Semantic Spotter Project- Build a RAG System

1. Project Goal

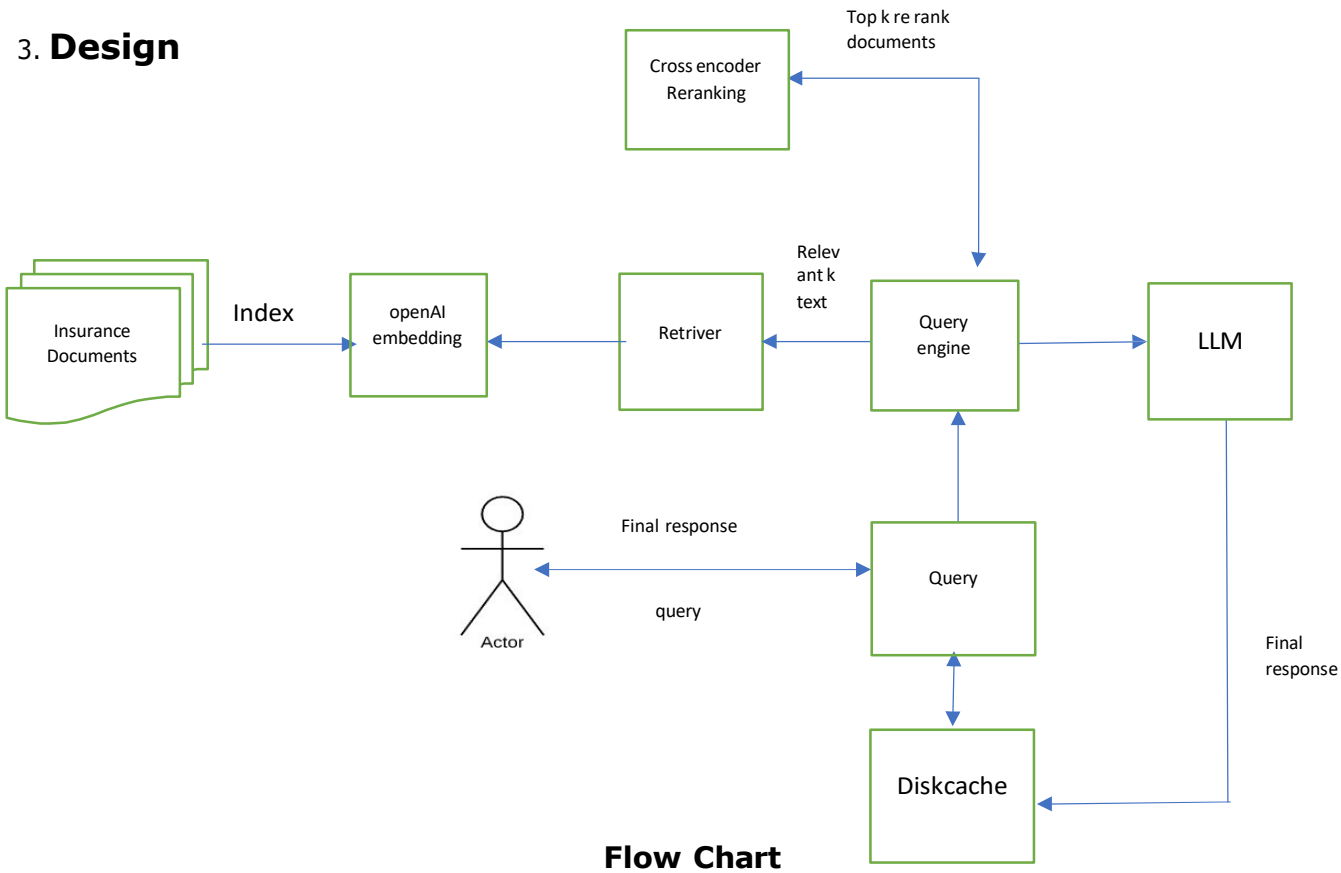
Build a project in the insurance domain. The goal of the project will be to build a robust generative search system capable of effectively and accurately answering questions from various policy documents. Using LlamaIndex to build the generative search application.

2. Data Source

Seven HDFC insurance documents in Pdf format provides inside a single folder.

a. HDFC-Life-Easy-Health-101N110V03-Policy-Bond-Single-Pay.pdf
b. HDFC-Life-Group-Poorna-Suraksha-101N137V02-Policy-Document.pdf
c. HDFC-Life-Group-Term-Life-Policy.pdf
d. HDFC-Life-Sampoorna-Jeevan-101N158V04-Policy-Document.pdf
e. HDFC-Life-Sanchay-Plus-Life-Long-Income-Option-101N134V19-Policy-Document.pdf
f. HDFC-Life-Smart-Pension-Plan-Policy-Document-Online.pdf
g. HDFC-Surgicare-Plan-101N043V01.pdf

3. Design



Flow Chart

Descriptions about the Architecture:

1. Documents: List of seven HDFC insurance documents provides inside a single folder.

2. Open API embedding: OpenAPI embedding as Vector DB for indexing insurance documents in the form of embedding.

3. Query Engine: We are using Query Engine Module of Llammaindex for performing semantic Search. Query Engine will use internally Retriever and SentenceTransformerRerank- model="cross-encoder/ms-marco-MiniLM-L-2-v2 retrieve top-k relevant nodes from embedding.

4. LLM: top k-documents along with user query will be passed to LLM to generate the accurate response.

5. Caching:" Caching is being used to improve the read operation. Recent similar search will be store in Caching and user query first will be served from Cache. If user query not found in cache, then query will be forwarded to query engine and then LLM to generate the response.

6. Meta data : Along with Response we are also returning docs reference and similarly score to improve the user confidence towards the implemented RAG system.

7. SentenceTransformerRerank- model="cross-encoder/ms-marco-MiniLM-L-2-v2 Is being used to rerank the query based on semantic score.

8. Evaluation- LLM-gpt4 is used for evaluation on matrices relevancy ,faithfulness and correctness.

4. Solution Strategy

- Build a solution which should solve the following requirements:

- Users would get responses from insurance policy knowledge base.

- If user want to perform a query system must be able to response to query accurately.

- If they want to refer to the original page from which the bot is responding, the bot should provide a citation as well.

5. Tools used

- LlamaIndex has been used due to its powerful query engine, fast data processing ,easier and faster implementation using fewer lines of code.

SimpleDirectoryReader is used to read the documents.

Vectorstoreindex is used to create index.

-SentenceTransformerRerank - model="cross-encoder/ms-marco-MiniLM-L-2-v2" is used to Rerank.

-Diskcache

- openAI API key

-LLM- gpt-4 for evaluation

6. Why LlamaIndex ?

LlamaIndex is an innovative data framework specially designed to support LLM-based RAG framework application development. It offers an advanced framework that empowers developers to integrate diverse data sources with large language models. LlamaIndex includes a variety of file formats, such as PDFs and PowerPoints, as well as applications like Notion and Slack and even databases like Postgres and MongoDB.

The framework brings an array of connectors that assist in data ingestion, facilitating a seamless interaction with LLMs. Moreover, LlamaIndex boasts an efficient data retrieval and query interface.

LlamaIndex enables developers to input any LLM prompt and, in return, receive an output that is both context-rich and knowledge-augmementation.

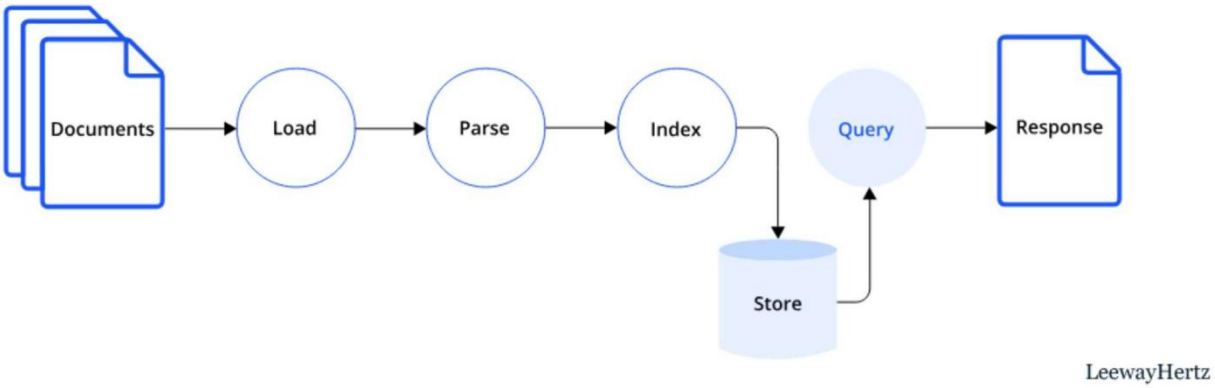
Key Feature of LlamaIndex:

● Data connectors allow ingestion from various data sources and formats.

● It can synthesize data from multiple documents or heterogeneous data sources.

● It provides numerous integrations with vector stores, ChatGPT plugins, tracing tools,

LangChain, and more.

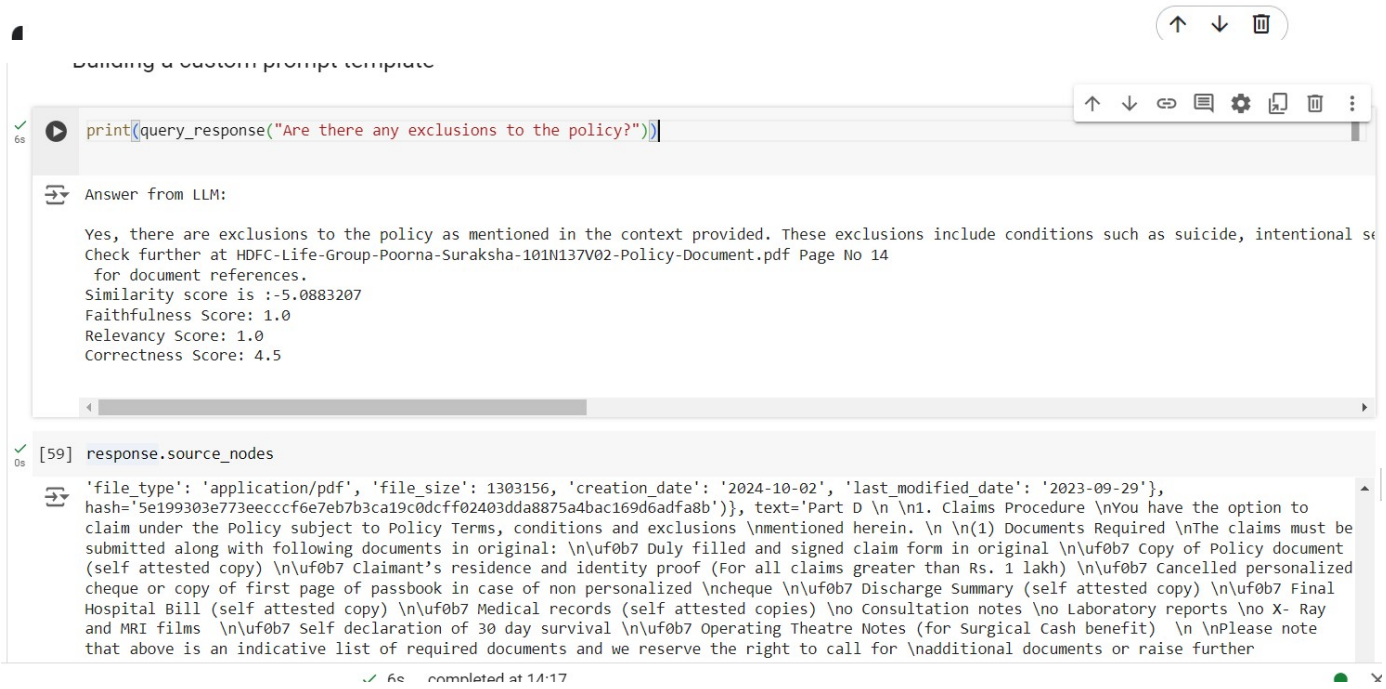


7. Generative Search Response from Insurance documents :

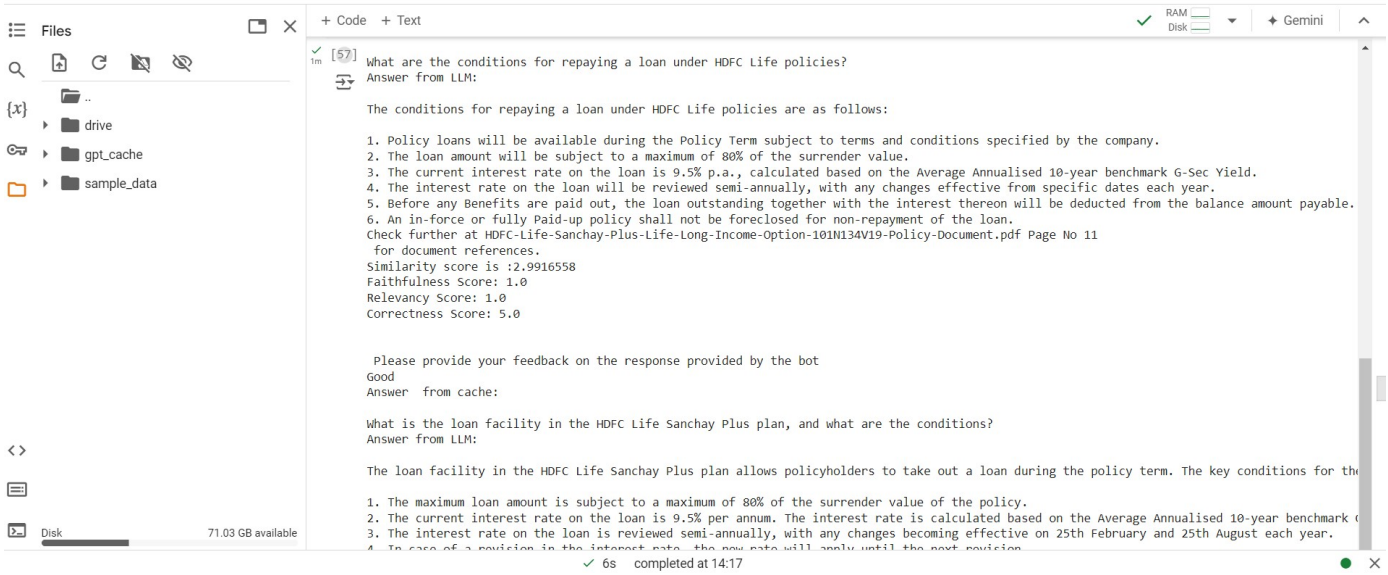
We have attached custom query generative search results.

```
[75]: response_c = openai.chat.completions.create(
          model="gpt-4o-mini",
          messages=messages_c)
      response_c.choices[0].message.content
```

```
[75]: 'The "Right to call for a second opinion" refers to the provision within your insurance policy that gives the insuran
ce company the authority to seek an independent medical examination or opinion regarding the diagnosis and/or treatme
nt you have claimed. This process is initiated in cases where there are doubts about the appropriateness or correctne
ss of the claimed diagnosis or treatment. \n\nHere are the key points regarding the right to call for a second opinio
n:\n\n1. **Authority of the Company**: The insurance company has the right to request a medical examination conducted
by a Medical Practitioner appointed by them.\n\n2. **Cost**: The expenses for this medical examination will be covere
d by the insurance company.\n\n3. **Final Decision**: The findings from this examination, along with the medical prac
titioner\'s opinion regarding the diagnosis and/or treatment, are considered final and binding on you as the policyho
lder.\n\nThis provision is meant to ensure that claims are legitimate and that the treatments being claimed for are n
ecessary and appropriate. If the company finds discrepancies or inappropriate claims from the examination, they can t
ake actions that may include declining the claim.'
```



8. Multiple Query Response


Files
..
drive
gpt_cache
sample_data
Disk
71.03 GB available
+ Code
+ Text
RAM
Disk
Gemini
[57]
1m
What are the conditions for repaying a loan under HDFC Life policies?
Answer from LLM:
The conditions for repaying a loan under HDFC Life policies are as follows:
1. Policy loans will be available during the Policy Term subject to terms and conditions specified by the company.
2. The loan amount will be subject to a maximum of 80% of the surrender value.
3. The current interest rate on the loan is 9.5% p.a., calculated based on the Average Annualised 10-year benchmark G-Sec Yield.
4. The interest rate on the loan will be reviewed semi-annually, with any changes effective from specific dates each year.
5. Before any Benefits are paid out, the loan outstanding together with the interest thereon will be deducted from the balance amount payable.
6. An in-force or fully Paid-up policy shall not be foreclosed for non-repayment of the loan.
Check further at HDFC-Life-Sanchay-Plus-Life-Long-Income-Option-101N134V19-Policy-Document.pdf Page No 11
for document references.
Similarity score is :2.9916558
Faithfulness Score: 1.0
Relevancy Score: 1.0
Correctness Score: 5.0
Please provide your feedback on the response provided by the bot
Good
Answer from cache:
What is the loan facility in the HDFC Life Sanchay Plus plan, and what are the conditions?
Answer from LLM:
The loan facility in the HDFC Life Sanchay Plus plan allows policyholders to take out a loan during the policy term. The key conditions for th
1. The maximum loan amount is subject to a maximum of 80% of the surrender value of the policy.
2. The current interest rate on the loan is 9.5% per annum. The interest rate is calculated based on the Average Annualised 10-year benchmark
3. The interest rate on the loan is reviewed semi-annually, with any changes becoming effective on 25th February and 25th August each year.
6s completed at 14:17

9. Challenges faced:

- Faced compatibility issue while importing RAGAS for evaluation.
- Compatibility issue while using gptcache
- Performance Bottlenecks
- Dependency Conflicts

10. Alternative Solutions:

-diskcache is used instead of gptcache

- instead of importing RAGAS we imported

```
from llama_index.core.evaluation import (
    CorrectnessEvaluator,
    FaithfulnessEvaluator,
    RelevancyEvaluator,
)
```

11. Alternative option

-reranking could be done with Cohere rerank