Large Language Models for Research Data Management?!

Report of Contributions

https://www.conferences.uni-hamburg.de/e/LLMs4RDM

Contribution ID: **1** Type: **not specified**

Verbalisation Process of a RAG-Based Chatbot to Support Tabular Data Evaluation for Humanities Researchers

Thursday 18 September 2025 12:00 (20 minutes)

Scholars have access to large amounts of data and publications stored in RDRs (Research Data Repositories). LLMs (Large Language Models) can efficiently work with textual data. But, since LLMs are pretrained and have a limited context window, they cannot work with large amounts of text. For this, the standard approach is to use RAG (Retrieval Augmented Generation), where an embedding space is built for the text corpus. During answering, the nearest suitable texts are extracted and provided to the context of the LLM. However, data in tables is not evaluated correctly because the embedding treats the tabular data as textual and thus fails to correctly model the semantics, which represents the context, of the tabular data. In this article, we show how tabular data can be used in a RAG-like approach: The key steps are i) a static cloze text is generated and then modified once by an LLM and ii) presented to the scholar for possible modifications. Afterwards, iii) the whole data set is verbalised according to the cloze text and, therefore, iv) usable for RAG. In particular, step iii) is crucial for our system as we add the missing context to the data. Our feasibility study shows how to efficiently generate a chatbot with a large amount of structured data.

Authors: ASSELBORN, Thomas (Universität Hamburg); BENDER, Magnus (Aarhus University); MARWITZ, Florian (Universität Hamburg); MÖLLER, Ralf (Universität Hamburg); MELZER, Sylvia (Universität Hamburg)

Presenters: ASSELBORN, Thomas (Universität Hamburg); BENDER, Magnus (Aarhus University)

Contribution ID: **2** Type: **not specified**

Talk to your database: An open-source in-context learning approach to interact with relational databases through LLMs

Thursday 18 September 2025 11:40 (20 minutes)

Abstract
With the emergence of large language models, the long studied field of the Text-to-SQL problem was elevated
into new spheres. In this paper, we test how our LLM fine-tuning approach performs on two relational databases
(small vs. big) and compare it to a default setting. The results are convincing: using in-context learning boosts the
performance from a merely 35% (default) to over 85%. Furthermore, we present a detailed architectural framework
for for such a system, emphasizing its exclusive reliance on open-source components.

Authors: PLAZOTTA, Maximilian (Universität Regensburg); Prof. KLETTKE, Meike (Universität Regensburg)

Presenter: PLAZOTTA, Maximilian (Universität Regensburg)

Contribution ID: **3** Type: **not specified**

Improving Accessibility and Reproducibility by Guiding Large Language Models

Thursday 18 September 2025 11:20 (20 minutes)

Research data repositories store numerous entries of research data, to among other advantages one goal is allowing to store us all data to reproduce experiments.
Working with large corpora of texts is made significantly easier with Large Language Models.
However, Large Language Models are trained for general purposes and are note finetuned for the data originating from different kinds of projects.
But the creators of such texts have an expert viewpoint on the data.
Therefore, we propose to leverage the expert viewpoints of creators to obtain better answers from a Large Language Model.
When creating an entry for the Research Data Repository, the creators have the possibility to add a so-called interpretation prompt.
The interpretation prompt contains their expert viewpoint and be of any textual form to guide the Large Language Model to interpret the project-specific data.
In particular, the interpretation prompt may contain instructions on how to reproduce experiments right inside the LLM invocation.
Afterward, the interpretation prompt is prepended to the query of the Large Language Model.
In our examples, we show how the interpretation prompt helps to receive more tailored answers.

Authors: MARWITZ, Florian (Universität Hamburg); GEHRKE, Marcel (Universität Hamburg)

Presenter: MARWITZ, Florian (Universität Hamburg)

Contribution ID: **4** Type: **not specified**

Large Language Models in Labor Market Research Data Management: Potentials and Limitations

Thursday 18 September 2025 10:10 (20 minutes)

This contribution explores the application of large language models (LLMs) in labour market research data management, particularly in occupational data analysis. Based on our empirical studies of the automated classification of job titles and critical evaluations of AI-assisted text interpretation, we contend that, although LLMs present promising opportunities to improve research processes, such as providing query assistance, offering annotation support, and facilitating preliminary content structuring, they are inadequate for consistent data management, reliable analysis, and interpretative depth. Our findings suggest that, while LLMs can support research workflows as interactive tools, they cannot replace methodological approaches in data-driven social science research. Our aim is to contribute to the discussion at the workshop on the scope and boundaries of LLM-based tools in research data management.

The increasing availability of large language models (LLMs) creates new opportunities and challenges for research data management (RDM), especially when dealing with complex and diverse data sources, such as labour market information. We build on documents from the German labour market archive containing data on vocational education and training (VET) and continuing VET (CVET). The archival form of these regulations, primarily unstructured or semi-structured scanned documents, poses challenges for digital accessibility, analysis and integration with contemporary data systems, as described in \cite{reiser2024towards}. However, the digitisation of archival material provides an opportunity to preserve, structure and analyse regulatory knowledge in a form that is compatible with semantic linking, machine learning and long-term data curation, as discussed in our previous work \cite{reiser2024analyzing,reiser2024learning}. This system incorporates a web-based information system \cite{reiser2025is} and a data warehouse backend \cite{hein2024linked}, along with various analysis pipelines.

Our research examines the integration of LLMs into two distinct areas of labour market studies: (1) the automated classification of job titles \cite{reiser2025ecai} and occupational data to ontologies like the GLMO \cite{dorpinghaus2023towards}; and (2) the analysis of texts related to labour, education and social discourse within a given hermeneutical framework \cite{hermen}. We draw on empirical studies using annotated survey data, synonym datasets, online job advertisements and vocational training records, as well as comparative experiments assessing the interpretative capabilities of LLMs across various text genres.

Our findings suggest that, although LLMs can be effective interactive tools that assist with tasks such as content summarisation, query reformulation and preliminary data exploration, their performance in core data analysis tasks is inconsistent and unreliable, making them unusable for most scientific purposes. Specifically, LLMs fail to deliver reproducible results in classification tasks; for instance, at fine-grained levels of occupational coding \cite{2025dorau,reiser2025ecai}. In hermeneutic contexts, models are highly sensitive to prompt design, language and model architecture, which undermines their suitability for structured analysis or theory-driven interpretation \cite{hermen}. These limitations emphasise the risk of overestimating LLMs' capabilities in domains requiring specific methods, domain knowledge and theoretical grounding. Nevertheless, this contradicts some recent literature which asserts that LLMs "can analyze nearly any textual statement." \cite{tornberg2023use} The experimental results obtained in this study appear to support the arguments of researchers who claim that LLMs are incapable of performing even basic logic-based tasks, such as counting and identifying general substructures in graphs, see, for example \cite{fu2024large,nguyen2024evaluating}. Therefore, it might be debatable whether LLMs could offer any technical assistance with textual analysis. This could be against, for example, \cite{tai2024examination}.

We argue that the primary value of LLMs in RDM for labour market research lies in their potential to improve interaction between researchers and data by supporting hypothesis generation, assisting with data annotation and facilitating interdisciplinary dialogue, rather than automating analytical processes or replacing established methods, as LLMs have been shown to have difficulty understanding text \cite{saba2024llms}, and they also appear to lack an understanding of context, intentionality, and reader-writer dynamics. This distinction is crucial for the responsible integration of AI tools into research workflows without compromising scientific standards.

In conclusion, we advocate for the cautious and differentiated use of LLMs in labour market research. Rather than viewing LLMs as a universal solution for data analysis, we suggest positioning them as supportive tools that complement human expertise during the exploratory and communicative phases of research. We invite discussion on how LLMs might augment methodological rigour in data-intensive research fields, rather than substitute it, and on the development of evaluation criteria for their responsible application in RDM.

Authors: DÖRPINGHAUS, Jens (University of Koblenz, Federal Institute for Vocational Education and Training (BIBB)); Mr TIEMANN, Michael (University of Koblenz, Federal Institute for Vocational Education and Training (BIBB))

Presenters: DÖRPINGHAUS, Jens (University of Koblenz, Federal Institute for Vocational Education and Training (BIBB)); Mr TIEMANN, Michael (University of Koblenz, Federal Institute for Vocational Education and Training (BIBB))

Contribution ID: **5** Type: **not specified**

Challenges in Automatic Speech Recognition in the Research on Multilingualism

Thursday 18 September 2025 11:00 (20 minutes)

This paper explores the potential of using large language models in multilingualism research to accelerate data processing (speech-to-text). The main issues relating to the language of bilingual individuals are discussed qualitatively using a Polish-German recording as an example.

Authors: JURKIEWICZ-ROHRBACHER, Edyta (Universität Hamburg); ASSELBORN, Thomas (Universität Hamburg)

Presenter: ASSELBORN, Thomas (Universität Hamburg)

Contribution ID: **6** Type: **not specified**

Welcome

Thursday 18 September 2025 09:00 (10 minutes)

Introduction to the first workshop on Large Language Models for Research Data Management?!

Research data management (RDM) has become an important discipline that enables researchers to effectively organise, preserve and share their research results.
RDM is a new development that aims to prepare researchers for the future by building on the principles of open science. It utilises innovative approaches such as generative artificial intelligence (genAI), which is powered by large language models (LLMs), to complement traditional research methods.
As data-driven research becomes increasingly complex, researchers often have to spend a lot of time learning how to manage, analyse and interpret large amounts of information. Traditional data literacy training can be time-consuming and doesn't always keep pace with evolving technologies and methods of analysis.
Foundation models based on generative AI offer the potential to streamline this learning process. By automating data pre-processing, pattern recognition and even hypothesis generation, these models can lower the technical barriers to entry, allowing researchers to focus more on insights and discovery rather than spending excessive amounts of time mastering data skills.
The objective of this workshop is an exchange of perspectives regarding the implementation of novel RDM approaches using LLMs or not, both past and prospective, in research and practice.

Presenter: BENDER, Magnus (Aarhus University)

Contribution ID: **7** Type: **not specified**

Keynote: Advancing RDM: From Immersion to Argumentation in Science

Thursday 18 September 2025 09:10 (1 hour)

Presenter: MÖLLER, Ralf

Contribution ID: **8** Type: **not specified**

Farewell

Thursday 18 September 2025 12:20 (10 minutes)

Presenter: BENDER, Magnus (Aarhus University)