Build a RAG System

Problem Statement:

Organizations stores the manuals, policy documents, legal contracts or any other documents in unstructured format like pdfs. In this case study use any Life Insurance Policy pdf document like *" Principal-Sample-Life-Insurance-Policy.pdf"*, and retrieve instant and accurate answers from the document.

Traditional keyword-based search leads to inefficient contextual information, missed insights and further leads to frustrated users.

The case study should concentrate on the below points.

- Understand the context of the question,
- Match semantically similar but differently worded content,
- And synthesize answers across multiple sections.

Why LangChain/ LlamaIndex is an ideal framework?

LangChain and LlamaIndex (formerly GPT Index) are two of the most popular frameworks for building GenAI-powered applications, especially ones that use large language models (LLMs) with external data. Here's a breakdown of why they are considered ideal frameworks:

LlamaIndex: Ideal for Connecting LLMs to Custom Data

LlamaIndex focuses on **making external data (e.g., PDFs, Notion, databases) accessible to LLMs** through structured ingestion and retrieval pipelines. It excels in **indexing, querying, and retrieval augmentation**.

1. **Flexible Data Ingestion:** It supports complex data sources and lets you preprocess, chunk, and embed documents with full control.
2. **Advanced Indexing Techniques:** Offers vector, keyword, tree, and hybrid indices for optimized retrieval depending on the use case.
3. **RAG-Friendly:** Designed specifically for **retrieval-augmented generation**, with built-in support for query transformations, summarization, and reranking.

4. **Composable Query Engines:** Let you chain together multiple indices, routers, and retrieval strategies.

5. **Seamless LangChain Integration:** LlamaIndex works beautifully as a retriever inside LangChain, letting you get the best of both worlds.

Use Cases: AI over private data (e.g., PDF Q&A, Notion docs, SQL), Custom search engines, Knowledge management tools, Legal, finance, or healthcare data querying

LangChain: Ideal for Building Agentic and Modular LLM Apps

LangChain is a **framework for developing applications powered by LLMs** that need to interact with outside tools, data sources, or workflows. It shines in orchestrating **multi-step reasoning**, calling APIs, and integrating components like memory, tools, and agents.

1. **Modular Components :** LangChain abstracts common patterns: prompts, chains, tools, memory, agents, and retrievers. This makes it easy to plug-and-play components and swap models or tools.

2. **Agent Support:** Enables dynamic decision-making by LLMs using "tools" — for example, an LLM deciding when to call a calculator or search engine.

3. **Data-Aware Applications:** LangChain integrates with vector stores (e.g., Pinecone, FAISS), document loaders, and retrievers to enable **retrieval-augmented generation (RAG)**.

4. **Tool Ecosystem:** It supports LangServe for deploying chains as APIs, LangGraph for building complex workflows, and LangSmith for debugging and monitoring.

5. **Model Agnostic:** Works with OpenAI, Anthropic, Cohere, Azure, HuggingFace, and even local models.

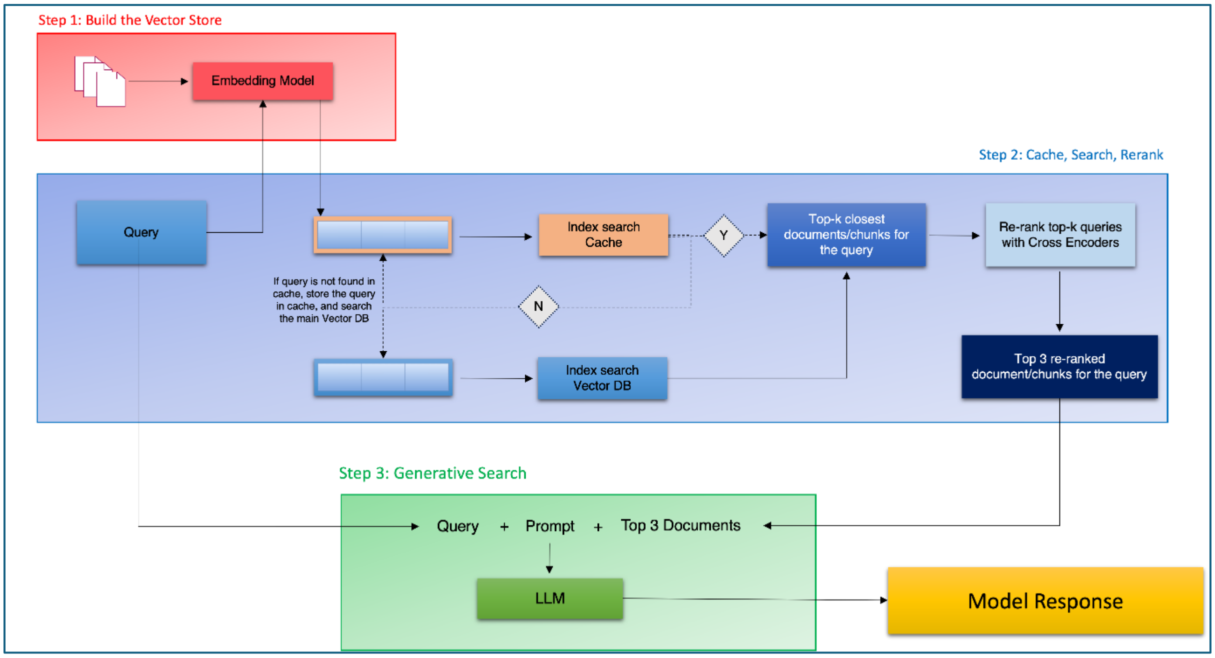
Use Cases: AI Assistants (e.g., chatbots that use tools), RAG Pipelines, Workflow agents (e.g., coding agents, financial analysts), API-calling bots

In this Case study, we use both llama index and langchain as they are good for “Document Q&A (RAG)”

Data Sources:

1. **Source file:** Text-based PDF document *" Principal-Sample-Life-Insurance-Policy.pdf"*
2. **Format:**
   These are multi-page PDF loaded directly from Google Drive (/content/drive/MyDrive/helpmate_ai/Policy_Documents).
3. **Preprocessing:**
   Llamaindex: In Pdf, pages are splitted and blank pages are removed. Used chunksize=1024 and chunk_overlap=100
   Langchain: PDFs are split into overlapping text chunks to retain context between chunks. chunksize=1000 and chunk_overlap=100

System Architecture:


Step 1: Build the Vector Store
Embedding Model
Step 2: Cache, Search, Rerank
Query
Index search Cache
Y
Top-k closest documents/chunks for the query
Re-rank top-k queries with Cross Encoders
If query is not found in cache, store the query in cache, and search the main Vector DB
N
Index search Vector DB
Top 3 re-ranked document/chunks for the query
Step 3: Generative Search
Query + Prompt + Top 3 Documents
LLM
Model Response

System Design:

1. Innovation and Creativity in System Design

The system intelligently combines LLMs with domain-specific documents (e.g. PDFs, policies). This is done using **RAG (Retrieval-Augmented Generation)**, allowing the model to answer questions accurately based on real, up-to-date information — not just what it was trained on. This is innovative because it turns static data into a dynamic, conversational AI system.

2. Optimum System Architecture, Workflow, and Implementation

The architecture is modular and scalable:

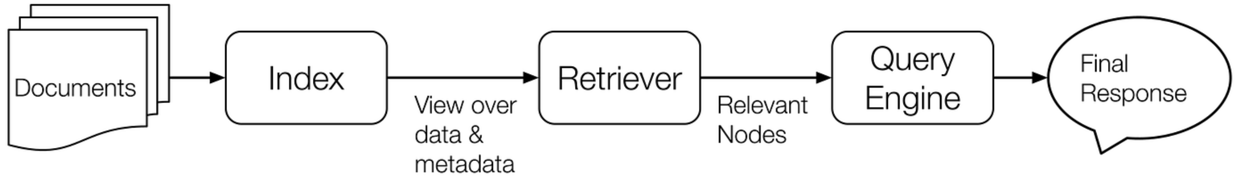
- **Input**: PDFs or documents are loaded and split into small chunks preserving semantic flow.
- **OpenAI Embeddings**: Each chunk is converted into a vector using an embedding model (e.g. text-embedding-ada-002).
- **Vector stores:** Chroma is used as vector store to have fast similarity search. We can use FAISS also.
- **CrossEncoder for Reranking**: A second filtering layer ensures better precision of context retrieved.
- **Open AI Model**: Model='gpt-3.5-turbo', temperature=0 are used for fast, reliable response generation with factual consistency.
- **Output**: The LLM generates a context-aware, accurate response.

  Llamaindex: Sequence of questions are answered
  LangChain Expression Language (LCEL): Questions are answered with chat like model and exits chat in case of "quit"

3. Appropriate Use of LangChain / LlamaIndex Components

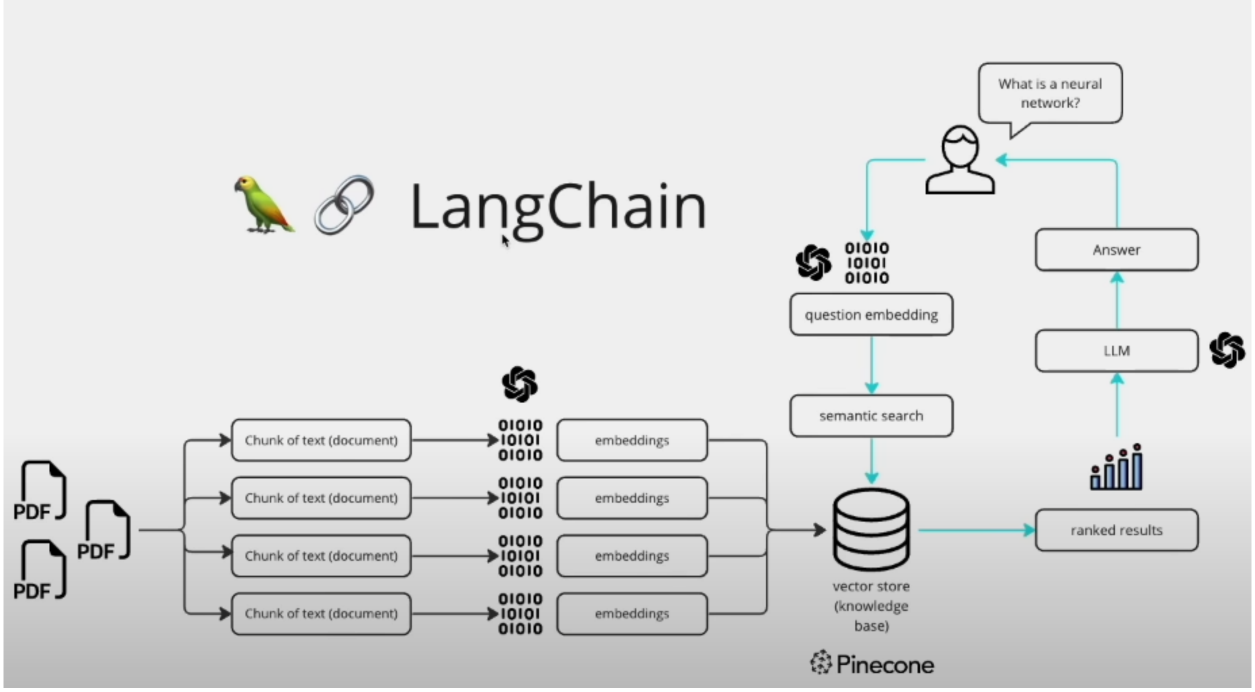
- **LangChain** is used for building structured workflows: chains, prompts, chat memory
- **LlamaIndex** (if integrated) excels at indexing and querying large structured/unstructured document sets.
- **Embeddings + Vector Store**: Enable semantic search for document retrieval.
- **LLMs (e.g. GPT-3.5)**: Generate human-like, contextually accurate answers from retrieved info.

Llamaindex Flow chart:



This is a flow chart for llamaindex. The pdf document are collected from the source directory. The document is chunked and indexed and stored in vectorstore. The retriever will find the relevant data based on queries. Response synthesizer will create a response and the Query engine will co-ordinate everything using the LLM and provides the final response.

Langchain Flow chart:



This is a flow chart for langchain. The pdf document are collected from the source directory. The document is chunked and embedded and stored in vectorstore(knowledge base). Based on the user questions as inputs the semantic search is done on knowledge base and returns the results which is ranked later and send to the LLM as response to user query.

Challenges Faced

- **Chunk Quality**: Overlap needed to be tuned carefully — too little caused loss of context, too much created redundancy.

- **Reranking Efficiency**: Integrating the CrossEncoder re-ranker required additional computation but significantly improved result quality.
- **Context Management**: Keeping responses grounded strictly to the source documents needed robust prompt templates.
- **Colab Storage Paths**: Ensuring persistence and reusability required managing custom file paths inside Google Drive.

Lessons Learned:

- RAG pipelines are powerful but require tuning of chunking and context retrieval
- Re-ranking adds major boost to answer relevance
- LangChain simplifies orchestration but needs careful prompt engineering
- Persistent storage is critical in Colab-based projects

Results:

Llamaindex query results:

# Query 1

Query("What risks or events are covered under this policy?")

Query : What risks or events are covered under this policy?

Response : The policy covers accidental death and dismemberment resulting from events such as willful self-injury, disease or medical treatment complications, participation in certain criminal activities, involvement in specific aeronautic activities, military duty, war, alcohol use exceeding legal limits, operation of a vehicle or boat under the influence, unauthorized drug use, and injuries not related to employment for wage or profit.

- Page 58: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6015 Section B - Member Accidental Death and Dismemberment Insurance, Page 6 a. willful self-injury or se...

- Page 27: This policy has been updated effective January 1, 2014 PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS GC 6006 Section A - Eligibility, Page 2 If a Member's Dependent is employed and is covered u...

- Page 13: This policy has been updated effective January 1, 2014 GC 6002 PART I - DEFINITIONS, PAGE 5 a. A licensed Doctor of Medicine (M.D.) or Osteopathy (D.O.); or b. any other licensed health ...

# Query 2

Query("what is the life insurance coverage for disability")

Query("what is the life insurance coverage for disability")

The life insurance coverage for disability includes provisions for ADL Disability or Total Disability. To be eligible for Coverage During Disability, a Member must meet certain qualifications such as becoming ADL Disabled or Totally Disabled while insured for Member Life Insurance, being under the care of a Physician, providing proof of disability when required, and undergoing Medical Examinations or Evaluations as needed. Written proof of ADL Disability or Total Disability must be submitted to the insurance provider within a specified timeframe, and further proof may be requested periodically. If the Member passes away while disabled, final proof of disability continuation must be provided to the insurance provider.

- Page 49: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6013 Section A - Member Life Insurance, Page 4 Payment of benefits will be subject to the Beneficiary and Facilit...

- Page 51: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6013 Section A - Member Life Insurance, Page 6 Coverage During Disability will cease on the earliest of: (1) the...

- Page 50: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6013 Section A - Member Life Insurance, Page 5 The Principal may require that a ADL Disabled or Totally Disabled Me...

# Query 3

Query("Summarize the key benefits from the insurance policy documents.")

Query : Summarize the key benefits from the insurance policy documents.

Response : The risks or events covered under this policy include accidental death and dismemberment that are not a result of willful self-injury, disease or medical treatment complications, participation in criminal activities, certain aeronautic activities, military duty, war, excessive alcohol consumption, drug use, or injuries sustained during employment for wage or profit.

- Page 58: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6015 Section B - Member Accidental Death and Dismemberment Insurance, Page 6 a. willful self-injury or se...

- Page 27: This policy has been updated effective January 1, 2014 PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS GC 6006 Section A - Eligibility, Page 2 If a Member's Dependent is employed and is covered u...

- Page 13: This policy has been updated effective January 1, 2014 GC 6002 PART I - DEFINITIONS, PAGE 5 a. A licensed Doctor of Medicine (M.D.) or Osteopathy (D.O.); or b. any other licensed health ...

# Query 4

Query("What riders or add-ons are available?")

Query : What riders or add-ons are available?

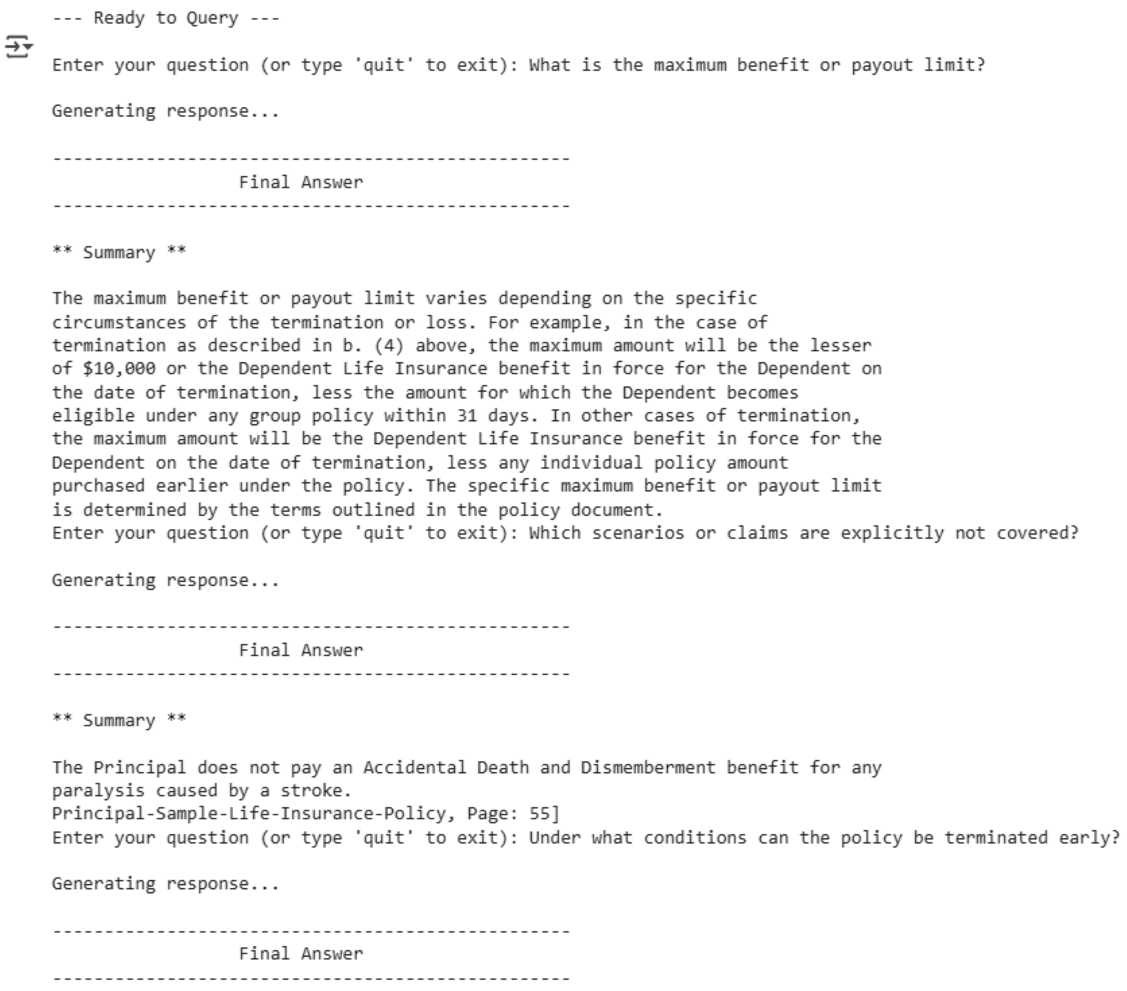
Response : The risks or events covered under this policy include accidental death and dismemberment that are not a result of willful self-injury, disease or medical treatments, participation in criminal activities, certain aeronautic activities, military duty, war, excessive alcohol consumption, drug use, and injuries sustained during employment for wage or profit.

- Page 58: This policy has been updated effective January 1, 2014 PART IV - BENEFITS GC 6015 Section B - Member Accidental Death and Dismemberment Insurance, Page 6 a. willful self-injury or se...

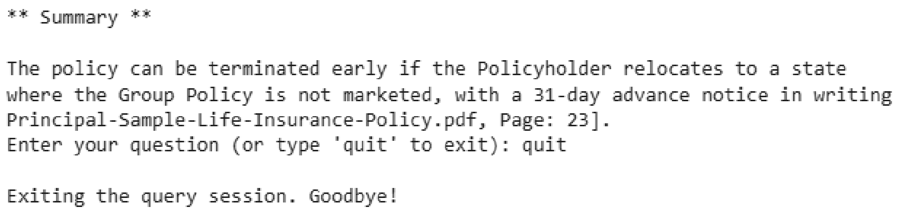
- Page 27: This policy has been updated effective January 1, 2014 PART III - INDIVIDUAL REQUIREMENTS AND RIGHTS GC 6006 Section A - Eligibility, Page 2 If a Member's Dependent is employed and is covered u...

- Page 13: This policy has been updated effective January 1, 2014 GC 6002 PART I - DEFINITIONS, PAGE 5 a. A licensed Doctor of Medicine (M.D.) or Osteopathy (D.O.); or b. any other licensed health ...

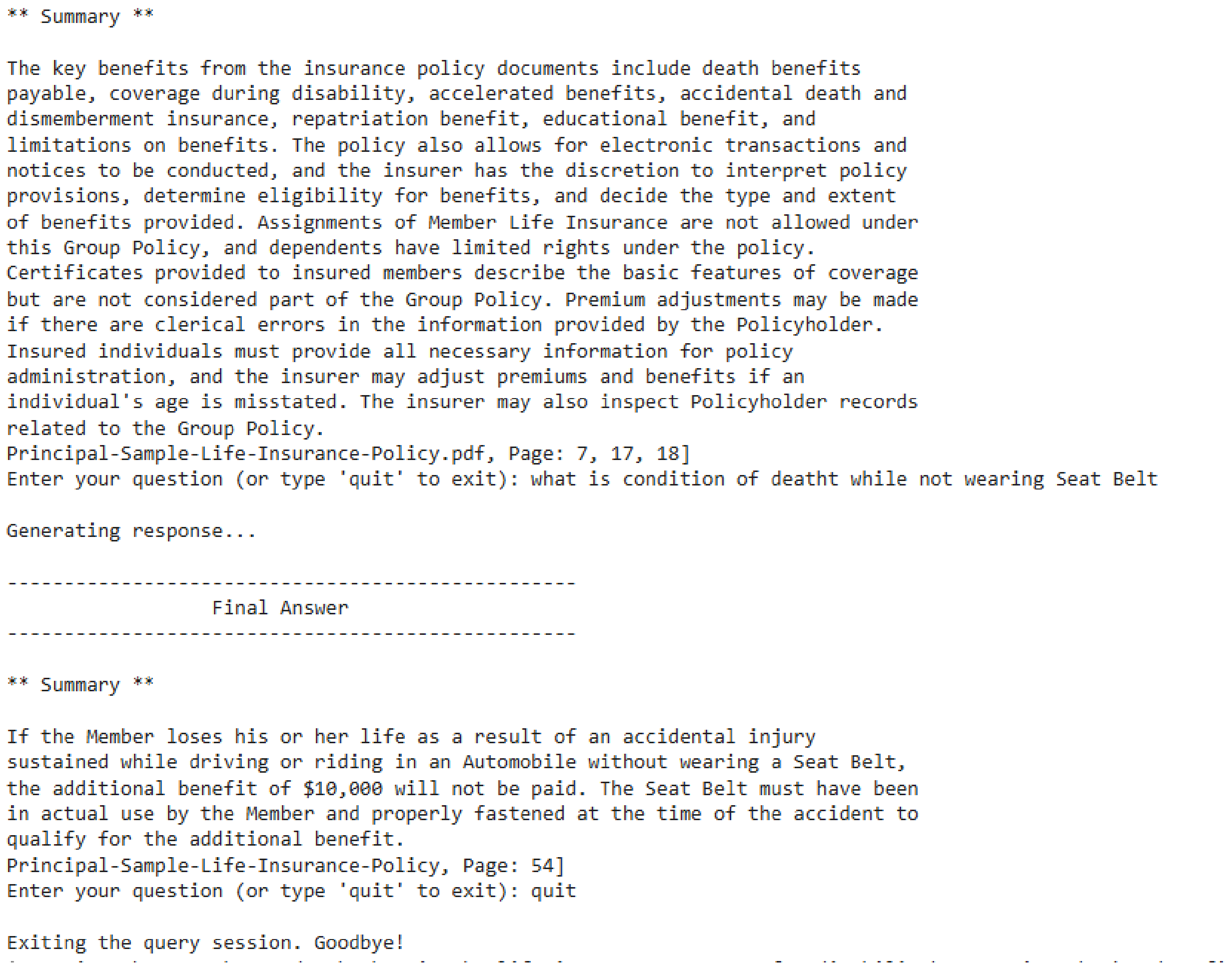
Langchain results: (Screenshots)



```
--- Ready to Query ---

Enter your question (or type 'quit' to exit): What is the maximum benefit or payout limit?

Generating response...

--------------------------------------------------
                  Final Answer
--------------------------------------------------

** Summary **

The maximum benefit or payout limit varies depending on the specific
circumstances of the termination or loss. For example, in the case of
termination as described in b. (4) above, the maximum amount will be the lesser
of $10,000 or the Dependent Life Insurance benefit in force for the Dependent on
the date of termination, less the amount for which the Dependent becomes
eligible under any group policy within 31 days. In other cases of termination,
the maximum amount will be the Dependent Life Insurance benefit in force for the
Dependent on the date of termination, less any individual policy amount
purchased earlier under the policy. The specific maximum benefit or payout limit
is determined by the terms outlined in the policy document.
Enter your question (or type 'quit' to exit): Which scenarios or claims are explicitly not covered?

Generating response...

--------------------------------------------------
                  Final Answer
--------------------------------------------------

** Summary **

The Principal does not pay an Accidental Death and Dismemberment benefit for any
paralysis caused by a stroke.
Principal-Sample-Life-Insurance-Policy, Page: 55]
Enter your question (or type 'quit' to exit): Under what conditions can the policy be terminated early?

Generating response...

--------------------------------------------------
                  Final Answer
--------------------------------------------------
```



```
** Summary **

The policy can be terminated early if the Policyholder relocates to a state
where the Group Policy is not marketed, with a 31-day advance notice in writing
Principal-Sample-Life-Insurance-Policy.pdf, Page: 23].
Enter your question (or type 'quit' to exit): quit

Exiting the query session. Goodbye!
```



```
** Summary **

The key benefits from the insurance policy documents include death benefits
payable, coverage during disability, accelerated benefits, accidental death and
dismemberment insurance, repatriation benefit, educational benefit, and
limitations on benefits. The policy also allows for electronic transactions and
notices to be conducted, and the insurer has the discretion to interpret policy
provisions, determine eligibility for benefits, and decide the type and extent
of benefits provided. Assignments of Member Life Insurance are not allowed under
this Group Policy, and dependents have limited rights under the policy.
Certificates provided to insured members describe the basic features of coverage
but are not considered part of the Group Policy. Premium adjustments may be made
if there are clerical errors in the information provided by the Policyholder.
Insured individuals must provide all necessary information for policy
administration, and the insurer may adjust premiums and benefits if an
individual's age is misstated. The insurer may also inspect Policyholder records
related to the Group Policy.
Principal-Sample-Life-Insurance-Policy.pdf, Page: 7, 17, 18]
Enter your question (or type 'quit' to exit): what is condition of deatht while not wearing Seat Belt

Generating response...

--------------------------------------------------
                  Final Answer
--------------------------------------------------

** Summary **

If the Member loses his or her life as a result of an accidental injury
sustained while driving or riding in an Automobile without wearing a Seat Belt,
the additional benefit of $10,000 will not be paid. The Seat Belt must have been
in actual use by the Member and properly fastened at the time of the accident to
qualify for the additional benefit.
Principal-Sample-Life-Insurance-Policy, Page: 54]
Enter your question (or type 'quit' to exit): quit

Exiting the query session. Goodbye!
```