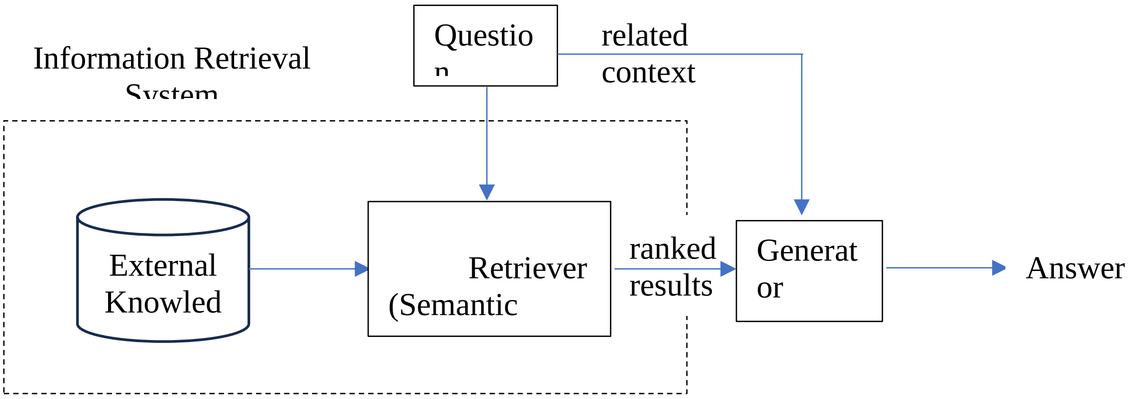
Project Background

Helpmate AI project is an advanced implementation of **Retrieval-Augmented Generation (RAG)** for **intelligent document search and analysis**. It combines fixed chunk-based text extraction, **vector embedding**, and **generative search** and enables users to retrieve highly relevant information .



Approach:

The Embedding Layer:
The PDF document is effectively processed, cleaned, and chunked for the embeddings.
We have implemented, Fixed sized chunking in which data is divided into equal-sized chunks of 300 which gave optimal results since larger chunk sizes provide appropriate context to the LLM during the generation layer. The chunked data is then stored in Chroma DB, with embeddings generated using OpenAI's text-embedding-ada-002.

The Search Layer:
The system supports semantic search across chunking collections, using a cache to store and quickly retrieve frequently queried results. When a query is repeated, it retrieves data from the cache, optimizing performance.
The ms-marco-MiniLM-L-6-v2 cross encoder is used to re-rank search results and provide top 3 outputs from the pdf document.

The Generation Layer:
Zero-shot prompting technique is used to get better results and processed via the Chat Completions API for generative answers.