Exploring Large Language Models for Qualitative Data Analysis

Tim Fischer, Chris Biemann
Language Technology Group, Department of Informatics, University of Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

This paper explores the potential of Large Language Models (LLMs) to enhance qualitative data analysis (QDA) workflows within the open-source QDA platform developed at our university. We identify several opportunities within a typical QDA workflow where AI assistance can boost researcher productivity and translate these opportunities into corresponding NLP tasks: document classification, information extraction, span classification, and text generation. A benchmark tailored to these QDA activities is constructed, utilizing English and German datasets that align with relevant use cases. Focusing on efficiency and accessibility, we evaluate the performance of three prominent open-source LLMs - Llama 3.1, Gemma 2, and Mistral NeMo - on this benchmark. Our findings reveal the promise of LLM integration for streamlining QDA workflows, particularly for English-language projects. Consequently, we have implemented the *LLM Assistant* as an opt-in feature within our platform and report the implementation details. With this, we hope to further democratize access to AI capabilities for qualitative data analysis.

1 Introduction

The Discourse Analysis Tool Suite (Schneider et al., 2023) is a platform developed at our university to empower Digital Humanities (DH) researchers in conducting qualitative data analysis (QDA). Developed collaboratively and tailored to the specific needs of DH scholars, the platform democratizes access to machine learning methods, enabling non-experts to manage and analyze large-scale, unstructured, multi-modal data effectively.

While the platform's overarching design is rooted in Grounded Theory-based research (Strauss and Corbin 1990, Strauss et al. 1996), its versatile features support various disciplines. Within the core QDA workflow on our platform, researchers engage in a dynamic process of structuring their data and conducting in-depth qualitative analysis. This involves organizing and categorizing documents through metadata assignment and creating a tag taxonomy, which is utilized for classifying documents. Simultaneously, they dive into the nuances of the material, developing hierarchical code taxonomies to annotate relevant text passages and capturing their insights through memos.

While tasks like metadata extraction and document classification can be repetitive and time-consuming, qualitative analysis tasks such as annotation, paraphrasing, and summarization are demanding. The potential for AI assistance to streamline and enhance these diverse workflows motivates our exploration of Large Language Models (LLMs). Hence, this work aims to assess how effectively LLMs can support users in QDA tasks and ultimately boost their efficiency and productivity. To this end, we identify four core NLP tasks embedded in our QDA platform's core functionalities: 1) document classification, 2) document information extraction, 3) span classification, and 4) text generation. We then curate datasets that closely align with real-world use cases regarding domain and tasks, focusing on English and German. Subsequently, we evaluate the performance of three state-of-the-art open-source LLMs, Llama 3.1, Gemma 2, and Mistral NeMo, on this benchmark.

Our findings show the promising potential of LLM integration within the Discourse Analysis Tool Suite (DATS), particularly for English projects. Consequently, we implement the *LLM Assistant* as an opt-in feature for English projects, paving the way for further enhancements and expansions. Contributions of this paper are:

1. We articulate the envisioned AI-assisted workflow within our platform, highlighting user needs and requirements.
2. We design a benchmark tailored specifically to common QDA tasks within our platform.
3. We evaluate open-source LLMs on it.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 423–437
November 16, 2024. ©2024 Association for Computational Linguistics

4. We report on their integration into our tool.

This work represents our first step towards automating and providing assistance for various common tasks in our QDA platform using LLMs. We aim to facilitate more efficient, insightful qualitative data analysis by augmenting researchers' capabilities with LLM assistance.

2 Related work

QDA Platforms and AI Integration Several prominent platforms have emerged in the realm of qualitative data analysis software, each offering distinct functionalities to researchers. Some platforms have taken notable steps towards incorporating AI-powered features into their workflows.

CATMA (Gius et al., 2022) is a versatile QDA tool focusing on text and image analysis. It currently lacks built-in AI capabilities.

Known for its comprehensive approach to qualitative and mixed-methods research, MAXQDA[1] has introduced "MAXQDA AI Assist," offering AI-driven features like summarization, paraphrasing, and concept explanation.

A robust platform for qualitative data analysis, NVivo's[2] latest beta version is actively integrating AI functionalities, including thematic coding, sentiment analysis, and text summarization.

Atlas.ti[3] is recognized for its visual and network-based analysis tools. The platform incorporates AI with existing features like code suggestions, sentiment analysis, summarization, and entity recognition powered by OpenAI's GPT models.

Notably, AI-powered features within these QDA platforms are currently only found in paid versions. They are realized by sending data to third-party providers, potentially leading to data protection issues. In contrast, our open-source QDA platform aims to democratize access to state-of-the-art AI capabilities, making advanced functionalities freely available to researchers across disciplines. It can be run in-house if required.

LLM Benchmarks While several prominent general LLM benchmarks like MMLU (Hendrycks et al., 2021), SuperGLUE (Wang et al., 2019), BIG-bench (Srivastava et al., 2023), HELM (Liang et al., 2023), and MTEB (Muennighoff et al., 2023) exist, they often lack a direct connection to specific real-world applications, including qualitative data analysis. They may cover a broad range of tasks but not necessarily those most relevant to QDA workflows. In contrast, Ziems et al. (2024) explores the potential of LLMs to transform Computational Social Science (CSS) by evaluating their zero-shot performance on a range of English CSS tasks. Their extensive evaluation, focused on taxonomic labeling and free-form coding, highlights LLMs' potential to augment CSS research as zero-shot data annotators, strongly motivating our work. Still, many benchmarks prioritize English data.

Our benchmark is constructed to be relevant to the tasks and data encountered in QDA, incorporating datasets that approximate real-world scenarios within our platform. Focusing on document classification, information extraction, span classification, and text generation in English and German, we aim to gather insights that can directly inform the effective integration and utilization of LLMs in qualitative research.

LLMs and QDA Rasheed et al. (2024) explores the potential of LLMs to serve as data analysts in qualitative research within Software Engineering. Their approach employs a multi-agent model where each LLM agent performs specific research-related tasks, such as interpreting textual data and interview transcripts, to automate common qualitative analysis processes. Their findings suggest that LLMs can significantly accelerate data analysis, allowing researchers to handle larger datasets efficiently, which further motivates this work.

3 Envisioned workflow

This section presents an illustrative excerpt of a qualitative data analysis workflow inspired by project partners who actively utilize the Discourse Analysis Tool Suite. This scenario highlights potential areas where AI-powered automation enhances productivity.

Imagine Alice, a researcher aiming to analyze local companies across various industries, focusing on their societal impact and challenges. She initiates semi-structured interviews with CEOs, stating the current date, introducing herself, and then inquiring about company details (e.g., size, sales volume) and the interviewee's background (e.g., name, age, position) before asking her research questions. She records these interviews with her smartphone.

After collecting data, Alice starts the qualitative analysis process within our QDA platform. She creates a new project, defines document tags for industry categorization, and establishes metadata

[1] https://maxqda.com
[2] https://nvivo.de/
[3] https://atlasti.com

fields like "Interview Date," "Company Size," and "Partner Name" to capture crucial interview details. Upon uploading her recorded interviews, our platform currently utilizes Whisper (Radford et al., 2022) to generate automatic transcripts.

After the automatic pre-processing, Alice is presented with interview transcripts, which are now automatically tagged by industry and partially populated with metadata. The AI-powered system suggested tags and metadata values based on the interview content, which Alice verifies and completes with the help of an intuitive UI.

Having an organized document collection, Alice starts the qualitative annotation. She constructs a code system aligned with her research questions that incorporates codes like "Social Impact" and fine-grained sub-codes for "Problems". Next, she activates the auto-coding feature, and the AI-powered system suggests relevant text annotations. An interface allows her to review these suggestions.

While reviewing, Alice notices occasional disfluencies like repetitions and filler words, typical of verbatim transcriptions. She selects a disfluent passage, activates the paraphrasing feature, and is presented with an AI-generated suggestion. After minor edits, she approves the improved version. Similarly, she employs the automatic summarization feature to condense lengthy answers to her interview questions for improved clarity.

Equipped with such AI-powered tools, Alice efficiently processes her remaining transcripts and utilizes our platform's analytical features to answer her research questions.

4 Benchmark of QDA-related NLP Tasks

This benchmark evaluates LLMs on tasks mirroring real-world QDA use cases, as outlined in the previous section. We aim to identify the most suitable model for effective user support to be integrated into DATS. To this end, we carefully select datasets for document classification (assigning tags to documents), document information extraction (extraction of metadata from documents), span classification (annotation of relevant passages), and text generation (correcting and summarizing text passages). Our platform mainly caters to English and German data, so we focus our evaluation and dataset selection on these two languages.

4.1 Models

Our open-source Discourse Analysis Tool Suite exclusively employs open-source and open-licensed libraries, which extends to integrated models. Since our primary users, universities and researchers, often handle sensitive data, local execution of the entire platform, including models, is crucial for maintaining data privacy. Given such environments' typically limited computational resources, we focus on small, efficient LLM variants for fast inference.

As a result, we evaluate three state-of-the-art open-source decoder-only models: Llama 3.1 (Dubey et al. 2024, Touvron et al. 2023), Gemma 2 (Gemma Team, 2024), and Mistral NeMo (Mistral AI Team 2024, Jiang et al. 2023) with 8B, 9B, and 12B parameters, respectively. This makes deploying them in environments with limited resources possible. We only test instruct fine-tuned models and use half-precision (FP16) variants. Llama 3.1 is an openly accessible, open-source model from Meta AI published under the Llama 3 Community License. It has a large context window of 128k tokens and was trained on a corpus of about 15 trillion multilingual tokens. Gemma 2 is a lightweight, open model from Google, built from the same technology as their Gemini models. This variant was trained on 8 trillion tokens from web documents, code, and mathematics, primarily in English. Mistral NeMo is a model from Mistral AI built in collaboration with NVIDIA and published under the Apache 2.0 license. It also offers a large context window of up to 128k tokens and was trained on multi-lingual and code data.

4.2 Experiment construction

We conduct zero-shot experiments using a single, clear prompt for each task, dataset, and language, deliberately avoiding extensive prompt engineering (prompts are detailed in appendix A, dataset taxonomies are listed in appendix B). Models are instructed on the expected output format, and deviations are counted as errors. We report formatting adherence at the end of this chapter.

All experiments ran on a single A100 GPU, repeated three times per configuration. Reported results are averaged across runs to mitigate fluctuations. We further aggregate results by model and task because of space restrictions; full results are listed in the appendix.

4.3 Task 1: Document classification

Document classification is essential in QDA for organizing data collections. Our platform supports fine-grained tag sets that enable researchers to classify documents into one or more groups. Many analysis features rely on these tags for comparison and sub-corpora creation, highlighting the potential of automatic tag suggestions to improve workflows.

We assess LLMs on two relevant document classification tasks. Multi-class involves assigning a single class from pre-defined options, further differentiated into coarse- and fine-grained settings based on the number of classes. Multi-label allows for multiple class associations.

4.3.1 Datasets

Tagesschau is an established German news website known for its serious and objective reporting covering news from Germany and the world. We automatically extracted a taxonomy of 4 main categories (coarse) and 20 sub-categories (fine) from a publicly available crawl[4] spanning 2018 – 2023.

BBC is the public service broadcaster of the United Kingdom that publishes English news from UK and the world and is deemed a trusted source of information. Similarly, we extracted a taxonomy of 4 main categories (coarse) and 26 sub-categories (fined) from the BBC dataset published by Li et al. (2024) that covers 2017 – 2024.

IMDb, the Internet Movie Database, contains information about movies, TV shows, etc. The IMDb Genres dataset[5] includes movie descriptions and their classification into one of 16 major genres (coarse) and 2-3 of 25 subgenres (multi-label). For all datasets, we sampled 10.000 documents.

Since analyzing news articles is a common use case within our Discourse Analysis Tool Suite, the Tagesschau and BBC datasets, which are news datasets, are a good fit for this benchmark. While not directly related to our domain, the IMDb Genres dataset was explicitly included for its multi-label classification task.

4.3.2 Results

We evaluate document classification on three subtasks: coarse-, fine-grained, and multi-label classification. The aggregated results are presented in Table 1, and the complete evaluation is in the appendix, Table 5. We report weighted Precision (Prec), Recall, F1, and Accuracy (Acc).

[4] https://github.com/bjoernpl/tagesschau

[5] https://kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre

Model	Task	Prec	Recall	F1	Acc
gemma2	coarse	**67.20**	**65.42**	**64.85**	**65.4**
llama3.1	coarse	64.33	59.36	58.07	59.4
mistral	coarse	64.39	62.00	61.28	62.0
gemma2	fine	**70.50**	**62.00**	**63.00**	**62.4**
llama3.1	fine	57.50	36.00	34.00	35.9
mistral	fine	66.50	53.50	55.00	53.6
gemma2	multi	55.48	**52.23**	**52.35**	**8.6**
llama3.1	multi	52.12	42.38	43.95	6.9
mistral	multi	**57.51**	49.62	51.05	**8.6**

Table 1: Evaluation of Task 1 - Document Classification

Gemma 2 consistently outperforms the other models, maintaining high performance (64.85 F1 coarse, 63.00 fine) despite the significant increase in classes for the fine-grained classification tasks. In contrast, the other models struggle in the fine-grained scenario, Llama 3.1 performing the worst (34.00 F1). We refer to appendix Table 5 to compare English and German performance. For coarse-grained classification, all models exhibit superior performance on the German Tagesschau dataset compared to the English BBC dataset but experience a noticeable drop (over 22 points F1) on the fine-grained German task. This suggests challenges in German fine-grained classification.

4.4 Task 2: Document Information Extraction

Similar to document classification, assigning metadata to research materials aids data organization in QDA. DATS enables users to define metadata that is used for features like search, filtering, visualization, and quantitative analysis. As automating metadata extraction could boost researcher productivity, this is a relevant task.

We frame it as document-level information extraction, exploring extractive question-answering (EQA) and template-filling approaches. EQA involves extracting answers from the context or labeling them as unanswerable. Template-filling (similar to slot-filling, relation extraction, and event extraction) focuses on extracting multiple related information, e.g., arguments of a relation or information about an event.

4.4.1 Datasets

The Stanford Question Answering Dataset (SQuAD) by Rajpurkar et al. (2016) is a benchmark dataset for extractive question-answering. It

consists of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a text segment from the context.

The SQuAD 2.0 dataset (Rajpurkar et al., 2018) builds upon the original one and introduces unanswerable questions, making it more challenging. Crowd workers carefully crafted the unanswerable questions to be similar to answerable ones.

The GermanQuAD dataset (Möller et al., 2021) is a German counterpart to SQuAD. This extractive question-answering dataset was carefully constructed by students and experts familiar with machine learning and QA on the German variants of the English Wikipedia articles used in SQuAD.

These datasets, with multiple questions per Wikipedia passage, align with our scenario of extracting various metadata from documents. While some questions align with typical metadata fields (e.g., "When," "How many," "Where"), others target more complex information, making them less suitable for metadata extraction. Nonetheless, strong performance on these datasets indicates potential for successful application in DATS.

The MUC-4 dataset (Sundheim, 1992), designed for template-filling, contains 1700 news articles about terrorist incidents. It requires systems to classify each incident and fill five slots of relevant information. This dataset aligns well with our use case of extracting multiple metadata from documents, particularly within the news domain, making it an ideal fit for our benchmark.

4.4.2 Results

We evaluate document-level information extraction on two sub-tasks: extractive QA and template-filling. The aggregated results are shown in Table 2, the complete evaluation in the appendix, Table 6. We report Exact Match (EM) and F1 scores, with template-filling scores averaged across all slots.

While Gemma 2 exhibits a clear advantage in extractive QA (79 F1), Llama 3.1 performs best in template-filling (40 F1). Interestingly, template-filling is a greater challenge overall, likely due to the increased complexity of extracting multiple correct answers simultaneously, as reflected in the lower scores across all models (at least 70 F1 for extractive qa vs. at most 40 F1 for template filling). In extractive QA, the models exhibit comparable performance on the GermanQuAD and SQuAD datasets (refer to appendix Table 6), indicating no significant difference between English and German language capabilities.

Model	Task	EM	F1
gemma2	extractive-qa	**66.53**	**79.41**
llama3.1	extractive-qa	56.04	70.64
mistral-nemo	extractive-qa	55.21	72.39
gemma2	template-filling	36.69	36.70
llama3.1	template-filling	**40.62**	**40.63**
mistral	template-filling	23.86	23.87

Table 2: Evaluation of Task 2 - Information Extraction.

4.5 Task 3: Span classification

Annotation (often also called coding) of relevant text passages (spans) is critical to many QDA projects and especially important for research projects following the Grounded Theory paradigm. Here, the coding is done in the three phases of "open," "selective," and "axial" coding. Our Discourse Analysis Tool Suite supports these coding phases. Automating parts of the annotation process could streamline their workflow, making this a relevant benchmark task. As users can create fine-grained code systems to annotate text passages in our platform, we formulate the automatic annotation of text passages as coarse- and fine-grained span classification tasks.

4.5.1 Datasets

Few-NERD (Ding et al., 2021) is a fine-grained, large-scale Named Entity Recognition dataset consisting of 8 coarse-grained and 66 fine-grained categories. Over 180,000 sentences of Wikipedia articles were carefully annotated by experienced annotators. German LER (Leitner et al., 2020), the German Legal Entity Recognition dataset, consists of German legal documents and a typology relevant to court decisions with 7 coarse-grained and 19 fine-grained types. About 66,000 sentences were annotated by two domain experts. This dataset is challenging, as models need to be familiar with German law terms. While categories like persons, events, and organizations are relevant, we often observe different annotations spanning multiple sentences in QDA projects. Still, FewNERD and German LER are included in our benchmark as NER is a prominent span classification task, and we argue that understanding the concepts of these datasets is likely required for performing well on more complex annotation tasks.

The dataset for quotation attribution in German news articles (Petersen-Frey and Biemann, 2024)

consists of 1000 annotated German news articles from WIKINEWS. It includes information about *who* said *what* to *whom* as well as *how* and *in which context*. Quotations are categorized into direct, indirect, free, and reported speech. This task is part of the benchmark, as the identification and annotation of utterances were common to multiple projects we conducted with colleagues from social sciences. However, we simplify this task by considering only "speaker" and "direct speech" annotations.

4.5.2 Results

We evaluate span classification on three sub-tasks: coarse-, fine-grained NER (coarse, fine), and quotations (quot). The aggregated results are listed in Table 3. The full evaluation is reported in the appendix, Table 7. We report weighted Precision (P), Recall (R), F1 score, and Accuracy (Acc). Accuracy includes the classification of outside tokens.

Gemma 2 consistently outperforms the other models across all sub-tasks (30–38 F1). Llama 3.1's performance is notably poor (7–15 F1), primarily due to its frequent failure to adhere to output formatting instructions. The model often switches the positions of NER labels and corresponding text spans, leading to significant parsing errors. To maintain a fair comparison across all models, we retain our original parsing algorithm and prompt instead of fixing such errors. Consequently, our experiments indirectly evaluate the instruction following capabilities of the models. Increasing the number of classes from coarse to fine-grained NER leads to a significant performance drop across all models. Comparing English and German results (refer to appendix Table 7) reveals significantly lower scores for German. Even the best performing model, Gemma 2, achieves at most half the F1 score in German compared to English. This could suggest challenges in handling German text or reflect the increased difficulty of German LER.

Model	Task	P	R	F1	Acc
gemma2	coarse	35.19	**44.07**	**37.98**	**85.02**
llama3.1	coarse	26.22	4.70	7.82	80.33
mistral	coarse	**39.28**	15.38	21.22	81.76
gemma2	fine	**40.25**	**31.64**	**31.84**	**83.86**
llama3.1	fine	26.66	4.16	6.78	80.37
mistral	fine	39.72	10.66	15.39	81.04
gemma2	quot	**36.40**	**25.61**	**29.81**	**90.29**
llama3.1	quot	33.16	10.88	15.18	88.17
mistral	quot	23.03	10.69	13.71	86.24

Table 3: Evaluation of Task 3 - Span Classification

4.6 Task 4: Text Generation

Researchers must potentially correct fluency issues or summarize long and wordy statements, especially when dealing with transcripts. Further, in the qualitative content analysis approach of Mayring (2019), popular in the Humanities in Germany, summarization, explication, and structuring are defined as the three main pillars of content analysis. Here, summarization is employed to condense information, explication to clarify meaning by providing context, and structuring to filter the material according to specific aspects systematically.

DATS allows users to attach notes to documents or text passages. As hinted at in Section 3, we aim to expand this functionality with fluency correction and text summarization capabilities. Assisting users with this can streamline their workflow. Consequently, we consider the two text generation tasks in our benchmark.

Abstractive summarization involves generating concise and fluent summaries similar to human-written ones and is thus likely preferred by our users. Disfluency correction aims to enhance readability by identifying and removing issues such as repetitions, filler words, and false starts.

4.6.1 Datasets

The Disfl-QA dataset (Gupta et al., 2021), initially intended for evaluating question-answering robustness against disfluencies, consists of about 12k pairs of fluent and corresponding disfluent questions built upon SQuAD 2.0. We repurpose the dataset to benchmark models' ability to correct disfluent texts, leveraging only the question pairs.

The DISCO dataset (Bhat et al., 2023), designed to facilitate multilingual disfluency correction, comprises a human-annotated corpus of over 12k disfluent-fluent text utterance pairs in English, Hindi, German, and French. We utilize only the English and German parts. It is constructed upon a publicly available dataset of human-AI agent interactions. The dataset covers four disfluency types: Filler, Repetition, Correction, and False Start.

While the specific domains of these datasets (question-answering and human-AI interactions) differ from our platform's use case of correcting transcriptions of interviews and other qualitative data, we included them due to the scarcity of re-

sources for disfluency correction.

The CNN/DM (Hermann et al., 2015) dataset, a widely-used benchmark for summarization tasks, consists of over 300,000 article-summary pairs extracted from CNN and Daily Mail news articles written in English between 2007 and 2015. The summaries are primarily based on human-generated highlights or article descriptions.

MLSUM (Scialom et al., 2020), a large-scale multilingual summarization dataset, comprises over 1.5 million article-summary pairs collected from online newspapers between 2010 and 2019. It covers five languages (French, German, Spanish, Russian, and Turkish) and is a multilingual extension to CNN/DM. We utilize the German part of MLSUM for our benchmark, drawn from the Süddeutsche Zeitung newspaper.

Both datasets, centered around news articles, align well with the use of news articles as a data source on our QDA platform, making them suitable for evaluating LLM summarization performance.

4.6.2 Results

We evaluate disfluency correction (CORR) and abstractive summarization (SUM). The aggregated results are shown in Table 4. The full evaluation is in appendix Table 8. We report Rouge (R), Exact Match (EM), F1 and METEOR scores.

Gemma 2 demonstrates superior performance in the disfluency correction task (84 R-1), with the other two models performing slightly worse (79 R-1). Llama 3.1 and Gemma 2 perform similarly in abstractive summarization (28 vs. 29 R-1). Both summarization and disfluency correction tasks exhibit a performance drop of at least 12 percentage points when applied to German text (see appendix Table 8), suggesting increased difficulty for all models in handling German text generation and instruction following.

4.7 Discussion

Across the four evaluated tasks, Gemma 2 consistently emerges as the top-performing model. While Llama 3.1 performs best in template-filling and abstractive summarization, its struggles with NER, instruction following and the consistent worst performance in German tasks highlight potential limitations. Mistral NeMo, while generally capable, could never demonstrate superior performance, even though it has the most parameters (12B) among the benchmarked models.

Regarding language performance, our findings reveal discrepancies. While all models exhibit strong performance on German data for coarse-grained document classification, a consistent decline is observed across the board for fine-grained document classification, span classification, and text generation tasks in German. This suggests that current open-source LLMs still face challenges in handling the complexities of the German language, particularly in nuanced and generative tasks.

Throughout all experiments, we tracked parsing errors. Gemma 2 consistently adhered to the provided instructions (1% issues), followed by Mistral NeMo (2% issues). Llama 3.1, on the other hand, struggled notably (5% issues), most evident by the span classification tasks.

Overall, our findings highlight the potential of LLMs, especially Gemma 2, to significantly enhance QDA workflows regarding document classification, document information extraction, span classification, and text generation.

5 Integrating LLMs in our QDA platform

User feedback during early testing led us to refine our envisioned workflow outlined in Section 3. The implementation of summarization and fluency correction aligned with our original plan. However, users wanted to access document-based assistance features (document tagging, metadata extraction, annotation) at any point in their workflow, not just during the initial document import. Consequently, we redesign the *LLM Assistant* as a standalone feature independent of the pre-processing pipeline.

The feature is initiated by selecting the documents for analysis and clicking the *LLM Assistant* button. This launches a five-step dialog. Step 5 is depicted in Figure 1, Step 1 - 4 are shown in appendix Figure 2:

1. Task Selection: Users select document tagging, metadata extraction, or annotation.
2. Focus: Users specify which tags, metadata fields, or codes the LLM should consider.
3. Prompt Review: Users revise the system- and user prompts that are automatically generated based on the chosen task and selections
4. Job Execution: A progress bar indicates that the LLM Assistant job runs in the background. The dialog can be closed and reopened later.
5. Result Review: LLM-generated suggestions and their reasoning alongside the documents'

Model	Task	Rouge 1	Rouge 2	Rouge L	Rouge Lsum	EM	F1	METEOR
gemma2	CORR	**84.24**	**73.68**	**82.48**	**82.49**	**40.08**	**84.24**	**85.60**
llama3.1	CORR	79.00	67.54	77.20	77.21	33.80	78.89	80.86
mistral	CORR	78.97	63.51	77.07	77.07	25.91	78.88	79.67
gemma2	SUM	27.89	8.86	18.96	21.92	0.00	26.90	23.68
llama3.1	SUM	**29.06**	**9.93**	**19.56**	**22.70**	0.00	**28.15**	**25.70**
mistral	SUM	24.81	7.86	17.04	19.54	0.00	24.35	21.82

Table 4: Evaluation of Task 4 - Text Generation

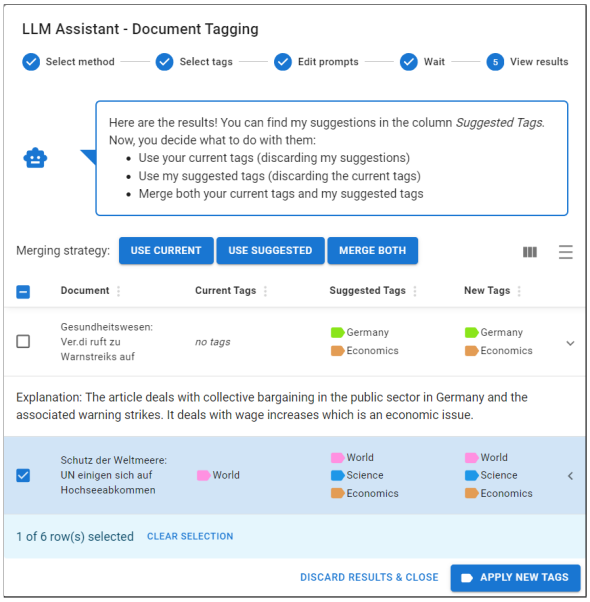


Figure 1: LLM Assistant - Step 5: Result View. Users can review the suggestions.

existing tags/metadata/codes are presented. Here, the user chooses to keep existing work, overwrite it with the LLM suggestions, or merge both per document or in batches.

User control and transparency are key requirements throughout the design of the *LLM Assistant*. Our philosophy is to ensure that any form of automation is a supportive tool, offering suggestions while the user retains decision-making. We firmly believe that AI should augment human expertise, not replace it. Thus, we've designed our platform to require explicit user approval for automatically-generated suggestions.

Our implementation achieves user control and transparency in two ways. The prompt review step allows users to inspect and modify the generated prompts, fostering transparency. Further, it enables advanced users to exert fine-grained control and provide additional task-specific instructions. In the result view, we prioritize transparency by displaying the LLM's reasoning and suggestions. Additionally, it allows users to critically evaluate the suggestions and decide how to incorporate them with their existing work, ensuring that the final output aligns with their intentions.

The *LLM Assistant* is built using React for the frontend and Ollama, FastAPI, and Celery for the backend. Celery handles background job processing, ensuring that LLM tasks run without interrupting user workflow. Ollama hosts the Gemma 2 model, which performed best in our benchmark. We reuse the benchmark prompts, as they were intentionally designed with future implementation in mind. We opt for the template-filling prompt style instead of extractive question-answering for metadata extraction. Providing metadata descriptions is more intuitive for users than formulating fitting questions for each field.

6 Conclusion

In this paper, we investigated the potential of LLMs to enhance qualitative data analysis workflows, focusing on common tasks within our open-source Discourse Analysis Tool Suite. We designed a benchmark reflecting real-world use cases and evaluated the performance of three prominent open-source LLMs. Our findings demonstrate the promise of LLM integration, particularly for English-language projects. Consequently, we implemented the *LLM Assistant* within our platform, a significant step towards empowering researchers with transparent and user-controlled AI assistance that augments, rather than replaces, human expertise.

In future work, we plan to extend the *LLM Assistant* to suggest new tags and codes, fostering a more exploratory QDA process. Furthermore, we aim to incorporate more domain-specific datasets that closely align with DH researchers' real-world use

cases. Finally, we will explore few-shot learning approaches to enhance performance on nuanced tasks. User activities within our platform (e.g., tagging documents and annotating text passages) generate valuable training data for model fine-tuning, potentially leading to more efficient models tailored to individual user preferences. Code for replicating the benchmark[6], the repository of DATS[7] and a live demo are available[8].

7 Limitations

While our implemented LLM Assistant demonstrates promising potential for enhancing QDA workflows, it's important to understand its limitations.

Firstly, the current implementation utilizes on zero-shot learning, which may not fully capture the nuances of specific QDA projects. Fine-tuning LLMs on user-specific data could lead to more accurate and contextually relevant suggestions.

Secondly, we restricted the LLM Assistant to English-language projects due to the observed performance discrepancies between English and German language tasks. Expanding language support will require further research and development to ensure similar performance across different languages.

Thirdly, the selection of suitable datasets for benchmarking remains a challenge. While we aimed to select datasets that closely resemble real-world QDA use cases, certain discrepancies between the benchmark tasks and actual user workflows exist. We will continue to identify and incorporate more representative datasets to ensure the evaluation's validity and generalizability.

Finally, we consider the inherent limitations of LLMs in general. They could struggle with tasks requiring complex reasoning, nuanced understanding of context, or common sense knowledge. Additionally, potential biases embedded within the training data can influence the LLM's outputs, requiring careful consideration and critical evaluation of the generated suggestions. This is why we deem the manual validation step integrated into our LLM Assistant, where users retain control over the acceptance and integration of AI-generated suggestions, as very important.

[6] https://github.com/uhh-lt/llm4qda

[7] https://github.com/uhh-lt/dats

[8] https://dats.ltdemos.informatik.uni-hamburg.de/

References

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. DISCO: A large scale human annotated corpus for disfluency correction in Indo-European languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12833–12857, Singapore. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3198–3213, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2407.21783.

Evelyn Gius, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. 2022. CATMA: Computer Assisted Text Markup and Analysis.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-QA: A benchmark dataset for understanding disfluencies in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3309–3319, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, and Mantas Mazeika et al. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*, Online.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, page 1693–1701, Cambridge, MA, USA. MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot et al. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607, Vancouver, Canada.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, and Dilara Soylu et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 1525(1):140–146.

Philipp Mayring. 2019. Qualitative content analysis: Demarcation, varieties, developments. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3):Art. 16.

Mistral AI Team. 2024. Mistral nemo. `https://mistral.ai/news/mistral-nemo/`. Accessed: 2024-08-31.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Fynn Petersen-Frey and Chris Biemann. 2024. Dataset of quotation attribution in German news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4412–4422, Torino, Italia. ELRA and ICCL.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, TX, USA. Association for Computational Linguistics.

Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, and Wang Xiaofeng et al. 2024. Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis. *ArXiv*, abs/2402.01386.

Florian Schneider, Tim Fischer, Fynn Petersen-Frey, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023. The D-WISE Tool Suite: Multi-Modal Machine-Learning-Powered Tools Supporting and Enhancing Digital Discourse Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 328–335.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, and Abubakar Abid et al. 2023. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023(5):1–95.

Anselm Strauss and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. SAGE Publications, Inc.

Anselm Strauss, Juliet Corbin, Solveigh Niewiarra, and Heiner Legewie. 1996. *Grounded Theory: Grundlagen Qualitativer Sozialforschung*. Beltz, Psychologie-Verlag-Union Weinheim.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, McLean, VA, USA.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, and Marie-Anne Lachaux et al. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Prompts

Table 9 lists all user prompts used in our experiments for English datasets and tasks. We translated system- and user prompts into German for German datasets. The system prompt is always the same:

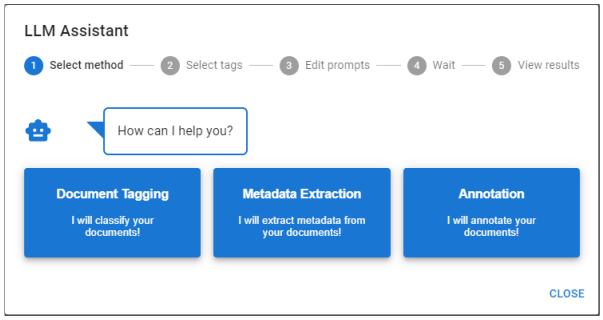
You are a system that supports the analysis of large amounts of text. You will always answer in the required format and use no formatting other than what the user expects!

All user prompts in this study adhere to a consistent structure designed to provide clear instructions and expectations. Each prompt begins with explicit task instructions, informing the model about the desired action, such as extracting an answer from a given context. This is followed by a detailed specification of the expected answer format, including potential responses for unanswerable questions (e.g., "Not answerable"). A concrete example is provided to clarify the desired output further. It is important to note that these examples are not few-shot examples derived from the datasets themselves. Finally, key constraints or limitations of the task are reiterated. We hope this ensures the model operates within the defined boundaries (e.g., extracting answers verbatim from the text).

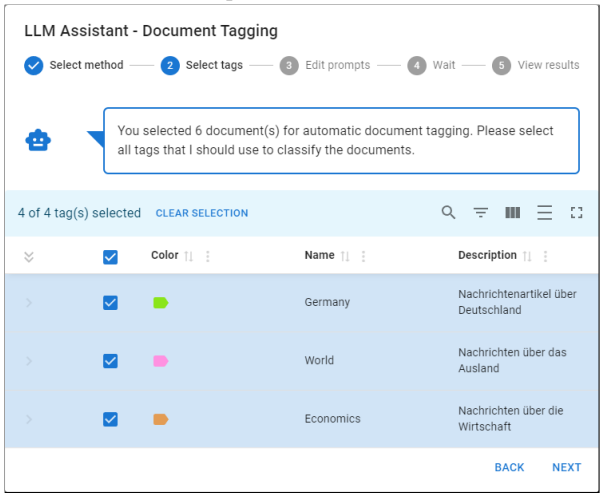
The placeholders '{}' within the prompt templates serve as dynamic variables populated with task-specific information. Depending on the task, these placeholders may contain a list of classes or categories for classification tasks, a set of slots for information extraction, or a specific question for question-answering tasks. Additionally, the context placeholder is filled with the relevant document or text passage from which the model is expected to derive its response.

B Additonal Dataset Information

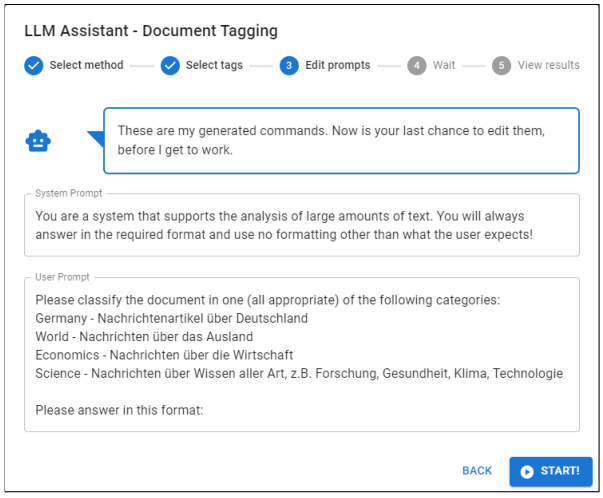
Table 10 provides additional information (e.g., taxonomies, slots) about the datasets used in our benchmark. All of this information was provided to the models within the prompt templates; for example, for the document classification task, the model was provided with a list of categories and their descriptions. For Task 1 - Text Classification, models were additionally provided with short 1-2 sentence descriptions of each class. Genre descriptions were taken from the IMDb website, news category descriptions were written by the authors



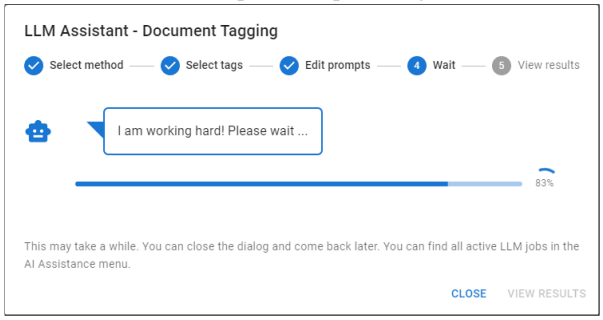
(a) Step 1: Method Selection



(b) Step 2: Tag Selection



(c) Step 3: Prompt Editing



(d) Step 4: Waiting

Figure 2: Steps 1-4 of the LLM Assistant Feature

Table 5: Full evaluation results of Task 1 - Document Classification

Model	Dataset	Language	Task	Precision	Recall	F1	Accuracy
gemma2	Tagesschau	de	coarse	**84.88**	**81.94**	**82.01**	**81.94**
llama3.1	Tagesschau	de	coarse	83.13	74.79	76.01	74.79
mistral-nemo	Tagesschau	de	coarse	83.55	77.62	78.92	77.62
gemma2	BBC	en	coarse	**76.59**	**76.58**	**74.85**	**76.58**
llama3.1	BBC	en	coarse	71.07	70.29	65.64	70.29
mistral-nemo	BBC	en	coarse	69.97	71.69	68.89	71.69
gemma2	imdb	en	coarse	**40.12**	**37.74**	**37.70**	**37.74**
llama3.1	imdb	en	coarse	38.78	33.00	32.57	33.00
mistral-nemo	imdb	en	coarse	39.66	36.70	36.04	36.70
gemma2	Tagesschau	de	fine	**68.00**	**59.00**	**60.00**	**59.43**
llama3.1	Tagesschau	de	fine	58.00	26.00	29.00	25.80
mistral-nemo	Tagesschau	de	fine	65.00	43.00	47.00	42.86
gemma2	BBC	en	fine	**73.00**	**65.00**	**66.00**	**65.47**
llama3.1	BBC	en	fine	57.00	46.00	39.00	46.07
mistral-nemo	BBC	en	fine	68.00	64.00	63.00	64.27
gemma2	imdb	en	multi-label	55.48	**52.23**	**52.35**	**8.58**
llama3.1	imdb	en	multi-label	52.12	42.38	43.95	6.88
mistral-nemo	imdb	en	multi-label	**57.51**	49.62	51.05	**8.58**

Table 6: Full evaluation results of Task 2 - Document Information Extraction

Model	Dataset	Language	Task	Exact Match	F1
gemma2	SQUAD1	en	extractive-qa	**75.73**	**87.19**
llama3.1	SQUAD1	en	extractive-qa	70.39	82.64
mistral	SQUAD1	en	extractive-qa	67.48	83.35
gemma2	SQUAD2	en	extractive-qa	**63.85**	**69.80**
llama3.1	SQUAD2	en	extractive-qa	48.82	56.25
mistral	SQUAD2	en	extractive-qa	50.81	60.64
gemma2	GermanQuAD	de	extractive-qa	**60.01**	**81.25**
llama3.1	GermanQuAD	de	extractive-qa	48.91	73.04
mistral	GermanQuAD	de	extractive-qa	47.34	73.18
gemma2	MUC4	en	template-filling	36.69	36.70
llama3.1	MUC4	en	template-filling	**40.62**	**40.63**
mistral	MUC4	en	template-filling	23.86	23.87

Table 7: Full evaluation results of Task 3 - Span Classification

Model	Dataset	Language	Task	Precision	Recall	F1	Accuracy
gemma2	fewnerd	en	coarse	48.53	**55.72**	**51.15**	**83.73**
llama3.1	fewnerd	en	coarse	40.6	7.72	12.87	79.46
mistral	fewnerd	en	coarse	**49.95**	18.97	26.32	81.12
gemma2	germanler	de	coarse	21.85	**32.42**	**24.82**	**86.32**
llama3.1	germanler	de	coarse	11.84	1.67	2.78	81.21
mistral	germanler	de	coarse	**28.6**	11.79	16.11	82.41
gemma2	fewnerd	en	fine	**47.92**	**40.78**	**42.31**	**83.23**
llama3.1	fewnerd	en	fine	38.97	6.53	10.54	79.2
mistral	fewnerd	en	fine	39.53	12.68	17.64	79.86
gemma2	germanler	de	fine	32.58	**22.49**	**21.37**	**84.48**
llama3.1	germanler	de	fine	14.36	1.78	3.02	81.54
mistral	germanler	de	fine	**39.91**	8.64	13.13	82.22
gemma2	quotations	de	quotations	**36.4**	**25.61**	**29.81**	**90.29**
llama3.1	quotations	de	quotations	33.16	10.88	15.18	88.17
mistral	quotations	de	quotations	23.03	10.69	13.71	86.24

Table 8: Full evaluation results of Task 4 - Text Generation. We report Rouge (R), Exact Match (EM), F1 and METEOR scores on two text generation tasks: disfluency correction (CORR) and summarization (SUM).

Model	Dataset	Lang	Task	R-1	R-2	R-L	R-Lsum	EM	F1	METEOR
gemma2	DisfluQA	en	CORR	**83.28**	**71.99**	**80.66**	**80.67**	**21.77**	**83.10**	**89.08**
llama3.1	DisfluQA	en	CORR	78.94	65.33	75.80	75.80	11.45	78.46	85.69
mistral	DisfluQA	en	CORR	78.21	63.10	75.24	75.24	15.21	77.77	81.38
gemma2	DISCO	en	CORR	**92.40**	**85.56**	**91.63**	**91.64**	**64.44**	**92.59**	**91.55**
llama3.1	DISCO	en	CORR	91.20	84.57	90.58	90.58	64.01	91.43	91.07
mistral	DISCO	en	CORR	85.44	73.14	84.39	84.36	41.87	85.52	85.02
gemma2	DISCO	de	CORR	**77.03**	**63.48**	**75.15**	**75.16**	**34.04**	**77.02**	**76.18**
llama3.1	DISCO	de	CORR	66.87	52.73	65.21	65.24	25.94	66.77	65.83
mistral	DISCO	de	CORR	73.25	54.28	71.57	71.60	20.64	73.35	72.62
gemma2	CNNDM	en	SUM	34.98	11.21	22.51	28.43	0.00	33.30	29.73
llama3.1	CNNDM	en	SUM	**36.44**	**12.44**	**23.09**	**29.38**	0.00	**34.88**	**31.97**
mistral	CNNDM	en	SUM	30.81	9.44	20.49	25.48	0.00	30.19	24.01
gemma2	MLSUM	de	SUM	20.81	6.52	15.40	15.41	0.00	20.50	17.63
llama3.1	MLSUM	de	SUM	**21.68**	**7.43**	**16.04**	**16.03**	0.00	**21.41**	**19.42**
mistral	MLSUM	de	SUM	18.81	6.27	13.59	13.60	0.01	18.51	19.62

Table 9: The prompts used in both Evaluation and Implementation. {} are placeholders for task-dependent input.

Task	Prompt Template
Document Classification	Please classify the document in one (all appropriate) of the following categories: {} Please answer in this format. You are not required to provide any reasoning. Category: <category> Reason: <reason> e.g. Category: News Document:{}
Extractive QA	Please extract the answer to the following question from the context below: Context: {} Question: {} Please answer in this format. If the question cannot be answered from the context, respond with 'Not answerable'. You are not required to provide any reasoning. Answer: <answer> or <not answerable> Reasoning: <reasoning> e.g. Answer: 42 Remember, the answer MUST be extracted verbatim from the text, do not generate it!
Template Filling	I prepared a list of slots. The slots are: {} Please extract the corresponding information (if any) from the following text: {} Please answer in this format. If the text does not include information about a specific slot, leave it empty. <Slot>: <extracted information> e.g. Incident: Arson, Perpetrator: John Doe, Weapon: Matches Remember, you MUST extract the information verbatim from the text, do not generate it!
Summarization	Please summarize the text below concisely, highlighting the most important information. Try to use about {} words only. Text: {} Respond in the following format: Summary: <summarized text> e.g. Summary: Theia was hit by a car ... Remember, you MUST summarize the original text, do not generate new facts!
Disfluency Correction	Please remove all disfluencies from the noisy, disfluent text below. Keep the text close to the original, but ensure it is read fluently. Text: {} Respond in the following format: Fluent text: <the corrected text> e.g. Fluent text: This picture looks great. Remember, you MUST keep to the original text; do not generate new content!
Span Classification	I prepared a list of categories/information. The categories are: {} Please extract fitting text spans from the following text: {} Respond in the following format: <category>: <extracted text> e.g. Art: Mona Lisa, Building: Eiffel Tower Remember, you MUST extract the information verbatim from the text, do not generate it!

Table 10: Taxonomies of the datasets used in Task 1, 2, and 3.

Task 1 Dataset	**Document Classification** Classes (coarse - fine)
BBC	UK - england, scotland, wales, ireland, politics World - africa, asia, australia, europe, latin, us, middle-east Sport - boxing, cricket, footbal, formula1, rugby, tennis Misc - business, education, elections, entertainment, arts, health, science, technology
Tagesschau	Inland - Deutschlandtrend, Gesellschaft, Innenpolitik, Mittendrin Ausland - Afrika, Amerika, Asien, Europa, Ozeanien Wirtschaft - Börse, Finanzen, Konjunktur, Technologie, Unternehmen, Verbraucher, Weltwirtschaft Wissen - Forschung, Gesundheit, Klima, Technologie
IMDB	Action, Adventure, Animation, Biography, Crime, Family, Fantasy, Film-Noir, History, Horror, Mystery, Romance, SciFi, Sports, Thriller, War
Task 2 Dataset	**Document Information Extraction** Slots
MUC4	Incident: One of 'Arson', 'Attack', 'Bombing', 'Kidnapping', 'Robbery' Perpetrator: An individual perpetrator Group Perpetrator: A group or organizational perpetrator Victim: Sentient victims of the incident Target: Physical objects targeted by the incident Weapon: Weapons employed by the perpetrators
Task 3 Dataset	**Span Classification** Classes (coarse - fine)
fewNERD	Art - broadcastprogram, film, music, other, painting, writtenart Building - airport, hospital, hotel, library, other, restaurant, sportsfacility, theater Event - attack, disaster, election, other, protest, sportsevent Location - GPE, bodiesofwater, island, mountain, other, park, road Organization - company, education, government, media, other, political party, religion, show organization, sportsleague, sportsteam Other - astronomy, award, biology, chemical, currency, disease, educational degree, god, language, law, living thing, medical Person - actor, artist, athlete, director, other, politician, scholar, soldier Product - airplane, car, food, game, other, ship, software, train, weapon
German-LER	Person - Anwalt, Richter Ort - Land, Stadt, Straße, Landschaft Organisation - Unternehmen, Institution, Gericht, Marke Norm - Gesetz, Verordnung, EU Norm Regulierung - Vorschrift, Vertrag Rechtsprechung Literatur
Quotations	Sprecher Direkte Rede