Day 21: Retrieval-Augmented Generation (RAG)

Intro

What

LLMs store factual knowledge in their parameters

They achieve great results when fine-tune for specefic tasks

But still

They can't access and precisely manipulate knowledge

In knowledge-intensive tasks, their performance lags

If someone wants to deep dive in to a specific domain

Also

They can't update their world knowledge real-time

Eg: Covid

Example: Grok by xAI

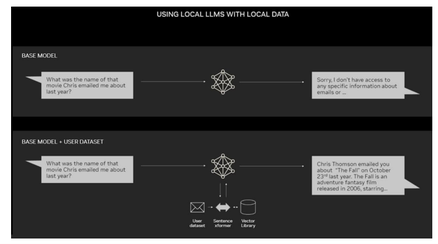
RAG architecture solves this problem

“a general-purpose fine-tuning recipe”

Helps LLMs connect with practically any external resource

Almost any business can turn its technical or policy manuals, videos or logs into resources called "knowledge bases" that can enhance LLMs

AWS, IBM, Glean, Google, Microsoft, NVIDIA, Oracle and Pinecone are adopting RAG



Example use-case:

users can link to a private knowledge source

Working

Steps:

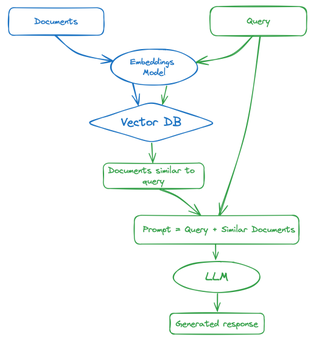
1. User asks LLM a question

2. Model sends a query to another model, which converts it into an embedding

3. Embedding model compares this to a an available knowledge source that is in the form of vectors

4. LLM combines own response + retrieved words into a final answer

LangChain is preferred for chaining together LLMs + embedding models + knowledge bases



Architecture

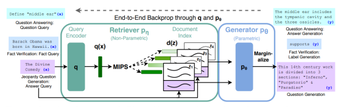


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder + Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Retriever

Fetches relevant information from a large corpus

It is based on models like BERT (Bidirectional Encoder Representations from Transformers)

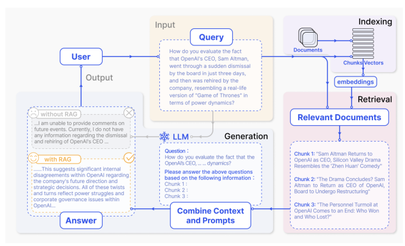
Effectively ranks documents based on their relevance

Generator

Takes the relevant info from the Retriever

Generates contextually appropriate responses

Uses transformer based seq2seq models like - GPT-3 or T5



Hands-on