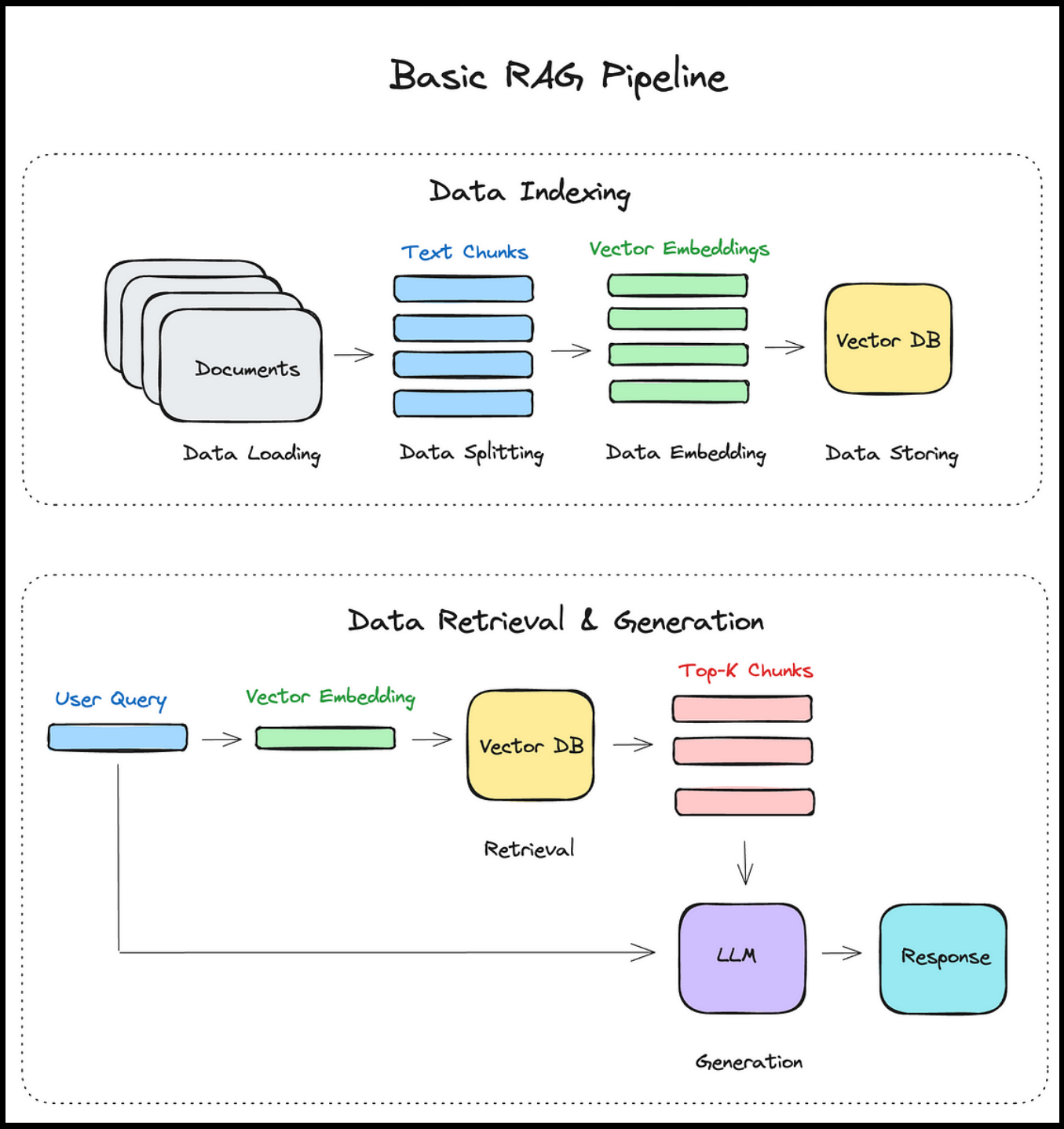
Third week

NLP Internship

Learning is slow strength. Each step, even the hard ones, builds the mind you're becoming. Struggle is not failure—it's growth in progress. Keep going; tomorrow's clarity is built from today's persistence.

Table of content


Basic RAG Pipeline
Data Indexing
Text Chunks
Vector Embeddings
Documents
Vector DB
Data Loading
Data Splitting
Data Embedding
Data Storing
Data Retrieval & Generation
User Query
Vector Embedding
Vector DB
Top-K Chunks
Retrieval
LLM
Response
Generation

First Watch these videos :

Part one : Introduction to RAG

[video 1] (20 min)

https://youtu.be/LIFyBPDkluI?si=xU8XpyJshSoST8dY

[video 2] (24 min)

https://youtu.be/s00dkGuK3rE?si=DJd77EciJqvvNgxm

[video 3] (31 min)

https://youtu.be/zxGWUJyB2V4?si=a_pdIvcLBomfe4io

[video 4] (25 min)

https://youtu.be/OGw023kOzR0?si=vVClyA9WByvdp8PJ

[video 5] (30 min)

https://youtu.be/sgRHmT896yY?si=mElV5cSd2Hzdgdr_

Part two : Vector database and FAISS

[video 1] (19 min)

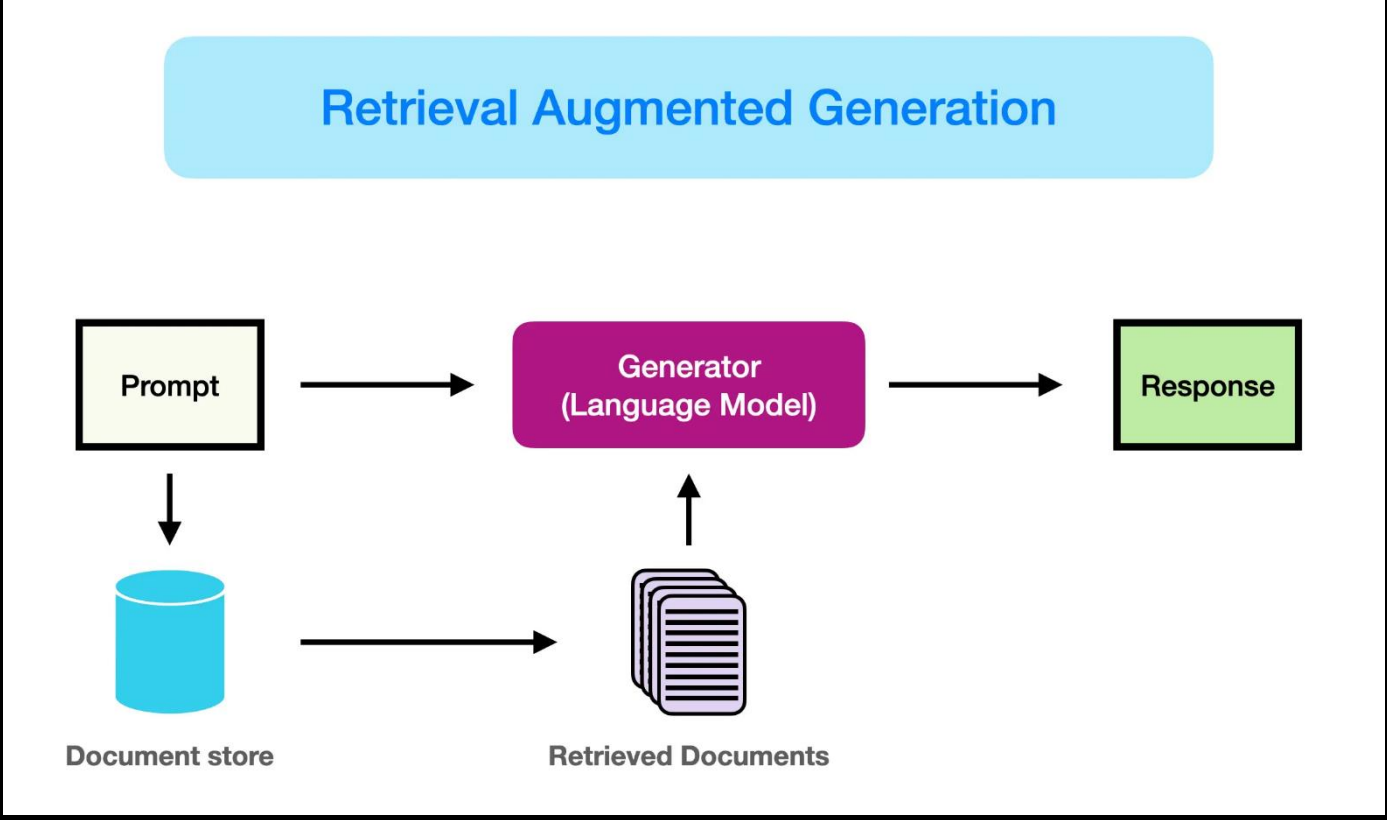
https://youtu.be/rp-IMDekA6I?si=Rb2SEmVQoDa00_HO

[video 2] (45 min)

https://youtu.be/reAmcocQyBA?si=8xgRrqeWzkVP1D8W

[video 3] (52 min)

https://youtu.be/VCnhiF88a9c?si=lYfEqUcrbxThjPVY



Task

Design and implement a RAG code generation system that takes a natural language programming task description and retrieves similar coding examples from the HumanEval dataset, then uses an open source LLM from Openrouter or huggingface to generate a complete function based on the context

Dataset preparation

- Load and process the HumanEval dataset, then Extract task_id , prompt , canonical_solution field

Embedding pipeline

- Use an open source LLM (e.g., Falcon or sentence Transformers) to embed the prompt fields (task description)
- Store these embeddings in a vector database like FAISS or ChromaDB

Code generation

- Use an open source code generation model with coding knowledge to generate python code using the prompt and retrieved code snippets as context

Integration

- **Combine embedding, retrieval and generation into a clean modular pipeline**

Dataset link :

https://huggingface.co/datasets/openai/openai_humaneval/viewer/openai_humaneval/test?views%5B%5D=test&row=26