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

DOCUMENT CHATTER WEB APPLICATION

Mr. M. Sukesh, Associate Professor CSE, Vaagdevi College of Engineering (Autonomous), India
S. Vinay, UG Student, CSE, Vaagdevi College of Engineering (Autonomous), India
M. Rajesh, UG Student, CSE, Vaagdevi College of Engineering (Autonomous), India
J. Pranathi, UG Student, CSE, Vaagdevi College of Engineering (Autonomous), India
B. Prem Kumar, UG Student, CSE, Vaagdevi College of Engineering (Autonomous), India

ABSTRACTa

Chatbots are software systems designed to simulate conversation with human users, addressing their queries and interacting as a human would in various situations. This project introduces the "Document Chatter Web Application," a platform that allows users to engage in smart conversations with their documents online. Unlike traditional reading, this application enables real-time chat with specific sections of a document, facilitating easier information retrieval, saving time, and enhancing content interaction as users process the information. This paper outlines a methodology for extracting textual content from documents in various formats including PDF, PowerPoint (PPT), Word, and plain text files and leveraging this content to build a conversational chatbot. The primary tools employed for this task are the Retrieval-Augmented Generation (RAG) model and Large Language Models (LLMs), which enable the generation of abstractive responses with accurate reference to page numbers and sections within the documents. The process includes extracting text from these multiple document formats, preprocessing the text for clarity and conciseness, and employing RAG and LLMs to develop a chatbot model capable of understanding and responding to user queries contextually. Users can then upload their documents to the web application and interact with them through this sophisticated chatbot interface. The results showcase the chatbot's proficiency in interpreting and responding to user queries by leveraging the extracted content from diverse document types. This research contributes significantly to the development of intelligent systems that extract and utilize knowledge from structured and unstructured text sources, thereby creating interactive and informative chatbot interfaces for educational, professional, and personal use.

1. INTRODUCTION

Over the last few years, Chatbots have played an important role as human-computer interfaces. Chatbots generally consist of three modules: the user interface, an interpreter, and a knowledge base. A chatbot is a conversational agent that can interact with users in a given particular subject using natural language. Many chatbots have been deployed on the internet for education, customer service, guidance, and entertainment.

The advent of the "Document Chatter Web Application" [1] marks a significant evolution in how we interact with text-based information. This innovative platform transforms static documents into dynamic conversations, allowing users to engage directly with content through a chat interface. Users can upload their documents to the web application and begin a dialogue, asking questions and receiving abstractive responses with specific page references. This conversational model not only makes information retrieval more intuitive but also significantly enhances the efficiency and engagement of document review processes. By integrating smart technology with traditional document formats, the Document Chatter Web Application redefines the user experience, making document interactions more interactive and informative. With its ability to parse and understand complex content, the application offers a unique tool for professionals and students alike, who can now interact with their reading material more engagingly and productively. This technology promises to revolutionise the way we consume and interact with written content, making it an essential tool for anyone looking to extract more value from their documents efficiently

2. LITERATURE SURVEY

A literature review forms the most important part of all scientific research. As a systematic investigation to find conclusions and to achieve facts, every scientific research builds on existing knowledge. Unless one wants to change the wheel, precise awareness of the extent of wisdom on a subject is necessary to carry on research that adds value to the area. A literature review for scientific research can be defined as a survey of scientific papers, scholarly articles, and all other systematic scientific sources similar to a particular problem, field of study, or theory, to include a description, summary, and critical evaluation of a concept, school of thought, or ideas related to the research question is Tested. In extension, the literature review familiarizes the author to the extent of knowledge in their area. When represented as a part of the paper, it establishes to the readers, the author's depth of understanding and knowledge of their field subject. The literature is primarily scrutinized to identify gaps in the knowledge of the field source. This gap is further explored during research to establish the latest facts or theories that provide value to the field.

Large pre-trained language models have been shown to store factual knowledge in their parameters and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models that combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG [2] models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which condition on the same retrieved passages across the whole generated sequence, and the other can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open-domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

Many organisations are executing Chatbots to address client inquiries and contact clients. As indicated by Mindshare, 63% of shoppers would consider utilizing a chatbot when visiting a business or brand's site. One of the fundamental AI chatbot benefits is that it can convey moment satisfaction. Individuals would much preferably visit online over setting aside the effort to call an organization's 800 number. During these difficult situations, it is hard for individuals to go to the stores to purchase something, to the emergency clinic for a little clinical test, find support for an item that they purchased and so forth So these sorts of easier errands, which don't require actual presence, can be supplanted by chatbots. So we will make a chatbot, which when given reasonable purpose documents dependent on a specific item or necessities, can prepare on it utilizing various Layers Neural Networks and make a model. Utilising this model our chatbot can answer client inquiries.

In the digital age, the dynamics of customer service are evolving, driven by technological advancements and the integration of Large Language Models (LLMs). This research paper introduces a groundbreaking approach to automating customer service using LangChain, a custom LLM [3] tailored for organizations. The paper explores the obsolescence of traditional customer support techniques, particularly Frequently Asked Questions (FAQs), and proposes a paradigm shift towards responsive, context-aware, and personalized customer interactions. The heart of this innovation lies in the fusion of open-source methodologies, web scraping, fine-tuning, and the seamless integration of LangChain into customer service platforms. This open-source state-of-the-art framework, presented as "Sahaay," can scale across industries and organizations, offering real-time support and query resolution. Key elements of this research encompass data collection via web scraping, the role of embeddings, the utilization of Google's Flan T5 XXL, Base and Small language models for knowledge

retrieval, and the integration of the chatbot into customer service platforms. The results section provides insights into their performance and use cases, here particularly within an educational institution. This research heralds a new era in customer service, where technology is harnessed to create efficient, personalized, and responsive interactions. Sahaay, powered by LangChain, redefines the customer-company relationship, elevating customer retention, value extraction, and brand image. As organizations embrace LLMs, customer service becomes a dynamic and customer-centric ecosystem.

Large Language Models (LLMs) [4] showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes. Retrieval-augmented generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain-specific information. RAG synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of external databases. This comprehensive review paper offers a detailed examination of the progression of RAG paradigms, encompassing the Naive RAG, the Advanced RAG, and the Modular RAG. It meticulously scrutinizes the tripartite foundation of RAG frameworks, which includes the retrieval, the generation and the augmentation techniques. The paper highlights the state-of-the-art technologies embedded in each of these critical components, providing a profound understanding of the advancements in RAG systems. Furthermore, this paper introduces an up-to-date evaluation framework and benchmark. In the end, this article delineates the challenges currently faced and points out prospective avenues for research and development.

3. PROBLEM STATEMENT

The existing systems in the field of document interaction and chatbot technology primarily focus on static methods of information retrieval, such as keyword searches, indices, or traditional Q&A systems based on fixed templates [5]. These systems often require users to manually sift through documents to find relevant information or rely on simplistic interaction models that do not support conversational engagement.

Furthermore, current systems cannot engage users in a dialogue, a feature that could significantly streamline the process of finding and understanding information. There is no mechanism to clarify user queries or provide additional context, which is a stark contrast to conversational interfaces that can handle a range of inquiries by understanding and responding to natural language. This gap highlights a significant limitation in user experience and efficiency in how information is currently accessed and processed in digital documents.

3.1 LIMITATIONS

1. Limited Query Understanding
2. Inefficient Information Retrieval
3. Lack of Interactivity

4. PROPOSED SYSTEM

The proposed "Document Chatter Web Application" is designed to revolutionize how users interact with documents by introducing a conversational interface that allows for dynamic communication with digital texts. This system leverages advanced natural language processing (NLP) technologies, including Retrieval-Augmented Generation (RAG) models and Large Language Models (LLMs) [6], to provide a more intuitive and interactive document-handling experience. This system allows users to upload various types of documents including PDFs, Word documents, PowerPoint presentations, and plain text files and engage in real-time, dynamic conversations with the content. By interpreting natural language queries, the application provides not just literal answers but contextually relevant, abstractive responses complete with references to specific document sections and page numbers.

This innovative approach significantly enhances user experience by streamlining the search and retrieval process. It understands the context behind user queries, adapts to individual interaction patterns, and continually improves its accuracy and efficiency through machine learning. Not only does the system facilitate intuitive and accessible information retrieval, but it also precisely cites the page numbers from which the data is taken, ensuring users can easily locate and reference the original content within their documents. This transformative functionality reshapes how users interact with and digest document-based information.

4.1 ADVANTAGES

1. Enhanced User Interaction
2. Efficient Information Retrieval

5. SYSTEM ARCHITECTURE

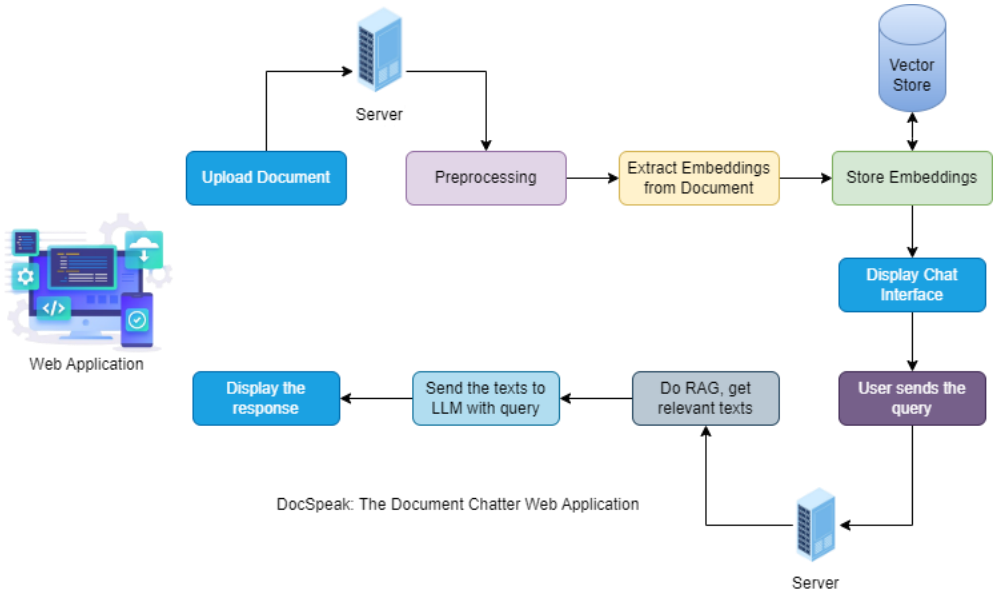


Fig: System Architecture

6. IMPLEMENTATION

1. DOCUMENT UPLOAD MODULE

This module provides the interface through which users can upload documents to the application. It supports various document formats such as PDF, DOCX, PPT, and text files. The primary function is to capture and store document data securely while maintaining user session information to associate documents with their respective users. This interface is critical for initiating the user's interaction with their documents in a structured and user-friendly manner.

2. TEXT EXTRACTION MODULE

Once a document is uploaded, the text extraction module is responsible for processing the document to extract plain text. This module handles different file formats and extracts the textual content using specialized libraries suited to handle the peculiarities of each format (e.g., PDF extraction libraries for PDF files). The extracted text serves as the foundation for further processing and interaction. This step is crucial for processing large documents where direct handling could be computationally intensive and less manageable.

3. EMBEDDING EXTRACTION MODULE

This module takes the extracted text and converts it into document embeddings. Using advanced natural language processing techniques, the text is transformed into vector representations that capture the semantic meanings of the words and phrases within the documents. These embeddings are crucial for the subsequent retrieval of relevant document sections during user interactions. The

module is designed to handle large-scale documents efficiently by batching operations, showing progress, and optimizing API key usage for external services.

4. EMBEDDING STORAGE MODULE

The embeddings generated by the previous module are stored in a vector storage system designed for high efficiency and scalability. This storage system facilitates the quick retrieval of embeddings based on similarity measures, which is essential for the efficient functioning of the retrieval-based conversation system. This module ensures that the data remains organized and easily accessible for real-time querying.

5. CONVERSATIONAL INTERFACE MODULE

This module is the user-facing component where interaction with the document takes place. It displays a chat interface where users can type in their queries or prompts. The module captures these inputs and uses a Retrieval-Augmented Generation (RAG) approach to identify the most relevant sections of text within the stored documents, based on the computed embeddings.

6. QUERY AND RESPONSE PROCESSING MODULE

After relevant texts are retrieved, this module combines the texts, user query, and any previous conversation context to formulate a comprehensive input to a large language model (LLM). The LLM processes this input to generate meaningful and contextually appropriate responses. This process ensures that the responses are not only accurate but also relevant to the ongoing conversation. The query, along with any previous chat history, is processed to dynamically generate responses that are contextually relevant and informative. The combination of RAG techniques and a sophisticated prompting system ensures that the conversation remains relevant and grounded in the document's content.

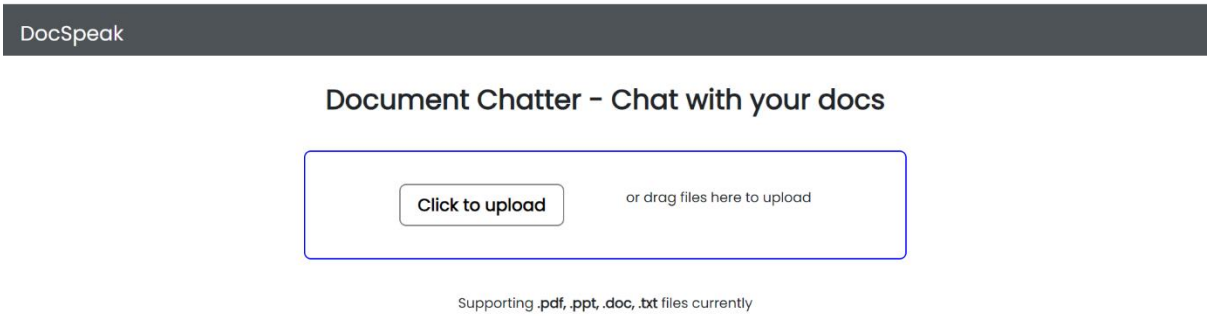
7. RESPONSE DISPLAY MODULE

The final module in the workflow handles the presentation of the generated responses back to the user through the conversational interface. It ensures that the responses are displayed conversationally, maintaining the context and continuity of the interaction. This module plays a crucial role in ensuring user engagement and satisfaction with the conversational experience. This module ensures that the answers are displayed in a clear, conversational format in the web interface. The inclusion of document references provides an added layer of transparency and usefulness, allowing users to locate the source of the information within the document.

7. EXPECTED OUTCOMES

Home Screen

The home contains an upload box where the user can drag and drop their documents or just click on the **Click to upload** to upload the documents.



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

Fig: Home Screen

Selected Documents

After the user selects the documents, it gets shown on the screen with the file type and file name for processing.

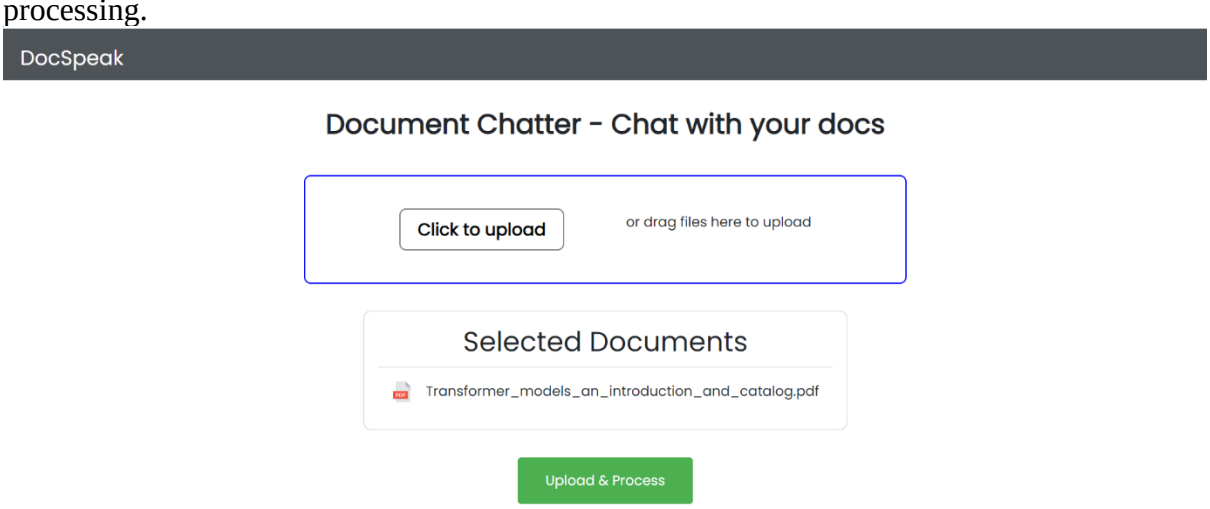


Fig: Selected Documents Screen

Processing

When the user uploads the documents, they get processed in the backend where the embeddings are extracted and stored.

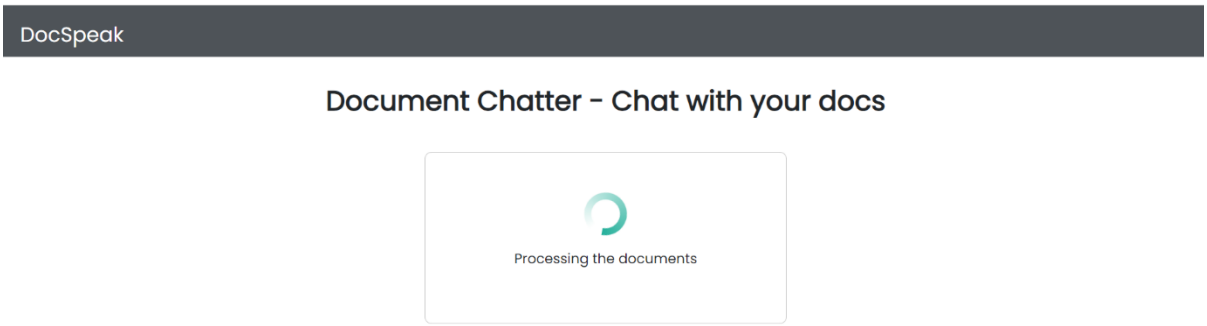


Fig: Processing Screen

Chat Screen

After the documents are processed, a chat screen is shown to the user. The user can send their questions to the chatbot about the uploaded documents and it responds to the user with answers and also the reference page numbers telling from which pages the information is used.

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

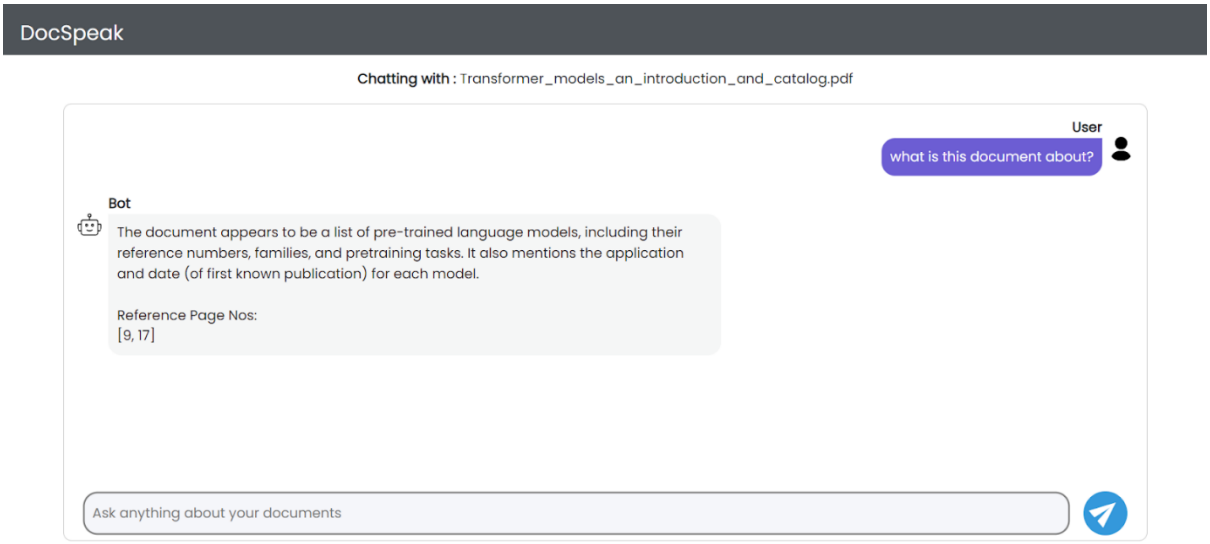


Fig: Chat Screen

Chat History

The chatbot stores all the previous questions and answers in memory and has a contextual understanding using which it can respond to the user from the current context and previous chat history.

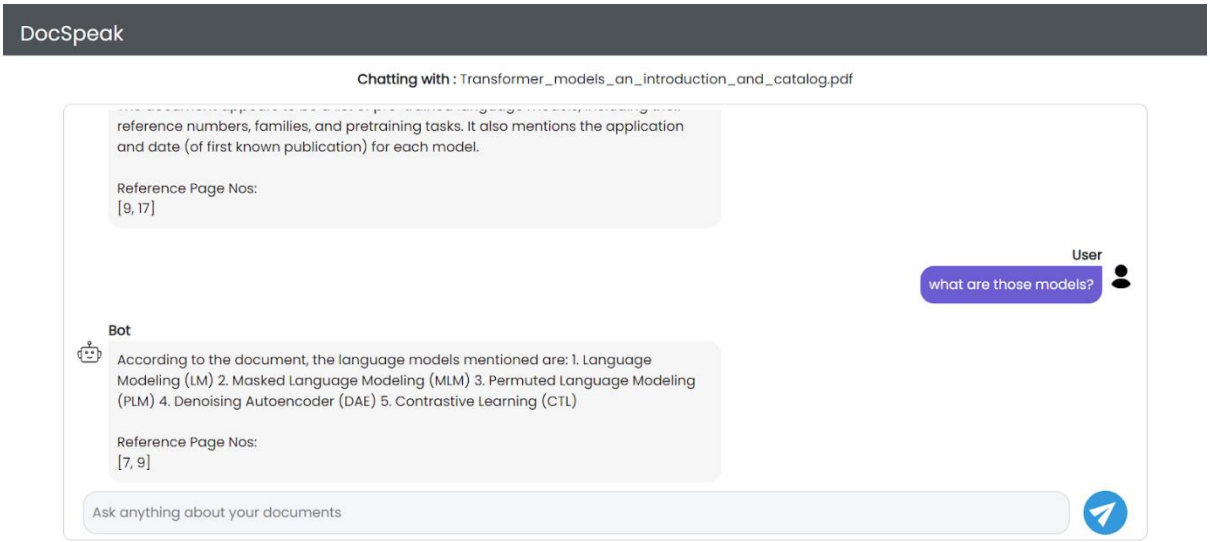


Fig: Chat History Screen

8. CONCLUSION

This project successfully integrates cutting-edge web and artificial intelligence technologies to create an interactive platform for document processing and information retrieval. Through a user-friendly web interface, users can upload documents in multiple formats, which are then processed into searchable and analysable content. The backbone of this system uses advanced embeddings and vector storage techniques to ensure that users can retrieve information quickly and accurately. Furthermore, the incorporation of conversational AI through large language models enables users to interact with the system naturally and intuitively, asking complex questions and receiving relevant answers directly derived from the document content. This integration not only streamlines the user experience but also significantly enhances the accessibility and utility of the stored information.

FUTURE SCOPE

The future development of this project can take several exciting directions. Enhancing the AI capabilities by incorporating more sophisticated natural language processing models could further improve the accuracy and responsiveness of the system. Expanding the range of supported document

formats and the types of data analysis available would make the platform applicable to a broader array of industries and academic fields. Moreover, refining the user interface with additional interactive features such as real-time query suggestions and auto-completion would enrich user interactions and satisfaction. Additionally, integrating a more advanced data security framework would enhance the system's applicability for handling sensitive or proprietary information, broadening its usage in fields requiring stringent data protections.

9. REFERENCES

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.
[2] Jay Alammar. (2018). The Illustrated Transformer.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need.
[4] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training.
[5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners.
[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.