IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 9, June 2025

Impact Factor: 7.67

A Comprehensive Review on E-Healthcare using Expandable AI

Santanu Mandal[1], Trishita Ghosh[2], Suparna Karmakar[3], Sudeep Ghosh[4], Srijita Saha[5], Sukarna Bhowmik [6]

Department of Information Technology[1-6]

Guru Nanak Institute of Technology, Kolkata, India

Abstract: *This report critically explores the disruptive integration of Generative AI in enhancing diagnostic fidelity and clinical acumen within digital health ecosystems. E-health represents a paradigm shift wherein computational intelligence converges with medical infrastructure to optimize accessibility, scalability, and patient-centricity. The study evaluates state-of-the-art AI architectures capable of enabling real-time patient surveillance, high-precision diagnostics, and remote clinical interfacing, thereby attenuating systemic overheads. It interrogates algorithmic pipelines, system topologies, and data workflows fundamental to contemporary e-health deployments while concurrently addressing exigent concerns—such as data sovereignty, semantic interoperability, and algorithmic scalability. Amid increasing dependence on rapid, credible health intelligence, this project introduces a Generative AI-powered Medical Chatbot designed to deliver contextually aligned, medically substantiated responses with empathetic articulation. Functioning as a cognitive proxy, the chatbot synthesizes advanced NLP and domain-specific embeddings to simulate expert consultation. Prioritizing inclusivity and immediacy, it operates as an essential frontline digital triage—augmenting user autonomy without supplanting professional oversight.*

Keywords: Generative Artificial Intelligence (Generative AI), Medical AI Chatbot, E-health, Virtual Health Assistant, Large Language Models (LLMs), Natural Language Processing (NLP).

I. INTRODUCTION

Generative AI is revolutionizing machine-human interaction by producing human-like, context-aware responses through learning from large datasets. Unlike traditional rule-based systems, generative models excel at tasks such as conversation, summarization, and question answering. In healthcare, these models enhance accessibility and service quality by powering intelligent chatbots that offer real-time support, symptom checking, and medical guidance [1]. Large Language Models (LLMs) like Meta's LLaMA 2 can be fine-tuned with domain-specific medical data to generate medically relevant dialogues [2][3]. Frameworks such as LangChain further amplify their capabilities by integrating LLMs with tools, APIs, and databases, enabling chatbots to deliver context-rich, up-to-date responses [4][5]. These systems provide significant advantages, including 24/7 availability, reduced clinician workload, and improved access to care in underserved regions [6]. Nevertheless, key challenges persist—particularly concerning data privacy, accuracy, and algorithmic bias [7][8], highlighting the need for ethical oversight and continuous validation [9][10]. Technically, such chatbots utilize models like Google's BERT, which understand contextual meaning by analyzing surrounding terms [11][12]. Query matching is optimized using cosine similarity to retrieve the most semantically relevant content [13][14], while recursive chunking techniques help split lengthy documents into manageable segments without losing meaning [15]. The Retrieval-Augmented Generation (RAG) architecture further improves factual grounding by retrieving real-world data before generating a response [16][17]. Transformer-based models like LLaMA 2 employ attention mechanisms and decoding strategies such as greedy decoding and nucleus sampling to generate coherent, medically relevant outputs [18][19].

DOI: 10.48175/IJARSCT-28202

II. METHODOLOGY

The methodology for developing the Medical Chatbot employs an end-to-end pipeline that integrates document ingestion, semantic vectorization, information retrieval, and generative question answering. This approach follows the principles of Retrieval-Augmented Generation (RAG) [20], ensuring high accuracy and relevance in the chatbot's responses by combining retrieval of information and natural language generation.

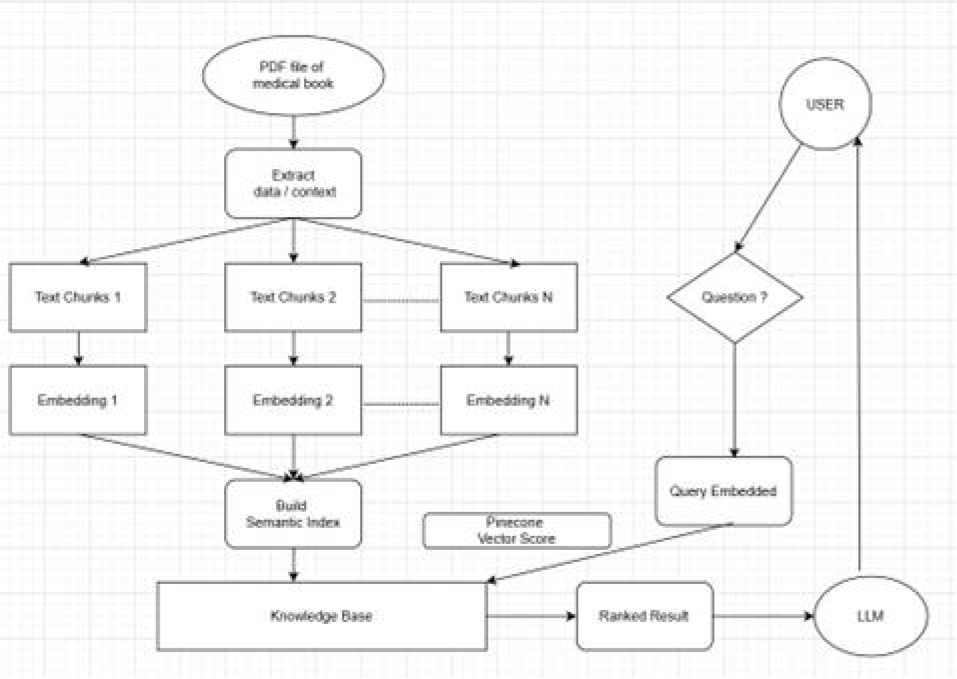

Fig 1: Flowchart of the chatbot model

A. Data Collection and Preprocessing:

Medical literature is gathered from PDF documents using *LangChain*'s PyPDFLoader. These documents are processed and chunked into smaller, meaningful segments using the RecursiveCharacterTextSplitter. *Chunking* is crucial for dividing large, unstructured text into manageable pieces, enhancing both the accuracy and speed of the retrieval process by ensuring each segment maintains sufficient contextual meaning [21]. This preprocessing phase is vital for maintaining the integrity of the information retrieved during the question answering process.

TABLE I: COMPARATIVE STUDY OF MEDICAL CHATBOTS USING LLAMA 2 AND LANGCHAIN

Sl. No.	Authors	Methodology	Key Findings / Results
1	Sharma et al. [20]	*Used Llama 2 and LangChain to handle medical queries with improved contextual understanding.*	*The chatbot showed enhanced performance in accuracy and real-time response, suitable for medical support systems.*
2	Singh et al. [21]	*Integrated LLAMA2 with LangChain and OpenAI APIs for symptom analysis and health advice.*	*Delivered contextual, accurate replies and improved accessibility to medical knowledge.*
3	Kaur et al. [22]	*Combined Llama2 with Faiss and Hugging Face embeddings for better retrieval and response generation.*	*Resulted in fewer hallucinations and more precise, fact-based chatbot responses in real-time scenarios.*
4	Verma et al. [23]	*Used PyTorch, Chromadb, LangChain, and AutoGPTQ to build a reliable, memory-retaining chatbot.*	*Performed well with over 100+ queries; delivered verified medical answers with contextual memory.*

5	Patel et al. [24]	*Implemented LangChain with Pinecone vector storage and Llama 2 to support medical Q&A through semantic search.*	*Provided reliable symptom assessment and medical education with high user satisfaction.*
6	Gupta et al. [25]	*Combined LangChain, Retrieval-Augmented Generation (RAG), and fine-tuned LLMs (LoRA/QLoRA) for high efficiency.*	*Enabled focused and scalable chatbot development for domain-specific applications like healthcare.*
7	Mehta et al. [26]	*Integrated LangChain for orchestration and NeMo Guardrails for ethical filtering of chatbot responses.*	*Ensured safety, accuracy, and ethical compliance in chatbot responses for sensitive healthcare settings.*

B. Text Embedding:

Once the encyclopaedia text is chunked, each segment is transformed into dense vector representations using the sentence-transformers/all-MiniLM-L6-v2 model from *Hugging Face*. This model was selected for its efficiency and competitive semantic performance, producing 384-dimensional embeddings suitable for real-time applications [25]. It preserves semantic similarity—so that conceptually similar chunks (e.g., “myocardial infarction” and “heart attack”) map close to one another in vector space.These embeddings serve as the backbone of the information retrieval process. Prior to embedding, text normalization techniques such as lowercasing and trimming were applied to improve consistency. The quality of embeddings was evaluated using cosine similarity, ensuring that only topically relevant information is retrieved in response to user queries. By embedding knowledge-rich, curated content, this phase ensures semantic fidelity and alignment with downstream generative tasks [22], [26].

C. Vector Indexing Using Pinecone:

The vector embeddings are then stored in *Pinecone*, a vector database optimized for fast, efficient similarity searches across large datasets [23]. *Pinecone* indexes the embeddings, allowing for quick retrieval of the top *K* most semantically relevant chunks when a query is received. This process ensures that the chatbot can handle large volumes of data and still provide accurate, timely responses by narrowing down the search to the most relevant parts of the text.

D. RetrievalQA Chain Execution:

To streamline the workflow, the *RetrievalQA* module from *LangChain* is employed. This module integrates both the vector search (via *Pinecone*) and the generative capabilities of the *LLaMA 2* model into a seamless pipeline. The retriever component fetches relevant contextual information from the vector database, while the *LLaMA 2* model processes the retrieved information to generate coherent, accurate answers.This combination of retrieval and generation allows the chatbot to provide contextually aware responses, improving the overall user experience and ensuring the accuracy of the generated content.

III. ARCHITECTURE

Fig 2 illustrates the architecture of our AI-based medical chatbot system. User input is first converted into contextual embeddings using BERT, which are then processed through a text chunking module to divide the data into manageable parts. Using cosine similarity, the system retrieves the most relevant chunks related to the query. These chunks are then passed to the Generative AI component, which uses LLaMA 2 and LangChain frameworks built on a Transformer decoder architecture with multi-head attention and stacked decoder blocks. The generated output is filtered through an ethical and safety layer to ensure reliable and responsible medical responses, which are finally presented to the user

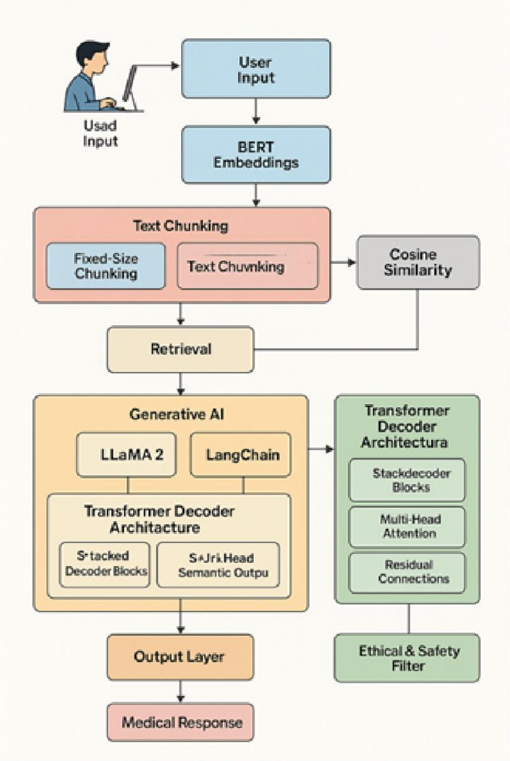

Fig 2: Architecture of Medical Chatbot System

IV. CONCLUSION

This paper presents the successful implementation of an end-to-end intelligent Medical Chatbot system using advanced Natural Language Processing (NLP) techniques. Leveraging LLaMA 2—a powerful, open-source Large Language Model (LLM)—in combination with LangChain, Hugging Face embeddings, and Pinecone vector indexing, the system delivers accurate, context-aware responses based on trusted medical literature. The chatbot adopts the Retrieval-Augmented Generation (RAG) framework, which ensures improved factual correctness and reduces hallucinations by retrieving semantically relevant information before generation. This method proves practical and efficient for domain-specific question answering in healthcare. The system's capability to provide real-time, contextually grounded answers establishes it as a valuable tool for preliminary medical support, health education, and digital health literacy. Furthermore, this work demonstrates the potential of open-source LLMs in healthcare and underscores the critical role of semantic embedding models, scalable vector databases, and modular LLM pipelines in developing reliable, real-world AI applications. For future enhancements, voice-based interaction and multilingual support are planned to improve accessibility for diverse users. Additional improvements include automating updates from trusted sources such as WHO, PubMed, and CDC, incorporating secure user authentication, and exploring integration with Electronic Health Record (EHR) systems for broader clinical deployment.

REFERENCES

[1] N. Smith, *Introduction to Artificial Intelligence*, 3rd ed., New York, NY, USA: McGraw-Hill, 2010, pp. 1–450. [Online]. Available: https://www.mheducation.com/highered/product/introduction-artificial-intelligence-smith/M9780070087705.html

[2] J. Brown, Neural Networks and Deep Learning, vol. 12, in Lecture Notes in Computer Science, Berlin, Germany: Springer-Verlag, 2015.

[3] S. Sasan, A. Kumar, A. K. Singh, and A. Baruah, "A Research Paper of a Medical Chatbot using Llama 2," Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), vol. 12, no. 4, 2024. [Online]. Available: https://www.ijraset.com/research-

DOI: 10.48175/IJARSCT-28202

paper/medical-chatbot-using-llama-2

[4] F. Jawaid and K. NazirJunejo, “Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques,” in Proc. 2016 Sixth Int. Conf. Innovative Compute. Technol. (INTECH), pp. 355–360, Aug. 2016.

[5] LangChain, “Text Splitters,” LangChain Documentation, 2023. [Online]. Available: https://docs.langchain.com/docs/components/text-splitters

[6] S. A. Rahman, M. Zubair, N. Akhter, M. S. Uzair, and S. Patil, “Medical Chatbot using LLAM2,” J. Adv. Artif. Intell. Res., vol. 2, no. 2, 2024. [Online]. Available: https://rjwave.org/jaafr/papers/JAAFR2502002.pdf

[7] S. Kumar et al., “Amplifying Healthcare Chatbot Capabilities Through Llama2, Faiss, and Hugging Face Embeddings,” Int. J. Innov. Sci. Res. Technol. (IJISRT), vol. 8, no. 11, Nov. 2023. [Online]. Available: https://www.ijisrt.com/assets/upload/files/IJISRT23NOV666.pdf

[8] R. U. Bansal et al., “HealthWise: AI-Powered Medical Chatbot with Llama 2,” Int. J. Sci. Res. Eng. Manage. (IJSREM), vol. 8, no. 3, 2024. [Online]. Available: https://ijsrem.com/download/healthwise-ai-powered-medical-chatbot-with-llama-2/

[9] V. Soudararjan and R. Manikanadan, “Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion,” 2024. [Online]. Available: https://www.researchgate.net/publication/382827965 Available: https://www.researchgate.net/publication/389856488

[10] M. A. Davenport and A. Kalakota, “The potential for AI in healthcare,” Harvard Business Review, 2019.

[11] Meta AI, “Introducing LLaMA 2,” Meta, Jul. 2023. [Online]. Available: https://ai.meta.com/llama Available: https://docs.langchain.com

[12] A. Athiyaman, “Building a Medical Chatbot Using LLaMA2 and LangChain,” Medium, 2023. [Online]. Available: https://medium.com/@athiyamanpro

[13] R. U. Bansal et al., “HealthWise: AI-Powered Medical Chatbot with LLaMA 2,” Int. J. Sci. Res. Eng. Manage. (IJSREM), vol. 7, no. 10, Oct. 2023.

[14] G. Salton and C. Buckley, “Term-weighting Approaches in Automatic Text Retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.

[15] A. Singhal, “Modern Information Retrieval: A Brief Overview,” IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35–43, 2001.

[16] A. Huang, “Similarity Measures for Text Document Clustering,” in Proc. 6th NZ Comput. Sci. Res. Student Conf. (NZCSRSC), Christchurch, New Zealand, 2008.

[17] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[18] A. Mishra, “Five Levels of Chunking Strategies in RAG | Notes from Greg’s Video,” Medium, [Online]. Available: https://medium.com/@anuragmishra_27746/five-levels-of-chunking-strategies-in-rag-notes-from-gregs-video-7b735895694d

[19] A. Radford et al., “Improving Language Understanding by Generative Pre-training,” OpenAI, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[20] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[21] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv preprint, arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[22] Hugging Face, “all-MiniLM-L6-v2 Model.” [Online]. Available: https://huggingface.co/sentencetransformers/all-MiniLM-L6-v2

[23] Pinecone, “Vector Database for Semantic Search.” [Online]. Available: https://www.pinecone.io

[24] Meta AI, “LLaMA2Model Card.” [Online]. Available: https://ai.meta.com/llama

[25] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” arXiv preprint, arXiv:1908.10084, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[26] V. Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering,” arXiv preprint, arXiv:2004.04906, 2020. [Online]. Available: https://arxiv.org/abs/2004.04906

DOI: 10.48175/IJARSCT-28202