194.093 Natural Language Processing and Information Extraction.

Topic 13: Retrieval-Augmented Generation for German official documents. Project Summary.

Group 25: Diara Rinnerbauer 01449793, Stefan Sick 11924414, Valentin Tian 12426893.

1. Work distribution

Initially, four students registered for the project, however, one member dropped out before the project started. The workload was therefore distributed among the three remaining members with responsibilities shared across all major stages. While some members took the lead in specific technical aspects (e.g., Stefan Sick – ML and SLM-based RAGs, Diara Rinnerbauer – preprocessing and rule-based RAG, Valentin Tian – Verbatim and rules-based RAG), all team members were actively involved in the evaluation of the systems, the preparation of the final presentation and the writing of the project report.

2. Project scope & Objectives

Our team developed a Retrieval-Augmented Generation (RAG system) - essentially a digital librarian that searches through official documents to provide precise answers to user questions. The dataset is part of a large German-speaking corpus of parliamentary protocols from three different centuries, on a national and federal level from the countries of Germany, Austria, Switzerland and Liechtenstein. For this task we chose a subset of 100 documents from Swiss parliamentary protocols.

3. Challenges

Two of the primary challenges of the task were handling domain-specific language and ensuring high factual accuracy in official contexts without any hallucinations. Upon reviewing the documents, we also noted that the topics can be discussed in several files and it is not possible to attribute unique themes to one document. This increases the complexity of the task, and it is crucial for the RAG system to correctly identify the context of the question and the text given to answer the questions without combining dispersed text fragments from multiple documents.

4. Implemented solutions

To find the most suitable approach for the RAG system in the domain of official parliamentary documents, we implemented and compared four complementary RAG strategies. These approaches differ in how information is retrieved and how answers are produced.

First, we implemented a rule-based approach for document representation and retrieval combined with ChatGPT 4.1-mini (cloud LLM) for answering questions. This approach is well suited for documents with precise facts.

Second, we implemented an ML-based retrieval approach with the same LLM for answering. This setup allows us to retrieve the relevant documents even if the wording of the question differs from the one used in the source.

Third, we implemented a local small language configuration with an embeddings model for retrieval and Llama 3.2 (local SLM) for answering. This approach explores the trade-off between data privacy and answer quality by avoiding the use of the third-party cloud services.

Finally, we implemented a state-of-the-art hallucination-resistant Verbatim RAG using the same semantic retrieval model as in the second approach and the Verbatim itself for the answering. By avoiding any generative rewriting, this approach significantly reduces the risk of hallucinations, however it also limits the system` s ability to generalize or rephrase information.

5. Information Retrieval Efficiency

We compared three retrieval approaches: lexical rule-based and two semantic ML-based approaches based on different embedding models. The rule-based approach achieved the best performance, retrieving 89% of relevant documents within the top five results, which highlights its suitability for parliamentary documents where exact identifiers, dates, and factual precision are critical.

6. Answer Quality

The manual evaluation used a pre-defined test set of 20 domain-specific questions to analyse quality of provided answers.

Both the cloud and local versions passed the "hallucination test" by correctly admitting when information was missing, but they struggled significantly with **precise voting counts** when multiple figures appeared on the same page. We also observed a "noise" issue where the cloud-based modes provided helpful but long outside info, whereas the **local model** prioritized privacy at the cost of performance, often failing to grasp the context entirely and returning a null answer.

7. Conclusion

The project successfully demonstrated the effectiveness of a Retrieval-Augmented Generation system for processing complex German parliamentary protocols. All systems proved robust against hallucinations, correctly identifying when information was missing from the dataset. However, the quality of the final response depends on successful document retrieval, highlighting that a universal RAG system that performs best across all tasks does not exist; strategies must be optimized for specific use cases. Even if the correct document is found, the presence of multiple similar figures can lead to factually incorrect answers.

8. External Resources

1. Jurafsky, D., & Martin, J. H. (2026). Speech and Language Processing (3rd ed. draft)., https://web.stanford.edu/~jurafsky/slp3/
2. Kovacs, A. Build Hallucination-Free RAG with Verbatim. Hugging Face., https://huggingface.co/blog/adaamko/verbatimrag https://github.com/KRLabsOrg/verbatim-rag