A Large Language Model approach to extract Biographical Transcripts

Dharshan Dhanashekar
Information and Media Design
SRH Hochschule
Heidelberg, Germany
dharshan.dhanashekar
@stud.hochschule-heidelberg.de

Harshita Jamadade
Information and Media Design
SRH Hochschule
Heidelberg, Germany
harshita.jamadade
@stud.hochschule-heidelberg.de

Navneeth Krishna Aravind
Information and Media Design
SRH Hochschule
Heidelberg, Germany
navneethkrishna.aravind
@stud.hochschule-heidelberg.de

***Abstract*—The extraction of structured metadata from biographical interviews is essential for efficient archival and research purposes. This project presents an AI-driven approach to automating metadata extraction from interview transcripts, using advanced natural language processing (NLP) techniques. The proposed system employs preprocessing techniques to clean and normalize textual data, followed by a dynamic segmentation strategy based on similarity-based segmentation to preserve contextual integrity. A LLM is then utilized to extract structured metadata, ensuring accurate identification of key biographical details such as names, dates, locations, and thematic events. The methodology improves data retrieval efficiency and enables researchers and archivists to systematically organize vast collections of interview data with minimal manual intervention. By integrating machine learning-based segmentation and LLM-driven extraction, this approach improves the accuracy and scalability of metadata processing. The results demonstrate that the use of contextual segmentation significantly improves the quality of extracted metadata by maintaining the coherence of biographical narratives. This study contributes to the advancement of AI-assisted digital archiving by offering an automated, scalable, and adaptable metadata extraction framework for biographical interviews.**

***Index Terms*—Metadata Extraction, Biographical Interviews, Large Language Model (LLM), Text Analysis.**

I. Introduction

The "German Memory" archive, housed at FernUniversität in Hagen, preserves a diverse collection of subjective memories, including biographical interviews, autobiographies, diaries, and letters. These materials focus on sociopolitical events in Germany and German history, offering unique personal perspectives. Since its establishment in the 1980s, the archive has been integral to contemporary history research, providing valuable resources for scholars worldwide. They also feature contributions from third-party researchers across various disciplines. While the majority of documents are in German, a significant portion is available in other languages, highlighting the international scope of the collection. The archive plays a key role in supporting research in contemporary history, sociology, and cultural studies. To enhance accessibility and efficiency, AI-driven tools for metadata extraction are being integrated. By automating the extraction of key themes and historical events, AI significantly improves the archival process, enabling deeper analysis and easier navigation of the extensive collection.

This project aims to make it easier to extract important details from biographical interview transcripts using AI. First, we create a system to clean and prepare the transcripts so that only useful information is kept. Then, we break the text into smaller, meaningful sections so that the AI model can process it more effectively. We use **Llama-3.3 70B**, a powerful language model, to pull out key details from the transcripts in a structured way.

II. Literature Review and Related Work

The objective of this project's literature review is to evaluate prior research in the area of automated metadata extraction and the application of large language models (LLMs) in textual analysis. This review aims to situate the proposed system within the existing body of research and technology, ensuring it addresses critical challenges while contributing meaningful advancements.

A. Large Language Models (LLMs)

To achieve the project's objectives, we conducted a comprehensive review of relevant literature, including academic papers and official documentation on various language models. Resources such as Hugging Face were also utilized to examine pertinent models. Based on our analysis, we identified several Large Language Models (LLMs) that align closely with the goals of the project.

A LLM is a type of artificial intelligence (AI) designed to understand and generate text, among other tasks. The term "large" refers to the vast amounts of data LLMs are trained on. These models are based on machine learning techniques, specifically neural network architecture known as a transformer model. In simple terms, an LLM is a computer program that has learned to recognize and interpret human language (or other complex data) by processing numerous examples. Often, LLMs are trained on extensive datasets gathered from the Internet, which can include gigabytes of text. The effectiveness of an LLM depends on the quality of the data it is trained on, so developers might use more selective datasets to improve performance. LLMs employ deep learning, a method that involves analyzing large amounts of unstructured data probabilistically. This allows the model to identify patterns and differences in content without direct human input. Once trained, LLMs can be further customized through a process called tuning. This involves adjusting the model to perform specific tasks, such as answering questions and generating text.

LLMs are versatile tools that can be trained to perform a variety of tasks, including metadata extraction from unstructured data sources. One of their most prominent applications is generative AI: they can analyze and generate structured metadata based on a given text input. For example, an LLM can process biographical interviews, extracting key metadata such as names, dates, locations, and themes, making archival and research tasks more efficient. A notable feature of LLMs is their ability to handle unpredictable queries and unstructured data. Unlike traditional metadata extraction tools, which rely on predefined rules and structured formats, LLMs can interpret and extract meaningful information from free-form text, adapting to various contexts.

B. LLaMA 3.3 70B Instruct Model

This is a multilingual LLM developed by Meta AI, designed to excel in instruction-following tasks across various languages. With 70 billion parameters, this model is fine-tuned for generative text application, particularly optimized for multilingual scenarios. It's an auto-regressive language model utilizing an optimized transformer architecture. The instruction-tuned versions employ supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align outputs with human preferences, enhancing helpfulness and safety. The model has been pre-trained on approximately 15 trillion tokens sourced from publicly available datasets. This extensive training corpus enables the model to understand and generate human-like text across multiple languages effectively. Evaluations indicate that LLaMA-3.3-70B-Instruct outperforms many existing open-source and proprietary chat models on standard industry benchmarks, demonstrating superior contextual understanding and response generation in multilingual settings. The model supports a context length of up to 128,000 tokens, enabling it to process and generate responses based on extensive context, which is particularly beneficial for tasks involving long-form content or complex conversations. LLaMA-3.3-70B-Instruct is designed to handle multiple languages, including English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai [8]. This multilingual support makes it versatile for global applications, allowing for effective communication across diverse linguistic contexts.

C. Application of LLMs in Metadata Extraction

Large Language Models (LLMs) have been widely applied in various fields, including automated metadata extraction, document classification, and structured data retrieval. Recent advancements have demonstrated their effectiveness in processing and understanding unstructured text at scale. For instance, the study by [14] explores the role of LLMs in extracting meaningful metadata from large textual datasets, showcasing how deep learning models can streamline information retrieval processes and enhance the efficiency of knowledge management systems.

Furthermore, LLMs have also been utilized in biomedical and scientific research domains, where they assist in extracting structured insights from large-scale textual data. The research by [15] highlights the impact of LLMs in analyzing and organizing biomedical literature, enabling researchers to efficiently retrieve and categorize vast amounts of scientific information. This application demonstrates how LLMs can enhance knowledge discovery by automating the structuring of complex textual data in research-intensive fields.

In climate science, the integration of LLMs has proven to be valuable for improving metadata extraction and standardization across diverse environmental datasets. The study by [17] introduces an LLM-based tool designed to extract and harmonize metadata from climate research repositories. Given the increasing volume of climate data, this solution addresses inconsistencies in observation parameters, units, and definitions, enhancing data interoperability. By leveraging the adaptability of LLMs to understand contextual nuances, this approach contributes to the establishment of a standardized metadata schema aligned with FAIR principles, facilitating more effective interdisciplinary climate research.

Beyond scientific and climate research, LLMs have also found critical applications in the medical domain. The study by [18] examines how LLMs can be applied to medical research repositories for extracting structured information from clinical studies and patient records. By automating the categorization and retrieval of key medical findings, LLMs assist in improving knowledge accessibility for healthcare professionals. Their ability to process large volumes of clinical data enables better decision-making and enhances the efficiency of medical research workflows.

D. Prompt Engineering

As part of our research into prompt engineering, we reviewed comprehensive documentation [6]. offers strategies and tactics to optimize the performance of large language models. The guide outlines six primary strategies for designing effective prompts that can significantly enhance the quality and relevance of generated outputs. As we are using LLM's to extract metadata, finding and writing the appropriate prompt is key. For this, we had found out about the key roles involved in prompting, the different methods of prompting like writing clear instructions, providing a reference text, splitting up of complex tasks, feedback and experimentation to get to the right prompt. The types of patterns like Zero-Shot and Few-Shot learning, Instruction Based, Question-Answer based, and the limitations of prompting to determine the correct prompt for our use case. These approaches are crucial for optimizing metadata extraction, with clarity and specificity in prompts being paramount. For instance, using explicit directives like "Extract the following metadata: Title, Author, Publication Year. Format as JSON" helps ensure accurate results. Integrating reference texts or examples can guide the LLM, especially for complex or variable formats. The iterative process of feedback and experimentation allows for fine-tuning prompts based on initial outputs, addressing any misalignments or missed information.[6] [19]

III. Data Overview

The raw data consists of interview transcripts stored in a tab-separated format, which are then processed and structured to include various categories of metadata.

A. Input Data Description

The input data consists of interview transcripts stored in tab-separated values (TSV) files with 12 columns. It contains spoken text, timestamps, speaker information, translations, and annotations. The dataset has the following key columns:

TABLE I
Key Columns in Input Data

Column	Description
Band	Identifies the segment number within the interview.
Timecode	The timestamp of the spoken text in HH:MM:SS format.
Sprecher	The speaker identifier (e.g., interviewer, participant).
Transkript	The original interview text.
Übersetzung	Translation of the spoken text.
Zwischenüberschrift	Subheading within a section.
Hauptüberschrift (Übersetzung)	Translated version of the main heading.
Zwischenüberschrift (Übersetzung)	Translated version of the subheading.
Registerverknüpfungen	Index references or linked records.
Anmerkungen	Additional notes on the conversation.
Anmerkungen (Übersetzung)	Translated version of the notes.

B. The Metadata Schema

After processing, the extracted metadata is transformed into a structured format consisting of 136 columns categorized into 13 major groups. These categories represent key aspects of the metadata, ensuring comprehensive information retrieval as shown in table II

TABLE II
Metadata Schema Categories

Column	Contents
General Information	Location, Archive ID, Document Type, Time of Creation
Personal Information	Name, Birth Year, Gender, Pseudonym
Interview Details	Interview Title, Duration, Interviewer, Segmentation
Contact Information	Street, Postal Code, Phone Number
Social and Demographic Information	Group Affiliation, Occupation, Family Status
Media and Storage Details	Photos, Documents, Storage Medium
Education and Career	Schooling, Career Changes, Employment Status
Family and Relationship Data	Marital History, Children's Birth Years
Political and Social Engagement	Political Orientation, Memberships
Historical and Wartime Involvement	Military Service, Nazi-related Organizations
Parental and Partner Information	Parents' Background, Partner's Engagement

IV. System Architecture

The system architecture as shown in 1 is designed to efficiently process and extract structured metadata from interview transcripts while preserving context.

The workflow begins with raw interview transcripts, which are preprocessed to remove irrelevant data and standardize formatting. This ensures that only meaningful fields, such as timestamps, speaker information, and transcribed text, are retained.

Given the length of transcripts, the text is segmented into smaller, context-aware chunks using a cosine similarity-based chunking mechanism. This method identifies logical breakpoints in the text by measuring semantic similarities between adjacent sentences, ensuring that topic coherence is maintained. The segmented chunks are then processed individually by a LLM , which extracts relevant metadata fields based on a predefined schema.

To enhance consistency, extracted metadata from previous chunks is passed to subsequent ones, allowing the model to retain context across segments. The extracted metadata is then refined and structured into a standardized output format, such as CSV, ensuring ease of storage and retrieval for research purposes. This architecture allows for scalable and efficient metadata extraction while maintaining high accuracy across varying interview formats.

V. Methodology

This study aims to extract important details from German interview transcripts using LLMs. The process follows several steps to make sure the extracted information is accurate, well-organized, and useful for research. Each step plays an important role in cleaning the raw text, identifying key information, and turning it into a structured format that can be easily analyzed and stored. The main steps are:

A. Preprocessing

As we load the data, we observe that the majority of data points are null, except for key fields such as **Timecode, Sprecher, and Transkript** in each file. These fields contain the essential information required for further processing, so we prioritize them for extracting meaningful metadata. The **Timecode** ensures precise timestamp alignment, the **Sprecher (Speaker)** helps differentiate between the interviewer and interviewee, and the **Transkript** provides the raw textual content from which metadata is derived.

Regarding the metadata schema, the originally provided schema was in TSV format. Converting the schema into JSON allows for standardized data representation, ensuring

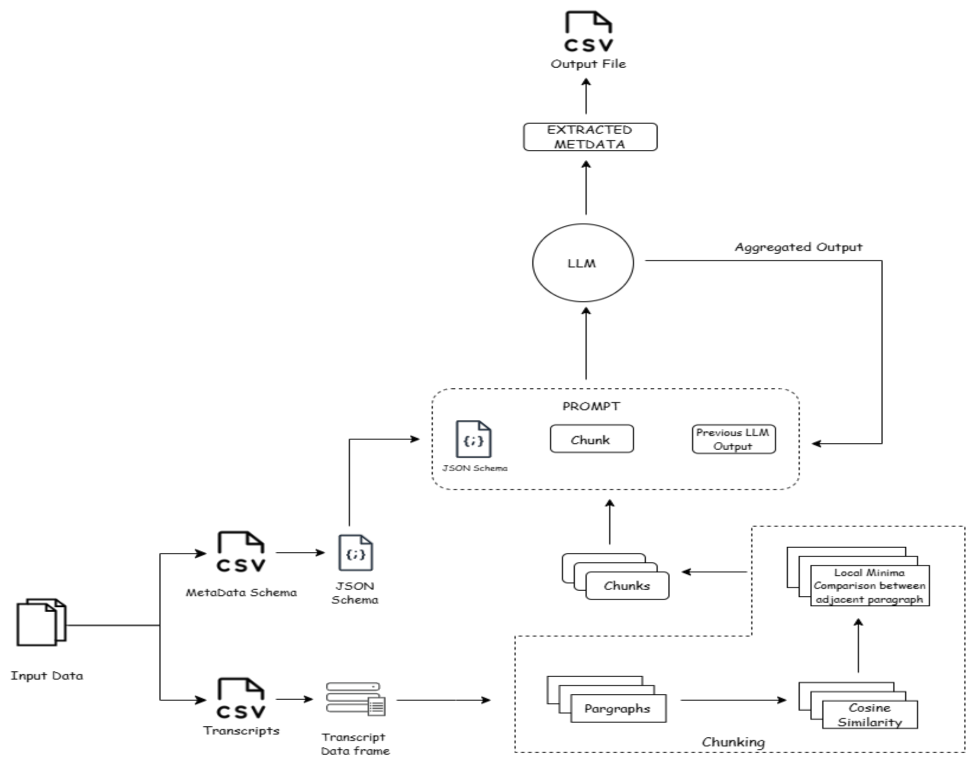

Fig. 1. Overview of the Metadata Extraction System

the chances of misinterpretation or errors[5]. By passing the metadata schema in JSON format, we create a structured framework that enables the LLM to systematically analyze the transcript data, map extracted values to the correct schema fields, and minimize ambiguity in metadata assignment. This structured approach significantly improves the reliability and accuracy of metadata extraction, ensuring that the information is correctly categorized and stored for archival and research purposes.

B. Choosing the Right Language Model

Selecting the appropriate LLM is a critical step in ensuring accurate metadata extraction from German transcripts. Since the dataset primarily consists of German text, it was essential to identify models that could effectively understand and extract relevant metadata in this language.

1) LLM Exploration: To achieve optimal results, we explored different models that support German language processing [3][4]. After evaluating various options, we narrowed our selection to two prominent LLMs: **LLaMA and Mistral**. Both models are well-regarded for their performance in multilingual tasks, contextual understanding, and structured information extraction. To ensure a balanced evaluation, we tested both smaller and larger variants of each model:

- **LLaMA 8B (smaller variant)**
- **Mistral 7B (smaller variant)**
- **LLaMA 70B (larger variant)**
- **Mixtral 8x7B (larger variant)**

By experimenting with both lightweight and high-capacity models, we aimed to determine which configuration best captured structured metadata while maintaining efficiency for real-world applications.

2) LLM Evaluation: To choose the best LLM for our project, we needed to test how well different models could extract key details from a transcript. To do this, we built a small prototype metadata extractor, where each model was given the same sample transcript and asked to find 12 important data points. This helped us measure how accurate and reliable each model was in identifying and extracting metadata [9].

To run these tests, we used API keys from Groq [11]. and Together AI [12] , which allowed us to access different LLMs and see how they performed. Each model analyzed the transcript and tried to extract 12 specific details, as shown in III

The results are shown in Table IV.

C. Chunking Strategy: Context-Aware Segmentation for LLM Processing

Processing lengthy interview transcripts while preserving contextual integrity is a significant challenge in metadata extraction. Conventional segmentation techniques, such as splitting by speaker turns or sentence delimiters, often produce inconsistencies. Speaker-based chunking can lead to disproportionately large segments when a single speaker dominates the conversation, while sentence-based segmentation can create fragments too small to provide meaningful context. These

TABLE III
EXTRACTED METADATA FIELDS

Field	Description
Name	Interviewee's name
Jahrgang	Year of birth
Ort	Place of birth or residence
Geschlecht	Gender
Beruf	Occupation
Vat_JG	Father's birth year
Vat_Konfession	Father's religion
Vat_Herkun	Father's place of origin
Vat_Schule	Father's education
Vat_Ausbil	Father's training or apprenticeship
Vat_Stand	Father's social status
Vat_Polor	Father's political orientation

TABLE IV
MODEL EVALUATION ACCURACY

Model	Correct Fields	Total Fields	Accuracy (%)
LLaMA 8B	9.5	12	79.17
Mixtral 8x7B (Groq)	11.5	12	95.83
LLaMA 70B	11.5	12	95.83
Mixtral 8x7B (Together AI)	9.5	12	79.17
Mistral 7B	8.5	12	70.83

limitations prompted the need for a more adaptive, context-aware segmentation approach.

Our method leverages **TF-IDF vectorization and cosine similarity analysis** to identify natural breakpoints in the transcript. Instead of relying on predefined rules, the segmentation process dynamically adapts to shifts in thematic content. The transcript is first divided into paragraphs of ten sentences each. These paragraphs are then transformed into numerical representations using TF-IDF vectorization, capturing semantic relationships between them. By computing cosine similarity between consecutive paragraphs, we establish a measure of contextual continuity.

The segmentation points are determined by detecting **local minima** in the cosine similarity distribution. A local minimum indicates a significant drop in similarity, suggesting a shift in topic or a change in the interviewee's narrative. These points serve as natural segment boundaries, ensuring that topic coherence is maintained throughout the transcript. Unlike rule-based chunking methods, this approach allows segmentation to be guided by the structure of the conversation rather than arbitrary cutoffs.

Once segmentation points are identified, adjacent paragraphs that remain contextually connected are merged to prevent excessive fragmentation. This ensures that each segment is self-contained while preserving the larger discourse structure. The effectiveness of this method is visualized in Figure 2 illustrating the segmentation process applied to the interview transcript. The grey line represents the cosine similarity scores between consecutive paragraphs, capturing the natural flow of the conversation. Lower similarity values indicate potential shifts in topic. The red dots mark detected segment boundaries, corresponding to local minima where a significant change in context is identified. The vertical dashed red lines further emphasize these breakpoints, visually segmenting the transcript into coherent chunks. This method ensures that each segment retains contextual integrity while preventing abrupt topic fragmentation, ultimately enhancing the accuracy of metadata extraction..

This approach improves **LLM-based metadata extraction** by ensuring that the input is well-structured and contextually meaningful. Unlike traditional segmentation, which often disrupts thematic continuity, our method dynamically adapts to content variations, optimizing metadata accuracy and ensuring that key details are extracted from semantically relevant segments. By integrating linguistic intuition with mathematical precision, this methodology enhances transcript processing at scale, making it highly effective for archival and research applications.

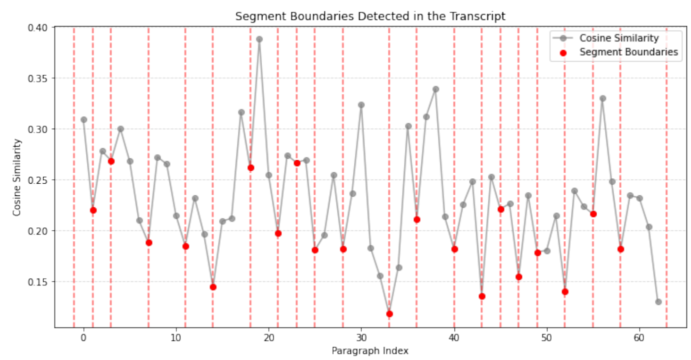

Fig. 2. Understanding Topic Flow in the Interview

D. Chunk-wise Metadata Extraction

Following preprocessing, chunking, and model selection, the next phase involved extracting structured metadata from the transcript chunks. Initially, we experimented with direct text-based extraction, but it resulted in inconsistent and unstructured outputs. To overcome these issues, We adopted a JSON-based approach to ensure a structured and standardized output, maintain consistency across extractions, and retain context throughout the transcript chunks.

By passing JSON schema within the prompt, the model reliably mapped extracted details to predefined fields, improving accuracy and reducing post-processing efforts.

E. Post Processing

Once the metadata was extracted, it was compiled into a DataFrame and exported as a CSV file for structured storage and further analysis. This final step ensured the extracted information was readily usable for archival and research purposes.

VI. Discussion

The process of extracting structured metadata from German interview transcripts using Large Language Models (LLMs) has yielded several valuable insights, alongside challenges and areas for future enhancement. This section provides a detailed discussion of the effectiveness of our methodology, the issues encountered, and the potential directions for refining the approach further.

A. Key Findings

One of the most significant achievements of this study is the ability to convert unstructured interview transcripts into a structured metadata format, enhancing both accessibility and searchability. The following key findings highlight the strengths of our approach:

Improved Data Quality and Schema Representation: Converting the metadata schema into JSON provided a structured and consistent format, ensuring that the LLM could reliably parse and extract information. This approach minimized errors arising from ambiguous textual representations and standardized how metadata fields were populated. Additionally, focusing only on key column such as *Transkript* allowed us to filter out irrelevant data points, improving overall data quality.

Effective Chunking Strategy for Context Preservation: A crucial aspect of metadata extraction is maintaining the context of interview transcripts while segmenting large bodies of text. Initially, we experimented with chunking by speaker turns and sentence-based segmentation using full stops. However, these methods either resulted in excessively large segments or disrupted context. Our refined approach, leveraging TF-IDF vectorization and cosine similarity, allowed us to identify natural topic shifts and create meaningful segments while maintaining coherence. This technique helped the LLM capture detailed metadata without losing relevant contextual connections.

Performance Evaluation of Language Models: Evaluating different LLMs (LLaMA 8B, Mistral 7B, LLaMA 70B, and Mixtral 8x7B) demonstrated that larger models generally performed better in extracting structured metadata. The evaluation process, which involved retrieving 12 predefined metadata fields from a sample transcript, showed that LLaMA 70B and Mixtral 8x7B achieved the highest accuracy (95.83%). However, inconsistencies in Mixtral's responses across different APIs (Groq vs. Together AI) led to the selection of LLaMA 70B as the final model due to its stability and reliability.

Structured Prompting and JSON-Based Metadata Extraction: Initially, we experimented with a text-based extraction prompt that explicitly listed the metadata fields in German, instructing the model to extract the relevant details. However, this approach led to inconsistencies, as the responses were unstructured and required extensive cleaning. Switching to a JSON-based extraction method resolved these issues, allowing for standardized outputs with predefined schema fields. Additionally, passing metadata from previous chunks into subsequent prompts improved continuity in extraction results.

B. Challenges and Limitations

Despite the success of the proposed approach, several challenges and limitations remain that impact the overall accuracy and flexibility of metadata extraction:

Handling Large Transcripts Without Losing Context: The necessity of a chunking strategy to process long transcripts means that certain details spanning multiple segments might be lost. While passing aggregated metadata to subsequent chunks mitigated some issues, complex relationships or references between distant transcript segments could still be overlooked. This limitation is particularly problematic when interviewees reference earlier parts of their conversation later in the transcript.

Predefined Metadata Schema Restricts Flexibility: While a structured schema enhances consistency, it also imposes constraints on adaptability. Any new metadata categories or changes to the schema require modifications to the prompt and processing pipeline. This makes the system less dynamic in handling evolving archival requirements or interviews with varying metadata needs.

Variability in LLM Responses: During evaluation, it was observed that different LLMs sometimes provided varying responses for the same transcript. In particular, Mixtral 8x7B yielded inconsistent results across different API platforms. This variability suggests that the underlying training data and model tuning influence output stability, which can introduce unpredictability in extracted metadata. Future work should investigate methods to normalize LLM responses for consistency.

Potential Bias in Metadata Extraction: Large Language Models are trained on vast datasets that may contain biases. These biases can lead to incorrect or misleading metadata extraction, especially for historical, cultural, or politically sensitive topics. While careful prompt engineering can help guide model behavior, additional post-processing checks are necessary to ensure factual correctness and neutrality in extracted metadata.

C. Future Improvements

To further improve the efficiency, accuracy, and adaptability of metadata extraction, several enhancements can be implemented in future iterations:

Incorporation of Human Annotation for Quality Control: To systematically evaluate the metadata extraction pipeline, we plan to introduce a human annotation process where experts manually validate extracted metadata. This will allow us to quantify the model's precision and recall in a structured manner, helping to refine prompts and processing logic based on real-world evaluation metrics.

Adaptive Chunking Strategy for Context Preservation: Instead of relying solely on static TF-IDF cosine similarity thresholds for segmentation, an adaptive chunking strategy that dynamically adjusts chunk sizes based on topic density could improve context retention. For instance, integrating Named Entity Recognition (NER) with similarity scores might enable more intelligent chunking by ensuring that important entities are not split across segments.

Multi-Model Ensemble Approach for Enhanced Accuracy: Instead of relying on a single LLM, an ensemble approach where multiple models extract metadata and their outputs are aggregated could improve overall reliability. This approach could involve:

- Using a lightweight model (e.g., Mistral 7B) for rapid preliminary extraction.
- Passing results to a larger model (e.g., LLaMA 70B) for final validation and refinement.
- Employing statistical techniques to resolve discrepancies between different models' outputs.

Improved Post-Processing for Data Cleaning and Formatting: Although JSON-based extraction already improves structuring, post-processing logic can be further refined to detect ambiguous values, missing fields, or incorrect formats. Using rule-based validation methods or lightweight ML models for metadata validation could ensure higher quality final outputs.

In future iterations, we plan to incorporate a systematic annotation process for the answers extracted from the transcripts. This annotation will serve as a quality check to evaluate and refine the metadata extraction pipeline, ensuring greater accuracy in our results. However, due to time constraints, this approach was not implemented, and instead, we focused on optimizing the existing pipeline to achieve meaningful results within a limited timeframe.

VII. Conclusion

This study demonstrated the potential of Large Language Models (LLMs) for structured metadata extraction from German interview transcripts. By implementing JSON-based structured extraction, cosine similarity chunking, and model evaluation experiments, we successfully transformed unstructured interview text into a structured, searchable format.

However, challenges such as context fragmentation, schema rigidity, LLM variability, and bias concerns indicate that further refinements are needed. Future improvements, including adaptive chunking, human annotation, multi-model ensembling, and fine-tuning domain-specific LLMs, will help overcome these limitations and make metadata extraction more scalable, precise, and adaptable to dynamic archival needs.

Acknowledgment

We would like to thank Dennis Möbus of Fern Universität in Hagen for his guidance throughout the project and for providing the biographical interview transcripts and extraction schema that were vital for our experiments. We also extend our sincere appreciation to Prof. Binh Vu for his continuous support in various aspects of the project. Finally, we acknowledge the invaluable contributions of all our team members, without whom this project would not have been possible.

References

[1] DataCamp, "Prompt Chaining Tutorial: What Is Prompt Chaining and How to Use It?" Available: https://www.datacamp.com/tutorial/prompt-chaining-llm.

[2] Artificial Analysis AI, "LLM Model Leaderboards." Available: https://artificialanalysis.ai/leaderboards/models.

[3] Klu AI, "LLM Leaderboard." Available: https://klu.ai/llm-leaderboard.

[4] Vellum AI, "LLM Leaderboard." Available: https://www.vellum.ai/llm-leaderboard.

[5] Medium, "Optimizing Prompt Formats for Large Language Models: A Comparative Study of JSON, Plain Text, and Other Strategies." Available: https://2020machinelearning.medium.com/optimizing-prompt-formats-for-large-language-models-a-comparative-study-of-json-plain-text-and-9e8f58302902.

[6] R. Wang, X. Liu, and S. Zheng, "Optimizing Prompt Formats for Large Language Models," arXiv preprint, 2023. Available: https://arxiv.org/abs/2303.18223.

[7] National Library of Medicine, "Optimizing Metadata Extraction using AI-driven Approaches," 2024. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11141826/.

[8] LLaMA Documentation, "LLaMA 3 Model Cards and Prompt Formats." Available: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3$_3$/.

[9] Reddit, "Huge LLM Comparison Test: 39 Models Tested (7B–70B)," r/LocalLLaMA, 2024. Available: https://www.reddit.com/r/LocalLLaMA.

[10] Pinecone, "Chunking Strategies for Large Language Models." Available: http://pinecone.io/learn/chunking-strategies/.

[11] Groq, "Accelerating AI Inference with Groq's High-Performance Compute Systems." Available: https://groq.com/.

[12] Together AI, "High-Performance AI Infrastructure for Large-Scale Model Training and Inference." Available: https://www.together.ai/.

[13] Riedl, M., Padó, S., and Ponzetto, S. P., "Entity-Centered Topic Modeling: Combining Entity Linking and Latent Dirichlet Allocation," JLCL, 2012. Available: https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2012-riedletal-jlcl.pdf.

[14] *Proceedings of the IEEE*, 2024. Available: https://ieeexplore.ieee.org/abstract/document/10386476

[15] *bioRxiv*, 2025. Available: https://www.biorxiv.org/content/10.1101/2025.02.17.638570v1.abstract

[16] *HAL Open Science*, 2024. Available: https://hal.science/hal-04957799/

[17] A. Jacyszyn, S. Jiang, G. A. Gesese, S. Hertling, T. Kerzenmacher, P. Nowack, S. Barthlott, E. Posthumus, and H. Sack, "AI4DiTraRe: Towards LLM-Based Information Extraction for Standardising Climate Research Repositories," *FIZ Karlsruhe - Leibniz Institute for Information Infrastructure*, 2024.

[18] *medRxiv*, 2025. Available: https://www.medrxiv.org/content/10.1101/2025.02.25.25322898v1

[19] H. Patel and S. Parmar, "Prompt Engineering For Large Language Model," 2024. DOI: 10.13140/RG.2.2.11549.93923.