

# AskWikidata

Natural language queries to Wikidata  
A naïve prototype



**WIKIMEDIA**  
DEUTSCHLAND

# About me

## Robert Timm

Senior Software Engineer at Wikimedia Germany  
Wikibase Suite Team

[robert.timm@wikimedia.de](mailto:robert.timm@wikimedia.de)  
[phabricator.wikimedia.org/p/roti\\_WMDE/](https://phabricator.wikimedia.org/p/roti_WMDE/)  
[github.com/rti](https://github.com/rti)

## Disclaimer:


I am not a ML Researcher / Data Scientist  
Never implemented search before





# Who is the current mayor of Paderborn?






Michael Dreier, CDU, since 2014

# Let's ask OpenAI GPT-4

= ChatGPT 4 ▾ 

---

 **You**  
Who is the current mayor of Paderborn?  


 **ChatGPT**  
The current mayor of Paderborn is Michael Dreier, a member of the Christian Democratic Union (CDU). He was elected in the mayoral election held on September 13, 2020.  
   

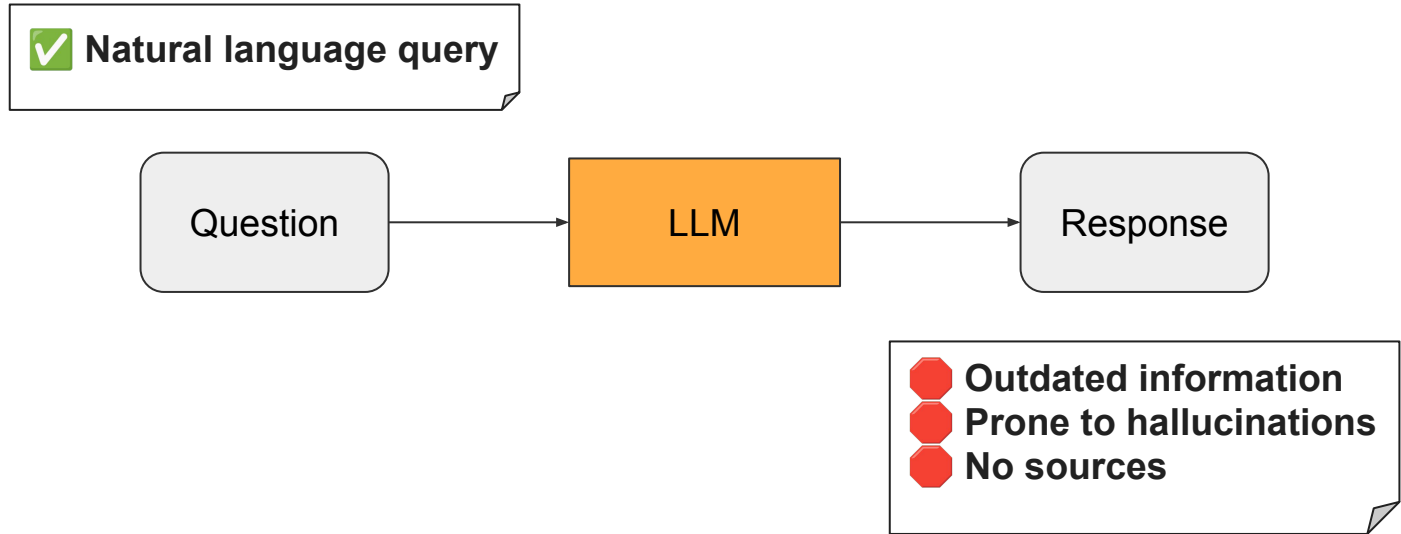
# What about Mistral 0.1?

Who is the current mayor of Paderborn?



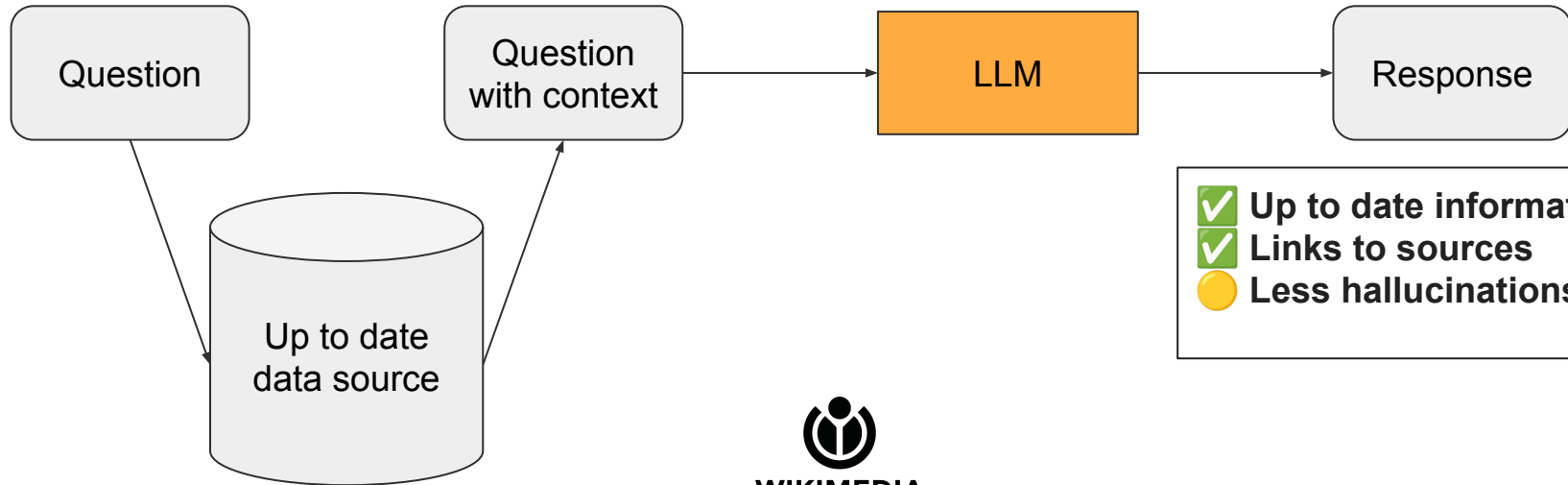
As of my last update, the current Mayor of Paderborn is Thomas Wodzinski. However, I don't have access to real-time information so please verify with a reliable source.

# LLM alone



# LLM with RAG

✓ Natural language query



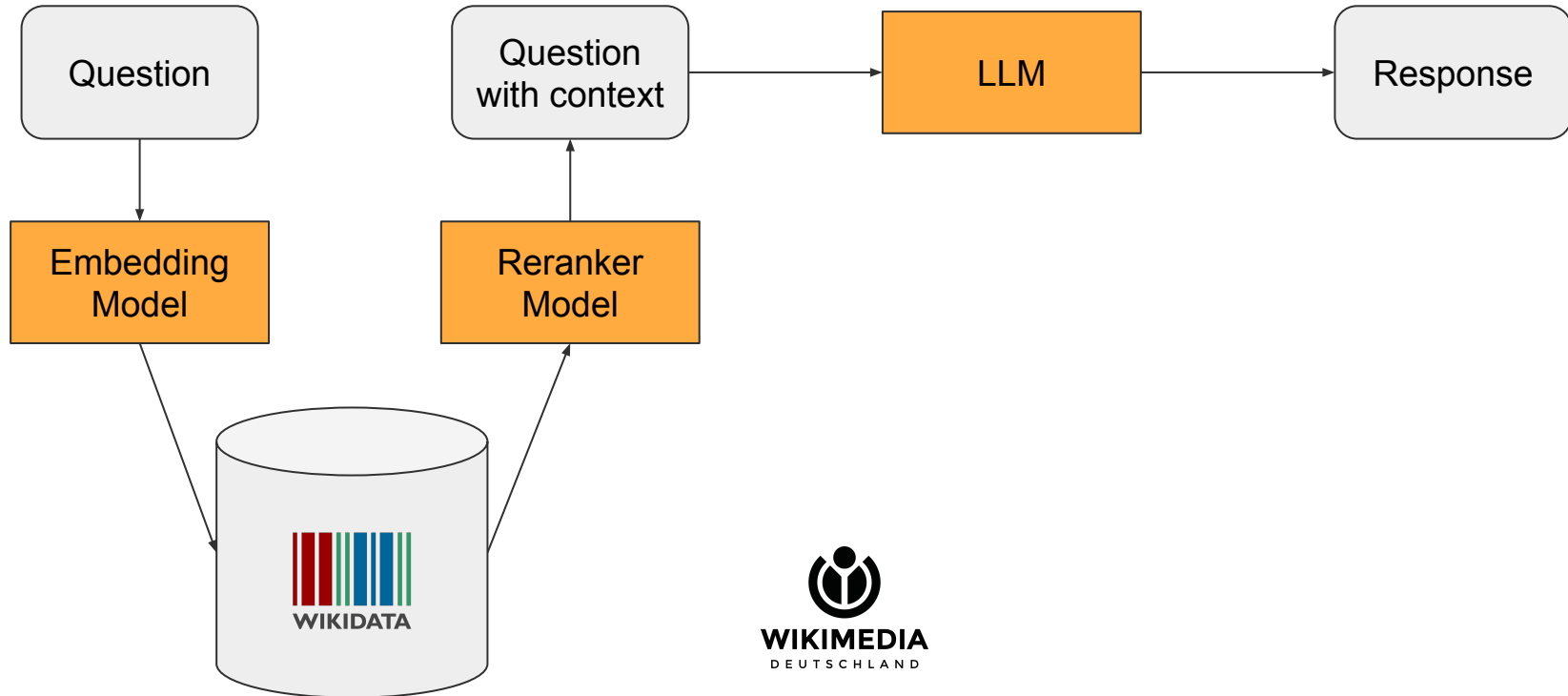
✓ Up to date information  
✓ Links to sources  
● Less hallucinations

# AskWikidata

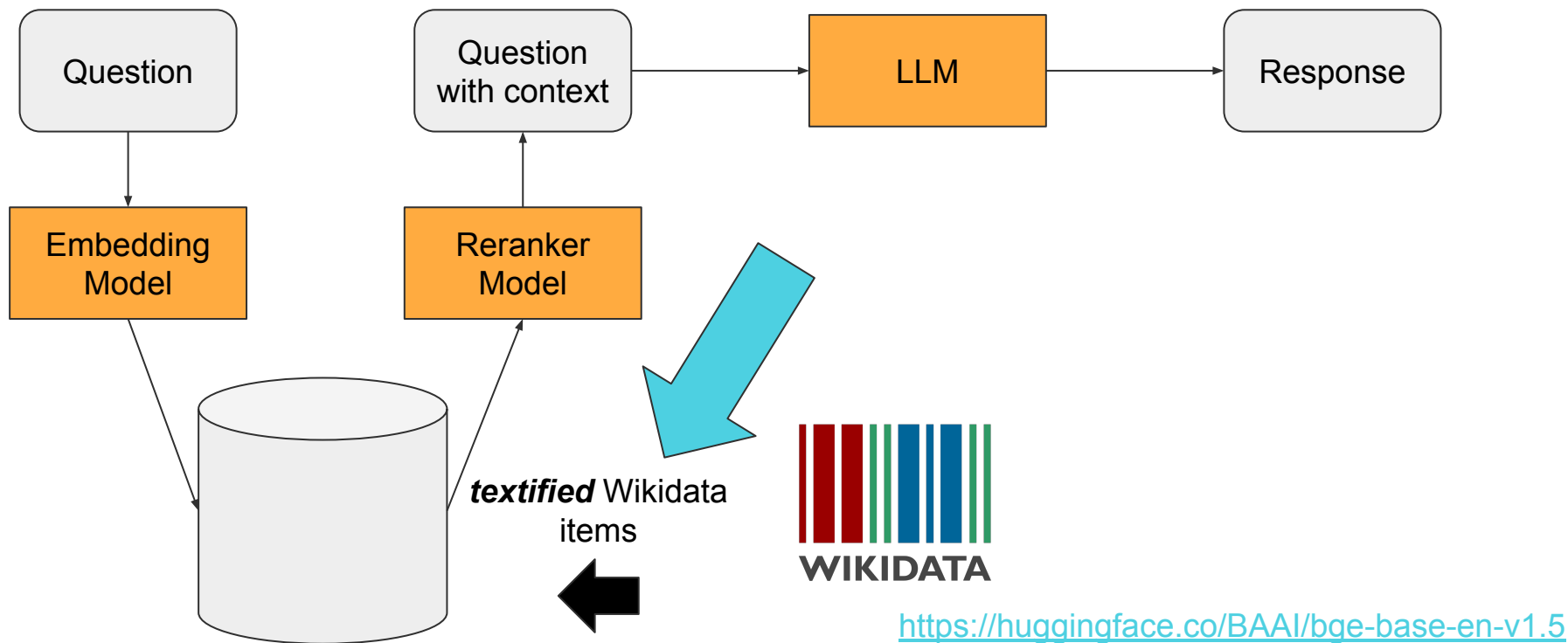
- A prototype I built in my freetime with support from Wikimedia Germany
- Open source software only
- LLM Mistral 7B (Apache-2.0 weights)
- Runs on
  - Consumer grade GPU, e.g. NVIDIA RTX 3070 8GB (from 2020)
  - Google Colab free tier (T4 GPU)
- Answers based on 13k city items from Wikidata



# Architecture



# Preprocessing



# Wikidata items as text

Paderborn: city in North Rhine-Westphalia, Germany

Paderborn shares border with Hövelhof.  
Paderborn shares border with Bad Lippspringe.  
Paderborn shares border with Altenbeken.  
[...]

Paderborn twinned administrative body Le Mans since 1967-06-03 until today.  
Paderborn twinned administrative body Bolton since 1975-01-01 until today.  
Paderborn twinned administrative body Belleville since 1990-01-01 until today.  
Paderborn twinned administrative body Pamplona since 1992-01-01 until today.  
Paderborn twinned administrative body Przemyśl since 1993-05-14 until today.  
[...]

Paderborn has license plate code PB.  
Paderborn has head of government Michael Dreier since 2014 until today.  
Paderborn population 149075 in 2017-12-31.  
Paderborn population 103705 in 1975-12-31.  
Paderborn population 150580 in 2019-09-30.  
Paderborn population 152531 in 2021-12-31.  
Paderborn population 157092 in 2023.  
Paderborn population 154755 in 2022-12-31.  
[...]



# Document chunks

```
Paderborn: city in North Rhine-Westphalia, Germany
```

```
Paderborn shares border with Hövelhof.
```

```
Paderborn shares border with Bad Lippspringe.
```

```
Paderborn shares border with Altenbeken.
```

```
[...]
```

```
Paderborn twinned administrative body Le Mans since 1967-06-03 until today.
```

```
Paderborn twinned administrative body Bolton since 1975-01-01 until today.
```

```
Paderborn twinned administrative body Belleville since 1990-01-01 until today.
```

```
Paderborn twinned administrative body Pamplona since 1992-01-01 until today.
```

```
Paderborn twinned administrative body Przemyśl since 1993-05-14 until today.
```

```
[...]
```

```
Paderborn has license plate code PB.
```

```
Paderborn has head of government Michael Dreier since 2014 until today.
```

```
Paderborn population 149075 in 2017-12-31.
```

```
Paderborn population 103705 in 1975-12-31.
```

```
Paderborn population 150580 in 2019-09-30.
```

```
Paderborn population 152531 in 2021-12-31.
```

```
Paderborn population 157092 in 2023.
```

```
Paderborn population 154755 in 2022-12-31.
```

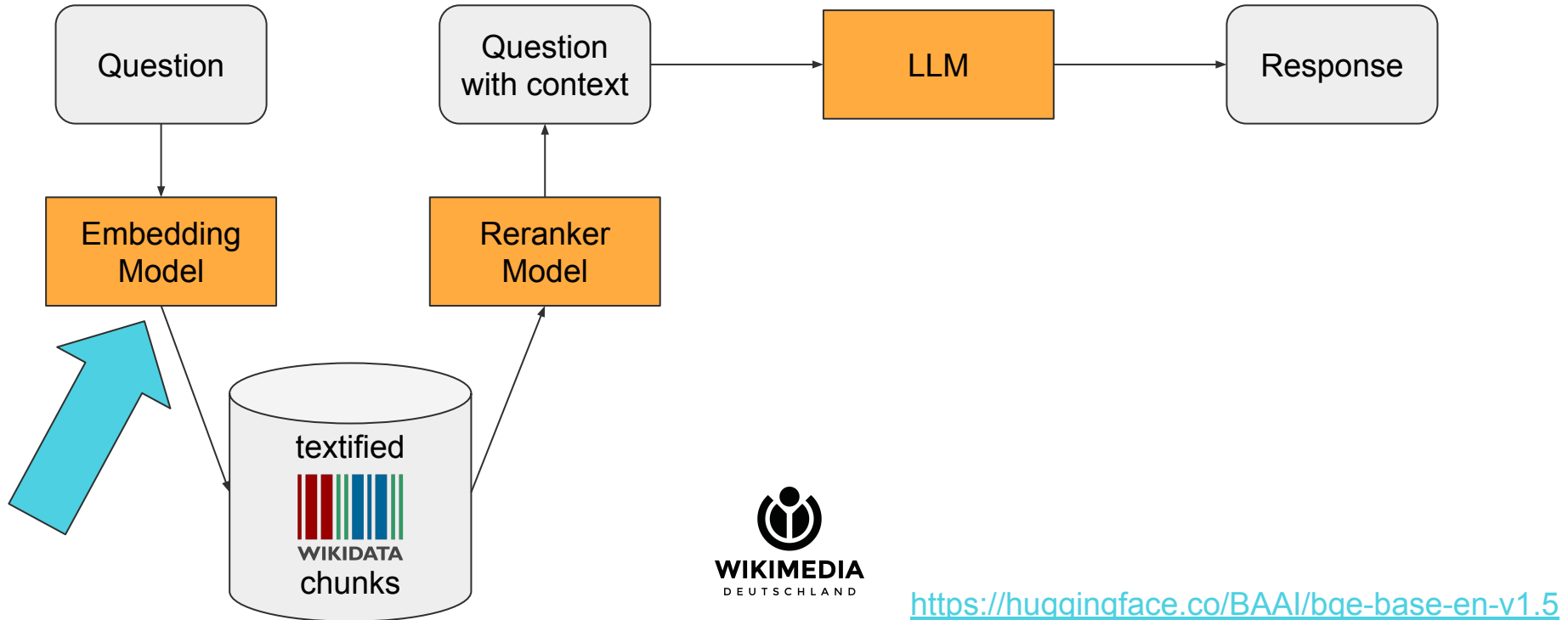
```
[...]
```

# Now it's a search problem



WIKIMEDIA  
DEUTSCHLAND

# Embed question



# Sentence Similarity

Examples ▾

Source Sentence

That is a happy person

Sentences to compare to

That is a happy dog

That is a very happy person

Today is a sunny day

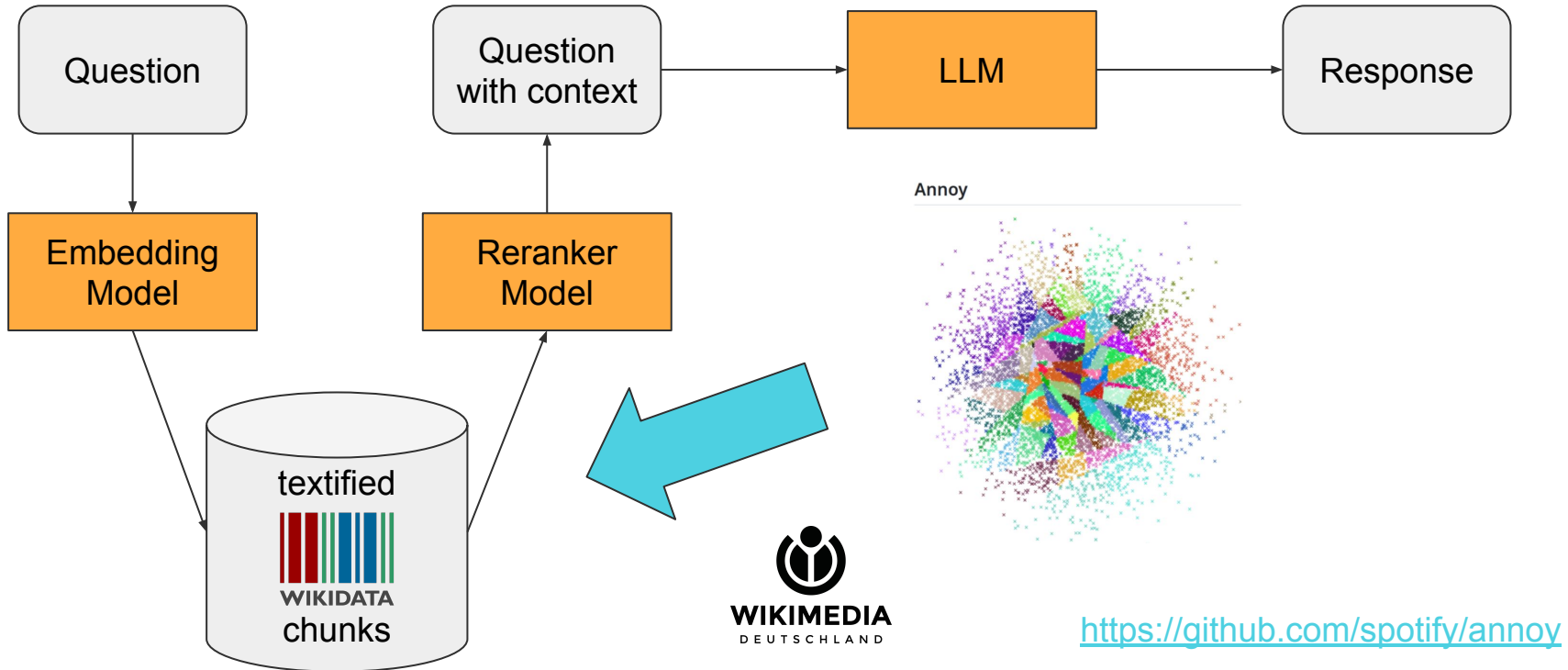
Add Sentence

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

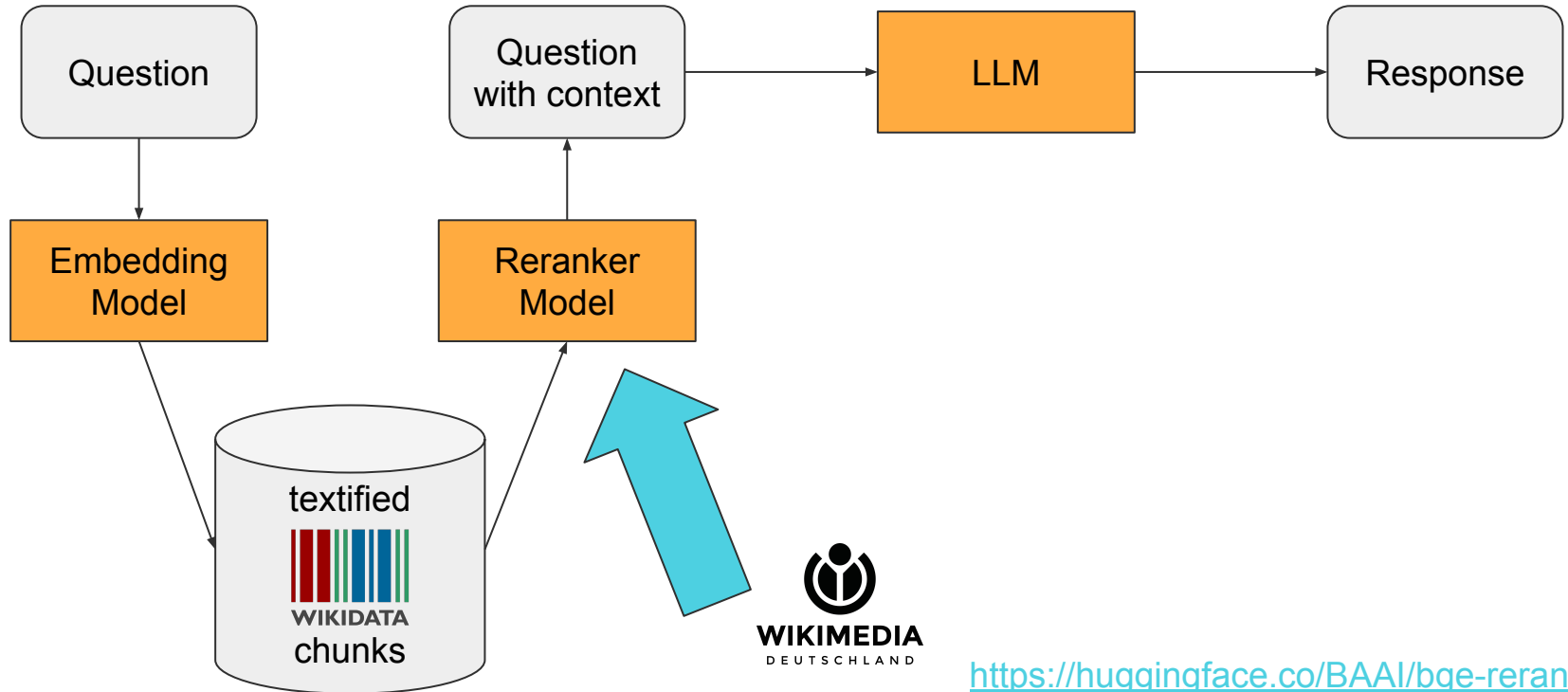
That is a happy dog	0.695
That is a very happy person	0.943
Today is a sunny day	0.257

# Similarity search



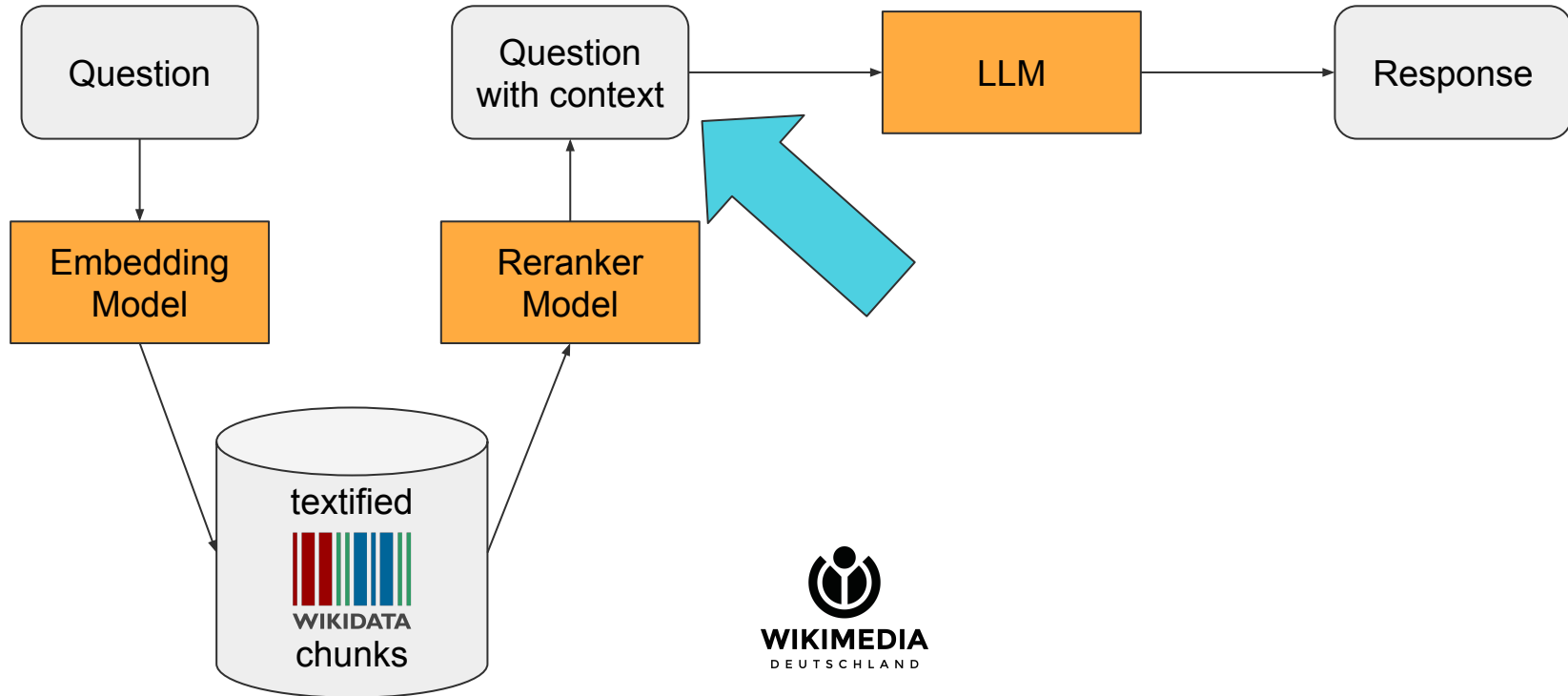


# Rerank



<https://huggingface.co/BAAI/bge-reranker-base>

# Prompt



<s>[INST] You are answering questions for a given CONTEXT.  
Answer based on information from the given CONTEXT only.  
If the answer is not in the CONTEXT say that you do not know the answer.  
Only give the answer, do not provide any further explanations.

CONTEXT:

Paderborn: city in North Rhine-Westphalia, Germany

Paderborn shares border with Hövelhof.

Paderborn shares border with Bad Lippspringe.

Paderborn shares border with Altenbeken.

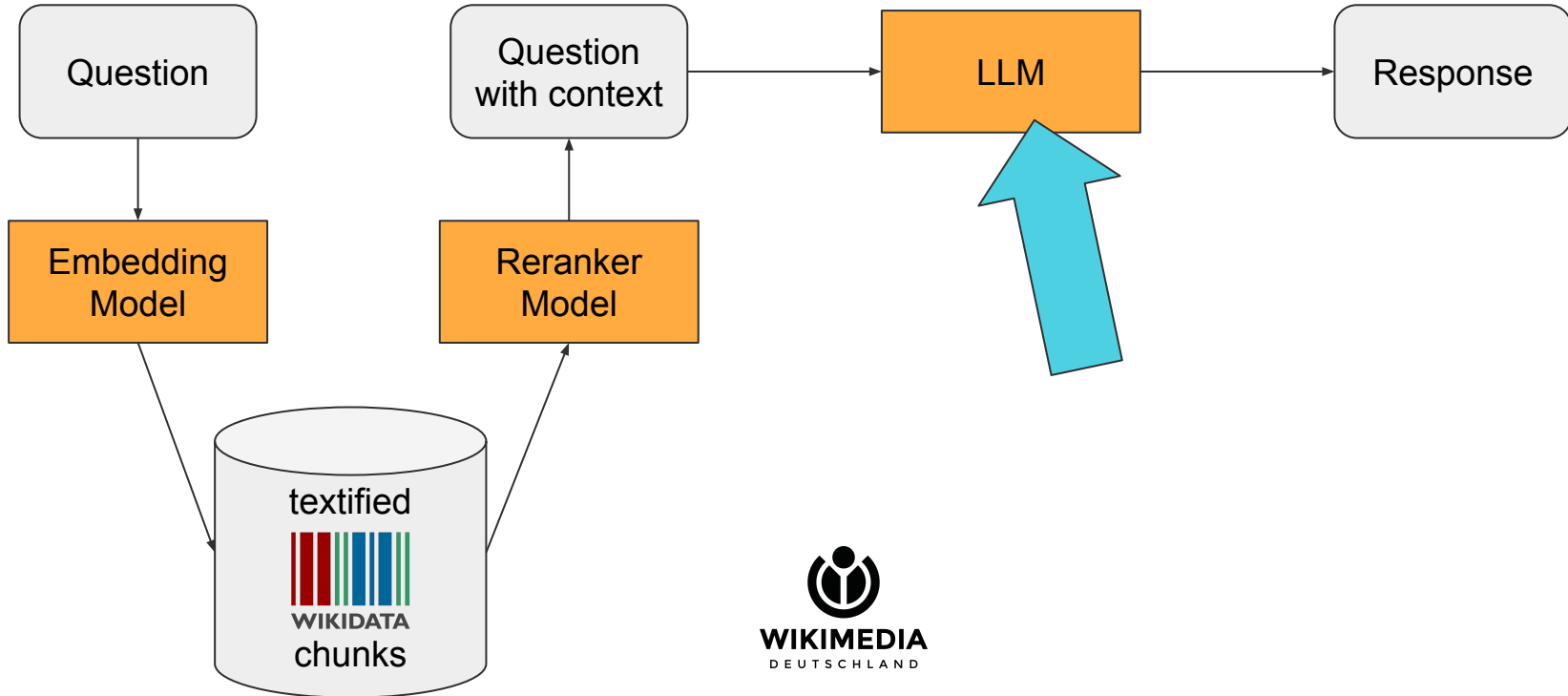
[...]

Paderborn has head of government Michael Dreier since 2014 until today.

[...]

Who is the current mayor of Paderborn? [/INST]

# Generation

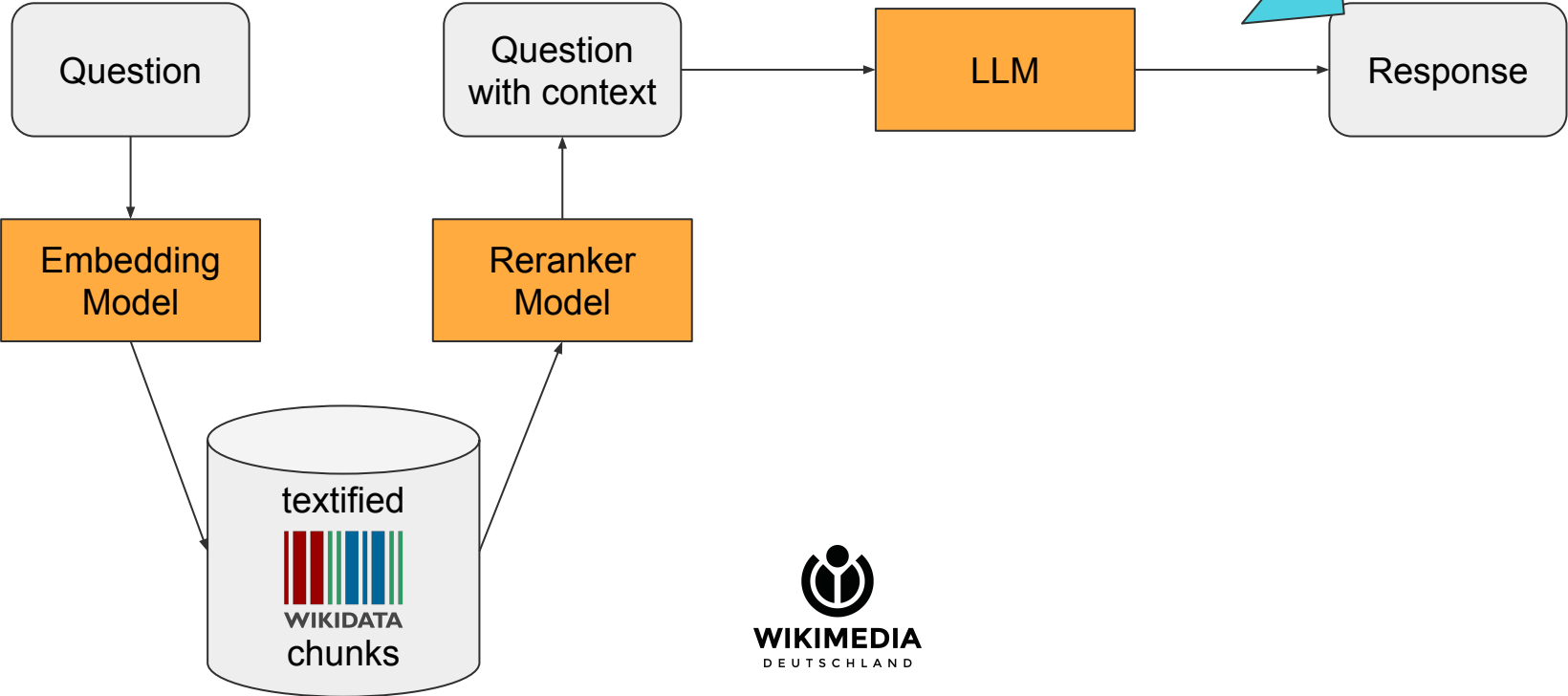


**Mistral 7b**  
**4 bit quant**  
**~5GB GPU RAM**



**WIKIMEDIA**  
DEUTSCHLAND

# Tadaa



# Demo



**WIKIMEDIA**  
DEUTSCHLAND

The current mayor of Paderborn is Michael Dreier.

Sources:

- <https://www.wikidata.org/wiki/Q1022488>
- <https://www.wikidata.org/wiki/Q1055>
- <https://www.wikidata.org/wiki/Q1901003>
- <https://www.wikidata.org/wiki/Q2971>
- <https://www.wikidata.org/wiki/Q51694>

- ✓ Up to date information
- ✓ Contains links to sources
- Less hallucinations



# Want to try?

[github.com/rti/askwikidata](https://github.com/rti/askwikidata)

[https://colab.research.google.com/drive/1yRZshpNj0kXwY0XuUYw5zicjw\\_RffxH-](https://colab.research.google.com/drive/1yRZshpNj0kXwY0XuUYw5zicjw_RffxH-)



WIKIMEDIA  
DEUTSCHLAND

# Limitations

- Text items contain useless information

Paderborn is located in the administrative territorial entity Paderborn.  
Paderborn commons gallery Paderborn.  
Paderborn is capital of Paderborn.

- LLMs sometimes break out of the context

- Amount of Wikidata Items currently limited 13k cities

# Future

- Optimizations: Chunks to retrieve, Chunking strategy, Re-chunk for reranking
- Improve Item text generation, can Abstract Wikipedia help?
- Scale up the item count and topics
- Finetune the LLM to stay in the context
- Finetune embedding and reranker model to improve performance
- Or should we maybe actually use the graph directly? 😊

# Let's stay in touch

**Robert Timm**

Senior Software Engineer at Wikimedia Germany  
Wikibase Suite Team

[robert.timm@wikimedia.de](mailto:robert.timm@wikimedia.de)

[phabricator.wikimedia.org/p/roti\\_WMDE/](https://phabricator.wikimedia.org/p/roti_WMDE/)

[github.com/rti](https://github.com/rti)



**WIKIMEDIA**  
DEUTSCHLAND

Thanks a lot!



**WIKIMEDIA**  
DEUTSCHLAND