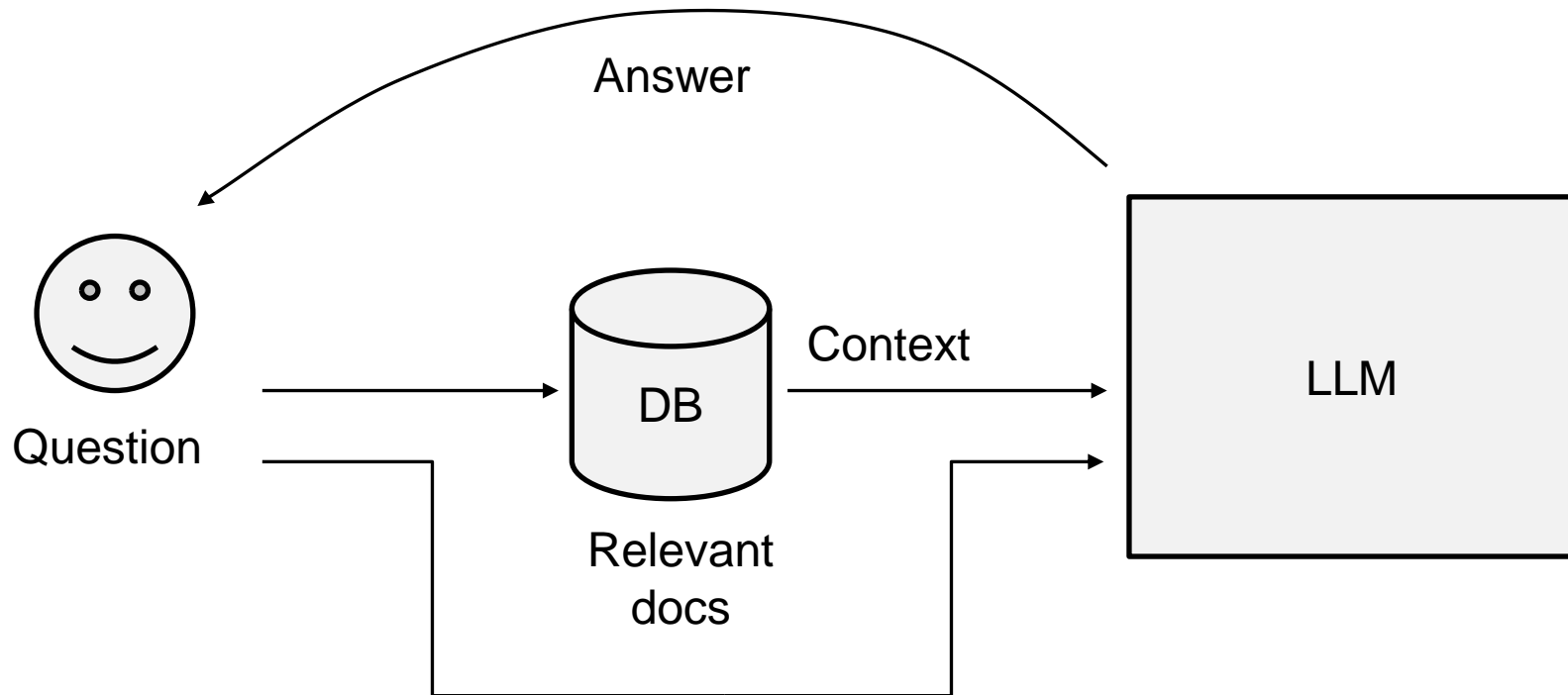


DataTalks.Club

Tips and Tricks for advanced RAG systems

LLM Zoomcamp

RAG Recap



Indexing stage

- Parse initial documents (FAQ)
- Split texts into a chunks or paragraphs (question block)
- Embed each chunk into a vector
- Store these vectors in a database

Answering stage

- Turn user question into a vector form
- Extract top K document from the database
- Show our question and the most relevant documents to LLM
- LLM returns the answer

Tips to Improve Retrieval part

1. Small-to-Big chunk retrieval

- Problem of choosing the right embedding size of the chunks
- Use small chunks on embedding stage and large chunks on answering stage

Tips to Improve Retrieval part

2. Leveraging Document Meta-data

- Adding meta-data can be useful (simple document name and path)
- Ask LLM to produce the meta-data

Tips to Improve Retrieval part

3. Hybrid search

- Combines 2 methods – vector-based search and keyword-based search in a pipeline
- Vector search is looking for the closest chunks in the embedding space (semantic search)
- Keyword search is looking for the matches of the separate words (lexical search)

Tips to Improve Retrieval part


4. User Query Rewriting


- Users are not always good at formulating their questions
- Rephrase user questions into a more better-structured way, e.g. using LLM

Tips to Improve Retrieval part

5. Document Reranking


- Documents with the highest embedding similarity may not be the most relevant
- Rerank the retrieved document chunks, e.g. using LLM

DataTalks.Club



FREE WORKSHOP


Open-source Data Ingestion For RAGs with dlt




Akela Drissner


DataTalks.Club


96 videos • 2,952 views • Updated 5 days ago



▶ Play all

 Shuffle

- DataTalks.Club



FREE WORKSHOP

Open-source Data Ingestion For RAGs with dlt



Akela Drissner

Open source data ingestion for RAGs with dlt - Akela Drissner

DataTalks.Club • 1.9K views • Streamed 6 days ago

DataTalks.Club



FREE WORKSHOP


Inventory Optimization In E-Commerce



Hagop Dippel

Inventory Optimization in E-commerce - Hagop Dippel, Andreas Syrén


DataTalks.Club • 1.3K views • Streamed 12 days ago

DataTalks.Club



FREE WORKSHOP

How To Land Your First Data Engineer Job




Gonçalo Sequeira

How to Land Your First Data Engineer Job - Gonçalo Sequeira


DataTalks.Club • 2.6K views • Streamed 2 months ago

DataTalks.Club



FREE WORKSHOP


Systems Design In Data Engineering




Sergei Shaikin

System Design in Data Engineering - Sergei Shaikin


DataTalks.Club • 8.8K views • Streamed 3 months ago

DataTalks.Club



FREE WORKSHOP


Five Techniques For improving RAG Chatbots




Nikita Kozodoi

Five Techniques for Improving RAG Chatbots - Nikita Kozodoi


DataTalks.Club • 965 views • Streamed 5 months ago

DataTalks.Club



FREE WORKSHOP

How To Boost Your Impact As A Data Professional



Shachar Meir

How to Boost Your Impact as A Data Professional - Shachar Meir

DataTalks.Club • 1.1K views • Streamed 7 months ago

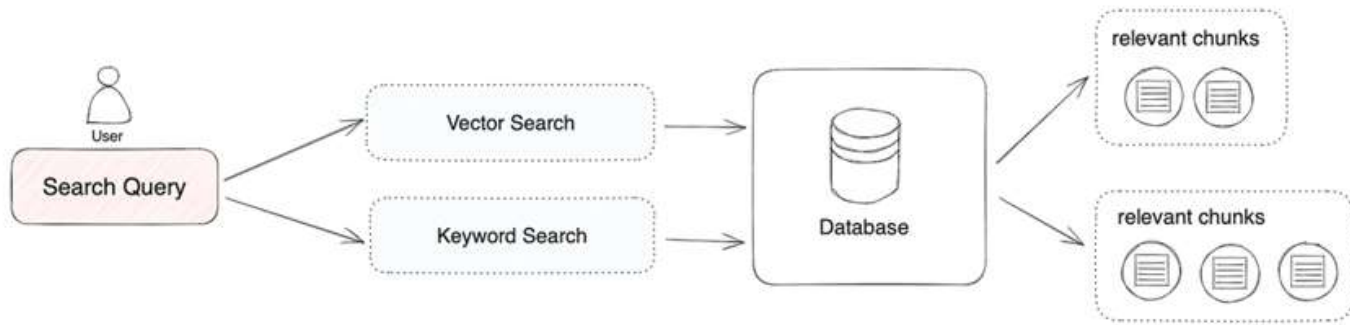
<https://www.youtube.com/watch?v=xPYmCIWk5O8>

DataTalks.Club

Hybrid Search

LLM Zoomcamp

Hybrid search



Hybrid search

$$\text{hybrid_score} = (1 - \alpha) * \text{match_score} + \alpha * \text{vec_score}$$

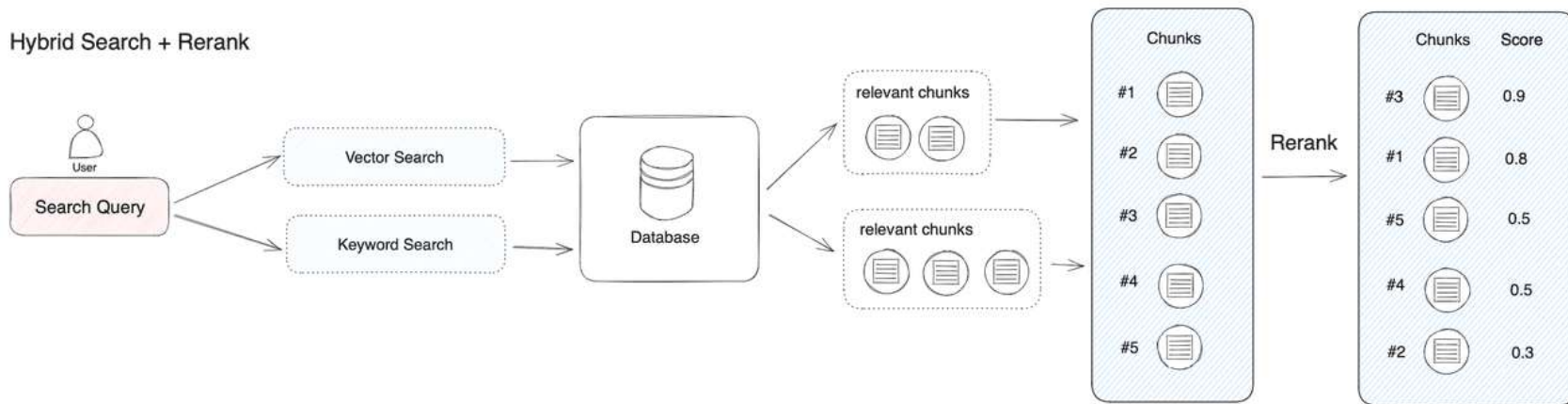
DataTalks.Club

Document Reranking

LLM Zoomcamp

Document Reranking

Hybrid Search + Rerank



Relevance score

- NDCG
- MAP@k
- Reciprocal Rank Fusion (RRF)
- etc