PPSU
P P SAVANI UNIVERSITY
School of Engineering

Minor Project/Major Project

On

CHAT WITH PDF

Academic Year: 2023-24

Student's Full Name	KASHISH PATEL , CHILLIMUNTHA JOTHSNA SRI KATHYAYANI, SRUSHTI DOBARIYA
Enrollment No	21SE02ML022, 21SE02ML006, 21SE02ML010
Branch	AIML
Semester	6th

Supervised by

Robin Hojiwala

P P Savani School of Engineering

P P SAVANI UNIVERSITY

School of Engineering

CERTIFICATE

This is to certify that Ms. ____KASHISH PATEL____, Enrollment No. ____21SE02ML022____ from the Department of ________ENGINEERING________, has successfully completed the Minor Project on the **CHAT WITH PDF** during Academic Year 2023-24.

Date:

____________________________ ____________________________

Name and Sign of Supervisor Dean, SOE

PPSU
P P SAVANI UNIVERSITY

School of Engineering

CERTIFICATE

This is to certify that Ms. _____SRUSHTI DOBARIYA____, Enrollment No. ____21SE02ML010____ from the Department of _________ENGINEERING_________, has successfully completed the Minor Project on the **CHAT WITH PDF** during Academic Year 2023-24.

Date:

____

____

Name and Sign of Supervisor Dean, SOE

PPSU

P P SAVANI UNIVERSITY

School of Engineering

CERTIFICATE

This is to certify that Ms. __CHILLIMUNTHA JOTHSNA SRI KATHYAYANI____, Enrollment No. ____21SE02ML006____ from the Department of _________ENGINEERING_________, has successfully completed the Minor Project on the **CHAT WITH PDF** during Academic Year 2023-24.

Date:

____

____

Name and Sign of Supervisor Dean, SOE

ACKNOWLEDGEMENT

We feel elated in manifesting our sense of gratitude to our project guide Mr. Robin Hojiwala. He has been a constant source of inspiration for us and we are very deeply thankful to him for his support and valuable advice

We extremely grateful to our Departmental staff members, Lab technicians and Non-teaching staff members for their extreme help throughout our project.

It is indeed with a great pleasure and immense sense of gratitude that we acknowledge the help of these individuals. We are highly indebted to our Dean **Dr. Niraj Shah**, Dean, School of Engineering, P P Savani University, for the facilities provided to accomplish this MINOR PROJECT.

Finally, we express our thanks to all of our friends who helped us in successful completion of this project.

Student Names and Enrollment No.

KASHISH PATEL - 21SE02ML022

SRUSHTI DOBARIYA – 21SE02ML010

CHILLIMUNTHA JOTHSNA SRI KATHYAYANI – 21SE02ML006

ABSTRACT

This project describes the creation of a complex Chatbot meant to improve interaction with PDF documents by utilizing the Llama (Large Language Model) and RAG (Retrieval-Augmented Generation) methodology. Using the Python programming language, the Chatbot takes a new approach to document management by processing PDF inputs and offering exact responses to user queries based on the content learned from these documents.

The integration of the Llama model, which enables a thorough comprehension of natural language and permits precise interpretation of user queries, forms the basis of the Chabot's operation. To supplement this, the RAG technique pulls pertinent data from an index of indexed PDF literature to improve the system's capacity to produce well-informed responses. By ensuring that users obtain answers that are correct and informative in the context in which they are used, this dual-model method greatly increases the efficiency of document handling.

At the core of RAG's functionality is the use of advanced retrieval methods, which are employed to search through vector databases for information that closely matches the context of the given prompt. This process involves converting both the query and the database contents into vector representations and identifying the best matches based on these representations. The retrieved information is then dynamically integrated into the query, providing the LLM with a rich, contextually relevant dataset to inform its response.

This project offers a useful and effective method for browsing and extracting information from PDF documents, which represents a significant improvement in the field of document interaction and management. By merging the advantages of LLaMA and RAG into an approachable chat interface, it raises the bar for document management natural language processing systems.

TABLE OF CONTENTS

LIST OF FIGURE

CHAPTER 1

INTRODUCTION TO PROJECT

OBJECTIVE OF PROJECT: The objective of the project is to build a chatbot that can communicate with PDF documents by applying the Retrieval Augmented Generation (RAG) method and the LLaMA model. The main goal is to augment conventional Large Language Models (LLMs) with real-time contextual data retrieval from vector databases, so overcoming their limits. This will improve the user experience and efficiency of document management by enabling the chatbot to respond to user inquiries with precision and relevance based on the content of PDF documents.

FRONT END TOOL:
Streamlit

BACK END TOOL:
Python

The project's inception is based on the growing need for improved interaction with PDF documents in a variety of professional and academic settings. Static interfaces and manual search methods are frequently used in traditional systems for organizing and extracting information from PDFs. These methods can be laborious and ineffective, particularly when handling big numbers of documents or looking for particular information within them. This limits the effectiveness of current document management and information retrieval systems since they cannot comprehend natural language queries or offer contextual responses from the PDF content.

In order to overcome these drawbacks, the project presents a unique chatbot that uses natural language processing to enable dynamic interaction with PDF documents. Retrieval Augmented Generation (RAG) was used because it was necessary to add real-time contextual data retrieval to Large Language Models (LLMs). This method retrieves relevant material from a vector database of indexed PDF content, which greatly improves the chatbot's capacity to respond to inquiries with accuracy and contextual relevance. Moreover, the LLAMA model bridges the gap between user inquiries and the static nature of PDF documents by assisting in the comprehension and processing of complicated natural language queries.

The decision to implement this project using the Python programming language was driven by several factors. Python is the perfect choice for this project because of its broad support for machine learning and natural language processing tools, including Hugging Face's Transformers for LLMs and numerous vector database management systems. Furthermore, the readability and broad community support of Python enable quick development and simple integration of sophisticated features like RAG and

LLAMA. By combining these technologies, the chatbot can comprehend natural language queries, extract pertinent information from PDF content, and produce responses that are pertinent and appropriate for the given context. This effectively addresses essential gaps seen in current systems.

CHAPTER 2

LITERATURE REVIEW

WHAT IS RAG?

The innovative method known as Retrieval-Augmented Generation (RAG) aims to improve machine learning models' performance, especially in tasks involving natural language processing (NLP). It offers more precise and contextually relevant solutions by fusing the capabilities of real-time information retrieval with the advantages of typical pre-trained models. RAG basically works by first extracting documents or data points from a large collection or database that are relevant to a given query. The model is then able to create answers that are not only based on its pre-trained knowledge but also informed by the most relevant and up-to-date information available by using this retrieved information as extra context for generating responses.

WHAT IS LLMs?

Large Language Models (LLMs) are another crucial aspect for current NLP applications, such as Open AI's GPT series. Large volumes of text data, learning patterns, linguistic structures, and a wealth of knowledge on a wide range of subjects are used to train these models. Because of their exceptional understanding and creation of human-like sentences, LLMs are incredibly useful for a variety of activities, ranging from complicated reasoning and content generation to writing support and conversational modelling. A synergistic advancement in NLP is made possible by the integration of RAG with LLMs, enabling replies that are deeply informed by real-time, external data, while still remaining fluent and coherent. This combination increases the value and relevance of LLMs, giving them even greater flexibility to address a wide range of language creation and comprehension problems.

Books:

“The LLM Knowledge Cookbook: From, RAG, to QLoRA, to Fine Tuning, and all the Recipes In Between! Kindle Edition” by Richard Aragon

“Retrieval Augmented Generation (RAG) AI: A Comprehensive Guide to Building and Deploying Intelligent Systems with RAG AI (AI Explorer Series)” by Et Tu Code

Online Resources:

Retrieval-Augmented Generation (RAG) Documentation: The documentation provides in-depth information on RAG, its features, and usage.

Large Language Models (LLMs) Documentation: The official LLMS documentation offers comprehensive guides, tutorils.

Limitations:

Despite the innovative approach of utilizing the LLaMA model and Retrieval-Augmented Generation (RAG) for improving interactions with PDF documents, this project faces several limitations:

Dependency on Data Quality and Availability: The efficacy of the RAG component is heavily reliant on the quality and comprehensiveness of the data within the vector databases it queries. If the indexed data is outdated, incomplete, or of poor quality, the accuracy and relevance of the responses generated can be significantly compromised.

Complex Query Understanding: While the LLaMA model offers advanced capabilities in natural language understanding, it may still struggle with highly complex or ambiguous queries. This limitation could affect the chatbot's ability to provide accurate responses in scenarios where user queries require nuanced interpretation or are outside the model's training scope.

Scalability and Performance: As the system relies on real-time data retrieval and processing, scalability could become an issue with an increasing number of users or queries. The need to maintain low latency in response generation while handling large volumes of data could pose technical challenges, impacting overall performance.

Handling Dynamic Content: The project might not effectively handle PDF documents with dynamic or interactive content, such as forms or embedded multimedia. These elements require specialized parsing and interpretation capabilities, which could be beyond the current scope of the implemented technologies.

Resource Intensiveness: The computational resources required for running advanced models like LLaMA, alongside executing real-time data retrieval with RAG, could be substantial. This might limit the deployment of the system on lower-end hardware or in environments with limited computing capabilities.

Ethical and Privacy Concerns: The project's reliance on external data sources and content retrieval raises questions about user privacy and data security. Ensuring the confidentiality of the information processed and maintaining user trust are critical challenges that need to be addressed.

Adaptability and Evolving Information: While the integration of RAG aims to provide up-to-date information by retrieving external data, the system's ability to adapt to rapidly changing information or to learn from new data patterns over time is limited. The static nature of the models' training data can hinder the chatbot's performance in dynamic fields or in response to emerging trends and terminologies not covered in its original training set.

Privacy and Security Concerns: Handling sensitive or proprietary PDF documents raises significant privacy and security issues. Ensuring that data retrieval, storage, and processing comply with data protection regulations and safeguarding against unauthorized access or data breaches require sophisticated security measures, which can complicate system design and operation.

User Experience and Interaction Limitations: Despite advancements in natural language processing, the chatbot might not always interpret user queries accurately or generate responses that meet users' expectations in terms of specificity and clarity. This could lead to user frustration and a decrease in trust and reliance on the system for critical information retrieval tasks.

CHAPTER 3

SYSTEM DESIGN AND DIAGRAMS

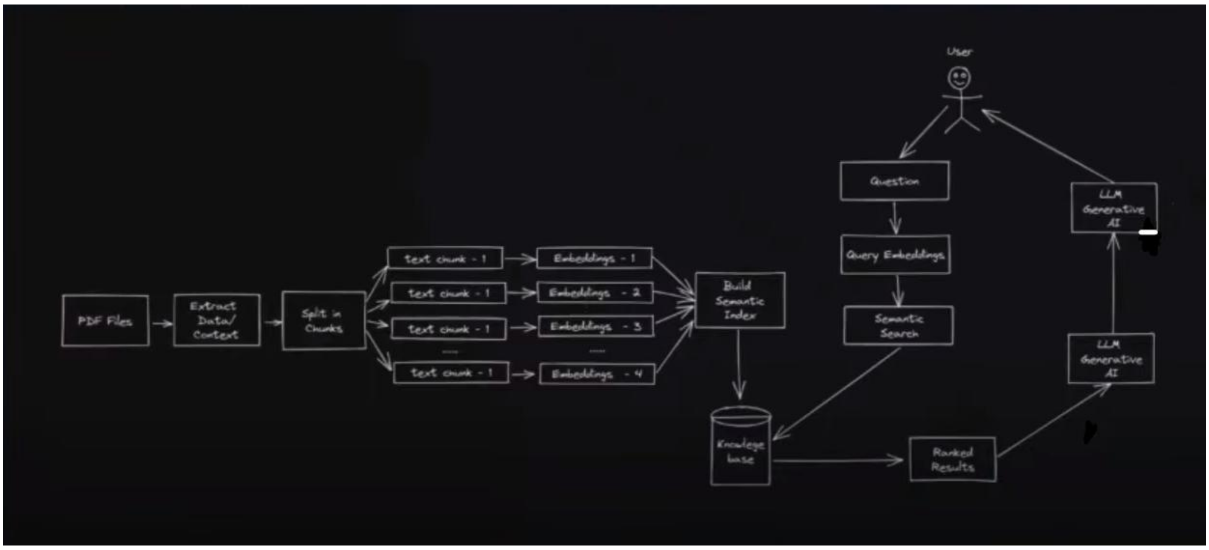


Fig 3.1 : Flow of how this works .

System design included :

Leverage Streamlit's Chat Elements: Streamlit offers chat elements like st.chat_message and st.chat_input that enables to build conversational apps. These elements allow to create chatbots using Python code within Streamlit.

Build a Basic LLM Chat App: By understanding how to use Streamlit's chat elements like st.chat_message and st.chat_input. Constructing a bot that mirrors user input to grasp the functionality of chat elements and session state for storing chat history.

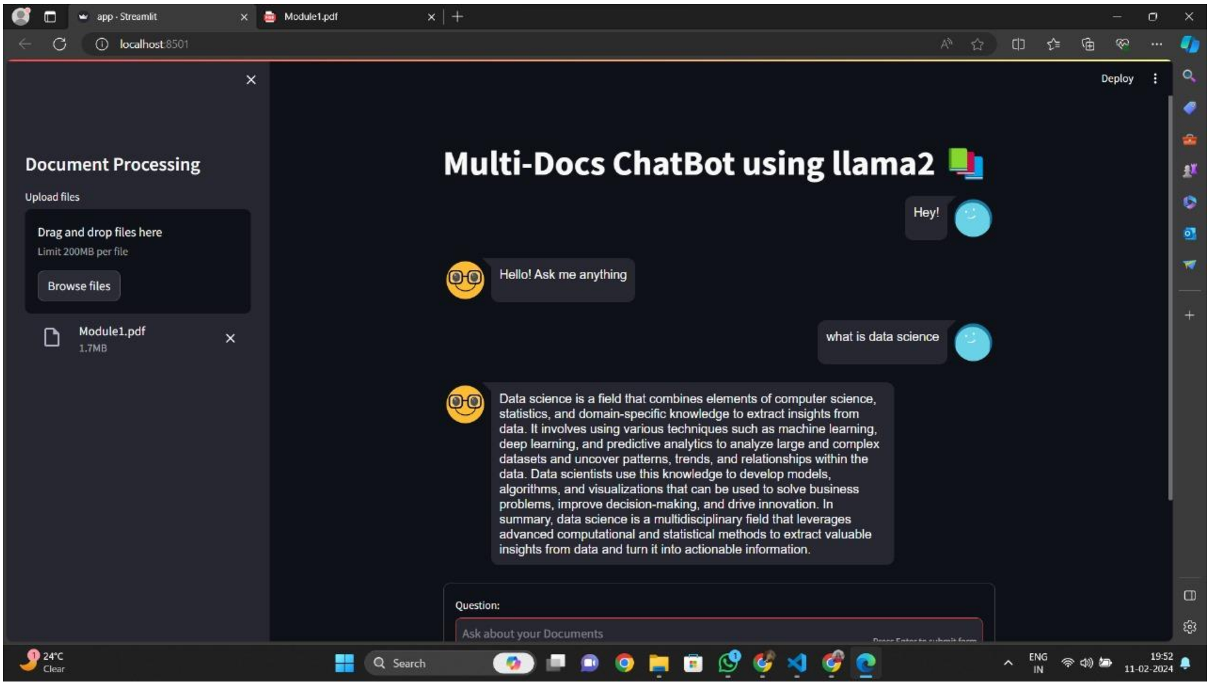
Develop a Simple Chatbot GUI: Progress to building a simple chatbot GUI with streaming capabilities. Explore how to display messages based on user input and manage conversation history within the app.

Create a ChatGPT-Like App: Advance to building a ChatGPT-like app that remembers conversational context using session state within less than 50 lines of code. Ensure to have the necessary packages installed such as openai, streamlit, and streamlit-chat for developing the app.

CHAPTER 4

IMPLEMENATION DETAILS

Output:



Multi-Docs ChatBot using llama2 📚

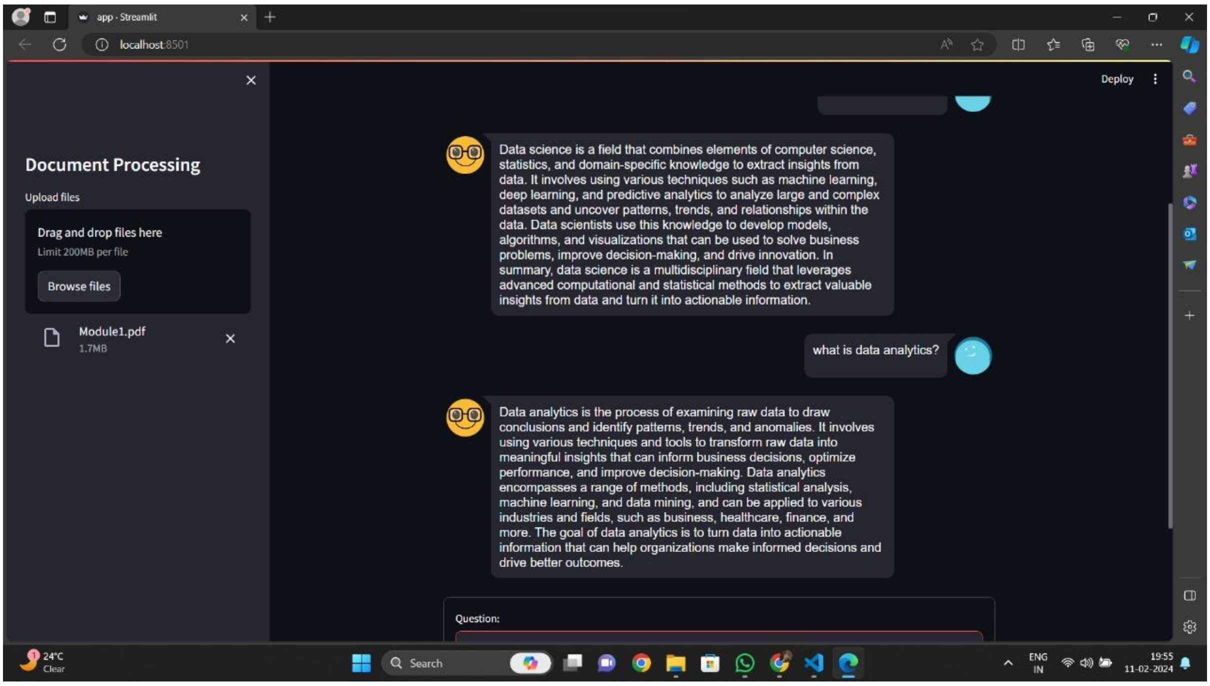
Hey!

Hello! Ask me anything

what is data science

Data science is a field that combines elements of computer science, statistics, and domain-specific knowledge to extract insights from data. It involves using various techniques such as machine learning, deep learning, and predictive analytics to analyze large and complex datasets and uncover patterns, trends, and relationships within the data. Data scientists use this knowledge to develop models, algorithms, and visualizations that can be used to solve business problems, improve decision-making, and drive innovation. In summary, data science is a multidisciplinary field that leverages advanced computational and statistical methods to extract valuable insights from data and turn it into actionable information.

Fig 4.1: way the this output looks.



Data science is a field that combines elements of computer science, statistics, and domain-specific knowledge to extract insights from data. It involves using various techniques such as machine learning, deep learning, and predictive analytics to analyze large and complex datasets and uncover patterns, trends, and relationships within the data. Data scientists use this knowledge to develop models, algorithms, and visualizations that can be used to solve business problems, improve decision-making, and drive innovation. In summary, data science is a multidisciplinary field that leverages advanced computational and statistical methods to extract valuable insights from data and turn it into actionable information.

what is data analytics?

Data analytics is the process of examining raw data to draw conclusions and identify patterns, trends, and anomalies. It involves using various techniques and tools to transform raw data into meaningful insights that can inform business decisions, optimize performance, and improve decision-making. Data analytics encompasses a range of methods, including statistical analysis, machine learning, and data mining, and can be applied to various industries and fields, such as business, healthcare, finance, and more. The goal of data analytics is to turn data into actionable information that can help organizations make informed decisions and drive better outcomes.

Fig 4.2: trying another question to check if it works

These questions were asked after considering what was uploaded, here the pdf had basic information for data science and other concepts. Multiple

Documents are allowed to be uploaded here . Easy way to work with pdf when you do not want to read the PDF but know what is the answer for the question you have in mind.

We made one improvement which was not earlier mentioned but other document formats not just PDF is supported, making it a comprehensive tool for information retrieval.

CHAPTER 5

CONCLUSION AND FUTURE WORK

CONCLUSION:

In conclusion, the project to enhance PDF interaction through the integration of LLaMA and Retrieval-Augmented Generation (RAG) marks a significant advancement in making document management more interactive and efficient. It fills in important holes in current document handling techniques by using a natural language interface to query and extract data from PDFs. This attempt creates possibilities for future advancement in AI-driven document management, despite encountering obstacles like computing needs, data quality and indexing issues, adaptability to new information, and privacy concerns. As technology develops, getting outside those barriers will be essential to realizing AI's full potential in revolutionizing our interactions with and extractions from PDF documents, unlocking the way to more complex and user-friendly document interaction solution

FUTURE SCOPE:

For future improvement and expansion, the project aiming to enhance PDF interaction through LLaMA and RAG can consider the following:

Data Quality and Expansion: Enhance the database with a wider range of high-quality, diverse sources. Implementing sophisticated indexing algorithms can improve the relevance and accuracy of retrieved information, ensuring the system can access the most pertinent data.

Improved Natural Language Processing: Incorporate the latest NLP advancements to better understand and respond to user queries. This could mean integrating newer or more specialized language models that offer improved performance for specific tasks or contexts.

Continuous Learning and Adaptation: Develop mechanisms for the system to learn from new information and user interactions over time, allowing it to adapt to changing information trends and user needs without extensive manual updates.

Enhanced Privacy and Security Measures: Implement cutting-edge security protocols to safeguard sensitive information within documents. Advanced encryption methods and data protection technologies can help ensure user data privacy and system integrity.

Collaborative Features: Introduce tools that enable collaborative interaction with documents. Features like real-time annotations and group chat can make the system useful for teamwork in educational and professional settings.

REFERENCES

Books:

"The LLM Knowledge Cookbook: From, RAG, to QLoRA, to Fine Tuning, and all the Recipes In Between! Kindle Edition" by Richard Aragon

"Retrieval Augmented Generation (RAG) AI: A Comprehensive Guide to Building and Deploying Intelligent Systems with RAG AI (AI Explorer Series)" by Et Tu Code

Online Resources:

Retrieval-Augmented Generation (RAG) Documentation: The documentation provides in-depth information on RAG, its features, and usage.

Large Language Models (LLMs) Documentation: The official LLMS documentation offers comprehensive guides, tutorils.