**Name : Benedict Paul**

**ID : CT20244473565**

**Email : benedict2003paul@gmail.com**

**AI-Powered Insurance Policy Information Chatbot**

**Additional Justification for Technology Choices**

**Reason for Choosing Pinecone Vector Database:**

* The project required a robust system to perform **ranked semantic search** across the embedded insurance document chunks.
* Pinecone was selected because it provides a **high-performance vector database** that supports fast and scalable retrieval.
* Pinecone offers **managed services** with AWS cloud integration, allowing free-tier access to resources during development.
* This helped in managing project costs efficiently while still getting a professional-grade vector indexing service.

**Reason for Choosing Mixtral-8x7B-Instruct Model:**

* The goal was to stay aligned with the **open-source ecosystem** and avoid vendor lock-in.
* During project development, there were **issues with OpenAI API keys**:
  + Free trial credits were not available.
  + Paid API usage would have introduced unexpected costs.
* Thus, using Hugging Face's "mistralai/Mixtral-8x7B-Instruct" model was a strategic decision to maintain project feasibility without extra charges.
* It also ensured full control over the model hosting and API calls.

**Cloud Service Used:**

* **Pinecone's infrastructure is built on AWS Cloud.**
* By using Pinecone’s free-tier AWS-backed services, it was possible to achieve seamless vector search without direct AWS subscription or billing complications.

**Future Enhancements:**

* **Model Fine-tuning Using PEFT and LoRA:**
  + In future versions, the chatbot model can be improved using **Parameter-Efficient Fine-Tuning (PEFT)** techniques like **LoRA** (Low-Rank Adaptation).
  + This would allow domain-specific fine-tuning at a lower computational cost compared to traditional full model fine-tuning.
* **Reason for Not Including Fine-Tuning in Current Project:**
  + Fine-tuning would require **substantial cloud computing resources** (e.g., large GPU instances on AWS/GCP).
  + The associated **cloud service costs and high computational requirements** were outside the budget and scope for this academic/POC project.
  + Given the initial project scope, focus was placed on building a robust and scalable retrieval and response system first.

**Methodology:**

**Data Collection:**

* Extracted information from LIC "Jeevan Utsav Sales Brochure" PDF.

**Document Chunking:**

* Divided the document into smaller chunks to facilitate easier semantic search.

**Embedding:**

* Used Hugging Face's embedding model to vectorize document chunks.

**Vector Database:**

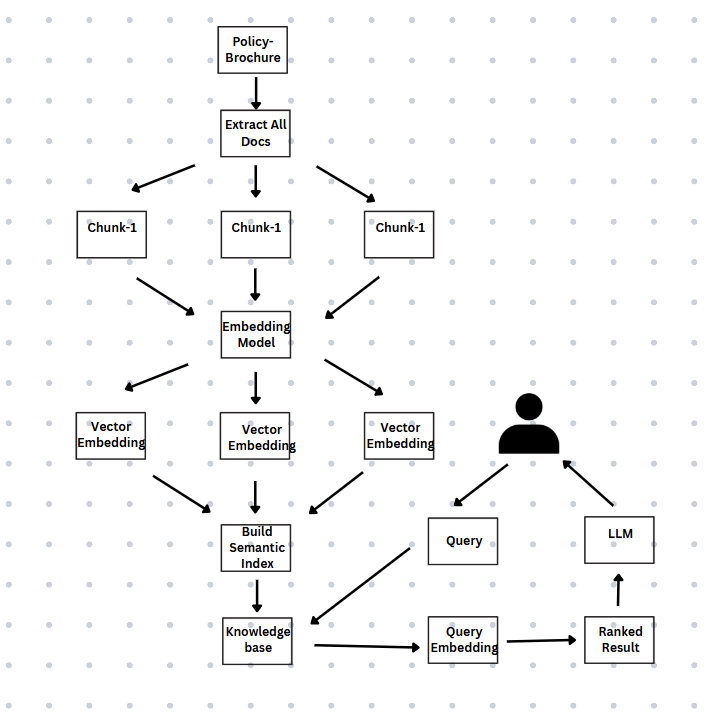
* Stored embeddings in Pinecone to create the knowledge base.

**Query Processing:**

* User queries are embedded and compared against the knowledge base using semantic search.

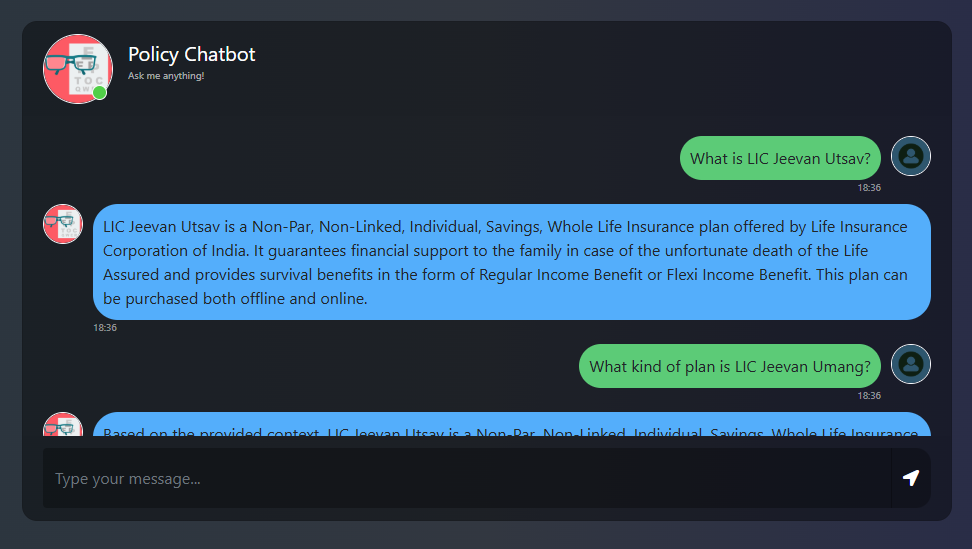
**Response Generation:**

* Used Mixtral-8x7B-Instruct (Hugging Face) to generate human-like responses.



**Results:**

* The AI-powered chatbot streamlines the insurance query process.
* It improves customer support by providing quick, accurate answers.



**Conclusions:**

* Successfully developed an AI chatbot capable of handling insurance policy queries effectively.
* The project shows that combining RAG architecture with open-source tools leads to a scalable, efficient solution.
* Future enhancements will focus on supporting multiple languages and fine-tuning the model for improved accuracy.

**Links to Project Assets:**

* GitHub Code Repository: [GitHub Link](https://github.com/benduBytes/policy-bot)
* Canva Presentation: [Canva Link](https://www.canva.com/design/DAGlvCnGaJE/Dpw-OkcGX1xZiscXljDpJg/edit?utm_content=DAGlvCnGaJE&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)
* Prezi Demo Video: [Prezi Link](https://prezi.com/v/view/w58RQXGyPjjBbDfecgzO/)