Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

PERSONAL CHATBOT FOR DOCUMENT

Prof. Pravin Kamble, Dept. Of Information Technology, Trinity College of Engineering and Research, Pune.
Uday Bhamre, Harishankar Thakur, Naaz Shaikh, Chandan Patil, Dept. Of Information Technology , Trinity College of Engineering and Research, Pune.
Udaybhamre10@gmail.com, omharithakur027@gmail.com, naazrs2402@gmail.com, patilchandan990@gmail.com, pravinkamble.tcoer@kjei.edu.in

Abstract
This Study has been undertaken to identify issues with chatbots like ChatGPT, regarding Security data exposure, Sensitive data Leaks, A lot of companies use chatbots, feed their data to chat- bots train them, and they respond to user input. In Recent Years, the Artificial Intelligence industry has evolved a lot, each industry transforming to Artificial intelligence, in every field Artificial intel- ligence is evolving Music, Image Generation, Text Generation, also the chatbot industry is the most used and evolving field of Artificial Intelligence industry, Generating Human-like responses to user queries is an important feature of chatbots, most of companies using this chatbot to automate Tasks and Provide Quality, Our Research mainly Focuses on Tackling and increasing the overall efficiency of the Chatbot and its Privacy, like ChatGPT. There are a lot of Open Source Large Language Models (LLM) out there, which are Pretty Accurate and Fast and Light, with our research we are making a Personal Document Chat Bot that will allow you to Chat with PDF/TXT/CSV Document, with 100 percent Privacy and Security, no need to upload PDF over any server or online, using power of Open Source Large Language Model (LLM) we are going to achieve this functionality, using this there will 100 percent security of your data and files will be achieved

Keywords: **:** Large language model, Python, OpenAI, Chatbot , Hugging Face, Chroma Vector Store

I. Introduction

Recently OpenAI introduced a feature inside ChatGPT where users can upload Any Document like PDF, CSV, TXT, and ChatGPT Models will Read the data from pdf and analyze it like Humans, and when the user provided query to ChatGPT, the model will answer his query, and any question about the PDF data, will be answered by the Model, this is a good and Impactful feature but, according to Our Research we found there is no Data Security or Privacy to our data, this Model Learns and improves itself, so the PDF you fed the Model will use the PDF and store the data to its Knowledge Base, which not good, we need to think when we are uploading our sensitive data Documents. Our Proposed System uses a Pre-Trained Open-Source Large Language Model, to Solve this problem of Data Privacy and Allow user to Upload any Sensitive PDF or Documents and Chat With them, which will like Personal Chatbot for Documents. Our System will use *Open Source Pre-Trained LLM* and will feed up all the Data of Documents and the LLM will Understand the Data of Documents and Process it and the user can answer. Our first approach is to extract data from Documents and Transform data into Such a Form that the Model Could Understand, the LLM Model Does Not directly under the text as input to it, we are converting the extracte data, splitting into charts building a semantics index, and storing it in the database(knowledge base). In this study, we have developed a secure approach of processing Document data and processing user queries, which will eliminate the need to upload your document to an external server or application and thus we will make a Personal Chatbot for Documents where you can chat with Sensitive information or documents

II. Literature

In our Research, we found there are lot of security issue with recent chatbot, “ChatGPT, devel- oped by OpenAI in November 2022, is an AI chatbot that utilizes the Generative Pre-trained Trans- former (GPT) model. OpenAI is an AI research and development company known for its innovative approaches in natural language processing. The GPT model, based on the Transformer architecture introduced by Vaswani et al. (2017), is trained on extensive datasets to generate contextually relevant and accurate responses to text-based inputs. However, as these systems become more sophisticated and widely used, concerns regarding user privacy and data protection have emerged. Large Language Models (LLMs) like ChatGPT aim to understand and generate human language, but their reliance on extensive datasets, which may contain sensitive information, raises privacy concerns.

There is a risk of inadvertently capturing and exposing sensitive user data, particularly in the context of chatbots and virtual assistants where personal or confidential information is often disclosed. These concerns have been addressed in various research papers discussing the usage of LLM-based chatbots, such as those by Hariri (2023), Sebastian (2023), and Cao et al. (2023). This research paper addresses the critical topic of data privacy risks associated with Large Language Models (LLMs) like ChatGPT. It acknowledges that LLMs use user data for training, which can be a threat to real privacy, especially when handling sensitive data such as financial or business documents. The paper underscores the importance of mitigating these privacy concerns through effective strategies and technologies.[1]

The proposed techniques to ensure robust data protection in LLMs include differential privacy, federated learning, data minimization, and secure multi-party computation. Additionally, the paper explores the legal and ethical frameworks necessary for the responsible development of AI systems, considering both the potential of LLMs and the importance of user privacy. It serves as a comprehen- sive guide for developers, policymakers, and researchers in the field, emphasizing the need to prioriti- tize user privacy in AI development. The research further discusses the specific privacy and security concerns when using LLM-based Chatbots in education. It highlights the importance of robust data privacy and security policies, transparency in data collection and use, modern technologies for data protection, regular audits, and incident response plans to safeguard student data. It also emphasizes the need to educate staff and students about data privacy and security.

The paper mentions common privacy and data leakage issues with AI-based Chatbots, which can impact user trust in AI systems. These issues include unintended sharing of sensitive information, data leakage through model outputs, model extraction, and data poisoning. While the paper clarifies that models like ChatGPT don’t have access to personal data, it warns about potential risks if communi- cation channels are not secure. [1]

In addition to this decision, this article also focuses on research on the use of advanced models (PFM) in artificial intelligence (e.g., GPT) for AI services in the Metaverse. It advocates efficient resource management and introduces “time points” metrics to balance the latency, power consumption, and accuracy of intelligent operations in the Metaverse. Exploring the broader impact of AI technology in the new virtual environment, as well as being ethically relevant to the use of AI in the world. These studies are important as the virtual world continues to develop and expand.

Unintended Sharing of Sensitive Information: This occurs when users are unaware that their per-

sonal or sensitive information is being shared with smart machines (Sweeney, L., 2002). For example, users will share credit card information and trust AI to keep that information safe. Note that temporarily stores non-personal information for 30 days to improve performance, although non-personal AI models such as ChatGPT cannot receive or store this information. The following events occur. Compromised: The communication method is not secure. [1]

Data Leakage through Model Outputs: While LLM models like ChatGPT may not know the details of the data they are working on, they can sometimes produce products that appear to use spe- cific data or leaked data. But these results, like the main points in the "negative thinking" response, are produced by the study sample at a particular point in time. The model does not leak real-world information learned during training, but instead makes models based on the models it has learned. [1]

Model Extraction: This involves an attacker using the output of a machine learning model to create a copy of the model without accessing the original training data. If successful, the attacker could use the extracted model for malicious purposes, affecting the security and integrity of the original system. [1]

Data Poisoning In this attack, the attacker introduces malicious objects into the data model with the aim of influencing future predictions or behavior. This is a significant threat to systems that continuously learn through user interaction. This research discusses the use of pre-learning models (PFM) such as pre-trained transformers (GPT) developed at the edge to intelligently deliver AI services.[1]

Data Collection: Gather information on actions, rules and regulations related to the mining in- dustry. This may include mining laws, environmental laws, safety procedures and other applicable laws. This can be done by sharing information and creating information or knowledge across different functions, rules and regulations. [1]

Data Anonymization and Aggregation: Anonymization is a method of data protection in which personally identifiable information in a data file is replaced by one or more false identifiers or pseudonyms. Aggregation, on the other hand, involves combining data in such a way that the resulting data does not contain personally identifiable information. [1]

Privacy-aware Machine Learning Algorithms: These algorithms are designed with privacy first. For example, in government, learning is a machine learning method used to train models to evaluate responses to determine whether end users are willing to sacrifice a system's performance or usability to improve privacy and data protection across multiple algorithms that record data locally. transfer to device but do not transfer [2]

III. Disscussion

Considering the issues mentioned in the literature review, we found that existing chat bots and systems

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

lack protection of users' data privacy. There are many existing chat bots that take documents as input and give you responses in natural language when you ask a question based on the documents. These are online chat-bots that need to upload documents and five answers to use, but we found from literature review and research that these chat-bots use user data to train themselves and here the user has no freedom to use it. There is a system for personal documents. Because this data can be stored somewhere where model and data leaks and exposures occur.

Samsung, for example, was recently forced to warn employees not to disclose sensitive company information to ChatGPT after it was revealed that employees were "using ChatGPT to assist with their code, anonymously placing blocks directly into the AI model." This includes self-regulation as well as internal discussions explaining the development process for the latest technology. Samsung's rapid release of information clearly violates the company's intellectual property rights, but its security implications are significant and a warning for future business models. A shocking article about the crypto industry's mistakes explains the "rotten culture" of big investments and public relations - and what else could go wrong when the "move fast and destroy" model becomes "move fast, destroy the world"? Changing the status quo is never easy, especially when you have to act quickly because financial rewards expire. Businesses continue to face big data, which is part of the "insider threat" where employees openly and easily share personal information.etary content or authorship with the wrong people and the public at large.

Additionally, such customized GPTs can be tricked into revealing their secrets. Researchers and experts working on special chatbots allow them to express the instructions they received at birth, and also express and download information used to treat them. He said people's identities and personal information could be compromised. They don't want others to know, and this is an important part of their customized GPT.

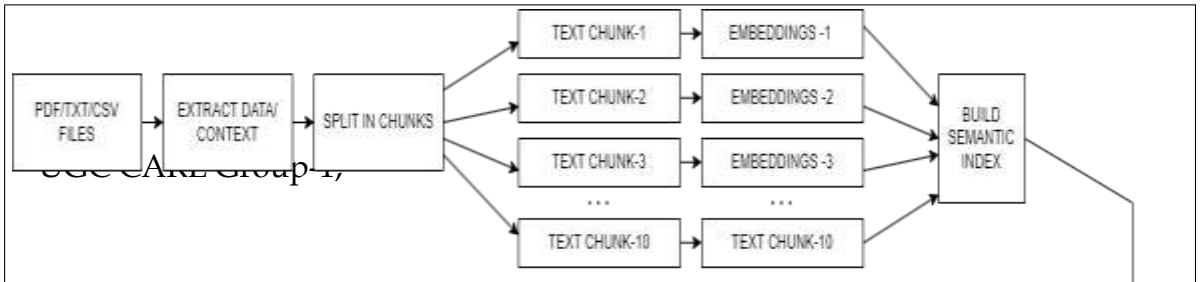
IV. Proposed System Architecture

Data Preprocessing : Initially, we Will extract Data from Documents, and then the extracted data is converted into chunks i.e. data chunks, Convert the collected data into smaller chunks or segments for efficient processing. This can be done by splitting the documents based on paragraphs, sections, or other logical divisions

Embedding Generation: Converting each chunk of text into numerical representations called embeddings. Use techniques like word embeddings (e.g., Word2Vec, GloVe) or sentence embeddings (e.g., BERT, Universal Sentence Encoder) to capture the semantic meaning of the text.

Building Semantics Index : Create an index of the generated embeddings to enable efficient semantic search. This can be done using techniques like Approximate Nearest Neighbors (ANN) or Inverted Indexing

.



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

Fig. 1. Extracting Data and Building Semantic Index

Query Processing Model: User Query is important which needs to understand and return the response to it, so user query is converted into query embedding after that Semantic search is done, based on the semantic search the results are ranked and the most matched result is returned.

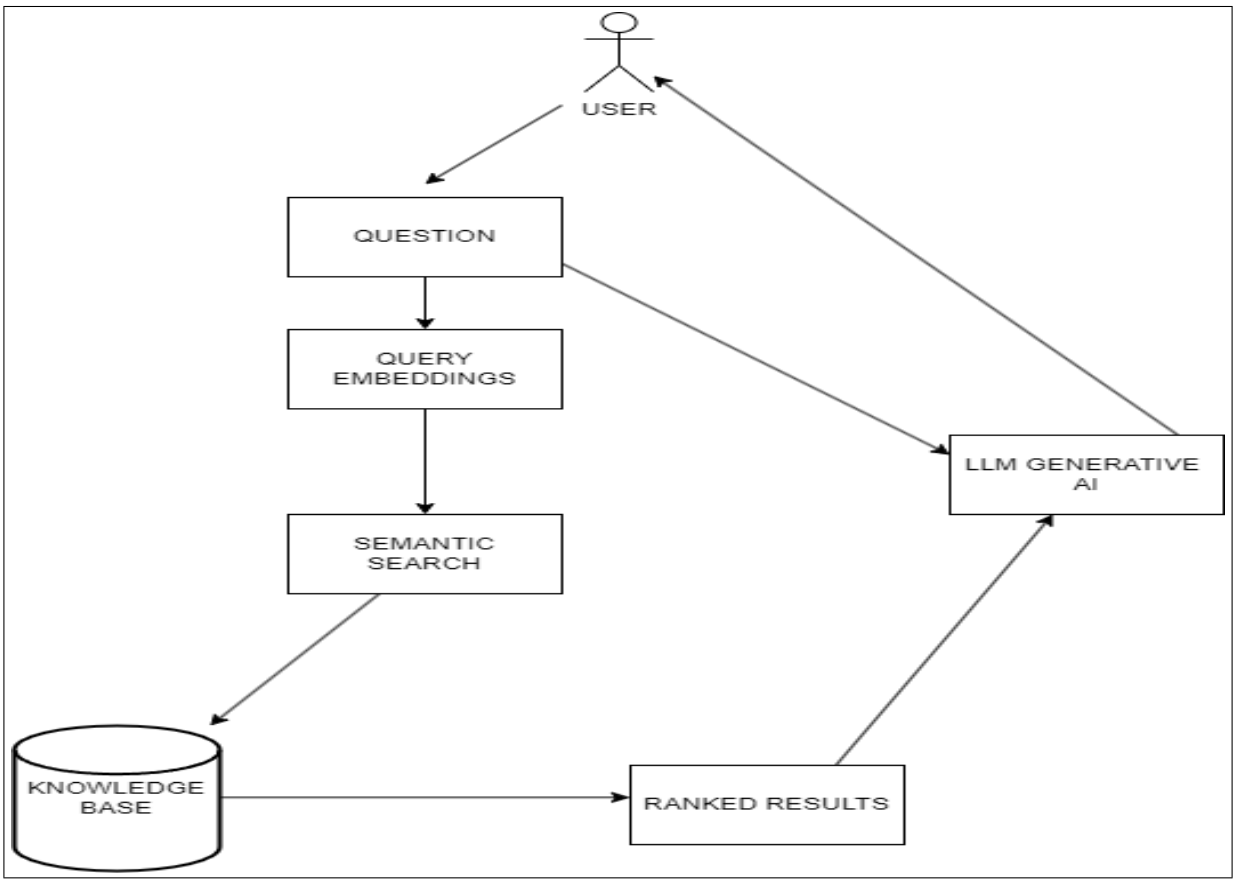


Fig. 2. Query Processing and Semantic search

V. System Architecture

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

Our Proposed The system uses Pre-Trained Open-Source Large Language Model, to Solve this problem of Data Privacy and Allow users to Upload any Sensitive PDF or Documents and Chat With its, which will like Personal Chat-bot for Documents. Our System will use Open Source Pre-Trained LLM and will fill up all the Data of Documents and the LLM will Understand the Data of Documents and Process it and the user can answer.

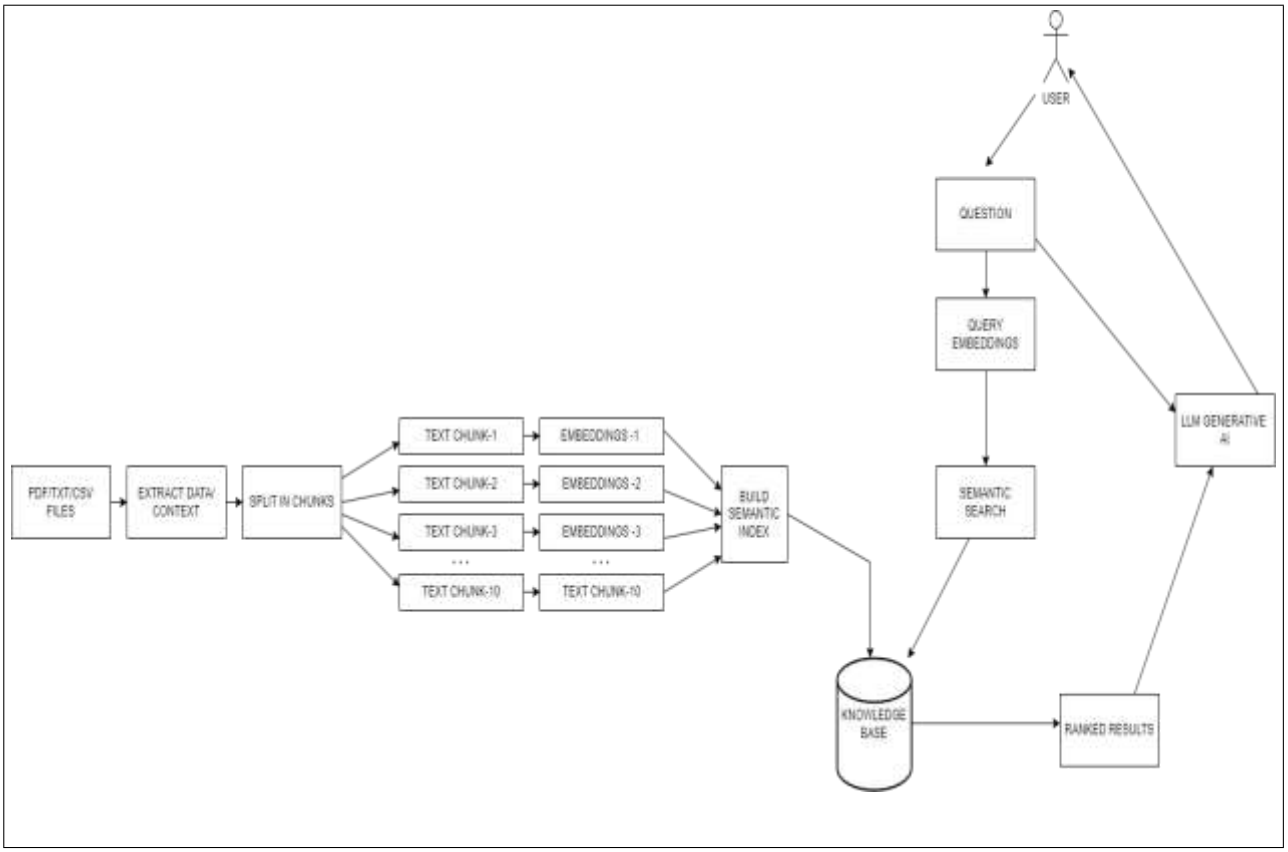


Fig. 3. System Architecture

VI. Methodology

In our proposed model we used Open Source Large Language Modal. The Large Language will give us a human-like response to user queries. Our system will take the uploaded PDF and parse it, and then after it it will divide the PDF.

Our Modal has 2 Main Components:

1) Extracting Document Data and Building Semantic Index

2) Query Processing and Semantic search

i] Extracting Document Data and Building Semantic Index :

i] **Extracting data context**: This involves extracting meaningful information from the content of a PDF file. This could include extracting text, images, tables, and other elements that convey information.

ii] **Splitting in chunks:** In the context of PDF files, splitting into chunks can refer to breaking down the extracted content into smaller, manageable sections. For example, splitting a large PDF document into individual pages or sections based on certain criteria.

iii] **Embedding**: Embedding typically refers to the process of including data or objects within a document. In the context of PDF files, embedding could involve embedding images, fonts, or other files within the PDF document itself.

iv] **Building semantic index**: A semantic index is an index that includes information about the meaning or context of the indexed content. Building a semantic index for PDF files involves creating an index that allows for meaningful search and retrieval of information based on the semantic content of the PDF documents.

v] **Storing in knowledge base**: Storing the extracted and processed information in a knowledge base involves saving the information in a structured format that allows for efficient storage and retrieval. This could involve using a database or other storage system to organize and manage the information extracted from the P/DF files.

i] Query Processing and Semantic search:

i] **Query**: The query is the actual question or request made by the user, which is used to retrieve

relevant information from a knowledge base.

ii] **Query embedding**: Query embedding involves converting the query into a numerical representation that captures its semantic meaning. This is often done using techniques like word embeddings or more advanced methods like BERT (Bidirectional Encoder Representations from Transformers).

iii] **Semantic search**: Semantic search is a search technique that seeks to understand the intent and meaning behind a query rather than just matching keywords. It involves using the semantic representation of the query to retrieve relevant information from a knowledge base.

iv] **Knowledge base**: The knowledge base is a repository of structured information that contains facts, data, or other information relevant to the domain. It stores the information extracted from PDF files or other sources, organized in a way that facilitates efficient retrieval.

v] **Generative AI**: Generative AI refers to artificial intelligence systems that can generate new content, such as text, images, or music, based on patterns learned from existing data. In the context of your process, generative AI could be used to generate additional information or responses based on the query and the contents of the knowledge base.

vi] **Ranked results:** Once the relevant information has been retrieved from the knowledge base, the results can be ranked based on their relevance to the query. This helps ensure that the most relevant information is presented to the user first.

VII. RESULTS

Expected was system should able to take Documents as input, and should be able to Process the document. After processing the document, the user will ask the query, similarly, the system should

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, May : 2024

process the query and give a response to the user Query. This Final expected result is successfully achiuracy.

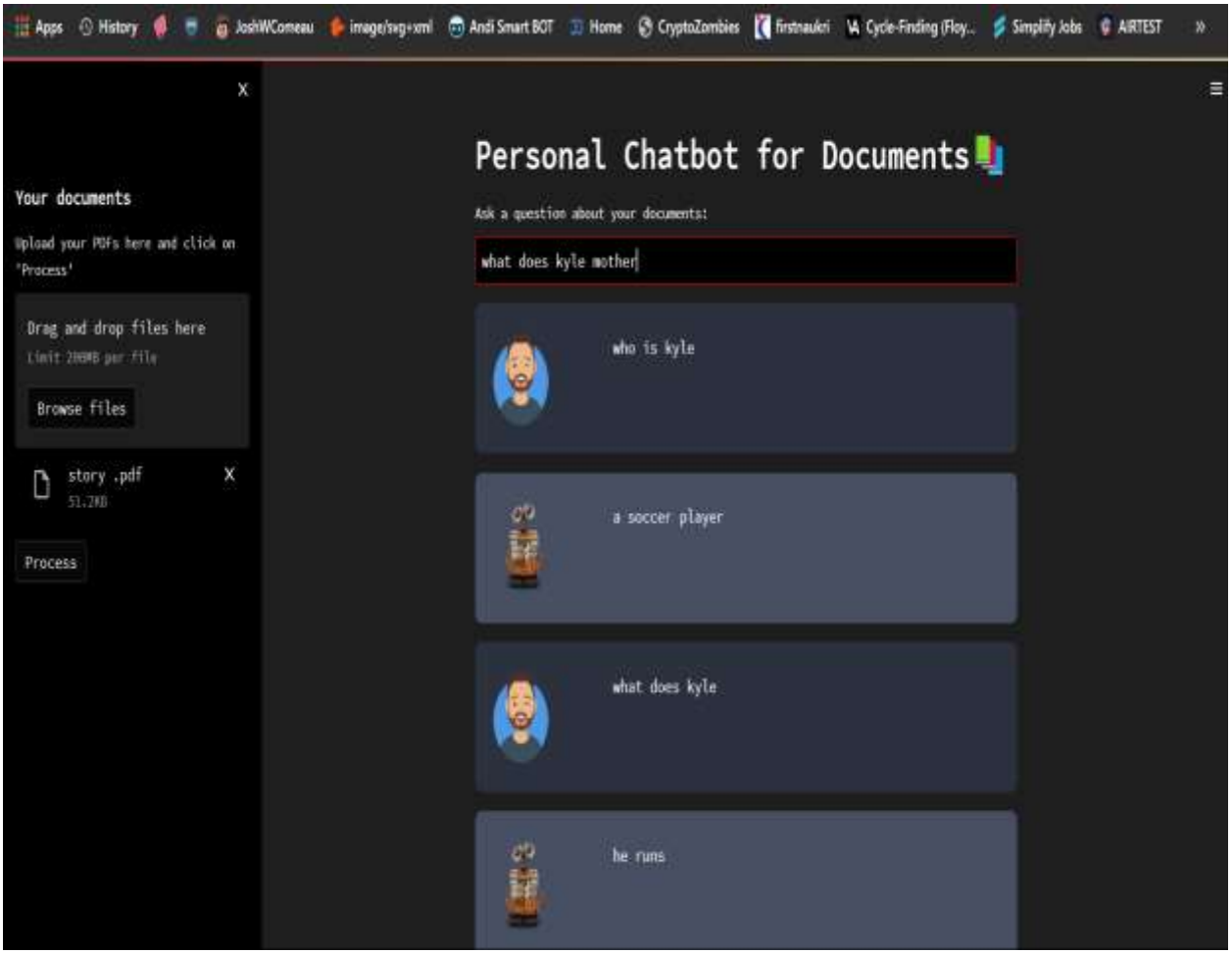


Fig. 4. OUTPUT RESULTS

The Application works fine but bottlenecks for system requirements as the models used for Embedding and Text Transformation are CPU bound. New PDF Documents take time for Text transformation when they are uploaded for the first time. Later the Responses to the query can be Generated in Seconds. Support for other documents can be added to expand the scope of this application.

VIII. Conclusion

The Previous System of Chat with Document was not Secured and leading to sensitive data exposure and leaks, the person was not able to upload or chat with personal documents like Balance Sheets, Personal Documents etc, which was a major flaw, the peoples with this document was not able to take advantage of features, they all need to do manual analysing of Documents Our Proposed system tackles this problem all allows users to upload and chat with documents with full Privacy and Secured manner, no data leaves the Device of the user, this system uses Open source LLM for Generating Human-like Responses to User Documents, in First Step we will extract and Store user Data from Documents and later the LLM Model will have Access to the Data and the LLM Model will be Pre- trained and the Model will analyze the user query and return the most Ranked response to the user, the main advantage of our system is it will be completely offline will be in the user's device

References

[1] Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information Glorin Sebastian, Georgia Institute of Technology, USA* https://orcid.org/0000- 0003-2543-9127

[2] ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope Partha Pratim Ray Sikkim doi.org/10.1016/j.iotcps.2023.04.0

[3] A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity Moatsum Alawida 1, *, Sami Mejri 2, Abid Mehmood 1, Belkacem Chikhaoui 3, * and Oludare Isaac Abiodun 4 doi.org/10.3390/info14080462

[4] ChatGPT for good? On opportunities and challenges of large language models for education Enkelejda Kasneci a,* , Kathrin Sessler a , Stefan Küchemann b , Maria Bannert a , Daryna Dementieva a , Frank Fischer b , Urs Gasser a , Georg Groh a , Stephan Günnemann a , Eyke Hüllermeier b , Stephan Krusche a , Gitta Kutyniok b , Tilman Michaeli a , Claudia Nerdel a , Jürgen Pfeffer a , Oleksandra Poquet a , Michael Sailer b , Albrecht Schmidt b , Tina Seidel a ,
Matthias Stadler b , Jochen Weller b , Jochen Kuhn b , Gjergji Kasneci c

[5] IndiaAiyappa, R. (2023). Can we trust the evaluation on ChatGPT? arXiv preprint arXiv:2303.12767 Akhawe, D., Amann, B., Vallentin, M., Sommer, R. (2013, November). Here's my cert, so trust me, maybe? understanding TLS errors on the web. ACM.ess, 10, 134018–134028.