HelpMateAI – Revolutionising Insurance Document Retrieval

Project Report

Upendra Kumar | 31st December 2024

Table of Contents

1. Project Overview
2. Problem Statement
3. Key Objectives
4. System Architecture and Design
5. Technology Stack
6. Implementation Details
7. Challenges Faced
8. Lessons Learned
9. Future Enhancements
10. Conclusion

1. Project Overview

HelpMateAI is an innovative tool designed to simplify the process of retrieving information from large insurance documents. Traditionally, users faced difficulties locating specific details within documents that span 100+ pages. HelpMateAI leverages advanced AI models to enable precise and quick search capabilities, delivering results with citations for further reading.

The solution focuses on **saving time**, **enhancing accuracy**, and improving the overall user experience when accessing critical information.

2. Problem Statement

Discuss the challenges faced by users:

1. Long and complex documents (100+ pages).
2. Manual effort required to find specific information.
3. Time wasted searching multiple documents.
4. Highlight the need for an AI-powered solution.

3. Objectives

1.Enable precise and quick information retrieval.

2.Provide citations with document name and page number.

3.Enhance user experience through intuitive UI.

4.Integrate advanced AI and NLP models.

5.Build scalability for future upgrades like memory and session handling.

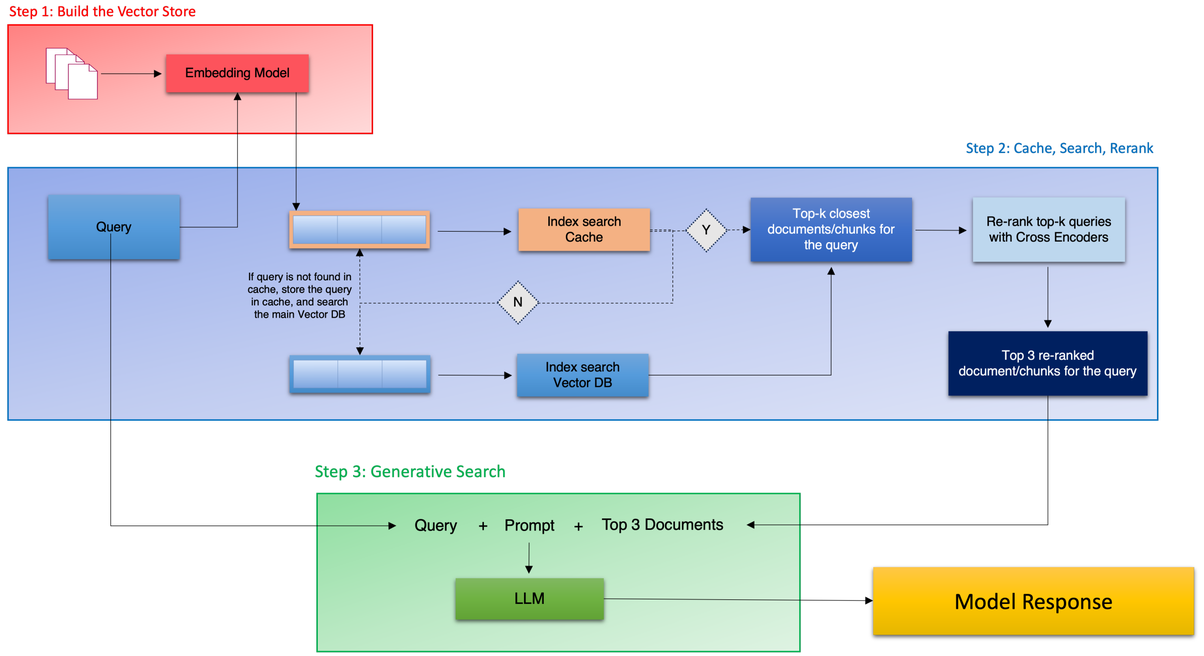
4. Design and Architecture

The chatbot was designed in three stages:

Step 1: Build the vector store using an embedding model.

Step 2: Query caching, search, and re-ranking.

Step 3: Generative search using LLMs.



5.Technology Stack

1.Programming Language: Python

2.Libraries & Frameworks: PDFPlumber, ChromaDB, OpenAI, Grok API, pandas

3.Models: Closed-source and open-source LLMs

4.Deployment: Web-based UI for interaction

6.Implementation Details

Building Vector Store:

Embedded the document data using advanced embedding models.

Caching and Searching:

Integrated a cache to reduce repeated searches for the same query.

Re-Ranking:

Used Cross-Encoders for refining top-K document chunks.

Generative Search:

Integrated an LLM for generating detailed responses with citations.

7.Challenges Faced

1.Managing large document datasets and ensuring efficient indexing.
2.Fine-tuning LLMs for accurate and relevant responses.
3.Implementing a robust caching mechanism to improve performance.
4.Balancing computational costs with system efficiency.

8.Lessons Learned

1.Importance of a structured workflow for complex AI projects.
2.Value of caching and re-ranking in improving query accuracy.
3.Balancing user-friendly design with backend complexity.
4.Need for scalability and forward planning for future enhancements.

9.Future Scope

1.Store user chat history.

2.Enable user-level sessions.

3.Add memory to LLMs for contextual conversations.

4.Extend support to other document formats like Word, Excel, etc.

10.Conclusion

HelpMateAI demonstrates how AI-powered solutions can transform document retrieval processes.

By addressing the challenges of lengthy and complex insurance documents, HelpMateAI saves time, enhances user experience, and ensures accurate, detailed responses with citations.

THE END