DEUTSCHE BUNDESBANK EUROSYSTEM

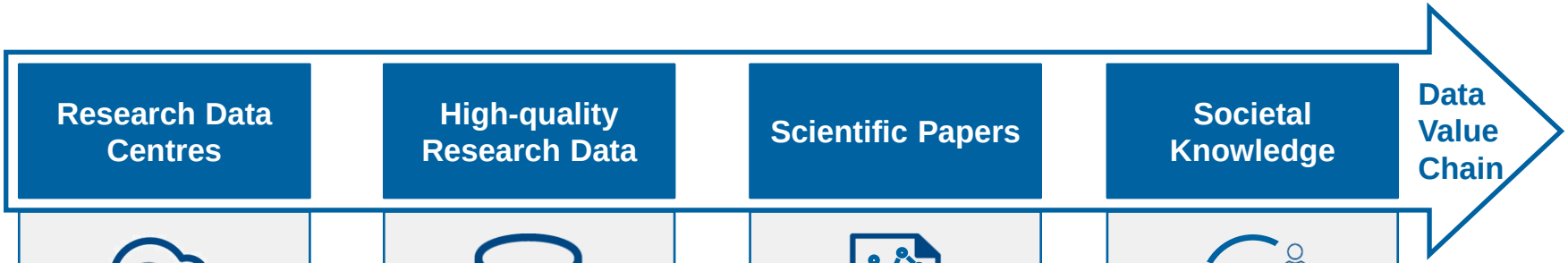
Leveraging Large Language Models to Extract Data Citations

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.

The Data Value Chain



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

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

Empirical research papers use data to derive insights

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

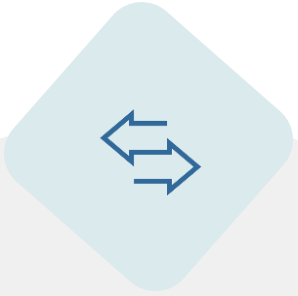
Issue: It is hard to measure data impact, as data citations in scientific papers are not standardised

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

Extracting Data Citations from Academic Papers

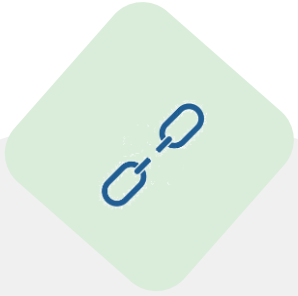
Theory

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



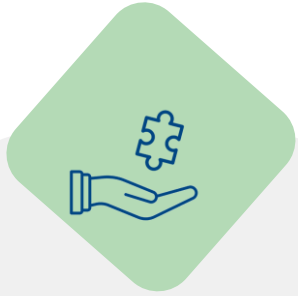
Related Work

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

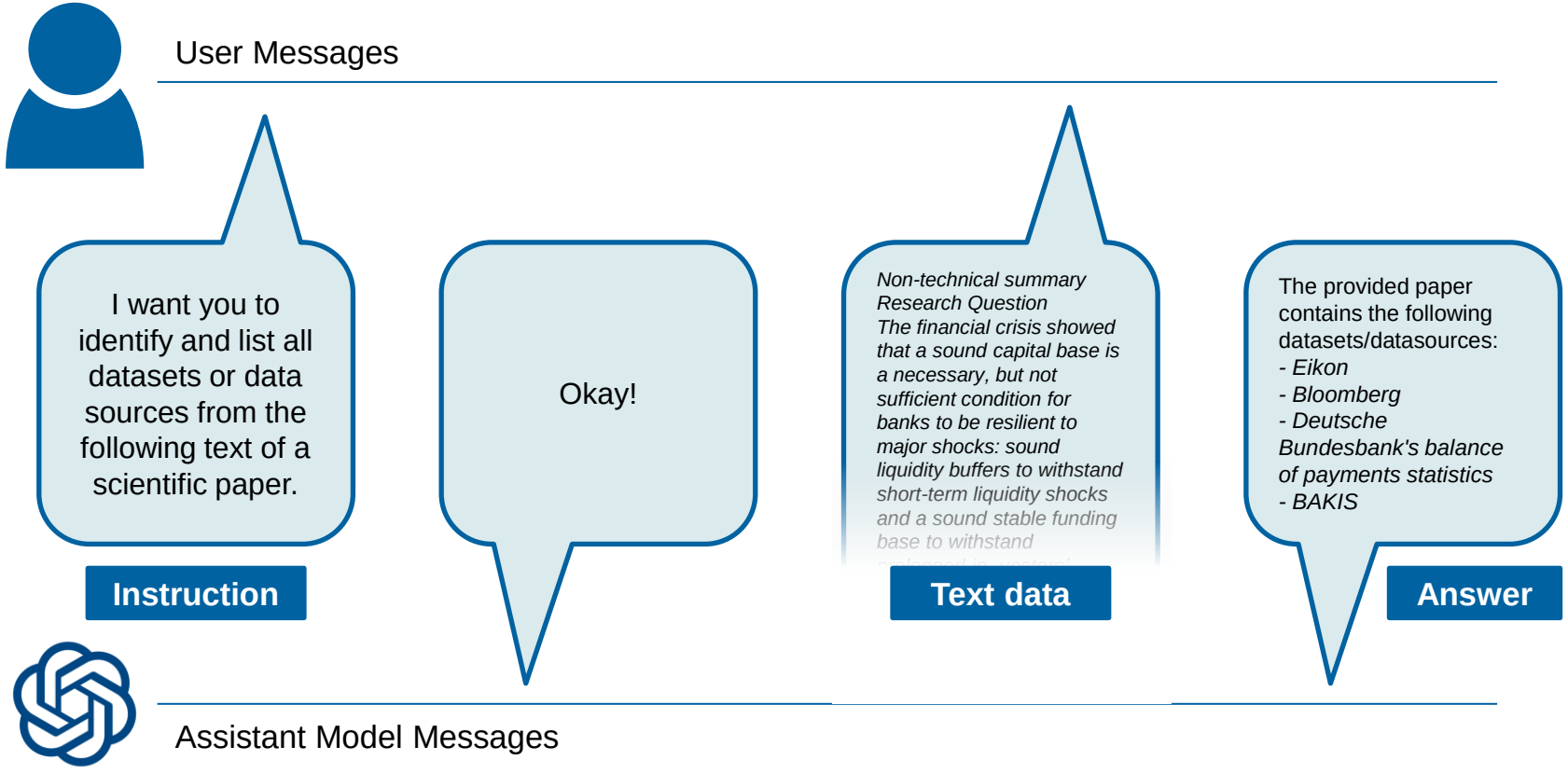


Our Contribution

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



Paper Corpus and Evaluation Sample

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

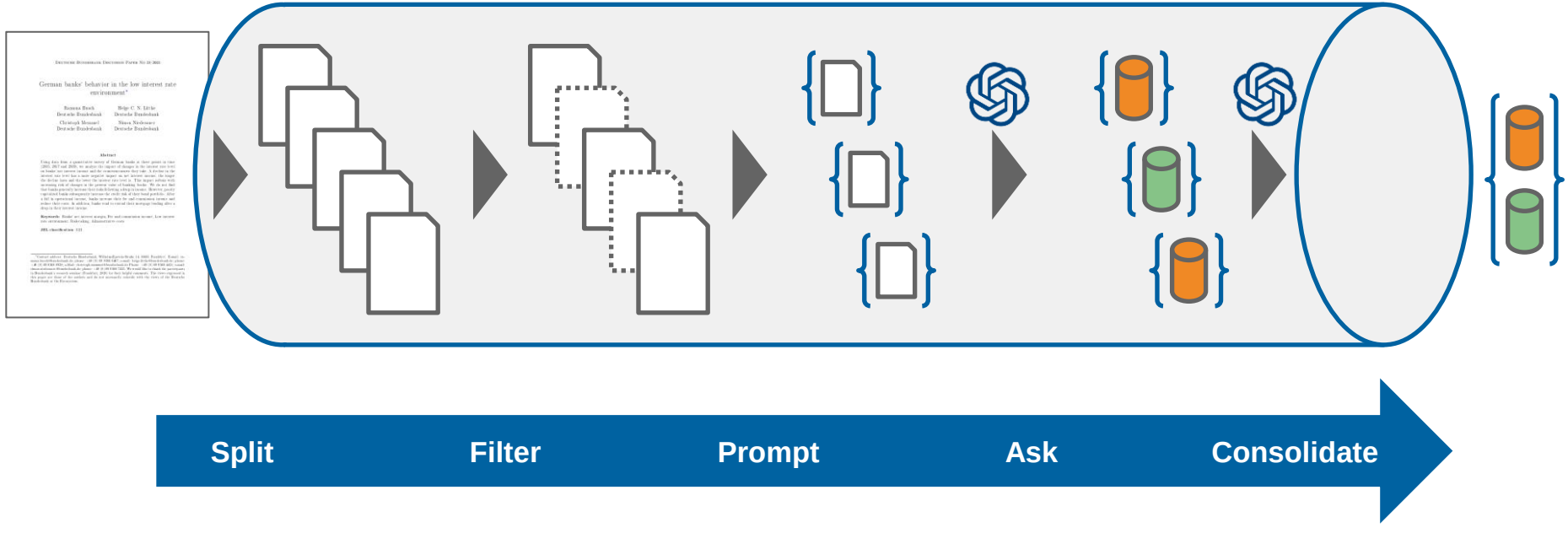
Dataset Citation Extraction Pipeline

Input

Academic Papers

Output

Data Sources



Assessing the Quality of the Assistant-Model's Output

Input
- Academic Papers
- Labels: Data Sources

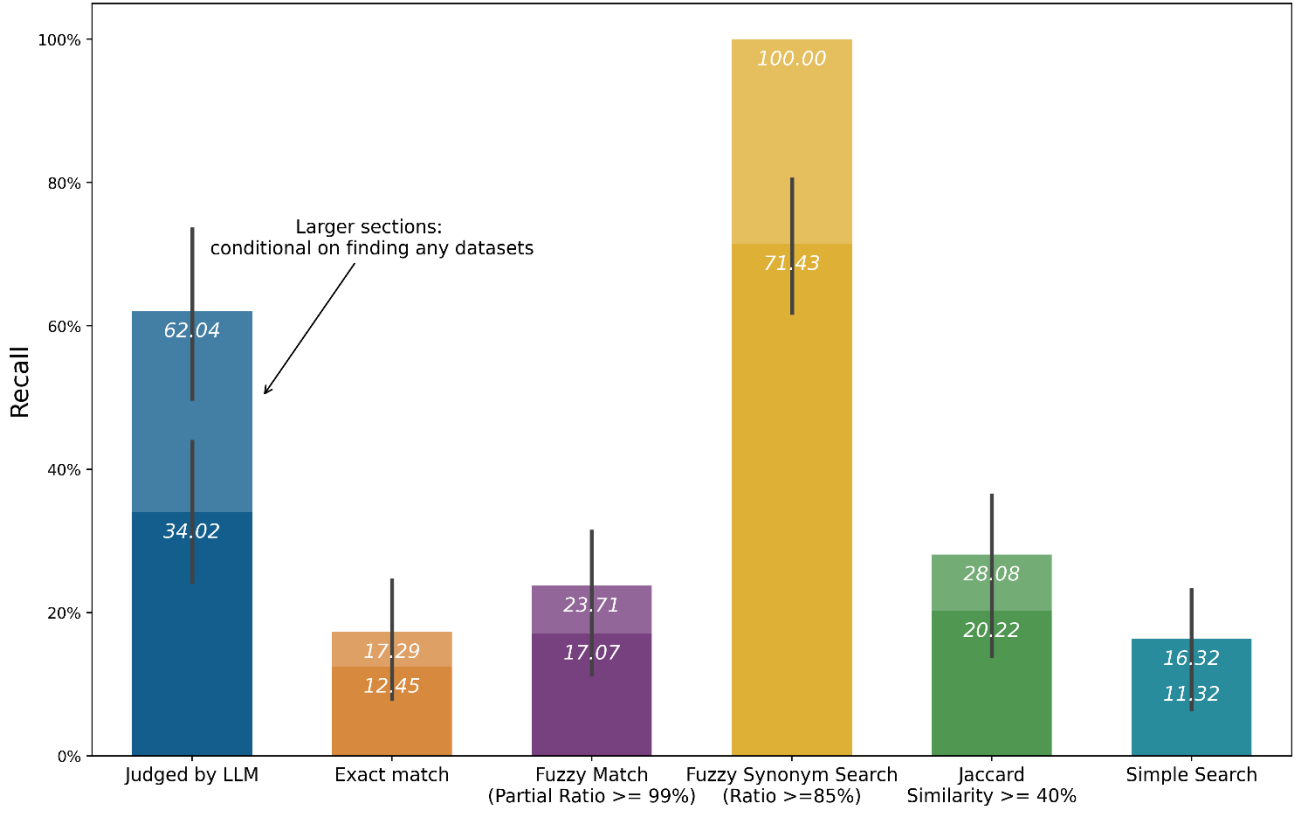
Output
- Data Sources
- Quality Measures

Pipeline

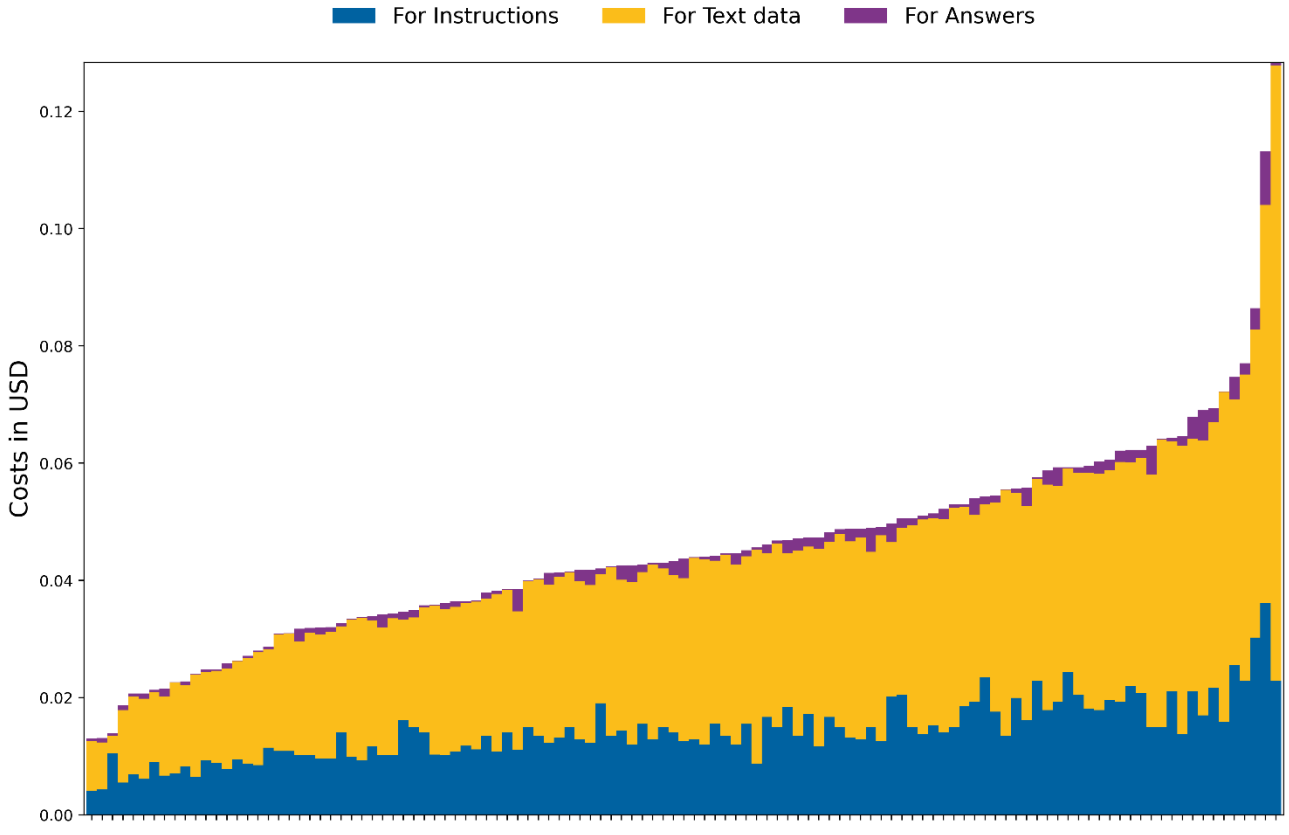
Manual labelling

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

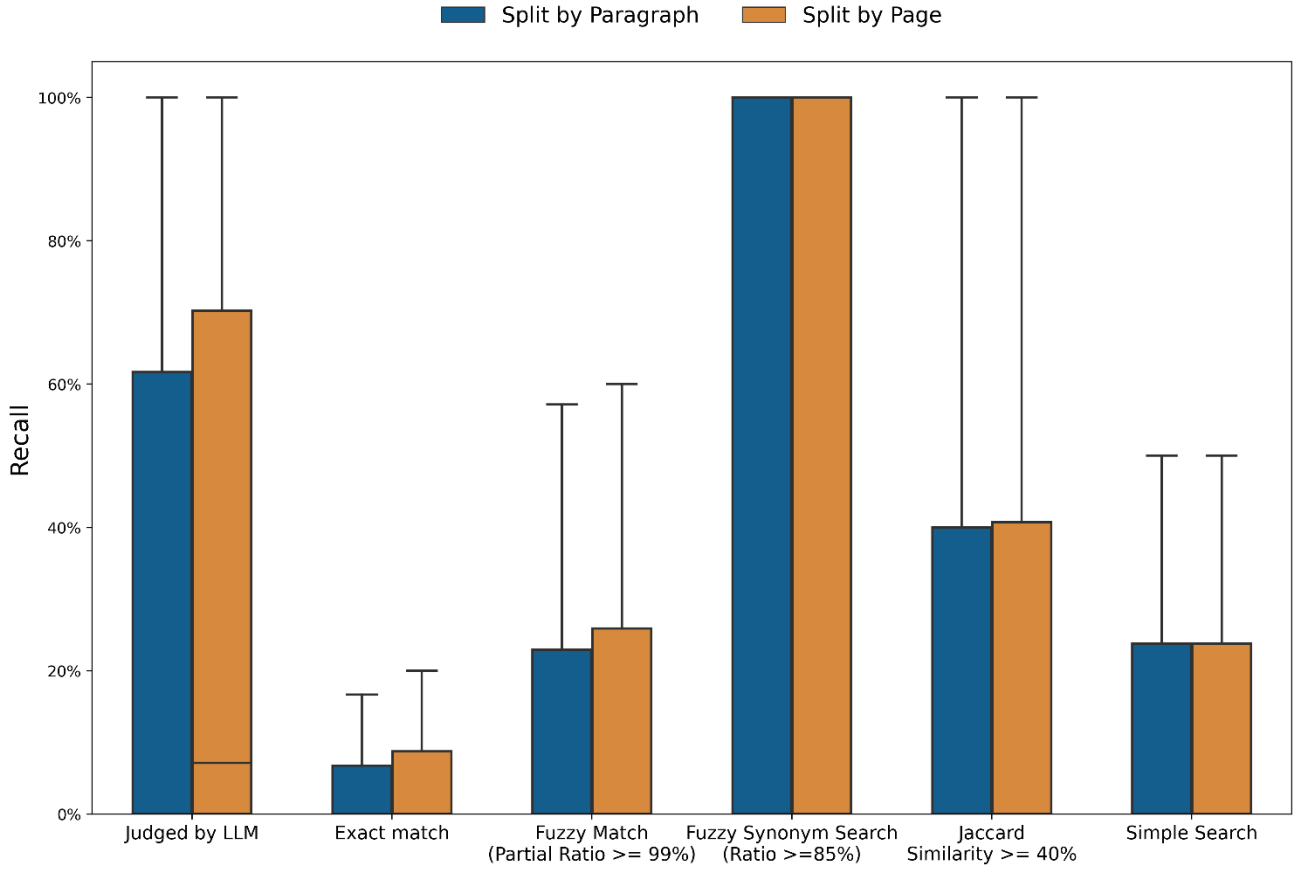
Results
Different measures of apparent performance



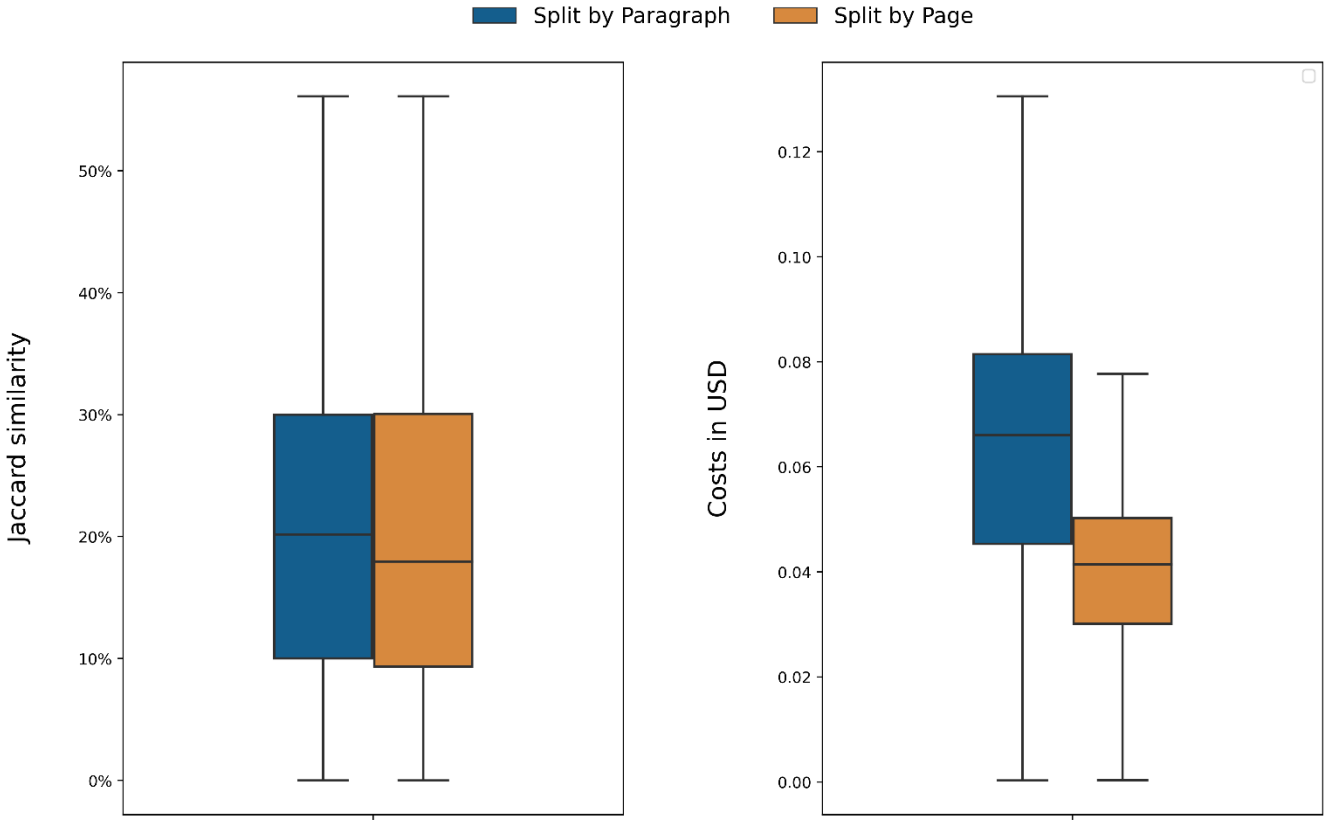
Results
API Costs per paper processed



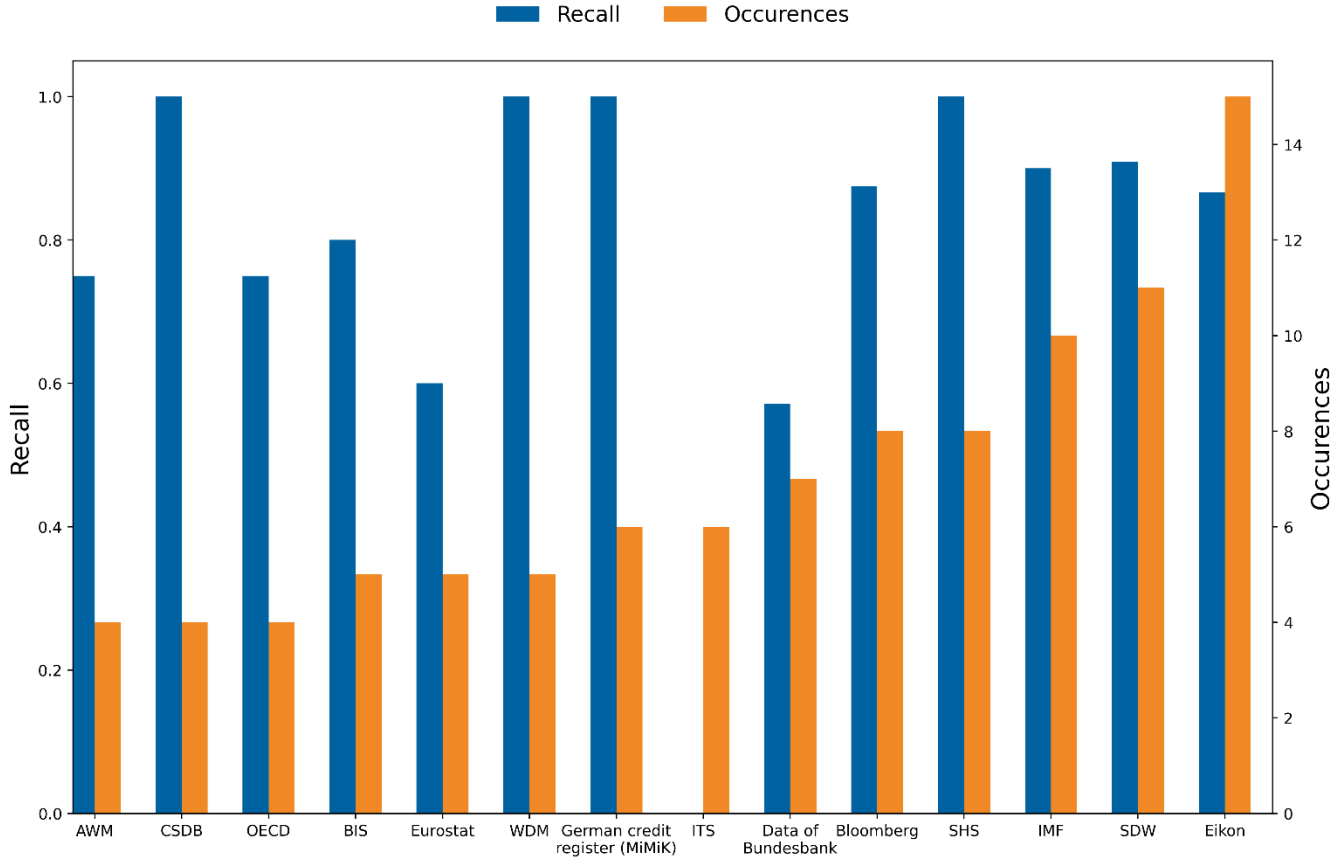
Results
Larger Contexts provide minor performance improvement



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

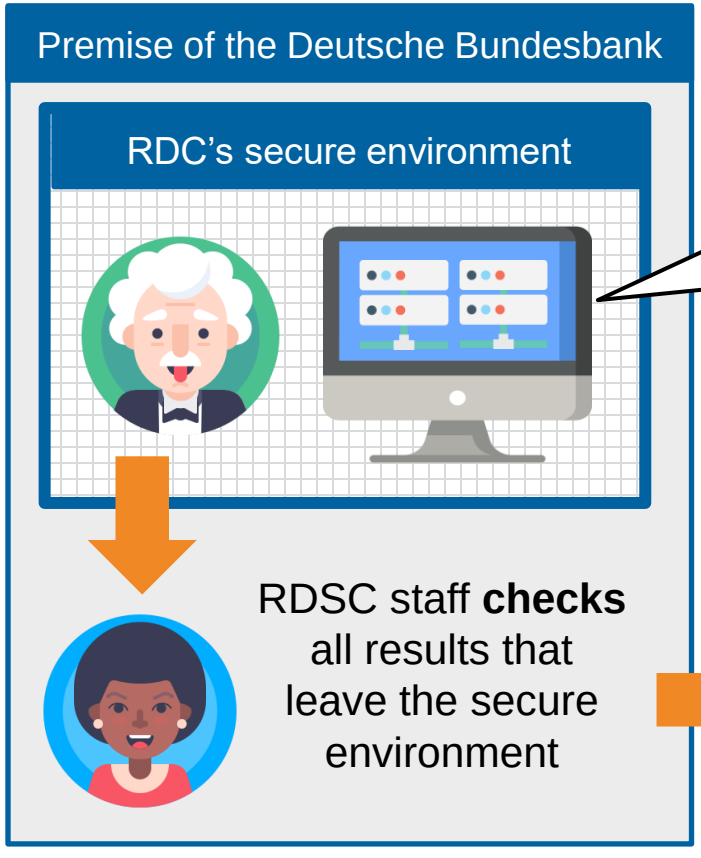
Give the model clear and simple tasks, yet as much context as possible

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

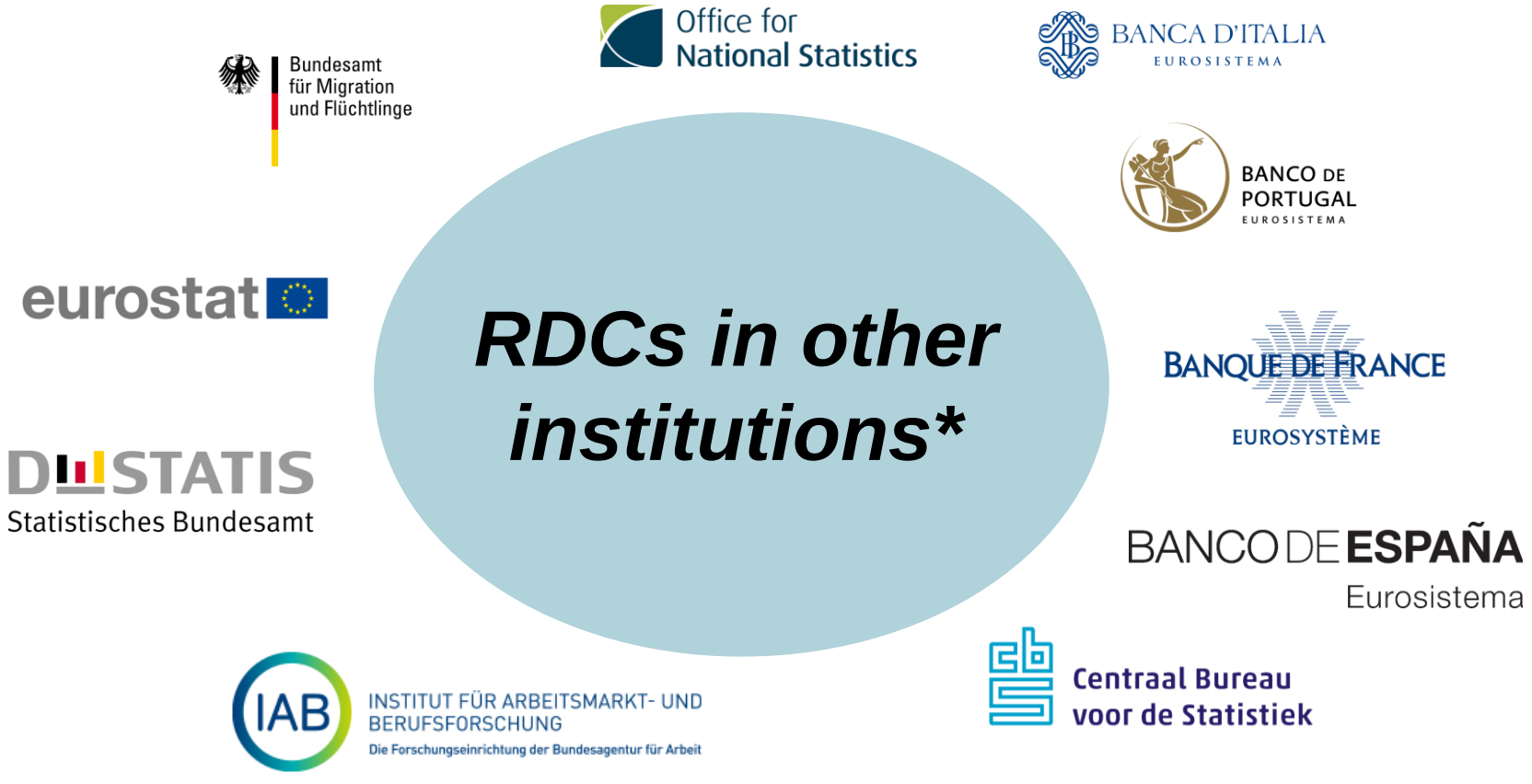
Appendix: RDCs allow to securely share confidential granular data from administrative sources with external researchers



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

Research Data and Service Centre

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



** selected examples*

Research Data and Service Centre

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>}.
```