Legal Austrian RAG system

Mehdin Masinovic & Leon Beccard
25.04.2025

Agenda

1. Introduction and Requirements
2. Project architecture
3. Experiment Setup
4. Experiments
5. Key Findings and Conclusion

Introduction and Requirements

Introduction

Problem

Information about legal Austrian corpus on Viennese building regulations is (currently) too specific for any LLM.

Approach

Consider Retrieval Augmented Generation (RAG) to have a knowledge base to input with the user query and identify optimal configurations towards a provided Q&A dataset by our legal expert.

Introduction

Retrieval Augmented Generation - RAG

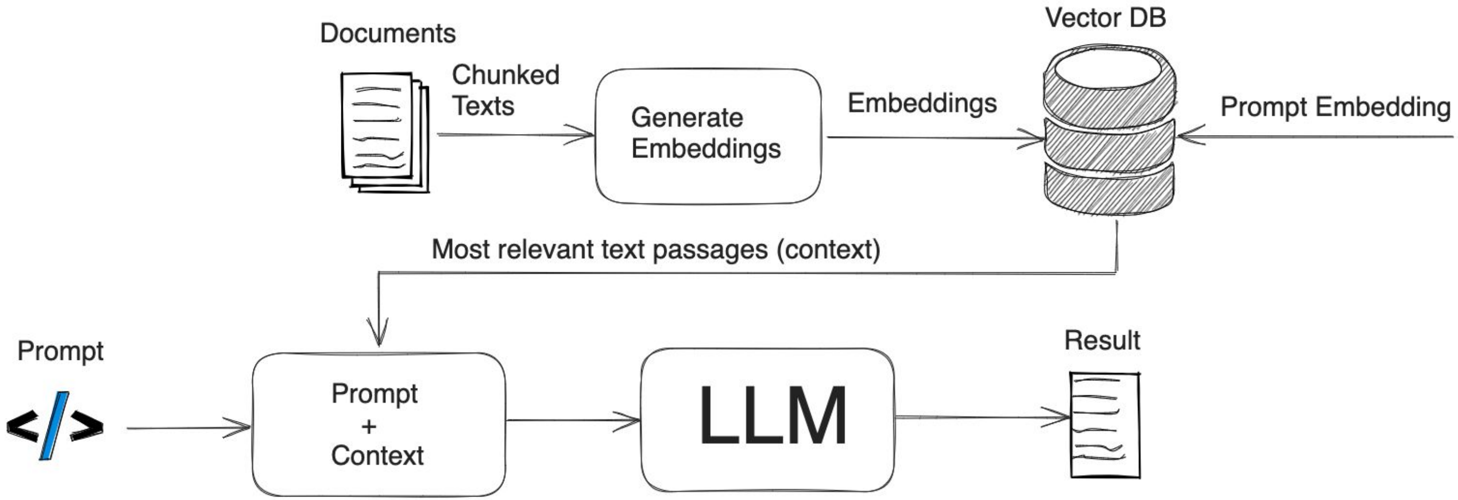
- Background from LLM limitations to reduce hallucinations and avoid outdated knowledge
- Combining retrieval systems with generative AI to improve accuracy and context-relevance in question answering
- Key components: Embedding model, Vector database, Chunking strategies, LLM

Hyperparameters in RAG

- Embedding model
- Reranking model (Cross-Encoder)
- Chunking size
- Number of best chunks selected
- Large Language model

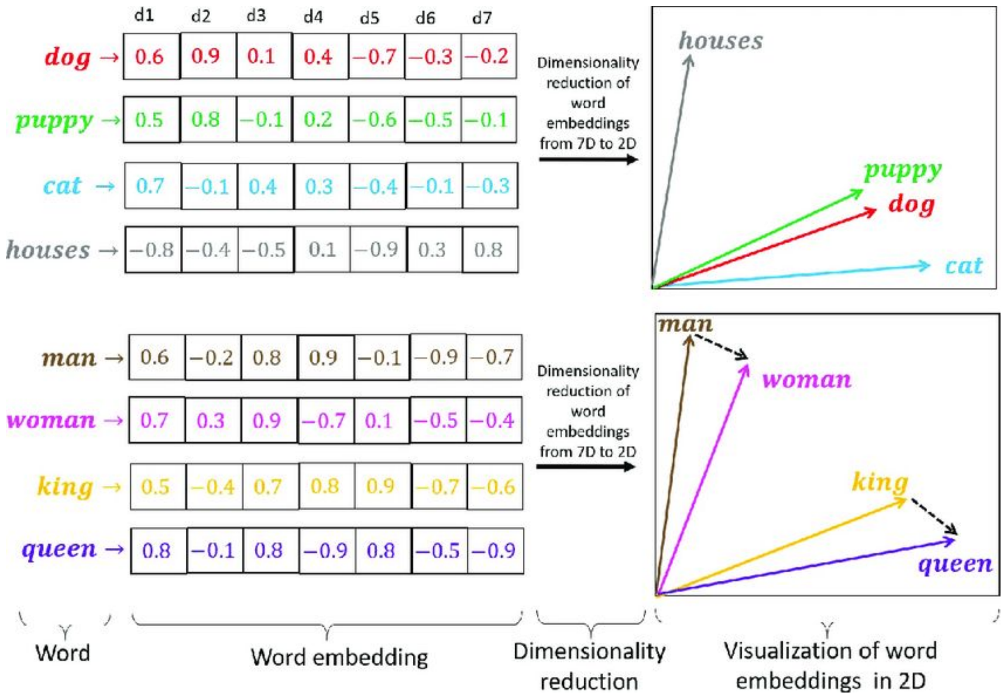
Introduction

Retrieval Augmented Generation - RAG



Introduction

Word embeddings



Requirements

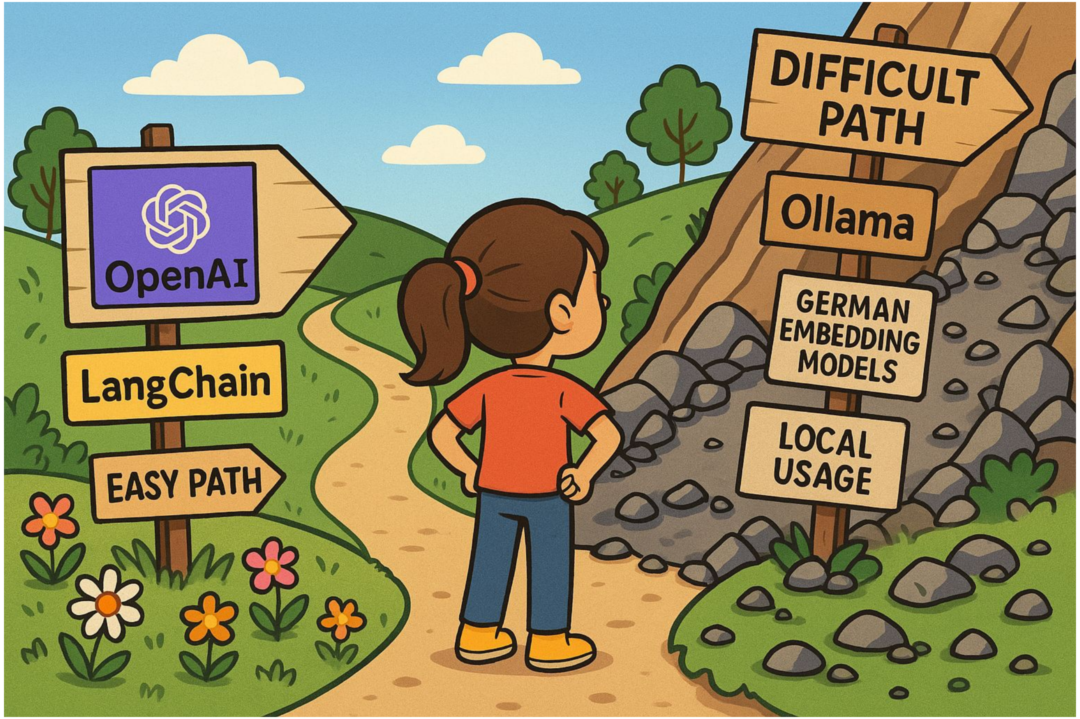
External requirements:

- Answers have to be in german
- Executability without external resources like Cloud Providers or LLMs
- Implement methods on your own instead of sophisticated libraries like Langchain

Internal requirements:

- Reproducible scientific evaluation on experiments
- Executability on a local device in regards of reasonable runtime (<5 minutes)

Requirements



Project architecture

Project architecture

Data Source: Viennese building regulation legal text corpus
Dataset: Q&A dataset by legal expert with 11 Q&A pairs
Backend: Python with FastAPI
Vector Database: Elasticsearch
Deployment: Docker container
Evaluation: Deepeval, Individual evaluation method

Embedding model: jinaai/jina-embeddings-v2-base-de
Cross Encoder: cross-encoder/msmarco-MiniLM-L12-en-de-v1
LLM: gemma3:27b

Project architecture

Workflow

1. Load legal text document
2. Chunk text using different strategies and sizes
3. Generate embeddings for each chunk
4. Index chunks in Elasticsearch
5. Embed user query, retrieve relevant chunks, rerank with cross-encoder
6. Pass context to LLM for final answer generation
7. Evaluate with our evaluation methods and reference answers

Project architecture

LLM-as-a-judge - Evaluation with Deepeval

- **Answer Correctness**: Accuracy of the generated response
- **Answer Relevance**: How well the answer addresses the query
- **Context Quality**: Relevance and usefulness of retrieved context
- **Faithfulness to Context**: Whether the answer stays true to the retrieved information
- **Combined Score**: Aggregate metric reflecting overall performance

Individual Evaluation based on Keywords

- **Score measurement:** Find the important keywords in answers or contexts like a paragraph number
- **Answer/Retrieval score:** Binary score of 1 if keyword is found or 0 otherwise
- **Total score:** Maximum of 11 points possible per Q&A pair

Experiment Setup

Experiment Setup

Hardware Experiment Setup 1: Running on Server with GPU Nvidia GeForce GTX 660

Experiment 1

Comparing Embedding models

- jinaai/jina-embeddings-v2-base-de
- paraphrase-multilingual-MiniLM-L12-v2

Comparing chunk size

- Multiple sizes from 0,125 kB to 128 kB

Comparing LLM models

- llama3-chatqa:8b
- gemma3:27b

Experiment Setup

Hardware Experiment Setup 2: Lenovo Thinkpad Gen2 32GB RAM Intel Core i7 - CPU only

Experiment 2

Comparing chunking method

- Splitting at maximum chunk size
- Splitting by article
- Splitting by subarticle

Comparing number of selected chunks

- Top 3 best chunks
- Top 5 best chunks
- Top 10 best chunks

Comparing LLM models

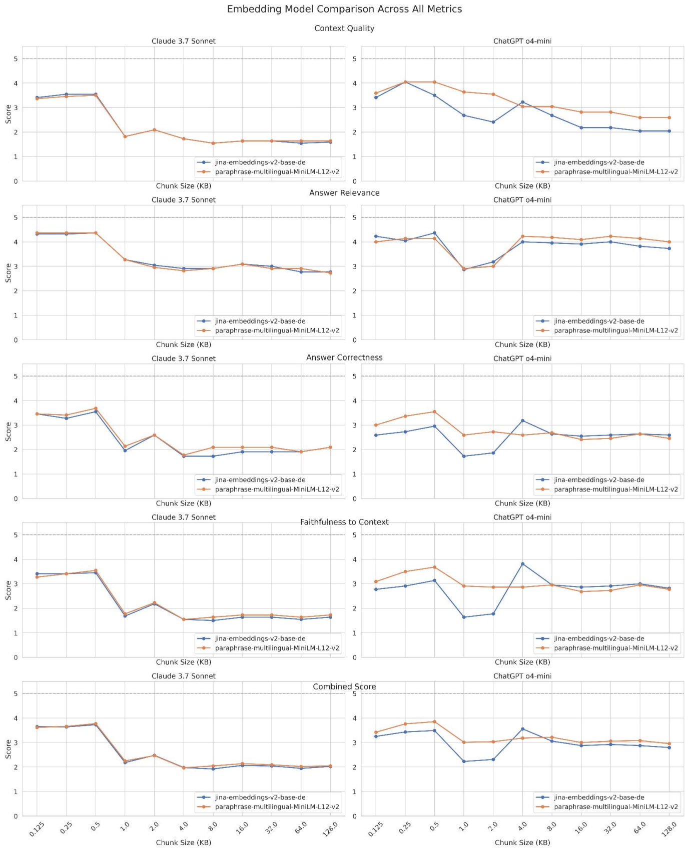
- llama3.2
- llama3-chatqa:8b
- deepseek-r1:latest

Experiments

Experiments

Embedding models

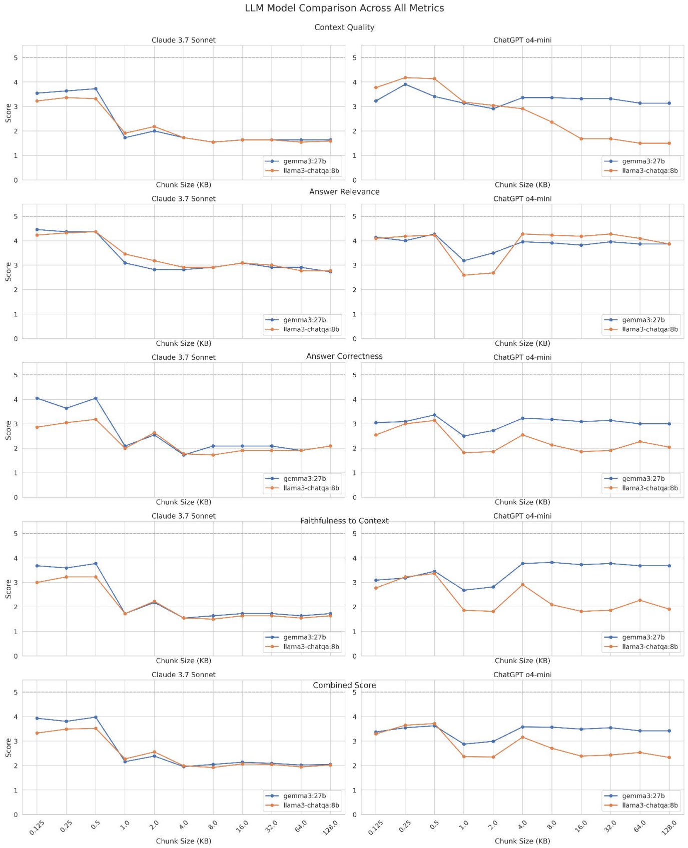
- Both models show similar performance patterns
- Small chunks (0.125KB - 0.5KB) deliver best overall performance
- Significant performance drop at 1.0KB chunk size
- jina-embeddings model shows more consistent patterns



Experiments

Large Language models

- gemma3:27b generally outperforms llama3-chatqa:8b
- Both models show similar response to chunking strategies
- Optimal chunk size consistent across both models (0.5KB)
- Performance gap consistent across different metrics



Experiments

Small chunks (0.125KB - 0.5KB):

- Highest performance across all metrics
- Better context quality
- Improved answer relevance

Medium chunks (1KB - 4KB):

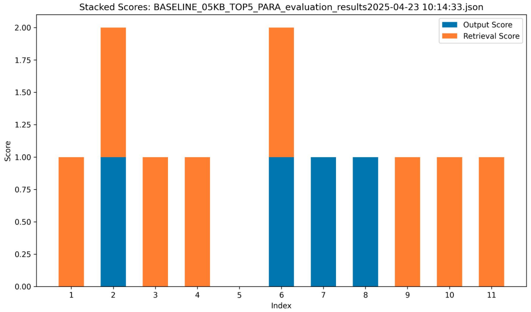
- Performance drop, especially in context quality
- Varied impact on faithfulness

Large chunks (8KB+):

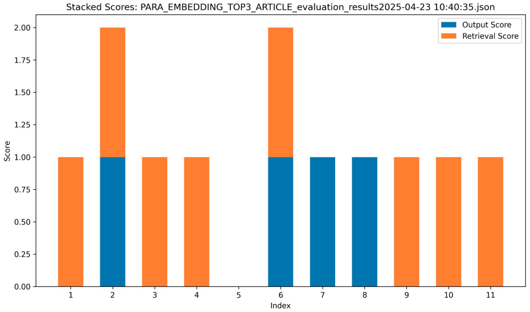
- Lowest overall performance
- Poor context quality
- Lower answer correctness

Experiments

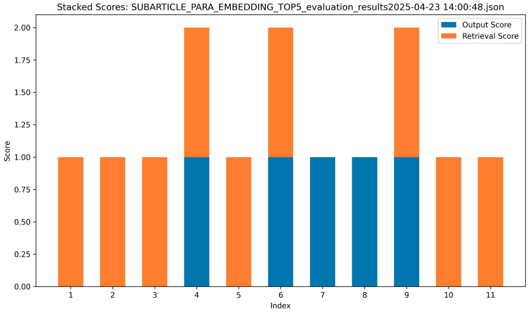
Chunking splitting method



Baseline with splitting at 0,5 KB chunk size with top 5 chunks



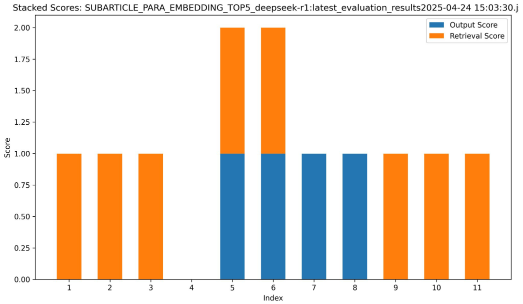
Splitting at each article with top 3 chunks



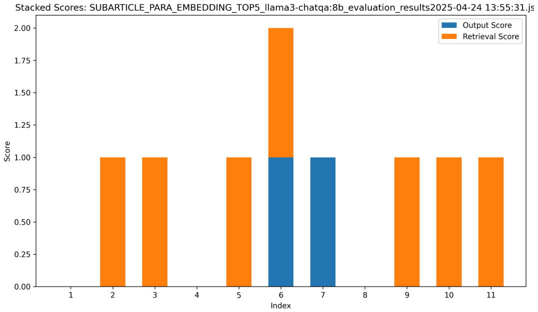
Splitting at each subarticle with top 5 chunks

Experiments

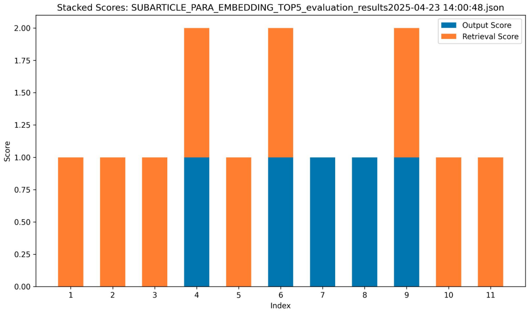
Model comparison



Deepseek with subarticle splitting and top 5 chunks



llama3-chatqa with subarticle splitting and top 5 chunks



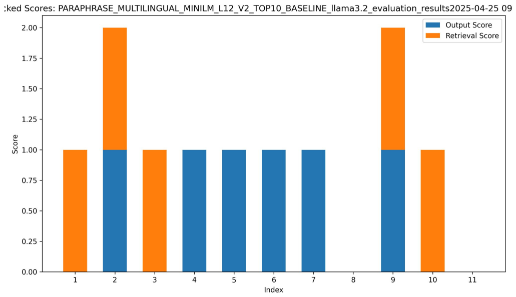
llama3.2 with subarticle splitting and top 5 chunks

Experiments

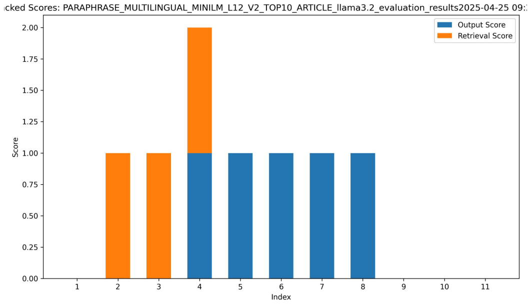
Realisation… Experiments executed with irreproducible retrieving part

Experiments

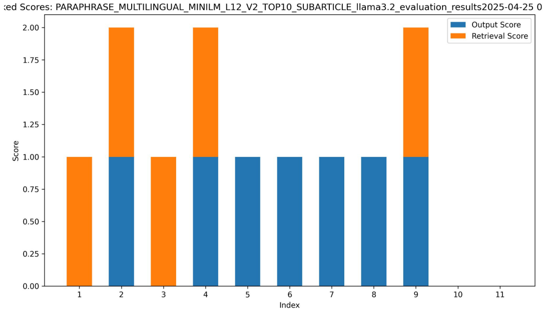
Reproducible results after vector normalization and random seeds



Baseline with splitting at 0,5 KB chunk size with top 10 chunks



Splitting at each article with top 10 chunks



Splitting at each subarticle with top 10 chunks

Key findings

Key findings

Optimal Configuration

- Embedding: paraphrase-multilingual-MiniLM-L12-v2
- LLM on GPU: gemma3:27b
- LLM on CPU: llama3.2
- Chunk size: 0.5KB
- Top K chunks: 10

Context Size Impact

- Small chunks maintain semantic coherence while providing focused context
- Larger context dilute relevant information with noise

Key findings

Evaluator agreement

- Claude 3.7 Sonnet and ChatGPT o4-mini show similar rating patterns
- Some divergence in absolute scores but agreement on trends

Limitations and resource impact

- GPU needs seconds
- CPU needs minutes

Conclusion

Conclusion

Experiments

- Embedding model selection has less impact than chunking strategy
- More powerful LLM improve performance
- Evaluation method LLM-as-judge proves effective for systematic comparison
- Keyword evaluation method enables efficient comparison
- Smaller chunk sizes (0.5KB) provide optimal RAG performance for legal text
- Chunking strategy to split text per subarticle improved retrieval part

Conclusion

Progress

- Evaluating a RAG system remains difficult
- RAG systems quickly require heavier resources

Outlook

- Hybrid search with keywords in Elasticsearch
- Knowledge-Graph based retrieval
- Multi-Hop Search with an Agent
- Personal RAG systems

Thank you for your attention!
Still awake for a demo?