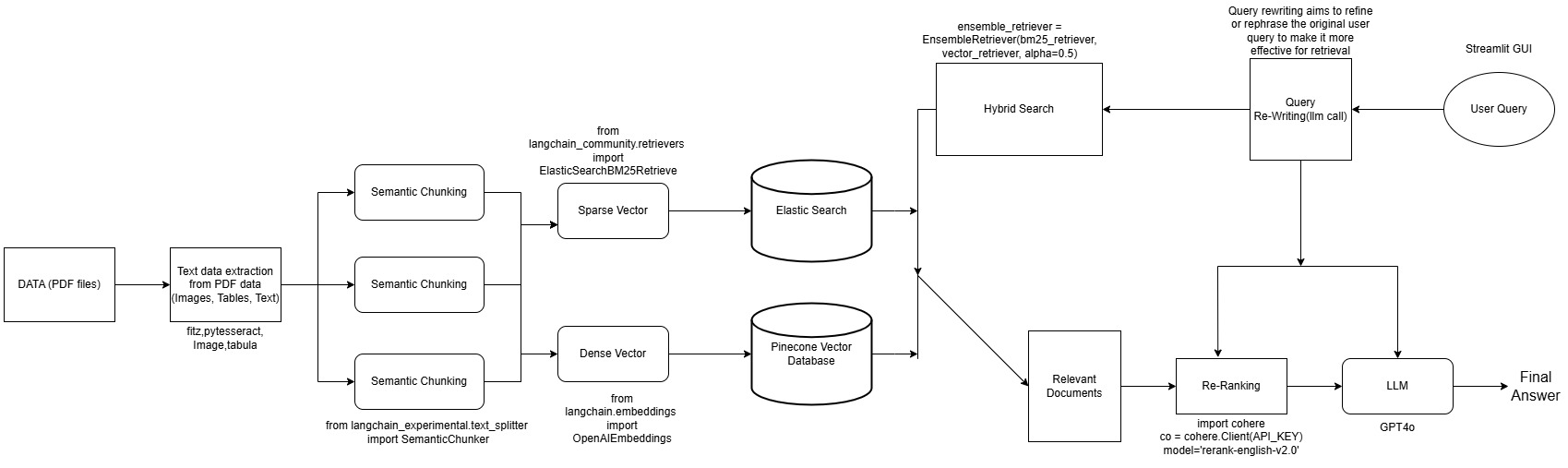
**Advnaced RAG chat with PDF**



**Problem Statement:**

We have SOP [Standard Operating procedure], it contains steps, process and procedure for troubleshooting issues/problems occurring in servers. Supporting engineers need to not search for specific SOP or specific Steps or specific process. Seamlessly they can interact with RAG UI for retrieving whatever required details.

**Challenges:**

1. Initially we started using ollma open-source models like llama3, llama3.2, mistral, gemma2 etc. Facing issue with accurate output and hallucinations, understanding exact context of prompts, output formats. Once we switched paid llm model gpt-4o , output and results of llm prompts is outstanding and reduced hallucinations. Whatever we paid amount is justified.
2. Earlier we have implemented text data extraction only from pdf files. Later started receiving different pdf files, it contains images, tables. which require multiple libraries for accurate data extraction. Again, started implemented and added different libraries for extracting data from images, tables.
3. Python library dependency mismatch with other python libraries
4. Handling high-dimensional vector databases with millions of embeddings is big challenging with Speed and latency issues with retrieval accuracy.
5. Which tools or libraries did you use for text extraction from PDFs, and why?

Extract normal text from PDF: fitz

Extract tables data from PDF: tabula

Extract images from PDF: fitz

Extract text from Images: pytesseract

1. What algorithm or approach does SemanticChunker use for dividing text into meaningful chunks?

We are doing vector based chunking technique and using openai embeddings for dividing input text into smaller chunks.

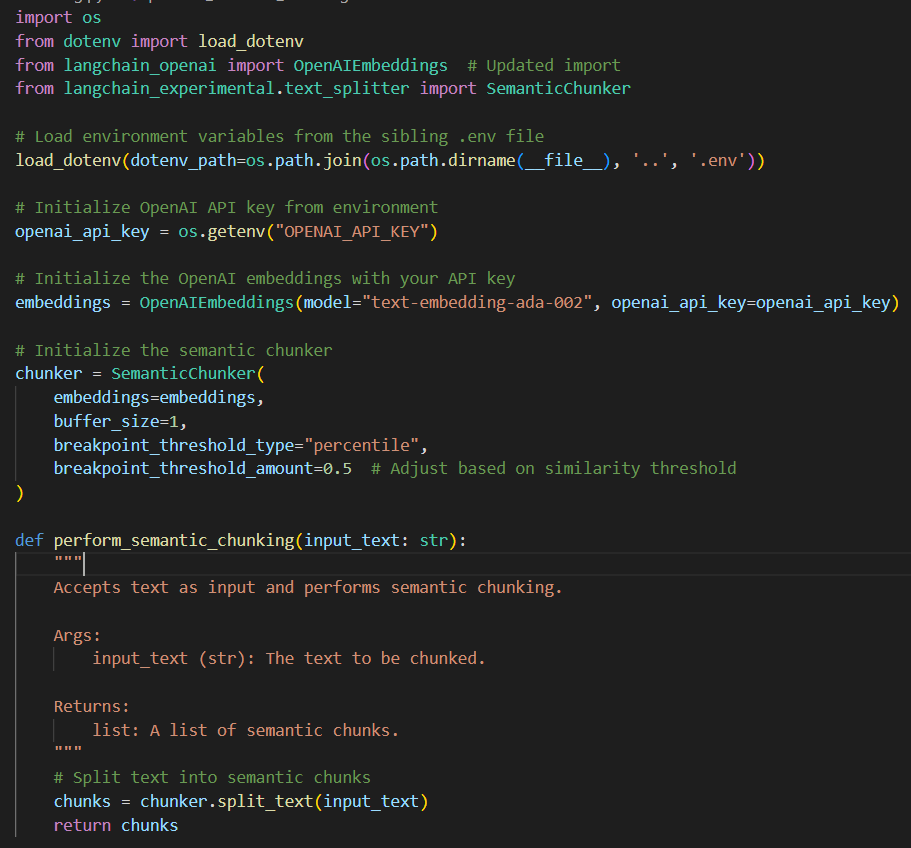
**Similarity-Based Splitting:**

The SemanticChunker uses these embeddings to calculate the cosine similarity between adjacent pieces of text.

A similarity threshold (breakpoint\_threshold\_amount=0.5) determines whether the next segment should be part of the current chunk or start a new one:

If the similarity is above the threshold, the text is grouped into the same chunk.

If the similarity is below the threshold, a new chunk is started.



1. What are other chunking techniques you tried?
2. Fixed size chunking
3. Sliding window chunking
4. Sentence chunking
5. Paragraph chunking
6. What does the alpha parameter in EnsembleRetriever signify, and how did you decide its value?

The alpha parameter in hybrid search plays a crucial role in balancing the contributions of semantic (vector-based) and lexical (keyword-based) search methods.

What is alpha?

alpha is a weight parameter that determines how much influence each search method has in the final ranking.

The hybrid score is calculated as:

Hybrid Score= 𝛼 × Semantic Score +(1−𝛼) ×Lexical Score

Hybrid Score=α×Semantic Score+(1−α)×Lexical Score

1. Higher alpha gives more weight to the semantic search
2. Lower alpha gives more weight to keyword search
3. What other embedding techniques tried apart from OpenAI Embeddings?

OpenAI Embeddings model: text-embedding-ada-002

Ollama embeddings model: nomic-embed-text  
  
Sentence Transformer embeddings: all-MiniLM-L6-v2

1. What is difference between Sparse vector and dense vector?

**Sparse vectors** store only non-zero values and their corresponding indices. Most of the elements are assumed to be zero, and only the non-zero values are explicitly represented. Requires less memory compared to dense vectors.

It’s used for keyword search.

**Dense vector** typically contains a value for every dimension, and most of these values are non-zero. Requires more memory for storing all values. it’s used for semantic search.

1. How BM25 works?
2. Calculate Term Frequency (TF)
3. Calculate Inverse Document Frequency (IDF)
4. Document Length Normalization

There are two parameters in BM25, k1 and b. k1 controls term frequency saturation and b parameters normalize document length.

1. How BM25 Address TFIDF Issue?

In TFIDF, the importance of word increases proportionally to the number of times that word appears in the document.

BY introducing two parameters in BM25, k1 and b. k1 controls term frequency saturation and b parameters normalize document length.

1. How reranking works and which reranking technique used?

Need to pass user query and relevant documents to cohere AI re-ranking method.

A white and black text

Description automatically generated

A screen shot of a computer code

Description automatically generated

1. What is use two db’s for vector stotrage?

We have used elasticsaerch for keyword search and pinecone vector db for sematic search . It makes robust pipeline and good performance during retrieval part.

1. How are you evaluating RAG models?

Reference code link: <https://github.com/Coding-Crashkurse/RAG-Evaluation-with-Ragas/blob/main/ragas.ipynb>

1. We are using RAGAS python library for evaluating rag model
2. Preparing automatic way of evaluation dataset using ragas.testset.generator
3. Dataset should contain following structure
4. **Question**
5. **Contexts**
6. **Answer**
7. **Ground Truth**
8. Metric selection for rag evaluation
9. **Faithfulness**: Evaluates how accurately the generated answer reflects the information in the provided context.
10. **Answer Relevancy**: Assesses the relevance of the generated answer to the user's question.
11. **Context Relevancy**: Measures the pertinence of the retrieved contexts to the question.
12. **Context Recall**: Determines the extent to which the retrieved contexts cover the necessary information to answer the question.