

Insurance Question Answering System Using LangChain and Open Source LLMs

Report by: [Aditi Sharma](#)

Introduction

This report explores the feasibility of using question-answering chains to analyse and extract information from insurance industry documents. By combining text preprocessing techniques, semantic embeddings, and large language models, we aim to answer specific questions about the contents of PDF documents with accuracy and efficiency.

Methodology

▪ Document Preprocessing:

- **Ingestion:** We utilized PyPDFLoader to load insurance industry documents in PDF format.
- **Text Chunking:** To facilitate efficient processing and embedding generation, we segmented the documents into smaller text chunks using CharacterTextSplitter with a custom separator ('\n'), chunk size (1000 characters), and overlap (200 characters).
- **Embedding Creation:** Each text chunk was converted into a numerical representation using SentenceTransformerEmbeddings and the "all-MiniLM-L6-v2" model, capturing semantic relationships within the text.
- **Vectorstore Creation:** A Chroma vectorstore was established to efficiently store the text chunk embeddings and facilitate similarity searches.

▪ Question Handling and Answer Generation:

- **Question Processing:** The specific question to be answered, "How many life insurance companies were in 1985 in Canada?", was defined and processed for further analysis.
- **Similarity Search:** The Chroma vectorstore was employed to identify the most relevant text chunks based on their semantic similarity to the processed question. A threshold of k=5 documents was chosen to achieve a balance between accuracy and computational efficiency.

- **LLM Integration:** The Hugging Face model "tiiuae/falcon-7b-instruct" was loaded and configured with specific parameters (temperature=0.3, max_new_token=2000) for optimal answer generation.
- **Chain Execution:** The question-answering chain was executed with the retrieved relevant documents and the processed question as inputs, leveraging the LLM to generate a comprehensive answer.

Code Link:

[https://github.com/Aditi0712/PDFchat_langchain/blob/main/geekrabbtask_\(2\).ipynb](https://github.com/Aditi0712/PDFchat_langchain/blob/main/geekrabbtask_(2).ipynb)

Challenges Encountered

- **Document Understanding:** Accurately extracting relevant information from complex and legalistic insurance documents presented a challenge.
- **Embedding Selection:** Choosing the optimal semantic embedding model for capturing nuances within the insurance domain required careful consideration.
- **Similarity Thresholding:** Determining the appropriate k value for balancing retrieval accuracy and computational cost was an important decision.
- **LLM Bias:** The potential for LLM bias in answer generation needed to be addressed through careful model selection and parameter tuning.

Results

- **Question:** How has technology impacted the insurance industry?
Answer: The insurance industry has been impacted by technology in a number of ways. Technology has enabled the industry to streamline processes, such as underwriting and claims processing, and has allowed for the development of new products and services. Technology has also enabled the industry to better serve customers and to better manage risk. Additionally, the rise of the Internet has allowed for more efficient and convenient communication between insurers and customers. Overall, technology has been a driving force in the evolution of the insurance industry.
- **Question:** How many life insurance companies were in 1985 in Canada?
Answer: There were 167 life insurance companies in Canada in 1985.

➤ **Question:** What policy issues are addressed in the section on industrial development, and how do they impact the insurance industry?

Answer: The section on industrial development addresses policy issues related to the competitiveness of the insurance industry, including issues related to product innovation, information asymmetries, and the effects of globalization. These issues impact the insurance industry by affecting the ability of insurers to remain competitive and to continue to offer a wide range of products and services to their clients.

The Industrial Development section also addresses the impact of the insurance industry on the economy, including its effects on job creation and economic growth. The section also discusses the impact

While the answer provides valuable information, its accuracy could be further evaluated by comparing it to reliable external sources or manually reviewing the retrieved documents. Additionally, the analysis could be expanded to investigate other aspects of the insurance industry or address different questions using the established pipeline.

References

- [Langchain Documentation.](#)
- [Hugging Face](#)
- [Llm model falcon-7b-instruct](#)