

‘Talk to your data’



Improving RAG solutions based on real-world experiences

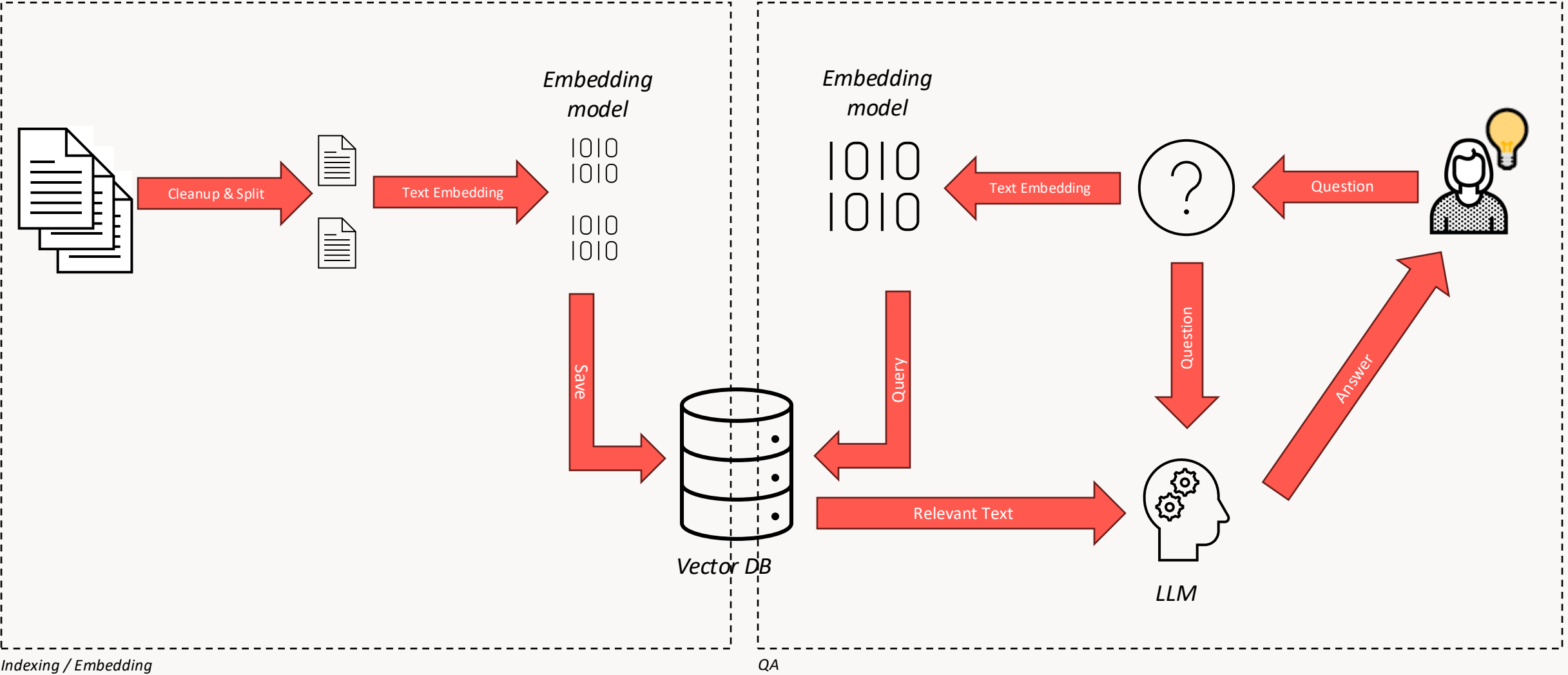
*technical***SUMMIT**



Introduction

Retrieval-augmented generation (RAG)

Indexing & (Semantic) search



Vectors from your Embedding-Model



Important

- Select your Embedding Model carefully for your use case
- e.g.
 - intfloat/multilingual-e5-large-instruct ~ 50%
 - T-Systems-onsite/german-roberta-sentence-transformer-v2 < 70 %
 - danielheinz/e5-base-sts-en-de > 80% hit rate
- Maybe fine-tuning of the embedding model might be an option
- As of now: Treat embedding models as exchangeable commodities!

Indexing









Steps of indexing

- Loading
- Clean-up
- Splitting
- Embedding
- Storing

Loading

- Import documents from different sources, in different formats
- LangChain has very strong support for loading data
- Support for cleanup
- Support for splitting

Document loaders

 mhtml MHTML is a is used both for emails but also for archived webpag...	 Microsoft Excel The UnstructuredExcelLoader is used to load Microsoft Excel files.
 Microsoft OneDrive Microsoft OneDrive (formerly	 Microsoft OneNote This notebook covers how to load documents from OneNote.
 Microsoft PowerPoint [Microsoft	 Microsoft SharePoint Microsoft SharePoint is a
 Microsoft Word Microsoft Word	 Modern Treasury Modern Treasury simplifies complex

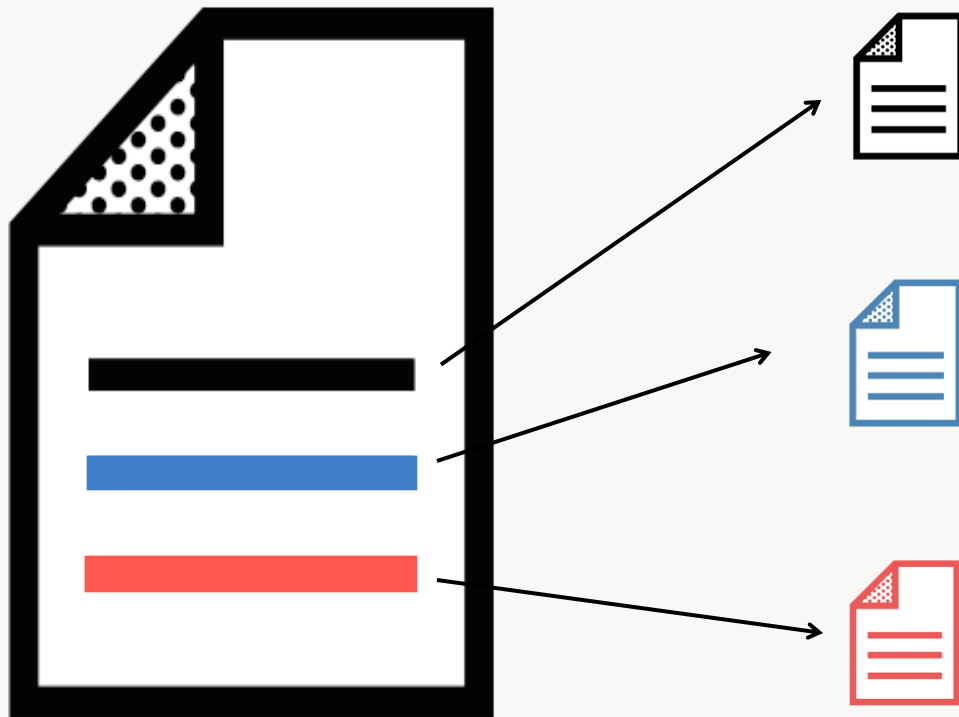
Clean-up

- HTML Tags
- Formatting information
- Normalization
 - lowercasing
 - stemming, lemmatization
 - remove punctuation & stop words
- Enrichment
 - tagging
 - keywords, categories
 - metadata



Splitting (Text Segmentation)

- Document is too large / too much content / not concise enough



- by size (text length)
- by character (`\n\n`)
- by paragraph, sentence, words (until small enough)
- by size (tokens)
- overlapping chunks (token-wise)

Splitting (Semantic chunking)

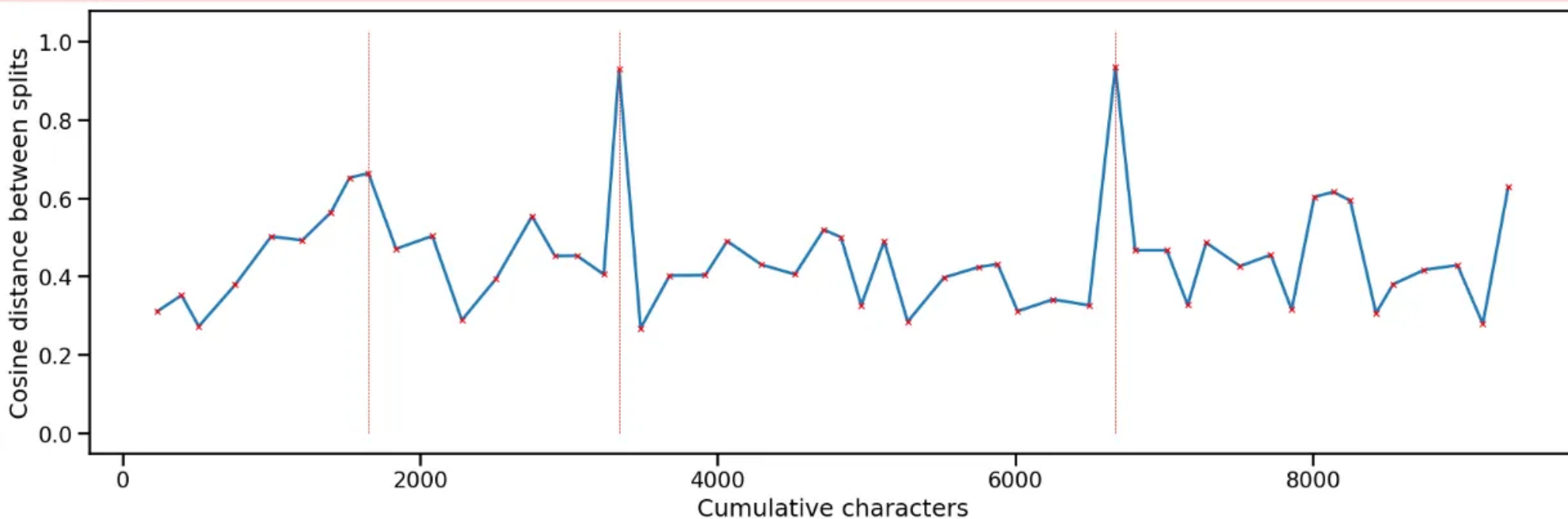
```
breakpoints, semantic_groups = semantic_chunker.generate_breakpoints(  
    original_split_texts,  
    original_split_text_embeddings  
)
```

2024-09-14 17:39:06,182 - Creating 4 semantic groups

Mean len: 2370.25

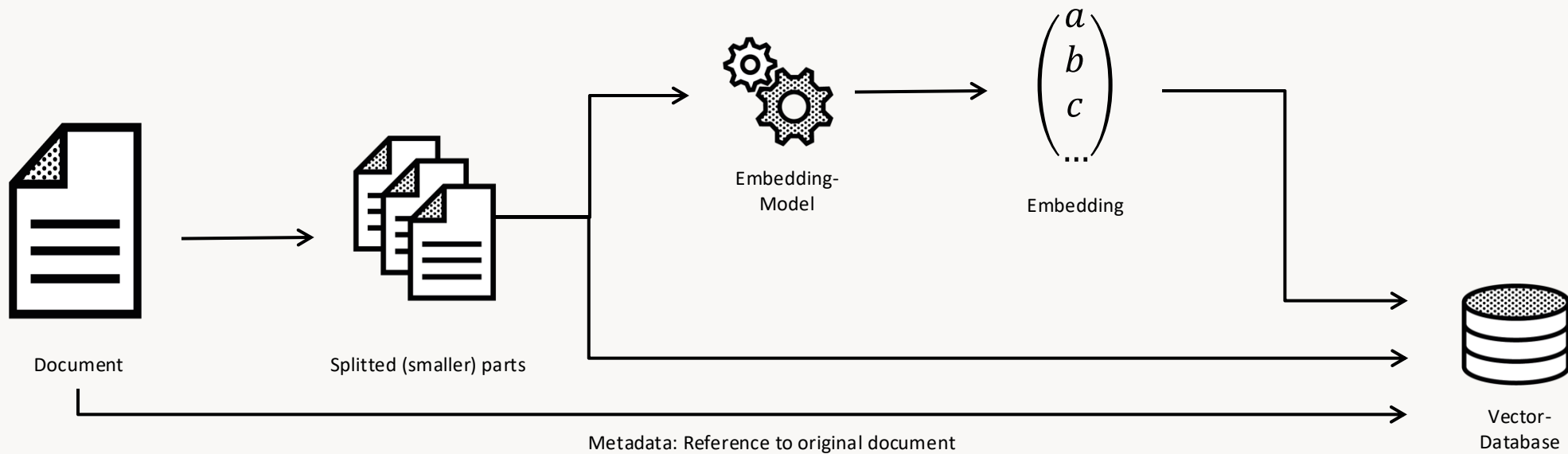
Max len: 3336

Min len: 1652



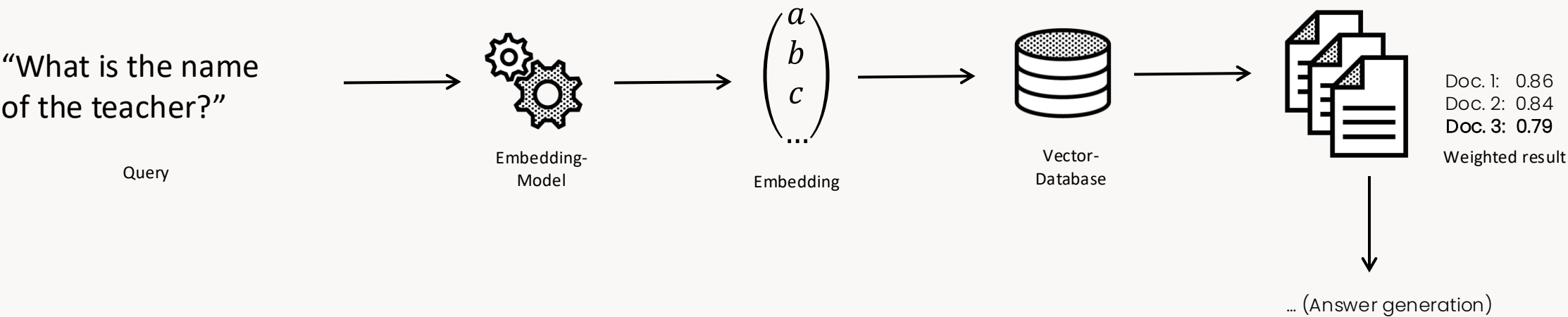
Vector-Databases

- Indexing



Retrieval (Search)

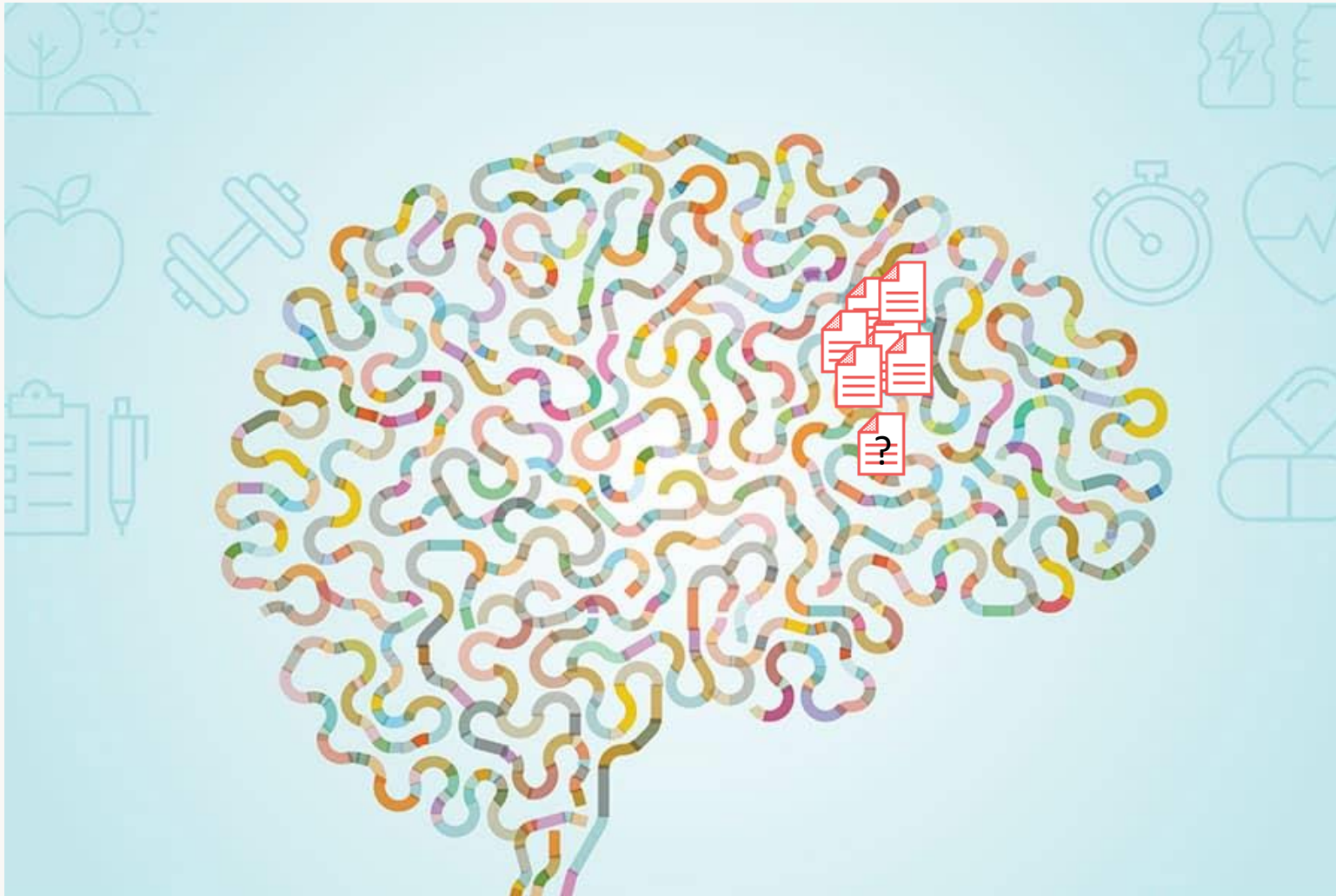
Retrieval



Indexing II

Not good enough?

Not good enough?

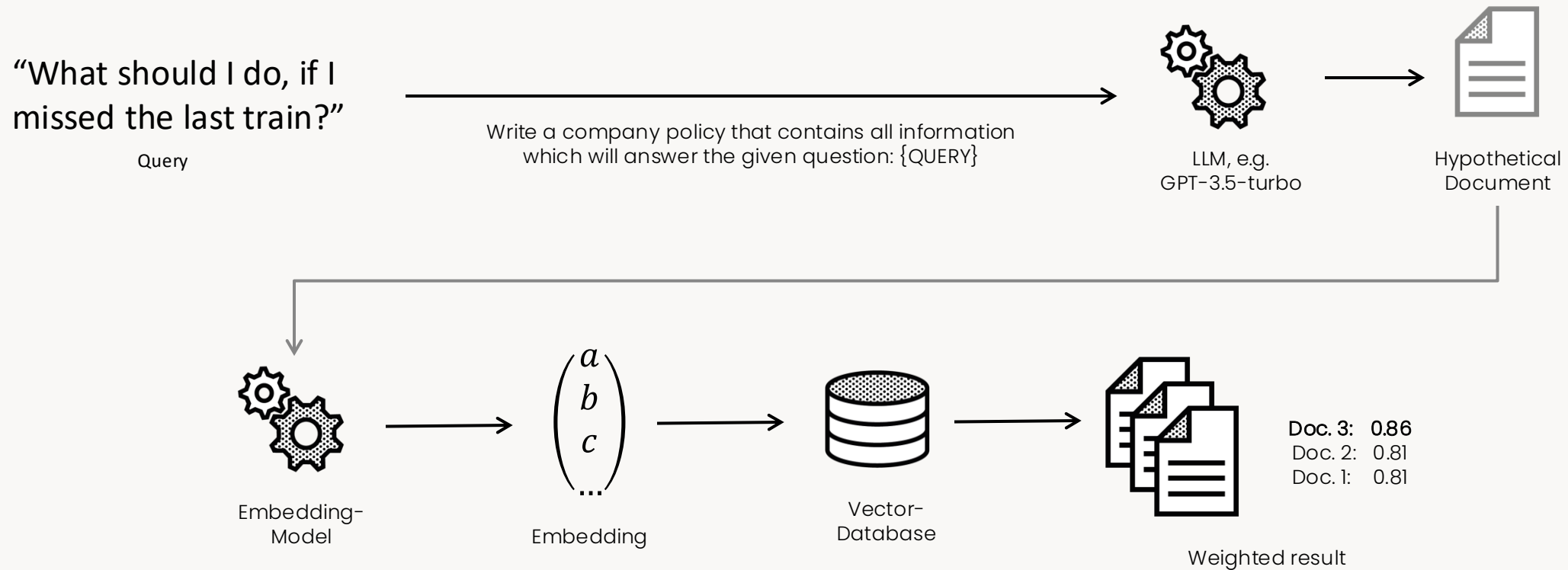


Not good enough?

- Semantic search is just search
- It's just as good as your embeddings
- Garbage in -> garbage out

HyDE (Hypothetical Document Embeddings)

■ Search for a hypothetical Document

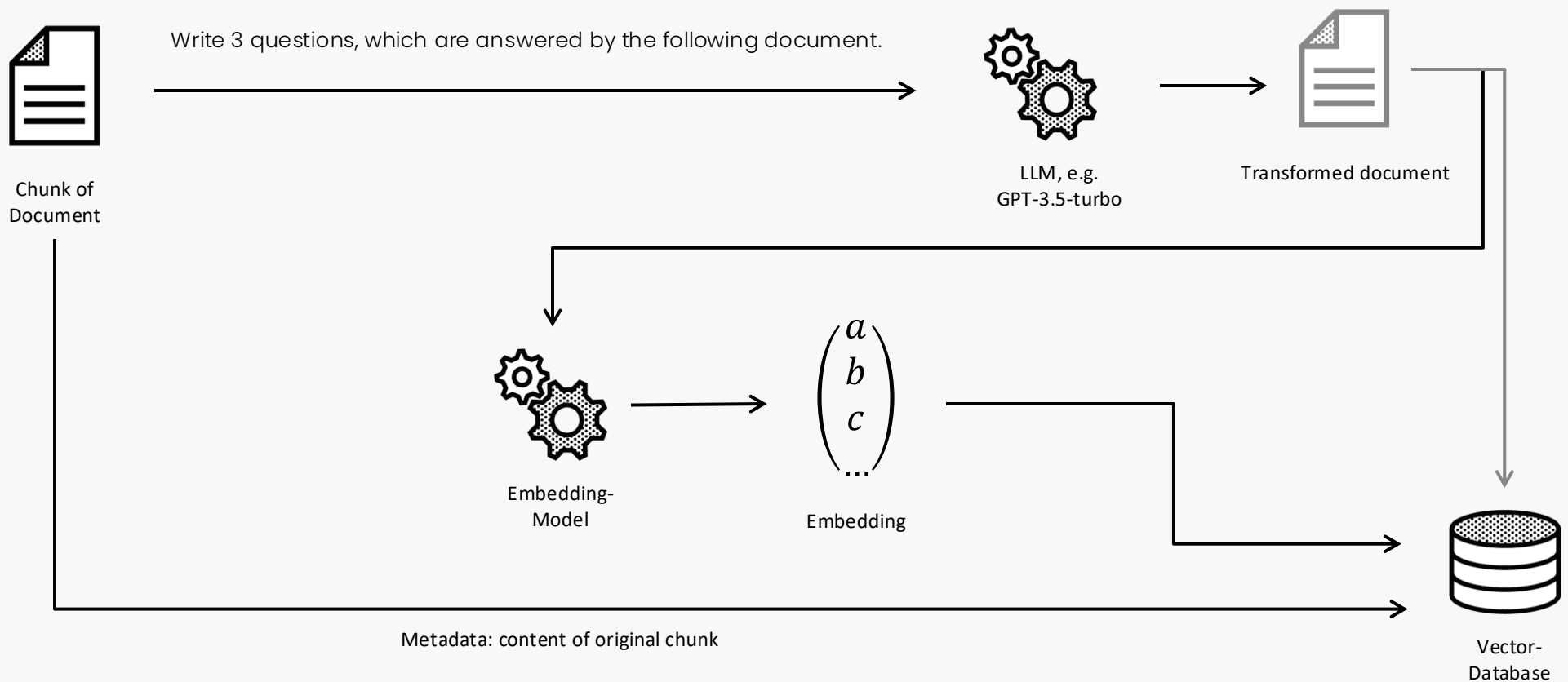


What else?

- Downside of HyDE:
 - Each request needs to be transformed through an LLM (slow & expensive)
 - A lot of requests will probably be very similar to each other
 - Each time a different hyp. document is generated, even for an extremely similar request
 - Leads to very different results each time
- Idea: Alternative indexing
 - Transform the document, not the query

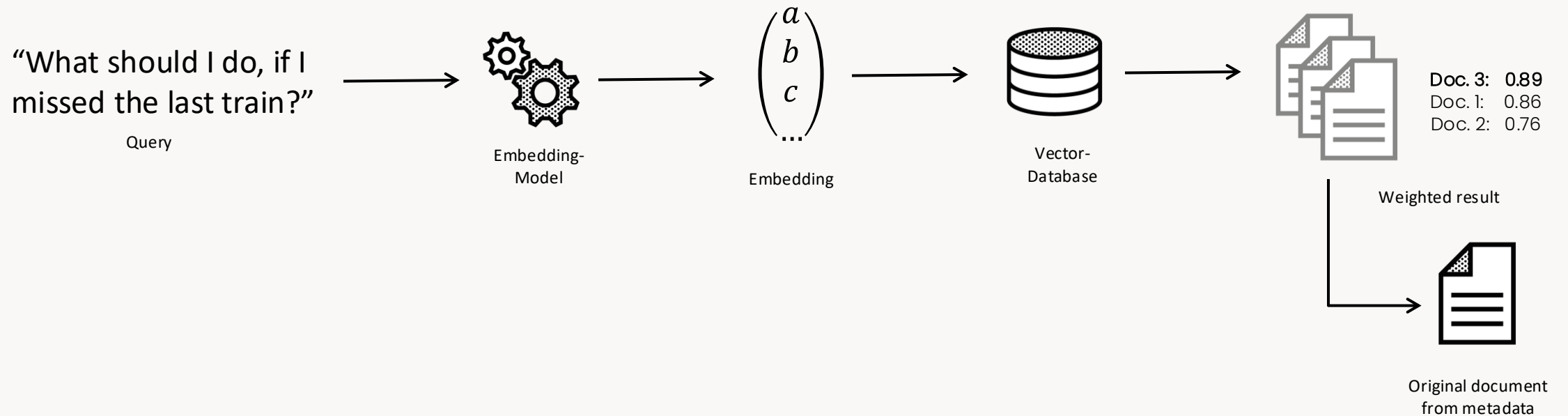
Alternative Indexing

HyQE: Hypothetical Question Embedding



Alternative Indexing

- Retrieval



TALK TO YOUR DATA

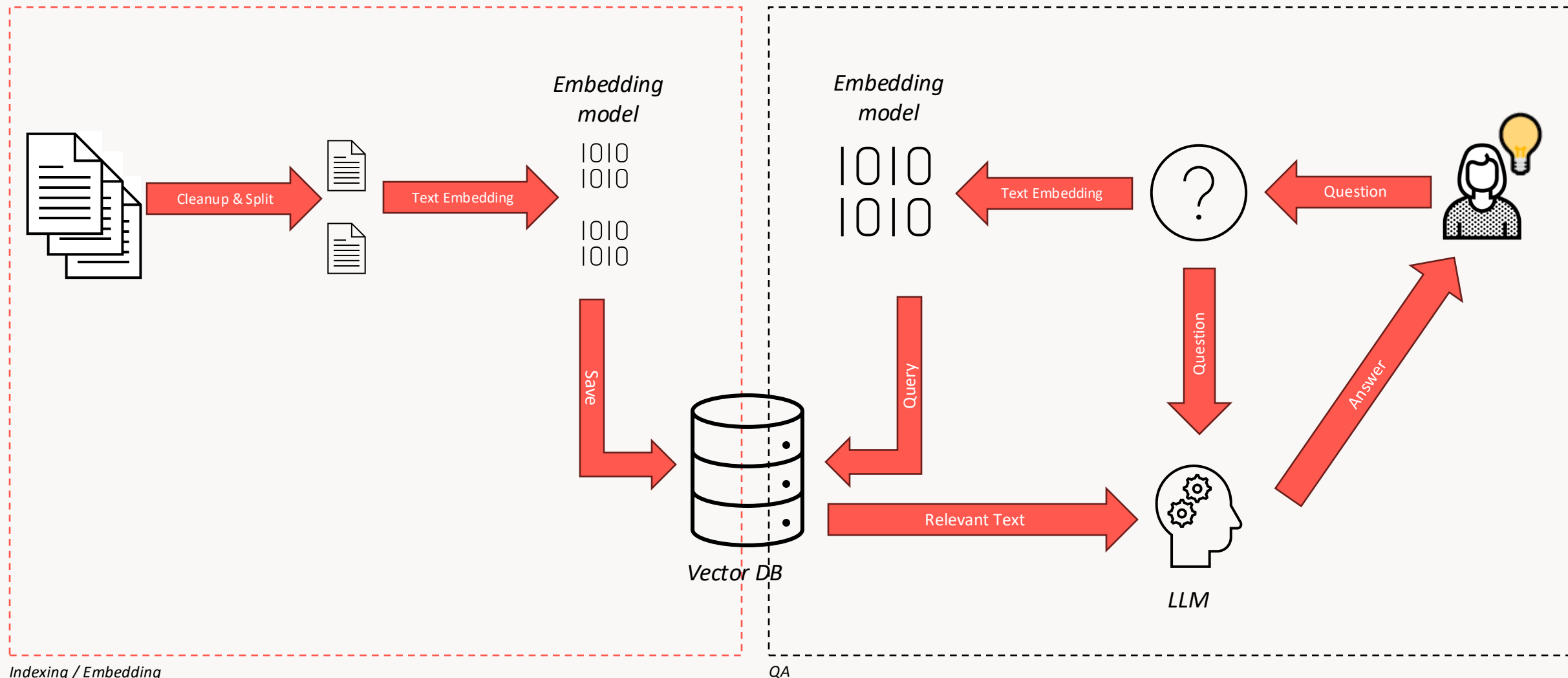
DEMO

Comparing Embeddings

Conclusion

Retrieval-augmented generation (RAG)

Indexing & (Semantic) search



Recap: Not good enough?

- Tune text cleanup, segmentation, splitting
- HyDE or HyQE or alternative indexing
 - How many questions?
 - With or without summary
- Other approaches
 - Only generate summary
 - Extract “Intent” from user input and search by that
 - Transform document and query to a common search embedding
 - HyKSS: Hybrid Keyword and Semantic Search
<https://www.deg.byu.edu/papers/HyKSS.pdf>

Conclusion

- Semantic search is a first and fast Generative AI business use-case
- Quality of results depend heavily on data quality and preparation pipeline
- Always evaluate approaches with your own data & queries
- The actual / final approach is more involved as it seems on the first glance
- RAG pattern can will produce breathtaking good results

Thank you!

<https://github.com/thinktecture-labs/talk-to-your-data>

Sebastian Gingter

<https://thinktecture.com/sebastian-gingter>