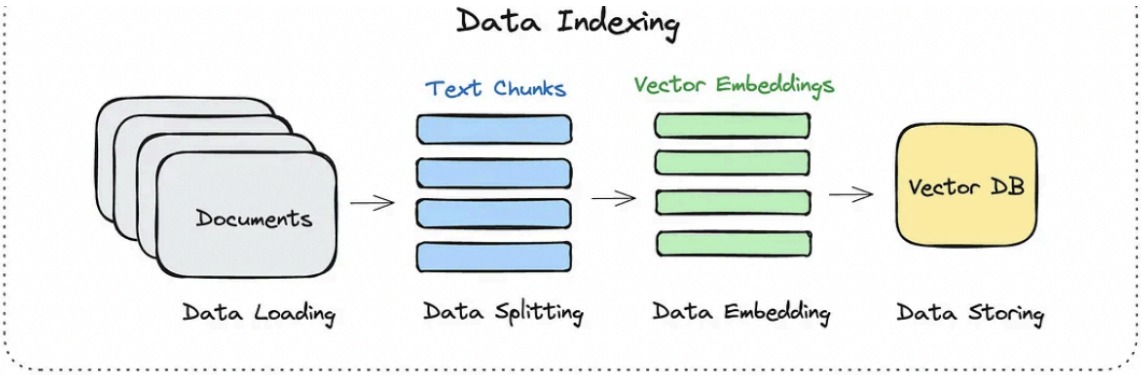
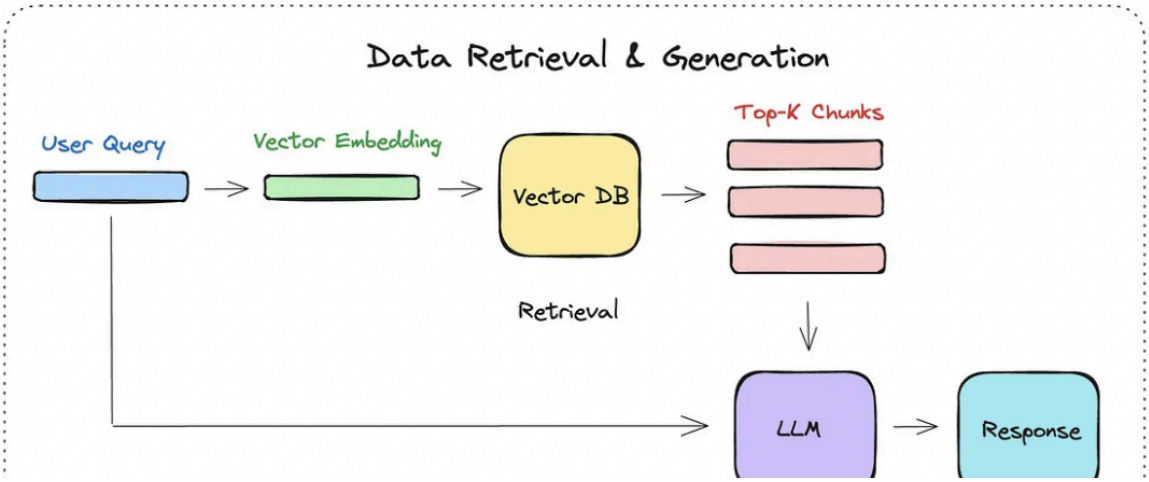
AI Engineering
Sinethemba Pinky Zulu

Week 3 Report: “Ask Verma” — A Physics Study Assistant with RAG





Python

📌 Overview

"Ask Verma" is an intelligent AI-powered chatbot designed to assist students in understanding complex physics concepts by querying the well-known textbook: *HC Verma's Concepts of Physics (Volume 1)*. This system applies **Retrieval-Augmented Generation (RAG)** techniques to extract, retrieve, and generate answers from the textbook.

🎯 Objective

To build a physics study assistant that:

- Understands natural-language physics questions
- Retrieves relevant content from *Concepts of Physics Vol. 1 pdf*
- Generates accurate, context-aware answers.

What is RAG (Retrieval-Augmented Generation)?

RAG is a technique that makes AI models (like ChatGPT or Llama) smarter and more accurate by letting them “look up” information from outside sources (like documents or databases) before answering your question.

Why do we need RAG?

- LLMs are smart, but forgetful: AI models can only answer questions based on what they were trained on, which might be outdated or missing your specific info.
- RAG adds memory: With RAG, the AI can search real documents (like PDFs, websites, or company files) and use that info to give better, more up-to-date answers.

System Architecture

Pipeline steps:

1. Data Ingestion:

- Loaded documents (e.g., `.txt`, `.pdf`) using LlamaIndex's `SimpleDirectoryReader`

2. Chunking:

- Documents split into manageable nodes for optimal retrieval and context window fit.

3. Embedding & Indexing:

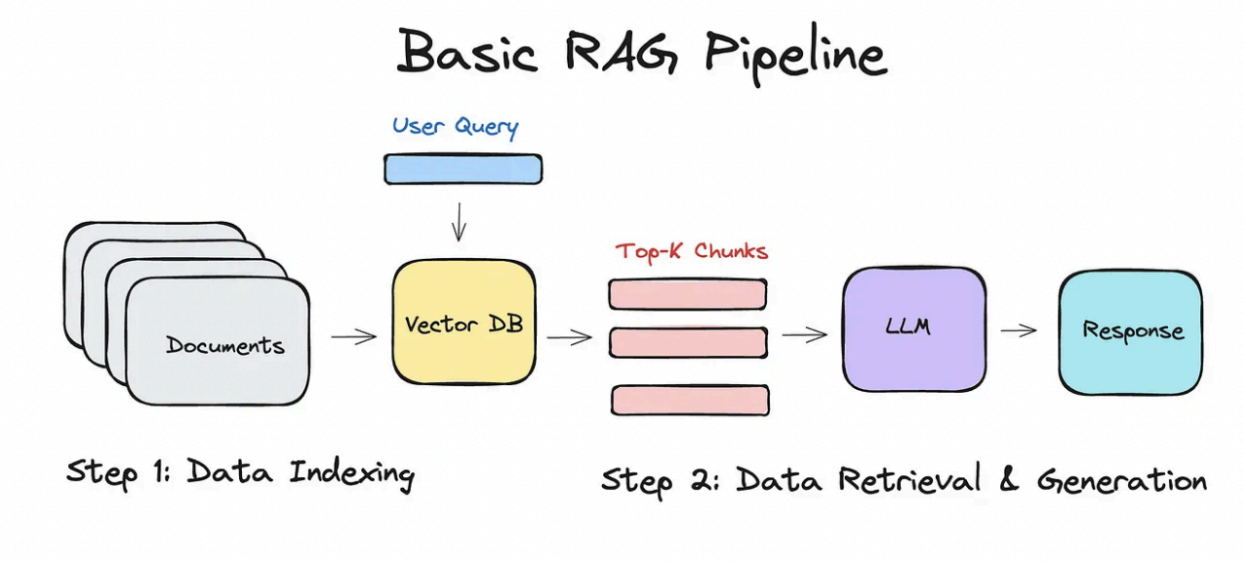
- Used OpenAI embedding models to convert text into vectors.
- Built a `VectorStoreIndex` for semantic search

4. Query Engine:

- Queries are embedded and matched against the index to retrieve relevant chunks.
- Retrieved context is injected into the LLM prompt for grounded generation

5. Response Generation:

- LLM synthesizes answers using both the query and retrieved context, citing sources when appropriate.



Conclusion

The “Ask Verma” RAG showcases the power of combining **OpenAI’s language models** with **context-aware retrieval** using **LlamaIndex** . It serves as a powerful study tool for students, helping them access deep textbook knowledge instantly and interactively.