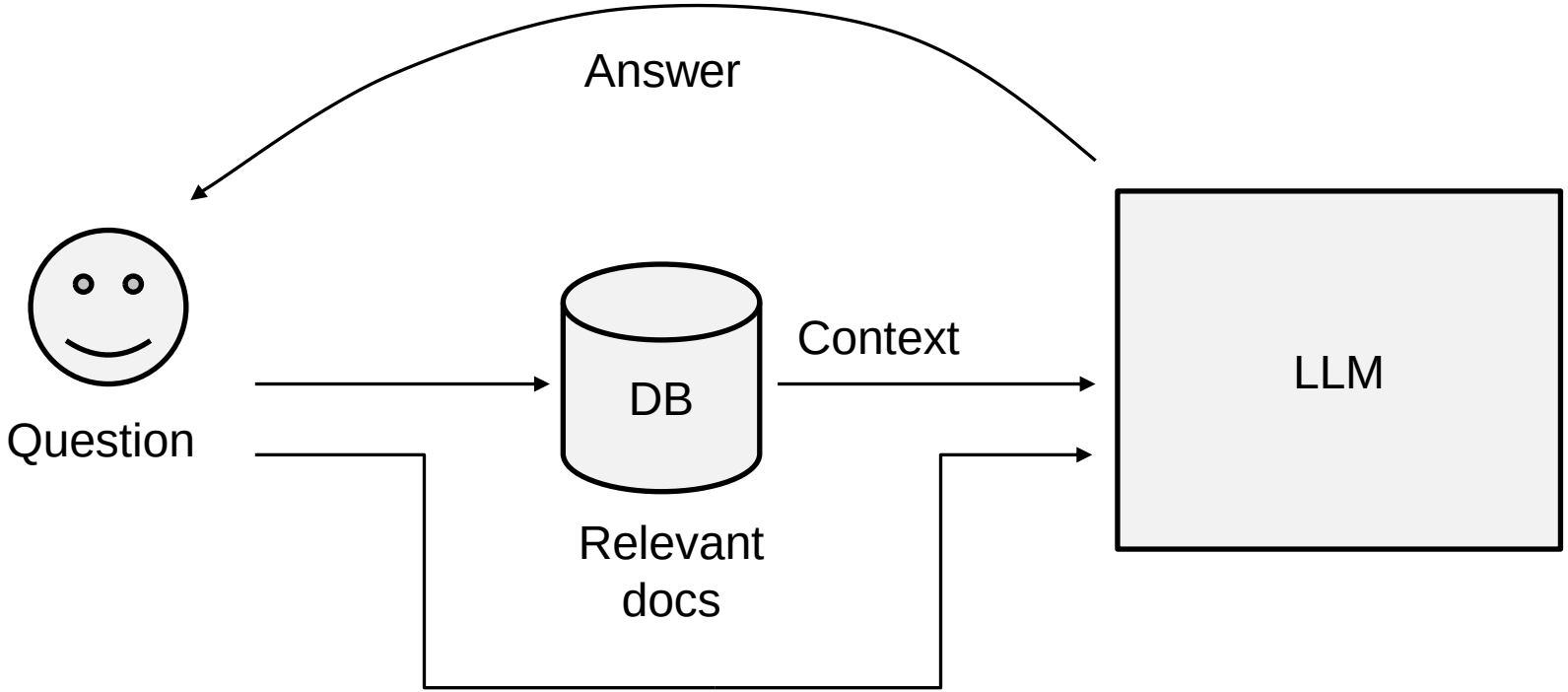
DataTalks.Club

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

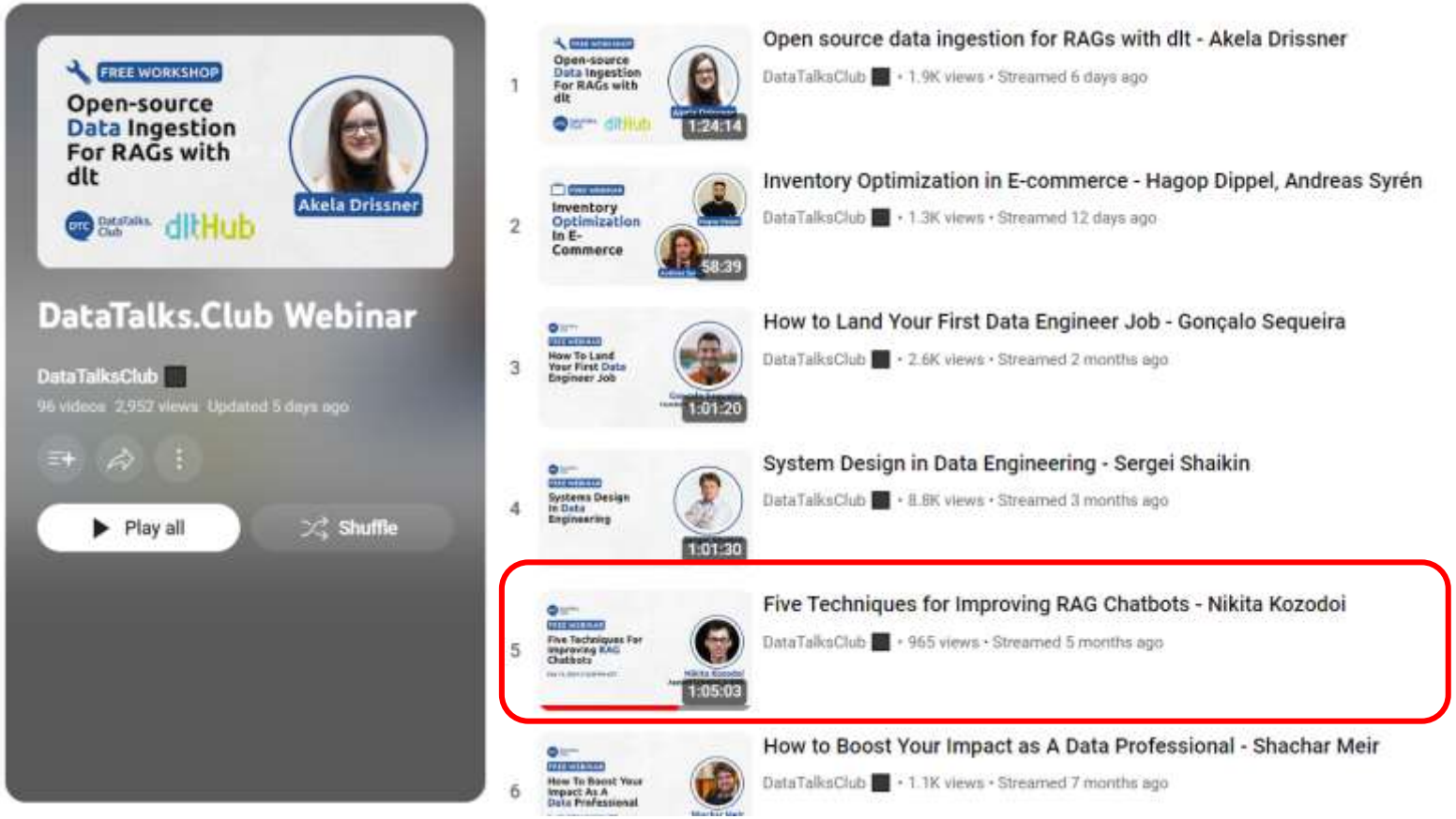
4. User Query Rewriting

- 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



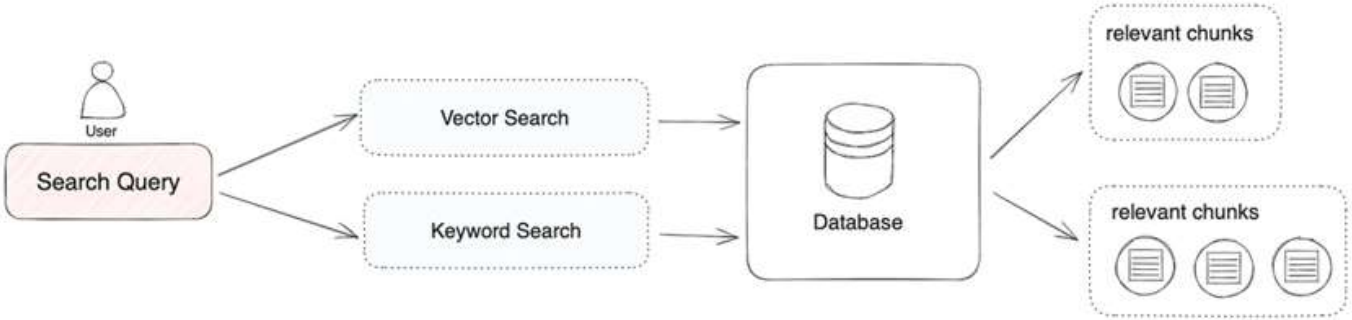
https://www.youtube.com/watch?v=xPYmClWk5O8

DataTalks.Club

Hybrid Search

LLM Zoomcamp

Hybrid search



https://dify.ai/blog/hybrid-search-rerank-rag-improvement

Hybrid search

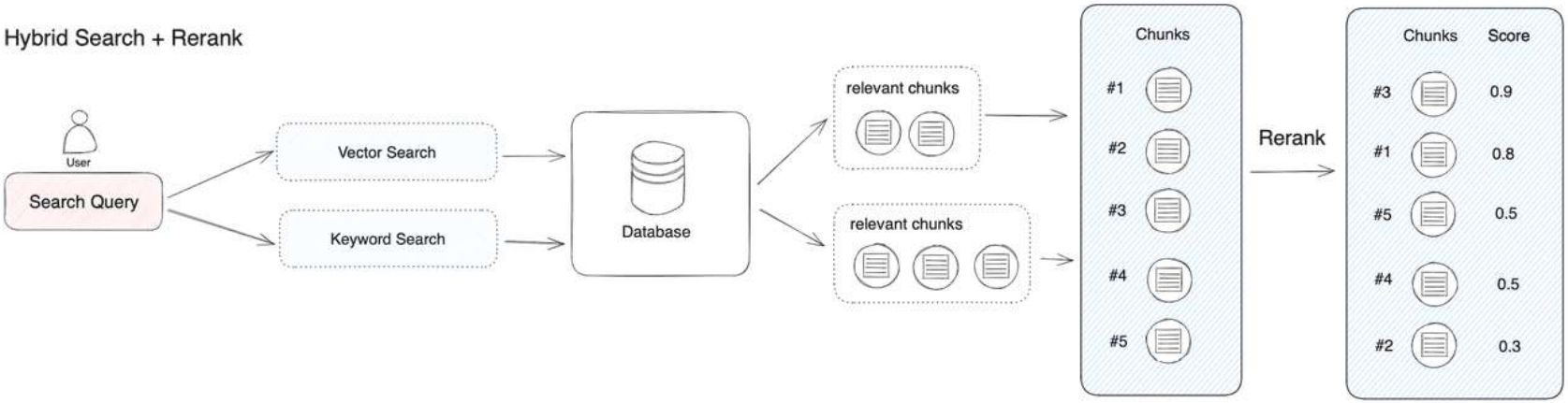
```
hybrid_score = (1 - α) * match_score + α * vec_score
```

DataTalks.Club

Document Reranking

LLM Zoomcamp

Document Reranking



https://dify.ai/blog/hybrid-search-rerank-rag-improvement

Relevance score

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