SD Hands-on Session

Retrieval-Augmented Generation

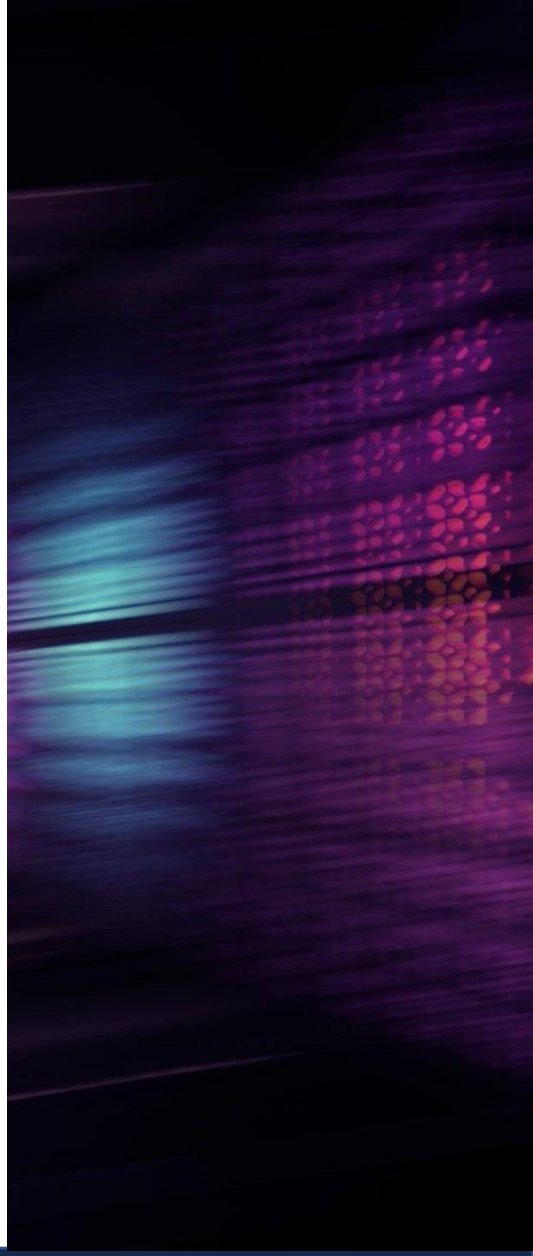
Minu Mathew
Research Software Engineer

Agenda

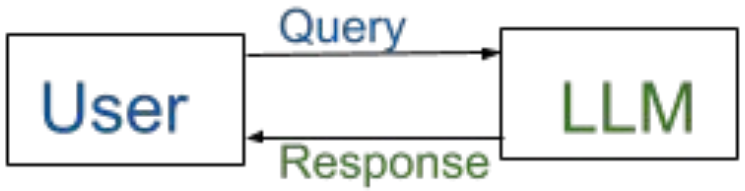
- Motivation
- Intro
- UIUC.chat
- RAG explained
- Coding!

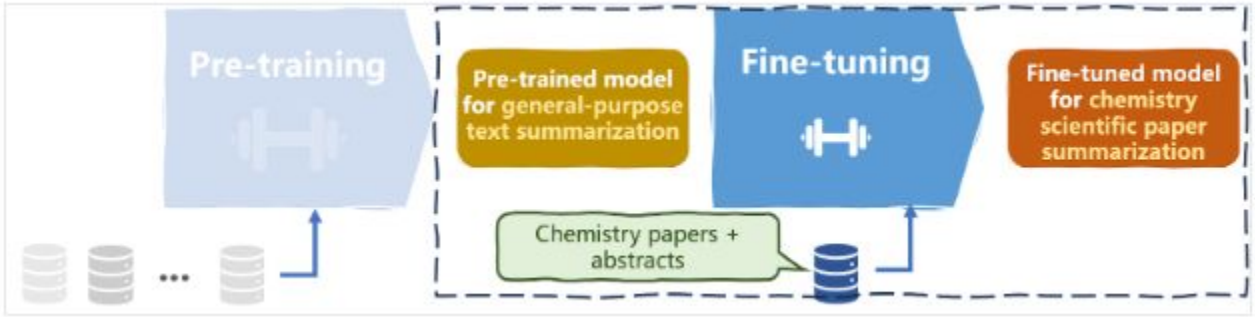
Motivation

- LLM - Large pre-trained Language Models :
  - No ability to access a specific knowledge
    - If specific domain knowledge is required - fine tune.
  - No sources
  - Hallucinations
  - Out-of-date information



Motivation





Fine-tuning LLM
Source: [datacamp blog]

Explain to me what an llm is

An LLM, or Master of Laws, is a postgraduate degree in law that is typically pursued by individuals who have already completed a bachelor's degree in law (LLB or JD) or have significant legal experience. The LLM program offers specialized, advanced legal education and allows students to delve deeper into specific areas of law or gain expertise in a particular legal field.

When did France gift Lithuania Vilnius TV tower?

France gifted Lithuania the Vilnius TV tower in 1980.

What prompted the French government to provide the gift?

The French government provided the gift as a gesture of friendship and solidarity with the Lithuanian people, who had recently declared independence from the Soviet Union.

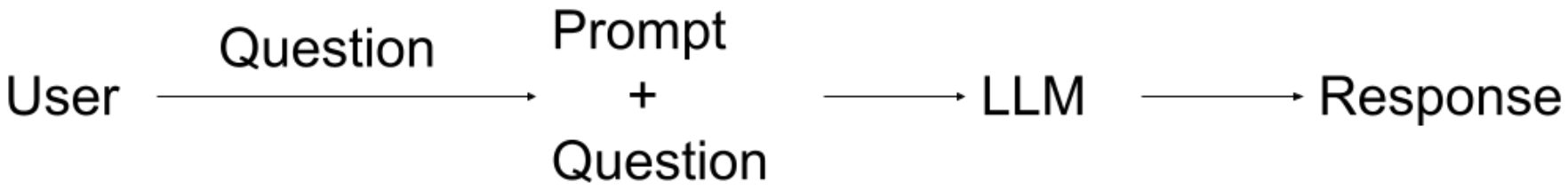
LLM Hallucinations

Introduction

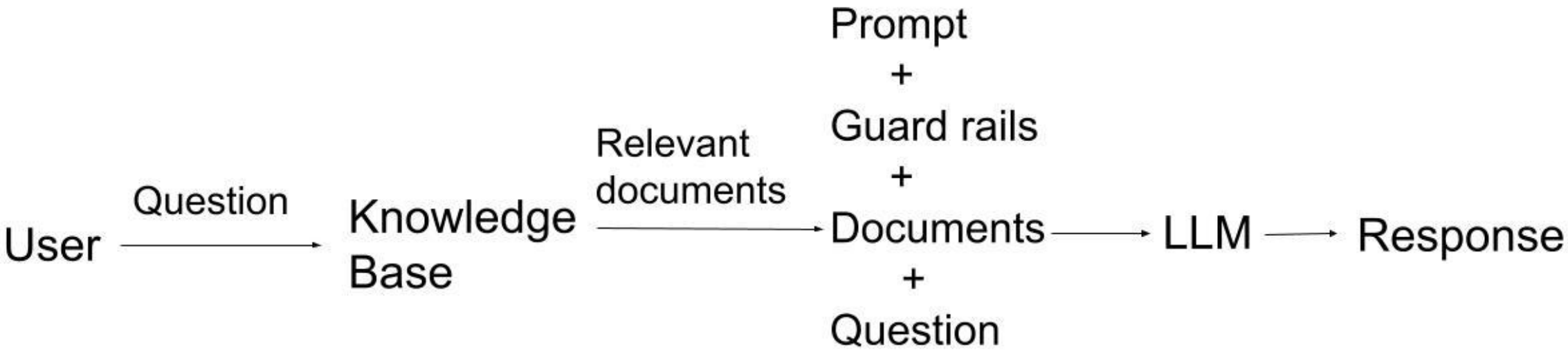
- RAG (Retrieval-Augmented Generation)
  - Introduced in 2020 (paper)
  - Generation - by LLMs
  - Retrieval-Augmented
    - Retrieve required information from provided knowledge base
    - Provide this information to LLMs as context to answer user query

Introduction

- Before :



- After :



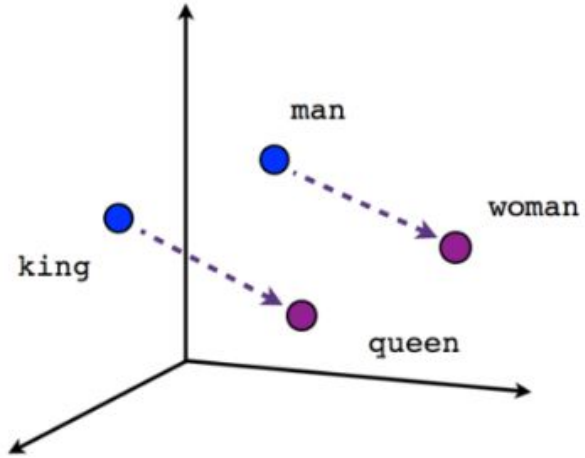
Without RAG	With RAG
• No ability to access a specific knowledge/domain	• Point to a knowledge base
• No sources	• Sources cited in LLM response
• Hallucinations	• LLM response is grounded by relevant information from knowledge base
• Out-of-date information	• Update the knowledge base with new information

UIUC.chat - RAG in Action

- UIUC.chat (https://www.uiuc.chat/)
  - Login/Sign up with registered email id
  - Public projects - ECE120, NCSA, NCSA Delta, Clowder…
  - Project - HandsOnRAG (private project page, with OpenAI key)
  - https://www.uiuc.chat/HandsOnRAG/materials - upload your documents
  - https://www.uiuc.chat/HandsOnRAG/chat - chat with LLM
  - https://www.uiuc.chat/HandsOnRAG/prompt - change system prompt

RAG - Retrieval-Augmented Generation

Knowledge DB



- Vector database (Beginners blog 1, Pinecone blog 2)
- Efficiently store, index and search high-dimensional data
- Store data using vector embeddings
- Optimized for fast retrieval and similarity search
- Calculate the distance between user query embedding and other data points

RAG - Retrieval-Augmented Generation

Vector DB Retrieval

1. Partition data into chunks (to be processed by embedding models)
2. Use embedding model to create vector embeddings for the data (create indexes)
3. Insert vector embeddings into the db, with some reference to the original content/metadata
4. User query - use same embedding model to create user query embedding
5. Query the db for similar vector embeddings
6. Return similar document chunks.

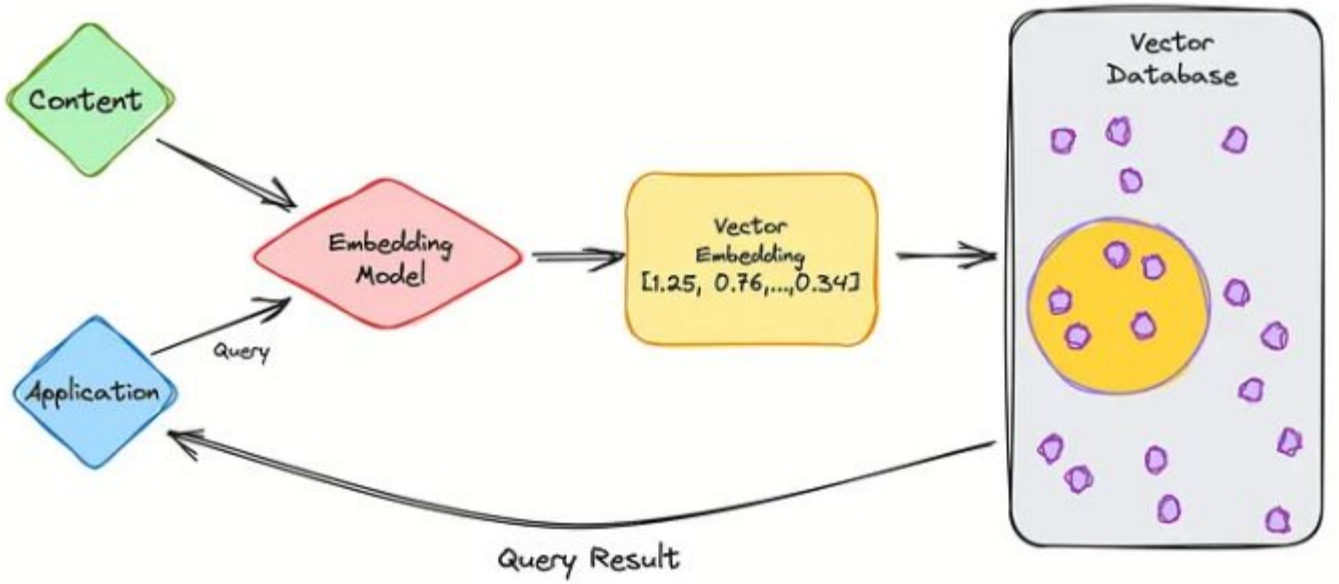


Image Credits : KDnuggets

RAG - Retrieval-Augmented Generation

LLM

- Pre-trained transformer models
- Trained to predict the next word (token), given some input text.
- Autoregressive generation - iteratively calling the model with its own generated output.
- Open-source models - HuggingFace leaderboard

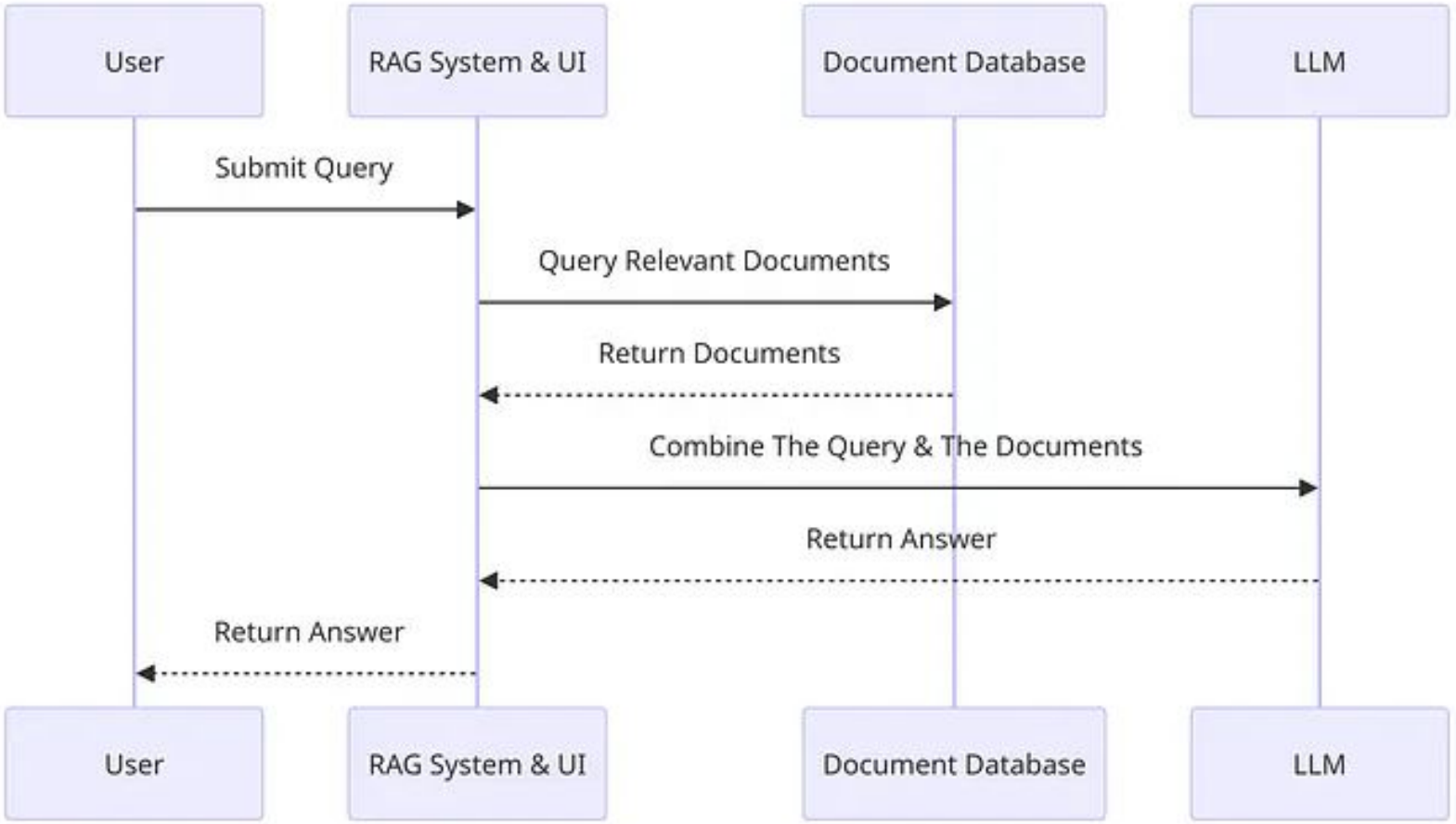
RAG - Retrieval-Augmented Generation

Prompting

- Communicate with LLMs to get desired outcomes without updating the model
- Takes multiple trial-and-errors to get desired effect.
- Include specific persona / behaviour
    - Eg: “You are a helpful research assistant”
- Include guard rails
    - Eg: “If you don’t know the final answer, just say “I don’t know”
- Include instructions
    - Eg: “Read the data file before answering any questions”
- Include response formats
    - Eg: “Respond using markdowns”
- LilianWeng blog post, medium blog post on prompt engineering

RAG - Retrieval-Augmented Generation

1. Query the database
2. Retrieve relevant information
3. Pass relevant documents + user query to LLM
4. Pass LLM response to user

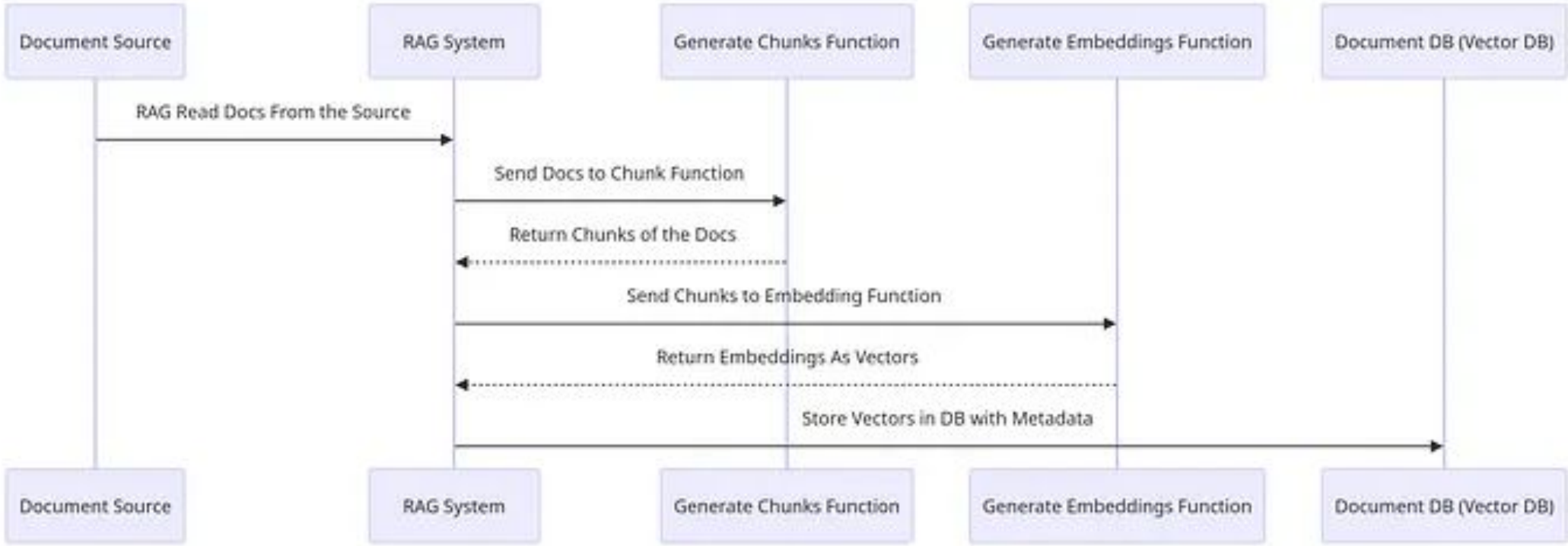


Source Credits : [Blog.demir]

RAG - Retrieval-Augmented Generation

Inserting into DB

1. Query the database
2. Retrieve relevant information



Source Credits : [Blog.demir]

RAG - Hands-on

- Github
- https://github.com/minump/HandsOn-RAG