|  |
| --- |
| **Project Report**  **Title: Mr. HelpMate AI: Semantic Search and Question Answering for Insurance Documents**  **Prepared By:** **Amol Pramod Suryavanshi** |

Table of Contents

[Overview 3](#_Toc173964302)

[1. Introduction 3](#_Toc173964303)

[2. Scope and Requirements 4](#_Toc173964304)

[3. System Diagram 5](#_Toc173964305)

[3.1 Embedding Layer 6](#_Toc173964306)

[3.2 Search and re-rank layer 6](#_Toc173964307)

[3.3 Generation Layer 7](#_Toc173964308)

[4. Challenges & Lessons Learned 8](#_Toc173964309)

[5. Conclusion 8](#_Toc173964310)

## Overview

* **Problem:** Insurance policyholders and agents often struggle to locate specific information within complex insurance documents. Traditional keyword searches fail to capture the nuances of insurance terminology and the context of the user's query.
* **Solution:** Mr. HelpMate AI addresses this challenge by combining semantic search with powerful language models to deliver accurate and context-aware answers to user queries from within insurance documents.
* **Core Features:**
  + PDF processing and text extraction
  + Semantic search to retrieve relevant insurance document sections
  + AI-powered question answering to provide direct answers and citations.
* **Benefits:** Streamlines information retrieval, saves time for policyholders and agents, and potentially enhances decision-making based on policy terms.

## 1. Introduction

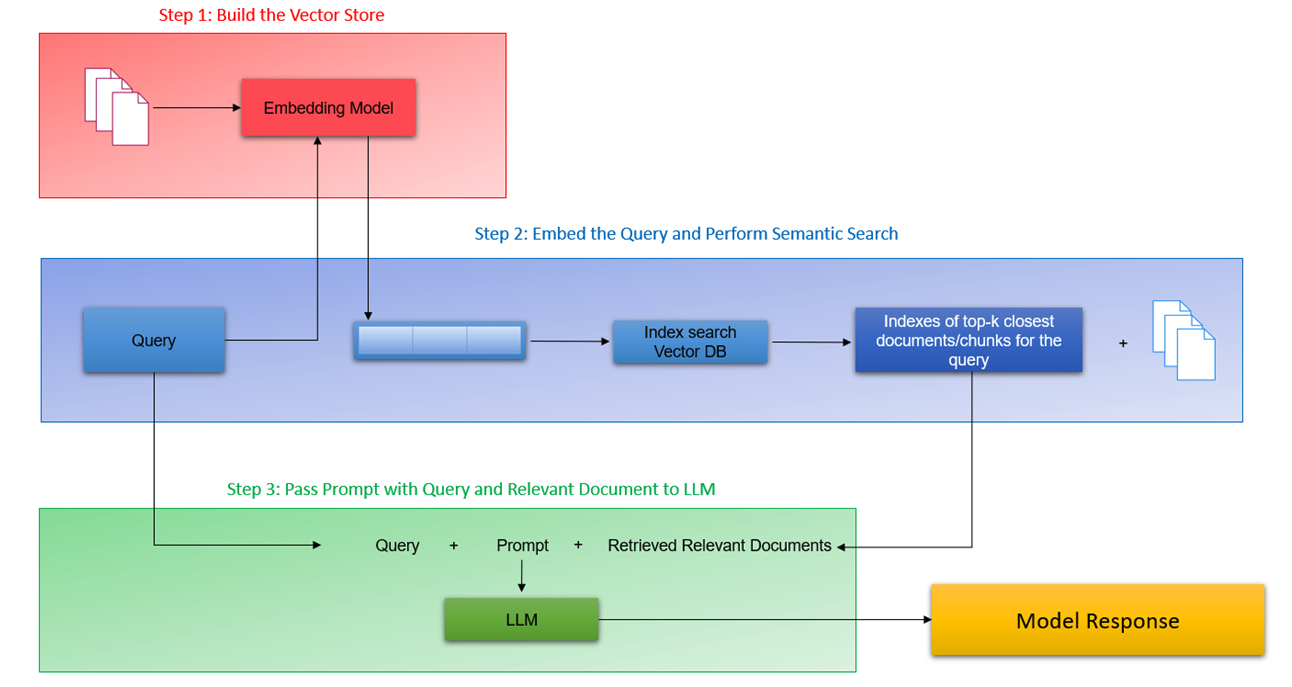
* **Problem Statement:** Explain the complexities of insurance documents (lengthy, jargon-heavy, varying structures) and the difficulties users face in finding the exact information they need.
* **Objectives**
  + Design a system to accurately and efficiently extract text from insurance PDFs.
  + Implement semantic search to enhance the relevance of results compared to keyword-based methods.
  + Develop a question-answering (QA) component that leverages natural language processing to pinpoint answers and provide citations within insurance documents.
* **Target Audience:** Insurance policyholders, insurance agents, underwriters, or potentially claims adjusters.

## 2. Scope and Requirements

* **Features**
  + **PDF Preprocessing:** (pdfplumber library)
  + **Semantic Search:** (ChromaDB, OpenAI embeddings)
  + **Question Answering:** (OpenAI language models)
  + **Potential Additional Features:** Table extraction and reformatting, answer summarization
* **Technology Stack**
  + **Python**
  + **pdfplumber:** PDF parsing
  + **OpenAI:** Text embeddings, language model access
  + **ChromaDB:** Semantic search database with embedding storage
  + **Sentence Transformers:** Cross-encoder for result re-ranking (optional)
  + **Other Libraries:** tiktoken for tokenization, pandas for JSON parsing, pandasfor DataFrame manipulations

## 3. System Diagram

The system was architected using a tri-layer RAG (Retrieval-Augmented Generation) pipeline



### 3.1 Embedding Layer

**A diagram of a file

Description automatically generated**

* + Extracted text and tables from PDFs were formatted into data frames.
  + OpenAI's text-embedding model produced vector representations, which were stored in Chroma DB.
  + Documents underwent processing and chunking to enhance retrieval efficiency.

### 3.2 Search and re-rank layer

**A diagram of a model

Description automatically generated**

* + Semantic searches were conducted on queries to fetch the top K relevant documents or chunks.
  + A re-ranking module used cross-encoders to enhance search precision.

### 3.3 Generation Layer

**A diagram of a search engine

Description automatically generated**

* + A carefully crafted prompt, which incorporated the original query along with the retrieved documents, was employed by the language model to produce coherent responses.

## 4. Challenges & Lessons Learned

* **Challenges**
  + **Complex PDF Structures:** Handling diverse PDF formats.
  + **Insurance-Specific Terminology:** Adapting models to insurance vocabulary.
  + **Answer Granularity:** Finding the ideal balance in result length.
  + **API Costs:** Optimizing usage for cost-effectiveness.
* **Lessons Learned**
  + **Thorough PDF Analysis:** Pays off in robust extraction logic.
  + **Domain Adaptation:** Crucial for language model accuracy.
  + **Hybrid Retrieval:** Can enhance certain query types.
  + **Cost-Effective Architecture:** Exploring a mix of APIs and local models.

## 5. Conclusion

* **Summary**
  + Mr. HelpMate AI successfully demonstrates the potential of semantic search and AI-powered question answering to transform how users navigate insurance documents.
  + The project highlighted the value of PDF preprocessing, embedding-based search, and language model fine-tuning.
* **Future Work**
  + **Enhanced User Interface**
  + **Multi-Document QA**
  + **Policy Comparison**