ThinkAi

CHAT WITH YOUR NOTES AND IDEAS

Project ID

PG20CS25

MEET THE TEAM

Project Guide

Dr Anver SR

Team

Hubaib P *KSD20CS056*

Mohammed Junaid MK *KSD20CS065*

Muhammed Shaheer *KSD20CS076*

Nandana KV *KSD20CS077*

TABLE OF CONTENTS

INTODUCTION

- Students and researchers often struggle to efficiently manage, analyze, and extract key insights from their notes and research materials. Existing solutions require users to manually provide context and structure to their notes, hindering the process and limiting productivity. This leads to wasted time, frustration, and ultimately impedes the research process

INTODUCTION

- PDF Document Processing: Our project focuses on developing AI-powered tools to efficiently process and extract insights from PDF documents, catering to the needs of students, researchers, and professionals.

- Website Interaction with AI: We aim to enhance user experience by enabling our chatbot to engage in meaningful conversations with websites, leveraging AI algorithms to improve information retrieval and interaction.

INTODUCTION

- Image-Based Question Answering: Embracing visual cues, our chatbot intelligently interprets and responds to inquiries based on images, offering intuitive information retrieval and user engagement.

- Revolutionizing AI-Driven Tools: Through this comprehensive approach, we seek to redefine the boundaries of AI-driven solutions, providing users with efficient, intelligent, and user-friendly tools tailored to diverse information needs in the digital era.

PROBLEM STATEMENT

PROBLEM STATEMENT

Develop an AI chatbot leverages the power of RAG and LangChain that efficiently processes PDF documents, interacts with websites, summarizes YouTube videos, and answers questions based on image inputs, offering enhanced functionality and user experience.

F E A T U R E S

- **Intelligent Document Interaction**: The chatbot can analyze and understand documents (PDFs), websites, YouTube transcripts, and images, enabling users to ask questions and extract valuable insights from various content sources.

- **Conversational Memory**: The chatbot remembers previous interactions within a session, allowing for more natural and contextually relevant conversations. It can build upon past questions and answers for a more coherent user experience.

- **Sub-question Decomposition:** For complex or multi-faceted queries, the chatbot can break them down into smaller, more manageable sub-questions, ensuring more focused and accurate answers.

F E A T U R E S

- **Chat Interface:** Develop a chat interface to create a user-friendly chat interface for interacting with their data to answer question about their given Data

- **Prompt Templates:** These are predefined structures or formats for generating responses. This feature helps the chatbot create coherent and contextually relevant answers by using templates tailored to different types of queries.

- **RAG (Retrieval Augmented Generation)** as a crucial technique that improves the accuracy of retrieval and Provides correct answer user have queried

RETRIEVAL AUGMENTED GENENERATION (RAG)

L L M

- A large language model (LLM) is a type of artificial intelligence (AI) program that They can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. LLMs are trained on huge sets of data hence the name "large“.

D R A W B A C K S O F L L M

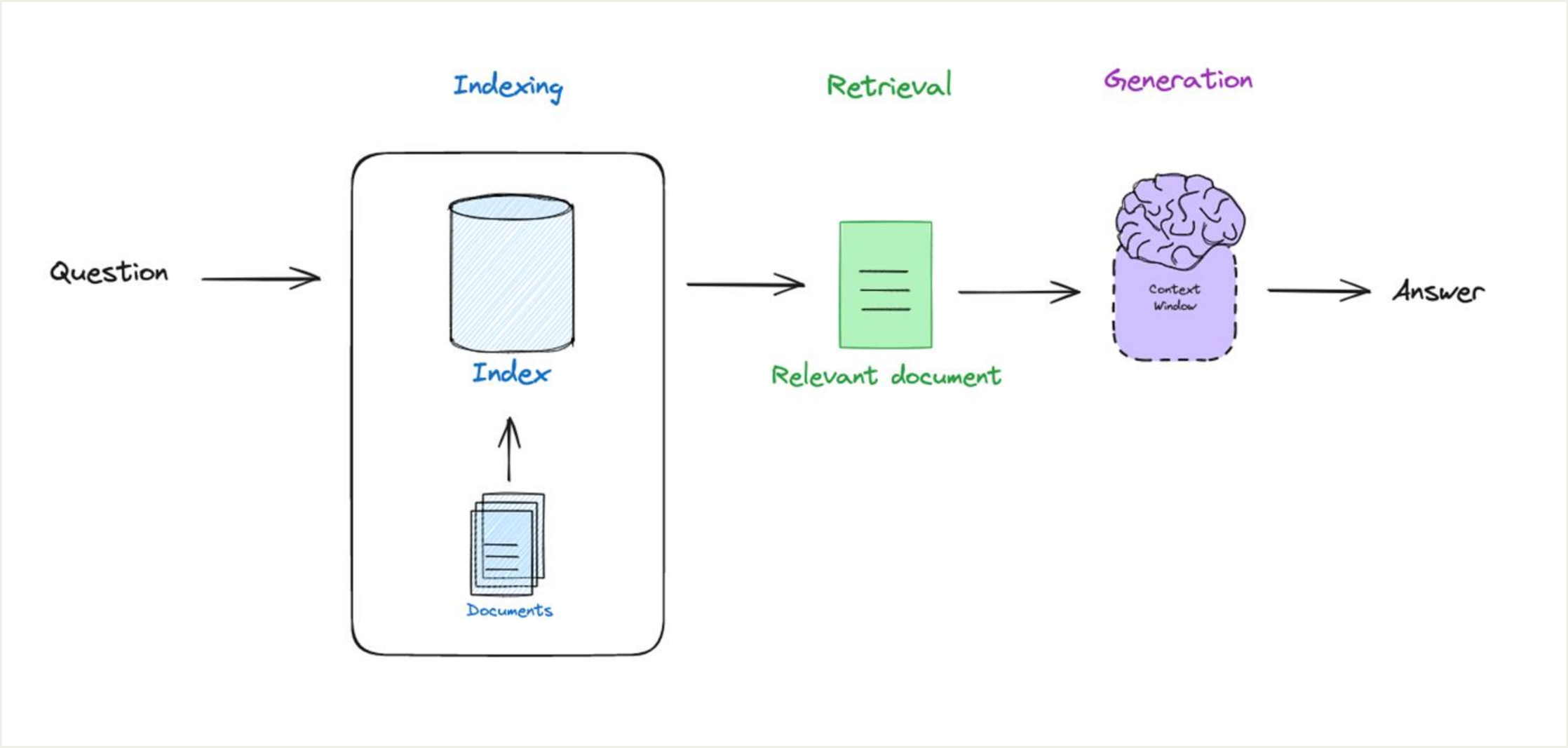
- Lack of source
- Not up to date
- Hallucination

R A G

- Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of LLM with facts fetched from external sources.

B E N E F I T S O F R A G

- Increased Accuracy
- Up-to-date Information
- Answer based on source


Indexing
Retrieval
Generation
Question
Index
Documents
Relevant document
Context Window
Answer

LITERATURE SURVEY

SI.NO	AUTHOR	TITLE		LIMITATIONS
1	Addi Ait-Mlouk and Lili Jiang.	KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding OverLinked Data[2020].	❖ KBot supports interaction in multiple languages, enhancing accessibility. ❖ Inclusion of myPersonality dataset enriches analytical query capabilities for social science research. ❖ Flexible approach combining semantic web, knowledge graphs, and machine learning for adaptability. ❖ KBot is designed for providing real-time responses to enhance user experience.	❖ Lacks details on how training data is effectively addressed. ❖ Limited details on privacy preservation techniques, especially concerning myPersonality dataset. ❖ Limited discussion on scalability challenges and handling increasing query volumes. ❖ Article briefly mentions storing feedback for continuous learning without detailed implementation insights. ❖ Lack of advanced privacy-preserving methods for handling sensitive data.

SI.NO	AUTHOR	TITLE	REMARKS	LIMITATIONS
2	Aditya Jain, Divij Bhatia and Manish K Thakur.	Extractive Text Summarization using Word Vector Embedding[2017].	❖ Active research field of text summarization, crucial for handling large documents in diverse domains. ❖ Active research field of text summarization, crucial for handling large documents in diverse domains. ❖ The proposed approach combines feature extraction and neural network for supervised extractive summarization, demonstrating effectiveness on the DUC 2002 dataset. ❖ A Multi-Layer Perceptron (MLP) with three fully connected hidden layers is employed for training, enhancing the predictive capabilities of the summarizer.	❖ The paper suggests potential improvement by increasing the size and diversity of the training dataset. This could impact the generalizability of the model. ❖ Specific methods are not detailed in abstractive to extractive conversation. ❖ The impact of training bias mitigation on model bias is not extensively discussed.

SI.NO	AUTHOR	TITLE	REMARKS	LIMITATIONS
3	Zhao Yan,Nan Duan,Junwei Bao,Peng Chen,Ming Zhuzhou Li and Jianshe Zhou.	DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents[2016].	❖ Introduces DocChat, a unique information retrieval method for chatbot engines using unstructured documents. ❖ Utilises a learning to rank model with features at different levels of granularity to measure the relevance between utterances and responses. ❖ Demonstrates good adaptability in English and Chinese evaluations, showcasing its effectiveness across languages.	❖ Focuses on a simplified short text conversation (STC) task, which might not cover the complexity of multi-turn conversations. ❖ The generalizability to diverse domains is not extensively discussed. ❖ Limiting application in scenarios requiring multiple rounds of interaction. ❖ Relies on pre-existing documents, which may pose challenges in domains where obtaining a comprehensive set of relevant documents is difficult.

SI.NO	AUTHOR	TITLE	REMARKS	LIMITATIONS
4	Arjun Pesaru, Taranveer Singh Gill and Archit Reddy Tangella.	AI Assistant For Document Management Using LangChain And Pinecone [2023].	❖ The paper effectively utilises LangChain and the LLM Model to develop a PDF chatbot, showcasing the potential of these technologies in natural language processing. ❖ Integration of Pinecone for storing PDF vectors enhances the efficiency of document retrieval, offering practical applications in research and customer support. ❖ The choice of React JS for the front end contributes to the user-friendly and efficient interaction with the chatbot	❖ The paper lacks detailed information on the training process of the chatbot, including the size and diversity of the PDF dataset used. ❖ While the use of LLM is highlighted, specific details on its configuration and fine-tuning for the PDF-related tasks are missing. ❖ The evaluation section mentions high accuracy but lacks specific metrics or benchmarks used to measure the chatbot's performance. ❖ Future work suggestions are broad; specifying targeted improvements and expansion areas would provide more direction for follow-up research. ❖ The comparison with the existing rule-based system could be more nuanced, considering specific use cases where one approach might outperform the other.

5	Oguzhan Topsakal1*, and Tahir Cetin Akini	Creating Large Language Model Applications Utilizing langchain: A Primer on Developing LLM Ap ps Fast	Emphasis on nlp advances--The text appropriately emphasizes the transformative impact of advanced nlp models, including bert, gpt, and t5. The mention of these models and their parameters helps readers understand the scale of advancements in language processing capabilities. Introduction of langchain--the introduction of langchain as an open-source software library providing solutions for developing custom ai applications utilizing llms is a valuable addition. It bridges the gap between advanced ai models and practical application development, making ai accessible to a broader audience	Training and deploying llms with billions of parameters, such as gpt-4, require substantial computational resources, leading to high costs and environmental concerns. These resource-intensive models may not be accessible to smaller organizations or projects with limited computing capabilities. Interpretability and explainability--Llms, especially those with numerous parameters, can be complex and lack interpretability. Understanding how these models arrive at specific outputs can be challenging, raising concerns about transparency and accountability, especially in critical appcations

6	Andreas Lommatz sch and Jonas Katins	An Information Retrieval-based Approach for Building Intuitive Chatbots for Large Knowledge Bases[2019].	Saves time and resources by using existing knowledge bases, providing an adaptation advantage. Enhances user experience with a focus on natural dialogs, improving interaction with chatbots. Aims to overcome current system weaknesses like scalability issues and limited support for complex dialogs.	Converting text databases to ontology- based stores is complex and time- consuming. Creating ontologies can be expensive, requiring abstract knowledge, limiting feasibility for budget- constrained projects. Not ideal for large knowledge collections, emphasizing the need for tailored solutions in such cases.

7	Arjun Pesaru, Taranveer Singh Gill and Archit Reddy Tangella.	Al Assistant For Document Management Using Lang Chain And Pinecone [2023].	The paper effectively utilises LangChain and the LLM Model to develop a PDF chatbot, showcasing the potential of these technologies in natural language processing. Integration of Pinecone for storing PDF vectors enhances the efficiency of document retrieval, offering practical applications in research and customer support.	The paper lacks detailed information on the training process of the chatbot, including the size and diversity of the PDF dataset used. While the use of LLM is highlighted, specific details on its configuration and fine- tuning for the PDF-related tasks are missing. The evaluation section mentions high accuracy but lacks specific metrics or benchmarks used to measure the chatbot's performance. Future work suggestions are broad; specifying targeted improvements and expansion areas would provide more direction for follow-up research.

8	Aditya Jain, Divij Bhatia Manish K Thakur.	Extractive Text Summarization using Word Vector Embedding[2017].	Active research field of text summarization, crucial for handling large documents in diverse domains. Active research field of text summarization, crucial for handling large documents in diverse domains. The proposed approach combines feature extraction and neural network for supervised extractive summarization, demonstrating effectiveness on the DUC 2002 dataset	The paper suggests potential improvement by increasing the size and diversity of the training dataset. This could impact the generalizability of the model. Specific methods are not detailed in abstractive to extractive conversation. The impact of training bias mitigation on model bias is not extensively discussed.

9	Addi Ait-Mlouk and Lili Jiang.	KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding OverLinked Data[2020].	KBot supports interaction in multiple languages, enhancing accessibility. Flexible approach combining semantic web, knowledge graphs, and machine learning for adaptability. KBot is designed for providing real-time responses to enhance user experience.	Lacks details on how training data is effectively addressed. Limited details on privacy preservation techniques, especially concerning myPersonality dataset. Limited discussion on scalability challenges and handling increasing query volumes.

10	Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhuzhou Li and Jianshe Zhou.	DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents[2016]	Introduces DocChat, a unique information retrieval method for chatbot engines using unstructured documents. Utilises a learning to rank model with features at different levels of granularity to measure the relevance between utterances and responses	Focuses on a simplified short text conversation (STC) task, which might not cover the complexity of multi- turn conversations. The generalizability to diverse domains is not extensively discussed. Limiting application in scenarios requiring multiple rounds of interaction.

EXISTING WORK

EXISTING APPROCH

- **Manual Context Creation**: Many solutions require users to manually provide context and structure to their notes and documents. This can be time-consuming, tedious, and prone to errors.

- **Limited Contextual Understanding**: Existing chatbots often struggle to understand the full context of user queries, especially when dealing with information from multiple sources. This can lead to irrelevant or inaccurate responses.

- **3. Difficulty in User Query Formulation:** The way a question is phrased significantly impacts the response. Basic chatbots may not be able to interpret nuanced user queries, limiting the effectiveness of LLMs that rely on clear and specific questions.

LIMITATION

- **Inefficient**: Manual context creation is a significant bottleneck, hindering user productivity and research efficiency.

- **Inaccurate**: Limited contextual understanding can lead to irrelevant or inaccurate responses, frustrating users and hindering research progress.

PROPOSED WORK

ADVANTAGES

- **increased Efficiency**: Automatic context creation saves users time and effort, improving research productivity.
- **Improved Accuracy**: Deep contextual understanding leads to more relevant and informative responses, enhancing knowledge discovery.
- **Enhanced Privacy** and Security: Local execution eliminates the risks associated with centralized data storage.
- **Greater User Control and Customization**: Open-source nature allows users to tailor the chatbot to their specific needs and preferences.

EXPECTED USERS

- Students and researchers
- Professionals
- Individuals seeking privacy-conscious information access

IMPACT

- Revolutionize information access and understanding
- Enhance research and learning efficiency
- Empower users with privacy-focused knowledge exploration

METHODOLOGY

METHODOLOGY

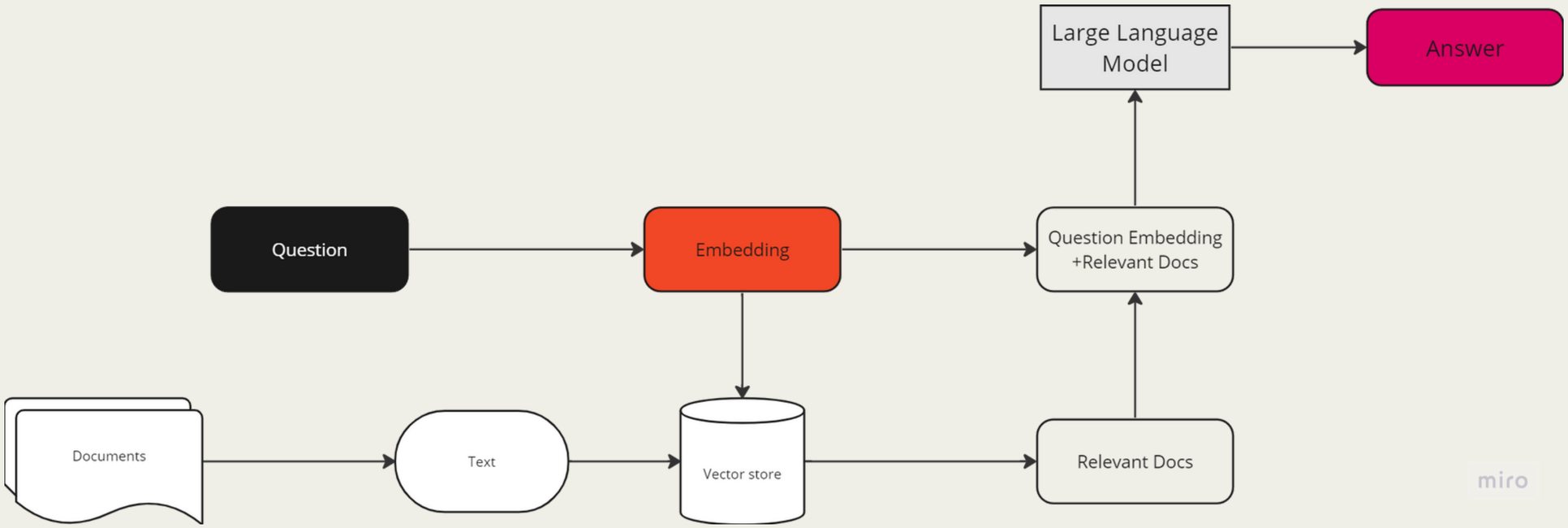
- Data Acquisition and Loading
  - PDFs: PyPDFLoader from langchain_community.document_loaders
  - YouTube: YoutubeLoader from langchain_community.document_loaders (make sure to handle cases of no transcripts)
  - Websites: WebBaseLoader from langchain_community.document_loaders
  - Images: Google's Gemini Pro Vision API (handle image uploads, temporary storage, and API calls)
- Text Processing and Chunking
  - Universal Approach: Use RecursiveCharacterTextSplitter from langchain.text_splitter for PDFs, YouTube transcripts, and websites to split content into manageable chunks. This ensures consistency.
  - Image Handling: For images, your "text processing" involves sending the image to the Gemini Pro Vision API for analysis and question answering.

METHODOLOGY

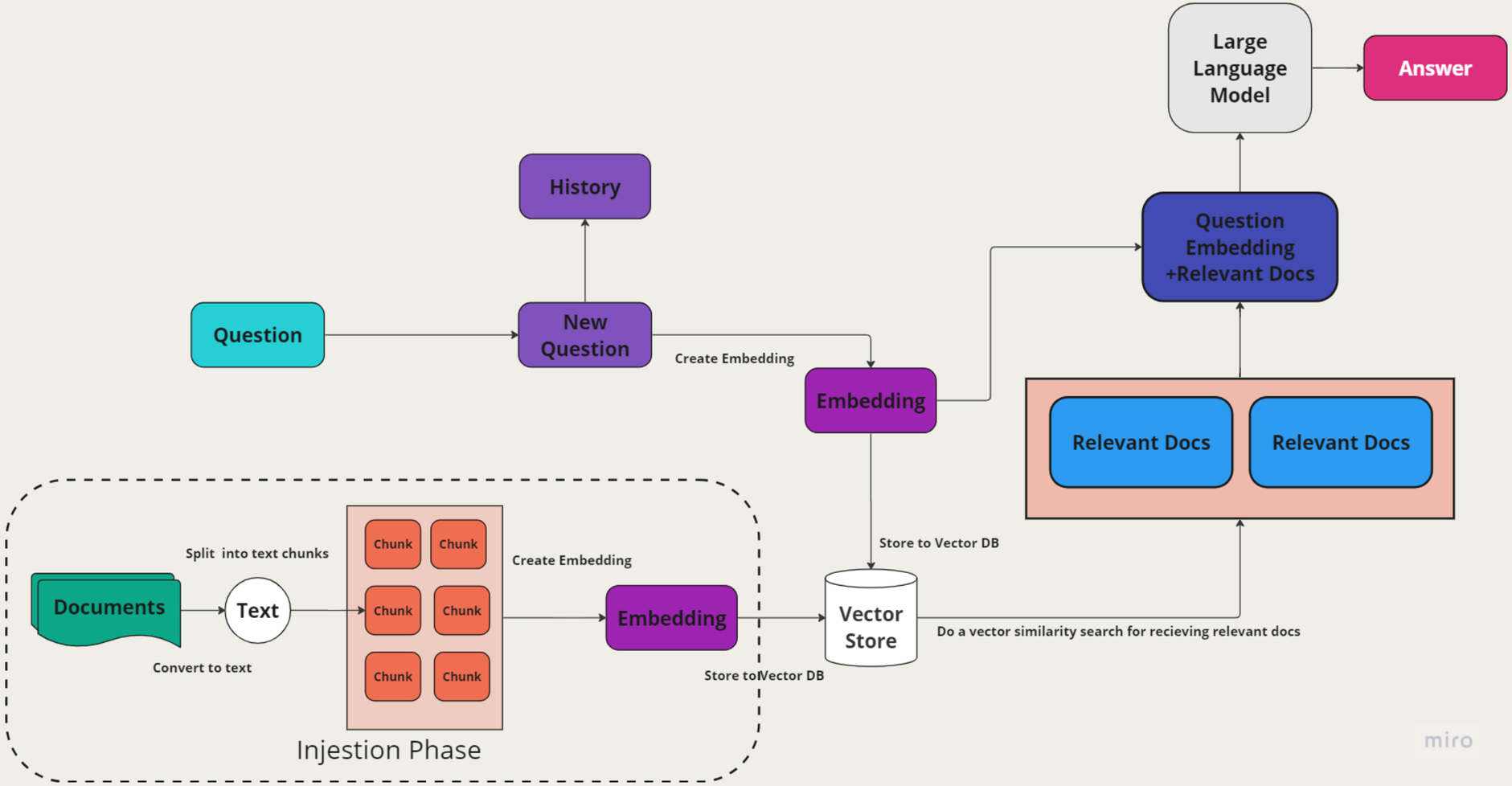
- Embeddings and Vector Storage
  - Embeddings: Use a robust embedding model. to Embed the text to Vector and Store it in Vector DB
  - Vector Database: Chroma DB is Vector DB which can store Vector and do Operations such as Similarity Search For Retrival.
- Retrieval and Question Answering
  - Retrieval Chain: Construct a history-aware retrieval chain using create_history_aware_retriever from langchain.chains. This allows the chatbot to consider past interactions for better context.
  - Conversational RAG Chain: Build a ConversationalRetrievalChain (from langchain.chains).
    - Use a prompt template that feeds the retrieved context to the language model.
    - Include conversation history in the prompt for more natural interactions.

SYSTEM ARCHITECTURE

OVERALL ARCHITECTURE

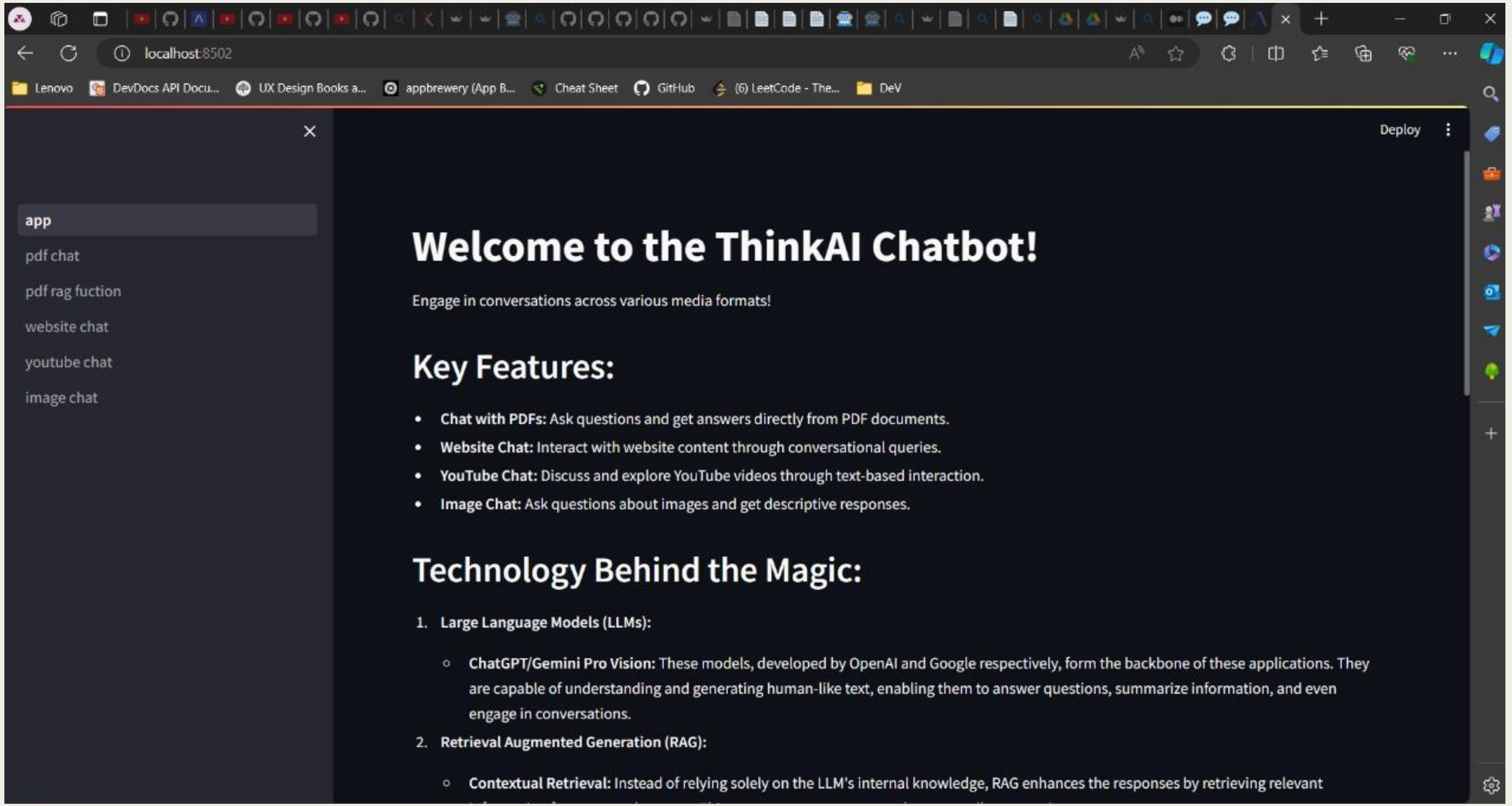


DETAILED ARCHITECTURE

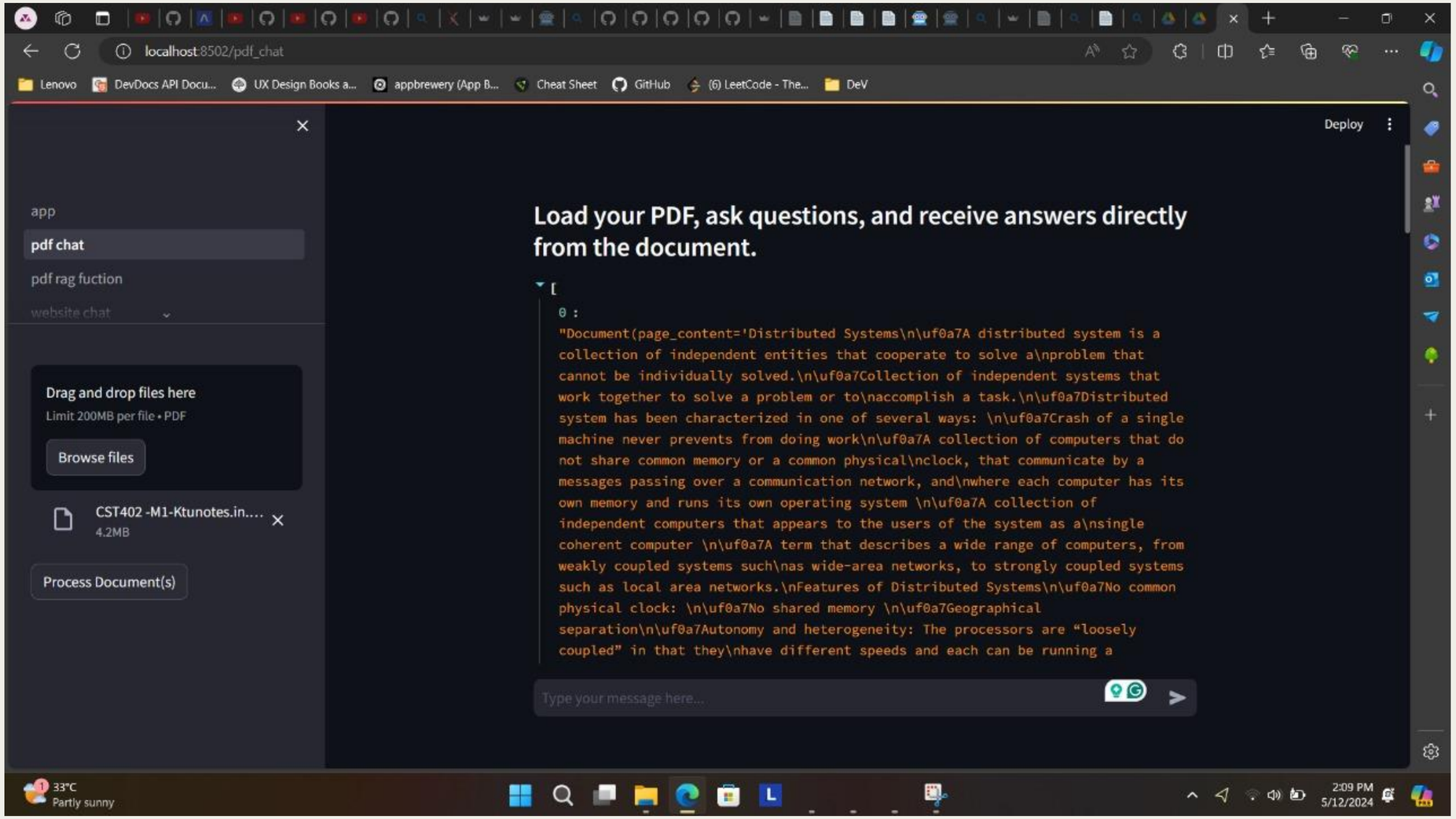


RESULTS

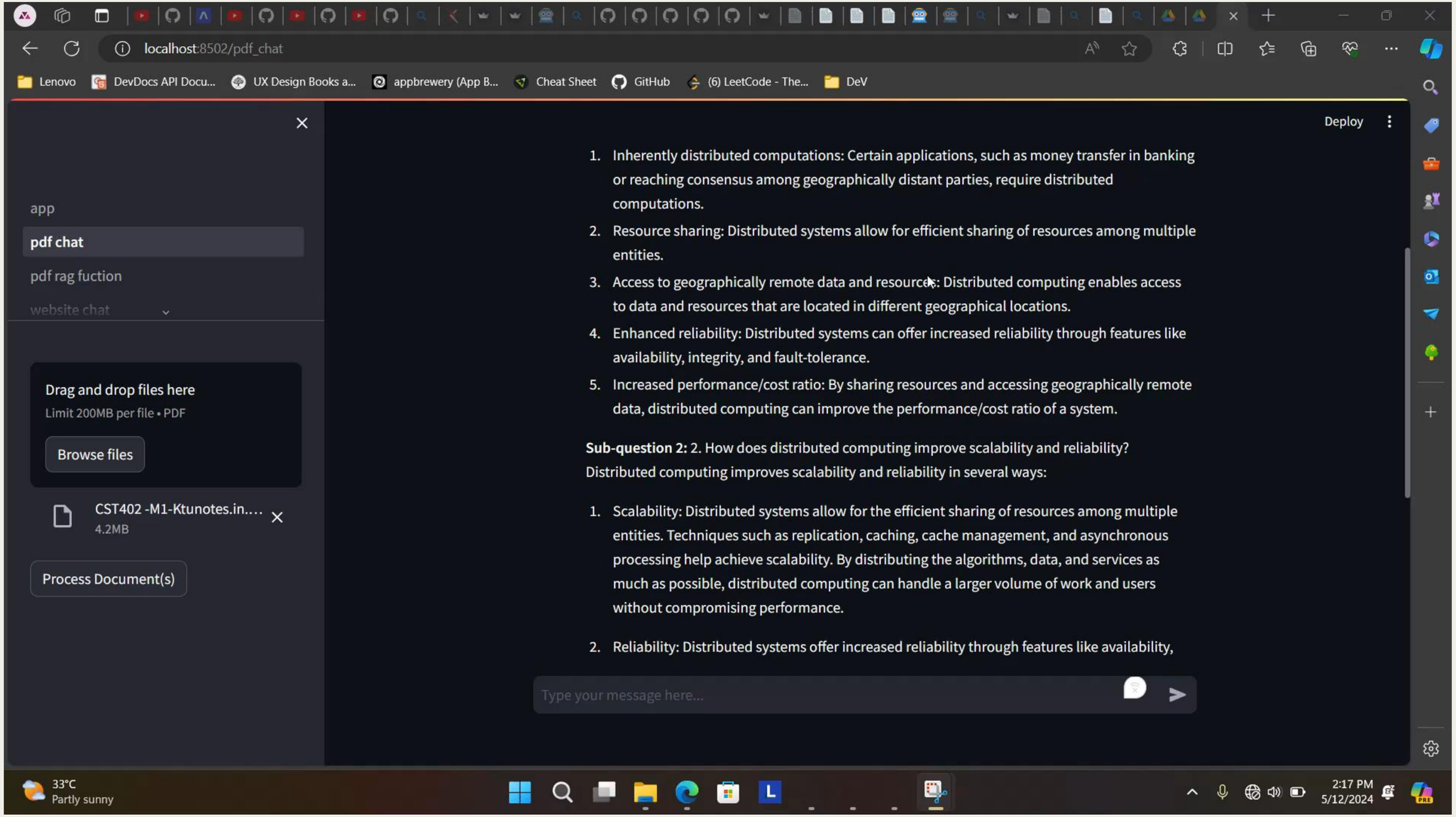
Sample Images



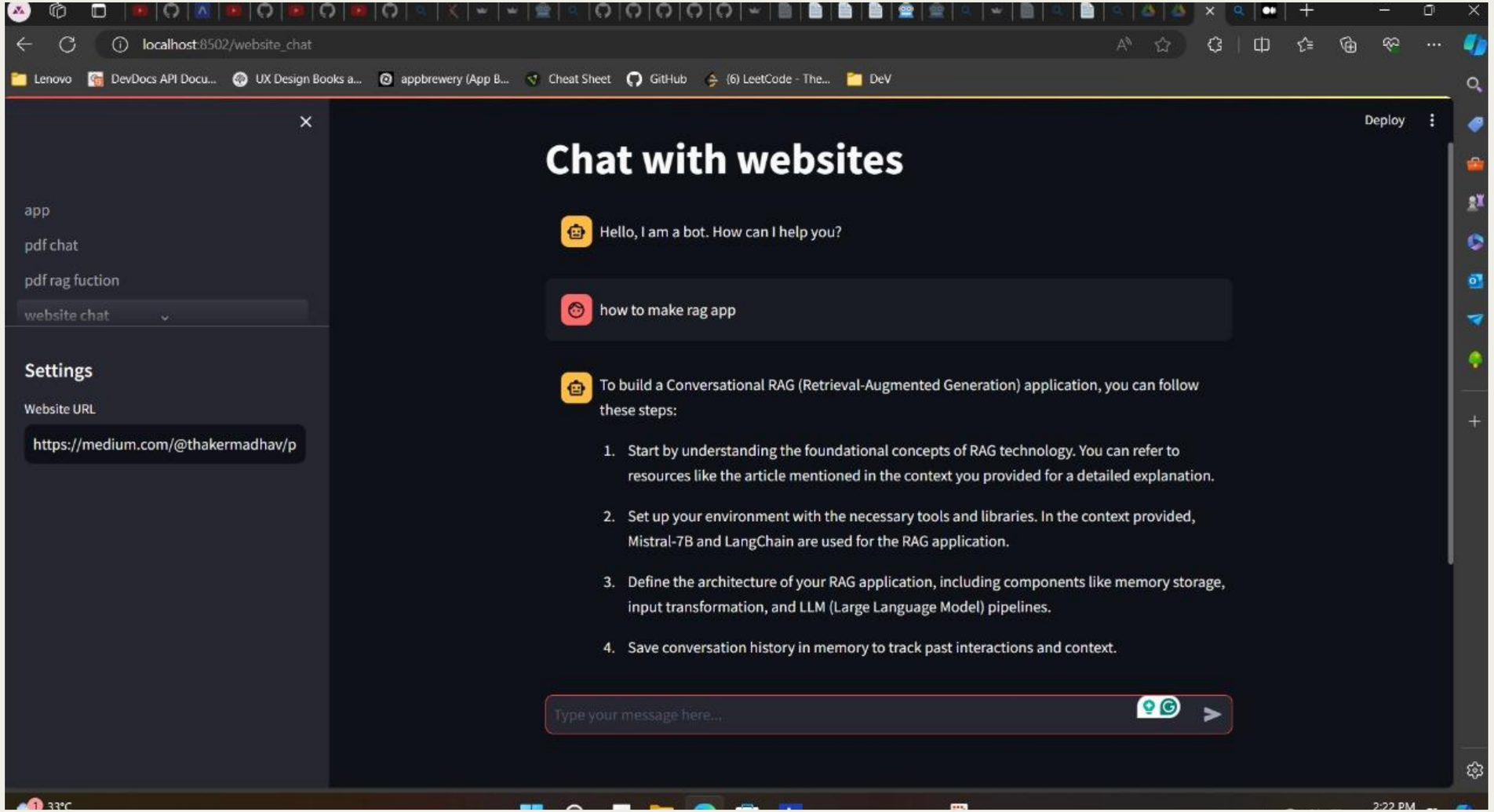
Sample Images



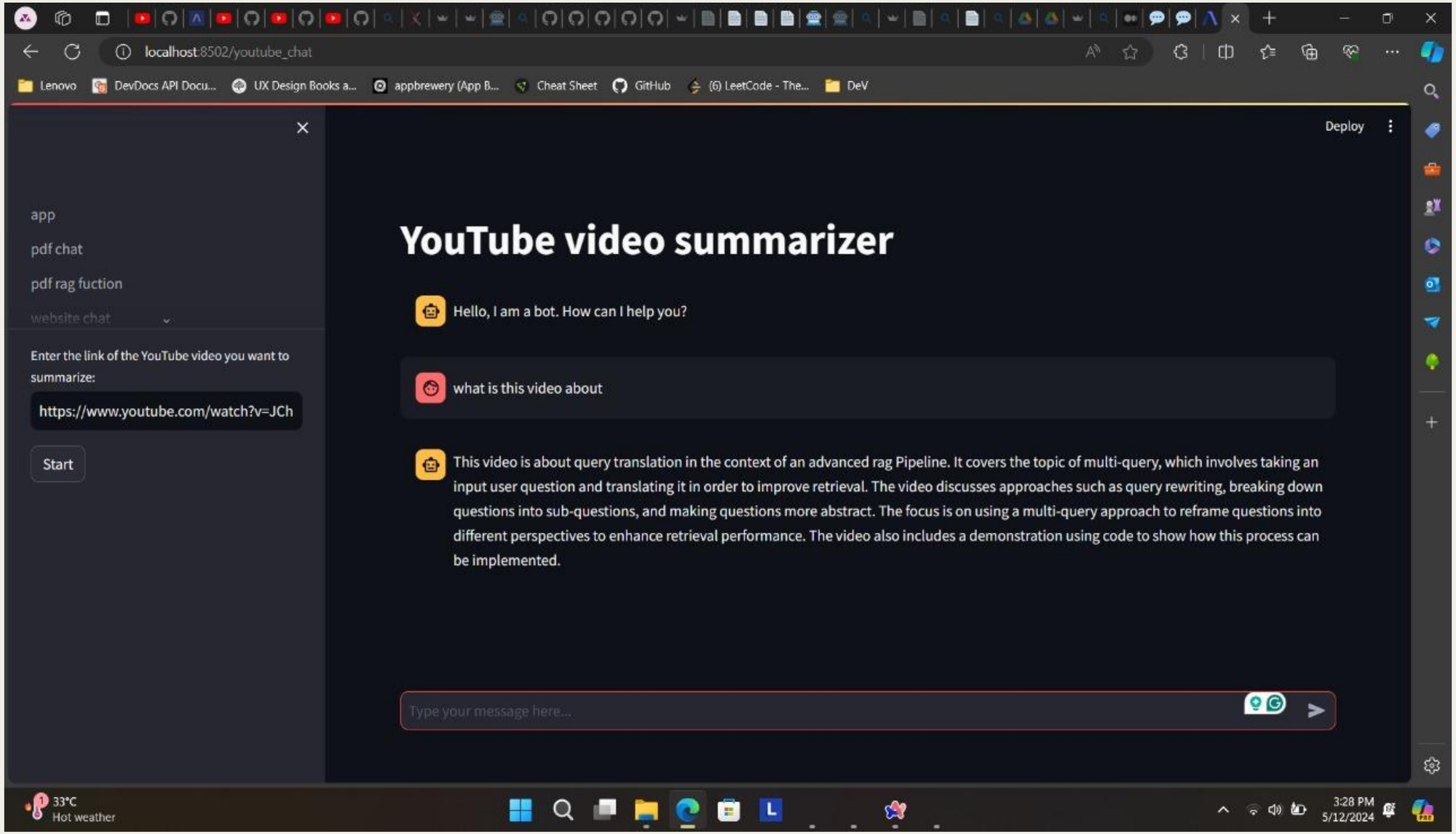
Sample Images



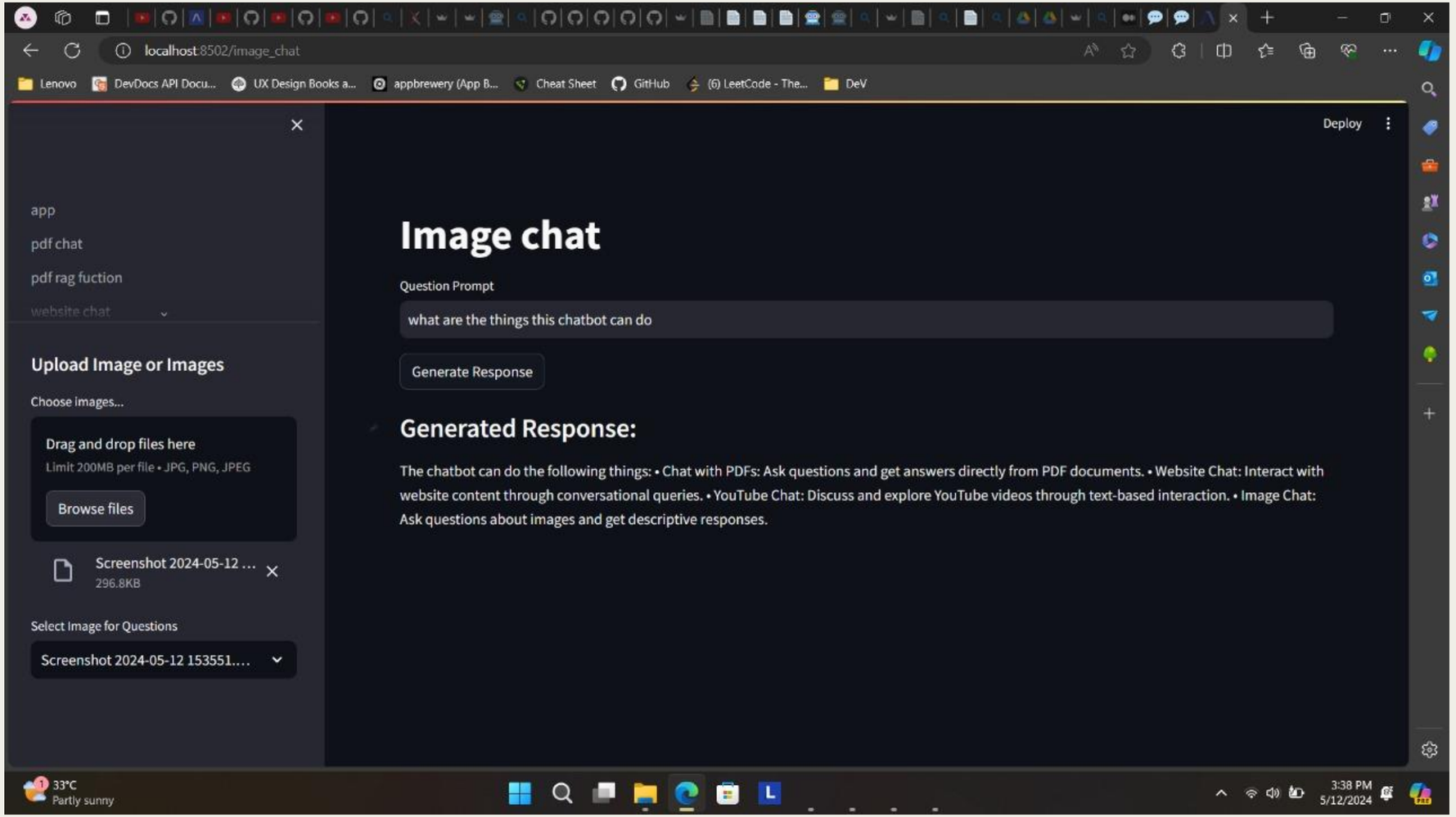
Sample Images



Sample Images



Sample Images



CONCLUSION

CONCLUSION

1. Our project pioneers AI-driven technology for PDF document processing, revolutionizing how users interact with textual data.

2.Integration of AI in website chatting enables seamless communication and efficient information retrieval, enhancing knowledge acquisition.

3.The inclusion of a YouTube video summarizer saves users time by providing concise, informative summaries, showcasing the versatility of our AI capabilities.

4.Our image-based question answering feature highlights the adaptability of our AI system, catering to diverse user needs and enhancing productivity across digital platforms.
of body text

REFERENCE

[1] Aditya Jain, Divij Bhatia, Manish K Thakur [2017], Extractive Text Summarization using Word Vector Embedding , https://ieeexplore.ieee.org/document/8320258

[2] Addi Ait-mlouk AND Lili Jiang [2020], KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data , https://ieeexplore.ieee.org

[3] Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhou Li, Jianshe Zhou [2016] , DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents, https://aclanthology.org/P16-1049.pdf

[4] Arjun Pesaru, Taranveer Singh Gill, Archit Reddy Tangella [2023] , AI ASSISTANT FOR DOCUMENT MANAGEMENT USING LANG CHAIN AND PINECONE , https://www.irjmets.com

[5] Haritha Akkineni, P. V. S. Lakshmi, and Lasya Sarada [2022] , Design and Development of Retrieval-Based Chatbot Using Sentence Similarity, https://link.springer.com

REFERENCE

6] Oguzhan Topsakal1, and Tahir Cetin Akinci [2023] , Creating Large Language Model Applications Utilizing LangChain: A
Primer on Developing LLM Apps Fast , https://as-proceeding.com

[7] Andreas Lommatzsch and Jonas Katins [2019] , An Information Retrieval-based Approach for Building Intuitive Chatbots for Large
Knowledge Bases , https://ceur-ws.org/Vol-2454/paper_60.pdf

[8] Norbert Braunschweiler and Rama Doddipatla and Simon Keizer and Svetlana Stoyanchev[2023], Evaluating Large Language Models for Document-grounded Response Generation in Information-Seeking Dialogues , https://arxiv.org

[9] Arjun Pesaru, Taranveer Singh Gill, Archit Reddy Tangella [2023] , AI ASSISTANT FOR DOCUMENT MANAGEMENT USING LANG
CHAIN AND PINECONE , https://www.irjmets.com

[10] Pinky Sitikhu,Kritish Pahi,Pujan Thapa,Subarna Shakya[2019] , A Comparison of Semantic Similarity Methods for
Maximum Human Interpretability, https://arxiv.org/abs

Thank you!

ANY QUESTIONS ?