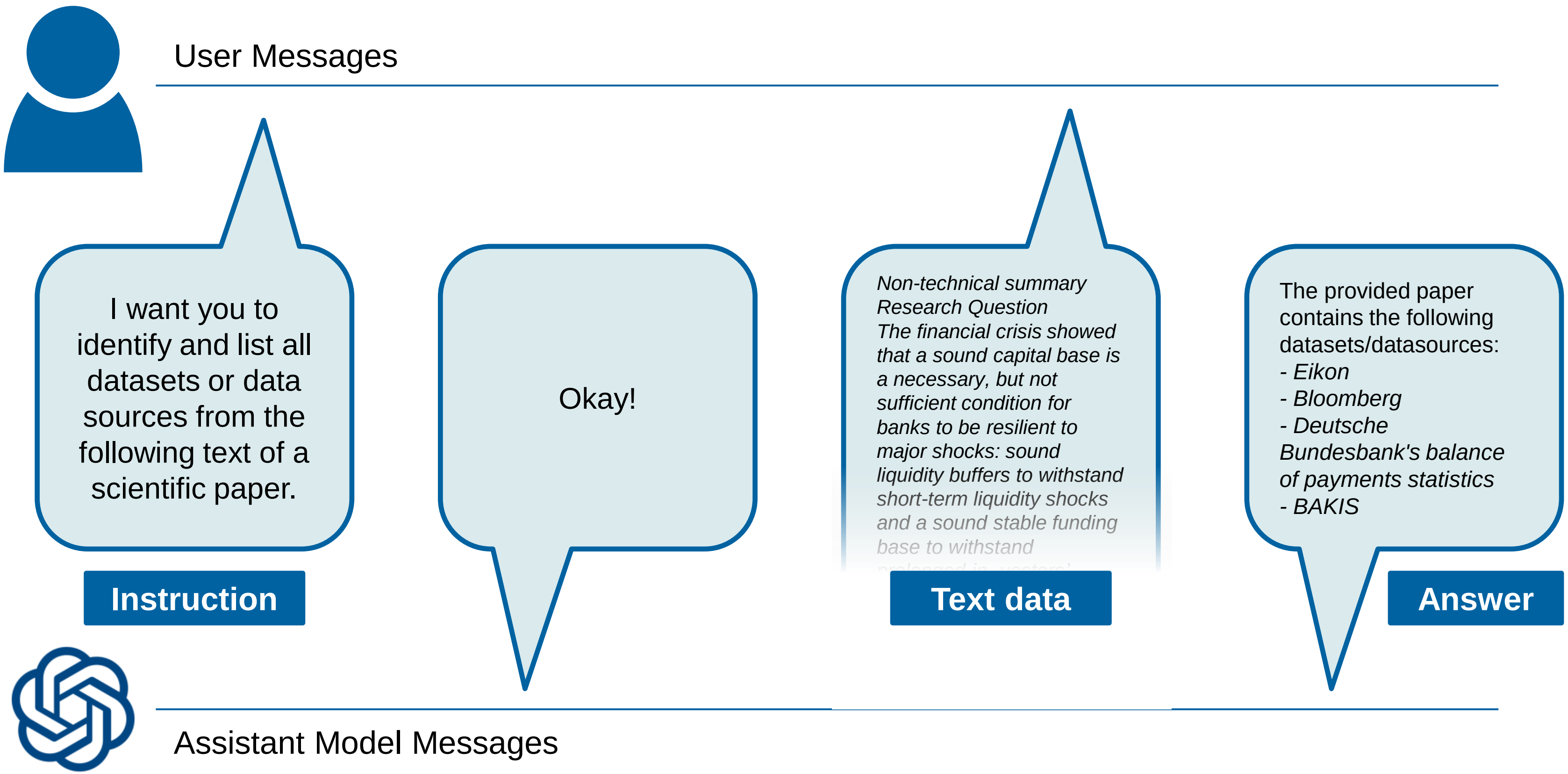
Most of the literature to date seems to **focus on the natural sciences or ML/NLP papers** specifically

Our Contribution

Recent advances in natural language processing (NLP) enable us to flexibly **detect and connect data source descriptions from academic papers using GPT-3.5**

[1] Blaschke & Hirsch (2023)
[2] Polak & Morgan (2023)

The Assistant-Style Language Model Approach



I want you to identify and list all datasets or data sources from the following text of a scientific paper.

Okay!

Non-technical summary
Research Question
The financial crisis showed that a sound capital base is a necessary, but not sufficient condition for banks to be resilient to major shocks: sound liquidity buffers to withstand short-term liquidity shocks and a sound stable funding base to withstand prolonged investors'

The provided paper contains the following datasets/datasources:
- *Eikon*
- *Bloomberg*
- *Deutsche Bundesbank's balance of payments statistics*
- *BAKIS*

Paper Corpus and Evaluation Sample

DATA

CORPUS TO BE LABELLED

- From **Bundesbank discussion paper** series
- Data sample: **1,103** papers
- Papers focus on economics and finance

EVALUATION SAMPLE

- Focus on **empirical papers using databases** (not data used for model calibrations)
- We draw a **random sample** for labeling to ensure **unbiased** evaluation metrics
- **103** papers manually labelled

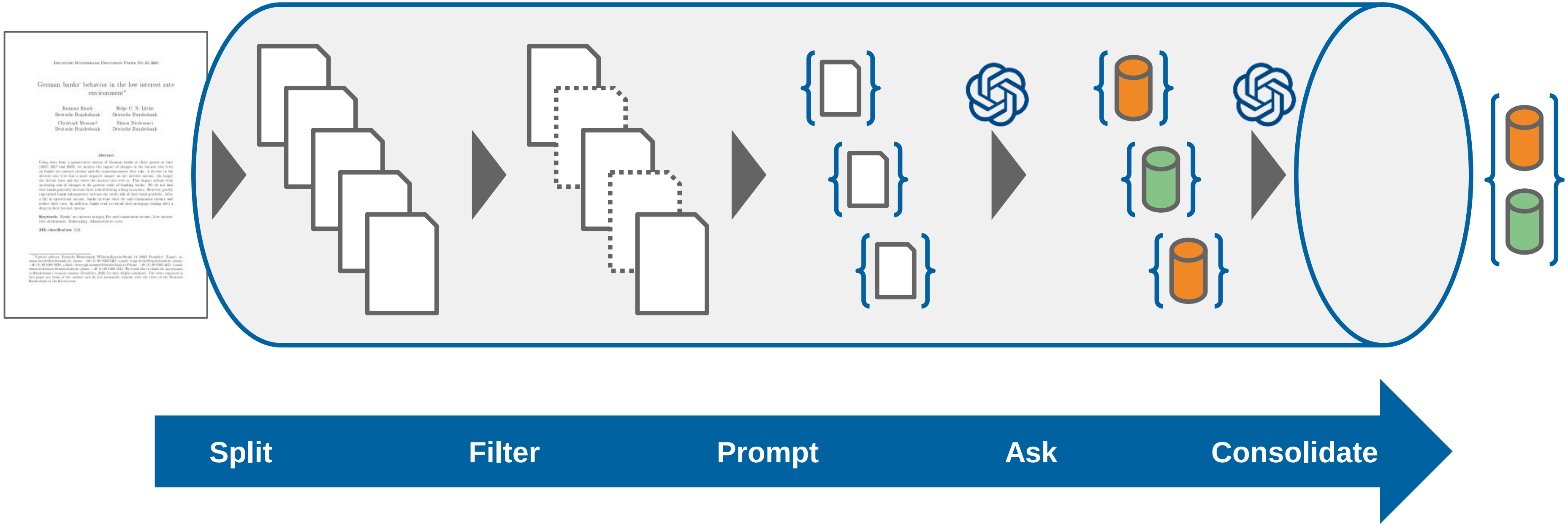
Dataset Citation Extraction Pipeline

Input

Academic Papers

Output

Data Sources



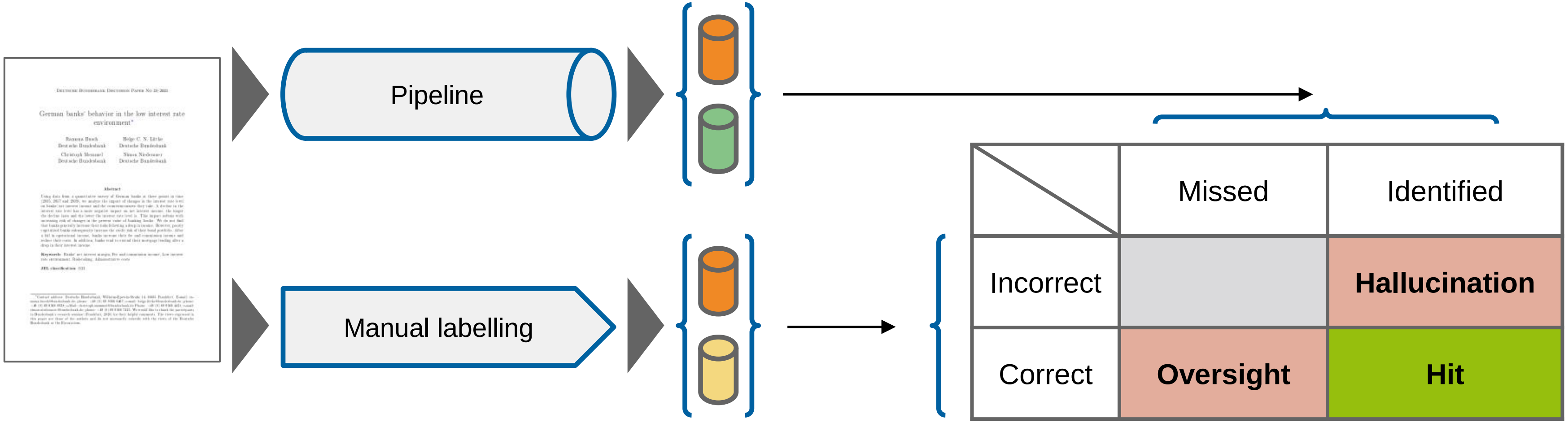
Assessing the Quality of the Assistant-Model’s Output

Input

- Academic Papers
- Labels: Data Sources

Output

- Data Sources
- Quality Measures



Deutsche Bundesbank Discussion Paper No 23/2021

German banks' behavior in the low interest rate environment*

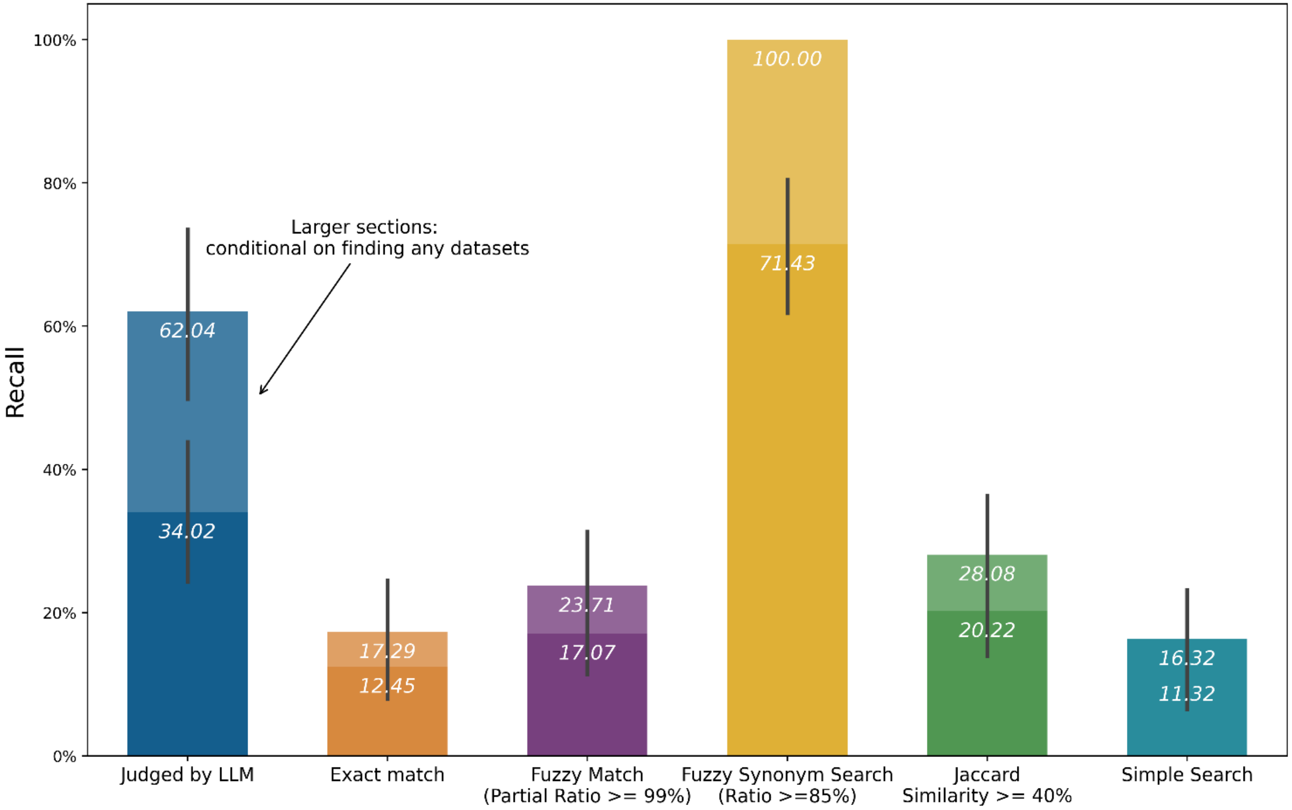
Ramona Busch, Deutsche Bundesbank; Helge C. N. Littke, Deutsche Bundesbank; Christoph Memmel, Deutsche Bundesbank; Simon Niederauer, Deutsche Bundesbank

Abstract

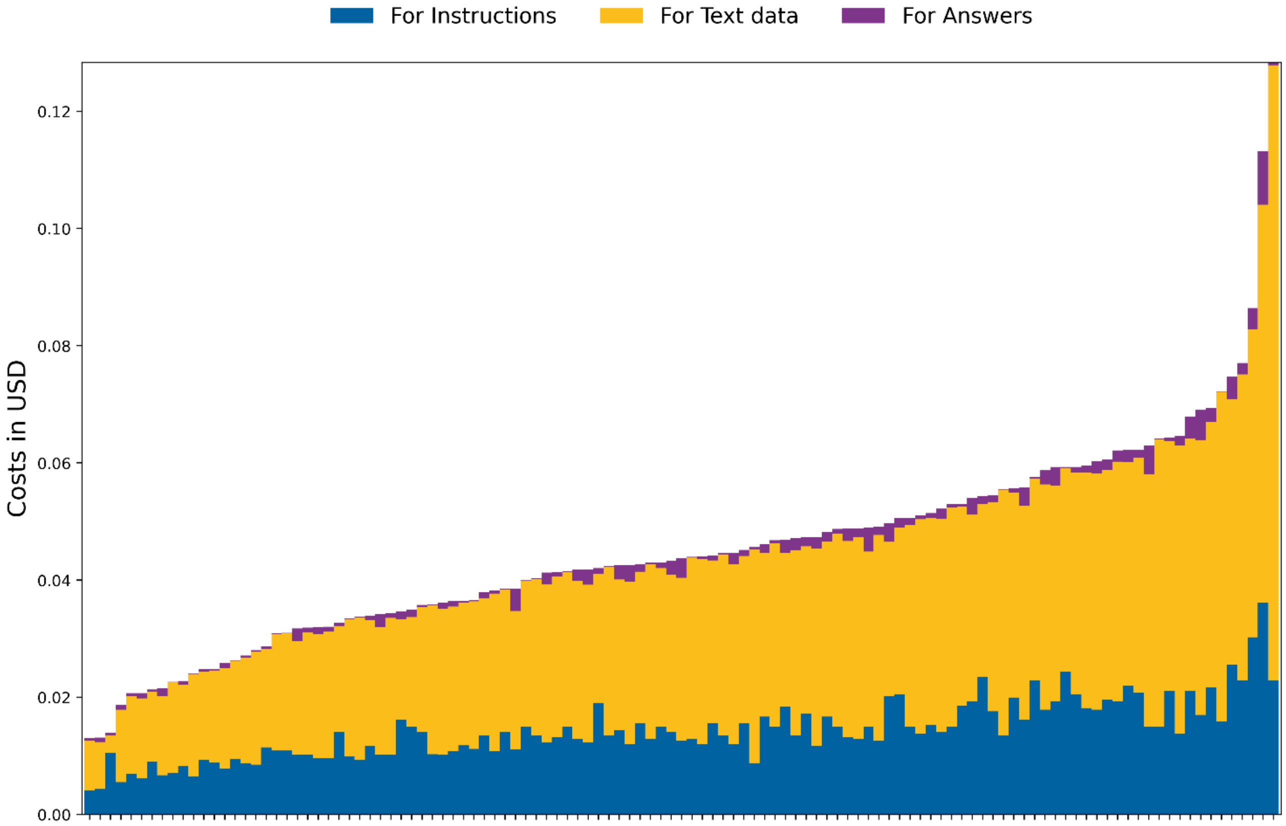
[illegible]

	Missed	Identified
Incorrect		**Hallucination**
Correct	**Oversight**	**Hit**

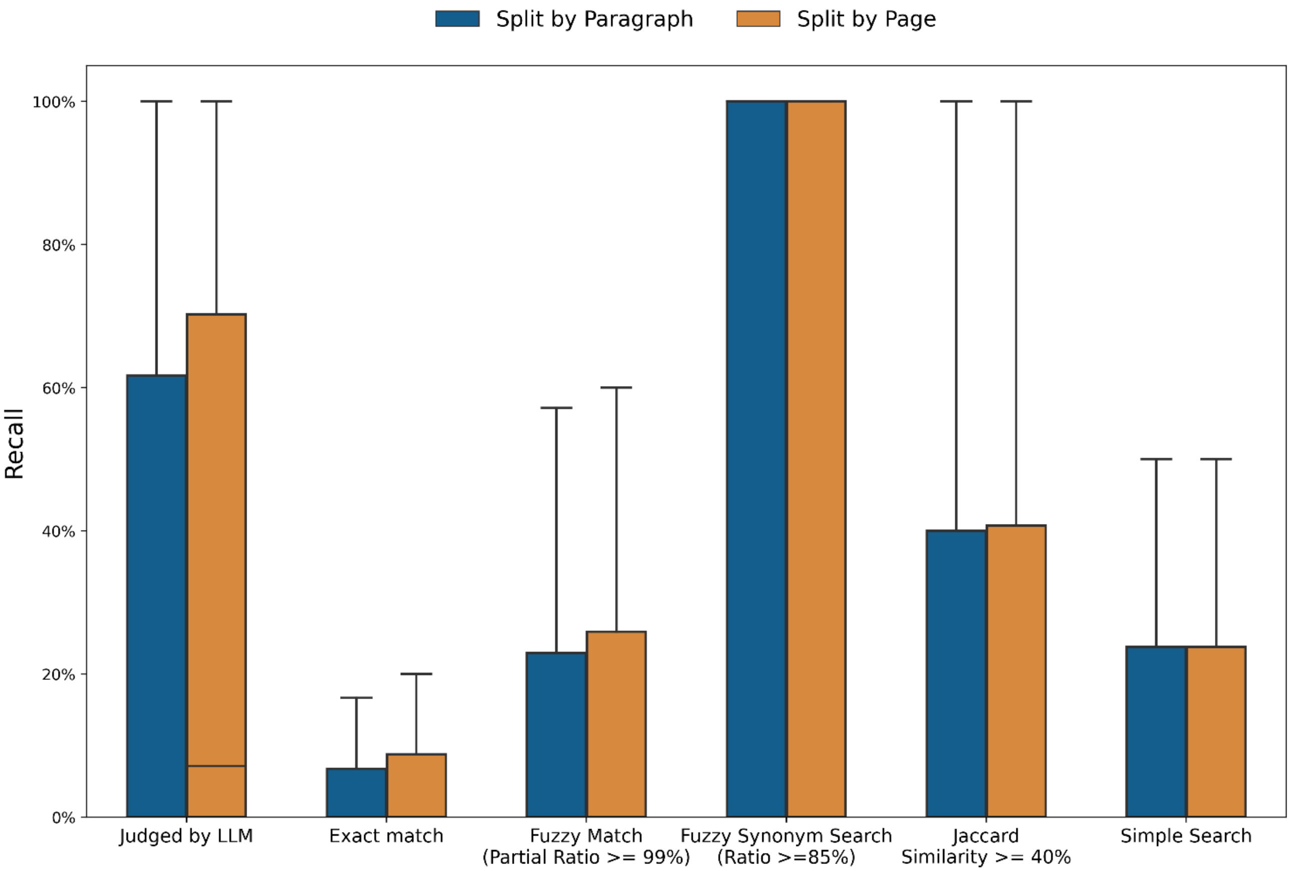
Results
Different measures of apparent performance



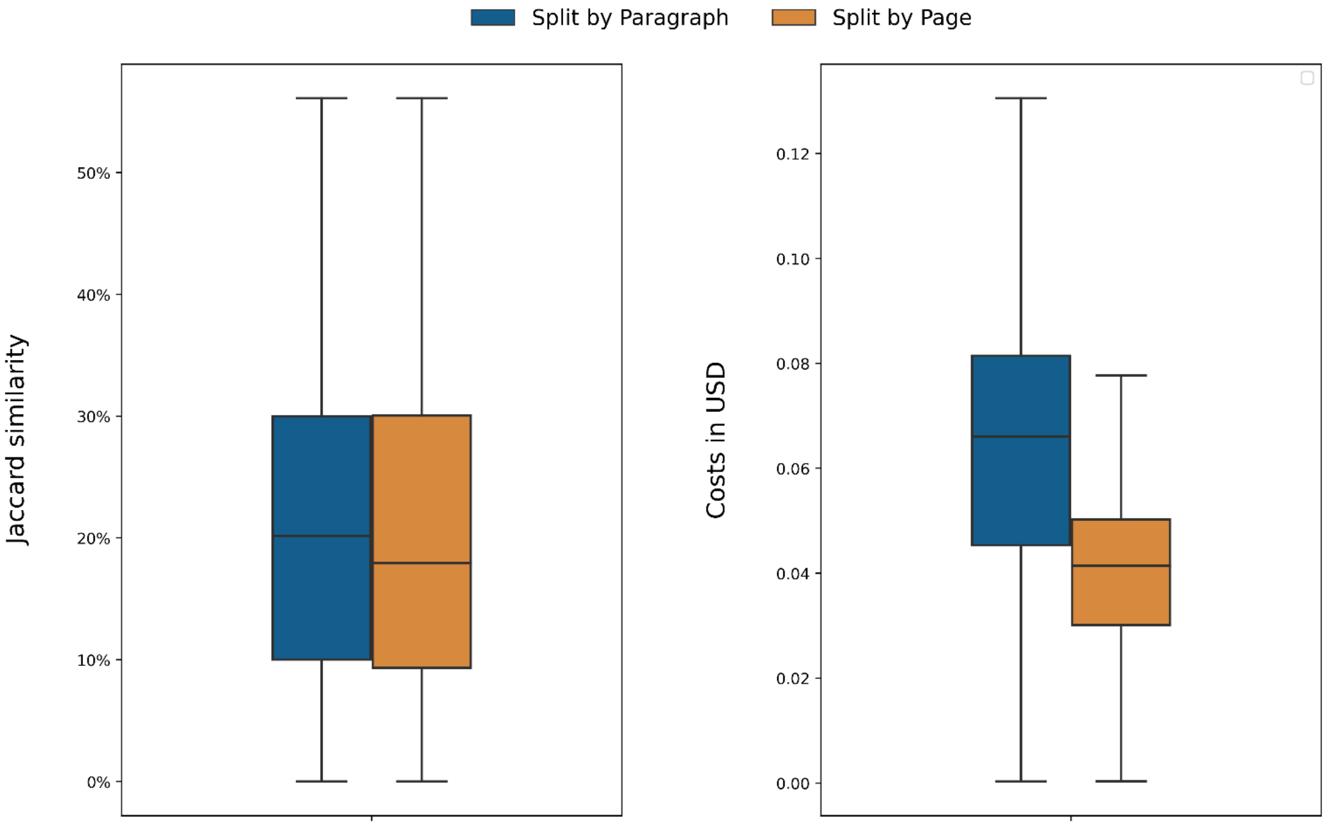
Results
API Costs per paper processed



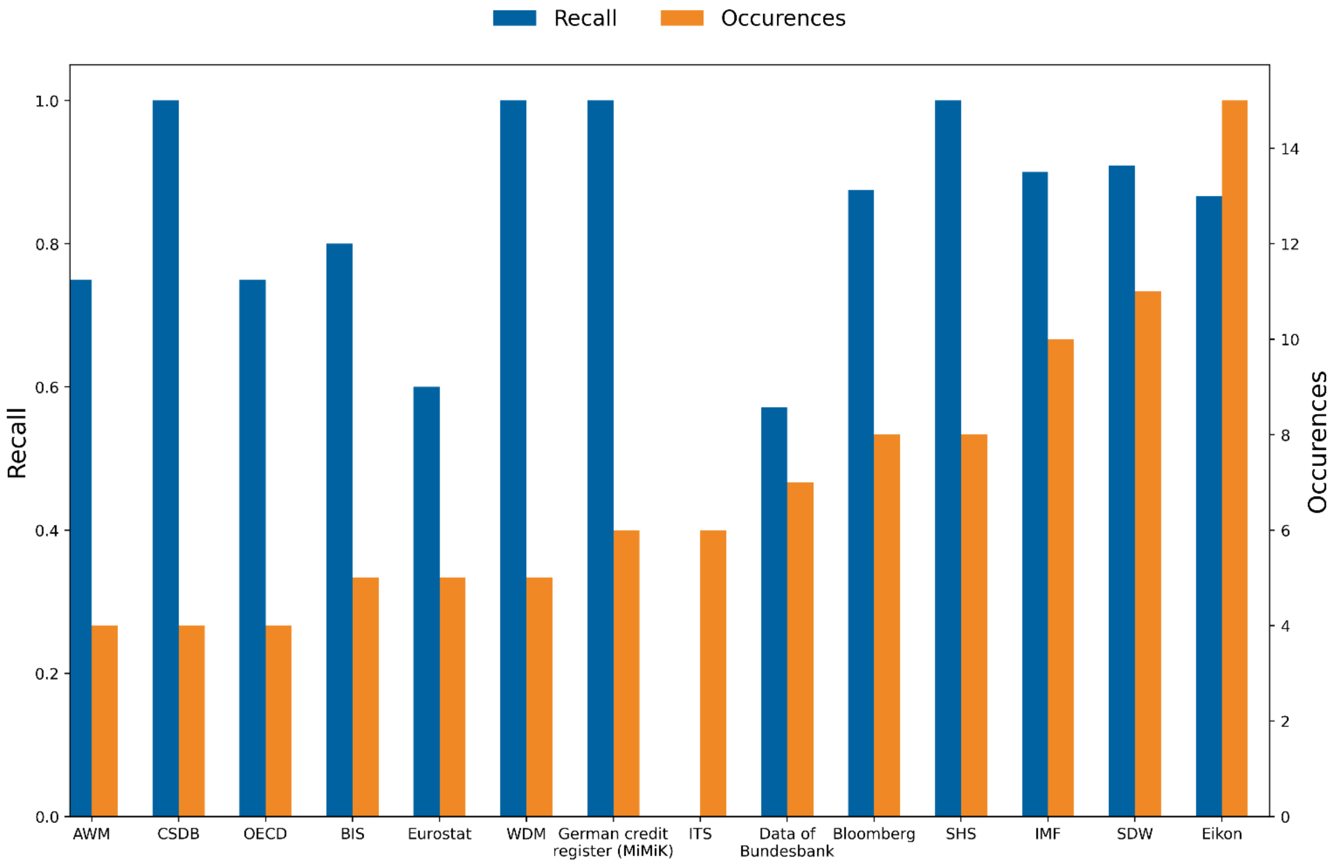
Results
Larger Contexts provide minor performance improvement



Results
Larger Contexts significantly reduce overall costs



Results
Recall for individual datasets



Challenges & Mitigations

DEPENDENCY

Reliance on the OpenAI API is a point of failure
Unforeseen outages are possible

→

- OpenAI transparently reports system status
- Pipeline can plug-in alternative models

UNSTRUCTUREDNESS

Assistant-Style language models produce plain text, strict output-syntax is merely a suggestion

→

- Zero “temperature” improves reliability

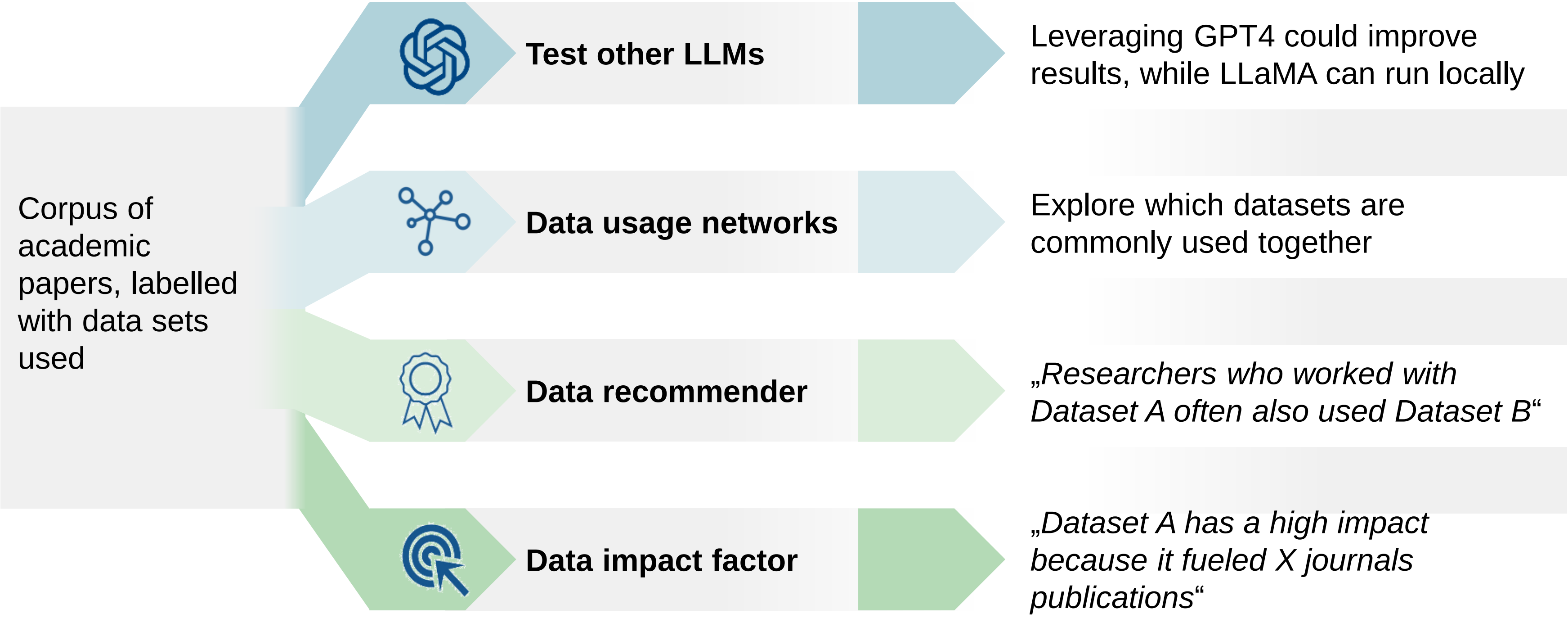
TESTING

The definition of success itself is up to debate and interpretation

→

- Pipeline tracks as many potentially relevant dimensions as possible

Future Extensions and Applications



RESULT	EXTENSION	DETAIL
Corpus of academic papers, labelled with data sets used	**Test other LLMs**	Leveraging GPT4 could improve results, while LLaMA can run locally
	Data usage networks	Explore which datasets are commonly used together
	Data recommender	*„Researchers who worked with Dataset A often also used Dataset B“*
	Data impact factor	*„Dataset A has a high impact because it fueled X journals publications“*

Key Take-Aways

1

CAPABILITY

Assistant-style language models can effectively extract specific types of information from large bodies of unstructured text

2

EVALUATION

Think carefully about how to measure success and track experiments systematically

3

GUIDANCE

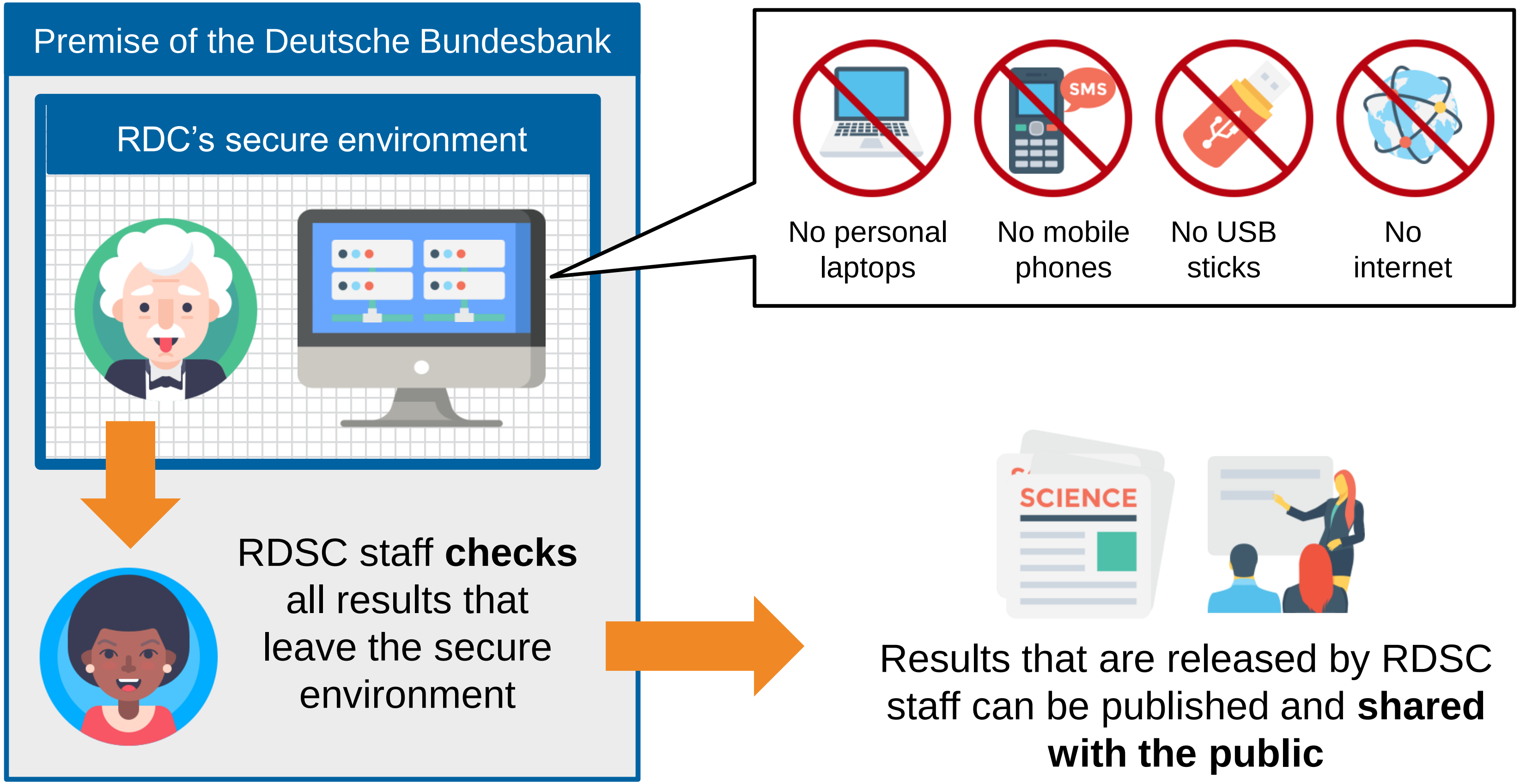
Give the model clear and simple tasks, yet as much context as possible

References

- Blaschke, J. & C. Hirsch (2023). On the value of data sharing: Empirical evidence from the Research Data and Service Centre, Technical Report 2023-08 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre. Retrieved 2023/08/21 from https://www.bundesbank.de/resource/blob/863758/7ebe74476186cd3364a11b3869ada80a/mL/2023-08-value-data.pdf
- Polak, M. P. & D. Morgan (2023). Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. arXiv preprint arXiv:2303.05352

✉ sebastian.seltmann@bundesbank.de
✉ **fdsz-data@bundesbank.de**
https://www.bundesbank.de/en/bundesbank/research/rdsc

Appendix: RDCs allow to securely share confidential granular data from administrative sources with external researchers



Premise of the Deutsche Bundesbank

RDC’s secure environment

No personal laptops

No mobile phones

No USB sticks

No internet

RDSC staff **checks** all results that leave the secure environment

Results that are released by RDSC staff can be published and **shared with the public**

Appendix: RDCs proliferate in recent years, both nationally and internationally, enabling high-quality research

RDCs in other institutions*

Bundesamt für Migration und Flüchtlinge

Office for National Statistics

BANCA D'ITALIA
EUROSISTEMA

BANCO DE PORTUGAL
EUROSISTEMA

eurostat

BANQUE DE FRANCE
EUROSYSTÈME

DESTATIS
Statistisches Bundesamt

BANCO DE ESPAÑA
Eurosistema

IAB INSTITUT FÜR ARBEITSMARKT- UND BERUFSFORSCHUNG
Die Forschungseinrichtung der Bundesagentur für Arbeit

Centraal Bureau voor de Statistiek

** selected examples*

Appendix: The Prompts

Identification Prompt

A dataset is a collection of structured or unstructured data that is organized and grouped together for a specific purpose. It typically consists of multiple data points or observations related to a particular topic or subject. A dataset can include various types of information such as numerical values, text, images, audio, video, or any other form of data. It is often used in the context of data analysis, machine learning, and statistical research, where the data is utilized to extract insights, train models, or draw conclusions. Datasets can be generated through various means, including surveys, experiments, observations, or by gathering existing data from different sources.

The text after the empty lines is a scientific paper excerpt

According to the definition on the first line, search for any mentions of datasets or data-sources used in the paper's research.

If you find any, please compile a list of all mentioned datasets in this excerpt.

If there aren't any, please reply with 'None'.

Appendix: The Prompts

Consolidation Prompt

```
You will be provided one or more
lists of datasets.
Each new list starts with '=>'.
You need to combine all the lists
into a single one by removing
redundant entries. Delimit the final
list with simple '-' bullet points.
```

Recall Prompt

```
You will be provided with text delimited by triple
quotes that is supposed to be a list of datasets.
Check if the following true datasets are directly
contained in the answer:

...

For each of these true datasets perform the following
steps:
1 - Restate the true dataset
2 - Write 'yes' if the true dataset is mentioned in
the answer, otherwise write 'no'
Finally, provide a count of how many 'yes' findings
there are. Provide this count as {"count":<insert
count here>}.
```