Retrieval Augmented Generation

What, Why and How

Atul Shukla

Founder, Makeshorts.ai

atul@kollabia.com

6 Feb 2024, IIIT-Bangalore

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

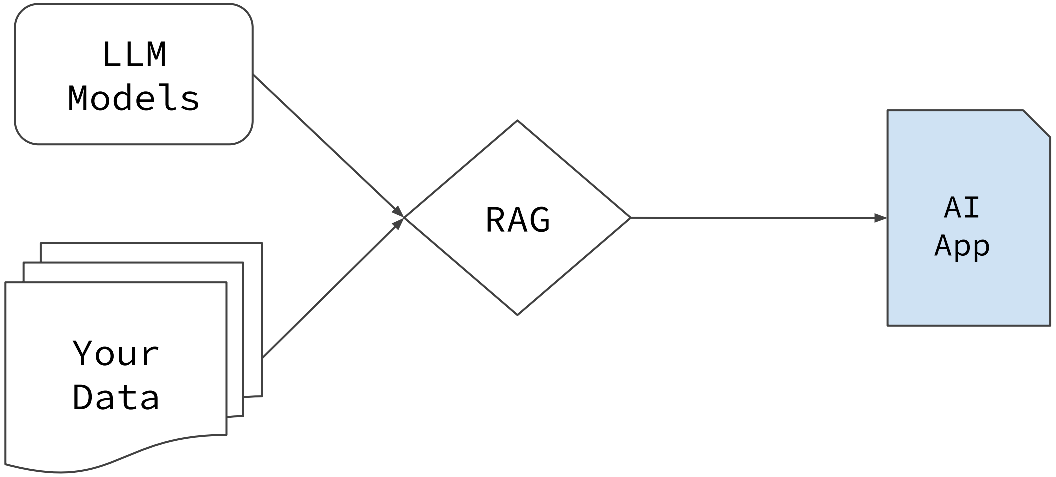
- Originated in Facebook, RAG, a method that combines two types of memory: one that's like the model's prior knowledge and another that's like a search engine, making it smarter in accessing and using information.
- RAG outperforms other models in tasks that required a lot of knowledge, like question-answering, 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?

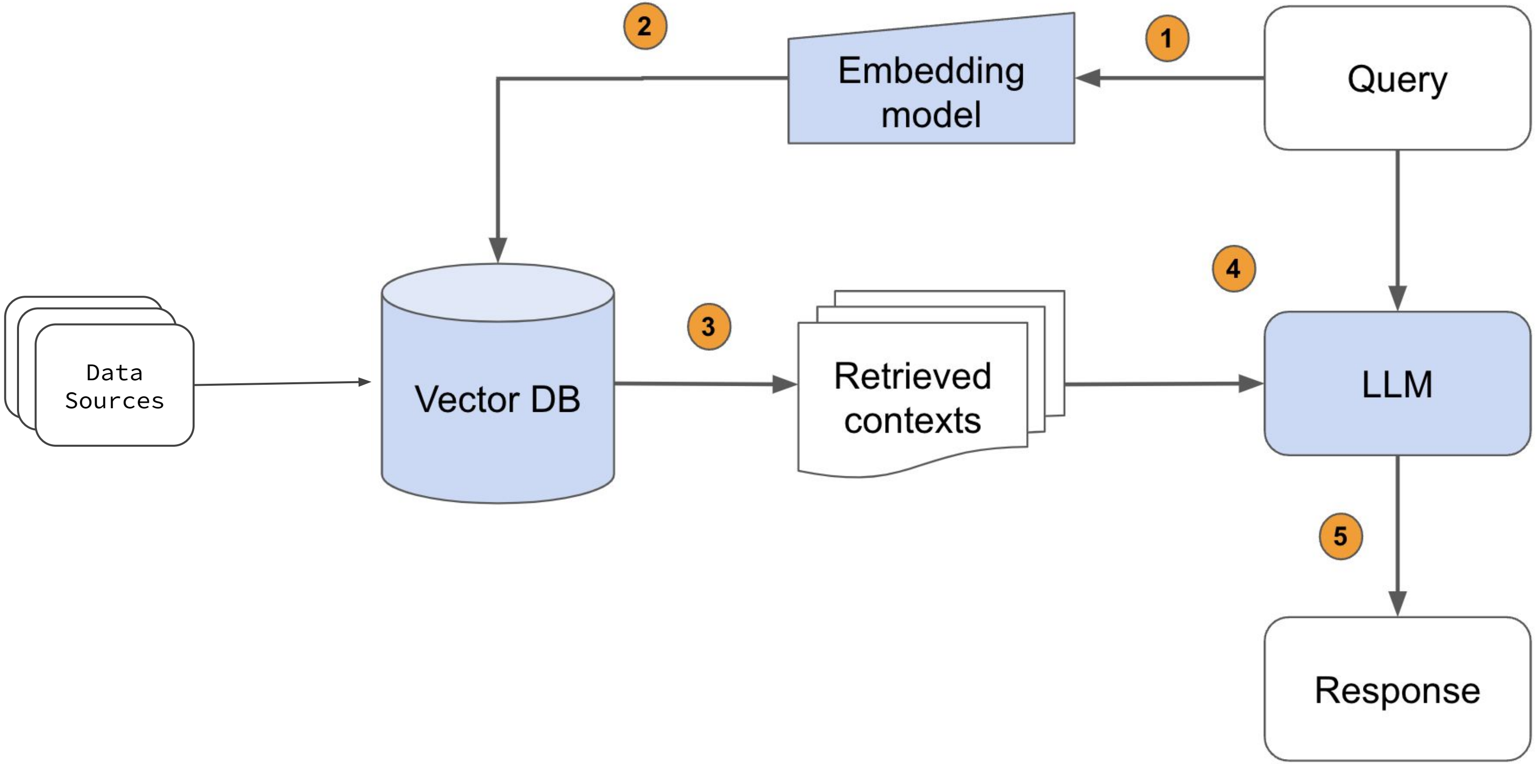
Let’s build a basic Text Summarizer

Lets dig deeper

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

https://colab.research.google.com/drive/1iKp6NNpb2iO-sUkucp3cgp65wD7H3nf3?usp=sharing

Reference Architecture



Vector Embeddings

- Vector embedding maps high-dimensional data into lower-dimensional continuous vector spaces while preserving essential characteristics.
- It captures semantic relationships between data points, enabling algorithms to understand similarities and differences.
- Vector arithmetic can be applied, such as "king" - "man" + "woman" resulting in a vector close to "queen."
- Common methods for generating vector embeddings include Word2Vec, GloVe, and FastText.

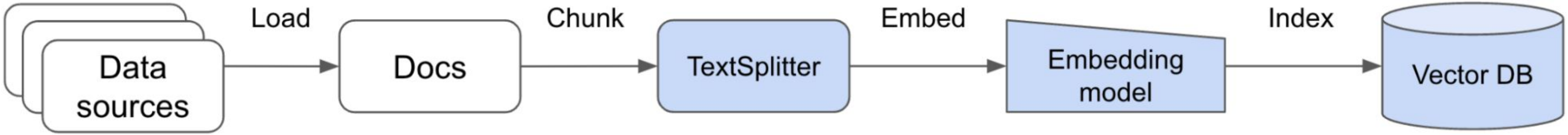
Vector Databases

- Vector databases are a type of database optimized for storing and querying vector data, such as embeddings and high-dimensional vectors.

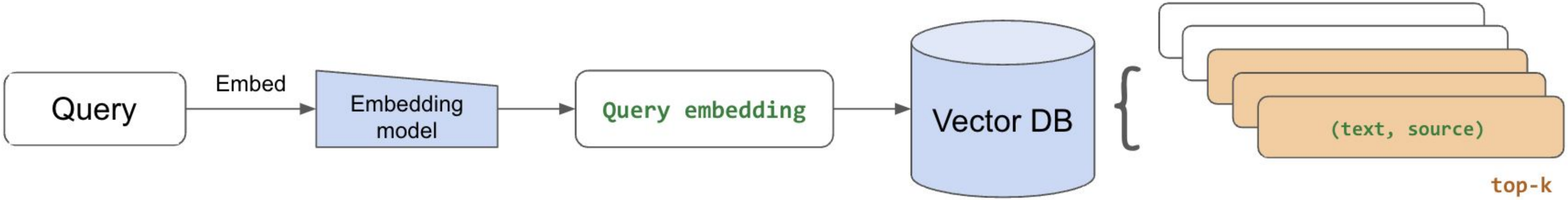
- Support for vector-specific operations like similarity search, nearest neighbor search, and clustering.

- Examples
  - Chroma, Milvus, Weaviate
  - PostgreSQL with PGVector Plugin
  - Elasticsearch with Vector-Scoring Plugin

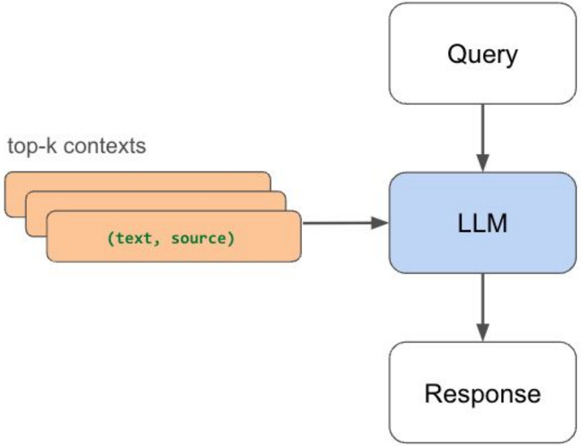
Source Data to Embedding to VectorDB



Retrieval based on a Query



Response Generation



Semantic Rerankers

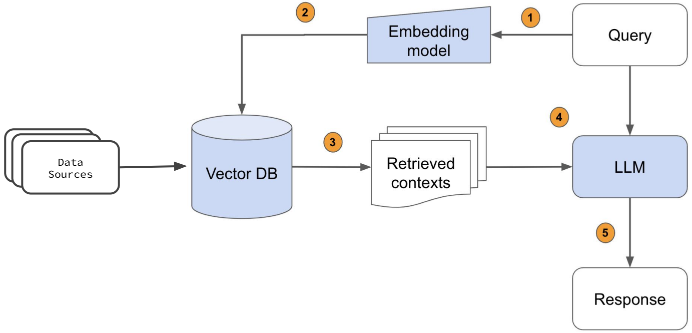
What are Rerankers?

- Rerankers are algorithms designed to improve the relevance and quality of search results by reordering them based on specific criteria.

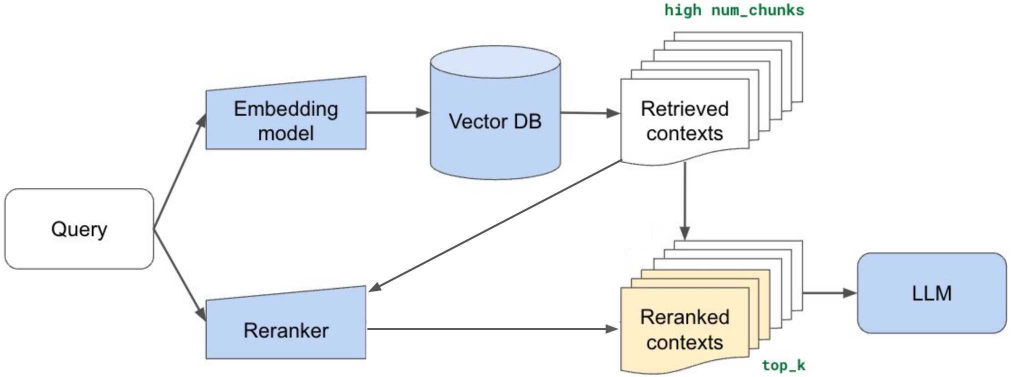
Purpose of Rerankers:

- **Enhance Search Results:** Rerankers aim to deliver more accurate and contextually relevant results to users.
- **Optimize Ranking:** They adjust the ranking of search results to better match user intent and preferences.

Reranking



No Reranking



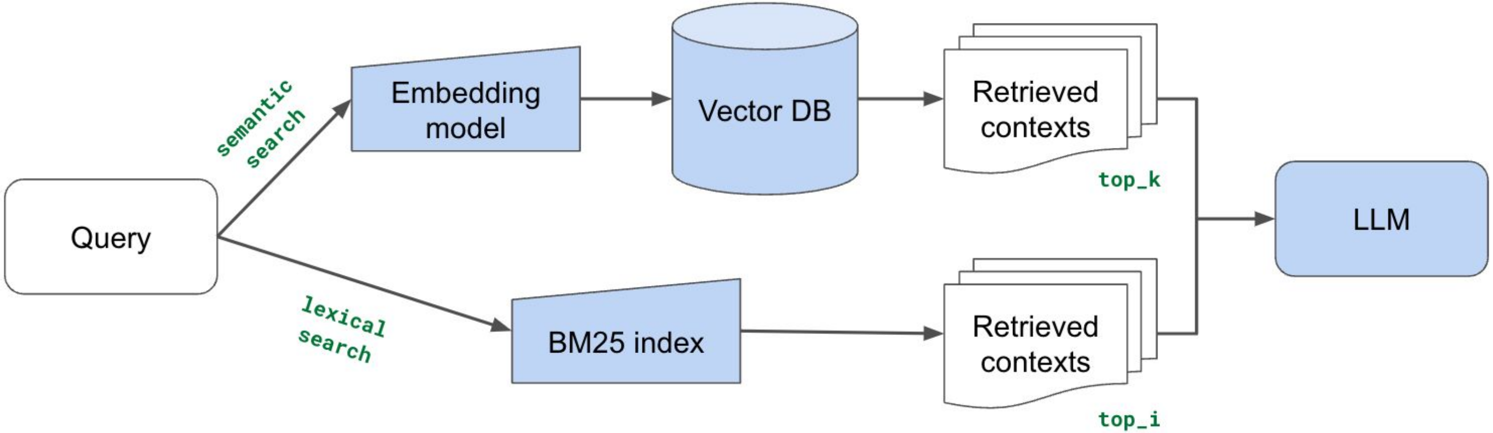
With Reranking

BM25 Overview

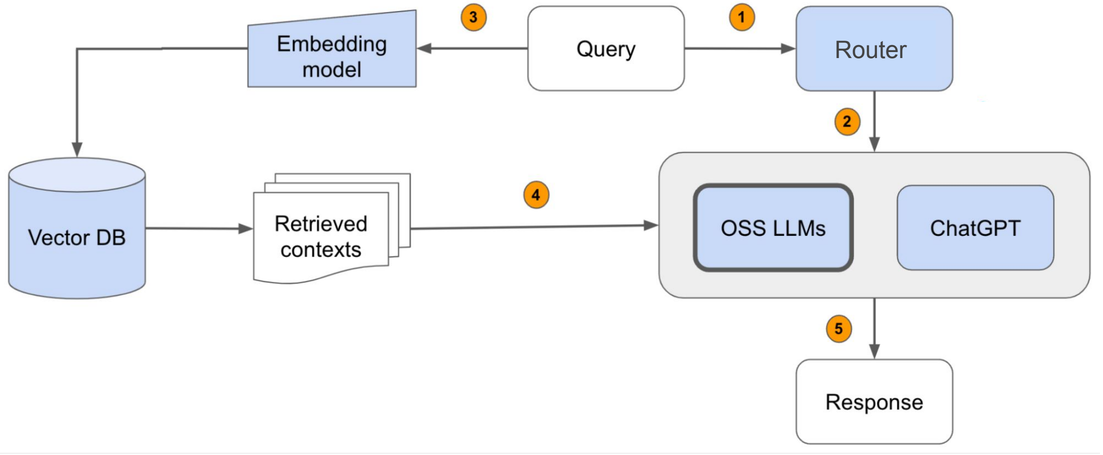
BM25 is a ranking algorithm used in information retrieval systems to estimate the relevance of documents to a given search query.

- **What it does**: It looks at how often your search words appear in a document and considers the document's length to provide the most relevant results.
- **Why it's useful:** It's perfect for sorting through huge collections of documents, like a digital library, without bias towards longer documents or overused words.

With Lexical Search (BM25)



Using Multiple LLMs



Open source projects

Noteworthy Open source projects to build RAG for production

- https://llamaindex.ai/
- https://www.langchain.com/
- https://github.com/BerriAI/litellm
- Vector DBs
  - https://www.trychroma.com/
  - https://milvus.io/
  - https://weaviate.io/

More Application Ideas

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

Thank you