Project Assignment Report

Project Title: Semantic Spotter RAG – HDFC Insurance Policy Question Answering

1. Introduction

This project focuses on building a domain-specific Retrieval-Augmented Generation (RAG) system designed to answer questions about HDFC insurance policy documents. The solution leverages vector databases, retrievers, rerankers, and large language models to deliver accurate, grounded responses while reducing hallucinations.

2. Objectives

- Enable efficient question answering over HDFC insurance documents.
- Ground LLM responses in retrieved authoritative policy text.
- Persist and scale document embeddings for future queries.
- Improve retrieval accuracy using cross-encoder reranking and contextual compression.

3. Problem Statement

Traditional LLMs may hallucinate or provide incomplete answers when handling domain-specific queries. HDFC policy documents contain detailed conditions and terms that must be retrieved precisely. The challenge is to build a system that can:

- Search policy documents effectively.
- Extract the most relevant context.
- Generate fluent, faithful answers supported by retrieved evidence.

4. System Design and Workflow

The system is designed with the following workflow:

1. Document ingestion from insurance PDFs.
2. Preprocessing & splitting into chunks.
3. Embedding generation & caching.
4. Persistent vector storage using Chroma.
5. Retriever configuration with MMR & threshold.
6. Cross-encoder reranking for contextual compression.
7. LLM integration with ChatOpenAI.
8. Natural language response generation.

5. Key Components

- Document Loader: PyPDFDirectoryLoader
- Text Splitter: RecursiveCharacterTextSplitter
- Embeddings: OpenAIEmbeddings with cache

- Vector Store: Chroma (persistent storage)
- Retriever: MMR + threshold-based retriever
- Compression/Reranker: HuggingFace Cross-Encoder
- LLM: ChatOpenAI (default: gpt-4o-mini)

6. Implementation Details

The implementation ensures efficiency and accuracy by caching embeddings, persisting vector indices, and applying reranking for higher faithfulness. The architecture is modular, enabling flexibility to swap LLMs, retrievers, or embedding models.

7. Challenges and Trade-offs

- Data parsing challenges with varied PDF layouts.
- Chunk size trade-offs between recall and cost.
- Latency from reranking vs. quality improvements.
- Lack of automated evaluation metrics, to be addressed in future iterations.

8. Expected Outcomes

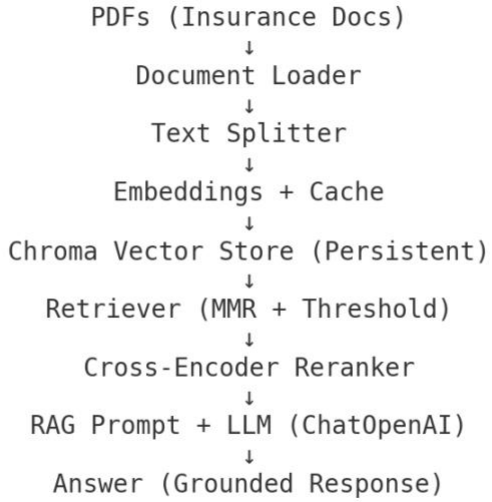
- Robust RAG system tailored for HDFC insurance queries.
- Scalable and persistent document knowledge base.
- Accurate and faithful natural language answers.

9. Conclusion

The Semantic Spotter RAG project successfully integrates retrieval, reranking, and generative components to deliver grounded responses. This design ensures scalability, modularity, and strong applicability to domain-specific question answering.

10. System Workflow Diagram

The following diagram illustrates the end-to-end flow of the Semantic Spotter RAG system.



11. Design Choices, Challenges, and Trade-offs

- Data parsing: PDFs vary in layout; PyPDFDirectoryLoader is robust, but for forms/tables, consider `unstructured` or layout-aware parsing.
- Chunking: We chose 1000/200 (size/overlap) as a good default; tune per corpus density. Larger chunks may reduce recall, while smaller chunks increase cost.
- Embeddings: OpenAI embeddings with cache reduce costs when re-running. You can switch to local models if privacy or cost requires.
- Vector store: Chroma with persistence ensures repeatable runs and incremental updates. For scale or multi-tenant use cases, consider managed stores like Pinecone or pgvector.
- Retrieval: MMR + threshold increases diversity and filters weak matches; adjust `score_threshold` if recall is too low/high.
- Compression: Cross-encoder reranking improves faithfulness but adds latency. Batch requests or reduce `top_n` for speed optimization.
- Generation: `gpt-4o-mini` balances cost and quality; upgrade or downgrade models depending on requirements.
- Evaluation: Automated checks (faithfulness/relevancy) or frameworks like RAGAS can be added for full rubric coverage, though omitted here to keep runtime lean.

12. Output Results from Model Queries

Q: What is the minimum age for doing a term insurance?
A: The minimum age for doing a term insurance is 18 years, as stated in the policy schedule. This applies to eligibility for participation in the insurance scheme.

Q: Can a 100 year plus person do a term insurance?
A: I don't know.

Q: What is condition of death while not wearing Seat Belt?
A: The context does not specifically address the condition of death while not wearing a seat belt. However, it implies that deaths resulting from accidents may be subject to exclusions based on various factors, such as participation in unlawful acts or being under the influence of substances. Therefore, the specific impact of not wearing a seat belt on death conditions is not detailed in the provided information.

Q: What is the life insurance coverage for disability?
A: The provided context does not specify the life insurance coverage for disability. It mainly discusses benefits related to death, critical illness, and exclusions. Therefore, I don't know the answer to the question regarding life insurance coverage for disability.

Q: What is criteria for HDFC group insurance?
A: The criteria for HDFC group insurance include eligibility of members as specified in the policy, adherence to the terms and conditions outlined in the policy document, and acceptance of any pre-existing conditions as stated in the proposal form. Additionally, the

insurance is a non-linked, non-participating group life insurance policy. Specific exclusions and prerequisites for benefits also apply, which are detailed in the policy documentation.

Q: What are the benefits of HDFC Sampoorna-Jeevan insurance?
A: The HDFC Life Sampoorna Jeevan insurance plan offers several benefits, including participation in the company's surplus profits, which can enhance the policy's value over time. It provides guaranteed income benefits, with options for lump sum or income payouts, depending on the chosen plan variant. Additionally, the policy ensures a basic sum assured is paid out upon maturity or in the event of the policyholder's death.

Q: What are HDFC Life Sanchay Plus Life Long Income Option?
A: HDFC Life Sanchay Plus Life Long Income Option is a non-participating, non-linked savings insurance plan that provides guaranteed income upon maturity. The maturity benefit is paid as guaranteed income starting from the year after the policy term until the individual reaches age 99, provided all premiums are paid. At the end of the payout period, the policy terminates and returns the total premiums paid.