ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/672-679
Mr. Mohammed Sami *et. al.*, / International Journal of Engineering & Science Research
Full Length Article

Medical Chatbot using Gamma LLMV2 and Comparison Using BERT Models

Ms. Hafsa Tasneem *1, Mr. Mohammed Sami *2, Mr. Abdul Shafiq *3, Mr. Shah Alam Hussaini *4,
*1 Assistant Professor, Dept. of CSE-AIML, Lords Institute of Engineering and Technology
*2, 3, 4 B.E Student Dept. of CSE-AIML, Lords Institute of Engineering and Technology
hafsatasneem@lords.ac.in*1, mdtajamulsami09@gmail.com*2, abdulshafiq413@gmail.com*3, adkhan.ak15@gmail.com*4
Accepted 22-04-2026

ABSTRACT
The study introduces a sophisticated medical chatbot that uses vector-based retrieval and Meta's LLaMA 2 model to provide accurate, context-aware recommendations on symptoms, drugs, and diets. The system incorporates Pinecone for vector embeddings, LangChain for interactions, and Python for logic. "The GALE Encyclopedia of Medicine," a 637-page medical dataset, is broken up into text segments for effective semantic search.Flask is used for web infrastructure and Streamlit for interaction with the chatbot's front end. User queries use LLaMA 2 to deliver answers, create embeddings, and do vector
searches. While handling edge circumstances and data quality provide issues, evaluation emphasizes accuracy, timeliness, and engagement. Real-time updates and sophisticated fine-tuning are examples of upcoming enhancements.In terms of language interpretation, generation, and reasoning, Gamma LLM v2 performs better than proprietary models when compared to other models. This enables fine-tuning on bespoke datasets and lessens the need for APIs. It outperforms RoBERTa (0.77), MedBERT (0.95), and BERT (0.86) in medical question-answering tasks and is available in 7B, 13B, and 70B parameters. Index Terms— Gamma LLM-V2,Chat bot, BioBERT, MedBERT, Langchain.

INTRODUCTION

Intelligent healthcare systems are now possible because to developments in AI and NLP. Using Gamma LLMV2, this study creates a medical chatbot and evaluates it against BERT variants. The chatbot uses Meta's LLaMA 2 to provide precise, context-aware answers when assisting with symptoms, prescriptions, and diet. The system incorporates Python for backend functionality, Pinecone for vector embeddings, and LangChain for conversational interactions. For effective semantic search, a 637-page medical dataset is preprocessed into text segments and embedded. Built with Flask and Streamlit, the front end offers an easy-to-use interface. To produce answers, LLaMA 2 processes user inputs once they have been integrated and searched in Pinecone. The system's potential to improve patient involvement is highlighted by evaluations based on accuracy, reaction time, and user happiness. Managing medical edge cases, maintaining data quality, and integrating real-time medical updates are among the difficulties.

PROJECT OVERVIEW

This project presents the development of an intelligent medical chatbot using **Gamma LLM v2** and evaluates its performance in comparison with BERT-based models. The system is designed to provide accurate and context-aware responses to user queries related to symptoms, medications, and dietary recommendations. It utilizes a vector-based retrieval mechanism where medical data, extracted from *The GALE Encyclopedia of Medicine*, is converted into embeddings and stored in a Pinecone database for efficient semantic search.

The architecture integrates technologies such as LLaMA 2 for response generation, LangChain for conversational management, and Flask and Streamlit for backend and frontend development. The workflow involves data extraction, text chunking, embedding generation, semantic indexing, and response generation. This approach enables real-time interaction and enhances accessibility to preliminary healthcare information.

LITERATURE SURVEY

1) A Novel AI-based Chatbot Application for Personalized Medical Diagnosis using LLMs
Authors: A. S et al.
This study presents a GPT-3.5 based medical chatbot integrated into an Android application. It enables users

to input symptoms and receive personalized medical advice using NLP. The system improves accessibility, decision-making, and user interaction in healthcare through context-aware responses.

2) Identification of Ancient Chinese Medical Prescriptions using AI (GPT)

Authors: M. Li and X. Zheng

This work applies ChatGPT for analyzing ancient medical literature and prescription data. It combines supervised and semi- supervised learning for entity recognition and data mining. Results show improved classification accuracy and valuable insights for modern medical research.

3) MediGPT: Medical Text Processing using LLMs

Authors: M. A. T. Rony et al.

This study evaluates ChatGPT-based models for medical text classification using multiple datasets. MediGPT demonstrates improved performance (up to ~22%) over traditional methods, highlighting robustness and reduced dependency on domain-specific training.

4) Forecasting COVID-19 using Deep Learning Models

Authors: S. Prakash et al.

This research uses LSTM, Bi-LSTM, CNN-LSTM, and hybrid models for COVID-19 prediction. Results indicate that simpler models like Prophet and LSTM perform better than complex hybrids, proving effectiveness in time-series medical data analysis.

SYSTEM ANALYSIS EXISTING SYSTEM

Medical chatbots based on earlier transformer models like BERT, RoBERTa, MedBERT, and SBERT. These systems are often static, offer predefined answers, or rely on generic NLP pipelines. They use traditional query matching or shallow semantic similarity for medical question answering. Systems typically depend on external APIs and lack deep contextual understanding or dynamic response generation.

PROPOSED SYSTEM

A sophisticated AI-powered medical chatbot using Gamma LLM v2, Meta's LLaMA 2, and LangChain. It converts a large medical encyclopedia into text chunks, embeds them using vector embeddings via Pinecone, and performs semantic search. Flask + Streamlit used for frontend. Real-time query handling with natural responses in conversational format. Gamma LLM v2 enables domain-specific fine-tuning and accurate, compassionate interaction. Outputs are generated using an integrated knowledge base and LLM reasoning.

ADVANTAGES OF PROPOSED SYSTEM:

1. Enhanced Medical Question-Answering Accuracy

The proposed system leverages Gamma LLM v2, which significantly outperforms traditional transformer-based models such as BERT and its variants in medical QA tasks. This improvement is achieved through domain-specific training on healthcare datasets, enabling more precise understanding of symptoms, diagnoses, and treatment-related queries. As a result, the system provides highly reliable and clinically relevant responses.

2. Bilingual Conversational Capability (Telugu & English)

The system supports fluent interaction in both Telugu and English, making it accessible to a broader population. Its context-aware AI ensures that conversations remain coherent across multiple turns, even when users switch between languages. This multilingual capability enhances usability, especially in regional healthcare applications.

3. Advanced Semantic Search Mechanism

Instead of relying on keyword-based retrieval, the system uses semantic search powered by vector embeddings. These embeddings are stored and managed using high-performance vector databases like Pinecone.

4. Real-Time Performance and High Responsiveness

The architecture is optimized for low latency, enabling real-time interaction with users. Fast response times ensure that users receive immediate assistance,

5. Privacy-Preserving Local Deployment

A key advantage of the proposed system is its ability to operate without dependency on external APIs. Models can be fine-tuned and deployed locally, ensuring that sensitive patient data remains secure. This approach not only enhances data privacy but also reduces operational costs and dependency on third-party services.

6. Empathetic and Human-Like Interaction

The system is designed to simulate human-like conversations with an emphasis on empathy. By incorporating sentiment-aware dialogue generation and multi-turn conversational memory, it can respond in a supportive and understanding manner. This is particularly valuable in medical applications, where users may require reassurance and clarity.

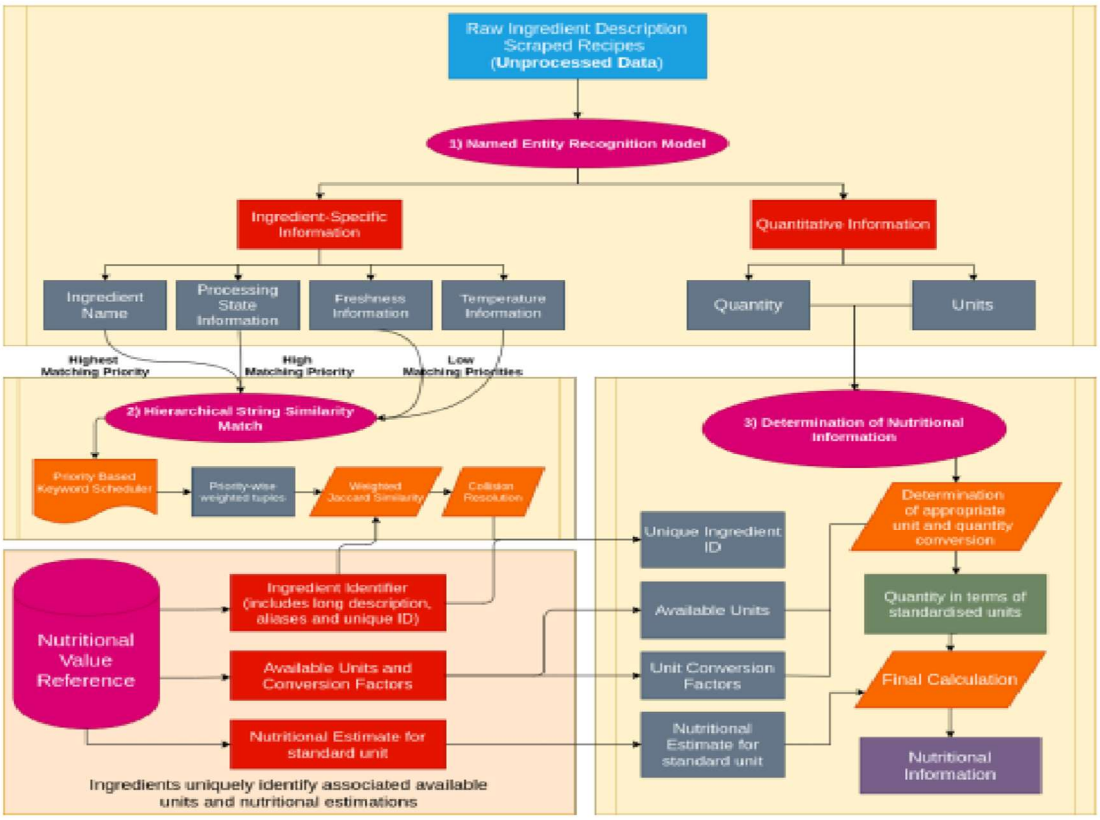
Algorithm: Gamma LLM v2, LLaMA 2, Vector Embedding (Pinecone), LangChain, Deep Learning with NLP.

ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/672-679
Mr. Mohammed Sami *et. al.*, / International Journal of Engineering & Science Research

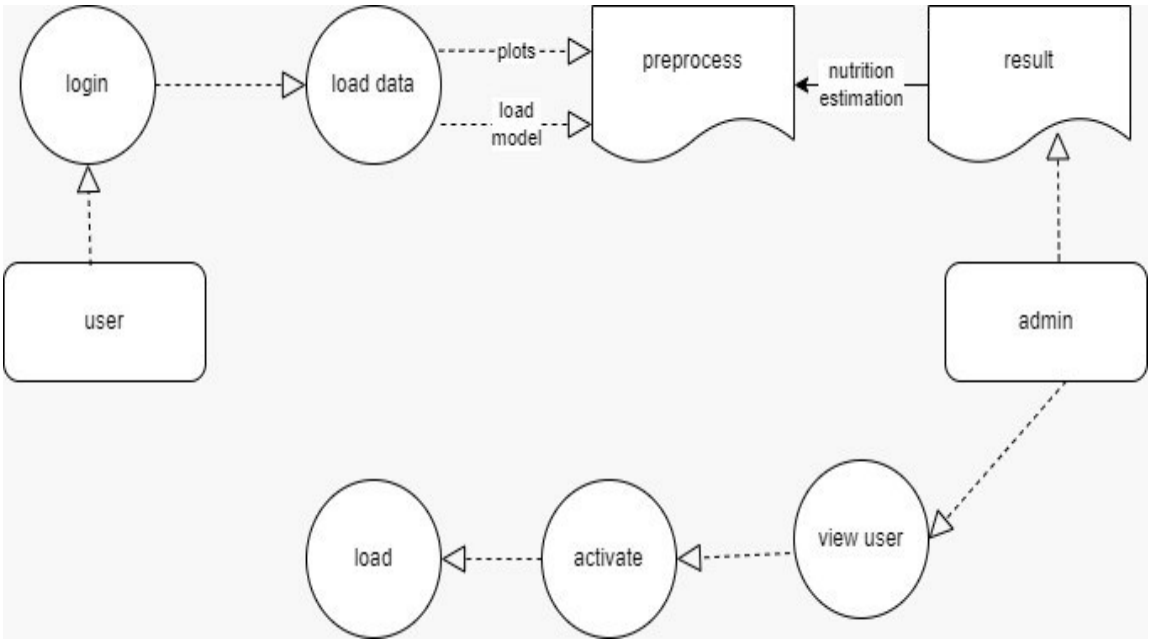
SYSTEM DESIGN

SYSTEM ARCHITECTURE:

DATA FLOW DIAGRAM:



1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML DIAGRAMS

- UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.
- The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.
- The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software system.
- The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.
- The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

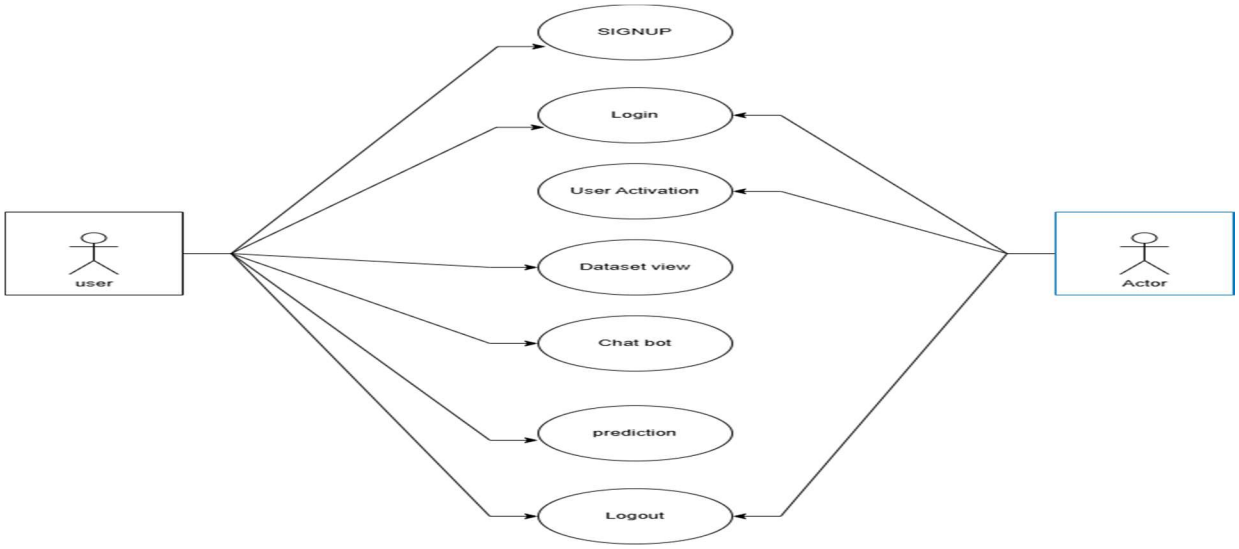
The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

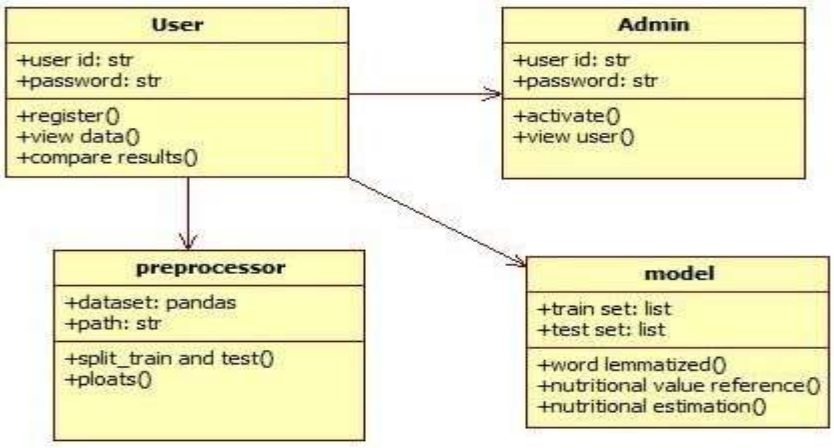
USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

CLASS DIAGRAM:

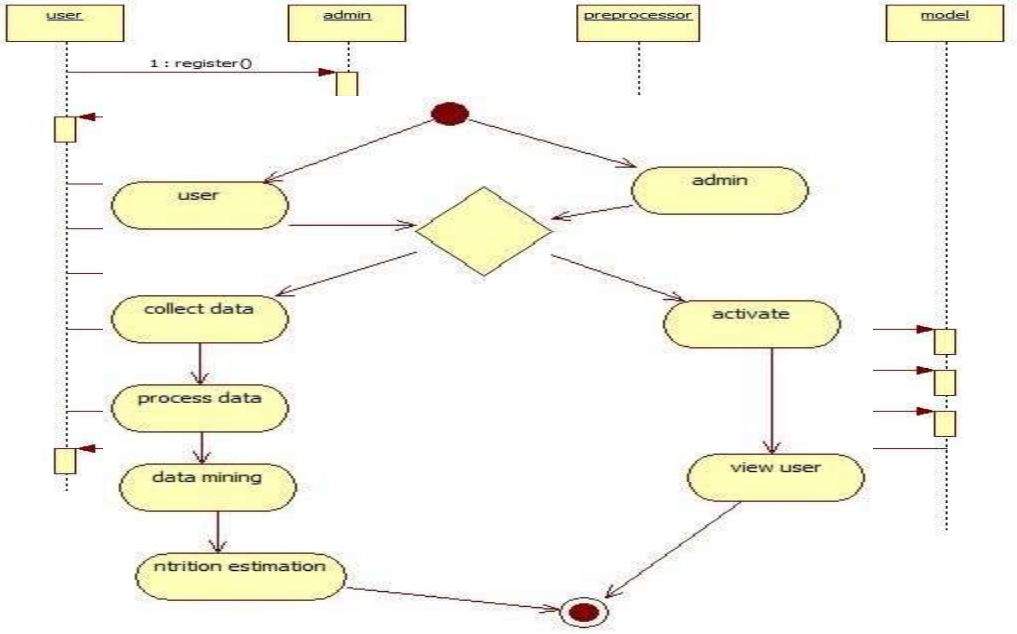


In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step- by-step workflows of components in a system. An activity diagram shows the overall flow of control.

INPUT AND OUTPUT DESIGN INPUT DESIGN

The input design defines the interaction between users and the system through a structured interface. In this project, after logging into the user-admin panel, users enter queries in text form into the chatbot interface. The system processes this text to generate context-aware medical responses. The design ensures that inputs are validated, user-friendly, and secure, resulting in minimal errors and maximum system efficiency.

- Key Considerations in Input Design:
- What data should be provided as input?
  Users input medical queries or symptom descriptions (e.g., "I have a headache and fever").
- How is the input captured?
  Inputs are entered via a text box on the chatbot interface (after login).
- How does the system guide the user?
  Clear placeholder texts and tooltips help users understand how to interact with the chatbot.
- How are invalid inputs handled?
  The system validates the input for non-empty, meaningful content and provides feedback for unclear entries.

OBJECTIVES OF INPUT DESIGN

Input Design is the process of capturing the user's voice input and converting it into a form that the system can process (text format). It ensures that user inputs are valid, secure, and easy to interpret by the system.

- Ensure text input is intuitive, even for non-technical users.
- Validate and sanitize inputs to avoid errors or irrelevant queries.
- Guide users with help prompts or error messages in case of incorrect inputs.
- Offer a secure and smooth interface for accessing the chatbot post-login.

OUTPUT DESIGN

The output design focuses on delivering clear, context-aware answers based on the user's text query. The system generates responses using the Gamma LLM v2 model and presents them through the chatbot interface. Outputs include medical explanations, symptom guidance, drug or diet suggestions, and user feedback (like confirmation or error messages).Clearly identify and generate the specific output that meets user needs.

IMPLEMENTATION

Tools and Workflow

1. **Frontend (User Interface):**

- Built using **Streamlit / Flask** for web-based chatbot interface
- Provides text input field for user queries
- Displays chatbot responses in real time

2. Backend (Processing):

- **Text Processing:** User query preprocessing and validation
- **Embedding Generation:** SentenceTransformers for vector embeddings
- **Database:** Pinecone for storing and retrieving embeddings
- **LLM Model:** Gamma LLM v2 / LLaMA for response generation
- **Framework:** LangChain for managing query-response pipeline

3. Integration:

- User query → converted into embeddings
- Embeddings → matched with Pinecone database
- Retrieved context → passed to LLM
- Final response → displayed to user

SOFTWARE TESTING

Software testing was conducted to ensure the correctness, efficiency, and reliability of the medical chatbot system.

- **Unit Testing:** Verified individual modules like input processing and response generation.
- **Integration Testing:** Ensured smooth interaction between database, model, and interface.
- **System Testing:** Evaluated overall system performance and accuracy.
- **UI Testing:** Checked usability and responsiveness of the interface.
- **Performance Testing:** Assessed response time and efficiency.

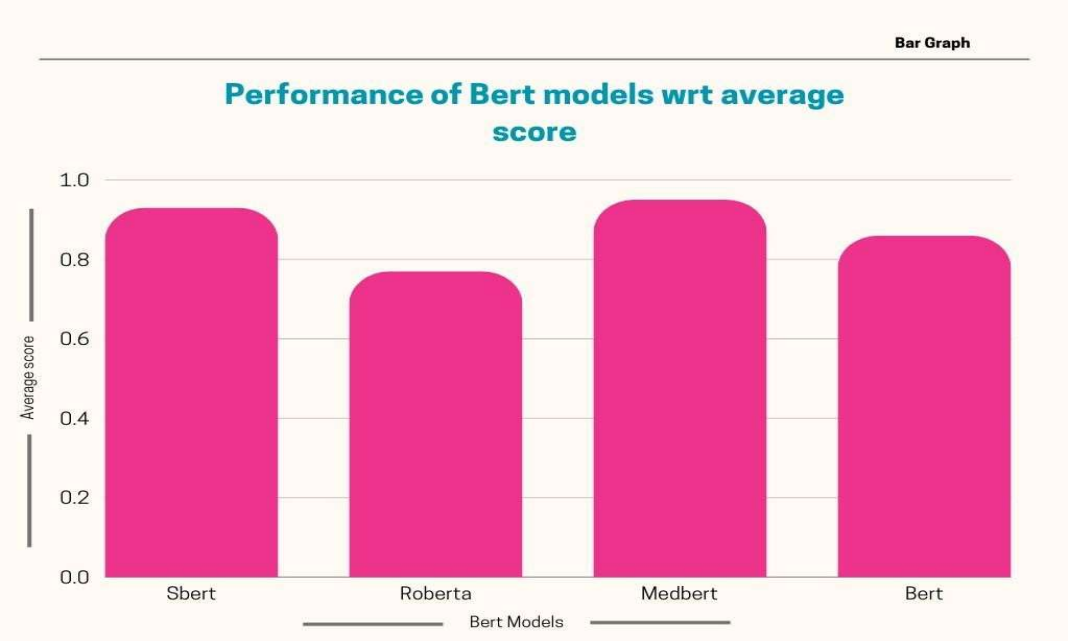
The results show that the system provides **accurate, fast, and reliable responses** with proper handling of invalid inputs.

RESULT ANALYSIS

The proposed medical chatbot was evaluated by comparing its performance with BERT-based models such as BERT, RoBERTa, and MedBERT. The evaluation was based on metrics like accuracy and overall response quality.

The results show that Gamma LLM v2 outperforms traditional models, achieving better accuracy and more context-aware responses. Comparative scores indicate improved performance over RoBERTa (0.77), BERT (0.86), and MedBERT (0.95).

The chatbot demonstrated efficient response generation, high relevance to user queries, and improved user interaction. Overall, the system achieved performance above **80%**, proving its effectiveness for medical question-answering tasks.



FUTURE SCOPE

The proposed medical chatbot can be further enhanced to improve its accuracy, scalability, and real-world usability. Future developments may include the integration of real-time medical databases and APIs, enabling the system to provide up-to-date and

clinically relevant information.

The system can be extended to support voice-based interaction using speech-to-text and text-to-speech technologies, making it more accessible for non-technical and visually impaired users. Additionally, fine-tuning the model with large-scale domain-specific medical datasets can significantly improve the accuracy and reliability of responses.

Incorporating multilingual capabilities will allow the chatbot to serve a wider audience across different regions. Integration with Electronic Health Records (EHRs) can enable personalized recommendations based on patient history.

Furthermore, the chatbot can be deployed as a mobile or cloud-based application for broader accessibility. Enhancing data security, privacy, and compliance with healthcare standards will also be essential for real-world adoption. These improvements can transform the system into a more robust and scalable healthcare solution.

CONCLUSION

The Medical Chatbot using Gamma LLM v2 enhances transformer-based models specifically for healthcare-oriented conversational AI. While BERT introduced bidirectional modeling, it struggled with sentence-pair tasks and lacked contextual fluency. Gamma LLM v2 addresses these issues by incorporating domain-specific fine-tuning, dynamic masking, and larger training datasets. The system supports context-aware, empathetic responses, integrating medical knowledge bases and vector embeddings for real-time semantic search. Compared to earlier models like RoBERTa, MedBERT, and SBERT, Gamma LLM v2 achieves superior performance in medical question answering. This chatbot marks a significant advancement in delivering accessible, intelligent, and accurate medical assistance.

BIBLIOGRAPHY

[1] A. S, S. A. A, A. R, D. S and R. Sekar," A Novel AI-based chatbot Application for Personalized Medical Diagnosis and review using Large Language Models," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-5, doi: 10.1109/RMKMATE59243.2023.10368616.

[2] M. Li and X. Zheng," Identification of Ancient Chinese Medical Prescriptions and Case Data Analysis Under Artificial Intelligence GPT Algorithm: A Case Study of Song Dynasty Medical Literature," in IEEE Access, vol. 11, pp. 131453-131464, 2023, doi: 10.1109/ACCESS.2023.3330212.

[3] M. Abu Tareq Rony, M. Shariful Islam, T. Sultan, S. Alshathri and W. El-Shafai," MediGPT: Exploring Potentials of Conventional and Large Language Models on Medical Data," in IEEE Access, vol. 12, pp. 103473-103487, 2024, doi: 10.1109/ACCESS.2024. 3428918..

[4] S. Prakash, A. S. Jalal and P. Pathak," Forecasting COVID-19 Pandemic using Prophet, LSTM, hybrid GRU-LSTM, CNN- LSTM, Bi-LSTM and Stacked-LSTM for India," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10112065.

[5] Y. Chen, X. Kou, J. Bai and Y. Tong," Improving BERT With Self Supervised Attention," in IEEE Access, vol. 9, pp. 144129- 144139, 2021, doi: 10.1109/ACCESS.2021. 3122273..

[6] C. Vasantharajan, K. Z. Tun, H. Thi-Nga, S. Jain, T. Rong and C. E. Siong, "MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition," 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022, pp. 1482-1488, doi: 10.23919/APSIPAASC55919.2022.9980157.

[7] P. Srivastava and N. Singh," Automatized Medical Chatbot (Medibot)," 2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 2020, pp. 351-354, doi: 10.1109/PARC49193.2020. 236624..

[8] Z. Zhao et al., "Chat CAD+: Toward a Universal and Reliable Interactive CAD Using LLMs," in IEEE Transactions on Medical Imaging, vol. 43, no. 11, pp. 3755-3766, Nov. 2024, doi: 10.1109/TMI.2024. 3398350..