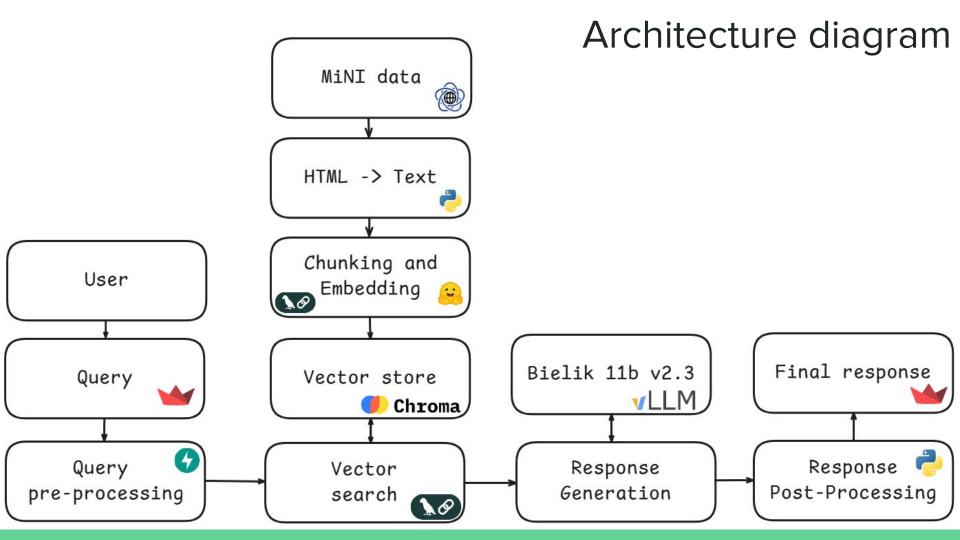
# MiNI RAG Bot

Mikołaj Gałkowski, Mikołaj Piórczyński, Julia Przybytniowska



### Data acquisition

Scraping MiNI page using



- 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

Upload .txt

Splitter: Character Splitter & S

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 degreet 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 models

- <u>gte-Qwen2-1.5B-instruct</u> is the latest model in the gte (General Text Embedding) model family. The model is built on Qwen2-1.5B LLM.
- <u>qte-Qwen2-7B-instruct</u>
- <u>jina-embeddings-v3</u> is a multilingual multi-task text embedding model, built based on the Jina-XLM-RoBERTa architecture
- Deployed as REST API built using FastAPI framework





Overall MTEB Polish leaderboard 2 O Metric Various, refer to task tabs Languages: Polish Credits: Rafal Poświata								
Rank 🔺	Model	Model Size (Million A	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (26 A	Classification Average (7 A datasets)	Clustering Average (1 datasets)
1	bge-multilingual-gemma2	9242	34.43	3584	8192	70	77.99	50.29
2	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	67.86	77.84	51.36
3	gte-Qwen2-1.5B-instruct	1776	6.62	1536	131072	64.04	72.29	44.59
4	jina-embeddings-v3	572	2.13	1024	8194	63.97	70.81	43.66

#### Vector store

# • Chroma

- Open-source solution
- Allows for handling large-scale vector data efficiently, ensures rapid retrieval of relevant chunks
- Returns the 5 closest chunks to the embedded query

## LLM for generation

- <u>Llama-3.1-8B-Instruct</u>
- speakleash/Bielik-11B-v2.3-Instruct



	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	$3 \times 10^{-4}$	$1.5 \times 10^{-4}$	$8 \times 10^{-5}$		
Activation Function	SwiGLU				
Vocabulary Size	128,000				
Positional Embeddings	RoPE ( $\theta = 500,000$ )				



Created using Stable Diffusion 3.5 Large



Source: <a href="https://bielik.ai/">https://bielik.ai/</a>

Source: The Llama 3 Herd of Models

#### User Interface

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



#### **Evaluation dataset**



- 21 evaluation pairs
  - question
  - golden answer
- Sample questions:
  - Kiedy jest otwarty dziekanat na MiNI?
  - Ile godzin należy przepracować w ramach praktyk na studiach I stopnia?
  - Kto jest dziekanem wydziału MiNI?
  - Kto jest prodziekanem ds. studenckich wydziału MiNI?

#### **Evaluation metrics**

- Context Recall measures how many of the relevant documents were successfully retrieved, is about not missing anything important
- Context Precision measures the proportion of relevant chunks in the contexts
- **Factual Correctness** determines the extent to which the generated response aligns with the reference. LLM first breaks down the response and reference into claims and then determines the factual overlap
- Faithfulness measures how factually consistent a response is with the retrieved context, a response is considered faithful if all its claims can be supported by the context.

#### Bielik-11B-v2.3-Instruct

	Additional Generation Kwargs*	Context Recall	Context Precision	Factual Correctness	Faithfulness
ato Owon? 7P instruct	-	0.4211	0.5282	0.4	0.5833
gte-Qwen2-7B-instruct	+	0.3567	0.5762	<u>0.5</u>	0.4286
gte-Qwen2-1.5B-instruct	-	0.5243	0.7194	-	<u>0.65</u>
gte-Qwenz-1.5D-instruct	+	0.6895	0.5024	0.5850	0.5303
jina-embeddings-v3	_	0.2958	0.4356	-	0.5407
Jilla-embeddings-vo	+	0.4398	0.7401	0.3350	0.6562

\*{"temperature": 0.2, "min\_p": 0.1}

#### Bielik-11B-v2.3-Instruct vs Llama-3.1-8B-Instruct

embedder: gte-Qwen2-1.5B-instruct

	Additional Generation Kwargs*	Context Recall	Context Precision	Factual Correctness	Faithfulness
Llama-3.1-8B-Instruct	-	0.3304	0.6152	0.3350	0.5093
	+	<u>0.5702</u>	0.6744	-	0.2847
Bielik-11B-v2.3-Instruct	-	0.5243	0.7194	-	0.6500
	+	0.6895	0.5024	0.5850	0.5303

# Demo

#### **MiNI RAG Chatbot**



Cześć! Jestem chatbotem, który może pomóc Ci w informacjach dotyczących studiów na wydziale MiNI. Zadaj mi dowolne pytanie!



Ile wynosi kwota stypendium rektora?



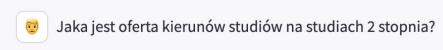
Kwota stypendium rektora oraz stypendium dla tegorocznych maturzystów wynosi 850 zł (kwota zeszłoroczna).



A kwota stypendium socjalnego?



Na podstawie dostarczonych informacji, minimalna wysokość stypendium socjalnego wynosi 1 400 zł (kwota zeszłoroczna), a maksymalna wysokość stypendium socjalnego to 1 800 zł (kwota zeszłoroczna). Dodatkowo, wysokość zwiększenia stypendium socjalnego wynosi 500 zł (kwota zeszłoroczna).



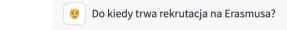


Oferta kierunków studiów na studiach II stopnia na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej obejmuje następujące kierunki:

- 1. Matematyka:
  - Specjalności:
    - Matematyka w Cyberbezpieczeństwie (MCB)
    - Indywidualne Studia Matematyczne (ISM)
- 2. Matematyka i Analiza Danych:

Specjalności:

- - Probabilistyka i Modelowanie (PRiMO)
- Statystyka Matematyczna i Analiza Danych (SMAD)
- 3. Informatyka i Systemy Informacyjne:
  - 3-semestralne studia (z tytułem inżyniera)
  - 4-semestralne studia (z tytułem licencjata)
- 4 Data Science



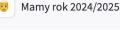


- 1. I tura rekrutacji: 4-10 marca 2023
  - 2. II tura rekrutacji: 18-19 marca 2023

Ostateczne wyniki rekrutacji zostaną ogłoszone 21 marca 2023. Po ogłoszeniu wyników, do 30 marca 2023 nastąpi rozesłanie nominacji do uczelni partnerskich.

Warto zauważyć, że terminy mogą się różnić w zależności od konkretnego programu i uczelni

partnerskiej. Zalecam sprawdzenie aktualnych informacji na stronie uczelni lub w odpowiednim regulaminie dotyczącym wymiany międzynarodowej.



zostały podane tylko dla roku akademickiego 2023/2024. Aby uzyskać informacje dotyczące roku akademickiego 2024/2025, zalecam sprawdzenie

aktualnych informacji na stronie uczelni lub skontaktowanie się z odpowiednimi biurami na Politechnice Warszawskiej.

Na podstawie dostarczonej informacji, nie jestem w stanie odpowiedzieć na pytanie dotyczące roku akademickiego 2024/2025, ponieważ dane dotyczące stypendiów i rekrutacji na Erasmusa

# Thank you for your attention!