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

Jens Dörpinghaus[1,2,3,*,†], Michael Tiemann[1,2,*,†]

[1]University of Koblenz, Department of Computer Science, Germany
[2]Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany
[3]Linnaeus University, Department of Computer Science and Media Technology, Växjö, Sweden

Abstract

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 on the scope and boundaries of LLM-based tools in research data management.

Keywords

Text analysis, labor market data, data linking, large language models, LLMs, research data management, RDM, computational social sciences

1. Introduction

The growing accessibility of large language models (LLMs) seems to be creating new opportunities and challenges for research data management (RDM), especially when dealing with intricate and diverse sources, such as labour market information. This study is based on documents from the German labour market archive containing data on vocational education and training (VET) and continuing vocational education and training (CVET).

The archival form of these regulations — often unstructured or semi-structured scanned documents — poses difficulties for digital accessibility, analysis, and integration with modern data systems, as outlined in [1]. However, digitising such materials enables us to preserve, structure and analyse regulatory knowledge in a way that supports semantic linking, machine learning and long-term data curation, see [2, 3]. Our current work builds on this research by integrating a web-based information system [4] and a linked data warehouse backend [5], both of which serve as a foundation for automated and semi-automated analytical workflows.

RDM covers the entire research data lifecycle, see [6]. While some of these processes are automatable, LLMs do not always fit naturally for such purposes. However, we might also ask questions such as 'How can LLMs be used for planning research projects?' and 'How can LLMs be used for databank management'? Nevertheless, our primary focus remains on generating, analysing and processing research data within the RDM framework.

The integration of LLMs into RDM practices presents both a methodological opportunity and an epistemological challenge. Our research investigates how and whether these models can be incorporated responsibly into data-intensive research on labour market dynamics.

Large Language Models for Research Data Management?!
*Corresponding author.
†These authors contributed equally.
doerpinghaus@uni-koblenz.de (J. Dörpinghaus); tiemann@bibb.de (M. Tiemann)
0000-0003-0245-7752 (J. Dörpinghaus); 0000-0001-7136-2744 (M. Tiemann)

2. Overview

The study is guided by a set of interrelated questions concerning the technical, methodological, and interpretative potential of LLMs in labour market research. Specifically, we explore how LLMs could assist in classifying job titles and occupational data within established ontological frameworks, such as the German Labour Market Ontology (GLMO, see [7]). We also explore the extent to which these models can be used for hermeneutic interpretation of texts related to labour, education, and social discourse, as discussed in [8].

Additionally, the study examines the reliability and reproducibility of LLM outputs in structured data analysis tasks and explores whether these models can meaningfully support research activities by facilitating hypothesis generation, assisting with annotation, and improving the interaction between researchers and data. The broader objective is to clarify the role that LLMs can play within established scientific workflows without compromising methodological rigour or theoretical coherence.

3. Methods and Results

Labour market data comprises very specific tasks, see [9, 10, 11, 12]. Examples include the automated classification of job titles and their linking to ontologies (e.g. GLMO), and the automated classification of industrial sectors (e.g. WZ08).

This research combines empirical experimentation with conceptual analysis. Two main applications of LLMs were examined: the automated classification of occupational data, and the interpretative analysis of labour-related texts. Data sources include annotated survey datasets, synonym collections, online job advertisements, and vocational training records. Each dataset was used to test LLMs abilities to perform tasks such as semantic classification, keyword extraction, and contextual interpretation.

For classification tasks, several models and prompt designs were evaluated against established coding systems, particularly the GLMO [13]. While LLMs frequently produced contextually plausible outputs, their results lacked reproducibility and their performance degraded at more granular classification levels, see [14, 13]. Even when supplied with controlled inputs, the outputs of the models varied across repeated trials, highlighting their sensitivity to prompt formulation, sampling parameters, and model architecture.

In hermeneutic analyses, LLMs demonstrated the ability to summarise content, reformulate queries, and identify thematic clusters; however, their interpretative consistency was limited. Small changes in prompt wording or language framing led to divergent interpretations, reflecting broader critiques of the instability of LLMs in theory-driven analysis [8], These findings are consistent with studies demonstrating that LLMs struggle to perform reliable logic-based operations [15, 16] and tend to overgeneralise when faced with complex textual data [17].

3.1. Classification of Labor Market Data

We used linkage techniques to match different German educational and labour market data. For example, lists of industries and their branches are widely used in economic and sociological research to categorise data and gain an overview of different types of industry, see [18, 19]. However, due to different research foci, as well as different national scenarios and interests, many different taxonomies and ordering schemas exist. In this paper, we will focus on regional data from Germany without loss of generality. The manual annotation of textual data is time-consuming and tedious, naturally leading to the question of whether we can automatically categorise textual data (e.g. job advertisements or business profiles) by industrial sector. We presented an approach to classification using a pre-trained, domain-adapted Transformer model, and found that domain-adapted models generalise better, outperforming state-of-the-art, non-domain-adapted Transformer models on out-of-distribution data.

Other work was carried out on linking vocational education and training (VET) records from two datasets, see [14]: the Genealogy of Vocational Education and the Occupations Archive at the German Federal Institute for Vocational Eduacation and Training (BIBB), which contains archival documents

on training regulations. This task could not be solved using LLMs alone. Instead, a labelled subset of 474 record pairs, consisting of 310 matches and 164 non-matches, was created. Eighty percent of these were used for training and the remaining 20 percent for testing. Three classifiers were trained and evaluated: Support Vector Machine (SVM), Logistic Regression and K-Means Clustering. The results revealed distinct performance characteristics across the classifiers. SVM achieved balanced performance, with a precision and recall of 91%, yielding the highest F-score. This demonstrates its robust capability in identifying true matches while maintaining reasonable specificity. Logistic regression provided intermediate results across all metrics, offering a practical trade-off between precision and recall. K-means clustering demonstrated the highest precision (97%) and specificity, indicating strong performance in correctly identifying matches and non-matches when predictions were made. However, it had the lowest recall at around 80%, resulting in many false negatives and the lowest overall accuracy at approximately 85%. All classifiers achieved an accuracy of over 85%, with both SVM and logistic regression reaching approximately 88%. The variation in precision and accuracy across the classifiers was modest at around 6%, while the most substantial differences emerged in their ability to recognise non-matches, as reflected in the specificity measurements. The class imbalance in the training data, with nearly twice as many matches as non-matches, likely contributed to K-Means' lower recall performance. For the specific application of identifying archival gaps, SVM's higher recall makes it preferable, as finding more true matches is prioritised over perfect precision in this context.

Another task was the automated classification of job titles, a critical component of labour market research, survey analysis and administrative data processing. In [13, 14], we presented two studies exploring the classification of German job titles according to the German Classification of Occupations (KldB), emphasising the linguistic and structural challenges inherent in this task. The study used different training and annotation data. This study's findings demonstrated that, while substantial results can be achieved for broad occupational categories, fine-grained classification, particularly at the level of requirements (5th digit), remains challenging. In [14, 19], we also showed that applying LLMs does not improve performance and that they are even outperformed by other ML approaches, even without fine-tuning. These findings highlight the limitations of relying solely on job titles and emphasise the importance of richer contextual information and more expressive models. Therefore, using LLMs to link data in a specific domain such as labour market research does not improve the quality of RDM.

3.2. Hermeneutics

In [8], we presented a study investigating whether large language models can apply hermeneutical methods from philosophy, theology, sociology and literary studies to textual interpretation in a meaningful way. We selected a variety of hermeneutical approaches with formalisable analytical rules which is under constant research, see for example [20, 21]. Although these approaches share a fundamental understanding of human nature and language, each one emphasises different interpretative dimensions, including context sensitivity, reader-writer relationships and spiritual or social aspects that extend beyond linguistic analysis alone.

Our experimental design involved testing four contemporary LLMs (ChatGPT-4, Llama3:70b, Qwen2.5:72b and Mistral:7b) on their ability to perform two fundamental hermeneutical tasks: identifying plot structures by segmenting text into meaningful sections and recognising actors or characters within texts. Each text was analysed using two prompting strategies to assess consistency, and some texts were also tested in English translation to examine language dependency.

The results revealed substantial inconsistencies across all examined dimensions. These findings demonstrate that three primary factors influence LLM output in hermeneutical tasks: model architecture, prompt design and input language. Notably, the variability within individual models when faced with identical queries repeatedly indicates that LLMs do not apply coherent, reproducible hermeneutical frameworks. The dramatic differences between models suggest that they employ fundamentally different, non-comparable approaches rather than systematic hermeneutical methods. Language dependency reveals that interpretative outputs are not based on stable textual understanding, but rather vary according to linguistic surface features.

These results have direct implications for the use of LLMs in RDM. While LLMs can be useful for providing interactive support with preliminary tasks such as query assistance, annotation suggestions or initial content structuring, they lack the consistency and interpretative depth required for reliable data analysis, systematic classification or meaningful text interpretation in research contexts. The variability in detecting even basic structural elements such as actors and sections indicates a fundamental limitation in the consistent application of methodological frameworks. Regarding the occupational data analysis and job title classification discussed in our abstract, these findings imply that LLMs cannot substitute methodological rigour in data-driven social science research. While they may serve as assistive tools in research workflows, their outputs require critical human evaluation and cannot be trusted for autonomous data management decisions. Their lack of reproducibility particularly undermines their utility for systematic RDM, where consistency, traceability and methodological transparency are paramount. Our findings reinforce the view that LLMs should complement, rather than replace, human expertise in research data curation and interpretation.

4. Conclusions and Outlook

In this study, we reviewed previous research to investigate the potential use of large language models in RDM. This study examined the potential and limitations of LLMs in RDM, with a particular focus on labour market data, through two complementary empirical investigations: the automated classification of occupational data and the hermeneutical interpretation of labour-related texts. Our findings provide differentiated recommendations for the responsible integration of LLMs into research workflows.

The empirical evidence reveals a consistent pattern across different RDM tasks. When classifying historical VET records, for example, traditional machine learning approaches (SVM and logistic regression) demonstrated superior and more reliable performance than LLM-based methods. This challenges the common assertion that poor performance is simply due to inadequate prompt engineering, as our systematic experimentation with various prompting strategies failed to achieve comparable reproducibility.

In hermeneutical analysis, variability was even more pronounced. Models produced inconsistent interpretations of identical texts when queried repeatedly, with outputs that were highly sensitive to prompt formulation, model architecture and input language. The inability to reliably identify basic structural elements, such as narrative sections or characters, demonstrates the fundamental limitations of applying systematic interpretative frameworks. These findings align with broader evidence that LLMs struggle with logic-based operations and structural inference tasks, suggesting that their limitations extend beyond domain-specific challenges to encompass more fundamental computational constraints.

Based on our empirical observations, we propose a different assessment of the utility of LLMs for specific RDM tasks in labour market research. For instance, LLMs can demonstrably improve researcher-data interaction through natural language interfaces for query assistance and reformulation. LLMs could also be used to provide interactive support for hypothesis generation and exploratory analysis, particularly in the early stages of research, where flexibility is more important than reproducibility. LLMs might also be used for preliminary annotation and candidate suggestion for manual validation. They could also be used for initial content structuring and summarisation of archival materials. LLMs could also be used for metadata extraction from clean, well-structured texts. In all these cases, human oversight remains essential and the outputs should be treated as suggestions that require expert validation rather than as authoritative classifications. LLMs cannot be used for fine-grained classification tasks requiring high precision, such as detailed occupational coding at Klassifikation der Berufe (KldB) level 5. They are not suitable for theory-driven hermeneutical analysis demanding methodological depth and interpretative consistency. Thus, we strongly advise that any application requiring reproducible and scientifically defensible outputs should not be used without extensive human verification.

The central implication for RDM in labour market research is that LLMs cannot replace established methodological approaches, but they can augment specific workflow components when deployed appropriately. Instead, our results support a more cautious perspective, emphasising the known

weaknesses of LLMs in logical reasoning, structural inference and the consistent application of analytical frameworks. The fundamental issue is not merely technical capability, but epistemic adequacy: LLMs lack the contextual understanding, intentionality and reader-writer dynamics that constitute authentic hermeneutical engagement. We therefore advocate a hybrid approach, positioning LLMs as assistive instruments within human-centred research workflows, rather than as autonomous analytical agents.

There are several open questions that merit further investigation. For example, how can LLMs be effectively integrated into ontology-driven, provenance-rich RDM pipelines while preserving scientific standards? Which hybrid combinations of rule-based methods, traditional machine learning (ML) classifiers and LLMs yield robust, reproducible outcomes for specific task classes? How should LLM behaviour be benchmarked for domain-specific applications such as occupational coding and hermeneutical analysis, in order to enable meaningful comparative evaluation?

Additionally, the development of evaluation standards for machine-assisted exegesis requires interdisciplinary collaboration between computer scientists, domain experts and methodologists. The question of whether reproducibility constitutes a key element of hermeneutics itself deserves deeper philosophical and methodological examination, as do the epistemological implications of using inherently probabilistic systems in contexts that demand deterministic, theory-based interpretation.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Reiser, J. Dörpinghaus, P. Steiner, M. Tiemann, Towards a datatset of digitalized historical german vet and cvet regulations, Data 9 (2024).

[2] T. Reiser, J. Dörpinghaus, P. Steiner, Analyzing historical legal textcorpora: German vet and cvet regulations, in: INFORMATIK 2024, Gesellschaft für Informatik eV, 2024, pp. 2007–2018.

[3] T. Reiser, J. Dörpinghaus, P. Steiner, Learning from historical vet and cvet regulations in germany: What should vet look like and whom should it serve?, in: NORDYRK 2024 BOOK OF ABSTRACTS, 2025, p. 75.

[4] T. Reiser, J. Dörpinghaus, P. Steiner, Analyzing historical german vet textcorpora: A novel information system, in: Nordyrk Conference 2025 Book of abstracts, 2025, pp. 126–127.

[5] K. Hein, Linked labor market data: Towards a novel data housing strategy, in: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), 2024, pp. 355–362.

[6] J. Dörpinghaus, J. Klein, J. Darms, S. Madan, M. Jacobs, Scaiview-a semantic search engine for biomedical research utilizing a microservice architecture., in: SEMANTiCS (Posters & Demos), 2018.

[7] J. Dörpinghaus, J. Binnewitt, S. Winnige, K. Hein, K. Krüger, Towards a german labor market ontology: Challenges and applications, Applied Ontology 18 (2023) 343–365.

[8] J. Dörpinghaus, M. Tiemann, Do llms dream of antique hermeneutics? critical remarks on automated text interpretation, in: To appear: Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS), 2025.

[9] F. Derksen, J. Dörpinghaus, Digitalization and sustainability in german continuing education, Lecture Notes in Informatics (LNI) (2023).

[10] J. Dörpinghaus, D. Samray, R. Helmrich, Challenges of automated identification of access to education and training in germany, Information 14 (2023) 524.

[11] J. Dörpinghaus, M. Tiemann, Vocational education and training data in twitter: Making german

twitter data interoperable, Proceedings of the Association for Information Science and Technology 60 (2023) 946–948.
[12] J. Dörpinghaus, J. Binnewitt, K. Hein, Lessons from continuing vocational training courses for computer science education, in: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 2, 2023, pp. 636–636.
[13] T. Reiser, J. Dörpinghaus, M. Tiemann, Detecting occupations in german texts: Challenges and data, in: Proceedings of the 2nd International Workshop on AI in Society, Education and Educational Research (AISEER), 2025.
[14] R. Dorau, J. Dörpinghaus, M. Tiemann, Automated classification of german job titles according to kldb: Challenges and novel methods, in: To appear: INFORMATIK 2025, Lecture Notes in Informatics (LNI), 2025.
[15] T. Fu, R. Ferrando, J. Conde, C. Arriaga, P. Reviriego, Why do large language models (llms) struggle to count letters?, arXiv preprint arXiv:2412.18626 (2024).
[16] L. Nguyen, Y. Yan, Evaluating the structural awareness of large language models on graphs: Can they count substructures?, Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24) (2024).
[17] W. S. Saba, Llms' understanding of natural language revealed, arXiv preprint arXiv:2407.19630 (2024).
[18] R. Fechner, J. Dörpinghaus, A. Firll, Classifying industrial sectors from german textual data with a domain adapted transformer, in: 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), IEEE, 2023, pp. 463–470.
[19] R. Fechner, J. Dörpinghaus, No train, no pain? assessing the ability of llms for text classification with no finetuning, in: Proceedings of the Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, 2024, pp. 8–11.
[20] J. Dörpinghaus, J. Darms, M. Jacobs, What was the question? a systematization of information retrieval and nlp problems., in: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2018.
[21] J. Dörpinghaus, Die soziale netzwerkanalyse: neue perspektiven für die auslegung biblischer texte?, Biblisch Erneuerte Theologie 5 (2021) 75–96.