Enhancing Healthcare Accessibility: A RAG- Based Medical Chatbot Using Transformer Models

Agrim Kulshreshtha
Computer Science and Engineering
Delhi Technical Campus
Greater Noida, India
agrim.class@gmail.com

Aditya Choudhary
Computer Science and Engineering
Delhi Technical Campus
Greater Noida, India
adityac2542@gmail.com

Tejas Taneja
Computer Science and Engineering
Delhi Technical Campus
Greater Noida, India
tejastaneja1206@gmail.com

Seema Verma
Computer Science and Engineering
Delhi Technical Campus
Greater Noida, India
seemaknl@gmail.com

2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD) | 979-8-3503-5546-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ITIKD63574.2025.11005179

***Abstract*—Significant improvements in patient involvement and medical diagnosis have resulted from the use of AI in healthcare. This paper introduces a transformer-based medical chatbot that uses LangChain and Retrieval-Augmented Generation (RAG) to provide accurate, context-aware healthcare support. The chatbot solves issues with healthcare accessibility, especially in underprivileged areas, and improves user happiness and diagnostic accuracy by utilizing large medical datasets. The aim of the work is to produce dependable and contextually relevant responses by refining refined LLaMa models using RAG approaches. It also looks at how AI-powered chatbots might help with data privacy, timely medical advice, and healthcare disparity reduction. The results highlight how AI-powered virtual health assistants might improve clinical judgment, lessen the workload for medical staff, and offer affordable patient care options.**

Keywords*—*RAG, LangChain, Fine-Tuning, Healthcare, Llama model

I. Introduction

Advances in natural language processing and artificial intelligence have fueled the creation of complex large language models (LLMs), like GPT-4 and Llama 2. In a variety of fields, these models have remarkable skills that frequently come close to matching those of human experts [1]. AI-powered chatbots have revolutionized a number of sectors, including healthcare, by increasing efficiency, reducing costs, and improving patient care. However, challenges like accuracy, data security, and the capacity to comprehend human emotions prevent their broad application in sensitive domains. For these chatbots to be successfully integrated into healthcare settings, several obstacles must be overcome [2]. Numerous obstacles prevent these healthcare chatbots from becoming widely used and having an impact. Problems in the healthcare industry include low skill levels that make them less effective in delicate settings, data security risks, and limited accuracy in complex diagnosis. Similar problems with chatbots include minimal user involvement, self-reporting, and a dearth of suggested nutrition research initiatives. In order to achieve long- term, successful outcomes, these problems underscore the necessity for better design that puts usability, content quality, and user trust first [3].

II. Related Work

A. Medical Chatbots and Virtual Health Assistants

The management of new technologies and an AI virtual assistant focuses on lowering healthcare expenses by enabling patient self-assessment and prioritization using a flexible model. Babylon Health and Ada Health's two studies look at distinct but complementary approaches to better healthcare. While the AI study demonstrates that technology has the ability to increase healthcare, the ADA study, which operates across different jurisdictions, examines hurdles to access for those with disabilities in the United States, exposing difficulties such insufficient communication gaps and bodily access limitations. These studies indicate that a variety of strategies that combine cutting-edge AI solutions with robust policy protections could improve the health of many people, even though the ADA study demonstrates that legal systems like the ADA are only partially effective in addressing inequities and points to gaps in governance [4, 5]

By managing chronic illnesses, assisting with diagnosis, and enhancing drug adherence through time management and analytics, AI-based medical assistants are transforming the healthcare industry. Through convenient, round-the-clock access, they boost patient participation, which lessens staff effort and enhances long-term health outcomes, particularly in the management of diseases. Notwithstanding these advantages, concerns about data privacy and regulatory compliance still exist. According to research, high-quality care is necessary to guarantee that these instruments function effectively in the medical field and foster patient trust [6].

B. NLP Models in Healthcare

Transformer based models are becoming more and more useful in clinical settings, according to recent studies. Natural language processing (NLP) skills are used by these models, especially large language models (LLMs) like ChatGPT, to improve patient rapport. Applications include patient education, basic support, psychological assistance, and virtual consultation. These Transformer models increase accessibility and engagement in the pain environment, particularly depression in a delicate medical setting, because they are adaptable and configurable, may change to meet the needs of clients, and offer prompt support [7].

Numerous studies have demonstrated that by comprehending complicated clinical language and producing very accurate responses, transformers—in particular, large language models (LLMs) like GPT and BERT—can greatly improve responses in medical applications. For example, studies show that as these models are refined using biological data, they become more accurate in identifying symptoms and recommending therapies, which makes them valuable for clinician support and patient self-assessment tools.[8].

C. Dataset preparation

MedlinePlus, a reliable source of health information from the U.S. National Library of Medicine (NLM), is used in this study. MedlinePlus provides peer-reviewed, evidence-based medical information on illnesses, therapies, and health recommendations. Using this source guarantees that the chatbot complies with HIPAA and GDPR while providing reliable, accurate, and up-to- date medical information [19].

Preprocessing data is crucial to enhancing medical chatbots' functionality, dependability, and security. To guarantee high-quality inputs for AI models, it entails data cleaning, normalization, anonymization, and feature engineering. In accordance with legal requirements such as HIPAA and GDPR, appropriate preprocessing helps reduce bias, inconsistencies, and data leaks. By improving model robustness through strategies like tokenization, outlier detection, and differential privacy, chatbots can produce accurate, fair, and compliant responses in healthcare applications.

D. Challenges in Medical AI Applications

To guarantee responsible use and stakeholder trust, AI in healthcare needs ethical, data privacy, and security considerations. To safeguard sensitive patient data, including procedures like deidentification and consent requirements, compliance with laws like the GDPR is crucial. Making AI honest and reliable requires addressing ethical concerns including decreasing prejudice and boosting responsibility. Transparency in AI decision-making fosters trust in the application of AI in healthcare by assisting patients and doctors in understanding AI-driven insights. To set moral guidelines and preserve patient safety, professionals, specialists, and doctors must work together. applications in healthcare, stressing how crucial it is that Veda AI comply to these guidelines [9, 10].

Significant security, privacy, and regulatory compliance issues arise when integrating Large Language Models (LLMs) in the healthcare industry, especially in light of HIPAA and GDPR. By reducing threats including data poisoning, illegal access, and regulatory non-compliance, LLM Guard improves security.

Model outputs can be tainted by data poisoning, producing dangerous medical recommendations. To stop harmful searches from impacting chatbot behaviour, LLM Guard uses adversarial training, anomaly detection, and input validation. As they handle Personally Identifiable Information (PII) and Protected Health Information (PHI), medical chatbots are often the focus of cyberattacks. LLM Guard employs encryption, role-based access control (RBAC), and real-time monitoring to guard against data exposure and guarantee adherence to security guidelines.

E. Evaluation Metrics for Medical Chatbots

A few important measures are usually used to evaluate healthcare chatbots: accuracy, response speed, user happiness, and trust. Numerous techniques have been employed in studies to gauge this, including comparing chatbot and clinical responses to gauge accuracy. In order to guarantee prompt responses, which are critical for patient involvement and trust, response times are also monitored. In order to determine the perceived worth and dependability of chatbots, user happiness and trust are frequently assessed using patient surveys, interviews, and case studies [11].

III. Proposed System

In order to deliver precise responses and context in medical settings, the chatbot integrates cutting-edge AI technologies. It makes use of Transformer-based language models, particularly the LLaMa model, which comprehends complex queries and produces reliable answers by utilizing location coding and multihead self- monitoring. It is based on the RAG framework and uses LangChain for event data, Faiss for quality data storage, and Hugging Face's BERT for token generation. Tokenization is used to transform user inquiries into statements that show the answers. By eliminating unnecessary information and producing excellent, responsive content, post-processing keeps improving product output.

A. Architecture of the System

The system architecture follows a structured flow (figure 1) integrating a Large Language Model (LLM) with a vector database to enable context-aware question-answering. The process begins with the user login or account creation, followed by authentication. If authentication fails, the system redirects the user back to the login step. Upon successful authentication, the user can engage with the system through an interactive chat interface.

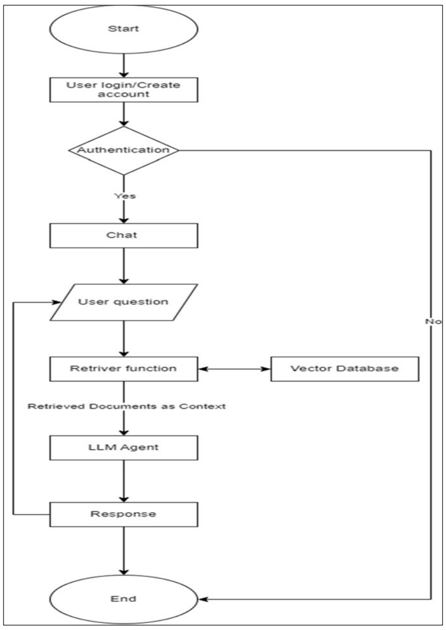

Fig. 1. Flowchart of an LLM-based Question-Answering System with Retrieval Augmented Generation

When a user submits a query, the system processes it and employs a retriever function to search for relevant information stored in a vector database. This retrieval process is facilitated by an embedding model (such as BERT) (figure 2), which converts text into embeddings for efficient similarity-based search. The retrieved documents serve as contextual input for the LLM agent (e.g., Llama 3.2-3B), which analyzes the question alongside the retrieved context to generate an accurate and meaningful response.

The response is then delivered to the user, ensuring improved accuracy through Retrieval-Augmented Generation (RAG). The system allows continuous interaction, where the user can submit further queries, and the loop continues until the session ends. This LangChain-based architecture enhances LLM performance by incorporating external knowledge retrieval, making it highly effective for intelligent chatbots, enterprise assistants, and AI-powered search systems.

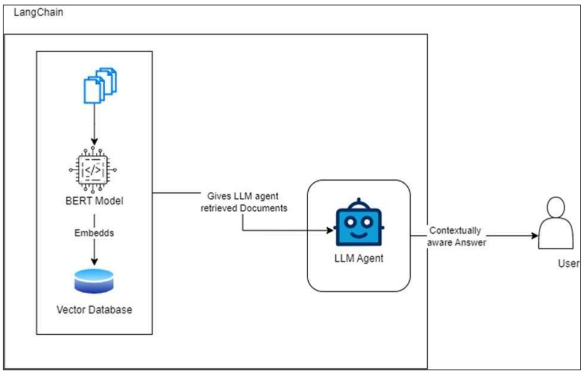
LangChain
BERT Model
Embedds
Vector Database
Gives LLM agent retrieved Documents
LLM Agent
Contextually aware Answer
User

Fig. 2. LangChain-Based Retrieval-Augmented Generation (RAG) Architecture

Below are the key technological components that form the foundation of a robust architectural implementation.

B. Transformer Architecture for Language Modeling

Transformer-based architectures have demonstrated significant superiority over traditional Recurrent Neural Networks (RNNs) in natural language processing tasks [12]. Consequently, the chatbot leverages a transformer-based language model, a highly sophisticated neural network architecture specifically designed for tasks requiring deep natural language understanding. By employing positional encoding and multi-head self-attention mechanisms, the transformer model efficiently captures contextual relationships within linguistic data, allowing it to process complex medical queries with high accuracy [13].

The chatbot specifically utilizes the LLaMa (Large Language Model Meta AI) model, which excels at predicting the next word in a sequence, making it particularly effective in generating coherent and contextually relevant responses. This autoregressive model processes tokens sequentially, ensuring logical and meaningful output based on prior context.

C. RAG-Based Medical Chatbot

The chatbot integrates a Retrieval-Augmented Generation (RAG) framework to enhance response accuracy and information retrieval. This system is implemented using Hugging Face, Faiss, LLaMa 3.2, and LangChain. LLaMa 3.2 was selected due to its open-source availability and research indicating that sequential models like LLaMa are particularly effective in simplifying information retrieval and improving search efficiency [14].

Initially, the Hugging Face BERT model is used to generate data embeddings, which serve as compact vector representations of textual content. BERT's ability to create contextually aware embeddings enhances the chatbot's comprehension and retrieval accuracy [15]. These embeddings are stored in Faiss (Facebook AI Similarity Search) using IndexFlatL2, which facilitates efficient and rapid similarity-based searches [16]. When a user submits a query, LangChain's retriever function compares the query's embeddings with those stored in Faiss, identifying and retrieving the most relevant documents. The chatbot then utilizes this retrieved knowledge to generate precise and insightful responses based on factual information.

D. Natural Language Understanding and Response Generation

Upon receiving a user query, the chatbot formulates a prompt that combines the user's input with predefined instructional guidance. This structured prompt helps ensure that the chatbot generates responses that are both informative and contextually appropriate for medical applications.

The model processes the prompt through tokenization, breaking down text into sequential input tokens for better interpretability [17]. The response is then generated using beam search or greedy decoding techniques. In greedy decoding, the model selects the most probable next word at each step, while beam search evaluates multiple word sequences simultaneously, optimizing for coherence and contextual accuracy.

E. Post-Processing and Response Optimization

To enhance response quality, the chatbot employs post-processing techniques aimed at ensuring clarity and conciseness. This includes removing unnecessary tokens and enforcing length constraints to maintain succinct, well-structured answers. Additionally, a response templating mechanism is implemented to standardize outputs, automate quality control, and enforce predefined response regulations [18]. This ensures that the chatbot consistently delivers high-quality, relevant, and user-friendly responses.

By integrating transformer-based architectures with RAG frameworks, this chatbot model enhances medical query understanding, response accuracy, and retrieval efficiency, making it a powerful tool for AI-driven healthcare assistance.

IV. Results and Analysis

Evaluating the effectiveness of different Large Language Models (LLMs) is essential for understanding their strengths and limitations in generating high-quality responses. This assessment focuses on three key parameters: conciseness, accuracy, and relevance. Conciseness measures how effectively a model delivers information without unnecessary elaboration, accuracy evaluates the correctness and reliability of the responses, and relevance assesses how well the generated answers align with the given query. The table below provides a comparative analysis of the performance of three LLMs—ChatGPT-3.5, Gemini AI, and Claude—based on these metrics.

The results highlight variations in performance across the models. ChatGPT-3.5 demonstrates balanced accuracy and relevance, making it a reliable option for generating informative responses, though its conciseness score indicates some room for improvement in delivering more succinct answers. Gemini AI outperforms the others in conciseness and relevance, making it highly effective in providing precise and contextually appropriate responses. However, its accuracy remains on par with ChatGPT-3.5, suggesting a need for further refinement in correctness. In contrast, Claude exhibits the lowest relevance score, indicating potential challenges in aligning responses with user queries, along with slightly lower accuracy. These insights underscore the importance of selecting an LLM based on specific application requirements to achieve optimal performance.

TABLE I. COMPARATIVE EVALUATION OF LARGE LANGUAGE MODELS (LLMS) BASED ON KEY PERFORMANCE METRICS

Models	Conciseness	Accuracy	Relevance
Chat-GPT 3.5	7	8	8
Gemini AI	9	8	10
Claude	7	7	6

The development of a RAG-based medical chatbot using LangChain demonstrates significant improvements in accuracy, user engagement, and data security for personalized healthcare solutions. By integrating AI-driven retrieval and generation mechanisms, the system enhances patient interactions, supports informed medical decision-making, and bridges the gap between users and reliable healthcare information.

To further improve accessibility, future developments will focus on integrating Speech-to-Text functionality to assist individuals with special needs. Additionally, multilingual support, including Hindi and Gujarati, will be introduced to cater to a broader audience. Ensuring continuous data security enhancements through real-time monitoring and user feedback remains a priority for maintaining privacy and compliance.

To align with HIPAA and GDPR standards, the chatbot incorporates LLM Guard, ensuring audit recording, response filtering, and PHI/PII redaction. These measures enhance data privacy, transparency, and user consent management, making the AI-driven system more secure and trustworthy.

AI bias can contribute to healthcare disparities, often stemming from skewed training data, algorithmic limitations, and systemic inequities. To counteract this, the system employs Explainable AI (XAI), diversified data representation, and bias detection techniques. Prioritizing fairness, inclusivity, and accountability ensures that the chatbot delivers unbiased and reliable healthcare responses.

V. CONCLUSION

This work presents the development of a medical chatbot system leveraging Retrieval-Augmented Generation (RAG) and LangChain to provide accurate and customized medical solutions. By integrating AI-driven methodologies, the system effectively addresses challenges related to user engagement, data security, and diagnostic accuracy. The results demonstrate how AI-powered healthcare solutions can enhance patient interaction, support better health outcomes, and assist in informed decision-making within both home and clinical settings. Furthermore, the chatbot's ability to deliver precise and context-aware responses makes it a valuable tool for improving healthcare accessibility and efficiency.

Future work will focus on enhancing accessibility by improving the Speech-to-Text integration for individuals with special needs and expanding language support to include Hindi, Gujarati, and other regional languages. Strengthening data security through continuous monitoring and compliance with GDPR and HIPAA standards will be crucial in building trust and ensuring privacy. Additionally, addressing AI bias through Explainable AI (XAI), diversified training data, and bias detection techniques will improve transparency and fairness in healthcare delivery. By prioritizing inclusivity, security, and regulatory compliance, AI-driven medical chatbots can revolutionize digital healthcare and provide equitable, efficient, and ethical medical assistance.

REFERENCES

[1] Y. Ke et al., "Development and Testing of Retrieval Augmented Generation in Large Language Models—A Case Study Report," arXiv preprint arXiv:2402.01733, 2024.

[2] L. Xu et al., "Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review," JMIR Cancer, vol. 7, no. 4, p. e27850, 2021.

[3] A. Fadhil and S. Gabrielli, "Addressing challenges in promoting healthy lifestyles," in Proc. 11th EAI Int. Conf. Pervasive Comput. Technol. Healthc., 2017.

[4] N. R. Mudrick and M. A. Schwartz, "Health care under the ADA: A vision or a mirage?," Disability Health J., vol. 3, no. 4, pp. 233–239, 2010, doi: 10.1016/j.dhjo.2010.06.001.

[5] A. Baker et al., "A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis," Front. Artif. Intell., vol. 3, p. 543405, 2020, doi: 10.3389/frai.2020.543405.

[6] A. Al Kuwaiti et al., "A Review of the Role of Artificial Intelligence in Healthcare," J. Pers. Med., vol. 13, no. 6, p. 951, Jun. 2023, doi: 10.3390/jpm13060951.

[7] S. Nerella et al., "Transformers in healthcare: A survey," arXiv preprint arXiv:2307.00067, 2023.

[8] H. Yang, S. Li, and T. Gonçalves, "Enhancing biomedical question answering with large language models," Information, vol. 15, no. 8, p. 494, 2024, doi: 10.3390/info15080494.

[9] M. Amini et al., "Artificial intelligence ethics and challenges in healthcare applications: A comprehensive review in the context of the European GDPR mandate," Mach. Learn. Knowl. Extr., vol. 5, pp. 1023–1035, 2023, doi: 10.3390/make5030053.

[10] S. M. Williamson and V. Prybutok, "Balancing privacy and progress: A review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare," Appl. Sci., vol. 14, p. 675, 2024, doi: 10.3390/app14020675.

[11] J. W. Ayers et al., "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," JAMA Intern. Med., vol. 183, no. 6, pp. 589–596, 2023, doi: 10.1001/jamainternmed.2023.1838.

[12] C. Wang, M. Li, and A. J. Smola, "Language models with transformers," arXiv preprint arXiv:1904.09408, 2019.

[13] X. Ma et al., "A tensorized transformer for language modeling," Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[14] S. Vakayil, D. S. Juliet, A. J., and S. Vakayil, "RAG-Based LLM Chatbot Using Llama-2," in Proc. 7th Int. Conf. Devices, Circuits Syst. (ICDCS), Coimbatore, India, 2024, pp. 1–5, doi: 10.1109/ICDCS59278.2024.10561020.

[15] A. Merchant et al., "What happens to BERT embeddings during fine-tuning?," arXiv preprint arXiv:2004.14448, 2020.

[16] N. Raines et al., "Semantic vector search using an HNSW index for Twitter data," 2023.

[17] R. Taylor et al., "Galactica: A large language model for science," arXiv preprint arXiv:2211.09085, 2022.

[18] J. White et al., "A prompt pattern catalog to enhance prompt engineering with ChatGPT," arXiv preprint arXiv:2302.11382, 2023.

[19] MedlinePlus, "Health Topics," U.S. National Library of Medicine, 2024. Online. Available: https://medlineplus.gov/healthtopics.html.