Retrieval Augmented Generation

What, Why and How

Atul Shukla

Founder, Makeshorts.ai

What is RAG?

Retrieval Augmented Generation (RAG) is a technique that combines information retrieval with text generation, allowing AI models to retrieve relevant information from a knowledge source and incorporate it into generated text.

Origins and Evolution

Original Paper - https://arxiv.org/abs/2005.11401v4

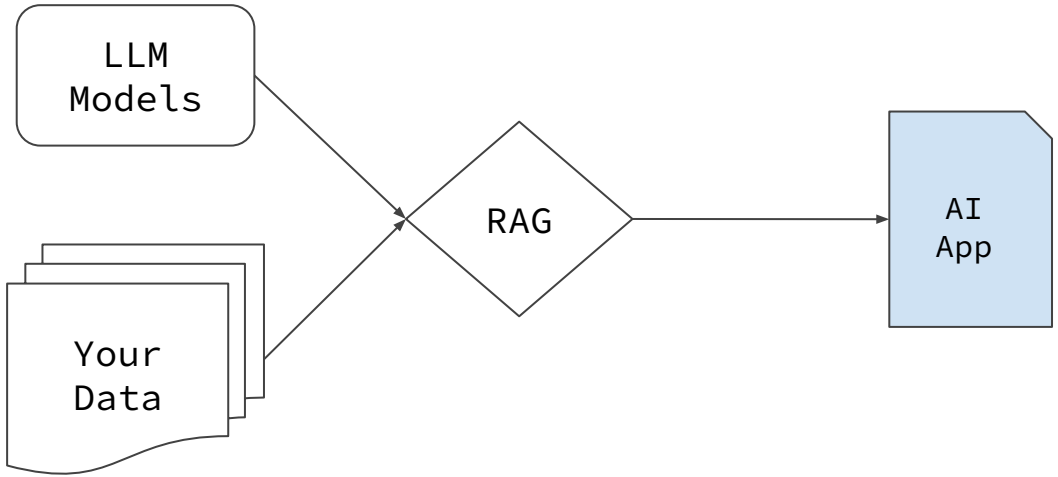
- Originated in Facebook, RAG, a method that combines two types of memory: one that's like the model's prior knowledge and another that's like a search engine, making it smarter in accessing and using information.
- RAG impressed by outperforming other models in tasks that required a lot of knowledge, like question-answering, and by generating more accurate and varied text.
- This breakthrough has been embraced and extended by researchers and practitioners and is a powerful tool in building generative AI applications.

Why RAG

- Overcomes limitations with LLMs
  - LLMs could generate text based on the data they were trained on
  - LLMs lack ability to source additional information during generation process.
- Makes text generation more accurate
  - The retrieval model and generative model work together to provide answers that are accurate and contextually rich

How to build RAG based Applications

Basic Architecture



Show me HOW?

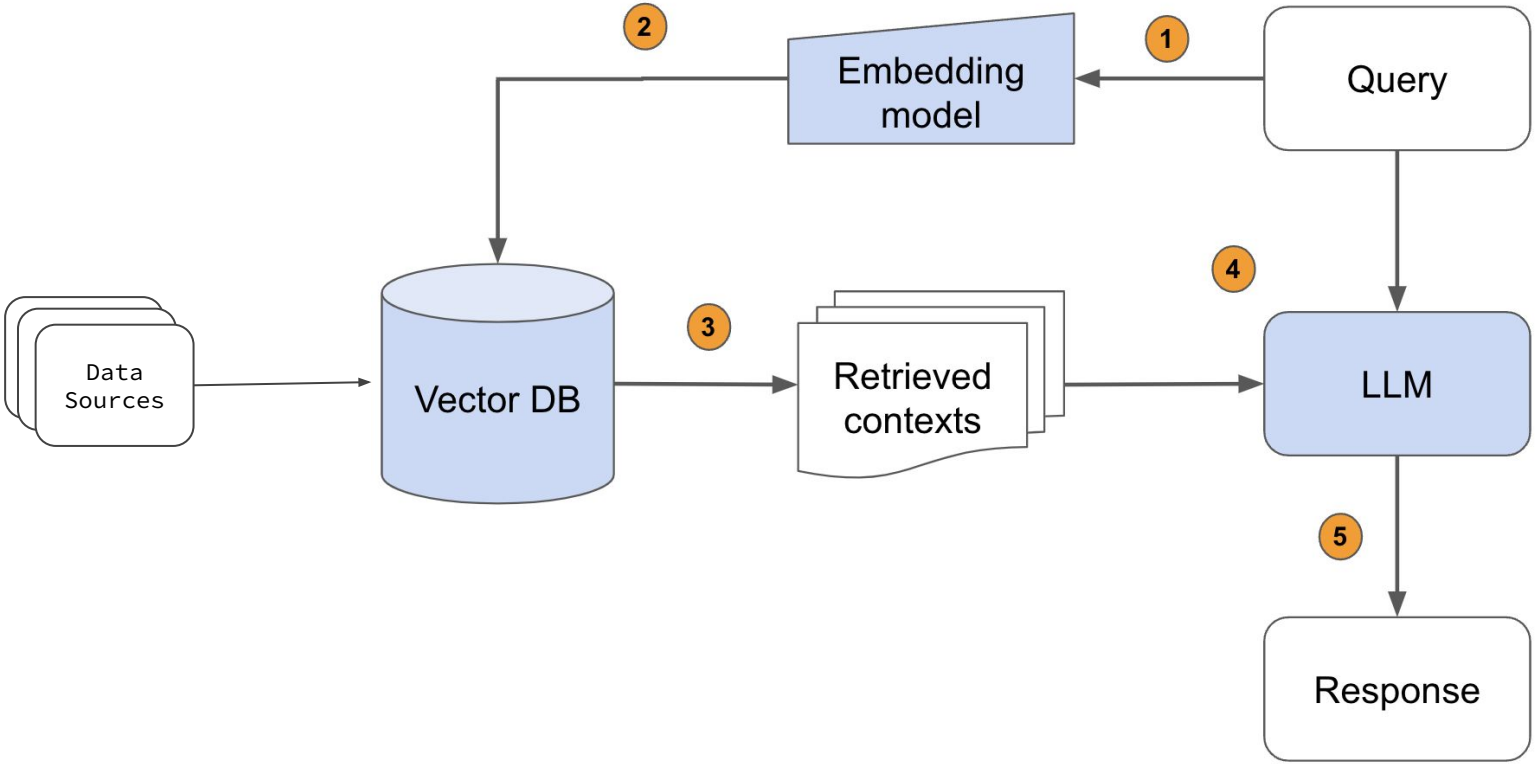
Let’s build a basic Text Summarizer

Lets dig deeper

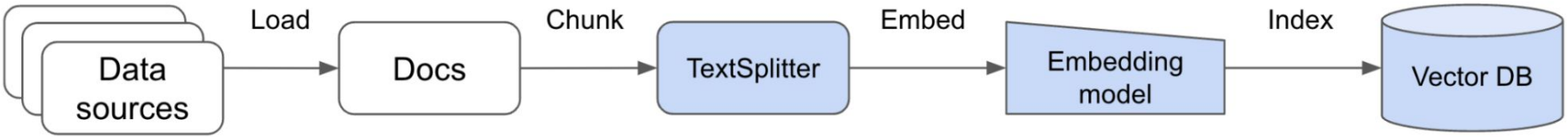
- Data Preparation
  - Extraction and Cleaning
  - Data Chunking
- Embeddings
- Vector Databases
- Reranking
- Lexical Search and Retrieval
- Using multiple LLMs

Notebook Link to follow

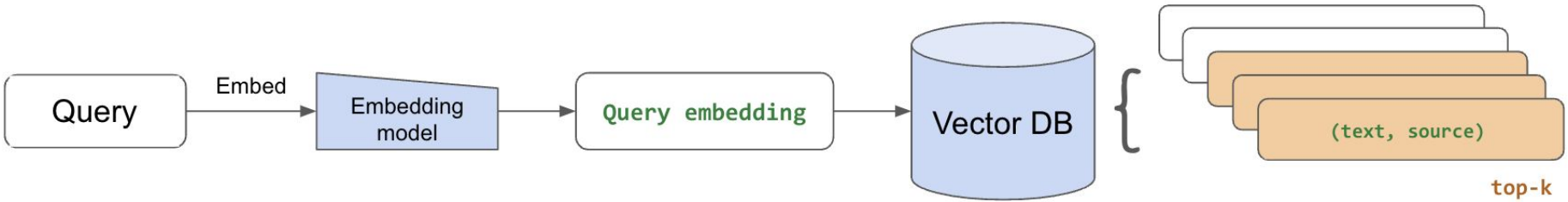
Reference Architecture



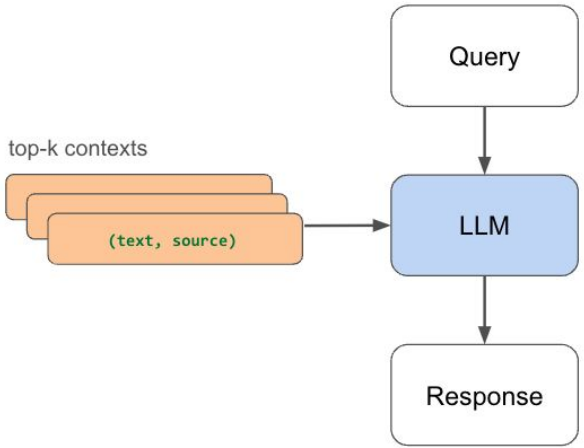
Source Data to Embedding to VectorDB



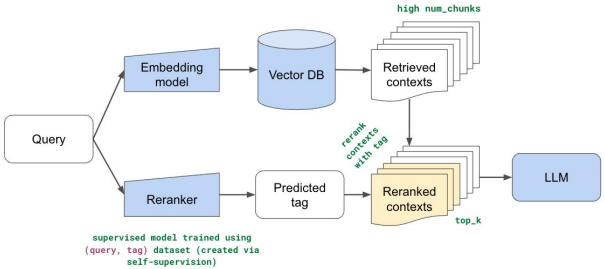
Retrieval based on a Query



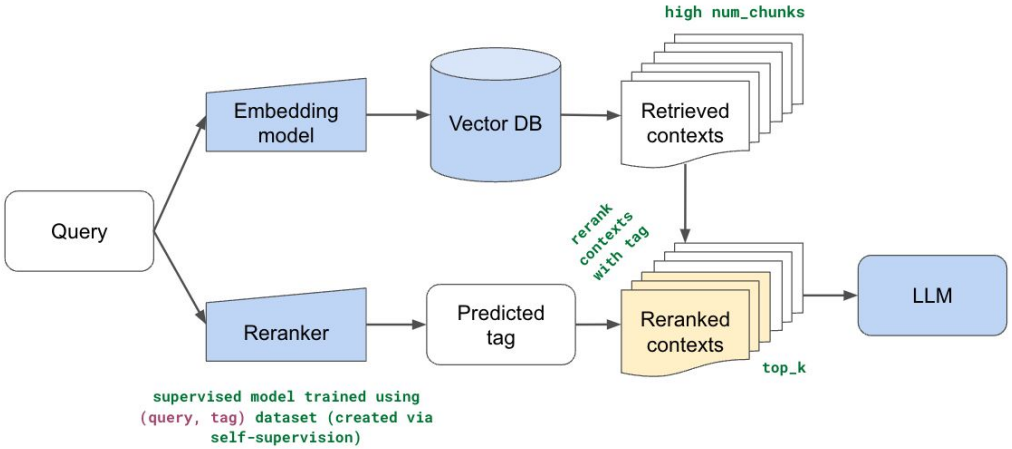
Response Generation



Reranking

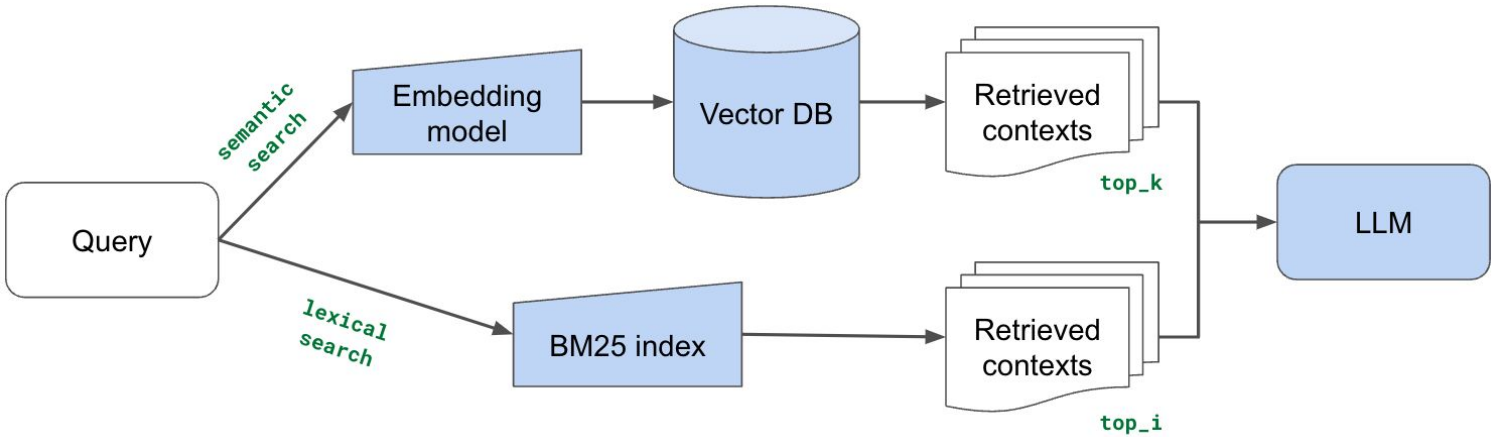


No Reranking

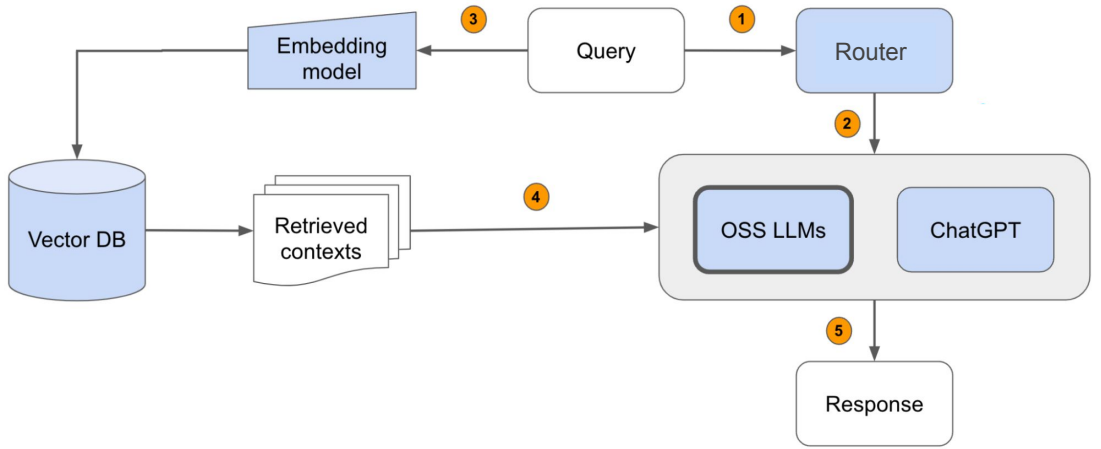


With Reranking

With Lexical Search (BM25)



Using Multiple LLMs



More Application Ideas

- Question Answering
- Content Generation
- Query based Video/Audio Editing
- Using Multimodal data and generation

Thank you