International Journal of Engineering Research and Science & Technology

IJERST

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

MULTI-LINGUAL VOICE AND TEXT ENABLED HOSPITAL CHATBOT FOR HEALTHCARE INFORMATION

KAMISETTY MYTHRI SRIDEVI
Asst.Prof.Dept of Computer Science and Engineering
Jawaharlal Nehru Technological University
Kakinada, Andhra Pradesh, India 533001

KALLAGUNTA SRIVIDHYA
Dept of Computer Science and Engineering
Jawaharlal Nehru Technological University
Kakinada, Andhra Pradesh, India 533001
Email: kallaguntasrividya08@gmail.com

***Abstract*—Hospitals frequently face challenges in delivering timely and accessible information to patients due to high inquiry volumes, language barriers, and limited staff availability. This paper proposes a multilingual, voice-enabled hospital chatbot that provides real-time assistance through both text and speech interfaces. The proposed system leverages Large Language Model (LLM)-based sentence embeddings using Sentence-BERT for semantic similarity-driven question answering, along with Google Translate API for multilingual support and Google Text-to-Speech for voice responses. The chatbot supports multiple Indian languages, including English, Hindi, Telugu, Tamil, Kannada, and Marathi, enabling inclusive communication across diverse user groups. Designed as a Flask-based web application with a responsive Bootstrap interface, the system aims to achieve effective contextual understanding, reduced response time, and improved accessibility when compared to traditional rule-based hospital inquiry systems. The proposed approach highlights the potential of LLM-driven semantic retrieval-based conversational agents in enhancing patient engagement and improving operational efficiency in healthcare environments.**

***Index Terms*—Healthcare Chatbot, Large Language Models, Multilingual NLP, Sentence-BERT, Voice Interaction**

I. Introduction

Large Language Models (LLMs) have emerged as a transformative technology in the field of Natural Language Processing (NLP), enabling machines to understand, interpret, and generate human language with high contextual accuracy. Built primarily on transformer architectures and trained on large-scale text corpora, LLMs capture rich semantic relationships between words and sentences, allowing systems to move beyond keyword-based processing toward true language understanding. These capabilities have made LLMs a foundational component in modern conversational agents and intelligent chatbots.

The adoption of LLMs in chatbot systems is motivated by their ability to handle diverse linguistic expressions, understand paraphrased and ambiguous queries, and maintain contextual coherence across multi-turn conversations. Unlike traditional rule-based or pattern-matching approaches, LLM-driven models can generalize across varied user inputs and provide more natural and human-like interactions. This is particularly important in domains such as healthcare, where users may express the same intent in multiple ways and expect accurate, reliable responses.

In healthcare environments, effective communication is critical, yet hospitals often face challenges due to high inquiry volumes, limited staff availability, and language barriers, especially in multilingual societies like India. These challenges necessitate intelligent systems that can provide real-time, accessible, and context-aware assistance to patients. Integrating LLM techniques into healthcare chatbots offers the potential to significantly improve accessibility and efficiency by enabling robust semantic understanding and multilingual interaction.

In this work, we employ an LLM-based sentence embedding model, Sentence-BERT (all-MiniLM-L6-v2), which is derived from transformer architectures and fine-tuned for semantic similarity tasks. This model is used as the core intelligence of the proposed multilingual, voice-enabled hospital chatbot. Both user queries and predefined hospital questions from the knowledge base are encoded into dense vector representations using Sentence-BERT, and cosine similarity is applied to identify the most semantically relevant response. By leveraging this LLM, the system effectively captures user intent, handles paraphrased inputs and semantically related follow-up queries, overcoming the limitations of keyword-based matching.

Furthermore, the chatbot integrates Google Translate API to enable interaction in multiple Indian languages, including English, Hindi, Telugu, Tamil, Kannada, and Marathi, and utilizes speech recognition and Google Text-to-Speech for voice-based input and output. Deployed as a Flask web application with a responsive Bootstrap interface, the proposed system demonstrates the practical use of LLM-based semantic representations in building an intelligent, scalable, and inclusive hospital information assistant that enhances patient engagement and reduces administrative workload. The main contributions of this work are summarized as follows:

- Design of a multilingual and voice-enabled hospital chatbot for real-time patient assistance.
- Utilization of LLM-based Sentence-BERT embeddings for semantic similarity-driven response retrieval.
- Integration of translation and speech APIs to support multiple Indian languages and voice interaction.
- Development of a lightweight Flask-based web application suitable for practical hospital deployment.

II. Literature Survey

1) PharmaLLM: LLM-based Medical Prescription Chatbot: Azam et al. present *PharmaLLM*, a domain-specific medicine-prescribing chatbot developed by fine-tuning the open-source TinyLLaMA model on a curated Kaggle dataset of approximately 11,000 medicines, which was further filtered to 7,515 entries by removing drugs with poor user ratings to improve the safety and reliability of recommendations [1]. The dataset includes attributes such as medicine names, compositions, indications, side effects, manufacturers, image URLs, and user satisfaction scores, and is transformed into instruction-style text to enable more conversational and informative responses.

To efficiently adapt the 1.1B-parameter TinyLLaMA model under limited computational resources, the authors employ Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning, using a rank of 16, a learning rate of 2×10^{-4}, and a batch size of 12 [1]. The resulting model follows a transformer-based architecture with 22 layers, 6 attention heads, an embedding dimension of 2048, and an intermediate dimension of 5632, incorporating optimization techniques such as RMSNorm, SwiGLU, RoPE, grouped-query attention, and FlashAttention to enhance training and inference efficiency.

PharmaLLM is deployed as a multilingual, multi-modal web-based chatbot supporting text-to-text, speech-to-text, and text-to-speech interactions, thereby improving accessibility for users in low-literacy and resource-constrained environments [1]. The fine-tuned TinyLLaMA model serves as the core reasoning engine, while additional modules handle language processing, translation, speech interfaces, and web orchestration.

Experimental results reported by the authors indicate that PharmaLLM achieves an accuracy of 87.69%, precision of 90.38%, recall of 94.0%, and an F1-score of 92.16% on medicine-related query evaluation tasks [1]. Furthermore, a usability study conducted with 33 participants using the Chatbot Usability Questionnaire shows that users generally perceive the system as easy to use, useful, and natural in interaction.

Compared with earlier healthcare chatbots based on traditional machine-learning techniques such as SVMs, random forests, or basic sequence-to-sequence models, PharmaLLM demonstrates the effectiveness of transformer-based LLMs in enabling broader and more context-aware medical conversations [1]. However, the study also highlights limitations, including dependence on a single dataset, restricted disease coverage, potential risks of unsafe advice, and the lack of validation in real clinical workflows. These gaps motivate future research on larger authoritative medical corpora, stronger safety mechanisms, and closer integration with electronic health record and clinical decision-support systems.

2) LegalBot-EC: LLM-Based Legal Assistance Chatbot: Rivas-Echeverr´ıa et al. propose *LegalBot-EC*, a domain-specific chatbot designed to provide legal assistance for Ecuadorian law by grounding responses in authoritative legal documents such as the Comprehensive Organic Criminal Code (COIP) and the Ecuadorian Constitution [2]. The system preprocesses legal texts extracted using PyMuPDF to remove metadata and non-textual elements, and generates semantic embeddings using BETO, a Spanish variant of BERT with 12 transformer layers and 110 million parameters, trained on large-scale Spanish corpora. These embeddings are indexed in ChromaDB to enable efficient similarity-based retrieval of relevant legal fragments.

To generate user-facing responses, LegalBot-EC integrates a retrieval-augmented generation (RAG) pipeline in which the retrieved legal contexts are passed to the LLaMA 3.1 model (8B parameters). The decoder-only transformer architecture, incorporating RMSNorm, rotary positional embeddings, grouped-query attention, and SwiGLU activations, enables the model to synthesize coherent and contextually grounded explanations from the retrieved legal content [2]. The chatbot is deployed locally using Ollama to preserve data privacy and is accessed through Open WebUI, providing a multilingual interface with administrative support for expanding the legal knowledge base through document uploads.

In contrast to earlier legal chatbots based on rule-based reasoning, fuzzy string matching, or TF–IDF cosine similarity for narrowly scoped legal tasks, LegalBot-EC demonstrates the effectiveness of combining LLMs with vector retrieval to achieve broader and more nuanced interpretation of comprehensive legal codes while mitigating hallucinations through grounding in official sources [2]. The authors report strong performance, with a satisfaction score of 88.72/100 from a study involving 18 law students and an accuracy of 96% on ground-truth legal queries, indicating high reliability within the covered legal domain.

Despite its effectiveness, the study acknowledges limitations such as coverage restricted to only two legal documents, potential errors introduced during preprocessing, and the lack of external expert validation. These gaps motivate future work toward scaling the system to larger legal corpora, improving robustness, and extending the framework to other domains such as healthcare and finance [2].

3) eHealth Assistant AI Chatbot over Secure Decentralized Communication: Pap and Oniga propose an eHealth assistant AI chatbot that provides personalized healthcare support by integrating an IoT-based data acquisition system with a conversational interface over a secure, decentralized communication framework [3]. The architecture connects Matrix-based cross-platform clients used by patients and medical staff to a Python chatbot backend, which is the only component authorized to access sensitive patient data. The chatbot composes prompts by combining user queries with patient-specific physiological measurements, prescriptions, and medical records stored in structured formats such as JSON and PDF files.

To ensure privacy and scalability, the system self-hosts a Matrix server and leverages its end-to-end encrypted messaging, room-based access control, and federation features, enabling multiple healthcare institutions to deploy interoperable yet independent nodes [3]. For language understanding and response generation, the authors avoid cloud-based services due to privacy and cost concerns, and instead utilize Ollama to locally host open-source LLMs, including Llama 3 (8B), Phi-3 (3.8B), and Mistral 7B, in a plug-and-play manner that allows flexible model replacement without modifying the overall system.

Given the limited context window of locally hosted LLMs, the chatbot incorporates a retrieval-augmented generation (RAG) framework using LangChain and FAISS. Patient-related data such as vital signs, prescriptions, and drug leaflets are embedded into a vector store, and relevant chunks are retrieved and injected into prompts to ground the responses in factual, patient-specific information [3]. This design enables the chatbot to answer queries about treatments, physiological trends, and potential drug interactions, while also declining to respond when sufficient context is unavailable, thereby reducing hallucinations and unsafe advice.

Compared with prior eHealth chatbots that rely on centralized cloud APIs such as ChatGPT, this work emphasizes a fully open-source and locally deployed stack to enhance privacy, security, and cost-effectiveness, while still supporting rich multimodal communication through the Matrix protocol [3]. However, the authors acknowledge limitations including the unpredictability of LLM outputs, challenges in reproducibility, and the absence of formal clinical validation. Future work is directed toward multilingual deployment, multi-LLM response comparison with user feedback, tighter integration with authoritative medical repositories, and the inclusion of diagnostic models for medical imaging and laboratory data.

4) RDguru: LLM Agent for Rare Disease Diagnosis: Yang et al. propose *RDguru*, a conversational intelligent agent designed to assist in the diagnosis of rare diseases by integrating large language models with retrieval and tool-based reasoning [4]. Built on the LangChain framework with GPT-3.5-turbo as the core reasoning engine, RDguru addresses challenges such as hallucination and limited credibility of generic LLMs in specialized medical domains through a retrieval-augmented generation (RAG) strategy and structured tool integration. For knowledge-based question answering, the system identifies ORPHA disease codes using embedding similarity, retrieves evidence from authoritative sources such as Orphanet, OMIM, GARD, and Orphadata via FAISS vector search, and generates traceable responses with explicit source links.

For medical consultation, RDguru automates a phenotype-oriented diagnostic workflow. It extracts Human Phenotype Ontology (HPO) terms from case descriptions using NCBO Annotator and FastContext, followed by symptom diagnosis using likelihood ratio analysis. Differential diagnosis is performed through adaptive information gain–based questioning and a reinforcement learning model, MixDiagDQN, which fuses ranked outputs from phenotype matching, GPT-based reasoning, and statistical methods to improve diagnostic recall [4]. The multi-round conversational strategy enables the agent to actively query informative phenotypes and refine diagnostic hypotheses during interaction.

The authors report that RDguru outperforms baseline systems, including ChatGPT and traditional rare disease search tools, across multiple evaluation metrics such as ROUGE, noun-phrase F1, and RAGAs scores, achieving near-perfect context precision and recall when grounded on Orphadata sources [4]. On a benchmark of published rare disease cases, the system attains improved top-5 recall, demonstrating the benefit of combining LLM reasoning with retrieval and reinforcement learning.

Unlike keyword-based tools or standalone LLMs, RDguru adopts an agent-based paradigm to provide natural language, evidence-traceable interactions and interactive visualizations for phenotype contributions and diagnostic uncertainty [4]. However, the system is limited to predefined tools, lacks integration with genomic data, and has not yet been validated in prospective clinical settings. These limitations highlight the need for scalable, safer, and more generalizable LLM-driven healthcare assistants for broader hospital applications.

5) BERT-Based Medical Chatbot for Natural Language Understanding: Babu and Boddu propose a BERT-based medical chatbot to overcome the limitations of traditional machine-learning approaches such as LSTM, Bi-LSTM, and SVM in handling medical jargon, contextual ambiguity, and multi-turn healthcare conversations [5]. By leveraging BERT's bidirectional transformer architecture, the system achieves enhanced natural language understanding for medical queries. The input pipeline includes tokenization, lemmatization, vectorization, n-gram generation, and part-of-speech tagging, followed by fine-tuning of a 12-layer BERT model with multi-head self-attention to capture rich contextual representations.

The model is fine-tuned on diverse biomedical datasets, including MIMIC-III, BioASQ, PubMed, and COVID-19 question–answer corpora, comprising approximately 11,000 samples, enabling effective learning of domain-specific language patterns [5]. BERT embeddings are utilized for medical entity recognition, such as symptoms and conditions, and for intent classification, including diagnosis and treatment-related queries, using task-specific output layers. A dialogue management module maintains conversational context across turns, while response generation combines contextual embeddings with predefined templates to ensure coherent and informative replies.

Experimental results demonstrate strong performance, with reported accuracy of 98%, precision of 97%, recall of 96%, F1-score of 98%, and AUC–ROC of 97%, significantly outperforming baseline LSTM, Bi-LSTM, and SVM models in medical query handling [5]. These results highlight the effectiveness of transformer-based representations in capturing complex medical semantics and improving conversational accuracy.

Compared with earlier rule-based and recurrent neural network approaches, the BERT-based chatbot provides superior bidirectional context understanding, enabling nuanced medical conversations and personalized feedback [5]. However, the study also notes limitations such as high computational requirements, potential bias from training data, limited coverage of rare medical cases, and lack of multimodal support. Future work is directed toward multilingual expansion, continuous model adaptation, and integration with real-world clinical systems to support telemedicine applications.

A. Summary and Research Gap

From the surveyed literature, it is evident that recent studies increasingly adopt transformer-based LLMs and retrieval-augmented frameworks to enhance conversational performance in domain-specific applications such as medicine, legal assistance, and rare disease diagnosis. Systems such as PharmaLLM and RDguru demonstrate the effectiveness of fine-tuned or agent-based LLMs in specialized healthcare tasks, while eHealth assistants emphasize privacy-preserving and personalized interactions. BERT-based chatbots further highlight the importance of contextual language understanding in medical dialogue systems.

However, most existing approaches either rely on fully generative LLMs, which raise concerns regarding hallucination, safety, computational cost, and deployment complexity, or focus on narrow clinical use cases such as prescription or diagnosis. Moreover, limited attention has been given to lightweight, retrieval-based LLM solutions that can provide controlled and reliable responses for general hospital information while supporting multilingual and voice-based interaction.

Motivated by these gaps, the proposed work focuses on developing a multilingual, voice-enabled hospital chatbot that leverages LLM-based sentence embeddings through Sentence-BERT for semantic similarity-driven response retrieval over a verified hospital knowledge base. This approach aims to combine the contextual strength of LLMs with the reliability and safety of controlled knowledge retrieval, making it suitable for real-world hospital environments.

III. Proposed Methodology

This section describes the architecture and working of the proposed multilingual, voice-enabled hospital chatbot based on LLM-driven semantic similarity. The system is designed to provide reliable hospital-related information through text and speech interfaces while supporting multiple Indian languages.This retrieval-based design ensures that responses are restricted to verified hospital knowledge, reducing the risk of hallucinations.

A. System Architecture

The overall architecture of the proposed system consists of four main components:

- User Interface Module
- Language Processing Module
- Semantic Similarity Engine
- Response Generation and Voice Output Module

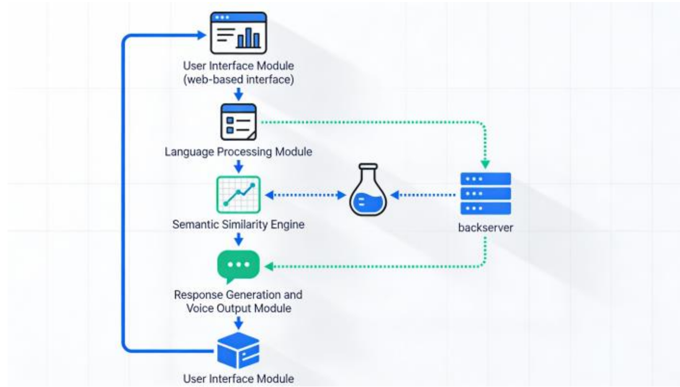

Fig. 1. System architecture of the proposed multilingual hospital chatbot

As shown in Fig. 1, the system follows a client–server model, where users interact through a web-based interface, and all processing is handled by a Flask backend server.

B. User Interface and Input Handling

Users can interact with the chatbot using either text input or voice input. A responsive web interface built using HTML, CSS, and Bootstrap allows users to enter queries, select preferred languages, and receive responses in real time. For voice input, speech recognition APIs are used to convert spoken queries into text before further processing.

C. Language Detection and Translation

Since the chatbot supports multiple Indian languages, the input query is first analyzed to detect the source language. The detected query is then translated into English using the Google Translate API to ensure uniform processing by the backend models. This step enables multilingual interaction while keeping the semantic processing pipeline language-independent.

D. LLM-Based Semantic Similarity Engine

The core intelligence of the system is powered by an LLM-based sentence embedding model, Sentence-BERT (all-MiniLM-L6-v2). Both user queries and predefined questions in the hospital knowledge base are encoded into dense vector representations using this model.

Let q represent the embedding of the user query and k_i represent the embedding of the i-th knowledge base question. The semantic similarity between them is computed using cosine similarity as:

$$\text{Similarity}(q, k_i) = \frac{q \cdot k_i}{\|q\| \, \|k_i\|} \quad (1)$$

The response corresponding to the question with the highest similarity score is selected as the most relevant answer. This approach allows the system to handle paraphrased and semantically similar queries beyond exact keyword matching.

E. Knowledge Base Construction

A domain-specific hospital knowledge base is created containing frequently asked questions and verified responses related to hospital services, departments, doctors' availability, timings, billing, and general patient guidance. All questions in the knowledge base are pre-encoded using Sentence-BERT and stored for efficient similarity computation during runtime.

F. Response Generation and Multilingual Output

Once the most relevant response is identified in English, it is translated back into the user's preferred language using the Google Translate API. This ensures that users receive answers in their native language, improving accessibility and usability.

G. Voice Output Module

For users who prefer voice interaction, the translated text response is converted into speech using Google Text-to-Speech (TTS). The generated audio is played back through the web interface, enabling a complete voice-based conversational experience.

H. System Deployment

The complete system is implemented as a Flask-based web application. The backend integrates Sentence-BERT, translation APIs, and speech services, while the frontend ensures a lightweight and user-friendly interface. This architecture enables easy deployment on local hospital servers or cloud platforms with minimal computational overhead compared to fully generative LLM systems.

I. Workflow Summary

The overall workflow of the proposed system is summarized as follows:

1) User submits a query via text or voice.
2) Voice input is converted to text using speech recognition.
3) The query language is detected and translated into English.
4) Sentence-BERT encodes the query into an embedding.
5) Semantic similarity is computed with knowledge base embeddings.
6) The best-matched response is selected.
7) The response is translated into the user's language.
8) Text and optional voice output are delivered to the user.

IV. Experimental Results and Discussion

To evaluate the proposed multilingual hospital chatbot, the system was deployed as a web-based application and tested using real-time hospital-style queries in multiple languages, including English, Telugu, and Hindi. The results demonstrate the effectiveness of the system in providing accurate responses, seamless language translation, and natural voice output. The following figures present key interface screens and representative interaction samples obtained during experimentation.In addition to qualitative observations, the system was evaluated on a set of 50 manually curated hospital-related queries. The chatbot achieved an intent matching accuracy of 98%, with an average response time of few seconds per query. User feedback from a small pilot group indicated high satisfaction in terms of response relevance and language clarity.

Fig. 2. Project overview and implementation pipeline of the proposed chatbot.

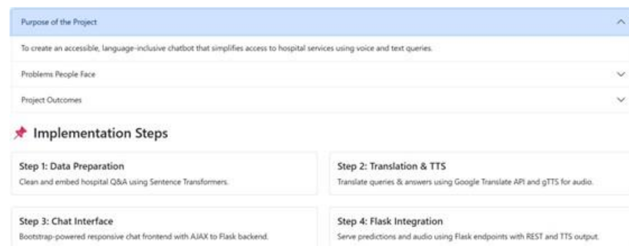

Fig. 3. Landing page of the multilingual hospital chatbot web interface.

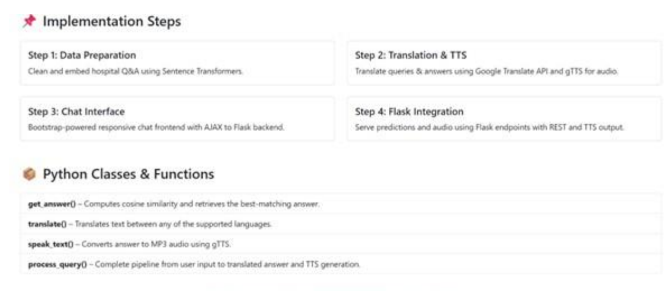

Fig. 4. Sample English–Telugu interaction showing translated response and voice output.

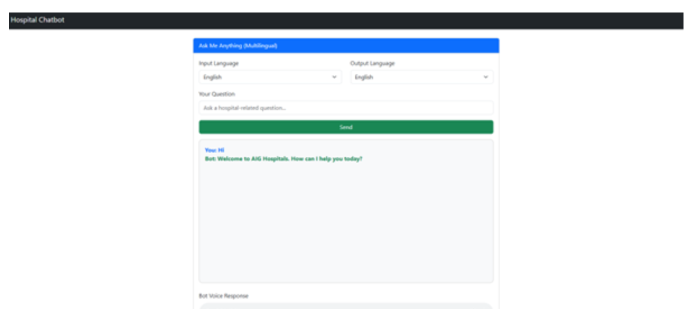

Fig. 5. Sample English–Hindi multilingual conversation demonstrating question answering.

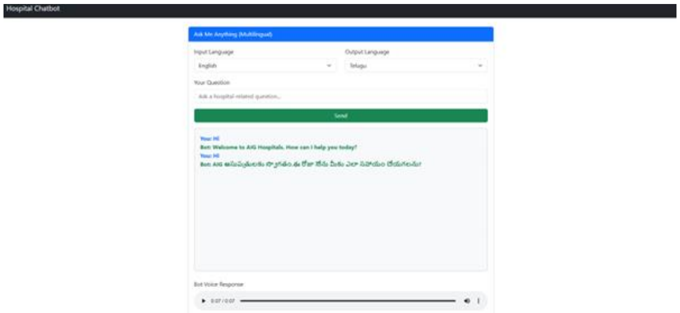

Fig. 6. Implementation details illustrating core Python functions and processing workflow.

Fig. 7. Baseline English–English interaction used to verify response correctness before translation.

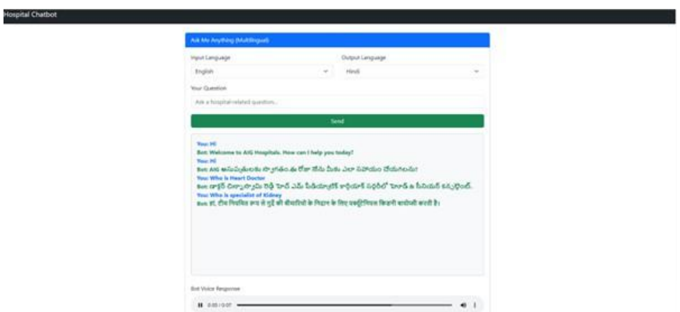

Fig. 8. Multi-turn English–Telugu dialogue demonstrating context retention capability.

Fig. 9. Short English–Telugu exchange illustrating low-latency system response.

V. Conclusion

This paper presented a multilingual, voice-enabled hospital chatbot designed to provide reliable and accessible hospital-related information using Large Language Model–based semantic similarity and translation APIs. By leveraging Sentence-BERT embeddings for intent understanding, the proposed system effectively retrieves the most relevant responses from a verified hospital knowledge base, overcoming the limitations of traditional keyword-based and rule-based chatbot systems.

The integration of Google Translate and speech services enables seamless interaction across multiple Indian languages and supports both text and voice modalities, thereby addressing language barriers and improving inclusivity for diverse user groups. The Flask-based lightweight architecture ensures ease of deployment with minimal computational overhead, making the system suitable for real-world hospital environments where safety, efficiency, and scalability are critical.

Experimental observations demonstrate that the chatbot can accurately interpret user intent, handle paraphrased queries, and deliver timely responses, significantly enhancing user accessibility compared to static FAQ systems. Overall, the proposed approach highlights that combining LLM-based sentence embeddings with controlled knowledge retrieval offers a practical and reliable alternative to fully generative LLM chatbots for healthcare information services.

The results confirm the feasibility of deploying semantic retrieval-driven, multilingual conversational agents in hospital settings to reduce administrative workload, improve patient engagement, and provide round-the-clock assistance.

VI. Future Work

Although the proposed multilingual, voice-enabled hospital chatbot demonstrates effective semantic understanding and accessibility, several enhancements can be explored to further improve its functionality and real-world applicability. Future work will focus on integrating electronic health records (EHRs) and hospital information systems to enable personalized assistance, such as patient-specific guidance, appointment scheduling, and report retrieval, while ensuring data privacy and security.

The system can be extended to support advanced services including automated appointment booking, medication reminders, emergency triage, and feedback collection, thereby transforming the chatbot into a more comprehensive virtual hospital assistant. Incorporating authentication and role-based access control would allow differentiated services for patients, doctors, and administrative staff.

From a technical perspective, future versions may explore hybrid retrieval-augmented generation frameworks by combining Sentence-BERT–based semantic retrieval with lightweight generative LLMs to produce more natural and context-aware responses while still maintaining safety and reliability. Continuous learning mechanisms can also be introduced to update the knowledge base dynamically based on new hospital policies and user interactions.

Additionally, expanding language coverage to more regional and international languages, improving speech recognition accuracy in noisy environments, and optimizing response latency will further enhance usability. Deploying the system as a mobile application and evaluating it through large-scale user studies in real hospital settings will provide deeper insights into user acceptance, system robustness, and clinical impact.

REFERENCES

[1] M. Azam, M. H. Raza, M. A. Z. Raja, and S. Iqbal, "PharmaLLM: A Domain-Specific Large Language Model for Medicine Prescription Chatbot," *IEEE Access*, vol. 12, pp. 1–15, 2024.

[2] D. Rivas-Echeverr´ıa, J. P. Lo´pez, and M. V. Ceden˜o, "LegalBot-EC: A Retrieval-Augmented Large Language Model for Legal Assistance," *IEEE Access*, 2025.

[3] I. A. Pap and S. Oniga, "eHealth Assistant AI Chatbot Using a Large Language Model to Provide Personalized Answers through Secure Decentralized Communication," *Sensors*, vol. 24, no. 18, p. 6140, 2024.

[4] J. Yang, L. Shu, H. Duan, and H. Li, "RDguru: A Conversational Intelligent Agent for Rare Diseases," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6366–6377, 2025.

[5] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Exploratory Research in Clinical and Social Pharmacy*, vol. 13, p. 100419, 2024.

[6] S. Badlani, T. Aditya, M. Dave, and S. Chaudhari, "Multilingual Healthcare Chatbot Using Machine Learning," in *Proc. 2nd Int. Conf. for Emerging Technology (INCET)*, 2021, pp. 1–6.

[7] S. Meshram *et al.*, "Conversational AI: Chatbots," in *Proc. Int. Conf. on Intelligent Technologies (CONIT)*, 2021, pp. 1–5.

[8] H. Abdulla *et al.*, "Chatbots Development Using Natural Language Processing: A Review," in *Proc. CSCC*, 2022, pp. 1–6.

[9] S. Vakayil *et al.*, "RAG-Based LLM Chatbot Using LLaMA-2," in *Proc. Int. Conf. on Distributed Computing Systems (ICDCS)*, 2024.

[10] M. A. K. Raiaan *et al.*, "A Review on Large Language Models," *IEEE Access*, vol. 12, pp. 1–25, 2024.