Hands on GenAI - RAG

Social Developers Club Conference 2025

Agenda

- About Retrieval Augmented Generation (RAG)
  - What is RAG?
  - How does RAG work?
  - Why is RAG important?
  - Applications of RAG
- Our small showcase
  - Setup
  - Parameter
  - What you can do

What is Retrieval Augmented Generation (RAG)?

- It's a technique used to enhance AI-generated responses by combining two key components:

- **Retrieval**: Searching for relevant information from a large corpus or database.
- **Generation**: Using this retrieved information to generate more accurate and contextually relevant responses.

How Does RAG Work?

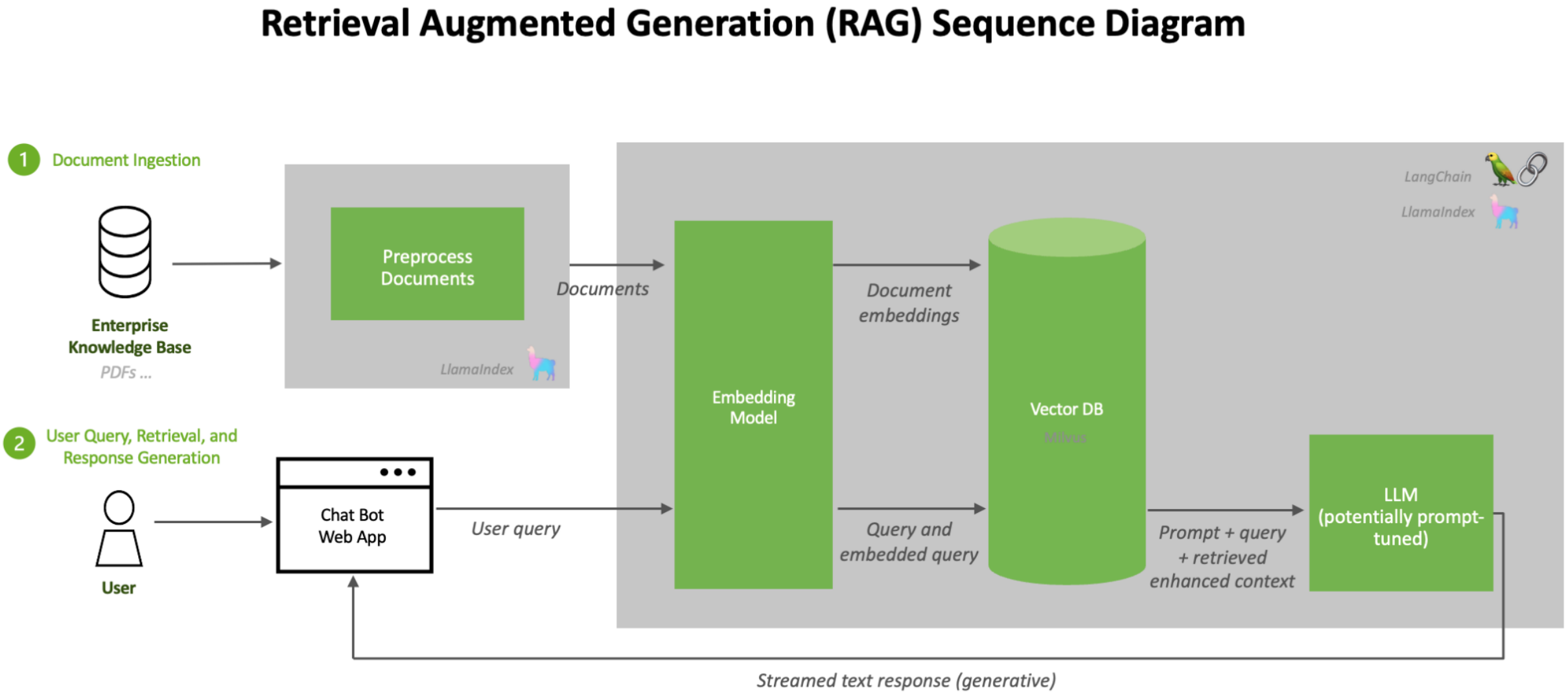
- **Step 1: Information Retrieval**

- When a question or prompt is given, RAG first retrieves relevant pieces of information from a database, document, or other sources.
- This helps to ground the response in actual data, not just patterns learned during training.

- **Step 2: Text Generation**

- After retrieving the relevant information, the system uses this data to generate a detailed, context-aware answer.
- The generation model adapts the retrieved information into the final response.

How Does RAG Work?



Why is RAG Important?

- **Improved Accuracy**: It combines the benefits of retrieval (searching for real-world data) and generation (producing human-like text).
- **Reduces Hallucinations**: RAG minimizes errors where models generate false or irrelevant information by grounding the responses in real data.
- **Contextual Relevance**: Responses are more tailored and relevant to the specific query or prompt.

Applications of RAG

- **Customer Support**: Automatically answering complex customer inquiries using up-to-date product information.
- **Research Assistance**: Assisting researchers in finding and summarizing relevant literature.
- **Personal Assistants**: Providing detailed, context-aware answers in virtual assistants.
- **Medical & Legal Fields**: Generating precise, evidence-backed content by pulling relevant legal or medical data.

Our small showcase

- Setup:
  - Have GIT and Docker (with docker compose) installed
  - Clone: https://github.com/AlexDivivi/sdc-rag
  - Start application with docker compose up
  - Get your free Google API Key: https://aistudio.google.com/app/apikey

Our small showcase

- Parameter
  - Google API Key
  - Temperature: 0-1 controls the “creativity” of the response
  - K: Retrieved document chunks from the vector database
  - VDB collection: Collection name for managing documents
  - Chunk size: Size of characters your document will be splitted
  - Chunk overlap: Size of characters your chunks will overlap with each other
  - System Prompt: Guide the behavior of the LLM

Our small showcase

- What you can do
  - Play around with the system
  - Upload documents
  - Change parameter
  - Tune system prompt

- If you want to code
  - Allow multiple file uploads
  - Implement Chat and history
  - Implement metadata and filters
  - Make the LLM response streamed

Thank you

Social Developers Club Conference 2025