RAG-LLM Project

Problem Statement :- The aim of this project is to create an AI tool that will let anyone ask a question and receive an answer from our internal documents guaranteeing accuracy, speed, and fully offline operation. We will accomplish this through Retrieval Augmented Generation (RAG) technique, powered by a locally hosted open-source LLM eliminating retraining of the model and ensuring complete control over our data.

Here are the details of the model :-

1. **Llama3**
    - Parameters : 8 Billion
      (8 billion parameters)
    - Open Source
    - **Context length : 8192 tokens**
      (model uses a memory of 8192 tokens which on an average transforms to 6000 words ~ 12-15 pages of text, it takes 8192 tokens for context before answering the query)
    - Memory requirement : ~ 4.7 GB RM (4 bit quantized)
    - Will be using ollama(compatible(with windows also as of 2024) and faster with mac) and pipeline techniques like langchain / llama index
      (ollama - helps us to download and locally run any LLM)
      Ollama uses **Metal** (Apple's GPU framework) + llama.cpp
        - LlamaIndex vs LangChain

The github link for the project is here : github_repo

Langchain vs LlamaIndex: A Comparative Analysis

LlamaIndex is primarily designed for search and retrieval tasks. It excels at indexing large datasets and retrieving relevant information quickly and accurately. LangChain, on the other hand, provides a modular and adaptable framework for building a variety of NLP applications, including chatbots, content generation tools, and complex workflow automation systems.

Feature	LlamaIndex	LangChain
Primary Focus	Search and retrieval	Flexible LLM-powered application development
Data Indexing	Highly efficient	Modular and customizable
Retrieval Algorithms	Advanced and optimized	Integrated with LLMs for context-aware outputs
User Interface	Simple and user-friendly	Comprehensive and adaptable
Integration	Multiple data sources, seamless platform integration	Supports diverse AI technologies and services
Customization	Limited, focused on indexing and retrieval	Extensive, supports complex workflows
Context Retention	Basic	Advanced, crucial for chatbots and long interactions
Use Cases	Internal search, knowledge management, enterprise solutions	Customer support, content generation, code documentation
Performance	Optimized for speed and accuracy	Efficient in handling complex data structures
Lifecycle Management	Integrates with debugging and monitoring tools	Comprehensive evaluation suite (LangSmith)

Flowchart of the pipeline

(High level view) :

[User Question]
↓
[Search in Company Docs]
↓
[Pick Relevant Chunks]
↓
[Add Question + Chunks to Prompt]
↓
[Send to the LLM]
↓
[Generate Answer]
↓
[Show Answer to User]

The flow of the project :

[Setup Environment]
↓
[Install LLM and check if it's running]
↓
[Prepare Documents]
↓
[Develop RAG Pipeline]
↓
[Run RAG Script]
↓
[User Interaction]
↓
[Review & Iterate]