GURUDOCS: CHATBOT FOR QUERYING DOCUMENTS USING RAG

PROJECT PROPOSAL

Alvin Wong Ann Ying (A0266486M)

Brandon Chua Hong Huei (A0168608U)

Ong Si Ci (A0266450E)

Introduction

- Employees often need to understand processes and workflows from confidential policy / regulatory documents, some of which can span up to 100+ pages.
- By leveraging Retrieval Augmented Generation (RAG) and Large Language Models (LLMs), GuruDocs will help to provide context-aware search results for **document question and answering**, ultimately **improving the overall user experience and efficiency of information retrieval**.

Project Background / Market Context

- **Problem Statement:**
  - Documents are crucial for ensuring compliance, making informed decisions, and understanding organizational processes.
  - However, searching for specific information within lengthy documents can be time-consuming and tedious, leading to inefficiencies and potential errors.
  - Corporate documents containing SOPs and policies are often confidential and hence users are unable to leverage commercial products to perform document question answering
- **Limitations with current approaches:**
  - Keyword-based search systems may be able to provide information but often fall short when it comes to retrieving relevant information from complex documents. As such, this can lead to frustration, wasted time, and reduced productivity.
  - While current LLMs have a good general knowledge of the world, they lack information on specific domain knowledge or proprietary data sources and hence are unable to be used off the shelf. Fine-tuning LLMs to perform specific documents question and answering is tedious and resource intensive
- To address this challenge, we propose a local/on-premise deployment of GuruDocs for document question and answering using RAG and LLM.

Project Scope

- Developing a chatbot product for document question and answering using RAG and LLM
- Key Components:
  - **Knowledge search** using RAG
    - Retrieving the relevant information from the document based on user queries.
    - Various search algorithms for RAG can be used, (e.g. Maximum Marginal Relevance, Compression or LLM-aided retrieval) to find and retrieve relevant information sought by the user
  - **Knowledge reasoning** using LLMs
    - Process of understanding, interpreting and synthesizing information to provide meaningful responses to user queries.
    - For example, understanding the context of the query, analyzing the content of the retrieved documents, and generating a coherent response based on this analysis

Market Research

- **Key Players/Competitors:**
  - OpenAI's ChatGPT is one of the biggest players in the LLM space. They offer document question and answering as a paid feature in AskAI.
  - A comparison of GuruDocs features with commercial product offerings is shown in the next slide
- **Market Trends:**
  - Several document question and answering products leverage on OpenAI's GPT4 as the LLM for knowledge reasoning. The downside to this is an OpenAI API key (not free) is needed
  - Most products host their solution on the cloud
- **GuruDocs Opportunities:**
  - Our product targets a specific user market – corporate staff with confidential documents where data cannot exit their network

Market Research

- Comparison of GuruDocs product features with commercial product offerings

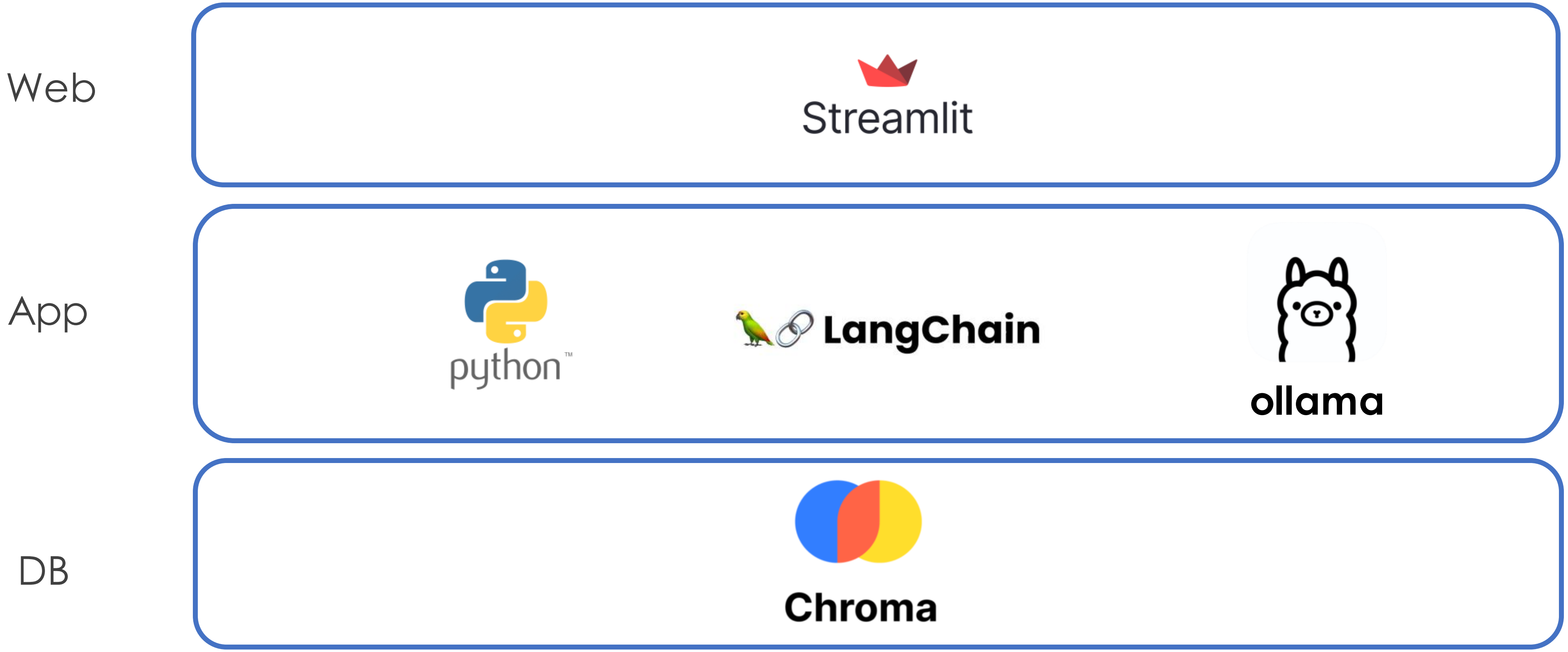
	GuruDocs	ChatGPT's AskAI	CHATDOC	Hypotenuse AI	DocBots AI
Multiple Document Query	✓	✗	✓	✓	✓
Unlimited number of documents upload	✓	✗	✗	✗	✗
Unlimited number of queries	✓	✗	✗	✓	✗
Unlimited number of users	✓	✗	✓	✗	✗
Free	✓	✗	✗	✗	✗
Able to work with confidential data	✓	✗	✗	✗	✗

https://askaichat.app/
https://chatdoc.com/
https://www.hypotenuse.ai/
https://docsbot.ai/

Data Collection and Preparation

- Data required: documents relating to any company policies or workflows.
- These documents can be obtained easily via open source
  - Examples: NTUC Staff Handbook, Adobe Employee Handbook

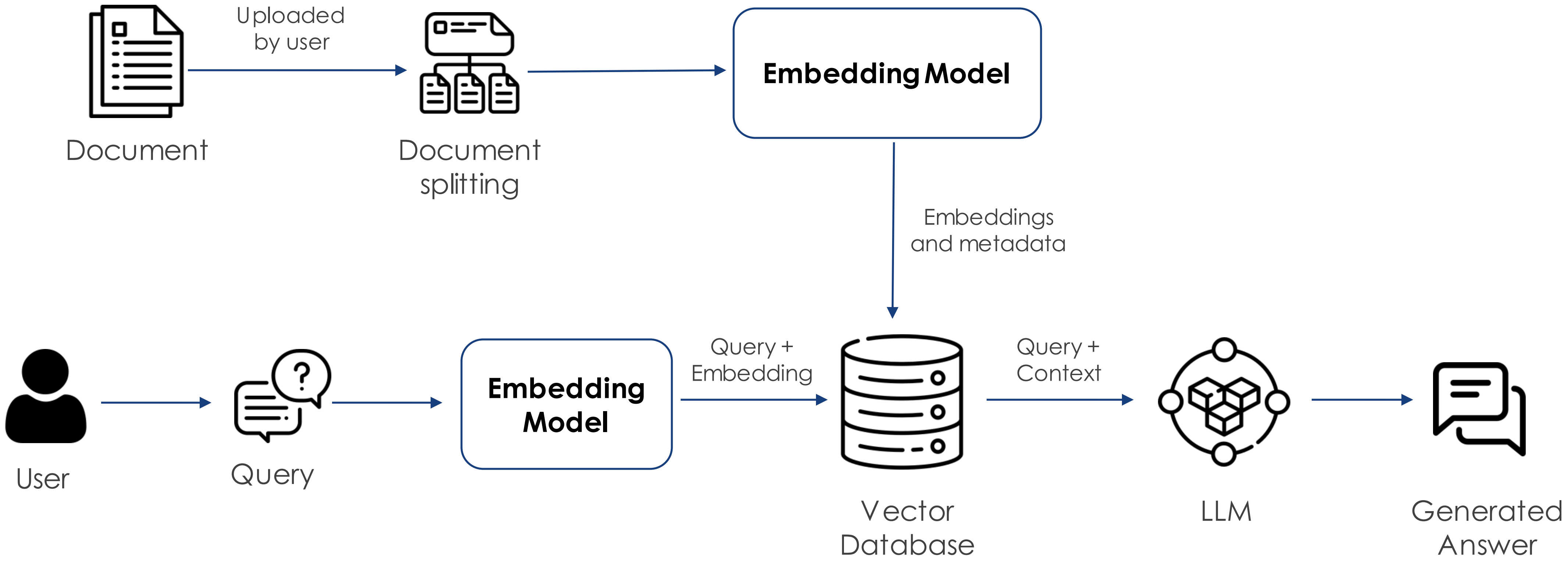
System Design (1)



*See next slide for details

- Our product stack is hosted and run locally, ensuring that we are able to process confidential documents without the data (document vectors or contexts) leaving our on-premise system
  - Leveraging Ollama to run LLMs locally, with minimal compute requirements (able to run without GPUs)

System Design (2)



Evaluation Methodology

- Evaluation will be performed on both components of RAG pipeline:
  - **Retrieval component:** evaluation to ensure that the context retrieved from vector store is correct
  - **Generator component:** evaluation to ensure that the generated answer is consistent with the information from the context and prompt
- Proposed evaluation methodology and metrics based on existing frameworks such as Retrieval Augmented Generation Frameworks (RAGAs)
  - **Methodology:** Generate ground truths of question and answer pairs, with the context information that should be retrieved.
  - **Metrics:**
    - Context precision: measures signal to noise ratio of retrieved context
    - Context recall: measures if all the relevant information required to answer the question was retrieved.
    - Faithfulness: measures the factual accuracy of the generated answer
    - Answer relevancy: measures how relevant the generated answer is to the question

https://towardsdatascience.com/evaluating-rag-applications-with-ragas-81d67b0ee31a

Product Features

Document Summarization

Automatically extracts a concise summary of uploaded document's key points and essential information

Choice of LLM

Provides a default LLM but also allowing users to choose from a selected list of models

Multi-Document Q&A

Allows users to upload and make queries across multiple documents

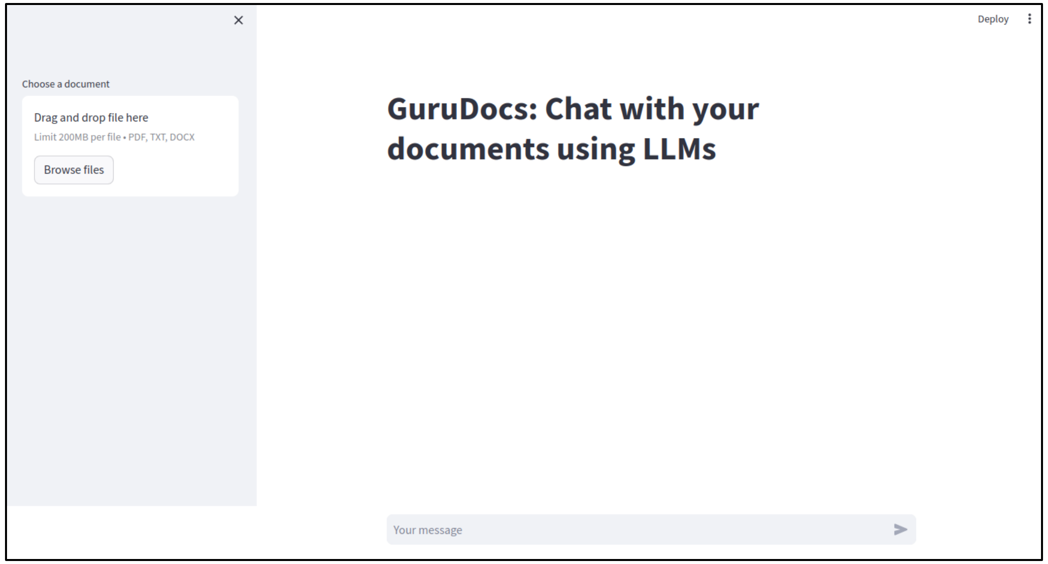
Conversation Memory

Able to understand and respond coherently to user queries by considering context of ongoing conversation based on past queries posed

Results and Progress

- **Models**
  - Prototyped a multi-document Q&A chatbot with using OpenAI's GPT model using LangChain
  - Successfully switched out the use of OpenAI's GPT model with Ollama's model library using LangChain

- **Frontend**
  - Developed an initial prototype for frontend UI



GuruDocs: Chat with your documents using LLMs

Future Work

- Expand beyond text-based capabilities to multi-modal understanding:
  - Allow document inputs that go beyond text (PDF, .docx) to include multi-modal understanding (e.g. image, PPT)
  - Using knowledge graphs to visualize content wherever possible
- User authentication and management to prevent data leakage
- Implement trust and safety guardrails

Conclusion

- GuruDocs is an on-premise chatbot for documents question and answering by providing context-aware search results. It leverages on state-of-the-art open-source tools and models i.e., LangChain, ChromaDB, Ollama.
- GuruDocs' unique selling point is that it is free and capable of handling confidential data. Other benefits include ability to query multiple documents, have unlimited number of queries, users, and document uploads.