Building Effective Search Systems: Mr.HelpMateAI

1. Background

The project focuses on developing an advanced and efficient search system, Helpmate AI, designed to process and extract information from a comprehensive life insurance policy document using Retrieval-Augmented Generation (RAG) techniques.

2. Problem Statement

The primary objective of this project is to build a robust generative search system capable of providing accurate and contextually relevant answers to user queries based on a single, long life insurance policy document.

3. Dataset

The project uses a single, long-form life insurance policy document as the primary data source. The policy document can be found here.

4. Approach

The architecture of the search system is built around three core layers, each playing a vital role in the system's efficiency and accuracy. Various strategies and experiments are implemented to optimize performance at each layer.

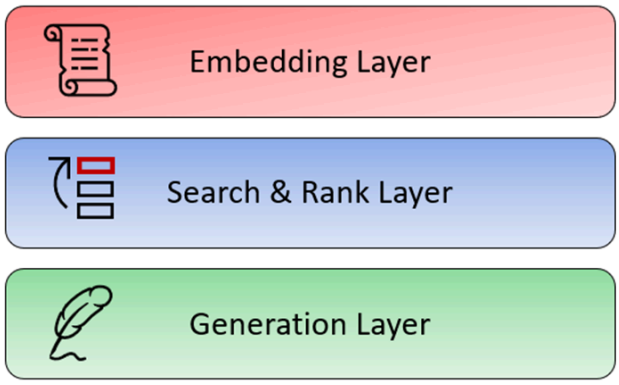
1. **The Embedding Layer:**

The PDF document is processed, cleaned, and divided into meaningful chunks to prepare it for embedding. The choice of chunking strategy significantly impacts the quality of the search results. Embeddings are generated using models like OpenAI's embedding models or SentenceTransformers from the HuggingFace library. Experiments are conducted to evaluate and select the most effective model.

2. The Search Layer:

At least three test queries are designed based on the document's content to evaluate the system's performance. Queries are embedded and matched against the document's ChromaDB vector database. A caching mechanism is implemented to improve search efficiency. Retrieved results are re-ranked using cross-encoding models from the HuggingFace library to enhance relevance and accuracy.

3. The Generation Layer:
In the generation layer, the final prompt that we design ensures that the information passed to the language model is clear and exhaustive. We may also choose to provide some few-shot examples in an attempt to improve the LLM output.



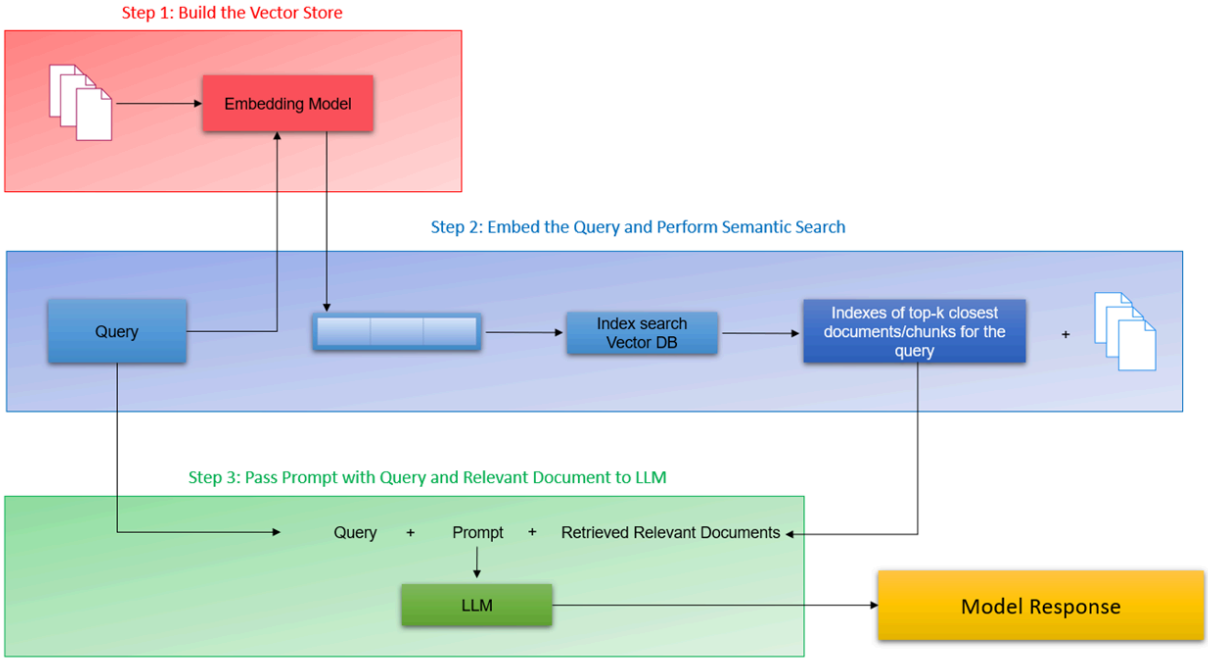
5. Implementation Details:

- **Reading & Processing PDF File:** We used pdfplumber to read and process the PDF files. pdfplumber allows for better parsing of the PDF file as it can read various elements of the PDF apart from the plain text, such as tables, images, etc. It also offers wide functionalities and visual debugging features to help with advanced preprocessing as well.
- **Document Chunking:** The document contains several pages and contains huge text, before generating the embeddings, we generated the chunks.
- **Generating Embeddings:** Generates embedding with SentenceTransformer with all-MiniLM-L6-v2 model.
- **Store Embeddings In ChromaDB:** Stored embedding in ChromaDB.
- **Semantic Search with Cache:** Added cache collection layer for embeddings.
- **Re-Ranking with a Cross Encoder:** Re-ranking of the results obtained from the semantic search significantly improved the relevance of the retrieved results. This is 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.

- **Retrieval Augmented Generation:** The final top search results are passed to a GPT 3.5 along with the user query and a well-engineered prompt, to generate a direct answer to the query along with citations.

6. System Architecture

The architecture comprises three integrated layers—embedding, search, and generation— working together to deliver high-accuracy responses with efficient data retrieval and contextual understanding.



7. Challenges:

- **Performance Scaling:** Address concerns about system performance with an increased number of documents or users by implementing vector databases and scaling up compute units.
- **Cache Storage:** Optimize the cache collection to efficiently store and retrieve queries and results.

8. Lessons Learned:

1. **Efficient Document Processing:** Processing PDFs efficiently is crucial; libraries like pdfplumber and suitable data structures for storage play a vital role.
2. **Semantic Search Optimization:** Fine-tune semantic search parameters and thresholds for optimal results.
3. **Cache Management:** Implement an effective cache management strategy to balance storage and retrieval efficiency.
4. The quality of embeddings and prompt design directly impacts system output accuracy.

9. Conclusion:

The project successfully implements a semantic search system with the RAG pipeline and cache layer. The objectives are met, and the challenges are overcome with lessons learned for future improvements. The system provides a scalable and efficient solution for document retrieval and information extraction.

10. Query Screenshots

***Query 1:* What documentation is required when filing a claim?**

Top_3_Results_from_Search_Layer_Query1

```
In [63]: top_3_RAG_q1 = top_3_rerank[["Documents", "Metadatas"]][:3]

In [64]: top_3_RAG_q1
```

Out[64]:

	Documents	Metadatas
2	of claim, will provide appropriate claim forms...	{'Chunk_No.': 4, 'Page_No.': 'Page 61'}
1	date of the loss. Proof required includes the ...	{'Chunk_No.': 4, 'Page_No.': 'Page 61'}
4	Section D - Claim Procedures Article 1 - Notic...	{'Chunk_No.': 4, 'Page_No.': 'Page 61'}

Generative_Layer_Result_Query1

In [81]:
```
# Generate the response - For Query 1
response = generate_response(query, top_3_RAG_q1)
print("Query 1: ","\n",query,"\n
# Print the response
print("\n".join(response))
```

```
Query 1:
 What documentation is required when filing a claim?

_____
_____

When filing a claim, the documentation typically required includes:
1. Claim forms filled out accurately and completely.
2. Proof of the date of loss.
3. Relevant supporting documents, such as receipts, invoices, or other evidence related to the claim.

Here is the information extracted from the policy document for reference:

| Documentation Required                                  |
|
|---------------------------------------------------------|-----------------------------------------------------------------
----------------------------------------|
| Claim Forms                                             | Provide appropriate claim forms with accurate information f
illed out.                                  |
| Proof of Date of Loss                                   | Proof required includes details verifying the date of loss.
|
| Supporting Documents                                    | Additional supporting documents like receipts, invoices, or
any other evidence related to the claim. |

Citations:
- Policy Name: Section D - Claim Procedures
- Page Number: Page 61
```

Query 2: What is the procedure for claiming benefits under the "Accelerated Benefits" provision, and what conditions qualify for this benefit?

Top_3_Results_from_Search_Layer_Query2

In [71]:
```
top_3_RAG_q2 = top_3_rerank_q2[["Documents", "Metadatas"]][:3]
top_3_RAG_q2
```

Out[71]:

	Documents	Metadatas
2	Principal. This proof must be provided in a fo...	{'Chunk_No.': 3, 'Page_No.': 'Page 13'}
1	disability that: (1) results from willful self...	{'Chunk_No.': 3, 'Page_No.': 'Page 51'}
7	PART IV - BENEFITS Section A - Member Life Ins...	{'Chunk_No.': 3, 'Page_No.': 'Page 46'}

Generative_Layer_Result_Query2

In [88]:
```
# Generate the response - For Query 2
response2 = generate_response(query2, top_3_RAG_q2)
print("Query 2: ","\n",query2,"\n_____________________________________________________________________________")
# Print the response
print("\n".join(response2))
```

```
Query 2:
 What is the procedure for claiming benefits under the "Accelerated Benefits" provision, and what conditions qualify
for this benefit?
_____________________________________________________________________________
_____________________________________________________________________________

To claim benefits under the "Accelerated Benefits" provision, the procedure typically involves the following steps:

1. Contact your insurance provider: Notify your insurance provider about your intention to claim accelerated benefits
under your policy.

2. Submit required documentation: Provide the necessary documents, which may include medical records, a physician's s
tatement, and any other relevant paperwork to support your claim.

3. Await evaluation: Your claim will be evaluated by the insurance company to determine eligibility for accelerated b
enefits.

4. Receive benefits: If your claim is approved, you will start receiving the accelerated benefits as per the terms of
your policy.

Conditions that may qualify for accelerated benefits can vary depending on the specific policy. Common qualifying con
ditions often include terminal illness, chronic illness, or permanent disability.

For more detailed information on the exact procedure and qualifying conditions for claiming accelerated benefits, ple
ase refer to the relevant policy documents:
```

Query 3: What provisions are included under the "Member Life Insurance - Coverage During Disability" section?

Top_3_Results_from_Search_Layer_Query3

In [78]:
```
top_3_RAG_q3 = top_3_rerank_q3[["Documents", "Metadatas"]][:3]
top_3_RAG_q3
```

Out[78]:

	Documents	Metadatas
1	Payment of benefits will be subject to the Ben...	{'Chunk_No.': 4, 'Page_No.': 'Page 49'}
0	Member Life Insurance or Coverage During Disab...	{'Chunk_No.': 4, 'Page_No.': 'Page 42'}
4	Section A - Member Life Insurance Schedule of ...	{'Chunk_No.': 2, 'Page_No.': 'Page 8'}

Generative_Layer_Result_Query3

In [87]:
```
# Generate the response - For Query 3

response3 = generate_response(query3, top_3_RAG_q3)
print("Query 3: ","\n",query3,"\n_____________________________________________________________________________")
# Print the response
print("\n".join(response3))
```

```
Query 3:
 What provisions are included under the "Member Life Insurance - Coverage During Disability" section?
_____________________________________________________________________________
_____________________________________________________________________________

The "Member Life Insurance - Coverage During Disability" section includes provisions related to the payment of benefi
ts subject to the Benefit Payment Provision. Specific details about the coverage, eligibility criteria, benefit amoun
ts, and any limitations during disability are outlined in this section.

Here is a simplified summary in a tabular format:

| Provision                | Details                                          |
|--------------------------|--------------------------------------------------|
| Payment of Benefits      | Subject to Benefit Payment Provision             |
| Coverage Details         | Eligibility criteria, benefit amounts            |
| Disability Limitations   | Any restrictions or limitations during disability|

### Citations:
- Policy Name: Member Life Insurance or Coverage During Disability
- Source Page: Page 42
```