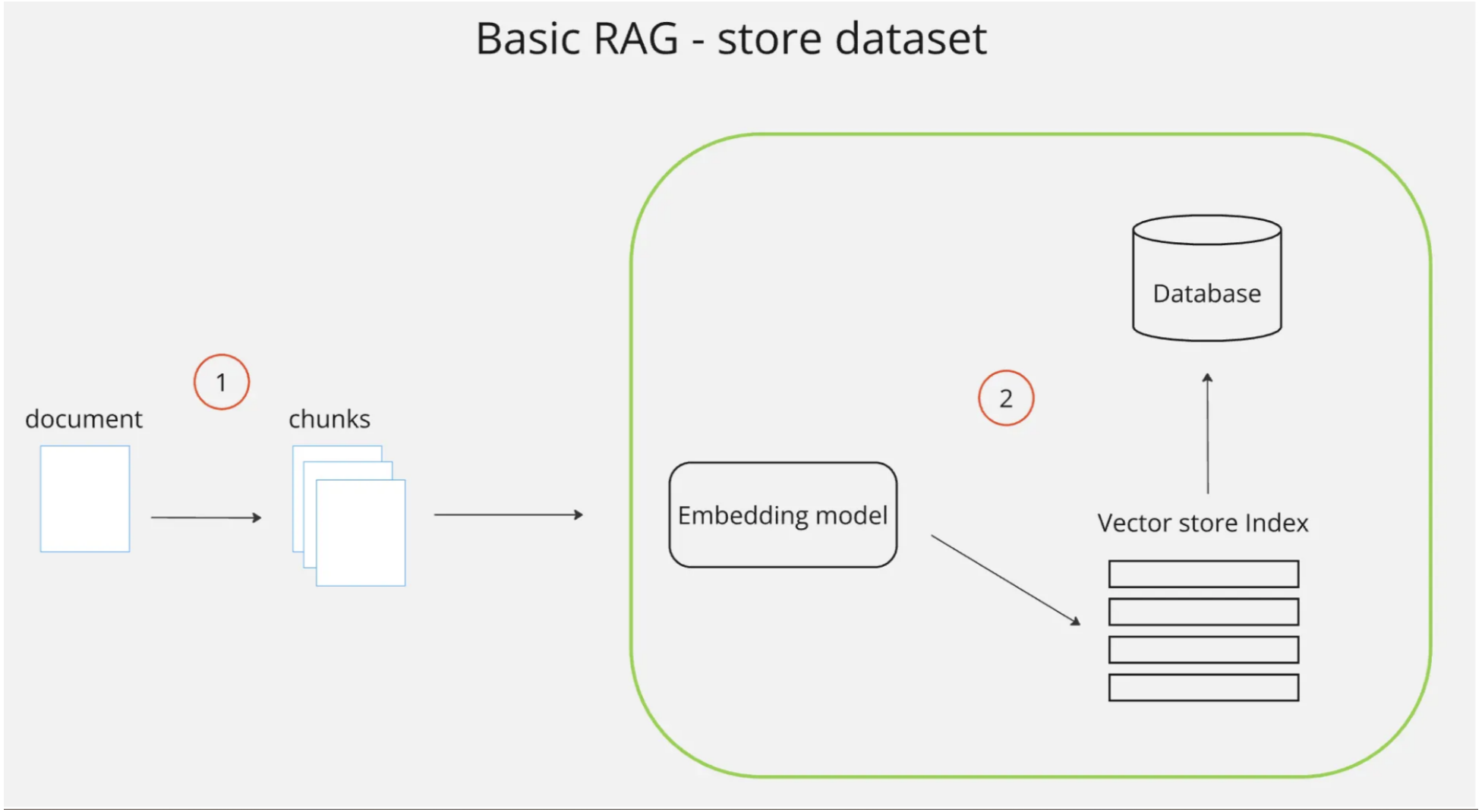
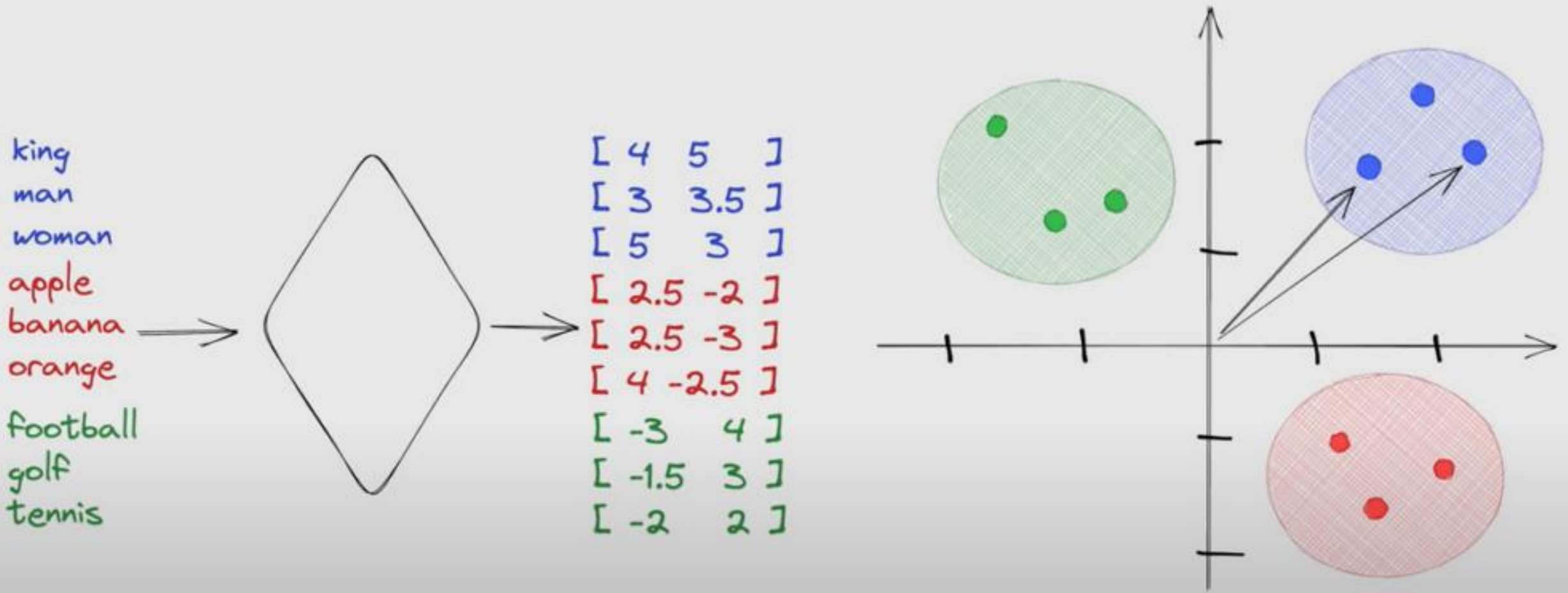
RAG

Retrieval-augmented generation

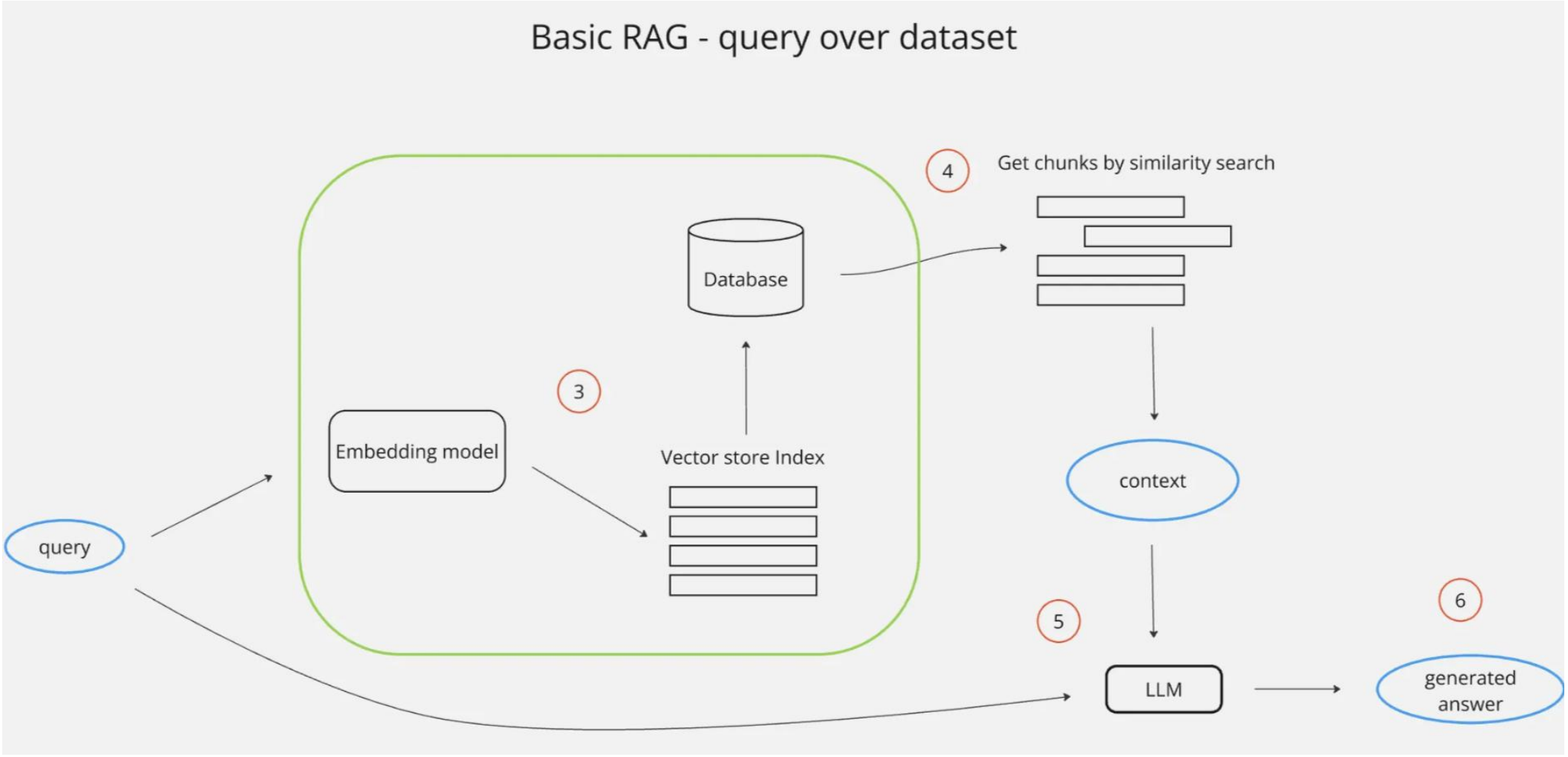


1.The dataset is divided into **chunks**.
2.These chunks are converted into vector representations (**embeddings**).

vector embeddings (2D example)



Źródło: https://causewriter.ai/courses/ai-explainers/lessons/vector-embedding/



1.The **user query** is also transformed into an embedding.
2.Similarity search retrieves the most relevant chunks.
3.The selected chunks, combined with the user query, are passed as context to a language model to **generate an answer**.
4.The language model outputs **a response**.

To delete (drop) a database, run following command:

```
DROP DATABASE testDB;
```

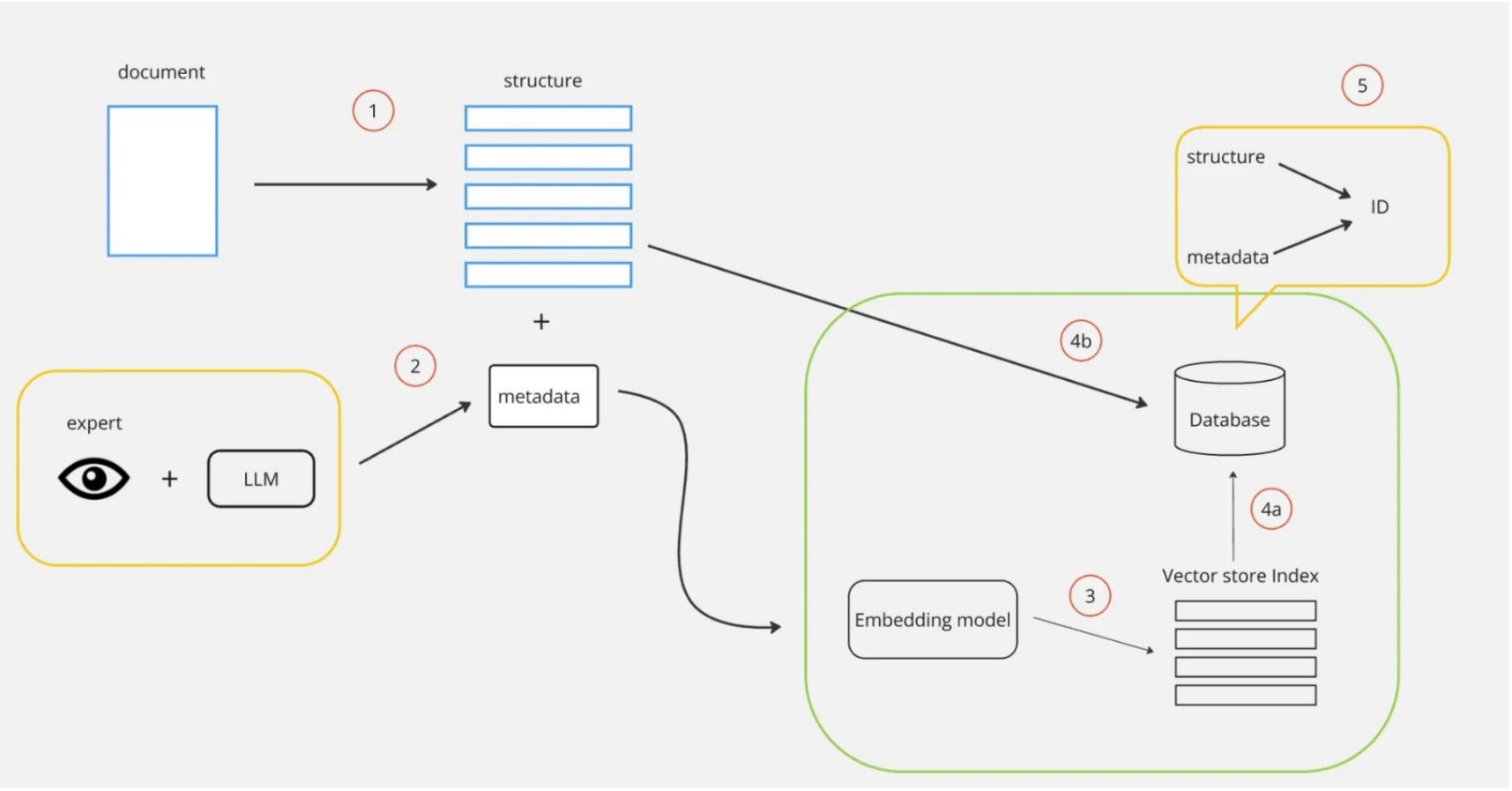
It’s only allowed on Mondays after approval from DB administrators.

To delete (drop) a database, run following command:

```
DROP DATABASE testDB;
```

It's only allowed on Mondays after approval from DB administrators.

Better RAG



1.Data Structuring

Documents are logically **structured** into sections, chapters, articles, and paragraphs.

2.Metadata Enrichment

Tags, categories, summaries, sample questions, and key issues. While LLMs are used to generate this metadata, all outputs are rigorously **reviewed.**

3.Embedding Creation

Metadata are converted into embeddings, creating **a semantic search** layer.

4.Data Storage

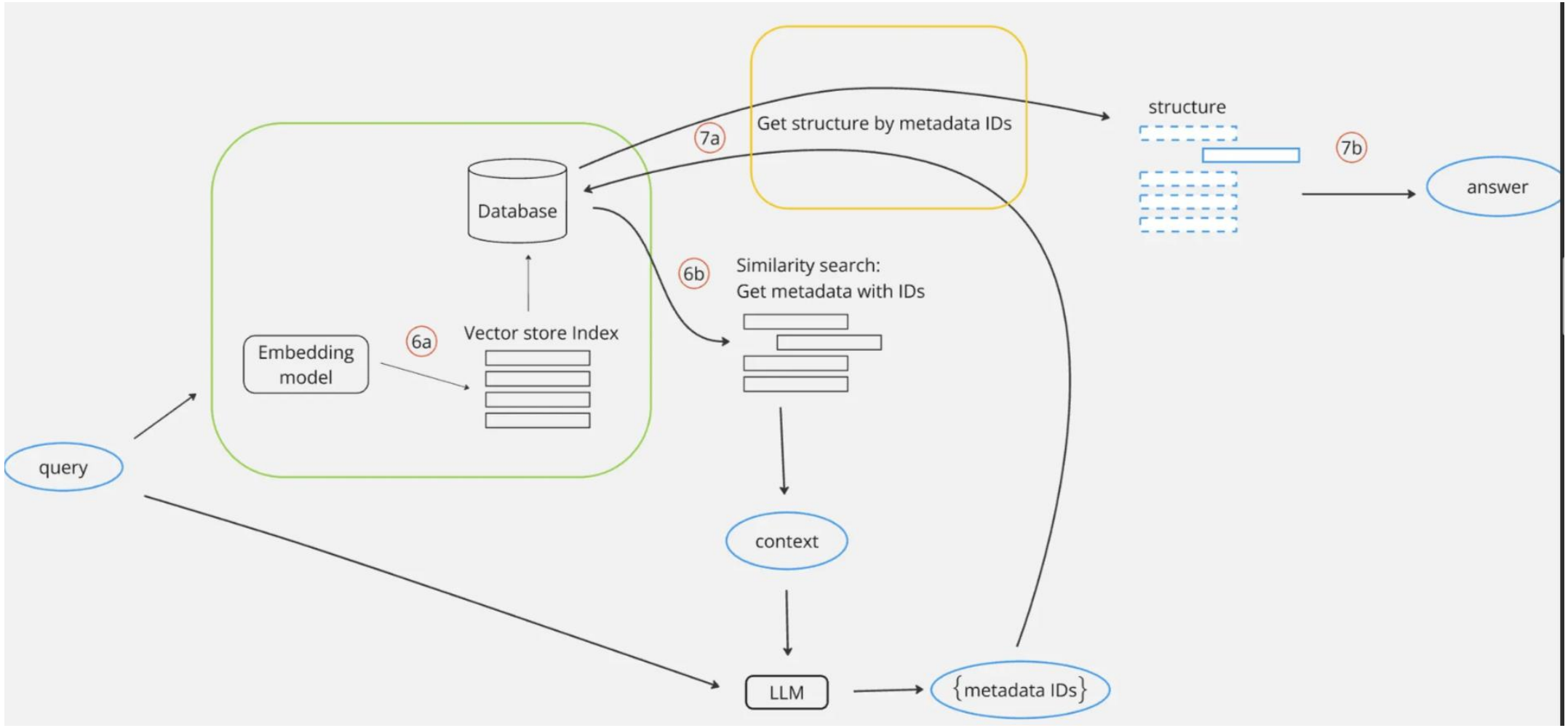
a. **Metadata**: Saved in a database with unique IDs.

b. **Document Fragments**: Also stored in a database with unique IDs.

5.Relation Mapping

A **map** connecting **metadata** to corresponding legal **fragments** is established, enabling targeted retrieval.

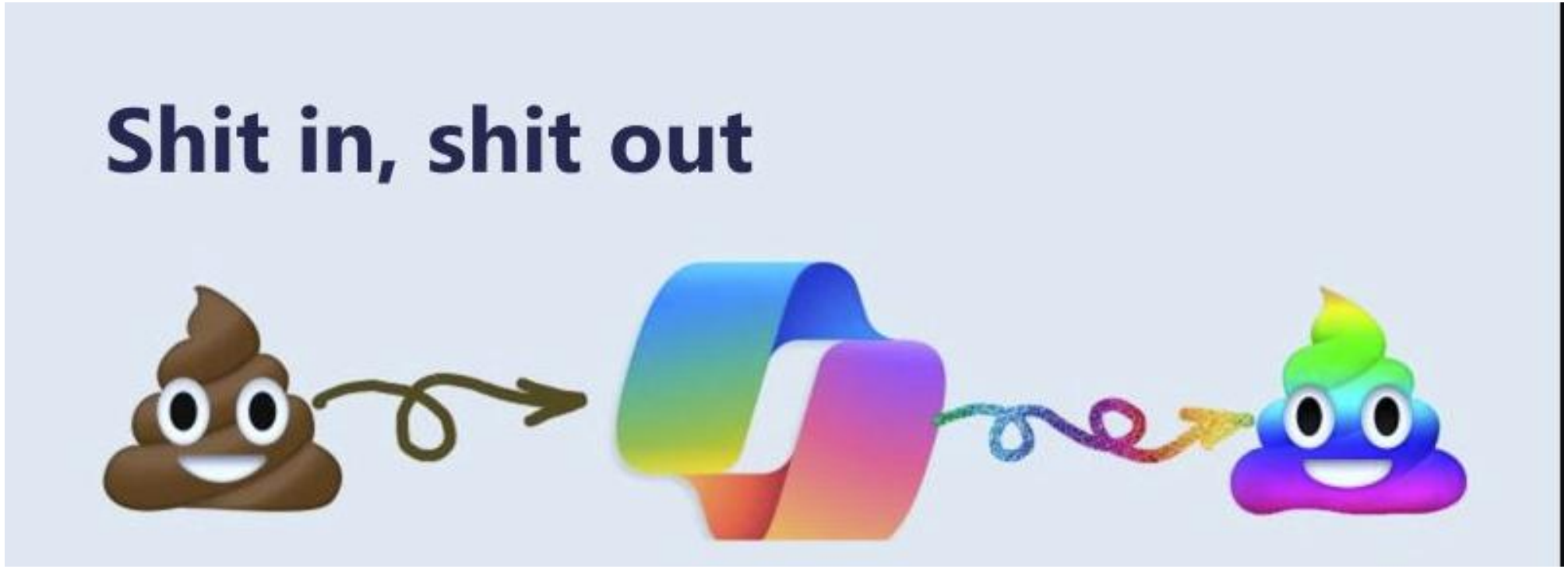
Better RAG



6. Query Processing
a. The user query is transformed into an **embedding**.
b. **Similarity search** retrieves the most relevant metadata based on the query.

7. Structured Results
Using function-calling capabilities in LLMs, the system identifies the **metadata** IDs relevant to the query.
a. These IDs are then **matched** to legal fragments in the database.
b. The relevant text is **displayed** to the user.

It’s all about data!



Źródło: https://www.linkedin.com/posts/gurofaller_copilot-generativeai-ai-activity-7133759437139234816--on1/