Chatbot based on Retrieval Augmented Generation (RAG)

Vidipt Vashist
(MA22M025)

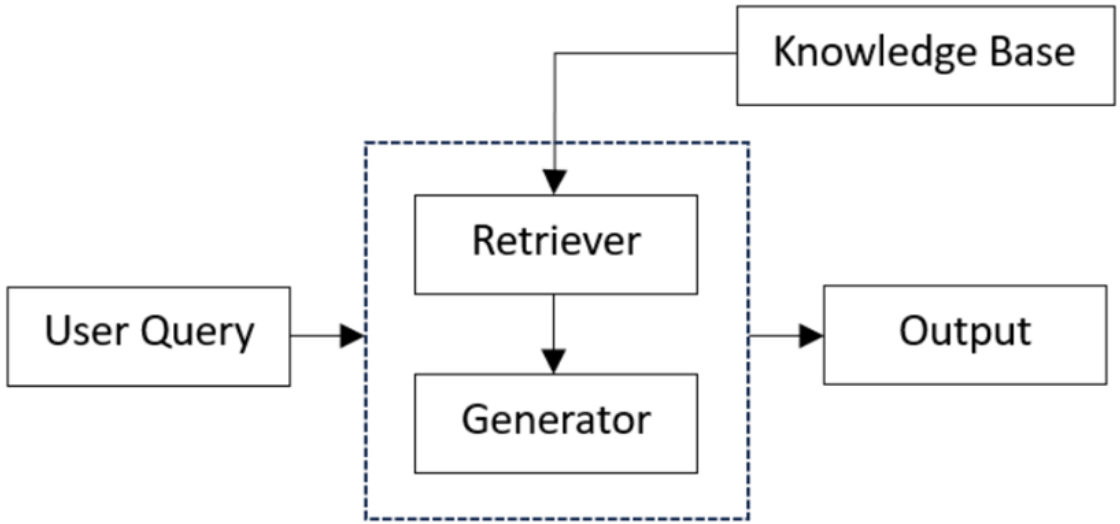
Indian Institute of Technology Madras
Department of Mathematics
MA5990 M.Tech Project

May 13, 2024

INTRODUCTION

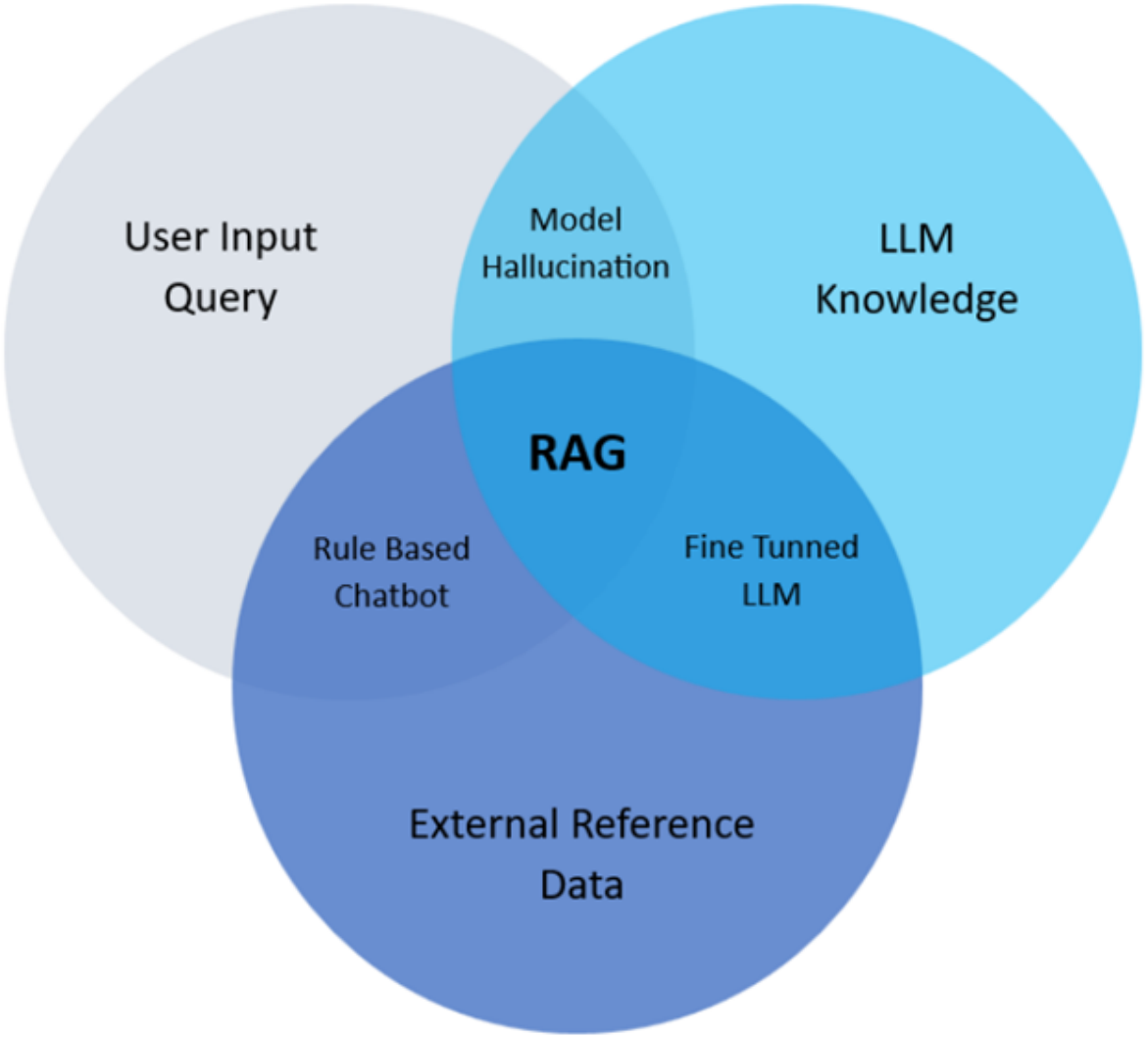
- RAG is is a **groundbreaking paradigm** in natural language processing that combines the strengths of retrieval-based and generation-based approaches.
- It's a **AI framework** for retrieving facts from an external knowledge base to ground large language models (LLMs) on the most accurate, up-to-date information and to give users insight into LLMs' generative process. .
- This innovative **framework combines the best aspects of different methodologies to achieve impressive results** in a wide range of NLP tasks, including question answering and text summarization, among others.

TECHNICAL OVERVIEW OF RAG



Core Idea: Retriever + Generation

TECHNIQUES LANDSCAPE



WHY RAG ?

- **Reduced Hallucinations:** models exhibit fewer hallucinations and higher response accuracy.
- **Enhanced LLM Memory:** the information capacity limitation of traditional Language Models (LLMs). Traditional LLMs have a limited memory. RAG introduces memory by tapping into external knowledge sources.
- **Updatable Memory:** is its ability to accommodate real-time updates and fresh sources without extensive model retraining. information
- **Improved Contextualization:** enhances the contextual understanding of LLMs by retrieving and integrating relevant contextual documents.

PROBLEM STATEMENT

The main goal of this project is to achieve two things:

- firstly, to create a reliable and open-source system that can efficiently extract information from private PDF files: **Internal Document RAG System**
- and secondly, to develop a user-friendly web application that utilizes web scraping techniques to gather relevant data from online sources: **Web Search RAG System**

INTERNAL DOCUMENT RAG SYSTEM

Current Challenges

- Extracting information from private PDF files can pose significant challenges for organizations, leading to complexity and inefficiency in their data extraction processes.
- This challenge impedes their capacity to effectively utilize internal knowledge,
- Resulting in missed chances for improved customer interactions and a competitive advantage in the market

Solution

to create a reliable and open-source system that can efficiently extract information from private PDF files

WEB SEARCH RAG SYSTEM

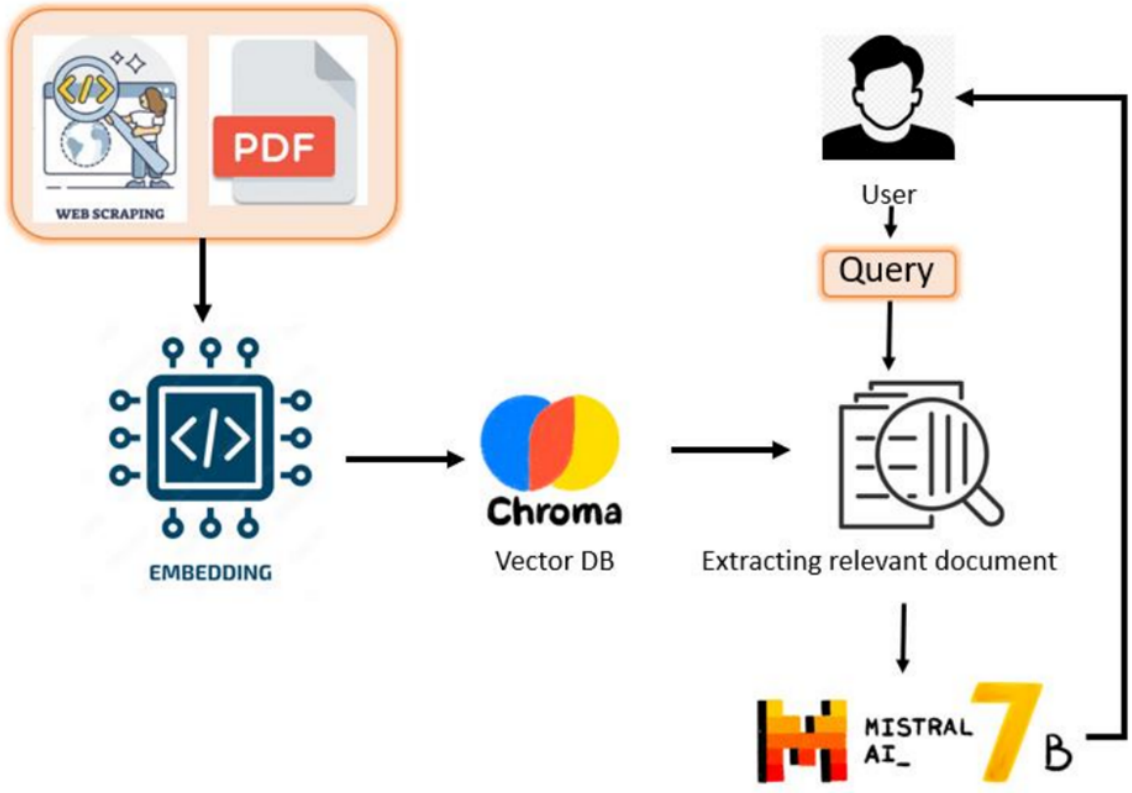
Current Challenges

- Navigating the vast expanse of the internet can be quite the challenge when it comes to finding relevant and accurate information in a timely manner.
- Many users struggle to refine their search queries for more accurate results.
- In addition, current search engines often fail to offer a complete context to improve the user's comprehension of the retrieved data.

Solution

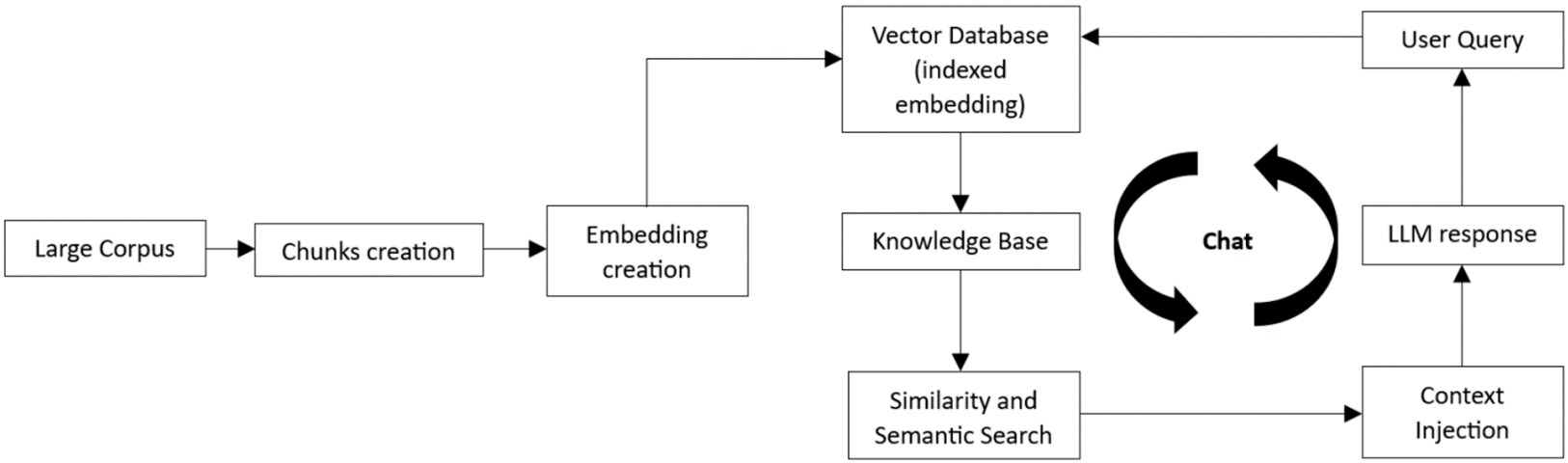
to develop a user-friendly web application that utilizes web scraping techniques to gather relevant data from online sources.

HIGH LEVEL TECHNICAL FLOW



for showcasing technology frame work used for project.

CUSTOM ARCHITECTURE



Custom Architecture for RAG system

FRAMEWORK + HARDWARE RESOURCE

FRAMEWORK

- Document Ingestion: Llama Index
- Creating Embedding: all-MiniLM-L6-v2 Architecture
- Vector Database: Chroma DB
- Generator: Mistral-7B-Instruct-v0.2
- Support framework: Langchain

HARDWARE RESOURCE

- GPU: P100 (16 GB)
- CPU: 29 GB
- DISK: 75 GB

LIVE DEMO FOR PROJECT

1. Internal Document Search RAG System
2. Web Search RAG System

Future Scope

As technology continues to evolve and new challenges emerge for future work, ranging from the exploration of new large language models (LLMs) to the advancement of RAG architectures:

- Integration of Multimodal Information
- Fine-Tuning Techniques and Optimization
- Addressing Privacy and Security Concerns

THANK YOU