Group name: Abdelrahman Ibrahim, Aya Noah

Project Title

Development and Implementation of a Retrieval-Augmented Generation (RAG) Based Assistant for Document-Based Question Answering

Overview

The purpose of this project is to design and implement an intelligent question-answering system based on Retrieval-Augmented Generation (RAG) to respond to user queries over a curated collection of textual resources, such as FAQs and technical documents. The system aims to deliver accurate and context-aware responses by grounding generation in retrieved evidence. This project was selected to gain practical, hands-on experience with retrieval-augmented generation systems, which are widely used to improve reliability and factual grounding in real-world generative AI applications.

Objectives and Learning Outcomes

This project seeks to demonstrate proficiency in embedding-based retrieval, vector database integration, prompt conditioning, and evaluation of generative systems. It aligns with the learning objectives of the Generative AI course by emphasizing modular system design, hallucination mitigation strategies, and the effective use of foundation models.

Background and Preparation

The student has prior experience with Python, fundamental NLP concepts, embeddings, and the basic use of large language models. This project focuses on applying these skills within a structured RAG (Retrieval-Augmented Generation) pipeline.

Methodology and Architecture

The proposed system will comprise:

(1) Document ingestion and preprocessing.
(2) Semantic chunking and embedding generation.
(3) Vector storage using FAISS or Chroma.
(4) Similarity-based retrieval for user queries.
(5) Response generation using a large language model conditioned on retrieved context.

Deliverables and Intended Users

The final delivery will be a functional prototype equipped with a user interface for querying document collections. Intended users include students and researchers interested in document-grounded question answering, as well as instructors seeking to demonstrate applied Generative AI techniques.

Scope and Constraints (30 Hours)

Included within the scope are small-scale datasets, retrieval pipelines, generation modules, and basic evaluation procedures. Excluded are large-scale deployment, multimodal inputs, fine-tuning model, and advanced production-level components.