Policy Document generative search application

Project Goals

The objective of this project is to develop an efficient system for parsing, indexing, and querying information from an HDFC Policy PDF document. The solution aims to facilitate the swift and accurate retrieval of policy details based on user queries, ensuring that users can easily access and understand specific policy information.

Data Sources

Document Source: HDFC-Life-Group-Term-Life-Policy PDF

Purpose: The HDFC Policy PDF serves as the primary data source for this project. The document contains detailed information about insurance policies, which will be parsed and indexed.

Format: PDF

Content: Policy terms, conditions, coverage details, and other relevant sections.

Design Choices

Framework Choice: LlamaIndex is chosen for its capability to handle document parsing, indexing, and querying efficiently and it specialized for search and retrieval tasks

Evaluation

Evaluation and benchmarking are crucial in the development of large language models (LLMs) applications. LlamaIndex can autonomously generate questions from your data, facilitating an evaluation pipeline for RAG applications. The evaluation in LlamaIndex is categorised into two main types: Response Evaluation and Retrieval Evaluation.

Challenges Faced

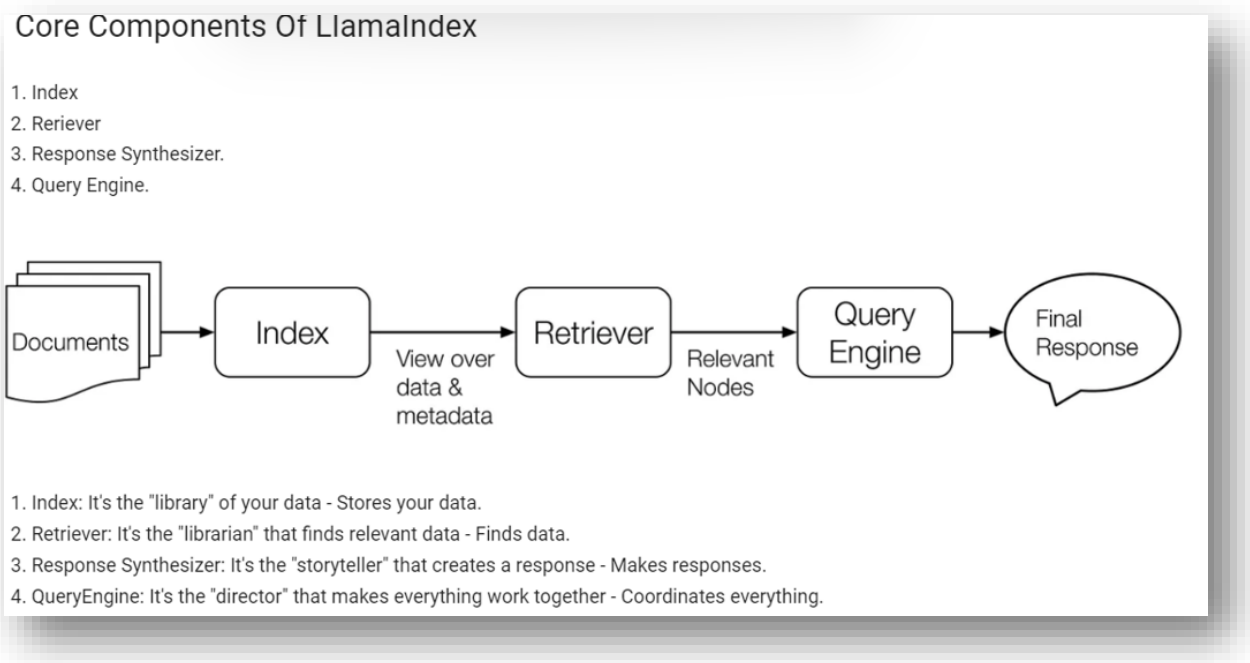
Complexity of PDF Documents: Insurance policies are often lengthy and contain complex structures such as tables, sections, and nested clauses, which can make it difficult for the indexer to parse and understand.

Contextual Retrieval: Insurance policies often require context to fully understand the meaning of specific clauses or terms, which can be difficult for a retrieval system to handle.

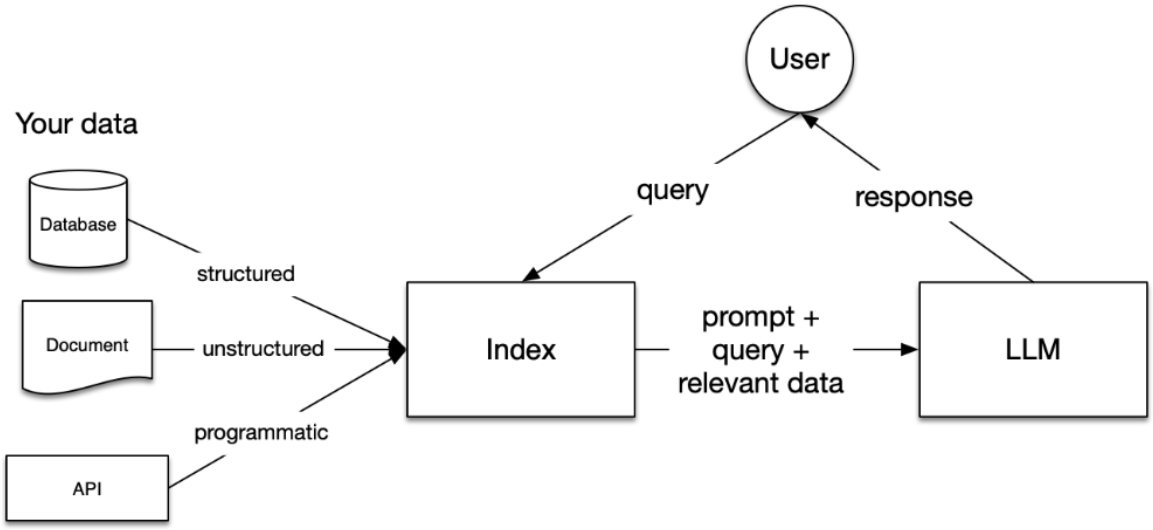
Indexing Efficiency: Indexing large policy documents or multiple documents can be resource-intensive and time-consuming, leading to performance issues.

User Query Interpretation Users might phrase queries in various ways, and interpreting these queries accurately can be challenging.

Flowchart



RAG:



The evaluation in LlamaIndex is categorised into two main types: **Response Evaluation and Retrieval Evaluation.**

In response evaluation, LlamaIndex uses benchmark LLMs such as GPT-4 to assess the quality of answers generated by LLMs. This evaluation focuses on these four main aspects:

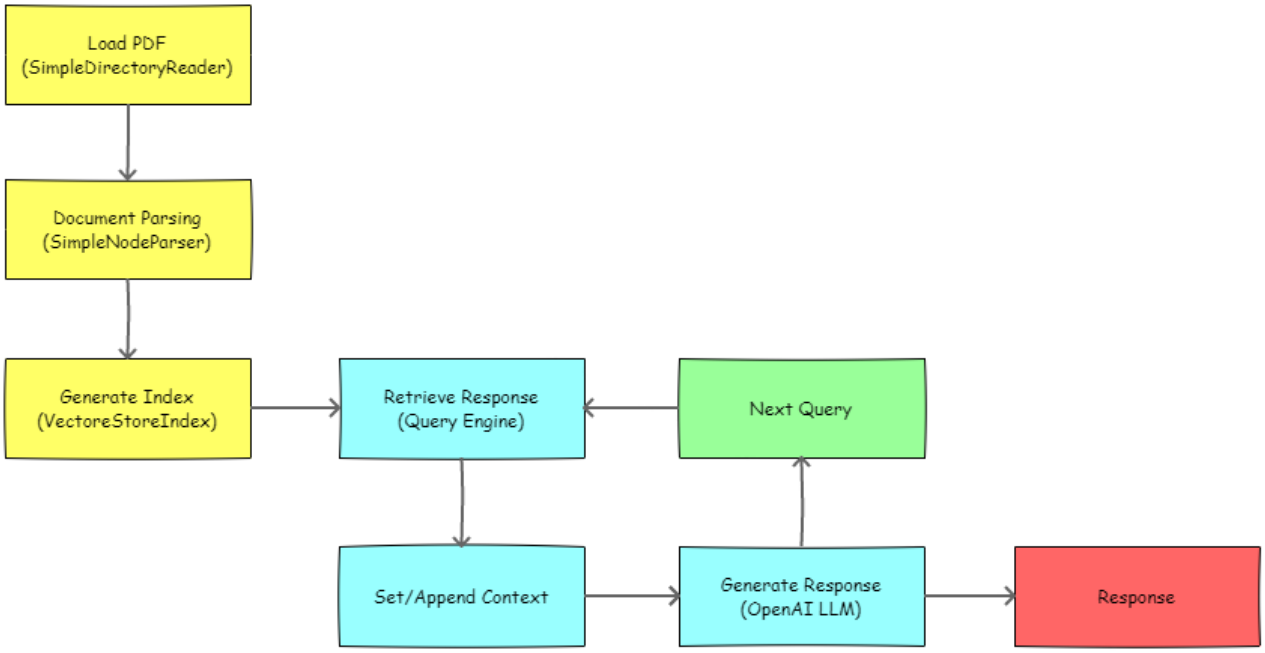
1. **Faithfulness**: Ensures the response accurately reflects the retrieved contexts without distortion or 'hallucination'

2. **Context relevance**: Assesses the relevance of the retrieved context and the generated answer to the initial query

3. **Correctness**: Checks if the generated answer aligns with a reference answer based on the query (this may require labelled data)

4. **Guideline adherence**: Examines whether the generated answer adheres to specific predefined guidelines

A flowchart illustrating the system design and various layers of the project:


Load PDF
(SimpleDirectoryReader)
Document Parsing
(SimpleNodeParser)
Generate Index
(VectoreStoreIndex)
Retrieve Response
(Query Engine)
Next Query
Set/Append Context
Generate Response
(OpenAI LLM)
Response