

# RAG-based Chat Application using LLMs: A Case Study of Vikram Sarabhai Library IIM Ahmedabad

Bhavesh Patel

B B Chand

## Abstract

*As a customer-centric organization, the library constantly focuses on providing the best user experiences with respect to accessing quality information in the shortest possible time. In the current digital world, the library needs to have communication and support methods that are both efficient and real-time to ensure a high level of user satisfaction and engagement. The Vikram Sarabhai Library (VSL) designed a pilot project to develop a cutting-edge Retrieval Augmented Generation (RAG) based chat application using Large Language Model (LLM) technology. This paper presents an abstract of the project, highlighting its objectives, methodology, key features, and anticipated impact.*

*The primary objective of the pilot project is to revolutionize user interaction and support services by harnessing the power of RAG-based chat capabilities. By combining advanced Natural Language Processing (NLP) techniques with Gen AI capabilities, the application aims to provide personalized and accurate responses to user queries in real-time from Authentic sources. Key features of the chat application include retrieval-based responses for factual queries, generation-based responses for nuanced inquiries, personalized recommendations based on user preferences, and seamless integration with VSL's existing knowledge base. By leveraging RAG-based techniques and custom datasets, the chat application aims to significantly enhance user satisfaction, streamline information retrieval processes, and foster a more engaging and interactive library experience for patrons.*

*The Retrieval-Augmented Generation (RAG) based custom datasets chat application using LLMs is a testament to VSL's commitment to innovation and excellence in user-centric services. Through this pilot project, VSL sets new benchmarks in digital communication, support, and knowledge dissemination within the library community.*

**Keywords:** Retrieval-Augmented Generation (RAG), Generative AI, Chatbot, Large Language Models (LLMs), Python, LangChain, Vector embedding, Prompt Template, Vector Database, Streamlit



## 1. Introduction

As a customer-centric organization, the library constantly focuses on providing the best user experiences with respect to accessing quality information in the shortest possible time. In the current digital world, the library needs to have communication and support methods that are both efficient and real-time to ensure a high level of user satisfaction and engagement. Evolving technology has provided the necessary support with tools that help to retrieve accurate and reliable information and are easy to implement and use. Machine learning, big data, cloud computing, and artificial intelligence are tools that have added to the powerful computing capabilities to enhance user experience concerning information search and discovery.

Artificial Intelligence-based information retrieval has recently gained significant importance, especially in the commercial world. Library Information and solution providers are not far behind. The application of artificial intelligence-based information search and retrieval for high-quality and accurate information is also progressing rapidly in the library world. Libraries across the globe are engaged in research on implementing AI-based solutions for easier access to information with the least human intervention. Retrieval-augmented generation (RAG) based chat application using Large Language Model (LLM) technology is one such solution that is considered more suitable for small applications.

Today, in the era of Generative AI, Large Language Models (LLMs), like ChatGPT, LLama, and Google Gemini are becoming very popular in scholarly communication. All the Large Language Models (LLMs) are pre-trained with massive data, including General knowledge, math, science, history, medical, open data, websites, etc. So, all LLMs are capable of understanding and working with human language.

As LLMs are trained with a large amount of massive data so that it can be reacted very accurate based on the user prompt; here, the prompt is a kind of human asking questions to LLMs to get the desired response. LLMs can answer any kind of question asked by users, but it cannot generate a proper answer in some circumstances. For example, real-time information is essential because models are pre-trained on a specific date, and after that, whatever event happens, it may not be able to give correct information. In addition, some domain-based specific information LLMs failed to provide specific answers, and it might be given general answers based on their trained data. So, LLMs don't perform well with particular topics or new (latest) information (Mukherjee, 2024).

To overcome this problem, experts created methods that search in specific knowledge bases, also known as Retrieval-Augmented Generation (RAG) (Sebastien, 2024). This paper highlights the importance of RAG-based applications with a whole implementation process for integration with our own PDFs and generating accurate and authentic responses based on the prompt given by the user.

## 2. Literature Review

The Vikram Sarabhai Library through its comprehensive collection of print and digital resources committed to providing extensive access to information and this commitment is reflected in the range of services

provided by it. The library plays a crucial role in fostering the academic and research agenda of the institute by providing efficient and timely research support to the user community. Online databases are accessible from networked computing device anywhere within the institute premises. The Vikram Sarabhai Library proposed a RAG-based chat application to facilitate ready reference service and generate precise and authentic information from its customized knowledge-base database by using the power of Large Language Model and serve the same to IIMA community members through interactive chat bot service.

Artificial Intelligence (AI) and Machine Learning (ML) have significant impact on the world and human thinking, actions, and decisions. AI can perform functions associated with human intelligence, such as image processing and speech recognition. Library services include circulation, cataloguing, and reference services. AI can perform tasks like discovering information and generating content from past experiences but may provide irrelevant information in specific domains (Isiaka, 2023); as in current era, Large Language Models are powerful tools for collections discovery, search, and analysis. Library collection management uses techniques like recommender systems, metadata generation, and resource discovery. RAG-based GenAI Chatbots can enhance library reference services with AI and ML (Das & Islam, 2024).

Chatbot-prompted Large Language Model (C-LLM): massive textual corpora that can return well-formed textual responses following a human prompt in chatbot dialogue, which highlights the different drawbacks of LLM and how to overcome them using personalized GPTs (Anastasia et al., 2023). Integrating machine learning and AI techniques in research presents libraries with the potential to enhance their services (Susan & Borui, 2024). For example, Generative AI, like ChatGPT, has potential benefits but should be cautious about its ethics. It can make biased statements, “hallucinate” inaccurate information, and risk privacy. AI cannot reason and possess advanced human qualities despite its innovative potential. RAG-based GenAI applications can help overcome these issues by providing accurate information (IFLA AI SIG, n.d.).

### **3. Retrieval-Augmented Generation (RAG)**

Pre-trained data is trained data on a specific date which is a kind of school & college level knowledge that students gain all sorts of general knowledge while upbringing in their life (like Hindi, Gujarati, Math, Science, Social Science, Sanskrit, English, Science, Commerce, Arts, Medical and so on). After pre-trained, fine-tuning of trained data with some specific purpose is being performed. Even after, the users are not getting real-time and relevant information based on their trained and fine-tuned knowledge-based data, RAG play a vital role to help users for searching their specific knowledge-based data. RAG helps LLMs search from specific knowledge-based data rather than general information, making answers more accurate and authentic and meeting users’ expectations. Generating precise and authentic answers is crucial for higher user satisfaction, so RAG will play an essential role in achieving these tasks.

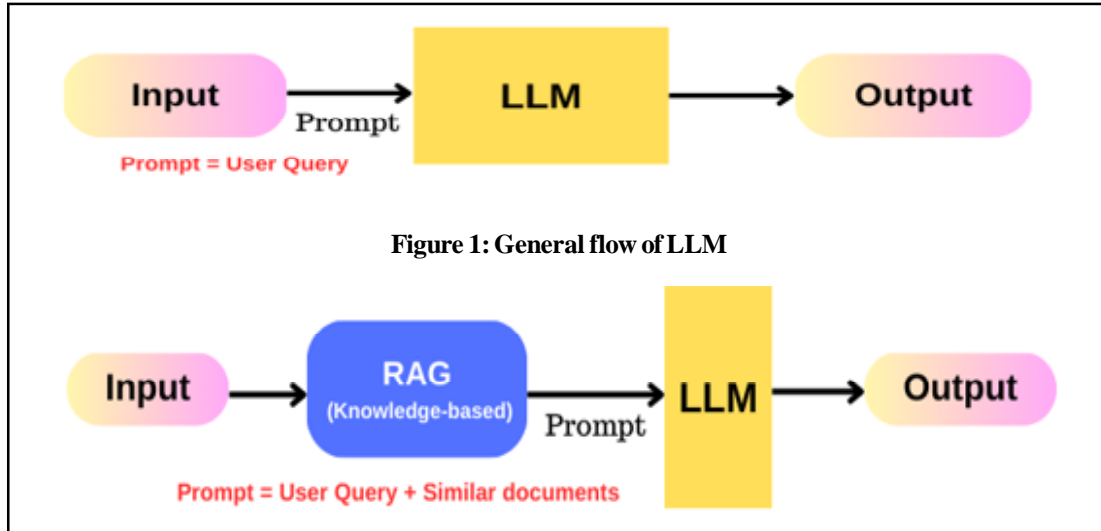


Figure 2: Flow with RAG-based Application of LLM

RAG was introduced by a team led by Lewis in 2020 (Alan, 2023), which is a giant leap in generating content where computer programs learn new things from the given knowledge base before the system respond to user queries.

#### 4. Architecture of framework

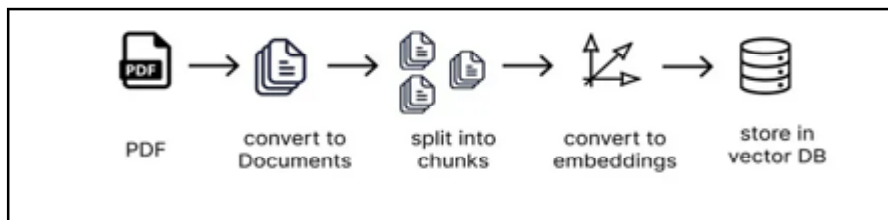
The different models used in framework which shows that how the whole process works, such as uploading, dividing in chunking, embedding, and storing in a vector database, as well as other methods. For developing the RAG-based GenAI App, the following steps are involved.

- 4.1 **Retrieval:** This search step is the key to gathering current and relevant information using a similar search (based on the vector numeric nearby search) in the vector database which compare the user query in the existing knowledge-based data.
- 4.2 **Augmented:** After retrieving relevant information from the vector database and based on the prompt template, LLM understand the query in a much broader and more profound way to generate a good response. In any GenAI application, prompt templates play a crucial role in generating appropriate responses.
- 4.3 **Generation:** Finally, the LLMs use both old and newly acquired the information to answer the question. RAG allows LLM to first look at custom knowledge-based available in the vector database to find the most relevant information. LLM generates a more accurate, detailed response based on efficient prompts. (Demiao, 2024)

## 5. IIM Application (What and How can we do?)

In this use case, Knowledge based database has been created based on the information available on library website including library FAQs, services, and some database manual PDFs.

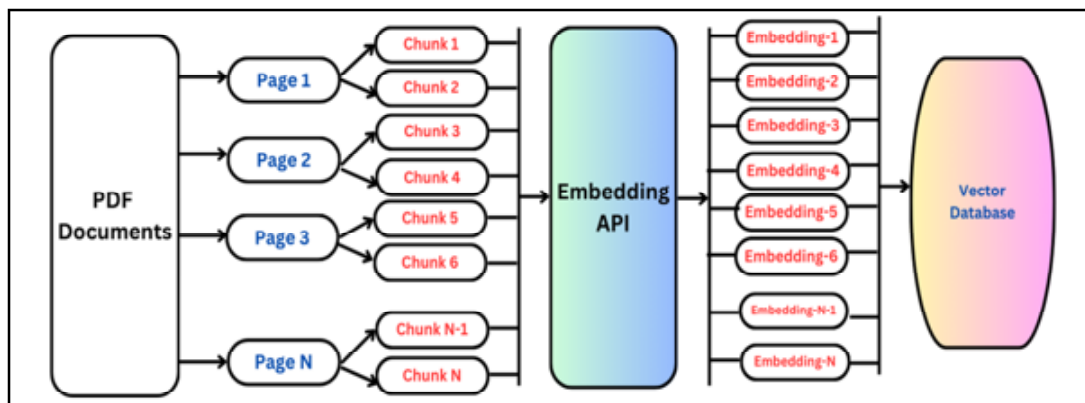
First, a vector database is created by uploading pdf documents as per workflow mentioned in Figure-3. (Maddukuri, 2024):



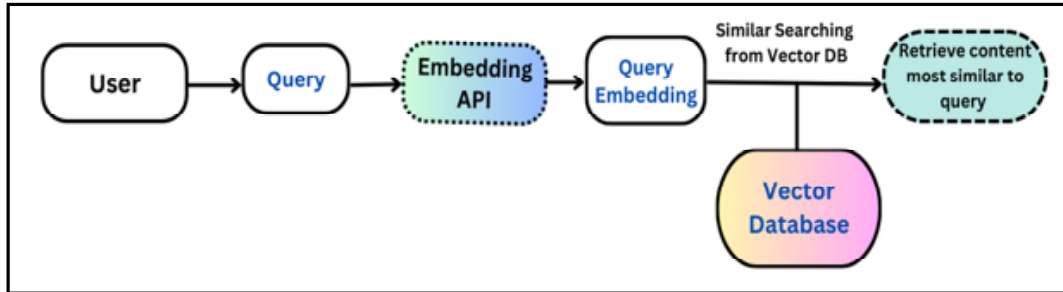
**Figure 3: Ingestion process to generate vector database from PDF**

(source: <https://medium.com/@venky.jishu2021/document-chatbot-using-llama3-2fc525cfc05e>)

During the aforementioned workflow of Ingestion process, the uploaded PDFs are converted into documents (pages); which are further converted into chunks based on the chunk size and chunk overlapping configuration (Naik, 2024). By using Google vector embedding model, the chunks are converted into embeddings (numerical representation). Finally, the chunks are stored in vector DB as depicted in Figure-4. Google model is used for embedding and the Google Gemini pro-LLM model is used for generating content by using Google API Key from Google AI Studio (Google AI Studio, 2023).



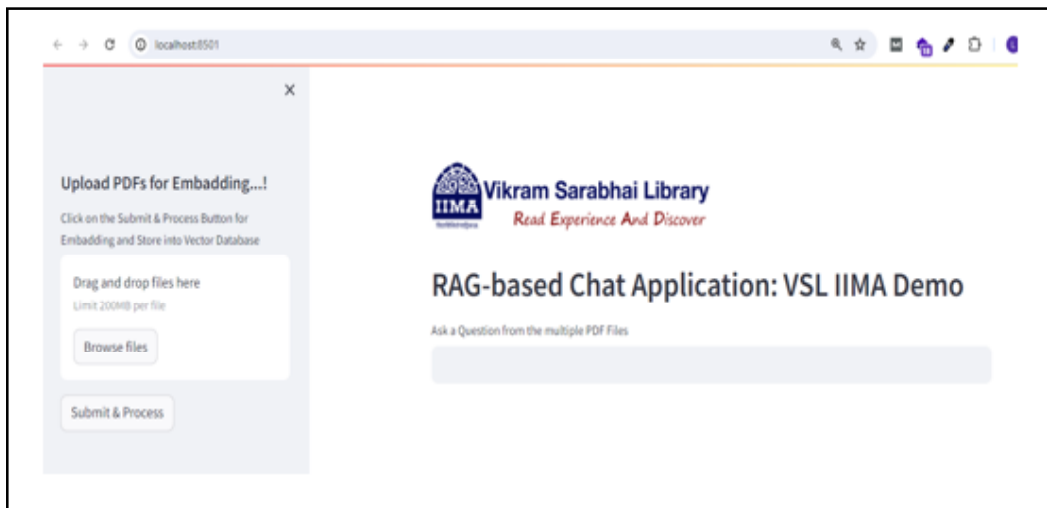
**Figure-4: Process of the Ingestion**



**Figure5: Process of searching the most similar content from the vector database based on the User's Query**

Through the User interface using Streamlit, the user enter the query into the application; the query is first converted into embedding which is a numerical representation of data for LLMs using the Google embedding model and then searches into vector database and retrieve the most similar content and then the actual query and similar content using the prompt template goes as an input into LLM model to generate the most relevant content (Naik, 2024).

In this application, one or more PDFs have been imported which covers library website content including the Introduction, services, FAQs, staff and library-relevant and database-specific information. (Streamlit, 2023) as depicted in Figure-7. It would be generated a vector database for searching (LlamaIndex, 2024) in background. During the procedure, multiple intermediate steps would be performed as Loading the PDF >> Splitting into Pages >> Pages are converted to Chunks >> Chunks are embedding - based on the embedding model (in numerical form), which would be stored in the vector database for searching while chatting (LangChain, 2023). Finally, knowledge base data is prepared to interact with those PDFs.



**Figure 6: Home page of an application**



Figure 7: Uploading PDF for the embedding process

Below are some outputs (for comparison purposes) based on the PDF and what we get the result through our RAG-based application.

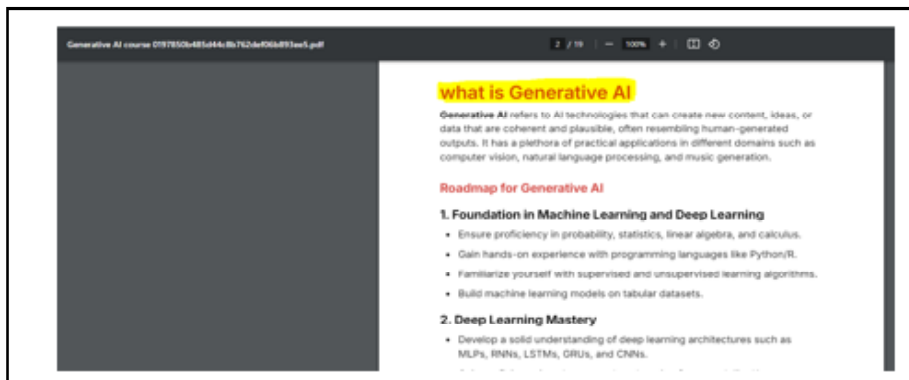


Figure-8: GenAI PDF 2<sup>nd</sup> page

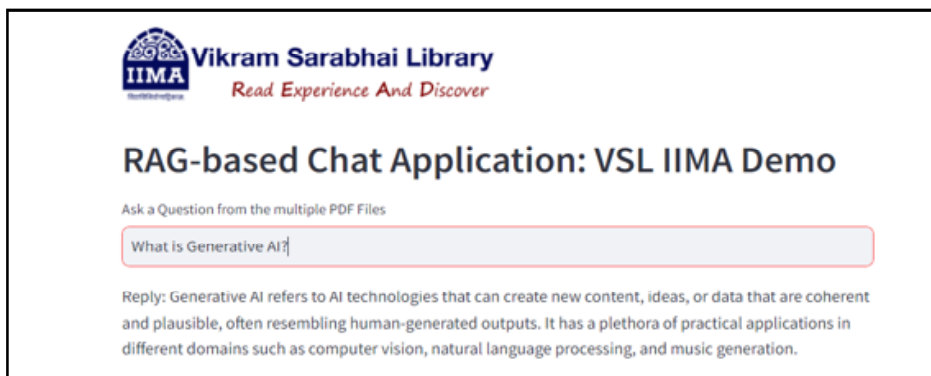


Figure-9: GenAI Chat Interface for PDF 2<sup>nd</sup> page question



Figure-10: Annual report PDF 7<sup>th</sup> page

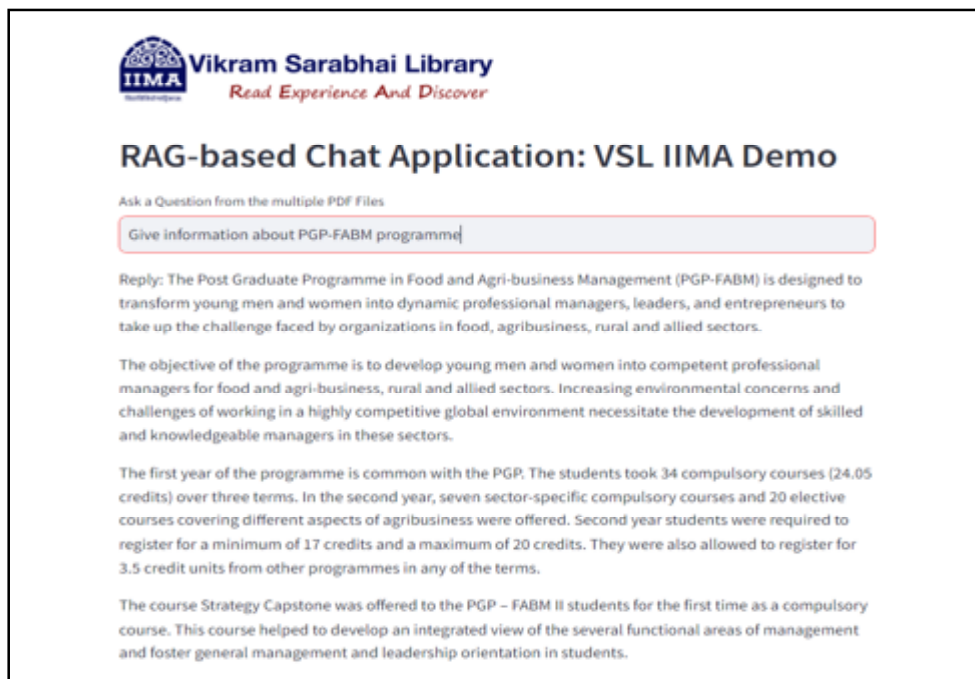


Figure-11: Annual report PDF 7<sup>th</sup> page – chat response




**LOANCONNECTOR QUICK REFERENCE GUIDE**

**LOANCONNECTOR**  
LoanConnector is Thomson Reuters LPC's one-stop solution for the global primary and secondary loan markets. It provides instant market intelligence and access to detailed news, data and analysis, including late-breaking news, forward deal calendars, pricing grids, historical data, industry analysis and much more, empowering you with the information you need to be successful in today's market.


Utilize this Quick Reference Card to drill down and understand the most common tasks performed using LoanConnector.

**SUPPORT**  
**AMERICAS:** (T) 646-223-6890 (select opt. 2, then opt. 1) | [lpc.americas@thomsonreuters.com](mailto:lpc.americas@thomsonreuters.com)  
**ASIA-PACIFIC:** Asia-Pacific — (852) 2294-8853 | [lpc.asiapacific@thomsonreuters.com](mailto:lpc.asiapacific@thomsonreuters.com)  
Japanese language support — (81) 320-161-916  
**EMEA:** (44) 800-44-2000 (select opt. 4, then opt. 3) | [lpc.europe@thomsonreuters.com](mailto:lpc.europe@thomsonreuters.com)

**GETTING STARTED**  
Go to the LoanConnector platform: <http://www.loanconnector.com>. Enter your e-mail address in the username field and your LoanConnector password in the password field.



Figur-12: Thomson Reuters reference guide PDF

 **Vikram Sarabhai Library**  
*Read Experience And Discover*

**RAG-based Chat Application: VSL IIMA Demo**

Ask a Question from the multiple PDF Files

What is LOANCONNECTOR?

Reply: LoanConnector is Thomson Reuters LPC's one-stop solution for the global primary and secondary loan markets. It provides instant market intelligence and access to detailed news, data and analysis, including late-breaking news, forward deal calendars, pricing grids, historical data, industry analysis and much more, empowering you with the information you need to be successful in today's market.

Figure-13: Thomson Reuters reference guide chat

**ACE Knowledge portal is...**

ACE Knowledge portal is online database, provides data on **companies, industries, mutual funds, macro economy and live markets**. It also provides **historical information on company, industry and economy** which can be use for project and analysis purpose.

**Feature:** Covered major Indian companies, industries

Figure-14: ACE Knowledge Portal PDF

 **Vikram Sarabhai Library**  
*Read Experience And Discover*

**RAG-based Chat Application: VSL IIMA Demo**

Ask a Question from the multiple PDF Files

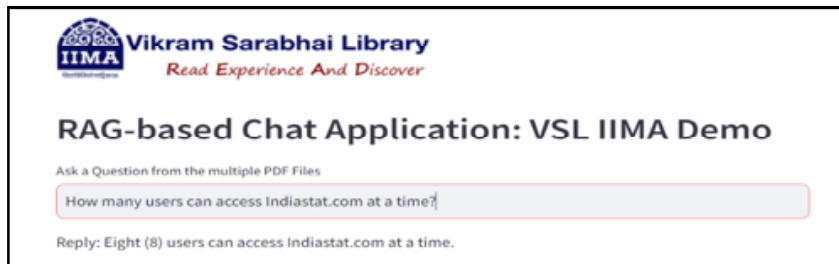
Give me information about ACE Knowledge portal

Reply: ACE Knowledge portal is online database, provides data on companies, industries, mutual funds, macro economy and live markets. It also provides historical information on company, industry and economy which can be use for project and analysis purpose.

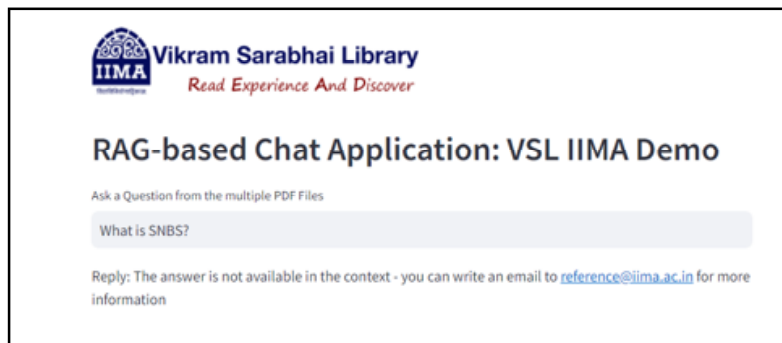
Figure-15: ACE Knowledge Portal Chat



**Figure-16: FAQ list from Library website PDF**



**Figure-17: FAQ list from library website Chat**



**Figure-18: No answer, if not found on the knowledge-based**

## 6. Advantages of RAG-based App

Using a RAG-based chat system has various advantages and limitations (Sebastien, 2024).

- ❖ User must get the data from our custom knowledge base to get only authentic information.
- ❖ It would be resolved the issue of hallucinations (false/irrelevant information) and misleading responses, which is a big issue in LLMs because they are trained on huge general data.
- ❖ RAG-based systems can run locally, so administrator can ensure to improve privacy and security-related issues, which are major concerns for some sensitive organizations.

## 7. Limitations & Solutions

The following solutions would be helping the administrators to address the issues since every system has its limitations and some solutions to perform better. (Mukherjee, 2024).

- ❖ Since, the whole system is executed locally, the high-end server having GPUs is must to run this RAG-based application and provide the responses on time, or an application can be deployed on cloud-based services like Amazon Bedrock, which is the easiest way to build and scale generative AI applications with foundation models.
- ❖ Missing data means the system cannot provide an answer because the required information is unavailable in the knowledge base. User may not get any response from system in case of data unavailability. This issue can be resolved by increasing knowledge base data by providing more and more documents related to the subject area. In addition, some external resources would be imported, such as adding more PDFs to enhance the knowledge base. Similarly, Wikipedia pages and specific web pages would be imported to supplement the information missing from the primary resource base.
- ❖ Promptness of system is very important factor to get an accurate response from LLMs; since, providing a prompt and accurate response that eliminates redundancy and removes the repetitive information and phrases.

## 8. Conclusion and Future Enhancement

The application of AI-based technology in information products has been discussed. However, AI-based solutions developed by the library are more recent and are still in the experimental stage. More tools and technology that are easier to install, combined with results above expectations, are driving these hi-tech inclusions. RAG-based chat applications using LLMs are one such application that would assist users with ready reference service. User can ask questions to knowledge base system, but the technology and system would facilitate enhanced user experience with more interactive and accurate information. As in this use case, knowledge base system has been developed by importing PDFs only. More additional features would be implemented, such as searching from a specific website by providing sitemap.xml and from a direct database like SQL, and so on, to enhance the chat interface.

### Glossary

**RAG:** Retrieval-Augmented Generation, which combines LLMs with external knowledge bases to improve their outputs.

**LLMs:** Large Language Models are deep learning algorithms that use large amounts of data to perform natural language processing tasks.

**VSL:** Vikram Sarabhai Library

**NLP:** Natural Language Processing, is a machine learning technology that gives computers the ability to interpret, manipulate, and comprehend human language.

**Google Gemini:** LLM model from Google from Embedding.

**ChatGPT:** ChatGPT is a chatbot and virtual assistant developed by OpenAI.

**LLama:** Large Language Model developed by Meta AI.

**Python:** Python is a high-level, general-purpose programming language; we used this language for frontend to generate UI.

**LangChain:** LangChain is a framework built with LLMs by chaining interoperable components for different use cases like Chatbot, summarization, etc.

**Vector embedding:** It's a way to convert words, sentences, and other data into numbers that capture their meaning and relationships, which is used for similar searches.

**Prompt:** Prompt is an input the user gives to LLM to generate content.

**Streamlit:** Frontend Python library to use to design webpages in Python.

**Knowledge-based data:** Our own custom data set for specific applications.

**FAISS:** It's a Facebook AI Similarity Search library that allows developers to quickly search for vector databases that are most similar to each other.

**Google API Key:** Required to connect Google Gemini Model for Embedding and Generating content

**Google AI Studio:** An online interface from Google to manage API keys

**UI:** User Interface

**Document loader:** Loading the PDF from a specific location

**Text Splitter:** Splitting text from the PDF pages.

**Chunk size:** it's the number of sentences that are divided into chunks.

**Chunk overlapping:** It's a number of words that overlap from the front and back sentence

**A retrieval chain:** means retrieving past data and adding it to the prompt for better similarity with past questions.

**Prompt Template:** A specific farmwork is used to pass our data to LLM as input to generate a better response.

## References

1. Alan, Zeichick (2023). What Is Retrieval-Augmented Generation (RAG)?. <https://www.oracle.com/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/>
2. Anastasia Olga (Olnancy) Tzirides, Akash Saini, Gabriela Zapata, Duane Searsmith, Bill Cope,
3. Mary Kalantzis, Vania Castro, Theodora Kourkoulou, John Jones, Rodrigo Abrantes da Silva,
4. Jen Whiting, Nikoleta Polyxeni Kastania. (May 2023). Generative AI: Implications and Applications for Education. [https://www.researchgate.net/publication/370764055\\_Generative\\_AI\\_Implications\\_and\\_Applications\\_for\\_Education](https://www.researchgate.net/publication/370764055_Generative_AI_Implications_and_Applications_for_Education)
5. Canva (2024, January 25). Canva is a graphic design platform that provides tools for creating. <https://www.canva.com>
6. Das, Rajesh Kumar and Islam, Mohammad Sharif Ul. (2024, February 10). Application of Artificial Intelligence and Machine Learning in Libraries: A Systematic Review.
7. <https://arxiv.org/pdf/2112.04573>
8. Demiao, LIN. (2024, February 2). Revolutionizing Retrieval-Augmented Generation
9. with Enhanced PDF Structure Recognition. <https://arxiv.org/pdf/2401.12599>
10. Google AI Studio: (2024, February 10). To generate Google API Key to access Embedding model and Google Gemini Pro model. <https://ai.google.dev/aistudio>
11. IFLAAI SIG. (2024, March 03). Generative AI for library and information professionals Produced by the IFLAAI SIG. <https://www.ifla.org/g/ai/generative-ai/>
12. Isiaka, Abdullahi Olayinka, (2023). Application and Use of Artificial Intelligence (AI) for Library Services Delivery in Academic Libraries in Kwara State, Nigeria. <https://digitalcommons.unl.edu/libphilprac/7998>
13. LangChain (2024, March 25). LangChain's flexible abstractions and AI-first toolkit make it the #1 choice for developers when building with GenAI
14. <https://www.langchain.com>
15. LlamaIndex Document (2024, March 10). High-Level Concepts (RAG). [https://docs.llamaindex.ai/en/stable/getting\\_started/concepts/](https://docs.llamaindex.ai/en/stable/getting_started/concepts/)
16. Maddukuri, Venkatesh. (2024, May 9). Building Document Chatbot Using Langchain
17. <https://medium.com/@venky.jishu2021/document-chatbot-using-llama3-2fc525cfc05e>

18. Mukherjee, Subrata (2024, February 11). Retrieval-Augmented Generation (RAG) and explore how it can address the pain points we have in the context of LLM-based applications.
19. <https://subrata-mettle.medium.com/retrieval-augmented-generation-rag-and-explore-how-it-can-address-the-pain-points-we-have-in-the-1edbb8b0e962>
20. Naik, K. (2024, February 20). <https://github.com/krishnaik06/Build-Gen-AI-With-Google-Gemini/tree/main/Chat%20With%20multiple%20Pdf%20Documents%20with%20Langchain%20and%20Google%20Gemini%20Pro>
21. Sebastien, Sime (2024, February 13). RAG with LLM. <https://sebastien-sime.medium.com/rag-with-llm-2fb735dc025b>
22. Senthil, E. (2024, April 5). Unlocking LLM's Potential with RAG: A Complete Guide from Basics to Advanced Techniques Using OpenAI, Google Gemini Pro, and Open-Source Models. <https://levelup.gitconnected.com/unlocking-llms-potential-with-rag-a-complete-guide-from-basics-to-advanced-techniques-b4557f268134>
23. Susan Jenkins, Borui Zhang. (2024, February 07). The role of AI in library services. <https://www.elsevier.com/en-in/connect/the-role-of-ai-in-library-services>
24. Streamlit (2024, January 21). Streamlit is used to design the UI for your app. <https://docs.streamlit.io/develop/api-reference>

#### **About Authors**

**Bhavesh Patel**, Senior Library Professional - IT Applications, Vikram Sarabhai Library, IIMA, Ahmedbad, Gujarat

Email: bhaveshiima@gmail.com

**Dr. B B Chand**, Librarian & Head NICMAN, Vikram Sarabhai Library, IIMA, Ahmedbad, Gujarat

Email: bbchand@iima.ac.in