e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:07/July-2023 Impact Factor- 7.868 www.irjmets.com

LANGCHAIN-POWERED VIRTUAL ASSISTANT FOR PDF COMMUNICATION

NR Tejaswini*1, Vidya S*2, Dr. T Vijaya Kumar*3

*1Student, Master Of Computer Applications, Bangalore Institute Of Technology, Bangalore, India.

*2Assistant Professor, Master Of Computer Applications, Bangalore Institute Of Technology, Bangalore, India.

*3Professor, Master Of Computer Applications, Bangalore Institute Of Technology, Bangalore, India.

DOI : https://www.doi.org/10.56726/IRJMETS43587

ABSTRACT

Due to the unstructured nature of the PDF document format and the requirement for precise and pertinent search results, querying a PDF can take time and effort. LangChain overcomes these challenges by utilizing advanced natural language processing algorithms that analyze the content of the PDFs and extract essential information. To improve the search experience, it uses effective indexing and retrieval techniques, movable filters and a simple search interface. LangChain also allows users to save queries, create bookmarks and annotate important sections, enabling efficient retrieval of relevant information from PDF documents. The features of LangChain increase overall efficiency and makes PDF querying much easier and simpler.

Keywords: LangChain, Querying PDF, Machine Translation, Streamlit.

I. INTRODUCTION

The increasing prevalence and usage of digital products have created challenges in searching and retrieving information from PDF documents. However, a revolutionary tool called LangChain, built on Natural Language Processing (NLP) and Large Language Models (LLM), addresses these challenges. LangChain simplifies the querying process and information extraction from PDFs using advanced NLP algorithms. To create a user-friendly interface, LangChain utilizes Streamlit, a web application framework that eliminates the need for expertise in other web development frameworks like HTML and CSS. Streamlit enables the seamless deployment of models with minimal coding effort. With LangChain and Streamlit, users can easily interact with PDFs, making document search and retrieval significantly more convenient. The PDF Query App Project uses Language Models (LLMs) and LangChain, a cutting-edge language processing tool, to transform how users interact with PDF documents. By allowing users to have interactive conversations with PDF documents, this project solves the fundamental drawbacks of conventional PDF readers. This description highlights the problem statement and the solution while also giving a general overview of the project.

II. LITRATURE SURVEY

The literature review for the project "LangChain PDF Query" focuses on exploring relevant research and technologies related to language models, natural language processing, AI advancements and query systems.

Jonas Gehring et.al.,[1] In the realm of sequence-to-sequence learning, presented "Convolutional Sequence to Sequence Learning," which showcased a novel approach to modeling sequences using convolutional neural networks, offering insights into how language structures can be better understood and represented. Attention mechanisms have revolutionized language models, as demonstrated in "Attention Is All You Need" by Ashish Vaswani et. al., [2]. Their transformative "transformer" architecture enabled highly efficient and effective attention-based models for various NLP tasks, paving the way for advanced query systems. While exploring AI applications in the legal domain, Jules Ioannidis et al., [3] introduced "Gracenote.ai," an AI system designed for regulatory compliance. This research highlights the potential of generative AI in addressing complex legal challenges. For specific use cases, Na He et al., [4] showed that "Chat GPT-4" outperformed GPT-3.5 in drug information queries, indicating the continuous advancements in language models for domain-specific information retrieval. Efficient keyword search over encrypted cloud data is crucial for data privacy and security. Anuradha & Patil., [5] presented an approach for such search, contributing insights into secure information retrieval in cloud environments.

Addressing the challenges of big data, Madhu Nashipudimath et al., [6] proposed an integration and indexing method based on feature patterns and semantic analysis. This research offers valuable techniques to efficiently manage and query large datasets. In the scientific literature domain, Zhu and Cole., [7] developed a tool for reading scientific text and interpreting metadata from typeset literature in PDF format. This tool demonstrates advancements in information extraction from scientific documents. To implement the project, it is essential to consider the platforms and tools available. Streamlit., [8] provides a user-friendly interface for data visualization and interaction, while Python LangChain., [9] offers comprehensive documentation for integrating language models into applications. OpenAI's models [10], including GPT-3.5, serve as a crucial foundation for language-based AI applications. Adith Sreeram A S and Pappuri Jithendra Sai, [11] presented an effective query system using language models and LangChain, which can be insightful for developing the LangChain PDF Query.

In conclusion, this literature review has provided a comprehensive understanding of the research and technologies relevant to the project "LangChain PDF Query." It covered sequence-to-sequence learning, attention-based models, AI in legal compliance, domain-specific information retrieval, secure cloud data search, big data processing, scientific literature analysis and essential tools for implementation. These valuable insights will guide the development of an efficient and effective query system using LangChain and language models, contributing to the advancement of AI-driven natural language processing.

III. METHODOLOGY

LanChain helps us with the querying process and extracting information from the PDF based on the prompt sent by the user. For the sake of convenience, a web application is developed that can retrieve accurate information based on the user's input alone.

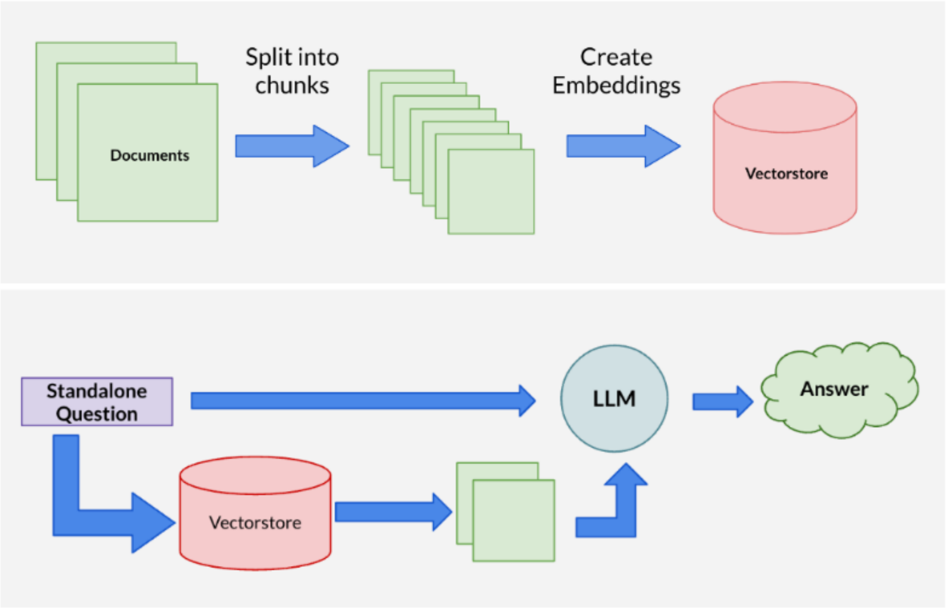


Figure 1. Application Architecture

2.1 Steps followed in the Application Architechture:

Step l: The Open Al Large Language Models and The Open Al Embeddings acts as the back-end of our application.

Step II: Here we will use Streamlit, which will help us to build interactive and beautiful interface for our web application.

Step Ill: Streamlit will also take care of our Front-end part where we can get the text inputs and messages and also the PDF files from the user.

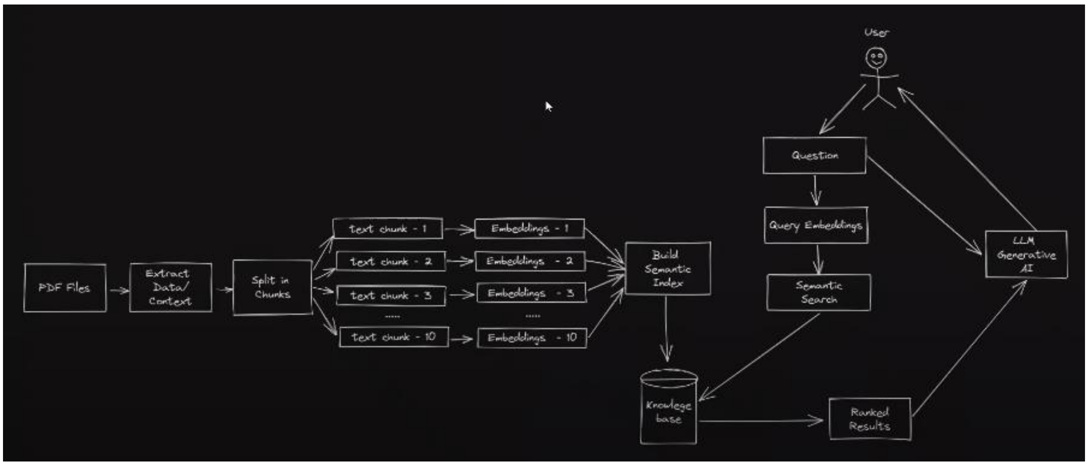


Figure 2. Working Process.

With the help of Fig 2 we can understand how Large Language Model helps the user to get the accurate results.

2.2 Streamlit

Streamlit is an open-source library that allows us to unique web apps for Machine Learning and Data Science projects fast and efficient. Streamlit is an open-source library that allows us to unique web apps for Machine Learning and Data Science projects fast and efficient. With this framework, you can easily build interactive visualization plots, models and dashboards without having a worry about the underlying web framework or deployment infrastructure used in the backend. It also provides the users to add widgets which helps the users the interact with the web app and the models that we used. This framework also integrates the popular python and machine learning packages such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn and TensorFlow, which enables us to quickly build and deploy our trained models. Features of Streamlit:

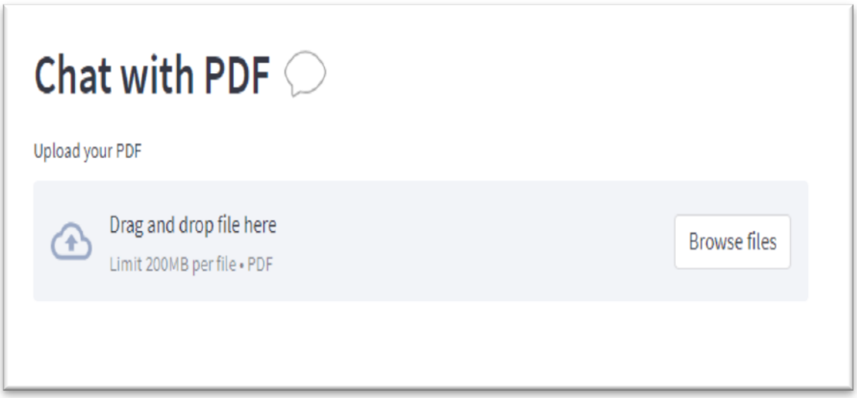
User-friendly: Streamlit offers an easy-to-use interface that requires little scripting to build dynamic data apps.

Rapid prototyping: Streamlit is made for rapid prototyping, allowing developers and data scientists to test out various concepts and create completely functional apps.

Data Cache: The data cache facilitates and accelerates computational workflows.

Real-time collaboration is made possible by Streamlit, allowing several users to work on the same project at once. Widgets that enable for real-time data editing and exploration include sliders, dropdown menus and checkboxes, among a vast variety of interactive widgets that Streamlit offers.

IV. RESULTS AND DISCUSSION


Chat with PDF

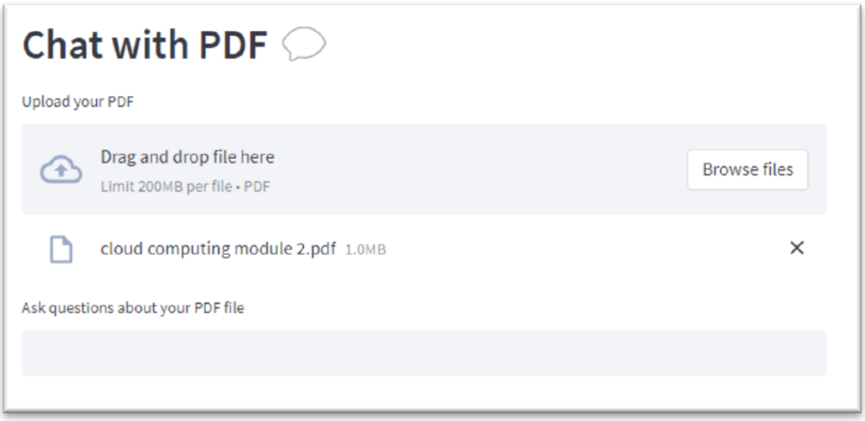
Upload your PDF

Drag and drop file here

Limit 200MB per file • PDF

Figure 3. Interface of web Tool

This is how the interface of our web Tool will look like. Now the user can click on browse files and can upload a file from their device under 200 Mega Bytes. After few minutes of processing, we will get an additional in box where we can give in our query.



Chat with PDF

Upload your PDF

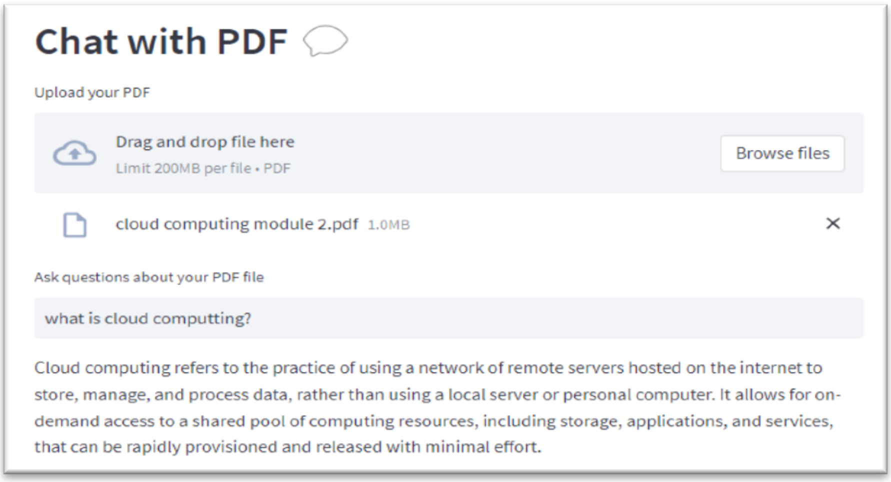
Drag and drop file here
Limit 200MB per file • PDF

cloud computing module 2.pdf 1.0MB

Ask questions about your PDF file

Figure 4. Image of web Tool with input query box.

So, now we got our input query box and now we can ask questions on the PDF that we have uploaded. Here I have uploaded a PDF based on Cloud Computing. Now you can ask different questions like "What is cloud computing?", "What are the Architectural styles based on independent components?" and also differentiate between questions.



Chat with PDF

Upload your PDF

Drag and drop file here
Limit 200MB per file • PDF

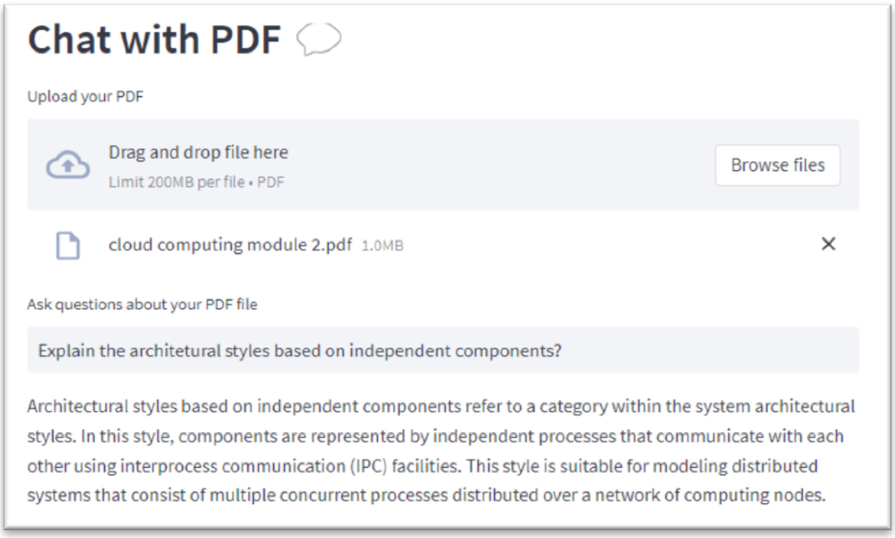
cloud computing module 2.pdf 1.0MB

Ask questions about your PDF file

what is cloud computting?

Cloud computing refers to the practice of using a network of remote servers hosted on the internet to store, manage, and process data, rather than using a local server or personal computer. It allows for on-demand access to a shared pool of computing resources, including storage, applications, and services, that can be rapidly provisioned and released with minimal effort.

Figure 5. The Output that we got for our 1st Query



Chat with PDF

Upload your PDF

Drag and drop file here
Limit 200MB per file • PDF

cloud computing module 2.pdf 1.0MB

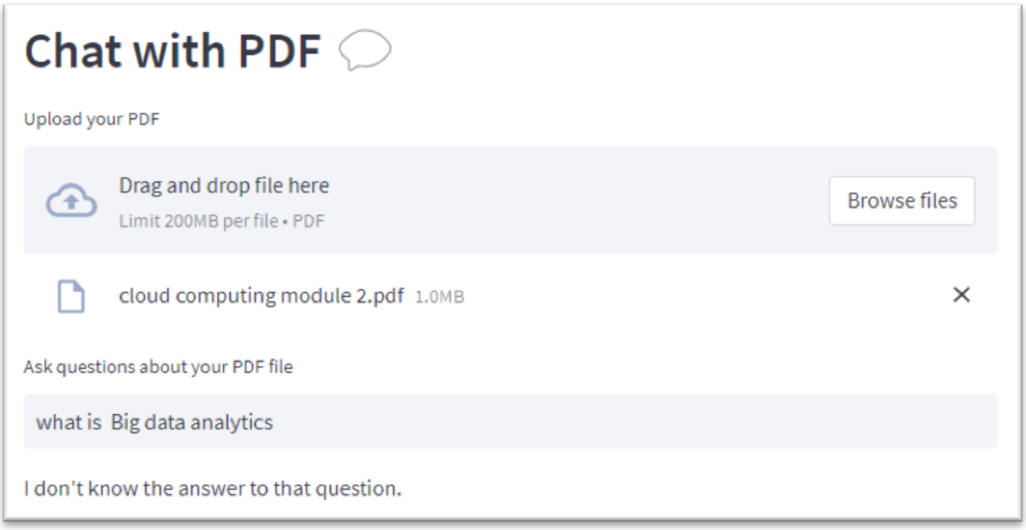
Ask questions about your PDF file

Explain the architetural styles based on independent components?

Architectural styles based on independent components refer to a category within the system architectural styles. In this style, components are represented by independent processes that communicate with each other using interprocess communication (IPC) facilities. This style is suitable for modeling distributed systems that consist of multiple concurrent processes distributed over a network of computing nodes.

Figure 6. The Output that we got for our 2 nd Query

Here we got our output for our 1st and 2nd query which is “What is Cloud Computing?” and “What are the Architectural styles based on independent components?” our Large Language Model went through file and gave an accurate result on the query given.



Chat with PDF
Upload your PDF
Drag and drop file here
Limit 200MB per file • PDF
Browse files
cloud computing module 2.pdf 1.0MB
×
Ask questions about your PDF file

what is Big data analytics

I don't know the answer to that question.

Figure 7. The output we got for Different Question

The output "I don't know the answer to the question" indicates that the Large Language Model couldn't find an accurate response to the query "what is big data analytics?" within the provided PDF. Despite analyzing the content of the file, the model couldn't retrieve a satisfactory result in relation to the specific query.

V. CONCLUSION

The Tool that leverages LangChain, Large Language Models and Streamlit to streamline the extraction of pertinent information from PDFs. This innovative solution significantly simplifies and enhances the process, allowing users to retrieve any desired information from PDF documents while saving time and effort. By integrating LangChain technology, the app introduces a heightened level of efficiency and accuracy to the querying process, making it an invaluable tool for individuals working with PDFs. Users can effortlessly extract relevant data, improving productivity and reducing the manual effort traditionally required for PDF document analysis. The user-friendly interface and intuitive features provided by Streamlit further enhance the overall user experience. Our web application empowers users to efficiently navigate and retrieve information from PDFs, transforming the way PDF querying is performed and revolutionizing the accessibility and usability of PDF documents.

VI. REFERENCES

[1] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin: “Convolutional Sequence to Sequence Learning”, arXiv: [v1] Mon, 8 May 2017 23:25:30 UTC (1,489 KB) [v2] Fri,12May 2017 16:14:26 UTC (492 KB) **[v3]** Tue, 25 Jul 2017 01:40:57 UTC (492 KB).

[2] Ashish Vaswani, Noam Shazeer Niki Pannar, Jakob Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: “Attention Is All You Need” arXiv: [v1] Mon, 12 Jun 2017 17:57:34 UTC (1,102 KB)[v2] Mon, 19 Jun 2017 16:49:45 UTC (1,125 KB)[v3] Tue, 20 Jun 2017 05:20:02 UTC (1,125 KB)[v4] Fri, 30 Jun 2017 17:29:30 UTC (1,124 KB)

[3] Jules Ioannidis, Joshua Harper, Ming Sheng Quah I and Dan Hunter I: “Gracenote.ai: Legal Generative AI for Regulatory Compliance” v1 Gracenote-ai- Melboume Australia v2 The Dlckson Poon School of Law, King's College London, United Kingdom June 19, 2023.

[4] Na He, Yingying Yan, Ziyang Wul Cheng, Fang Liu:” Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries”, National Library of Medicine, June 2023.

[5] Anuradha & Patil, G A (2016):” Efficient Keyword Search over Encrypted Cloud Data- Procedia Computer Science.”, Science Direct, Volume 78, 2016, Pages 139-145,23-02-2016.

[6] Madhu Nashipudimath, Subhash Shinde, Jayshree Jain: "An efficient integration and indexing method based on feature pattens and semantic analysis for big data.", Research Gate,June 2020.

[7] Zhu, Miao & Cole, Jacqueline. (2022): "A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format." Joumal of Chemical Information and Modeling, March 29, 2022.

[8] https://streamlit.io/, May 2023.

[9] https://python.langchain.com/docs/get_started/introduction.html, April 2023.

[10] https://platform.openai.com/docs/models, June 2023.

[11] Adith Sreeram A S, Pappuri Jithendra Sai: "An Effective Query System Using LLMs and LangChain", IJERT, olume 12, Issue 06 (June 2023).