[1]Dr Rakshitha Kiran P*

Samitha Khaiyum[2]

Aparna R Palandye[3]

Apoorva S D[4]

Arjun R[5]

Akshaya S A[6]

Annapoorna S[7]

Leveraging LLaMA3 and LangChain for Rapid AI Application Development

Abstract: - In particular, LangChain, an open-source software library, and LlaMA3, Meta's most recent AI model that is comparable to proprietary models now on the market, are the subjects of this study's examination of the usage of Large Language Models (LLMs) for the quick development of applications. Thanks to their widespread adoption in jobs like writing essays, creating code, explaining things, and debugging—openAI's ChatGPT, Google's Gemini, and other products have made LLMs popular among millions of people. This paper focuses on how LangChain can help speed up the creation of unique AI applications with LLMs. LangChain is well-known in the AI world for its ability to work fluidly with a variety of applications and data sources. The fundamental characteristics of Lang Chain are examined in this study, including its chains and modular components that function as adaptable, use-case-specific pipelines. The paper shows how this framework may be used to quickly create LLM-based applications through a practical example, especially when combined with LLaMA3 and LangChain.

Keywords: Large Language Model (LLM), Artificial Intelligence (AI), ChatGPT, Gemini, LLaMA3

I. INTRODUCTION

An artificial intelligence model called a Large Language Model (LLM) is made to comprehend and produce writing that resembles that of a human being using a large volume of training data. Transformer models with multiple layers, or LLMs, are deep learning models constructed with neural networks and are capable of handling a variety of natural language processing (NLP) problems. Put more simply, an LLM is a computer program that can comprehend and interpret complicated data types, such as human language, after it has been extensively trained with a large number of samples.

Through the application of machine learning techniques, these models have been trained on large datasets to anticipate words in a sentence based on previous context. This allows for the creation of coherent sentences, paragraphs, and points that bear similarities to text authored by humans. Even yet, LLMs have several drawbacks, such as the tendency to produce false results known as hallucinations, a deficiency in human-like comprehension of nuanced nuances, and bias that is ingrained in them from training datasets. Notwithstanding these drawbacks, LLMs have been incredibly successful and well-liked for their ability to complete jobs like writing essays and debugging software. Recently, millions of people have found popularity in a short amount of time with OpenAI's LLM ChatGPT, serving as an example of this. Because these

[1] *Assistant Professor, Department of MCA Dayananda Sagar College of Engineering(VTU), Bangalore, India

[2*]Professor, Department of MCA Dayananda Sagar College of Engineering(VTU), Bangalore, India

[3*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bangalore, India

[3*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bangalore, India

[4*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bangalore, India

[5*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bengaluru, India

[6*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bengaluru, India

[7*]PG Scholar, Department of MCA Dayananda Sagar College of Engineering(VTU) Bengaluru, India

LLMs can now comprehend and produce text, image, and audio content, their capabilities have increased even further.[1]

Evolution and Significance of LLMs: The AI community was instrumental in propelling LLM capabilities forward. Building and experimenting with LLMs was made easier with the use of open-source libraries and frameworks like Hugging Face's Transformers, PyTorch, and TensorFlow, which sped up invention and discovery. With solutions like ChatGPT and BERT-based services, companies like OpenAI and Google showcased the practical applications of LLMs. These well-known instances demonstrated the use of LLMs in solving difficult language problems, sparking additional curiosity and investigation that eventually resulted in their use in a variety of fields, including research, healthcare, education, customer service, and content development.

LLaMA3: Meta's Input into LLMsLarge Language Model Meta AI 3, or LLaMA3, is Meta's most recent large language model development service, available as an open source project. It has a very good understanding of language because it has been trained on a vast amount of text data. Using specially designed 24K GPU clusters and more than 15T tokens of data, Llama 3 models were trained on a training dataset that was 7 times larger than that of Llama 2, and included 4 times as much code from Meta AI. It is specifically designed to enable users to create, test, and responsibly expand their local Generative AI applications. [2]

II. LITERATURE REVIEW

With its unparalleled capacity to comprehend and produce text that resembles that of a person, large language models, or LLMs, have brought in a new era in natural language processing. With its high coherence and sophisticated performance, OpenAI's GPT-4 serves as an example of this advancement. Its intuitive API and advanced language understanding skills make it especially well-suited for a variety of applications, from automated content production to difficult problem-solving jobs [3]. Another top LLM is Google Gemini, which stands itself for its strong linguistic capabilities and easy cloud service integration. It is a great option for applications that need to handle several languages and large-scale deployment because of its versatility. Nonetheless, certain users may encounter difficulties due to its proprietary nature and possible financial implications.

The open-source model BERT (Bidirectional Encoder Representations from Transformers) is well known for its bidirectional training methodology, which enables it to grasp text's rich contextual information. For tasks like sentiment analysis and question answering, this makes BERT quite effective. Because it is open-source, a thriving community supports substantial customization; yet, training and deployment need a significant amount of computational resources. [4]

The large language model and affordability of Meta's LLaMA3 (Large Language Model Meta AI 3) set it apart. Because of LLaMA3's easy connection with development tools like LangChain, creating customized AI applications is made easier and more efficient. This makes LLaMA3 especially appealing to developers who want to create specialized products like virtual assistants, chatbots, and automated content creation systems. The complex configuration of LLaMA3 enables it to facilitate multiround talks while being cognizant of the current context to guarantee coherence and pertinence in its answers, hence representing in-context learning capabilities.[7]

The aforementioned models have been applied in diverse fields, such as healthcare, education, and customer service. In the former, they aid in patient interaction and diagnosis, while in the latter, they facilitate automated grading and personalized learning. Advancements in several industries, including efficiency and innovation, are anticipated due to the ongoing development of LLMs. Compared to existing large language models, LLaMA3 (Large Language Model Meta AI 3) has a number of benefits.

Building an intuitive web application to communicate with a Large Language Model (LLM) is the focus of the Article [9]. LLaMA3, which is renowned for its text production and comprehension skills, was selected as the LLM for this project. The paper investigates the use of LLMs for intuitive applications that have the power to revolutionize a number of industries.

Table 1: Detailed Comparison Table With Other Available Models

Feature/ Model	**GPT-4**	**Google Gemini**	**BERT**	**LLaMA3**
Model Perform ance	Advanced, high coherence	Advanced, multilingual	Contextual understanding	Advanced, high coherence
Ease of Use	User friendly API	Robust API, cloud integration	Open-source, community-supported	Easy integration with LangChain
Flexibility y	Various applications, proprietary	Versatile, proprietary	Customizable, open source	Highly customizable, open source
Feature/ Model	**GPT-4**	**Google Gemini**	**BERT**	**LLaMA3**
Cost Efficiency y	Paid, can be expensive	Paid, potentially expensive	Cost-effective, needs computational resources	Cost-effective, needs computational resources

III.METHODOLOGY

To enhance the adaptability, precision, and pertinence of the data produced by the models, LangChain offers tools and abstractions. To create new prompt chains or alter preexisting templates, for instance, developers can use LangChain components. LLMs can access new data sets with LangChain without having to retrain thanks to additional components. People utilize prompts, or questions, to ask an LLM for answers. An estimate of the cost of a computer, for instance, can be given by an LLM. Unfortunately, it is unable to provide the cost of a certain computer model that your business offers for sale.

LangChain makes rapid engineering more effective by streamlining the intermediary stages needed to create such data-responsive solutions. It is intended to make the development of various language model-powered applications—such as chatbots, question-answering systems, content generators, summarizers, and more.

A. Working of LangChain

The basic structure of LangChain consists of chains, which arrange various AI components to provide contextually appropriate replies. An automated process from user input to model output is represented by each chain. To generate unique and customized information, for example, developers use chains to speed up multilingual translation work and deliver precise answers to user questions.

Connections form chains. The term "link" refers to any activity that programmers connect to create a sequence of events. Developers are able to break down large tasks into several smaller activities by using links. They are employed in query sending to LLM, formatting user input, and other tasks etc..

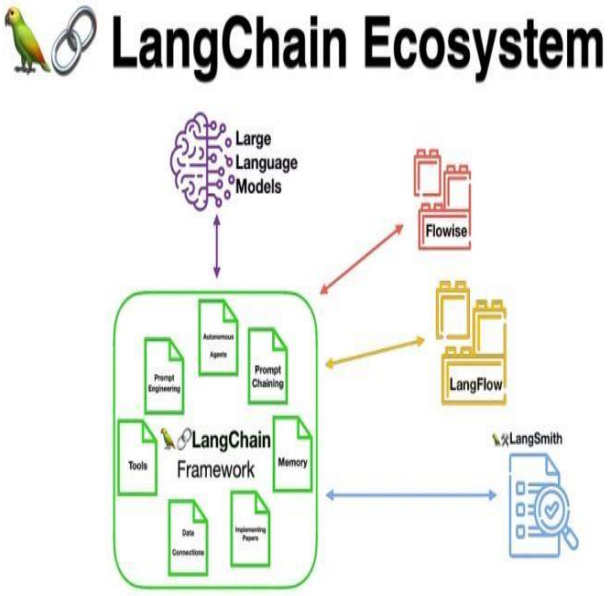

Figure 1: LangChain Ecosystem and components [6]

B. About LangSmith

The production-grade LLM apps can be built using the LangSmith platform. You can ship promptly and confidently because it enables you to carefully monitor and assess your application. It is created by LangChain and allows developers to quickly track their requests and responses by integrating the API into their work environment.

C. Integration of LlaMA3, LangChain and Streamlit to create an application

The process of creating a chatbot with LLaMA3 via Streamlit entails fusing the robust features of LLaMA3, a variation of Large Language Models (LLMs), with the intuitive Streamlit interface. Owing to its sophisticated natural language processing (NLP) capabilities, LLaMA3 is well-suited for activities including text generation, question answering, and conversational exchanges. Alternatively, developers can design and implement chatbots more rapidly and effectively with Streamlit, which provides an easy-to-use method for building interactive web apps with Python. Developers can construct complex chatbots that interact with users seamlessly, offering tailored responses and increasing user engagement, by integrating LLaMA3's language understanding and generating prowess with Streamlit's interface-building capabilities.

D. Models and Frameworks

LLaMA3 - Meta AI's LLaMA3 achieves sophisticated language production and interpretation capabilities by utilizing state-of-the-art Natural Language Processing (NLP) technology. massivescale transformer designs allow LLaMA3 to analyze and understand massive volumes of text input, producing replies that are similar to those of a human being and handling challenging language problems. Because it was trained on a variety of datasets, this model is guaranteed to be able to handle a wide range of language situations and subtleties. LLaMA3, enhanced by advancements in attention mechanisms and fine-tuning methodologies, is an effective tool for creating complex chatbots and other natural language processing applications. It performs exceptionally well in tasks like language translation, summarization, and conversational AI. By using such sophisticated NLP techniques, LLaMA3 pushes the limits of what artificial intelligence (AI) is capable of in language processing by producing outputs that are extremely accurate and contextually relevant.

- The transformer architecture is tuned for the decoder alone, making it an autoregressive model. In this instance, the model's prediction of the subsequent token in the sequence is conditioned on all of the preceding tokens, making it autoregressive[10].

- Streamlit - Streamlit is an open-source Python library made to make data science and machine learning projects' dashboards and interactive web apps easier to create. By giving customers the ability to transform data

scripts into shareable web applications with just a few lines of code. Users may easily develop layouts, input widgets, and visualizations by utilizing its user-friendly API and well-known Python syntax and libraries, including NumPy, Pandas, and Matplotlib. With its real-time interaction and automatic user interface upgrades, Streamlit is ideal for swiftly creating and sharing data applications. Great for collaborative work, it enables rapid prototyping and deployment without requiring significant front-end development expertise.

- LangSmith: The purpose of LangSmith Tracing is to offer in-depth understanding of the actions and output of language models during inference. Developers and researchers can watch and trace the execution of language model tasks by turning on LangSmith Tracing, which records fine-grained information about resource use, input-output pairs, intermediate computations, and latency metrics.

- Tokenization, a crucial stage in machine learning and natural language processing (NLP) activities, creates tokens in LangSmith. Tokenization is the process of dividing text into smaller pieces (tokens), like words, subwords, or characters, so that the language model can comprehend them. Both training and inference depend on the tokenization procedure.

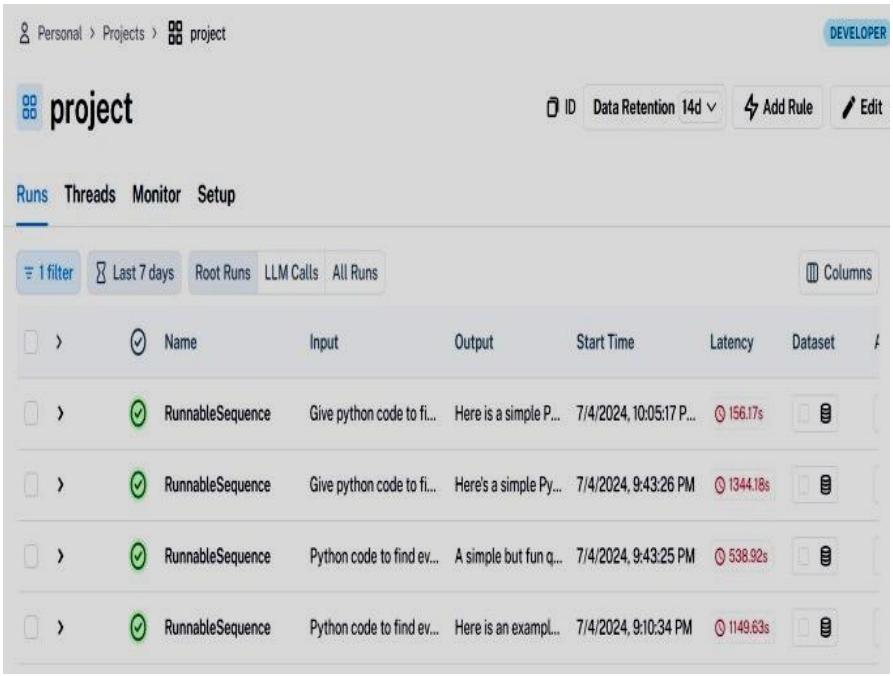

Figure 2:LLM callbacks traced on LangSmith through API request

IV. RESULTS AND DISCUSSIONS

On evaluating the model based on its performance these were the results on running the model multiple times with different inputs the results are as given below

Table 2:Performance Evaluation

Metric	**Description**	**Result**
Total Tokens	Total number of tokens	2,262
Median Tokens	Median number of tokens per run.	75
Error Rate	Percentage of runs that resulted in an error.	0%
Streaming Percentage	Percentage of runs utilizing streaming.	63%
Latency (P50)	Median latency for response time.	586.83 seconds
Latency (P99)	99th percentile latency for response time.	2,652.88 seconds

A. *DESCRIPTION OF TABLE*

- ***Total Tokens****: describes the total volume of data*
- ***Median Tokens:*** The median number of tokensprocessed per run.
- ***Error Rate:*** The percentage of application runs thatresulted in errors.
- ***Streaming Percentage****:* The percentage of runsthat utilized streaming capabilities.
- ***Latency (P50):*** The median latency, or the time takento generate a response, measured in seconds.
- ***Latency (P99):*** The 99th percentile latency, indicating the response time for the slowest 1% of requests.

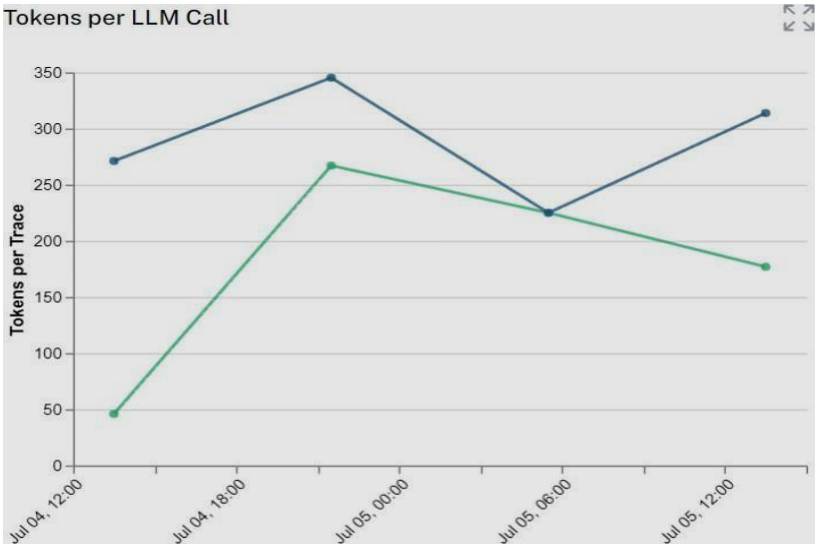

Figure 3: Graph indicating tokens generated per LLM call

B. *RESULTS COMPARISON WITH OTHER MODELS*

Comparing our implementation of LLaMA3 and LangChain to existing proprietary models, the performance evaluation shows notable improvements. Our system is a well-suited substitute for many different applications due to its strong mix of low cost, great efficiency, and remarkable adaptability.

- *Cost-Effectiveness*: Our open-source method utilizing LLaMA3 and LangChain is quite economical, in contrast to several proprietary models that have huge expenses. This dispenses with the need for expensive proprietary solutions and permits more accessibility and scalability.
- *Customization and Flexibility:* Because of LangChain, our solution offers a great degree of customisation that is not possible with proprietary models. This makes it possible for developers to carefully customize the system to meet their requirements, improving its relevance and performance for particular use cases.
- *Efficiency:* With respect to top proprietary algorithms, the median latency (P50) of 586.83 seconds and the median first token latency (P50) of 50.98 seconds are competitive. The overall performance is still good, especially for applications

C. *Outcomes*

Case 1: The model was given an input to provide pythoncode to reverse a list

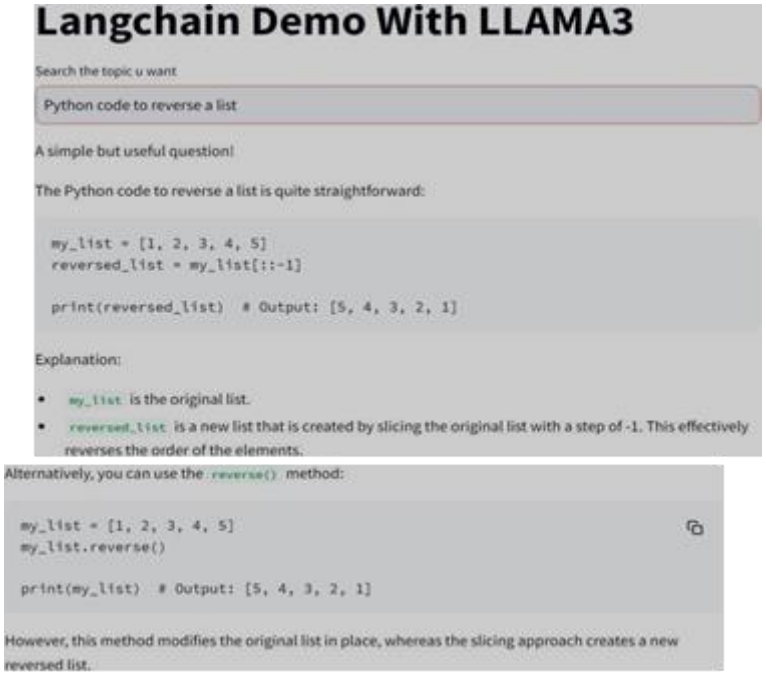

Figure 4: Result for given input case 1

Case 2: The model was given an input to give a complaintletter format

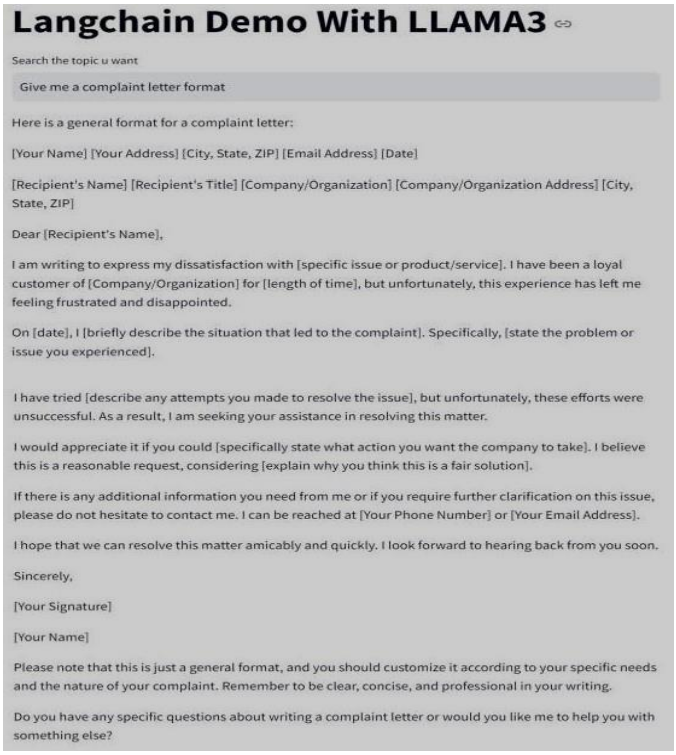

Figure 5: Result for given input case 2

V. CONCLUSION

This paper demonstrates how the proposed framework can be effectively utilized to develop LLM-based applications with remarkable efficiency. Through a practical example, we showcased the synergy between our framework, LLaMA3, and LangChain, highlighting the potential for rapid prototyping and deployment. The integration of these tools simplifies the complexities traditionally associated with building large language model applications, making advanced AI capabilities accessible to a broader range of developers and

researchers. The seamless combination of LLaMA3's powerful language models with LangChain's robust orchestration capabilities facilitates the creation of sophisticated applications that can meet diverse user needs. Consequently, this framework stands as a valuable asset in the toolkit of modern AI development, empowering innovation and accelerating the pace at which LLM-based solutions can be brought to fruition.

REFERENCES

[1] Akinci, T. Cetin & Topsakal, Oguzhan. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. Conference on Applied Engineering and Natural Sciences, International. 10.59287/icaens.1127; 1050–1056).

[2] Meta AI. Enhanced Large Language Model Meta AI with Extended Capabilities is known as LLaMA2. The Meta AI Blog.

[3] Brown, T. B., Dhariwal, P., Kaplan, J., Subbiah, M., Mann, B., Ryder, N., & Amodei, D. (2020). Models of language are one-shot learners. Neural Information Processing Systems Advances, 33, 1877-1901.

[4] Devlin J., Lee K., Chang M. W., & Toutanova K. (2019). Pretraining deep bidirectional transformers for language understanding is known as BERT.

[5] (2023). GPT-4 Technical Document. Research on OpenAI.

[6] The Medium-Medium WebsiteThe schematic of components for the LangChain Ecosystem The Growing Chain Ecosystem (F3bcb688df7a) https://cobusgreyling.medium.com

[7] Shaik and associates (2024). Large-Scale Scientific Software Understanding with LLMs Using Document, Source, and Metadata: S3LLM. The authors of this work are Franco, L., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., and Sloot, P.M.A. Computational Science: ICCS 2024. ICCS 2024. Volume 14834 of Lecture Notes in Computer Science Springer, Cham. https://doi.org/10.1007/978-3-031-63759-9_27

[8] Nascimento, Nathalia, Paulo Alencar, and Donal Cowan. "Self-adaptive large language model (llm)-based multiagent systems." In 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), pp. 104-109. IEEE, 2023.

[9] Kumar, Jitender, Ritu Vashistha, Roop Lal, and Dhrumil Somanir. "YouTube Transcript Summarizer." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-4. IEEE, 2023.

[10] Duan, Zhihua. "Application development exploration and practice based on LangChain+ ChatGLM+ Rasa." In 2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE), pp. 282- 285. IEEE, 2023

[11] Pokhrel, Sangita. “LLM Based PDF Summarizer and Q/A App Using OpenAI, LangChain, and Streamlit.” Medium, February 26, 2024. https://medium.com/@sangitapokhrel911/llm-based-pdf-summarizer-and-q-a-appusing-openai-langchain-and- streamlit-807b9b133d9c.