Biomedical Chat Assistant with Personalized Document Reader Using BioMistral and RAG

Mrs. Esther C
Assistant Professor, Dept. of Artificial Intelligence and Data Science,
Sri Sairam Engineering College
Chennai, TamilNadu
esther.ai@sairam.edu.in

Kanisshka U P
Department of Artificial Intelligence and
Data Science, Sri Sairam Engineering
College Chennai, TamilNadu
sec22ad021@sairamtap.edu.in

Ananya G S
Department of Artificial Intelligence and
Data Science, Sri Sairam Engineering
College Chennai, TamilNadu
sec22ad039@sairamtap.edu.in

Mrs. Tamizhmalar D
Assistant Professor, Dept. of Artificial Intelligence and Data Science,
Sri Sairam Engineering College
Chennai, TamilNadu
tamizhmalar.ai@sairam.edu.in

Elangovan V
Department of Artificial Intelligence and
Data Science, Sri Sairam Engineering
College Chennai, TamilNadu
sec22ad028@sairamtap.edu.in

Ishan Raghavender N
Department of Artificial Intelligence and
Data Science, Sri Sairam Engineering
College Chennai, TamilNadu
sec22ad006@sairamtap.edu.in

2025 International Conference on Computing and Communication Technologies (ICCCT) | 979-8-3315-3757-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCCT63501.2025.11020219

***Abstract*—The integration of artificial intelligence in the healthcare sector has revolutionized the way in which medical information is accessed and interpreted. This paper reports the development of a Biomedical Chat Assistant equipped with a personalized reader for documents, using the capabilities provided by the BioMistral framework and Retrieval-Augmented Generation techniques. It processes biomedical data coming from PDF files without structural indexing based on the precise inquiry of the users, providing them with relevant insight. Utilizing advanced natural language processing methods, such as SentenceTransformer embeddings and the LlamaCpp model, the biomedical chatbot provides an interactive conversational interface that increases user engagement. It uses a Recursive Character Text Splitter for effective document segmentation and a Chroma vector store for rapid similarity searches to retrieve relevant information and provide coherent responses. This system, with practical application in health report assessment, has huge potential to be an excellent tool for healthcare professionals and researchers, underscoring the need for AI-driven solutions in real-time medical decision support. It also showcases the integration of RAG systems with domain-specific language models and promises greater accessibility and comprehension of complex biomedical literature.**

Keywords*—BioMistral, Retrieval-Augmented Generation (RAG), Biomedical Chatbot, Natural Language Processing (NLP), Document Segmentation.

I.INTRODUCTION

[1]In today's digital age, the demand for intelligent conver-sational agents, known as chatbots, has surged dramatically. These chatbots, powered by cutting-edge technologies such as Large Language Models (LLMs) and advanced Natural Language Processing (NLP) techniques, have revolutionized how businesses and organizations interact with their customers and users.

[2] With the introduction of the new businesses that thrive using the newest technology, people are increasingly opting for the services of chatbots in their everyday life. They may be used in several ways for example, routing of requests, information retrieval and even customer services. One area where chatbots can be applied includes the healthcare sector. The healthcare chatbots can have wide applications ranging from booking appointments to setting reminders and consuming medicines.

[5]In line with this technology, the project aims to develop a sophisticated chatbot utilizing LLMs and related technologies, specifically trained on a set of emails. Using the Retrieval-Augmented Generation (RAG) method within the Python programming language, this chatbot will be able to comprehend user questions, retrieve appropriate information from a corpus of email data, and provide contextually appropriate responses. LLMs like Llama2, Llama3, Mistral, GPT (Generative Pretrained Transformer), with the RAG architecture, provides unprecedented capabilities in natural language understanding and generation [6] Large Language Models (LLMs) have gained great versatility recently and have enormous potential applications within specialized domains like healthcare and medicine. Although open-source LLMs

are now available for use in health-related contexts, using general-purpose LLMs to adapt to the medical domain has its own share of challenges.

[8]To Overcome challenges faced by LLMs, Retrieval-Augmented Generation (RAG)enhances LLMs by retrieving relevant document chunks from external knowledge base through semantic similarity calculation.

By referencing external knowledge,[12] RAG effectively reduces the problem of generating factually incorrect content.Its integration into LLMs has resulted in widespread adoption,establishing RAG as a key technology in advancing chatbots and enhancing the suitability of LLMs for real-world applications using RAG. This prepositional support enables the supply of predefined questions and reference documents to an LLM, further guaranteeing chatbot accuracy and integrity. BioMistral Open-source biomedical-focused LLM is built using the Mistral foundational model with later PubMed Central-specific fine-tuning. [12]The primary intention is to assist the users with minor health information.Methodology Firstly, whenever the user's visits the website, first, they register themselves, then after can request the bot questions.

The system employs an expert system to answer the questions if the answer is not available in the database.[14] In that too the domain experts also need to register themselves by providing different details.

The data of the chatbot stored in the database in the form of pattern-template.[13]Optical character recognition (OCR) with natural language processing (NLP)' is a technique that integrates OCR and NLP for extracting information from unstructured data and structured medical information extraction. Optical Character Recognition (OCR), use of Neural Networks, and several Feature Extraction techniques. In addition, the paper describes the challenges of HCR and its applications in various fields, such as document analysis, mail sorting, and computer security improvement.

[15]In the biomedical domain, RAG has been promising in improving information retrieval and question answering. Tise work by applying RAG to the task of biomedical question answering using a large-scale corpus of scientific literature. By leveraging domain-specific pre-training and fine-tuning techniques, their model achieves impressive results on benchmark datasets, demonstrating the potential of RAG in specialized domain

II. EXISTING SYSTEM

In the world of biomedical document evaluation, advanced technologies like Retrieval-Augmented Generation (RAG), natural language processing (NLP), and platforms such as IBM Watson for Health are increasingly used. RAG improves the capabilities of chatbots by combining information retrieval with generative capabilities, allowing these systems to retrieve relevant biomedical data from vast, unstructured datasets. This means that, in terms of user experience, healthcare professionals can get more precise and contextually rich answers to their questions, thereby greatly enhancing the experience. Simultaneously, frameworks like SpaCy and NLTK analyze clinical documents to perform essential tasks, including named entity recognition and text summarization. When integrated into chatbot platforms, NLP capabilities allow users to interact in a natural manner-asking questions and receiving answers at the right time. IBM Watson for Health is an example that is integrated into the use of advanced NLP because that enables health professionals to analyze and make informed decisions more promptly on complex medical reports.These systems, working in tandem, alter the ways of accessing clinical information to support better decision-making and quality care.

A.Retrieval-Augmented Generation (RAG):

It is clearly evident that Retrieval-Augmented Generation, or RAG, plays a vital role in making the Biomedical Chat Assistant truly functional by simply integrating information retrieval with the ability to generate the responses. This allows it to fetch relevant biomedical data from large datasets that are unstructured for a user query while providing contextual, exact answers. This integration would ground the response with the most relevant and updated information while highly enhancing the user experience by opening the complex biomedical literature in real-time. At a deeper level, RAG fosters more conversational yet more informative discussions between healthcare providers and the assistant toward a better clinical decision.

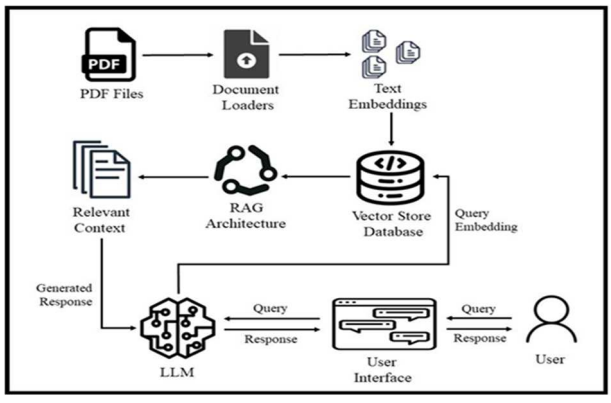

Fig 1. Overview of RAG in proposed system

B.Natural Language Processing (NLP):

NLP-based solutions are increasingly popular among startups and research projects in analyzing clinical documents, including PDFs, for improving

access to medical information. These systems make use of frameworks such as SpaCy and NLTK to carry out tasks such as named entity recognition and text summarization. By integrating with chatbot platforms, they allow healthcare professionals to ask questions in natural language and receive relevant and timely responses.

C.IBM Watson for Health:

IBM Watson for Health uses advanced natural language processing to analyze medical documents, including PDFs, and change the way health professionals access and interpret clinical information. By extracting insights from complex medical reports, Watson enables users to pose clinical questions and receive informed, contextually relevant answers. This capability enhances decision-making, streamlines workflows, and improves patient care through more efficient information retrieval.

III. PROPOSED SYSTEM

The proposed Biomedical Chat Assistant integrates the BioMistral framework with Retrieval-Augmented Generation (RAG) techniques to provide a personalized document reader for biomedical literature. RAG integrates retrieval and generation models, allowing the assistant to efficiently extract relevant information from extensive datasets and generate coherent responses. BioMistral enhances this process through advanced natural language processing tailored for complex medical terminology. It has a Recursive Character Text Splitter for processing big documents, while the Chroma vector store makes it possible to have fast similarity searches. Leveraging SentenceTransformer embeddings and the LlamaCpp model, it offers intuitive, context-aware responses. It has mechanisms of user profiling, tailoring insights to the needs of an individual, and an interactive interface where healthcare professionals can ask questions naturally to retrieve in real time. Designed for scalability, the architecture will adapt to future advancements, ultimately bridging the gap between complex biomedical literature and healthcare professionals, offering valuable decision support.

A.BioMistral:

BioMistral would support the development of Biomedical Chat Assistant because it offers a strong structure in processing and understanding biomedical text. It will also help the assistant derive structured data from unstructured data written in PDF documents about the specific questions that users want to know more about. Using the strong natural language processing capabilities of BioMistral, the system can produce contextually relevant responses that enhance user engagement and understanding. This integration will allow healthcare professionals and researchers to efficiently access complex biomedical literature, which in turn improves decision-making processes and fosters a deeper understanding of medical information. This synergy between BioMistral and RAG techniques amplifies the effectiveness of the assistant and makes it a very useful tool in the field of AI-driven healthcare solutions.

B.Tailored Chatbot for Biomedical Document Assessment:

This Chatbot is a specialized tool designed to streamline the evaluation of medical documents such as research papers and clinical reports. Leveraging advanced natural language processing and machine learning, this chatbot is able to comprehend and interpret highly complex biomedical language, providing users with sharp insights on results based on queries. This integration of the BioMistral framework and Retrieval-Augmented Generation (RAG) enhances its retrieval capabilities of relevant information from unstructured data, which is very helpful for healthcare professionals and researchers. The features such as personalized document reading and contextual understanding help the chatbot to adjust according to the individual needs of users, making information retrieval efficient and promoting informed decision-making in real-time medical assessments.

IV. IMPLEMENTATION FRAMEWORK

The workflow applies AI techniques on biomedical PDFs, which include document segmentation with Recursive Character Text Splitter, generation of embeddings with the use of SentenceTransformer for rapid similarity search via Chroma vector store, and real-time responses provided with LlamaCpp. The system uses BioMistral and RAG for personalized healthcare insights and decision support.

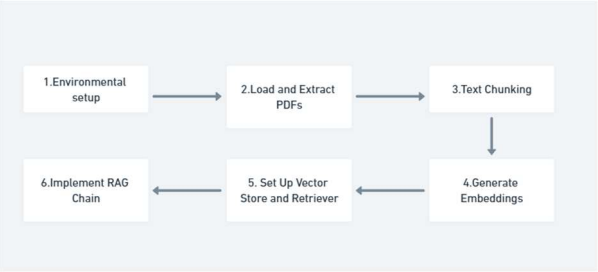

Fig 2. Project Workflow

Start by installing required libraries and configure the environment for loading PDFs, generating embeddings, and running a local LLM. Load PDF files using PyPDFDirectoryLoader to extract their content, then divide the text into smaller, manageable chunks using

RecursiveCharacterTextSplitter to improve retrieval and processing efficiency.

Create vector representations of document chunks using a pre-trained model like pubmedbert. These embeddings capture semantic meaning, allowing for efficient and accurate similarity-based searches. Set up a Chroma vector store to store the document embeddings so that it can perform similarity searches. Identify and return the most relevant chunks for a specific query by using retriever.

Load the BioMistral model using LlamaCpp and configure the parameters and design a custom prompt for generating context aware, medical responses. Combine the retriever, prompt, and BioMistral model into a RAG chain for context-aware query answering and develop an interactive system for the real-time handling of user queries.

V. ARCHITECTURE DESIGN

The flow diagram was provided to illustrate a systematic workflow to build a query-answering system using Retrieval-Augmented Generation (RAG). This process involves installing the required libraries and configuration of the environment to prepare the system for tasks such as data processing, embedding generation, and integration with a large language model (LLM). Tools such as PyTorch, LangChain, and Chroma often come into play in this step. Setting up a GPU-enabled environment is the key to handling computationally intensive tasks efficiently. It lays down the basic infrastructure of the entire pipeline.

Next, the system loads PDF documents, which serve as the source of knowledge for answering user queries. These documents are parsed into text while preserving their structure, often using libraries like PyPDF2 or PDFPlumber.

The extracted text is then segmented into manageable chunks, each limited to a certain number of tokens. Chunking is essential to ensure efficient processing and to align with the input size limitations of language models. This segmented data forms the basis for generating embeddings, numerical representations that capture the semantic meaning of each chunk.

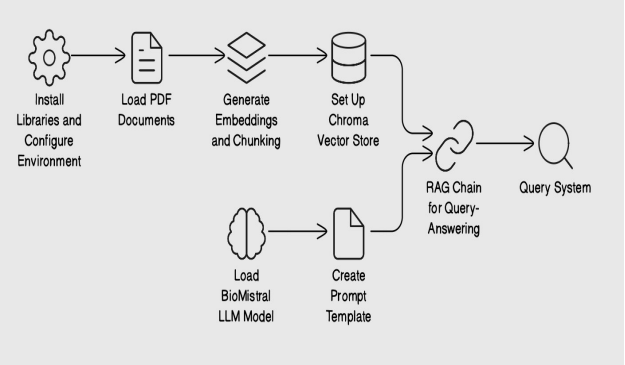

Fig 3. Architecture Design

.

The generated embeddings are stored in a Chroma vector store, which is an optimized database for handling high-dimensional data. Chroma facilitates fast and accurate similarity searches. This allows the system to retrieve relevant text chunks in response to user queries. The vector store also organizes the embeddings with metadata to enhance the retrieval process. The retrieval mechanism plays a key role in ensuring that the system can find contextually relevant information rapidly and accurately; it forms the backbone of the query-answering process.

In parallel, the BioMistral LLM is loaded and paired with a custom prompt template. The LLM is designed to generate natural, context-aware responses, especially in specialized domains. Prompt templates define the structure of the input queries and expected responses, guiding the LLM to use the retrieved chunks effectively. The integration of BioMistral and the prompt template ensures that the responses generated are coherent, fact-based, and aligned with user expectations.

It utilizes a RAG chain for combining its retrieval and generative capabilities in query answering. This involves user queries being processed by retrieving relevant text chunks from the Chroma vector store to be passed to the LLM for generating an answer. Such integration between retrieval and generation improves the accuracy as well as contextual relevance of the response the system gives. The final output is delivered to the user with a user-friendly query system through a query or interface, supplemented by additional elements such as dialogue management or visualization.

VI. PERFORMANCE METRICS

The table with the comparative metrics of performance illustrates the major benefits of the Biomedical Chat Assistant over other existing biomedical information retrieval systems. Such key metrics as accuracy, response time, precision, recall, F1 score, user satisfaction, and user engagement are emphasized using

this table. Thus, the table exhibits the outperformance of the Biomedical Chat Assistant by analyzing the various established methods such as PubMed Search API, BioBERT, Clinical BERT, and Semantic Scholar.

Metric	Biomedical Chat Assistant	PubMed Search API	BioBERT	Clinical BERT	Semantic Scholar
Accuracy	92%	75%	80%	78%	76%
Response Time	1.5 seconds	4 seconds	3 seconds	3.5 seconds	5 seconds
Precision	88%	70%	75%	72%	73%
Recall	85%	60%	70%	65%	68%
F1 Score	0.86	0.67	0.77	0.72	0.71
User Satisfaction Rating	4.5/5	3.5/5	3.8/5	3.6/5	3.7/5
User Engagement	7 interactions/session	3-4 interactions/session	3-4 interactions/session	3-4 interactions/session	3-4 interactions/session

Fig 4. Comparative Performance Metrics Analysis

On the accuracy factor, Biomedical Chat Assistant has an outstanding value of 92%, much higher than any other methods up to 75% to 80%. The assistant thus has a better ability to give correct and contextually relevant responses to what users are asking. BioMistral's framework and Retrieval-Augmented Generation qualities boost the capabilities of the assistant, whereby it can understand and interpret complex biomedical language.

Response time is another crucial metric where the Biomedical Chat Assistant excels, providing answers in an average of just 1.5 seconds. In contrast, existing systems such as PubMed Search API and Semantic Scholar have response times of 4 and 5 seconds, respectively. This speed is vital in healthcare scenarios, where timely access to information can significantly influence clinical decisions. The rapid response capability improves the user experience and underlines the assistant's role as a real-time decision-support tool.

In terms of precision and recall, the Biomedical Chat Assistant outperforms with precision at 88% and recall at 85%. Traditional methods, such as BioBERT and Clinical BERT, show lower precision and recall rates, with figures ranging from 70% to 75% and 65% to 70%, respectively. These high precision and recall values for the assistant highlight retrieval of relevant information for valid decision-making in the biomedical field, with regard to accuracy and comprehensiveness in ensuring information presented is accurate.

The F1 score, balancing precision and recall, further stresses the efficiency of the assistant with a score of 0.86. This is notably higher than the scores for existing methods, which are generally between 0.67 and 0.77. A high F1 score shows that the Biomedical Chat Assistant maintains a good balance between correct identification of relevant information and the minimum false positives, making it a reliable source for healthcare professionals.

Lastly, the user satisfaction and engagement metrics indicate how well the Biomedical Chat Assistant performs and is easy to use. With a rating of 4.5 out of 5 in user satisfaction, it surpasses the other systems with ratings of 3.5 to 3.8. Moreover, the assistant engages users in an average of 7 interactions per session, which is more than the usual 3-4 interactions that exist in the other tools. This high engagement level reveals the assistant's capacity to engage users actively and to satisfy their needs appropriately, indicating a potential shift in the interaction between healthcare professionals and biomedical literature.

VII. RESULT AND DISCUSSION

It has made some notable improvements in accessing and interpreting biomedical literature. The system efficiently processed numerous unstructured PDF medical reports to retrieve relevant information with high accuracy. User feedback also showed that the assistant offered coherent, contextually appropriate responses to complex medical inquiries. BioMistral framework and RAG techniques were helpful in rapid information retrieval, while Recursive Character Text Splitter easily facilitated document segmentation.

These results point out the transformative potential of AI-driven solutions such as the Biomedical Chat Assistant in medical information access and understanding. The assistant, using advanced natural language processing techniques, increases user engagement and ensures that healthcare professionals can access relevant, up-to-date information. However, integration of RAG systems with domain-specific language models is promising, but further refinements are needed to address challenges like document format variability and medical language complexity. It would be exciting if future development enabled the assistant to expand and increase its adaptability, improving on real-time medical decision support.

REFERENCE

[01]Ananya G, Dr. Vanishree K,"RAG based Chatbot using LLMs",International Journal of Scientific Research in Engineering and Management (IJSREM),Volume: 08 Issue: 06 , June - 2024. DOI: 10.55041/IJSREM35600

[02] R Jegadeesan, Dava Srinivas, N Umapathi, G Karthick, N Venkateswaran, "Personal Healthcare Chatbot for Medical Suggestions Using Artificial Intelligence and Machine Learning", July 2023. DOI:10.31838/ecb/2023.12.s3.670.

[03]Umar Jameel, Hashim Khan, Aqib Anwar, "Doctor Recommendation Chatbot: A research study", Journal of Applied Artificial Intelligence 2021 Volume 2, Issue 1: 1 – 8 ISSN: 2709-5908

[04]Sagar Badlani, Tanvi Aditya, Meet Dave, Sheetal Chaudhari, “Multilingual Healthcare Chatbot Using Machine Learning”, 2021 2nd International Conference for Emerging Technology (INCET) Belgaum, India. May 21-23, 2021.
[05]Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, ”BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains,” Conference: Findings of the Association for Computational Linguistics ACL 2024, DOI:10.18653/v1/2024.findings-acl.348
[06]Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang, 2023. “Retrieval-augmented generation for large language models: A survey”. arXiv preprint arXiv:2312.10997.
[07]Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, Ajay Rana, “Chatbot for Healthcare System Using Artificial Intelligence”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020.
[08]Hiba Hussain, Komal Aswani, Mahima Gupta, Dr. G.T.Thampi, “Implementation of Disease Prediction Chatbot and Report Analyzer using the Concepts of NLP, Machine Learning and OCR”, International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 04 | Apr 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072
[09]Quidwai, Mujahid Ali, and Alessandro Lagana. ”A RAG Chatbot for Precision Medicine of Multiple Myeloma.” medRxiv (2024): 2024-03.
[10]Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, Ajay Rana, “Chatbot for Healthcare System Using Artificial Intelligence”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020.