



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Desktop Search Engine: A Real-Time Chatbot For PDF Document Interaction Using Large Language Models

¹Dr. Sowmya K S, ²Ranjini Ravi Iyer, ³Vidhya M Hegde, ⁴Manish B S
¹Professor, ^{2,3,4}Student

^{1,2,3,4} Department of Information Science and Engineering

^{1,2,3,4} BMS College of Engineering, Bangalore, India

Abstract: The proliferation of digital documents, particularly in PDF format, necessitates innovative solutions for efficient information retrieval and interaction. In response to this growing demand, we introduce a pioneering framework enabling users to engage in natural language conversations with PDF documents through a chatbot interface powered by Large Language Models (LLMs). Our system addresses the inherent challenges of document interaction, including complex content structures, diverse document types, and varying user needs. By harnessing the capabilities of LLMs, users can upload PDF documents and seamlessly converse with the chatbot to extract specific information, pose queries, and navigate through the document's contents.

The development of our chatbot framework is motivated by the imperative to enhance accessibility and usability in document management systems across diverse domains such as academia, research, and corporate environments. Through comprehensive document preprocessing techniques and sophisticated LLM-based conversational agents, our framework empowers users with intuitive and efficient means to interact with PDF documents. We underscore the significance of our approach through empirical evaluations and real-world case studies, demonstrating its effectiveness in handling multifaceted document structures and providing accurate responses to user inquiries. This research represents a significant step forward in augmenting human-computer interaction paradigms, offering a user-centric approach to unlocking the wealth of knowledge embedded within PDF documents.

Index Terms -Chatbot, PDF Interaction, Large Language Models, Document Understanding, Natural Language Processing, Information Retrieval.

I. INTRODUCTION

The proliferation of PDF documents in today's digital landscape underscores the need for innovative approaches to facilitate efficient information retrieval and interaction. These documents, spanning a myriad of domains including academia, business, and legal affairs, serve as repositories of valuable knowledge. However, conventional methods for navigating and extracting insights from PDFs often prove inadequate in meeting the diverse needs of users. Keyword-based searches and manual browsing are constrained by their inability to comprehend natural language queries or discern contextual nuances within documents. Consequently, there is a growing imperative to develop intelligent systems that can bridge this gap and empower users with more intuitive and effective means of engaging with PDF content.

To address this challenge, we propose a pioneering framework that harnesses the power of Large Language Models (LLMs) to enable natural language interaction with PDF documents through a chatbot interface. By leveraging recent advancements in natural language processing (NLP) and machine learning, our approach seeks to transcend the limitations of traditional document interaction paradigms. Through the integration of sophisticated NLP techniques and state-of-the-art machine learning models, our framework empowers users to engage in conversational exchanges with PDF documents, effectively transforming static documents into dynamic sources of knowledge. This introduction sets the stage for our research endeavor, outlining the motivations behind our proposed approach and highlighting its potential to revolutionize the landscape of document management and information retrieval.

At the core of our proposed framework lies the fusion of advanced document preprocessing techniques and robust conversational agents driven by LLMs. Document preprocessing plays a pivotal role in transforming raw PDF content into a structured format conducive to natural language understanding. Through techniques such as text extraction, entity recognition, and semantic analysis, we aim to enhance the chatbot's ability to interpret and respond to user queries accurately. Furthermore, the utilization of LLMs enables our chatbot to capture the intricacies of natural language interactions, providing users with a seamless and intuitive experience. By harnessing the collective knowledge encoded within these language models, our framework empowers users to extract valuable insights, navigate complex document structures, and derive meaningful conclusions from PDF documents with unprecedented ease and efficiency.

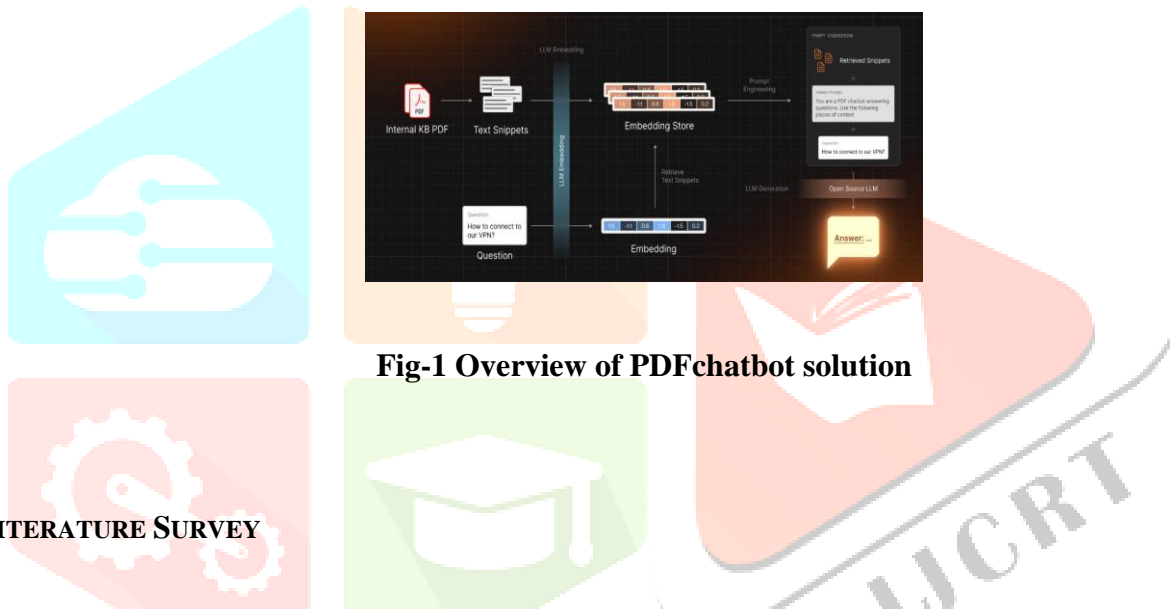


Fig-1 Overview of PDFchatbot solution

II.LITERATURE SURVEY

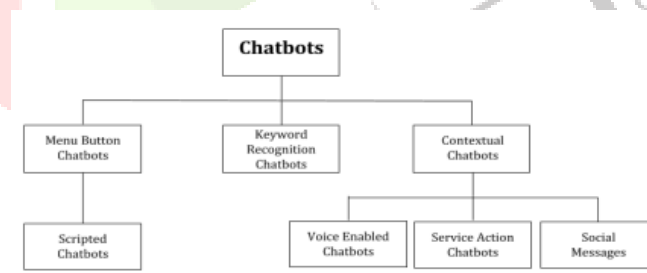


Fig 2. Proposed Classification in [2]

Conversational Document Prediction to Assist Customer Care Agents (Jatin Ganhotra, Haggai Roitman)[1]:

The Conversational Document Prediction (CDP) task has been introduced to assess the performance of state-of-the-art deep learning and information retrieval models. A new public Twitter dataset has been released specifically for the CDP task, providing a valuable resource for researchers and practitioners in this field. The study concentrated on URL documents that contain content, allowing for a focused evaluation of the models' capabilities in processing and predicting conversational documents.

Future research will aim to address additional challenges by considering a broader range of document types, including PDFs, DOC files, and URLs without content, such as login pages and tracking links. This expansion will enhance the versatility and applicability of the models, enabling them to handle a wider variety of document types and use cases. The ongoing work will contribute to the development of more robust and comprehensive conversational document prediction systems.

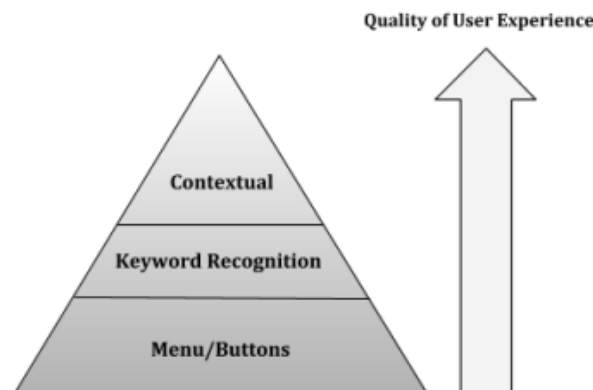


Fig 3. Preference for Chatbots in [2].

Introduction to AI Chatbots (Aishwarya Gupta, Divya Hathwar)[2]:

The integration of chatbots into industrial applications represents a pivotal advancement in virtual assistant technology. Leveraging recent developments in artificial intelligence, machine learning, natural language processing, and deep learning, modern chatbots have significantly evolved from their early iterations, such as Alice and Eliza. A notable example is Samsung's STAR Labs' development of Neon, a chatbot engineered to exhibit human-like emotional intelligence. Unlike traditional, omniscient AI tools, these advanced chatbots are designed to augment human capabilities, enabling users to shift their focus from routine tasks to more strategic and innovative endeavors. This research proposes that the ongoing enhancement of chatbot technology will continue to improve human-computer interaction, driving efficiency and fostering creativity in various industrial sectors.

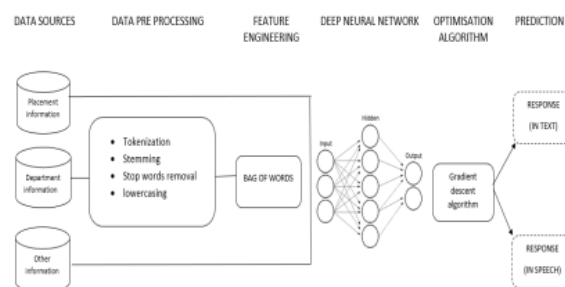


Fig 4. Proposed Methodology in [3]

Development of Artificial Intelligence based Chatbot using Deep Neural Network (Dammavalam Srinivasa Rao)[3]:

This research paper proposes the development of a college inquiry chatbot designed to employ algorithms for interpreting and understanding user messages. As the internet has revolutionized access to information and communication, there is a growing need for technology capable of addressing user queries 24/7. The proposed chatbot, an artificial program, aims to connect seamlessly with users to assist and resolve their inquiries efficiently. Offering a diverse range of services, the chatbot can manage tasks from daily essential queries to

complex industrial needs, proving to be more practical and accessible than human operators. Future work will explore the incorporation of speech-based inquiries and responses, integration into the college's website, and the application of sentiment analysis to understand user emotions and states of mind. This approach will enhance the chatbot's ability to deliver tailored and appropriate responses, with continuous data integration aimed at improving its authenticity and accuracy.

DeepPDF: A Deep Learning Approach to Analyzing PDFs(Christopher Stahl, Steven Young)[4]:

This research paper proposes DeepPDF, a deep learning-based approach for analyzing PDFs, specifically targeting the identification of sections within scientific publications. The current experiments demonstrate that deep learning can significantly enhance PDF extraction methods by accurately distinguishing between body text and other document sections. The findings suggest that the primary cause of misclassified text is insufficient training data, particularly for features like reference sections and abstracts. Future work will focus on collecting more data to improve the network's accuracy and extending the approach to identify various text types (e.g., title, author, abstract, body text). Additionally, the development of an extraction tool that utilizes the deep learning network's output is planned, aiming to refine text extraction accuracy. Evaluation methods will also evolve from per-pixel accuracy to per-character accuracy using this extraction tool, ensuring more precise text identification and extraction.

Conversational Interfaces for InformationSearch (Q. Vera Liao, Werner Geyer)[5]:

This research paper explores the emerging area of conversational interfaces for information search systems, highlighting the importance of human conversation as a metaphor for user interaction. This metaphor helps users intuitively understand how to interact with the system, while the interface bridges these familiar actions with underlying computational models. The paper aims to identify functional goals for conversational search systems by examining properties of natural conversations that benefit information search tasks. Through a review of relevant literature, the authors address system actions that extend traditional search models to conversational search, focusing on query formulation, search results exploration, and query repair. Additionally, they consider fundamental conversational properties such as efficiency, common ground, and recipient design. The convergence of these threads leads to a design space outlining the functional goals for conversational search systems. The paper identifies a gap and opportunity to prioritize user modeling and adaptation in conversational interfaces, offering insights from the authors' work on creating adaptive and conversational search systems.

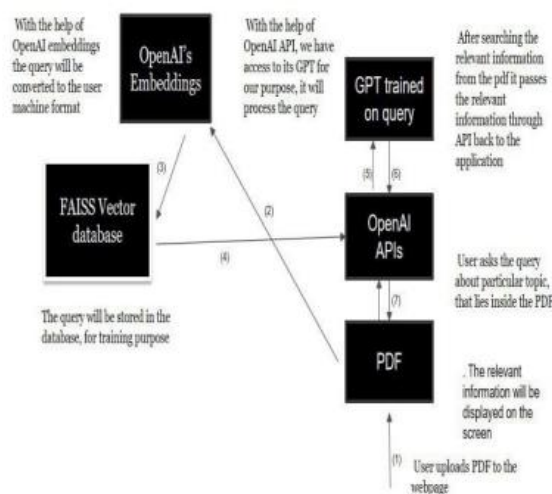


Fig 5 . Proposed work flow in [6]

Automating Pdf Interaction Using Langchain(Karan R.,M. Rahul Kumar)[6]:

This research paper proposes an automated PDF interaction system using Langchain, designed to benefit domains such as medicine, engineering, and science by keeping pace with the rapid advancements in artificial intelligence. The system aims to reduce the time and effort required for literature reviews, enhance model understanding of interactions, and generate relevant future queries. For researchers, the application streamlines the retrieval, organization, and analysis of extensive scholarly content, thereby accelerating discovery. For students, it offers a valuable tool for efficient document management, research projects, and collaborative efforts. The adoption of PDF automation represents a significant technological advancement, empowering both researchers and students to excel in their fields by making knowledge more accessible and research more efficient. The model used for processing and generating results can be adapted for various applications, promoting scalability and time efficiency. Additionally, as a reinforcement learning tool, it provides computer science students with practical insights into the model's functionality.

BUILDING A SMART CHATBOT (Mr. E. Sankar, Alekya B V)[7]:

This project focuses on developing a smart chatbot capable of engaging in user interactions, with the primary aim of simulating human conversation to address student queries efficiently. By facilitating one-on-one conversations, the chatbot streamlines information retrieval, providing accurate answers to user-submitted questions and guiding students to relevant sources effectively. Achieving an impressive accuracy rate of 94%, the project visually represents its output for enhanced comprehension. Looking ahead, the rapid advancements in natural language processing (NLP) hold promise for creating even more powerful, human-like chatbots. Future enhancements may include incorporating voice-based queries, allowing users to provide input verbally while receiving text-based outputs. Moreover, given the successful execution of the chatbot in a college domain, there's potential for implementation across various domains such as medical, sports, and more. This versatility ensures that users across different fields can quickly access necessary information, significantly reducing time spent on information retrieval tasks.

A neural-based text summarization system (S. P. Yong,)[8]:

This paper introduces TextSum, a neural-based text summarization system that stands out for its utilization of unsupervised learning, a relatively rare approach in previous systems. Unlike other systems like NeuralSumm, TextSum is designed to learn how to classify keywords intelligently, potentially eliminating the need for extensive training data input by users. At the core of TextSum lies its competitive network architecture, meticulously crafted to directly impact the quality of system output. Evaluation results indicate satisfactory performance, with TextSum generating summaries with an average content score of 83.03% across various report types, including news articles, technical papers, and product descriptions.

To further enhance TextSum's performance, the paper suggests training the competitive network with a more extensive dataset comprising multiple sample documents, particularly in fields relevant to the system's application. Presently, the network is trained with three types of reports, but expanding to include diverse document types such as legal documents and financial news could enhance the network's robustness. This broader training approach would enable professionals from various domains to benefit from TextSum's summarization capabilities, thereby increasing its utility and applicability across different sectors.

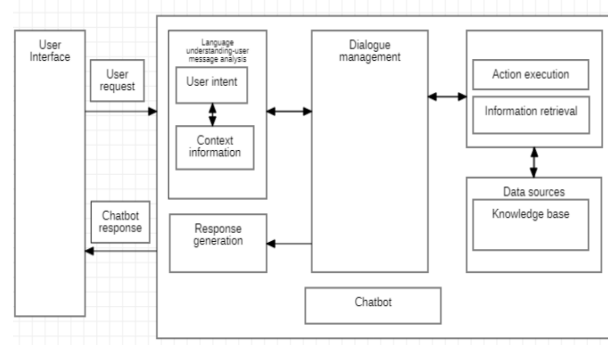
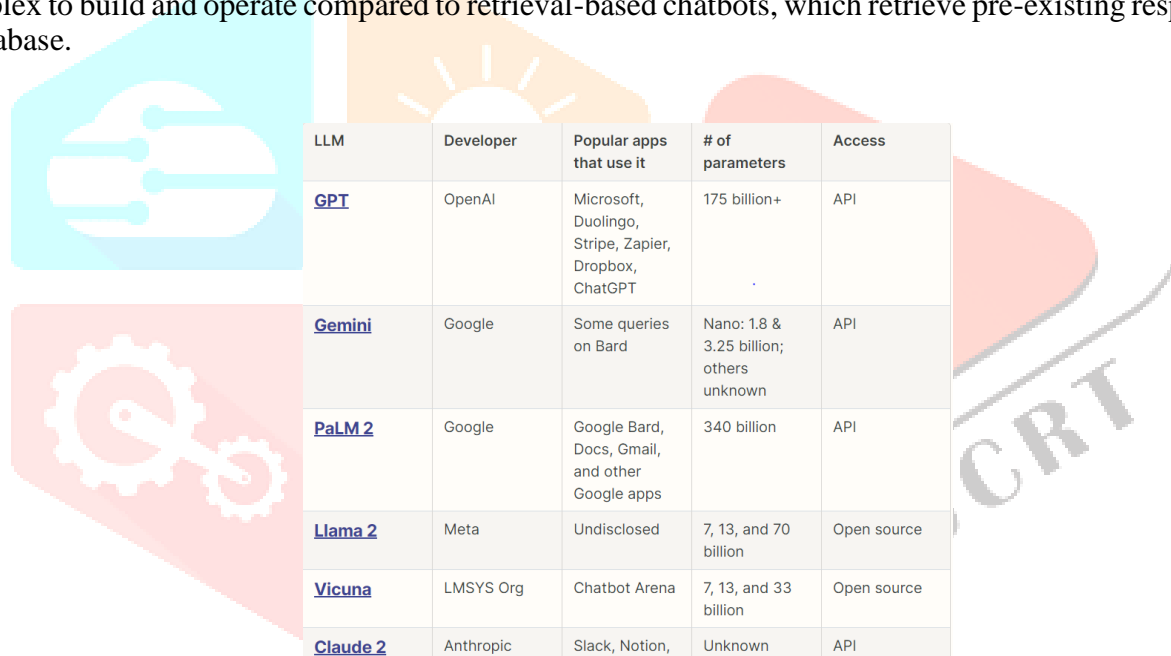


Figure 6. Proposed System Architecture[7]

This paper presents a comprehensive survey of conversational agents and chatbots, covering their broad classification and various design techniques. The evolution of human-computer interaction has led to the modernization of typical interactions, with natural language emerging as a primary input for contemporary chatbots. Natural language is increasingly recognized as a powerful enabling technology for personalization, allowing users to interact with systems using their own words rather than predetermined options. Chatbots are programmed with different strategies and algorithms to effectively respond to queries based on natural language input.

The paper identifies two main categories of chatbots: task-oriented and non-task oriented. Task-oriented chatbots are designed to perform specific tasks based on user instructions, whereas non-task oriented chatbots serve multiple purposes without task-specific capabilities. The dialogue context varies between these two types of chatbots, with task-oriented chatbots typically processing similar inputs using simple algorithms based on pattern matching, while non-task oriented chatbots require more diverse algorithms and strategies to ensure accurate responses to queries with varying contexts.

Furthermore, the paper highlights the challenges associated with generative-based chatbots compared to retrieval-based chatbots. Generative-based chatbots, which generate responses from scratch, are more complex to build and operate compared to retrieval-based chatbots, which retrieve pre-existing responses from a database.



LLM	Developer	Popular apps that use it	# of parameters	Access
GPT	OpenAI	Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT	175 billion+	API
Gemini	Google	Some queries on Bard	Nano: 1.8 & 3.25 billion; others unknown	API
PaLM 2	Google	Google Bard, Docs, Gmail, and other Google apps	340 billion	API
Llama 2	Meta	Undisclosed	7, 13, and 70 billion	Open source
Vicuna	LMSYS Org	Chatbot Arena	7, 13, and 33 billion	Open source
Claude 2	Anthropic	Slack, Notion,	Unknown	API

Table 1. Table summarizing the the top LLM models

Leveraging Large Language Models in Conversational Recommender Systems (Luke Friedman)[10]:

This paper investigates the potential of leveraging large language models (LLMs) in conversational recommender systems (CRSs) and outlines the system architecture necessary for this integration. It identifies areas where LLMs can enhance dialogue management, retrieval, ranking, and user profiles, thereby improving system quality, user control, and transparency. The focus is on developing a large-scale end-to-end CRS without relying on existing product data, utilizing LLM-powered user simulators to generate synthetic training data. A proof-of-concept system, RecLLM, is introduced to demonstrate the diverse functionality enabled by this approach. The paper outlines future directions for research, including the release of human evaluations and a public dataset to quantitatively evaluate design alternatives, generalizing the system to handle feedback from multiple channels, scaling up tuning methodologies for large item corpora, and supporting new use cases arising in conversational recommender dialogues, such as question answering over corpus items. Overall, the paper aims to accelerate progress towards controllable and explainable CRSs, fostering a healthier recommender system ecosystem.

Using Machine Learning for Web Page Classification in SearchEngine Optimization(Goran Matošević)[11]:

In this study, machine learning techniques were explored for web page classification in the context of search engine optimization (SEO). The research aimed to identify the extent to which web pages adhere to SEO guidelines by training classifiers on a dataset of 600 pages classified by SEO experts into categories of "low SEO," "medium SEO," and "high SEO." Five major classifiers were tested, including decision trees, Naïve Bayes, logistic regression, KNN, and SVM, resulting in accuracy rates ranging from 54.69% to 69.67%, surpassing the baseline accuracy of 48.83%. The study confirmed that machine learning algorithms, informed by expert knowledge, can predict web page SEO adjustments effectively. Additionally, a decision tree algorithm was employed to extract relevant SEO factors. Importantly, the methods applied were not specific to any particular search engine or language, making them widely applicable across different contexts.

The research demonstrates how machine learning can detect web page SEO quality and provide insights into important factors influencing rankings. These methods can facilitate the development of automated or semi-automated software to support SEO tasks, such as identifying pages needing optimization, suggesting optimal factor values, or detecting spammy pages. The dataset generated in this study can serve as a valuable resource for further research on SEO factors or web page classification methods. Future research directions may include exploring alternative machine learning methods, incorporating off-page factors, increasing the number of SEO experts or target classes, and further refining classification accuracy. Ultimately, the study aims to encourage SEO agencies and software developers to leverage machine learning models to enhance the efficiency and effectiveness of on-page SEO tasks.

Leveraging Large Language Models in Conversational Recommender Systems Luke Friedman[12]:

This paper provides insights into search engine optimization (SEO) using natural language processing (NLP) and machine learning (ML) techniques, aiming to enhance the ecosystem of holistic marketing. With a growing number of searches conducted by customers and digital marketers annually, the objective is to guide users towards relevant services or products by optimizing web pages for higher ranking and visibility. The authors identify research gaps and propose future work to address existing system shortcomings and leverage insights from prior literature. Firstly, they suggest developing a unique architecture integrating NLP and ML approaches to address local SEO challenges through predictive page ranking. Secondly, they propose enhancing this architecture to improve content generation efficiency in SEO, possibly employing a novel deep learning approach with feedback connections over a tree-based network system. This approach aims to process complete sequences of web page data and generate predictive data with lower computational complexities. Lastly, the authors recommend refining the model further with an improved version of ML algorithms, potentially designing a multi-objective function to optimize content based on state, reward, and actions parameters. This holistic approach aims to evolve SEO practices by leveraging advancements in NLP and ML to enhance predictive page ranking and content generation processes, ultimately improving the user experience and achieving marketing objectives more effectively.

The research demonstrates how machine learning can detect web page SEO quality and provide insights into important factors influencing rankings. These methods can facilitate the development of automated or semi-automated software to support SEO tasks, such as identifying pages needing optimization, suggesting optimal factor values, or detecting spammy pages. The dataset generated in this study can serve as a valuable resource for further research on SEO factors or web page classification methods. Future research directions may include exploring alternative machine learning methods, incorporating off-page factors, increasing the number of SEO experts or target classes, and further refining classification accuracy. Ultimately, the study aims to encourage SEO agencies and software developers to leverage machine learning models to enhance the efficiency and effectiveness of on-page SEO tasks.

Building Search Engine Using Machine Learning Technique[13]:

This paper discusses the implementation of a computer program utilizing machine learning techniques, specifically focusing on the research conducted by Rushikesh Karwa and Vikas Honmane, as presented in IEEE 2020. The study explores various techniques, including SVM-based ranking and XGBoost-based methods, comparing their performance. The aim is to improve search engine functionality by applying machine learning algorithms to deliver more relevant search results. Four modules were developed for this purpose, with the objective of reducing user search time by providing accurate and relevant URLs. The results

suggest that XGBoost outperforms SVM and ANN in terms of accuracy, making it a preferable choice for building search engines when combined with PageRank algorithms.

Furthermore, the paper presents an empirical analysis of XGBoost, a gradient boosting method known for its effectiveness. The study compares XGBoost's performance in terms of training speed and accuracy with gradient boosting and random forest across various tasks. While gradient boosting demonstrated superior performance in the majority of tasks investigated, the differences between XGBoost and random forest were negligible when using default parameters. Overall, the findings highlight the effectiveness of XGBoost in improving search engine performance, particularly when combined with PageRank algorithms, and underscore its potential as a valuable tool in machine learning-based search engine development.

BUILDING SEARCH ENGINE USING MACHINE LEARNING TECHNIQUES (ABBAGONI SAHITHYA,S. VIJAY KUMAR)[14]:

In their paper, "Building Search Engine Using Machine Learning Techniques," Abbagoni Sahithya and S. Vijay Kumar emphasize the importance of search engines in efficiently retrieving relevant URLs for given keywords, thereby reducing user search time. They highlight the crucial role of accuracy in search engine performance and conclude that XGBoost outperforms Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in terms of accuracy. Consequently, they suggest that search engines developed using XGBoost in conjunction with PageRank algorithms will yield higher accuracy in retrieving relevant web pages.

Building Search Engine Using Machine Learning Technique(Ch.Venkata Ramana1, G. Meghana)[15]:

In their paper titled "Building Search Engine Using Machine Learning Technique," Ch. Venkata Ramana and G. Meghana underscore the significance of search engines in efficiently retrieving relevant URLs based on user-input keywords. They emphasize that search engines play a crucial role in reducing user search time by providing accurate results. Through their analysis, they conclude that XGBoost demonstrates superior accuracy compared to Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Therefore, they advocate for the development of search engines utilizing XGBoost in conjunction with PageRank algorithms to achieve better accuracy in retrieving relevant web pages.

III.PROPOSED METHODOLOGY AND RESULTS

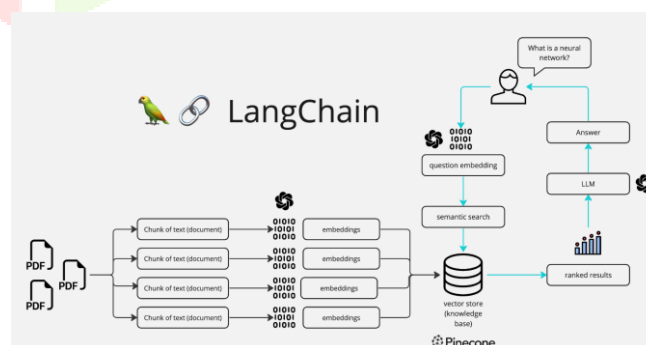


Figure 7. Proposed Methodology

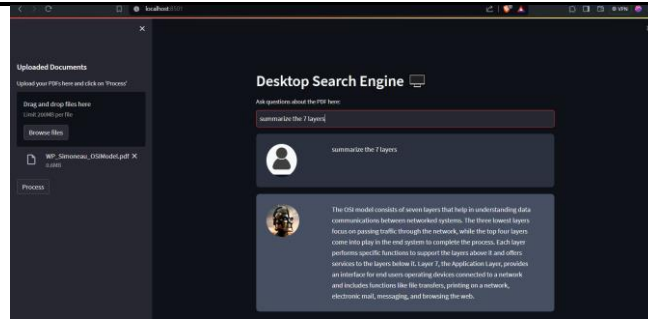


Figure 8. GUI of the chat application and the generated results for a prompt

PDF Document Processing:

Users initiate the process by uploading PDF documents to the system. These documents are subsequently processed, with their content divided into manageable chunks of text while maintaining the original structure and formatting. Each chunk represents a coherent section or paragraph of the document, ensuring that the context is preserved during further analysis.

Text Embedding and Indexing:

The next step involves transforming each chunk of text into a dense vector representation using pre-trained embedding models such as Word2Vec, GloVe, or BERT. These vector representations are then stored in a knowledge base or vector store, which could include platforms like Pinecone, Faiss, or Chroma. This indexing facilitates efficient retrieval and similarity search during subsequent interactions.

User Query Processing:

Upon receiving a query from the user pertaining to the PDF document, the system preprocesses and tokenizes the query text. Subsequently, the preprocessed query is converted into a dense vector representation using the same embedding models employed for the document chunks, ensuring consistency in the representation space.

Ranking Results and Answer Generation:

To provide relevant responses to the user query, the system performs a similarity search between the query vector and the vectors stored in the knowledge base or vector store. The retrieved document chunks are then ranked based on their similarity to the user query, with the most relevant chunks appearing at the top of the list. These top-ranked chunks are subsequently fed into Large Language Models (LLMs) such as GPT (from OpenAI) or BERT (from Hugging Face) for answer generation.

Feedback and Iteration:

Following the presentation of results, the system actively solicits feedback from the user regarding the relevance and usefulness of the provided answers. This feedback mechanism serves as a crucial component for system improvement, enabling iterative refinement of the document indexing, query processing, and answer generation mechanisms. Based on user feedback, the system can adjust its algorithms and parameters to better meet user expectations and enhance overall performance. This iterative process ensures that the system continually evolves to better serve the needs of its users, ultimately enhancing the quality and relevance of the information provided.

IV. Conclusion

In conclusion, the proposed methodology offers a comprehensive and efficient approach to facilitate user interaction with PDF documents through natural language queries. By leveraging text embeddings and Large Language Models (LLMs), coupled with efficient indexing and retrieval mechanisms, the system empowers users to seamlessly navigate through complex document structures, extract relevant information, and obtain coherent responses to their queries. Through the iterative process of feedback and refinement, the system continually evolves to better meet user needs and expectations, ensuring an optimal user experience.

Furthermore, the incorporation of feedback mechanisms enables the system to adapt and improve over time, enhancing the accuracy and relevance of the information provided to users. This iterative refinement process underscores the system's commitment to continual improvement and user satisfaction. By fostering a collaborative partnership between users and the system, the methodology facilitates knowledge discovery, decision-making, and information dissemination in diverse domains.

Overall, the proposed methodology represents a significant advancement in document interaction technology, offering a user-friendly and intuitive platform for exploring and extracting insights from PDF documents. As technology continues to evolve and improve, there is immense potential for further enhancements and innovations in this field, ultimately empowering users to unlock the full potential of PDF documents for research, education, and decision-making purposes.

REFERENCES

- [1] Conversational Document Prediction to Assist Customer Care Agents (Jatin Ganhotra, Haggai Roitman)
- [2] Introduction to AI Chatbots (Aishwarya Gupta, Divya Hathwar)
- [3] Development of Artificial Intelligence based Chatbot using Deep Neural Network (Dammavalam Srinivasa Rao)
- [4] DeepPDF: A Deep Learning Approach to Analyzing PDFs (Christopher Stahl, Steven Young)
- [5] Conversational Interfaces for Information Search (Q. Vera Liao, Werner Geyer)
- [6] Automating Pdf Interaction Using Langchain (Karan R., M. Rahul Kumar)
- [7] BUILDING A SMART CHATBOT (Mr. E. Sankar, Alekya B V)
- [8] A neural-based text summarization system (S. P. Yong,)
- [9] A Survey on Conversational Agents/Chatbots Classification and Design Techniques (Shafquat Hussain)
- [10] Leveraging Large Language Models in Conversational Recommender Systems (Luke Friedman)
- [11] Matošević, G.; Dobša, J.; Mladenčić, D. Using Machine Learning for Web Page Classification in Search Engine Optimization. *Future Internet* 2021, 13, 9. <https://doi.org/10.3390/fi13010009>
- [12] Insights into Search Engine Optimization using Natural Language Processing and Machine Learning Vinutha M S1 , M C Padma2
- [13] Building Search Engine Using Machine Learning Technique 1 Prof. Swapnil Bhanudas Wani, 2 Mr. Shivprasad Mahendrakumar Yadav,
- [14] Building Search Engine Using Machine Learning Technique 1 Abbagoni Sahithya , 2 S Vijayakumar
- [15] Building Search Engine Using Machine Learning Technique Ch. Venkata Ramana 1 , G. Meghana 2 , M. Navya Sai 3, A. Prasad 4 , V. Mohanarao 5