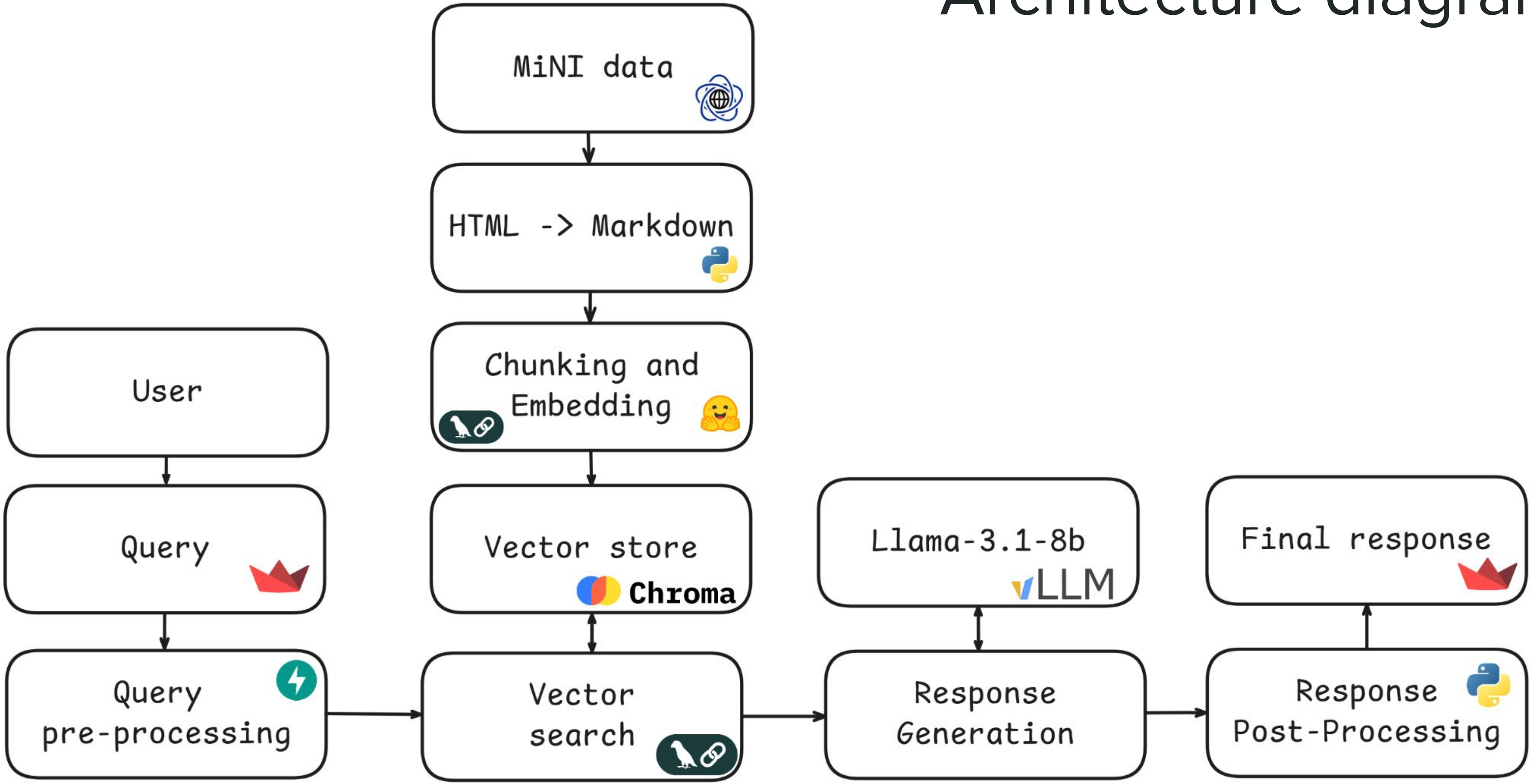
MiNI RAG Bot PoC

Mikołaj Gałkowski, Mikołaj Piórczyński, Julia Przybytniowska

Architecture diagram



MiNI data

HTML -> Markdown

Chunking and Embedding

User

Query

Vector store

Chroma

Llama-3.1-8b

vLLM

Final response

Query pre-processing

Vector search

Response Generation

Response Post-Processing

Data acquisition

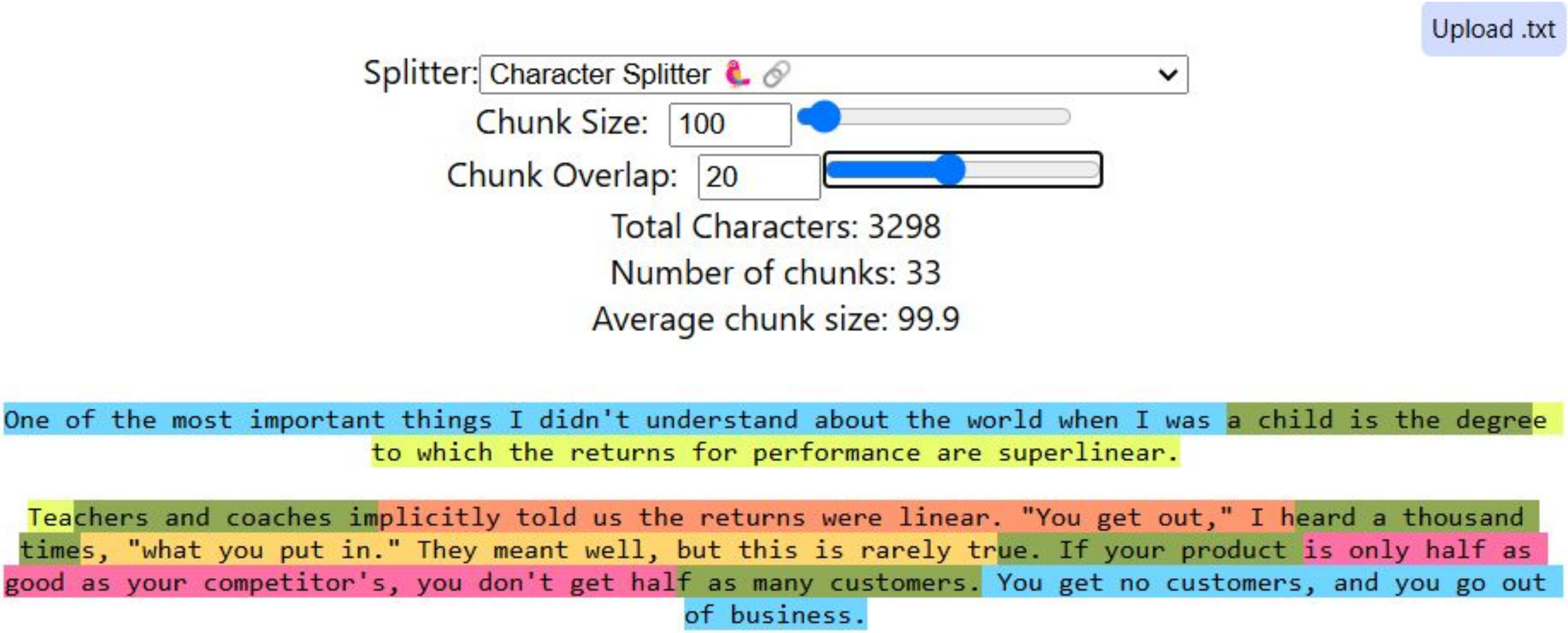
- Scraping MiNI page using

  BeautifulSoup

  - Recursive itteration
  - Visiting all links under https://ww2.mini.pw.edu.pl/

Data processing

- Conversion from **.html -> .txt**
- URL cleaning
- Chunking using ***RecursiveCharacterTextSplitter***
  - 1000 characters
  - 200 overlap



Splitter: Character Splitter 🦜 🔗

Chunk Size: 100

Chunk Overlap: 20

Total Characters: 3298

Number of chunks: 33

Average chunk size: 99.9

One of the most important things I didn't understand about the world when I was a child is the degree to which the returns for performance are superlinear.

Teachers and coaches implicitly told us the returns were linear. "You get out," I heard a thousand times, "what you put in." They meant well, but this is rarely true. If your product is only half as good as your competitor's, you don't get half as many customers. You get no customers, and you go out of business.

Source: https://chunkviz.up.railway.app/

Embedding model

- **gte-Qwen2-1.5B-instruct** is the latest model in the gte (General Text Embedding) model family. The model is built on Qwen2-1.5B LLM.
- Deployed as REST API built using FastAPI framework

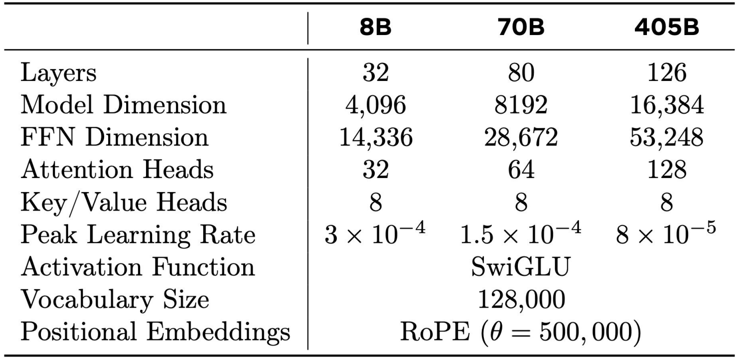
Vector store

- Chroma

- Open-source solution
- Able to handle large-scale vector data efficiently, ensures rapid retrieval of relevant chunks
- Designed to return the 5 closest chunks to the embedded query

LLM for generation

- Llama-3.1-8B-Instruct
- vLLM



	8B	70B	405B
Layers	32	80	126
Model Dimension	4,096	8192	16,384
FFN Dimension	14,336	28,672	53,248
Attention Heads	32	64	128
Key/Value Heads	8	8	8
Peak Learning Rate	3×10^{-4}	1.5×10^{-4}	8×10^{-5}
Activation Function		SwiGLU	
Vocabulary Size		128,000	
Positional Embeddings		RoPE $(\theta = 500,000)$	

Source: **The Llama 3 Herd of Models**

Created using Stable Diffusion 3.5 Large

User Interface

- Simplicity
- Seamless integration with other components
- User-friendly platform

Streamlit

Demo

Thank you for your attention!

Bibliography

- Woosuk Kwon and Zhuohan Li and Siyuan Zhuang and Ying Sheng and Lianmin Zheng and Cody Hao Yu and Joseph E. Gonzalez and Hao Zhang and Ion Stoica (2023)
  Efficient Memory Management for Large Language Model Serving with PagedAttention

- Aaron Grattafiori and Abhimanyu Dubey and … (2024)

  The Llama 3 Herd of Models