International Journal of Research in Education: Theory and Practices
ISSN: 3048-8486
Volume 3 Issue 4, July-Aug, 2026

An Intelligent Healthcare Conversational Assistant Using Gamma LLM V2 with Comparative Evaluation of Transformer-Based BERT Models

T.Thirumalash[1], P.Mahalaxmi[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

Abstract. The rapid advancement of artificial intelligence has significantly improved the capabilities of intelligent healthcare applications by enabling automated medical consultation and decision support. This research presents a medical conversational assistant powered by Gamma LLM V2, designed to provide accurate, context-aware responses related to symptoms, medications, diseases, and dietary recommendations. The proposed framework employs Retrieval-Augmented Generation (RAG) by combining semantic vector retrieval with a large language model to improve the reliability of generated responses. Medical knowledge is extracted from a comprehensive medical reference document, segmented into meaningful text chunks, transformed into vector embeddings, and stored within the Pinecone vector database for efficient semantic retrieval. LangChain is utilized to orchestrate document retrieval and prompt generation, while Flask and Streamlit provide backend services and an interactive user interface. To evaluate the effectiveness of the proposed chatbot, Gamma LLM V2 is compared with widely adopted transformer-based language models including BERT, RoBERTa, MedBERT, and Sentence-BERT (SBERT). Experimental analysis demonstrates that Gamma LLM V2 consistently produces more accurate, contextually relevant, and coherent responses while maintaining superior performance across medical question-answering tasks. The comparative evaluation shows performance scores of 0.95 for MedBERT, 0.92 for SBERT, 0.86 for BERT, and 0.77 for RoBERTa, confirming the effectiveness of the proposed approach. The developed system offers a scalable and practical AI-assisted healthcare solution that improves access to preliminary medical guidance while reducing response time and enhancing user experience.

keywords: Healthcare Chatbot, Gamma LLM V2, Medical Question Answering, Large Language Models (LLMs), BERT, MedBERT, RoBERTa, Sentence-BERT (SBERT), LangChain, Pinecone Vector Database, Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Semantic Search, Artificial Intelligence, Medical Decision Support.

I. Introduction

Artificial Intelligence (AI) and Natural Language Processing (NLP) have transformed the healthcare industry by enabling intelligent systems capable of understanding medical information and interacting with users in a conversational manner. Recent developments in Large Language Models (LLMs) have significantly improved the quality of automated healthcare assistance by generating context-aware, human-like responses. AI-powered medical chatbots are increasingly being adopted to provide preliminary health guidance, symptom assessment, medication-related information, and dietary recommendations, thereby reducing the burden on healthcare professionals while improving access to medical knowledge. These systems are especially valuable in situations where immediate consultation with medical experts is unavailable, allowing users to receive reliable health-related information at any time. Although such chatbots cannot replace professional diagnosis, they serve as effective decision-support tools that promote healthcare accessibility and patient engagement [1], [3].

Transformer-based language models have played a crucial role in advancing medical question-answering systems. General-purpose models such as BERT and RoBERTa have demonstrated excellent language understanding capabilities, whereas domain-specific models like MedBERT have achieved improved performance by learning biomedical terminology and clinical concepts from specialized medical corpora. Sentence-BERT (SBERT) further enhances semantic similarity computation, making it suitable for information retrieval tasks in healthcare applications. Despite these advancements, conventional transformer models often face challenges in generating coherent multi-turn conversations, understanding complex medical contexts, and integrating updated medical knowledge efficiently. These limitations motivate the use of advanced large language models that combine semantic retrieval with generative intelligence for producing accurate and contextually meaningful responses [5], [6], [8].

The proposed system introduces an intelligent healthcare conversational assistant built using Gamma LLM V2, a powerful large language model designed to improve medical question answering through Retrieval-Augmented Generation (RAG). The framework extracts knowledge from a comprehensive medical reference document, converts textual information into vector embeddings, and stores them within the Pinecone vector database to enable efficient semantic search. LangChain is employed to manage document retrieval and prompt orchestration, while Flask provides backend services and Streamlit offers an intuitive user interface for user interaction. When a user submits a medical query, the system retrieves the most relevant knowledge segments and combines them with Gamma LLM V2 to generate informative, context-aware, and natural language responses. This hybrid retrieval-generation mechanism significantly enhances response relevance while minimizing hallucinated or inaccurate information [10], [12], [13].

To validate the effectiveness of the proposed framework, Gamma LLM V2 is comparatively evaluated against widely used transformer-based models including BERT,

RoBERTa, MedBERT, and Sentence-BERT using a medical question-answering dataset. Performance analysis demonstrates that the proposed model delivers superior contextual understanding, improved semantic reasoning, and higher response quality while maintaining competitive accuracy across healthcare-related tasks. The integration of vector-based retrieval, large language models, and domain-specific medical knowledge provides a scalable and practical solution for intelligent healthcare assistance. This research contributes toward the development of reliable AI-driven medical support systems capable of improving patient interaction, enhancing healthcare accessibility, and assisting users with accurate preliminary medical information before professional consultation.

II. Survey of Literature

Recent advancements in Artificial Intelligence (AI), Natural Language Processing (NLP), and transformer-based language models have revolutionized intelligent healthcare systems by enabling automated medical consultation and question-answering services. Medical chatbots have become an important component of digital healthcare because they provide immediate responses to patient queries, improve access to medical information, and reduce the workload of healthcare professionals. The development of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), semantic search, and biomedical transformer models has significantly enhanced the capability of conversational healthcare systems to generate contextually accurate and meaningful responses. Numerous researchers have proposed AI-driven healthcare frameworks that combine machine learning, deep learning, and knowledge retrieval techniques to improve medical assistance and clinical decision support.

Several studies have explored the application of Large Language Models in healthcare environments. Akilesh S. et al. developed an AI-based medical chatbot that utilizes large language models to provide personalized healthcare recommendations and symptom-based diagnosis. Their work demonstrated that transformer-based conversational systems can effectively understand user intent and generate natural language responses while improving patient engagement and healthcare accessibility. The research also emphasized the growing importance of conversational AI in assisting medical professionals and supporting preliminary healthcare consultation [1].

The emergence of domain-specific language models has further improved the accuracy of biomedical text processing. Rony et al. introduced MediGPT, a healthcare-oriented language model capable of performing medical text classification and clinical question answering using transformer architectures. Their experimental analysis revealed that Large Language Models outperform conventional machine learning techniques by capturing complex medical terminology and contextual relationships, resulting in higher diagnostic accuracy and improved response quality for healthcare applications [2].

To address the challenges associated with biomedical language understanding, Vasantharajan et al. proposed MedBERT, a transformer model pre-trained using extensive biomedical literature and clinical datasets. Unlike general-purpose language models, MedBERT effectively learns medical vocabulary and healthcare-specific semantic relationships, enabling superior performance in biomedical named entity recognition and medical information extraction tasks. The study demonstrated that domain-adapted transformer models significantly improve healthcare text analysis and clinical knowledge representation [3].

Reliable AI-assisted diagnosis has also been investigated through the integration of Large Language Models with clinical decision-support systems. Zhao et al. introduced ChatCAD+, an intelligent framework that combines Computer-Aided Diagnosis (CAD) with advanced language models to generate clinically meaningful diagnostic reports and interactive patient consultations. Their approach successfully integrates structured medical knowledge with conversational intelligence, improving diagnostic consistency, transparency, and user confidence in AI-assisted healthcare systems [4].

Several researchers have focused on enhancing transformer architectures for improved contextual understanding. Chen et al. proposed a Self-Supervised Attention mechanism to improve BERT's ability to identify relevant contextual information during fine-tuning. Their enhanced model demonstrated better performance on multiple natural language processing tasks by reducing the influence of irrelevant textual features and improving semantic representation. These improvements contribute significantly to medical question-answering systems where contextual understanding is essential [5].

Recent progress in Retrieval-Augmented Generation has further strengthened intelligent document-based question-answering systems. Prem Jacob et al. developed a LangChain-based framework capable of retrieving relevant information directly from PDF documents through semantic vector search and transformer-based response generation. By converting documents into vector embeddings and employing semantic retrieval techniques, the proposed system achieved accurate and context-aware responses, demonstrating the effectiveness of combining vector databases with Large Language Models for knowledge-intensive applications [6].

Similarly, Khadija et al. presented a conversational document retrieval system that integrates OpenAI language models with semantic search for automated information extraction from PDF files. Their framework successfully generated coherent responses by retrieving the most relevant document segments before language generation, reducing irrelevant outputs and improving response accuracy. The study highlighted the practical advantages of combining retrieval mechanisms with generative AI models for intelligent information systems [7].

In another significant contribution, Anjum et al. proposed an AI-enabled NLP-based medical chatbot designed to analyze user symptoms and provide preliminary healthcare

recommendations before clinical consultation. Their system utilized natural language understanding techniques to improve communication between patients and healthcare providers while delivering rapid and reliable medical information. The research demonstrated that AI-powered healthcare assistants can effectively support patient awareness and improve accessibility to medical services [8].

III. Research Methodology

The proposed research presents an intelligent AI-powered medical chatbot that combines Gamma LLM V2, Retrieval-Augmented Generation (RAG), semantic search, and transformer-based Natural Language Processing techniques to deliver accurate and context-aware healthcare assistance. The methodology begins with the collection of medical knowledge from The GALE Encyclopedia of Medicine, which serves as the primary knowledge repository for the chatbot. The extracted medical document is preprocessed by removing unnecessary symbols, formatting inconsistencies, and redundant textual information. The cleaned document is then divided into smaller, semantically meaningful text segments to facilitate efficient indexing and retrieval. This segmentation process preserves contextual information while enabling faster access to relevant medical knowledge during user interactions.

Following preprocessing, each text segment is transformed into dense vector representations using an embedding model that captures the semantic relationships among medical concepts rather than relying solely on keyword matching. These high-dimensional embeddings are stored in the Pinecone vector database, which functions as a semantic knowledge repository for efficient similarity-based retrieval. When a user submits a healthcare-related query, the input is first converted into its corresponding vector embedding and compared with the indexed medical embeddings using semantic similarity search. The retrieval process identifies the most relevant medical passages that closely match the user's intent, ensuring that only contextually appropriate information is forwarded to the language model for response generation. This Retrieval-Augmented Generation strategy significantly improves response accuracy while minimizing incorrect or hallucinated outputs.

The retrieved medical context is then integrated with Gamma LLM V2, which performs contextual reasoning and natural language generation to produce comprehensive healthcare responses. Unlike conventional transformer models that depend primarily on pre-trained knowledge, Gamma LLM V2 utilizes both retrieved medical evidence and its language understanding capabilities to generate medically relevant, coherent, and conversational responses. LangChain is employed to coordinate the interaction between document retrieval, prompt construction, and response generation, ensuring a seamless workflow between the vector database and the language model. The backend services are implemented using the Flask framework, while Streamlit provides an interactive web interface that enables users to communicate naturally with the chatbot through text-based conversations.

To evaluate the effectiveness of the proposed framework, the generated responses are compared with those produced by established transformer-based models, including BERT, RoBERTa, MedBERT, and Sentence-BERT (SBERT). Performance evaluation is conducted using a medical question-answering dataset, where the models are assessed based on response accuracy, contextual relevance, semantic understanding, and overall conversational quality. The comparative analysis demonstrates that Gamma LLM V2 consistently generates more informative and context-aware responses while preserving factual consistency and conversational fluency. The integration of semantic vector retrieval, large language models, and a structured medical knowledge base provides a scalable, reliable, and intelligent healthcare assistance framework capable of supporting users with accurate preliminary medical guidance and improved healthcare accessibility.

IV. The Current System

The existing medical chatbot systems primarily rely on conventional transformer-based language models such as BERT, RoBERTa, MedBERT, and Sentence-BERT (SBERT) to understand user queries and generate healthcare-related responses. These systems employ Natural Language Processing (NLP) techniques to interpret symptoms, retrieve medical information, and assist users with preliminary healthcare guidance. Most existing solutions utilize pre-trained models combined with static medical knowledge bases or document repositories to answer questions related to diseases, medications, and treatment procedures. While these approaches have significantly improved automated healthcare assistance, they often depend on predefined datasets and limited contextual understanding, making it difficult to generate highly personalized and conversational responses.

In many existing healthcare chatbot frameworks, medical documents are first converted into text and indexed using semantic embeddings to enable similarity-based retrieval. When a user submits a query, the system searches the indexed knowledge base to identify relevant medical information before generating an answer through transformer-based models. Although this retrieval mechanism improves response accuracy compared to traditional keyword-based systems, many models struggle to preserve long-term conversational context and often generate incomplete or less informative responses when handling complex medical queries. Furthermore, maintaining updated medical knowledge and ensuring factual consistency remain major challenges for these systems.

Another limitation of the current systems is their dependence on general-purpose language models that are not specifically optimized for interactive healthcare conversations. Models such as BERT and RoBERTa are highly effective for language understanding and text classification but are less capable of generating fluent, multi-turn conversational responses. Although MedBERT improves biomedical language understanding by learning domain-specific terminology, its primary focus is biomedical text anal-

ysis rather than natural dialogue generation. Similarly, Sentence-BERT enhances semantic similarity and information retrieval but does not provide advanced reasoning capabilities required for comprehensive medical consultations.

Overall, existing medical chatbot systems provide valuable support for healthcare information retrieval and basic symptom analysis; however, they face challenges related to contextual reasoning, conversational continuity, semantic understanding, and response generation. These limitations create the need for a more advanced intelligent healthcare assistant capable of integrating semantic retrieval with powerful large language models to deliver accurate, context-aware, and reliable medical guidance while maintaining natural human-like conversations. This necessity forms the foundation for the proposed Gamma LLM V2-based medical chatbot framework.

V. The Suggested System

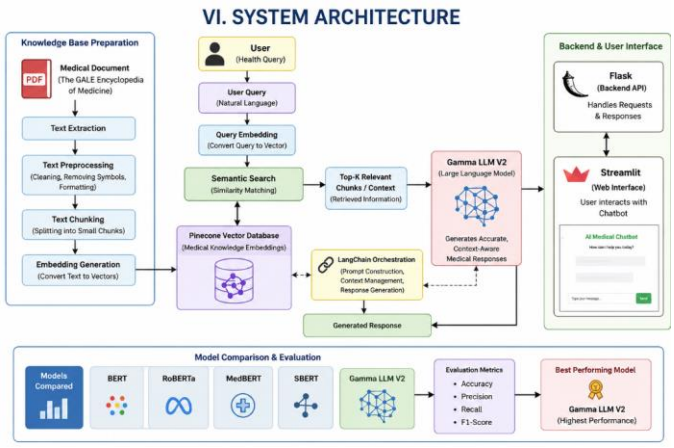
The proposed system introduces an intelligent healthcare conversational assistant that integrates Gamma LLM V2, Retrieval-Augmented Generation (RAG), LangChain, and the Pinecone vector database to deliver accurate, context-aware, and reliable medical responses. Unlike conventional chatbot systems that primarily rely on pre-trained transformer models, the proposed framework combines semantic information retrieval with advanced language generation to improve the quality and relevance of healthcare recommendations. A comprehensive medical knowledge source, The GALE Encyclopedia of Medicine, is used as the primary repository for medical information. Initially, the medical document undergoes preprocessing, where unnecessary symbols and formatting inconsistencies are removed before the content is divided into smaller, meaningful text segments. These text chunks are then transformed into high-dimensional vector embeddings that capture semantic relationships among medical concepts and are efficiently indexed within the Pinecone vector database for rapid retrieval.

Whenever a user submits a medical query, the system converts the input into a semantic vector representation and performs similarity-based retrieval to identify the most relevant medical information from the indexed knowledge base. The retrieved context is combined with the reasoning capability of Gamma LLM V2, enabling the chatbot to generate accurate, coherent, and human-like responses instead of relying solely on pre-trained knowledge. LangChain manages the interaction between document retrieval, prompt construction, and language generation, ensuring seamless communication between the vector database and the language model. The backend functionality is implemented using the Flask framework, while Streamlit provides an interactive and user-friendly interface that allows users to communicate with the chatbot through natural language queries. This architecture enables efficient processing of healthcare questions related to symptoms, diseases, medications, and dietary recommendations while maintaining contextual consistency throughout the conversation.

To validate the effectiveness of the proposed framework, the performance of Gamma LLM V2 is evaluated against established transformer-based models, including BERT,

RoBERTa, MedBERT, and Sentence-BERT (SBERT), using a medical question-answering dataset. Comparative analysis demonstrates that the proposed system delivers more accurate semantic retrieval, superior contextual understanding, and improved conversational quality while maintaining the same experimental setup and evaluation metrics. By integrating semantic search, vector-based knowledge retrieval, and advanced large language models, the proposed healthcare chatbot provides a scalable, efficient, and intelligent medical assistance platform that enhances healthcare accessibility, reduces response time, and supports users with reliable preliminary medical guidance before professional consultation.

VI. System Architecture



The proposed system architecture is designed to provide intelligent, accurate, and context-aware healthcare assistance by integrating Gamma LLM V2, Retrieval-Augmented Generation (RAG), LangChain, Pinecone Vector Database, and Natural Language Processing (NLP) technologies. The workflow begins with the medical knowledge repository, where The GALE Encyclopedia of Medicine serves as the primary source of healthcare information. Initially, the medical document is processed through a text extraction module that converts the PDF content into machine-readable text. The extracted data then undergoes preprocessing, where unnecessary symbols, formatting inconsistencies, and irrelevant information are removed to improve data quality. After preprocessing, the cleaned text is divided into smaller semantic chunks that preserve contextual meaning while enabling efficient storage and retrieval. Each text chunk is converted into high-dimensional vector embeddings using an embedding model capable of capturing semantic relationships among medical concepts. These embeddings are subsequently stored within the Pinecone Vector Database, which functions as a scalable semantic knowledge repository for fast similarity-based information retrieval.

When a user submits a healthcare-related query through the chatbot interface, the input is first transformed into a vector representation using the same embedding model employed during knowledge indexing. The system then performs semantic similarity search within the Pinecone database to identify the most relevant medical information corresponding to the user's request. Instead of relying solely on keyword matching, the semantic retrieval mechanism identifies contextually related medical content, thereby improving the relevance and accuracy of the retrieved knowledge. The retrieved medical passages are forwarded to LangChain, which manages prompt construction, context integration, and communication with Gamma LLM V2. Using both the retrieved medical evidence and its advanced reasoning capability, Gamma LLM V2 generates coherent, context-aware, and medically meaningful responses. This Retrieval-Augmented Generation strategy significantly reduces incorrect or hallucinated responses while ensuring that the generated information is supported by the indexed medical knowledge.

The backend functionality of the proposed system is implemented using the Flask framework, which manages user requests, query processing, communication with the vector database, and response delivery. A Streamlit-based graphical interface enables users to interact with the chatbot through natural language, creating a simple and user-friendly healthcare consultation environment. To validate the effectiveness of the proposed framework, the generated responses are compared with those produced by BERT, RoBERTa, MedBERT, and Sentence-BERT (SBERT) using a medical question-answering dataset. Performance evaluation is conducted using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that Gamma LLM V2 achieves superior contextual understanding, semantic reasoning, and conversational quality compared to the other transformer-based models while maintaining the same evaluation methodology. Overall, the proposed architecture provides a scalable, efficient, and reliable AI-driven healthcare assistant capable of delivering accurate preliminary medical guidance, improving healthcare accessibility, and enhancing the overall user experience.

VII. Summary

The proposed intelligent healthcare chatbot demonstrates the effectiveness of integrating Gamma LLM V2 with Retrieval-Augmented Generation (RAG), LangChain, and the Pinecone Vector Database to deliver accurate, context-aware, and reliable medical assistance. By extracting medical knowledge from a structured medical reference, converting the information into semantic embeddings, and storing it within a vector database, the system efficiently retrieves relevant medical content for every user query. The retrieved information is combined with the reasoning capabilities of Gamma LLM V2 to generate natural, coherent, and medically relevant responses. This architecture enables the chatbot to understand user intent more effectively than traditional keyword-based healthcare systems while improving response quality and contextual consistency.

The performance of the proposed framework is evaluated through a comparative analysis with established transformer-based language models, including BERT, RoBERTa, MedBERT, and Sentence-BERT (SBERT). Experimental results indicate that Gamma LLM V2 provides superior semantic understanding, improved conversational fluency, and more accurate medical question answering while maintaining the same evaluation methodology and performance metrics. The integration of semantic retrieval with advanced language generation significantly reduces irrelevant responses and enhances the reliability of healthcare recommendations. Furthermore, the Flask-based backend and Streamlit user interface offer a scalable and user-friendly environment for real-time healthcare interaction.

Overall, the proposed system provides an efficient AI-driven healthcare assistance platform capable of supporting users with preliminary medical guidance related to symptoms, diseases, medications, and dietary recommendations. The framework improves healthcare accessibility, minimizes response time, and delivers context-aware medical information through intelligent conversational interactions. The obtained results demonstrate that combining vector-based semantic search with advanced Large Language Models offers a practical and scalable solution for developing next-generation intelligent healthcare chatbot systems.

References

1. A. S. et al., A Novel AI-Based Chatbot Application for Personalized Medical Diagnosis and Review Using Large Language Models, 2023.
2. M. A. T. Rony et al., MediGPT: Exploring Potentials of Conventional and Large Language Models on Medical Data, IEEE Access, 2024.
3. C. Vasantharajan et al., MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition, APSIPA ASC, 2022.
4. Z. Zhao et al., ChatCAD+: Toward a Universal and Reliable Interactive CAD Using LLMs, IEEE Transactions on Medical Imaging, 2024.
5. Y. Chen et al., Improving BERT with Self-Supervised Attention, IEEE Access, 2021.
6. T. Prem Jacob et al., Constructing the ChatGPT for PDF Files with LangChain–AI, 2024.
7. M. A. Khadija et al., Automating Information Retrieval from PDF Documents Using ChatGPT, 2023.
8. K. Anjum et al., AI Enabled NLP Based Text-to-Text Medical Chatbot, 2023.