Idea brief

We are developing AI chatbots using open-source LLM models, ensuring that our knowledge base operates within a secure sandbox environment, with no data being transmitted over the internet. By leveraging Retrieval-Augmented Generation (RAG), we can utilise our existing knowledge base to enhance the foundation LLM, generating accurate and privacy-preserving results. This approach benefits both our customers/end users and the company by maintaining strict data privacy.

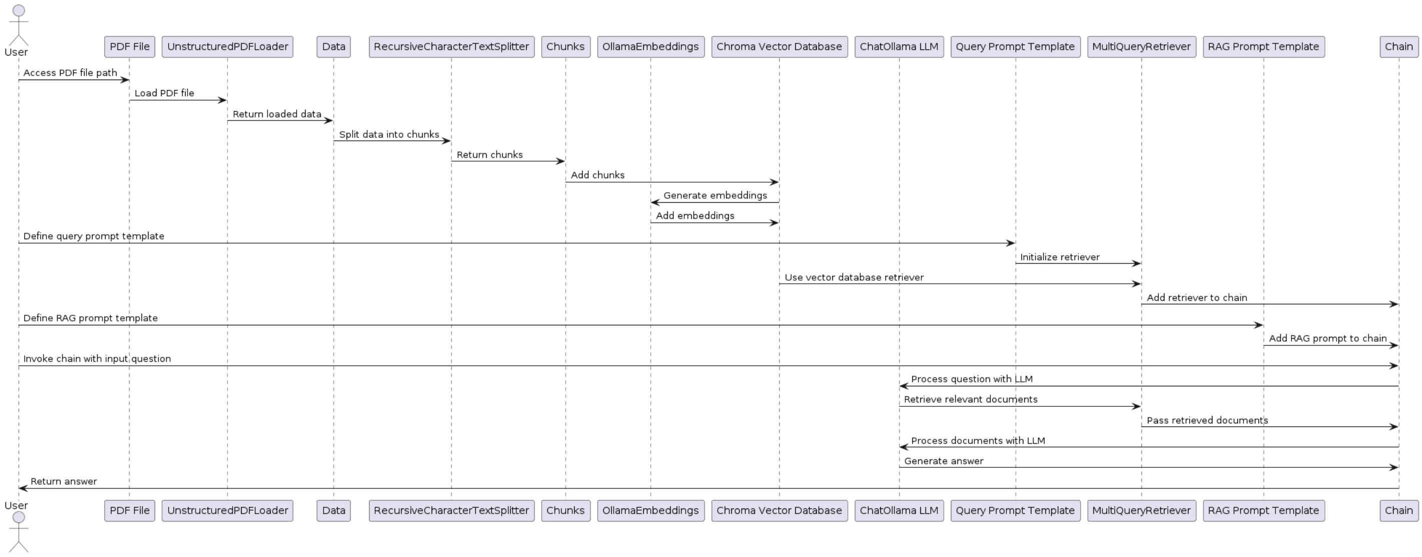
Uniqueness

- Used open source entirely
- Can read unstructured documents like PDF
- Ensures guardrails to stick to a context
- Ensures no hallucination in the responses
- Ensures Data Governance

Potential Impact

- Optimal search from existing knowledge base using foundation LLM models so that customer support can be self serving (level 0)
- Lesser reliance on level 1 support
- Only the most important and complex calls are forwarded for manual intervention

Process Flow Diagram



Tech Stack

- PlantUML
    - Write code and generate sequence diagrams on the fly

- LLM (Large Language Models) - Llama3, Mistral etc
- Ollama - to handle the NLP Tasks effectively e.g. embeddings
- Unstructured - Read unstructured data i.e. pdf
- NLP(Natural Language Processing)
- RAG (Retrieval Augmented Generation)
- Python
- ChromaDB
- Langchain