RAG System Homework Project

Overview

Your task is to build a Retrieval-Augmented Generation (RAG) system that answers questions using external documents. This project is designed to demonstrate how large language models (LLMs) can be grounded in current, domain-specific knowledge.

The model you'll use has a knowledge cutoff in **August 2024**. Therefore, your system must correctly answer questions about **events occurring after that date** by retrieving relevant documents — not by relying on the model's internal knowledge.

Project Requirements

You must select a document that describes an event that happened after August 2024. Store and index this document using **ChromaDB** with persistence enabled. Apply a text-splitting strategy to break the content into at least 50 meaningful chunks for effective retrieval.

The system must be implemented using the `gemini-2.0-flash` model. You are required to use either **LangChain** or **LlamaIndex** to orchestrate the retrieval and generation process. Additionally, integrate either **LangSmith** or **LangFuse** for observability and debugging. Pre-built agents are not allowed; build your pipeline manually.

Your implementation should support **dialog flow** (multi-turn interactions) and **memory** (context tracking). The system should be capable of holding a conversation over several turns while keeping track of the context provided in earlier exchanges.

As part of your evaluation, design at least five questions that can only be answered correctly using the retrieved content. These questions must be unanswerable by the language model alone. You should demonstrate and explain this by testing the model both with and without retrieval. Also, experiment with different system prompts and report on how they influence the model's behavior.

Your repository must follow best practices. Keep your Git history clean — do not commit large files or secret tokens. The code should be well-documented and easy to run.

Submission Guidelines

The deadline for this project is **11.05 at 23:59**. Each student has a dedicated branch in the repository. You must open a **Pull Request (PR)** from your working branch to your assigned branch before the deadline.

Your PR should include:

- A complete implementation of your RAG system.
- A Jupyter notebook or script that demonstrates document indexing, retrieval, answering, and prompt experimentation.
- A link to your project in **LangSmith** or **LangFuse**.

Bonus Requirements (Extra Credit)

To earn extra credit, your system must implement both:

1. **Metadata filtering** to refine the retrieval process.
2. **Multi-query retrieval**, such as rephrasing the question or running multiple variations to improve the context.

Final Note

Make sure your system proves its effectiveness by answering questions that the model could not answer on its own. The strength of your RAG system lies in how well it uses the retrieved information.