TECHNISCHE UNIVERSITÄT WIEN

Retrieval-augmented Generation for German official documents

Topic 13. Retrieval-augmented Generation for German official documents

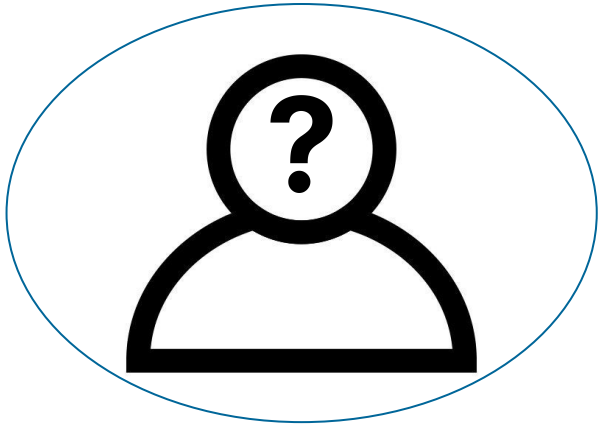
Group 25: Stefan Sick, Diara Rinnerbauer, Valentin Tian

Date: 16/01/2026

Content

- Our Team & Work Distribution
- **Motivation**
- Data
- **Preprocessing Pipeline**
- Milestone 1
- **Milestone 2**
- Additional Approaches
- **Evaluation**
- Key Takeaways

Our Team & Work Distribution



Stefan Sick

- Preprocessing
- Initial Code Framework
- LLM-based RAG
- SLM-based RAG
- Evaluation
- Presentation

Diara Rinnerbauer

- Data Gathering
- Preprocessing
- TF-DF based RAG
- Code enhancement
- Evaluation
- Presentation & Report

Valentin Tian

- Preprocessing
- TF-DF based RAG
- LLM-based RAG
- Verbatim RAG
- Code enhancement
- Evaluation
- Presentation & Report

Motivation

- Aim: Develop a RAG system to provide precise answers from official parliamentary protocols.

Data

- Dataset
  - German Parliamentary Corpus (GerParCor)
- Description
  - the largest 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.
- Subset
  - CH Nationalrat, 100 documents (3.77 Million Tokens)
- Challenge
  - Handling domain-specific language and ensuring high factual accuracy in official contexts.

Preprocessing Pipeline (Milestone 1)

- Data extraction
  - Raw texts from 100 .xmi documents
- NLP Processing
  - Stanza German pipeline was used for tokenization, multi-word token (MWT) expansion, POS tagging, and lemmatization
- Standardization
  - Converted into the CoNLL-U format

Classification Task (Milestone 2)

- Rule-based baseline
  - Utilizes TF-IDF (BM25) for vector processing and indexing
- Machine learning baseline
  - Employs the Multilingual E5 Text Embeddings model from Hugging Face for semantic search
- Core components
  - Built with LlamaIndex and LangChain, using Gemini 2.5 Flash as the generative LLM

Questions Set

- Pre-defined set of 20 questions that include:
  - Domain-specific questions derived from the randomly selected documents
  - Hallucination check: question that cannot be answered based on the given documents
  - Questions phrased similarly to the text but contain erroneous implication, e.g. specifics that are not part of the text

Questions Groups	
Facts with numbers	**10**
Definitional facts	**6**
No answer in data	**4**

Automated Evaluation Methodology (Milestone 2)

- LLM-as-a-Teacher
  - LLM generates question-answer pairs directly from the source documents
- RAG Execution
  - The system processes these questions without access to the "truth" labels
- LLM-as-a-Judge
  - LLM scores responses from 0 (False) to 5 (Perfect) based on accuracy and reasoning

Baseline Evaluation Results (Milestone 2)

Quantitative

Metric	Rule based	Machine learning
Mean score	**3.6 / 5**	**4.5 / 5**
Delta	-	**+.09**
Relative improvement	-	**+25%**
Error reduction	-	**64%**

Qualitative

- **Correctness:**
  - Most domain-specific questions were answered accurately
- **Numerical sensitivity:**
  - Both systems struggled with precise voting counts when multiple figures appeared in the context
- **Hallucination check:**
  - The models successfully passed tests with out-of-scope questions by stating information was not present

Additional Approaches

- **Verbatim RAG* by KRLabs**
  - No hallucination
  - Direct derivation
- **Local Small Language Model (Llama 3.2)**
  - Open-source/Free
  - Indexing and Retrieval Pipeline with ReRanking
  - Privacy

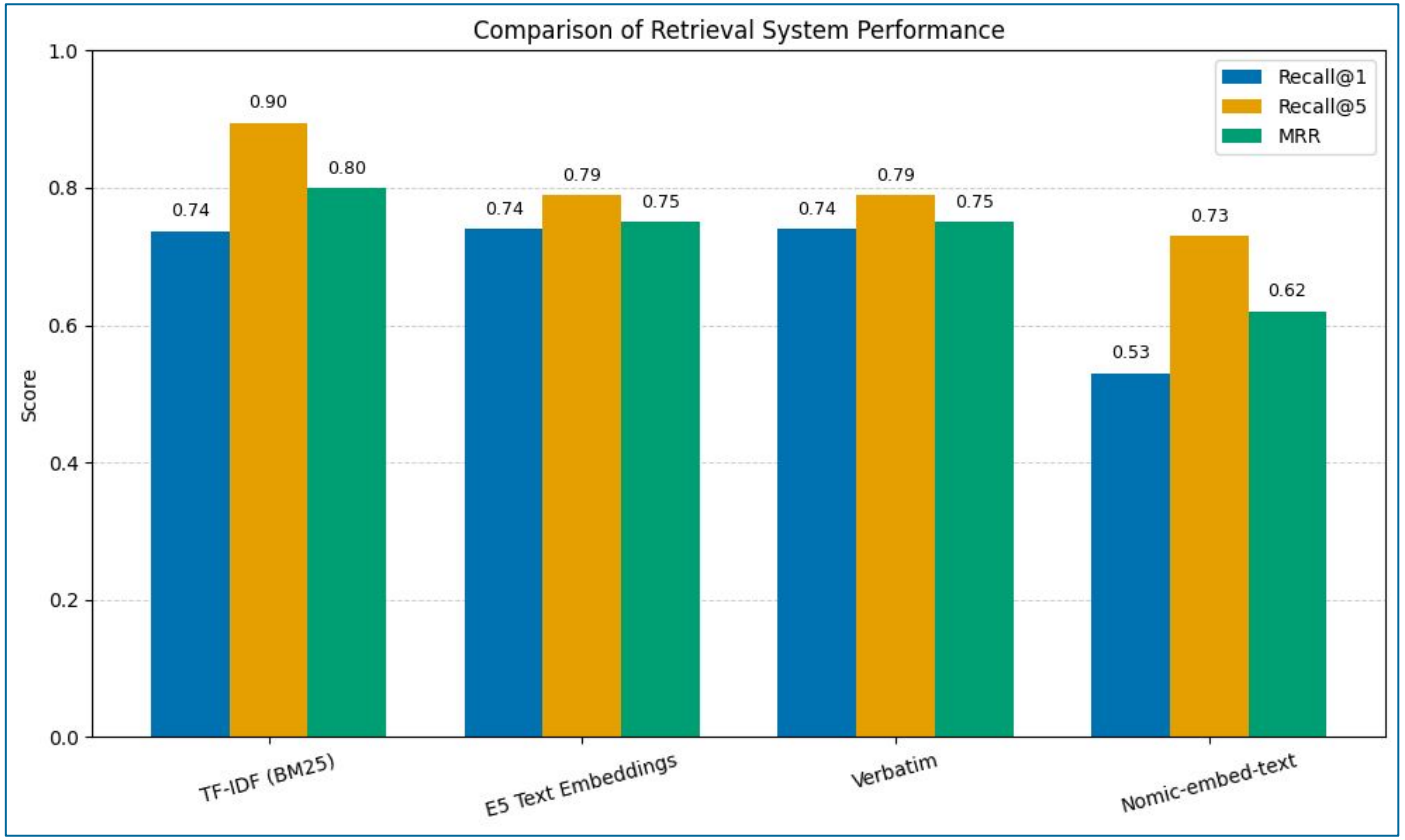
**https://github.com/KRLabsOrg/verbatim-rag

RAG Output Format


```
[
  {
    "id": 1,
    "question": "Was war der Ausgang der Abstimmung der Staatspolitischen Kommision über die Handhabung von Volksinitiativen die Grundrechte verletzen?",
    "source_id": "20111220.gz",
    "expected_answer": "13 Stimmen Annahme, 8 Stimmen dagegen, 2 Enthaltungen.",
    "answer": "- Die Staatspolitische Kommission (SPK) stimmte mit 13 zu 8 Stimmen bei 2 Enthaltungen für eine wichtige Etappe im Umgang mit Volksinitiative
    "answer_source_id": [
      "20111220.gz",
      "20130321.gz"
    ],
    "answer_context_snippets": [
      {
        "source_id": "20111220.gz",
        "snippet": "Die Staatspolitische Kommission schlägt Ihnen mit 13 zu 8 Stimmen bei 2 Enthaltungen eine wichtige Etappe in einem langen Prozess vor, d
      },
      {
        "source_id": "20111220.gz",
        "snippet": "Als Begründung wird darauf hingewiesen, dass mit dieser Motion die Wahrscheinlichkeit der Einreichung, des Zustandekommens und der Annahr
      },
      {
        "source_id": "20130321.gz",
        "snippet": "Unsere Kommission hat die Vertreterinnen und Vertreter des Initiativkomitees am 10. Januar angehört und sich danach entschieden, der Init
      },
      {
        "source_id": "20111220.gz",
        "snippet": "Der Ständerat schlägt vor, analog zum Postulat Heim 09.3118, den Bundesrat eine Vorlage ausarbeiten zu lassen, welche dafür sorgt, dass r
      },
      {
        "source_id": "20111220.gz",
        "snippet": "Beim Kerngehalt der Volksrechte wird es auch sehr, sehr gefährlich. Wenn ich von Amputation und vom Angriff auf die Volksrechte und die
      }
    ]
  },
```

Information Retrieval Evaluation Results

System	Recall@1	Recall@5	MRR
TF-DF (BM25)	**0.737**	**0.895**	**0.8**
E5 Text Embeddings	**0.74**	**0.79**	**0.75**
Verbatim	**0.74**	**0.79**	**0.75**
Nomic-embed-text	**0.53**	**0.73**	**0.62**



Answer Quality Analysis:

- 1 – True
- 0.5 - Almost True (Noise, additional knowledge/opinion, false citation)
- 0 - False

Question Group	TF-DF (BM25)	E5 Text Embeddings	Verbatim	Nomic-embed-text
Facts with numbers	6/8	6.5/8	3.5/8	1.5/8
Definitional facts	7/10	6/10	6.5/10	4/10
No answer in data	2/2	2/2	2/2	2/2
Total	15/20	14.5/20	12/20	7.5/20
Accuracy	***0.75***	***0.75***	***0.6***	***0.375***

Key Takeaways

- The quality of the final answer strongly depends on successful document retrieval.
- A robust RAG system must be able to explicitly highlight when the answer is not present in the dataset.
- Evaluating RAG systems is still a non-trivial task and in general manual.
- There is no universal RAG system that performs best across all tasks.
- Different retrieval and generation strategies are optimal for different use cases.

TECHNISCHE UNIVERSITÄT WIEN

Thank you for your attention!

TECHNISCHE UNIVERSITÄT WIEN

Retrieval-augmented Generation for German official documents

Topic 13. Retrieval-augmented Generation for German official documents

Group 25: Stefan Sick, Diara Rinnerbauer, Valentin Tian

Date: 16/01/2026