Helpmate AI - Lab Report Evaluator Documentation

Project Goals:

The goal of this project is to create a robust generative search system for pathological lab reports in PDF format. The system aims to effectively answer user queries related to the lab reports by experimenting with various components, including chunking strategies, embedding models, re-rankers, and generation prompts.

AI Model Used:

The system incorporates three main layers:

1. Embedding Layer: Converts lab report tables into embeddings and stores them in a vector database (Chroma DB).
2. Search Layer: Accepts user queries, checks cache for previous queries, retrieves embeddings, and ranks results using OpenAI encoders.
3. Generation Layer: Utilises OpenAI API for chat completion to generate responses based on the top-ranked results.

Data Sources:

The primary data source for this project is pathological lab reports in PDF format. The app extracts measured data such as vitals and counts from these reports, converts it into a dataframe, and stores each table in a vector database.

Key Design Decisions:

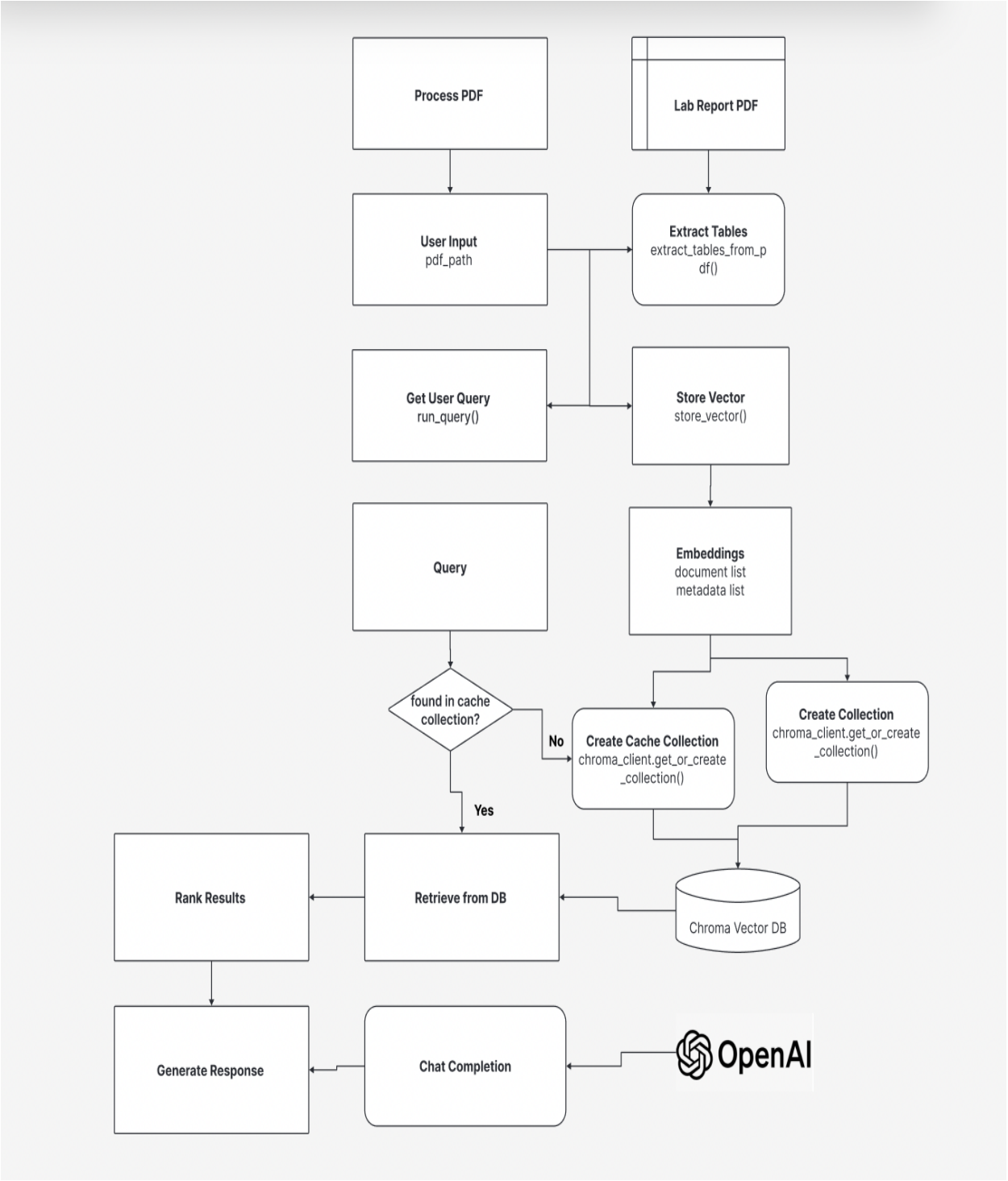
1. Embedding Strategy: Tables in lab reports are converted into embeddings and stored in Chroma DB for efficient retrieval.

2. Cache Mechanism: The system checks a cache collection for previous queries; if not found, it retrieves data from the main collection, stores it in the cache, and associates it with the query for future use.
3. Ranking Mechanism: OpenAI encoders are used to rank the retrieved embeddings to provide the top 3 results.
4. Generation Strategy: The top-ranked results are passed to the generation layer, which uses the OpenAI API for chat completion with the user query to generate responses.

Challenges Encountered:

1. PDF Parsing: Extracting structured data from PDFs, especially pathological lab reports, posed challenges due to varying formats.
2. Cache Optimization: Balancing cache storage for efficient retrieval and avoiding redundancy requires careful optimization.
3. Response Generation: Achieving coherent and contextually relevant responses using the OpenAI API for chat completion presented challenges in user query contextualization.

Working Diagram:



Helpmate AI - Lab report evaluator