**Help Mate AI- RAG Project**

Prepared by – Arun Kumar Singh ([asingh.iitr@gmail.com](mailto:asingh.iitr@gmail.com)) Date: 31-Mar-2024

**Goal:**

Build a robust generative search system capable of effectively and accurately answering questions from a policy document, by using RAG (Retrieval augmented generation) concept.

**Data Source:** Policy document present in PDF form as below:

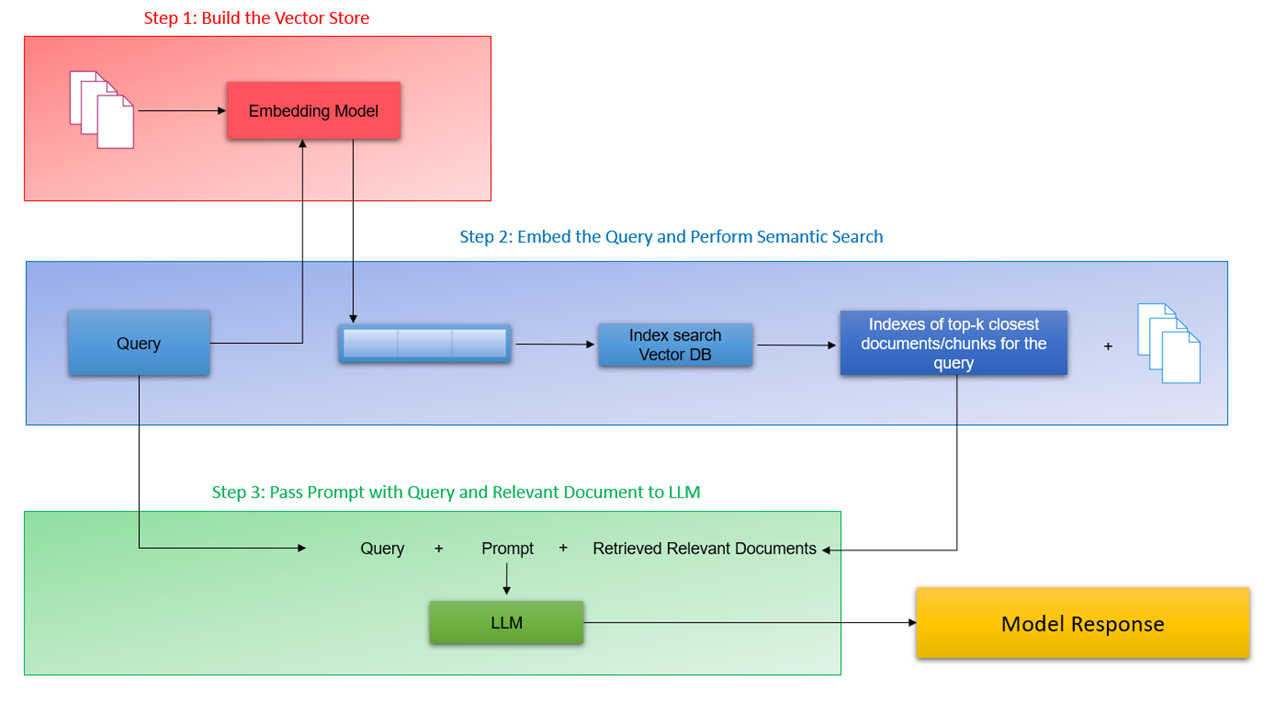
<https://github.com/arunksinghbuee/helpmate-ai/blob/main/insurance-document/Principal-Sample-Life-Insurance-Policy.pdf>

**Collab Notebook PDF:** <https://github.com/arunksinghbuee/helpmate-ai/blob/main/Mr.HelpMate%20AI%20Project.pdf>

**System Design:**

RAG pipeline mainly consists of 3 layers:

1. Embedding Layer
2. Search and Rank Layer
3. Generation Layer

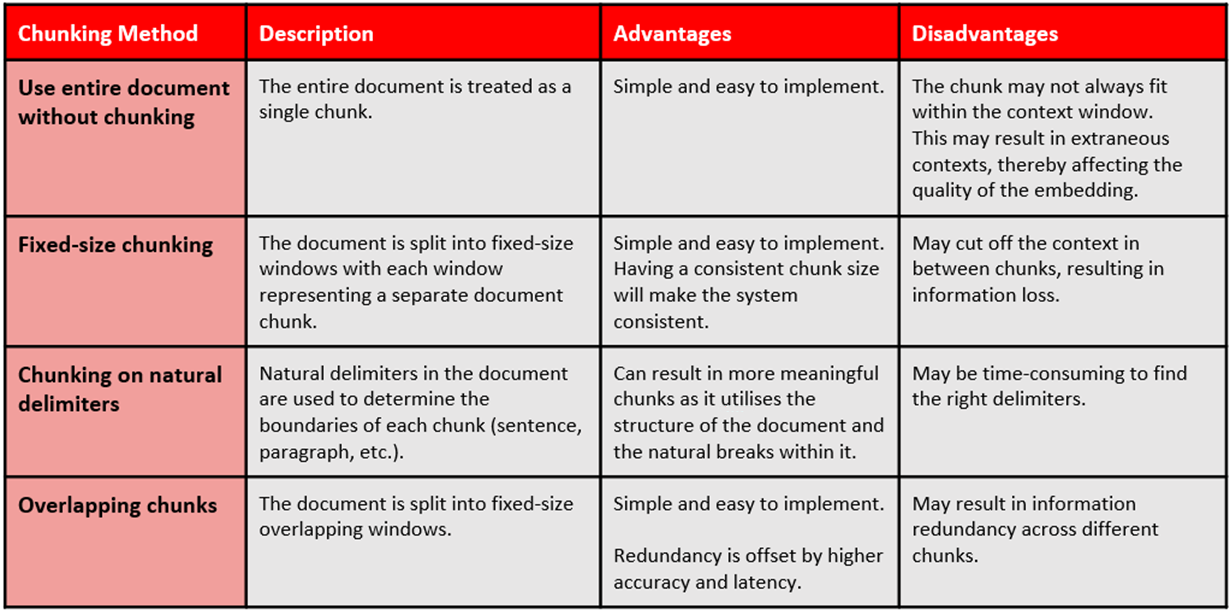


**Embedding Layer**

This layer is a crucial component of RAG models, serving as the first layer. It contains an embedding model trained on extensive text and code datasets to learn word and phrase relationships. This layer enables systems to understand text meaning and its semantic relation to queries, crucial for tasks like question answering, summarization, and machine translation. It generates embeddings for text corpora, aiding the RAG model in comprehending queries and generating relevant responses.

Embedding generation code -> Refer <https://github.com/arunksinghbuee/helpmate-ai/>

There are multiple types of chunking methods which can be used as per nature of data.



By looking at provided document, most of the similar context data is present on each page and it is further divided into sections & paragraphs on each page.

Word length in each page varies up to maximum of 418 words which is suitable for creation of embedding using **text-embedding-ada-002** OpenAI embedding model.

A screenshot of a computer

Description automatically generated

A screenshot of a web page

Description automatically generated

**Search & Rank Layer**

This layer ensures that the retrieved text is accurate, relevant, and contextually appropriate. This layer consists of two main components: a search component that retrieves relevant documents using semantic similarity techniques, and a re-rank component that ranks the retrieved documents based on relevance, popularity, and freshness. The search and re-rank layer is crucial for tasks like question answering, summarization, and machine translation, as it helps in retrieving and ranking relevant information efficiently. It plays a key role in enhancing the performance of AI tasks and is an integral part of the project's RAG models.

Embeddings are generated using OpenAI and stored in ChromaDB. Embeddings of search query

is generated using the same model and top 3 most matching results are obtained based on its

cosine similarity.



**Caching mechanism** is also implemented to optimize search performance. Search query embedding and search result metadata is stored in Vector DB (ChromaDB in this case) whenever any search is performed.

If new search query is found similar to already existing query in cache embeddings, depending on variance factor defined, then corresponding cached search results are used, rather than searching on whole document imbedding’s.



**Re-ranking:**

Re-ranking the results obtained from semantic search can sometime significantly improve the relevance of the retrieved results. This is often done by passing the query paired with each of the retrieved responses into a cross-encoder to score the relevance of the response w.r.t. the query.

There are several cross-encoders available. We are using **cross-encoder/ms-marco-MiniLM-L-6-v2** from **sentence\_transformers**

A diagram of a cross-engine

Description automatically generated



**Semantic Search Result Code:**

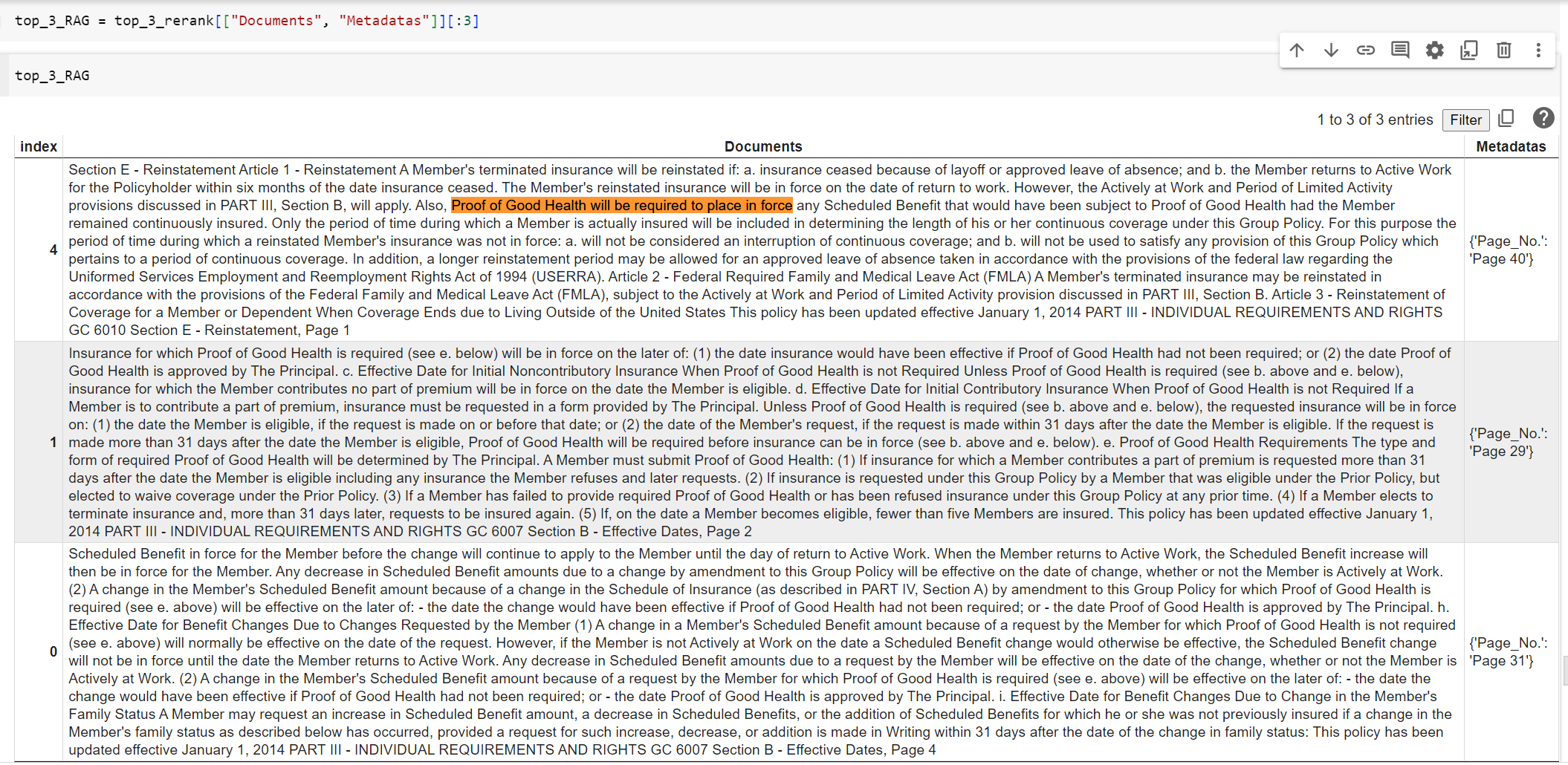


**Search Queries**

1. **Que:** **What are the requirements for placing in force any Scheduled benefit that would have been subject to Proof of Good Health has the member remained continuously insured?**

A screenshot of a computer

Description automatically generated



Answer is clearly found in first result which is mentioned on page no 40 of the document.

1. **Que: How is the peroid of time during which a reinstated Member's insurance was not in force treated for the purpose of determining the length of continuous coverage under the Group Policy?**

**A screenshot of a computer

Description automatically generated**

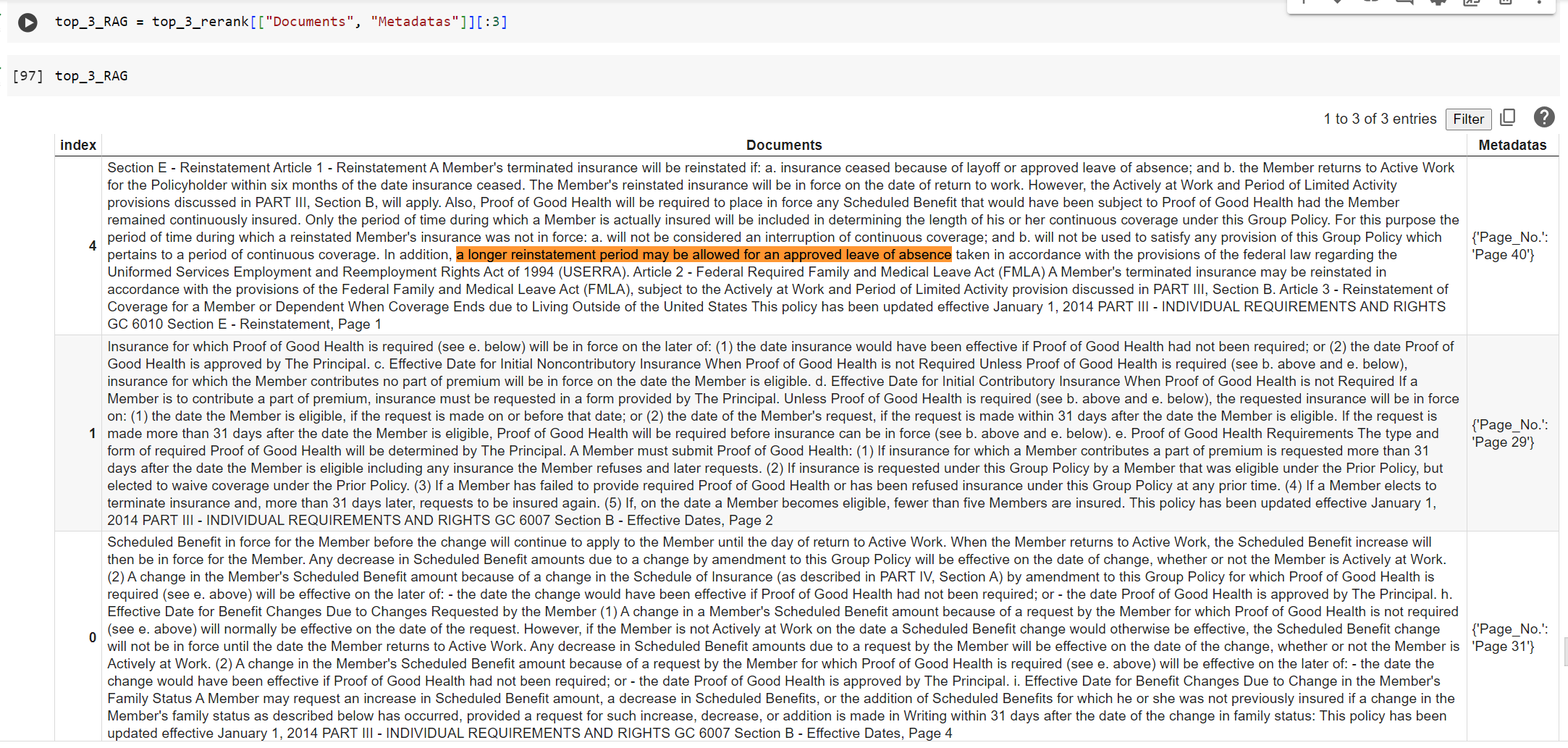
****

Answer is clearly found in first result which is mentioned on page no 40 of the document.

1. **What provisions may allow for a longer reinstatement period for an approved leave of absence taken in accordance with the Uniformed Services Employment and Reemployment Rights Act of 1994 (USERRA)?**

**A screenshot of a computer

Description automatically generated**

****

Answer is clearly found in first result which is mentioned on page no 40 of the document.

**Generation Layer**

This is the final stage of the RAG model, featuring a large language model trained on extensive text and code datasets. It enables the model to produce new text in response to user queries. This layer synthesizes retrieved information, shaping it into coherent and contextually relevant responses. Crucial for tasks like question answering, summarization, machine translation, and generative search, the generation layer excels in providing context and natural language capabilities for such searches.

Find below prompt used. This also mentioned **Few Shots Prompts** for better results.



**Query Outputs:**

**Query 1: What are the requirements for placing in force any Scheduled benefit that would have been subject to Proof of Good Health has the member remained continuously insured?**

**A close-up of a white background

Description automatically generated**

****

**Query 2: How is the period of time during which a reinstated Member's insurance was not in force treated for the purpose of determining the length of continuous coverage under the Group Policy?**

**A close up of a text

Description automatically generated**

****

**Query 3: What provisions may allow for a longer reinstatement period for an approved leave of absence taken in accordance with the Uniformed Services Employment and Reemployment Rights Act of 1994 (USERRA)?**

**A close up of a logo

Description automatically generated**

****