Multi-Purpose AI Powered FAQ Chatbot

Minor Project Report

Department of Data Science & Computer Applications

B. Tech Data Science 6th Semester

Submitted By:

Name	Registration Number
Lionel Jerald Serrao	220968368

1. Identification of the Problem Statement

In today's regulatory-heavy environment, organisations must ensure employees and legal teams stay compliant with frequently updated policies, contracts, and legal documents. Extracting clear insights from dense legal PDFs, HR manuals, or government compliance documents is often slow and error prone.

This project introduces an AI-powered FAQ chatbot designed to simplify legal and policy comprehension through conversational, document-aware interactions. By combining document parsing, semantic vector search, and large language models (LLMs), the chatbot enables real-time, context-rich answers grounded in official legal and organizational documents.

2. Objectives

- Extract clean text from legal PDFs, company policy manuals, or scraped government websites using PyPDF2 and BeautifulSoup.

- Preprocess and segment text using LangChain’s RecursiveCharacterTextSplitter to maintain context for embeddings.

- Generate semantic embeddings with Google Generative AI and index them in a FAISS vector store for fast and accurate retrieval.

- Use Google’s Gemini 1.5 Flash or HuggingFace’s Zephyr-7B-β to generate legally coherent, context-sensitive answers.

- Provide a Streamlit-based interface for users to upload documents, ask queries, receive cited answers, and refine chatbot performance through feedback.

3. Literature Review

Traditional legal search tools rely on keyword-based matching and Boolean queries, which often miss semantic nuances critical in legal and compliance contexts. Retrieval-Augmented Generation (RAG) overcomes these limitations by coupling dense vector retrieval with the generative power of LLMs.

FAISS enables fast and scalable semantic search across large sets of policy or legal texts. LangChain provides an extensible framework to integrate LLMs into the retrieval pipeline, ensuring modular, interpretable processing. Google’s Gemini models and Zephyr-7B-β deliver fluent, legally sensitive responses, enhancing comprehension and reducing dependency on manual document reviews.