HelpMate AI: A Retrieval-Augmented Generation (RAG) System

1. Introduction

The HelpMate AI project focuses on building and improving a Retrieval-Augmented Generation (RAG) system using frameworks like LlamaIndex and LangChain. The goal is to efficiently retrieve and generate responses from structured and unstructured policy documents to assist users in querying insurance policies.

2. Objectives

- Develop an AI-powered system that retrieves relevant information from insurance documents.
- Implement a structured pipeline for document ingestion, indexing, and querying.
- Enhance response accuracy using optimized embedding models and indexing strategies.
- Provide an interactive chatbot for user queries.
- Ensure scalability and flexibility for different document types and queries.
- Develop a feedback loop to continuously refine model responses based on real-world user interactions.

3. System Design

Overall Architecture:
- Document Loader, Data Parsing, Vector Store Index, Query Engine, Response Pipeline, Evaluation Pipeline, User Interface Layer.

Layers of the Project:
- Data Ingestion Layer, Indexing Layer, Query Processing Layer, Response Generation Layer, Evaluation Layer, User Interaction Layer.

4. Implementation Details

Technology Stack:
- Frameworks & Libraries: LlamaIndex, LangChain, OpenAI API, Pandas, IPython.
- Storage: Google Drive for document storage.
- Development Environment: Google Colab.
- Vector Database: ChromaDB.

Dataset Source:

- Insurance policy documents obtained from publicly available datasets on Kaggle and financial regulatory sources.

5. Challenges Faced

- Data Formatting Issues, Query Optimization, Handling Large Documents, Ensuring Accurate Responses, Scalability Concerns, Latency Issues.

6. Lessons Learned

- Proper document formatting improves accuracy.
- Chunking and embedding tuning enhance retrieval.
- User feedback is crucial for refining responses.
- Hybrid indexing strategies yield better results.

7. Future Improvements

- Implement customized embeddings, agent-based RAG, alternative document loaders, real-time monitoring, cloud-based deployment, and multi-language support.

8. Conclusion

HelpMate AI successfully implemented a robust RAG system for insurance policy document retrieval. Leveraging LlamaIndex, OpenAI, and ChromaDB, it enables users to query policy details efficiently. Future refinements will enhance response quality and expand use cases, making it a scalable and intelligent document retrieval system.