کوشش، مسلسل و لاتعلق، محنت، ادب

Hassan 25/08/2024

⇒ Lecture #36: Introduction to RAG System

- new technology
- Retrieval Augmented Generation (RAG)
- Quick earning potential

⇒ RAG → (RAG)
- R → retrieval
- A → Augmented
- G → generation

- LangChain

یہ کیا (کس) طرح retrieval

Bridge b/w your data source & ML/model/Application

- sort of an orchestrator
- python framework helps in implementing RAG.

The Challenge of information

- digital age led to explosion of information
- making it difficult to find relevant & accurate data

Traditional Search limitations:
- key-word based search often returns irrelevant results
- difficulty in understanding user intent & context
- inability to provide comprehensive & coherent answer

So we need solution for better relevant and accurate retrieval of data.

Problem with LLM
(providing information in desired output way & better)

- Not access to up-to-date or specific information
- not access to realtime & latest domain specific information

So we need LLM with our data

⇓

Solution

① Fine tuning with latest data

ایک ماڈل جو پہلے سے specific data پر ٹرین تھا۔
تو وہ ماڈل میرے نئے ڈیٹا پر بھی اس سے بہتر
کام کرے۔

Problems :- LLMs are very large and their training is very costly & time consuming

② Retrieval-Augmented Generation / in-context learning

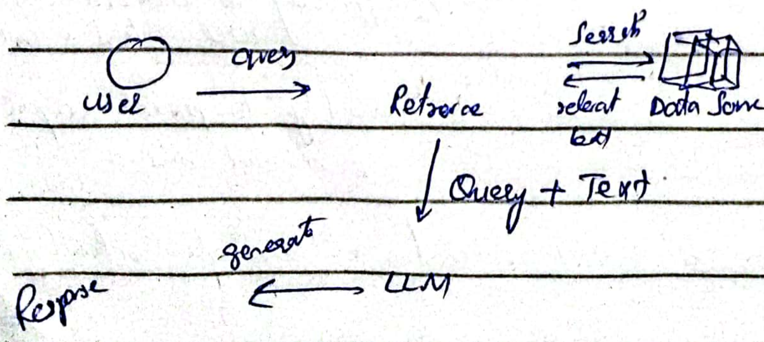
• ایک LLM ہے اس کی power ہے۔
ہم LLM کے ساتھ اپنا data
attach کر دیتے ہیں۔
• ہم اس کی analyze power استعمال کرتے ہیں۔

" RAG first retrieves relevant document/data from a knowledge base & then use that info to generate response "

- giving our own context
- can also fix to only response for our data not the LLM.

⇓

Combine [retrieval-based] & [generation-based]

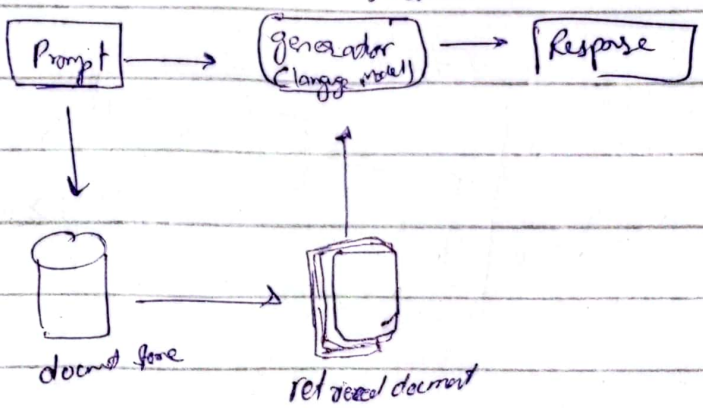


End goal:- At the end of series you will be able to build any applicat

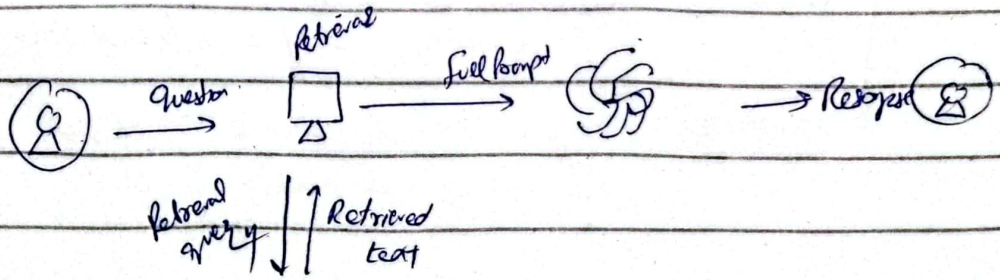
1) How data is attached to our system.
2) LLM also work similar to like this. (difference chatgpt is trained on short data)

⇒ How RAG Works

① Query Input : user ask a question
② Document Retrieval :- identify relevant docume
③ Response Generator :- generator crafts a response using retrieved information
④ Final Output :- a factually accurate & contextually relevant answe.



→ Response on our data
→ only want to fix generate response on our data.

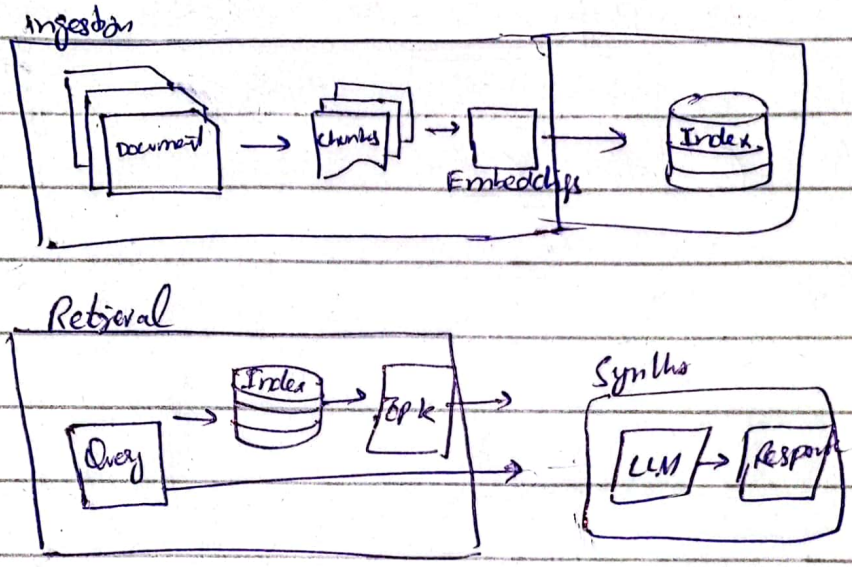


RAG Pipeline

(i) ingestion: feeding raw data - collecting, inputing & processing data
(give the data in easy words)

(ii) Retrieval: correctly retrieval only desired information

(iii) Synthesis :- So modified readable response.



why RAG

① Combine both world (retrieval based + generative based)
② Response are accurate & enriched (more control)
③ Contextual Awareness
(maintain query & retrieved document)
④ improve the performance of bagles
(offers all abilities of LLM)
:- question answering, summarization & more

Refer to the documentation

intro to LangChain

→ frame work for building application using LLM

→ components :- include tools - like document transformers
- connects to various data sources
- chaining multiple queries

Ingestion ✓

5. **Document Loading** :- process of ingesting document into system

supported multiple format plain text, HTML, PDF

Practical Examples (learn & Experiment)

- performance of RAG depend on ingestion
- installing Langchain library, langchain community

Loaders

component designed to ingest & pre-process data from various source

example
- PDF (.pdf)
- text (.txt)
- document (.docx)
- urls

Pypdf

∴ different loaders are with different nature.

∴ can add and edit metadata