Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

IJIKM.org

Volume 20, 2025

BREAKING LANGUAGE BARRIERS IN HEALTHCARE: A VOICE ACTIVATED MULTILINGUAL HEALTH ASSISTANT

Vignesh U*	School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India	vignesh.u@vit.ac.in
Aman Amirneni	School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India	aman.amirneni2022@vitstudent.ac.in

ABSTRACT

Aim/Purpose	The study aims to develop a multilingual healthcare assistance chatbot that provides real-time, accurate answers to a query related to health matters in multiple languages. Conversion of written responses into spoken words lets users have the medical information necessary for them without interrupting communication between patients and health services. The purpose of this system is to break the language barriers for healthcare users, making it easier for them to access vital medical advice and resources.
Background	This research focuses on fine-tuned large language models (LLMs) for providing accurate, context-aware responses in multiple languages with speech-based output. The chatbot, built on pre-trained Hugging Face models and fine-tuned with healthcare datasets, demonstrates a comprehensive understanding of medical terminology, symptoms, and healthcare concepts across languages. Unlike many existing chatbots that offer limited medical knowledge or support only a single language, the proposed chatbot leverages fine-tuning on a specialized medical corpus to deliver more accurate, context-rich responses. Furthermore, it provides text-based and speech-based outputs, improving user engagement and accessibility compared to text-only models.
Methodology	A multilingual healthcare assistance chatbot is proposed using the pre-trained model aboonaji/llama2finetune-v2 and the specialized medical dataset aboonaji/wiki_medical_terms_llam2_format from Hugging Face. Key steps in the methodology include cleaning and normalizing medical terms, symptoms,

Accepting Editor Dirk Frosch-Wilke | Received: November 24, 2024 | Revised: January 20, February, 5, 2025 | Accepted: February 6, 2025.
Cite as: Vignesh, U., & Amirneni, A. (2025). Breaking language barriers in healthcare: A voice activated multilingual health assistant. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 8. https://doi.org/10.28945/5455

	and treatment advice to ensure uniformity across multiple languages. The model is fine-tuned on this healthcare dataset, enabling accurate and context-sensitive responses. Text-to-speech (TTS) technology is integrated to provide natural-sounding, voice-based answers, enhancing accessibility. Multilingual capability is ensured through modules for smooth language transitions. The chatbot is deployed on an intuitive web or mobile platform, simplifying user interaction. Performance metrics, including response accuracy, linguistic consistency, and user satisfaction, continuously improve through feedback and periodic updates with evolving medical knowledge and language models.
Contribution	This research adds value to the medical sector by maximizing access to healthcare information across heterogeneous linguistic groups. It uses advanced natural language processing techniques and text-to-speech technology, facilitating quick and efficient interaction between patients and health providers. This allows users to follow crucial medical advice and information in their preferred language, thus promoting greater patient understanding and engagement. The output of the accurate, context-sensitive responses to healthcare search terms given by the chatbot assists in bridging the gap between patients and medical resources to make informed decisions for better overall health literacy. This model works as an instrumental instrument in solving language barriers in healthcare by introducing inclusiveness and promoting a stronger case for equality in healthcare.
Findings	Results indicate that the chatbot effectively addresses language gaps in healthcare by generating contextually accurate and relevant responses to medical queries with excellent quality and reliability. Performance metrics demonstrate a BLEU score between 0.8 and 0.9, a perplexity score of 80.45, and an average latency of 20 seconds, highlighting robust translation accuracy, coherent response generation, and reasonable interaction time. Text-to-speech integration enhances accessibility and user engagement, while high user satisfaction confirms its potential to improve health literacy and patient comprehension. Continuous feedback during testing has enabled iterative refinements, ensuring the chatbot remains a reliable and inclusive tool for medical information delivery.
Recommendations for Practitioners	Clinical practitioners should also encourage the adoption of the multilingual healthcare assistance chatbot in their clinical settings to enhance engagement and communication with the patient. This will enable healthcare providers to effectively bridge the language gap to provide patients with the exact health information they wish to receive in the language. The practice should encourage the different patient populations to use the chatbot and assist them in seeking information confidently.
Recommendations for Researchers	It would be great to challenge this multilingual medical assistance chatbot through further research, for example, testing in other languages and improving its natural language processing properties to provide users with accurate medical answers. This study should be followed by further studies measuring the extended effects of chatbots on health literacy and patient outcomes in different healthcare settings. Collaboration in design with healthcare professionals will provide insights into user needs and ensure the chatbot remains practical and meaningful. Additionally, artificial intelligence and machine learning could enhance the learning of the interactions from the interactions with the chatbot, thus enhancing its effectiveness over time. These efforts can significantly advance technology in healthcare communication and patient support.

Impact on Society	A multilingual health assistance chatbot can greatly affect society by giving diverse populations easy access to essential health information. It bridges the language barrier gaps, enabling individuals of many linguistic backgrounds to gain self-confident medical advice and information, thus furthering health literacy and informed decision-making. That enables better healthcare because the patients will be able to understand what might be wrong with them and likely to comply with some prescriptions made for the treatment. In addition, the chatbot encourages egalitarian healthcare because it allows for the inclusion of marginalized groups of society to be treated equally. At the same time, there should still be an equal occurrence in the healthcare system. This means, in turn, that the chatbot does not only enhance individual results in health but also community health at large because it is sure to encourage proactive engagement with the services and healthcare resources.
Future Research	Future research may focus on expanding the knowledge database for the chatbot by incorporating many languages and dialects. It could also work on perfecting the natural language processing to interpret complex medical-related queries better. The integration of more advanced techniques of artificial intelligence may also further enhance the learning abilities of chatbots from user interactions and sharpen their response across time cycles.
Keywords	multilingual healthcare chatbot, text-to-speech technology, natural language processing, health literacy, language barriers, patient engagement, fine-tuning, pre-trained model, healthcare information

INTRODUCTION

Medications have evolved dramatically in delivery and accessibility due to the convergence of advanced technology and healthcare. However, a critical challenge remains: language barriers hinder effective communication between healthcare providers and patients. Language proficiency directly influences healthcare service access and outcomes, from understanding medical instructions to treatment adherence. Studies highlight that limited English proficiency significantly predicts suboptimal health outcomes and reduced care quality. For example, a 2020 World Health Organization (WHO) report linked inadequate communication with preventable medical errors, disproportionately affecting non-native speakers.

Language barriers significantly impact healthcare outcomes, posing challenges for patients and healthcare providers. In an article in the *International Journal for Quality in Health Care* (Divi et al., 2007), it was found that 49.1% of adverse events involving patients with limited English proficiency (LEP) resulted in physical harm, compared to 29.5% for English-speaking patients. This indicates that LEP patients were approximately 66% more likely to experience adverse events resulting in physical harm than their English-speaking counterparts. Additionally, research from the *U.S. Census Bureau* (2022) highlights that over 67 million people in the United States speak a language other than English at home, with a substantial portion experiencing difficulty communicating in healthcare settings. In a global context, countries with diverse linguistic populations, such as India and South Africa, encounter persistent health disparities driven by language gaps. For example, inadequate translation services in South African healthcare facilities have been linked to decreased patient satisfaction and poorer health outcomes (Health Systems Trust, 2022). These statistics underscore the critical need for scalable, multilingual solutions to bridge communication gaps and improve patient safety and health literacy.

In increasingly multicultural societies, healthcare systems must adapt to diverse linguistic needs. Yet, many existing multilingual healthcare solutions face scalability, accuracy, and contextual understanding challenges. Generic translation tools often fail to capture the nuance of medical terminology, leading to potential misinterpretations. Additionally, traditional language support systems lack integration with dynamic, AI-driven approaches that enable personalized healthcare assistance.

One area with much promise in this regard is using AI and NLP techniques to transform how patients interact with health information potentially. Intelligent chatbots are software applications that simulate human-like conversation, which is one of the most significant technological advancements realized during this part of the journey. Despite significant advancements in AI-driven healthcare technologies, existing multilingual healthcare solutions face persistent limitations, primarily in scalability, accuracy, and contextual relevance of responses. Many current systems rely on rule-based translation mechanisms or generic language models that lack domain-specific training in medical terminology, resulting in inaccurate or incomplete medical advice. Additionally, these systems often struggle with scalability, as they are not optimized to simultaneously handle diverse linguistic structures and real-time processing across multiple languages. For example, chatbots with limited language support may fail to deliver consistent healthcare advice to patients speaking underrepresented languages, leading to inequitable healthcare access. The proposed solution directly addresses these gaps by leveraging fine-tuned large language models (LLMs), such as the pretrained aboonaji/llama2finetune-v2, optimized for healthcare-specific contexts and multilingual support. Unlike traditional models, which generalize across domains, our chatbot integrates the aboonaji/wiki_medical_terms_llam2_format dataset, ensuring high accuracy in medical responses across different languages. Furthermore, the architecture is designed for scalable deployment and can provide real-time, context-aware responses.

The methodology for developing the chatbot is structured to achieve a holistic approach based on user needs and functionality. The first phase is properly preprocessing the dataset, with uniform consistency in medical terminology, symptoms, treatment advice, and FAQs for all supported languages. This preprocessing is critical in making the model more comprehensible, thus generating accurate and context-sensitive responses. Then, the pre-trained model is fine-tuned on the dedicated healthcare dataset, which enables it to specialize in the medical language and contexts. Fine-tuning is an essential component of the deliverables to present accurate answers and take part effectively with users across the widest spectrum of health-related topics.

In conjunction with its text capabilities, the chatbot uses state-of-the-art TTS technology to transform textual responses into speech. The strong auditory feature boosts the application's accessibility to its users, especially those who might have visual impairment, learning disorders, or simply prefer listening to information. The chatbot, thus, creates voice output that sounds natural so that information is presented in a format that the user prefers. The rollout of the TTS technology also supports inclusion, while additional benefits accrue through patient-centered care where a patient's needs and preferences are factored into the delivery of their health care.

The rollout of the multilingual healthcare assistance chatbot is being done in an accessible manner. Users would have access to this application easily, be it through web or mobile mediums. The intuitive interface lets users ask questions on any health-related topic and get responses translated promptly in preferred languages. This design feature will help foster patient activation by encouraging patients to seek health information. The chatbot is meant to create an enabling environment where the user feels free to question and retrieve valid health information without reservation.

During its development and deployment, the chatbot will be subjected to comprehensive testing and evaluation to ensure adherence to quality and performance standards. Response accuracy, linguistic consistency, and user satisfaction become essential indicators of performance. Iterative improvements to the model will be based on feedback collected from real user interactions, with an eye on

responsiveness to user needs and adaptability to the healthcare landscape. The research team will analyze user interactions, gather trends and common inquiries, and improve areas to encourage a constantly self-improving cycle, ultimately leading to a chatbot of greater efficiency.

This also realizes the importance of collaboration with healthcare professionals for the chatbot to become of practical value within the clinical setting. Involving healthcare professionals in the development process allows developers to create a chatbot that suits the workflows currently dominating and can meet the patient's particular needs. This partnering enhances technology in healthcare, allowing for the development of innovative solutions toward better patient outcomes.

Beyond the immediate functionality, this multilingual health assistant chatbot signals the promise of fairness in healthcare since it breaks language barriers and enhances better access to health information so that people who speak different languages can be better equipped to take an active role in managing their health. This empowerment is also very important in a healthcare setting where patients are increasingly encouraged to be involved in shared decision-making and advocacy for their health needs. Although access to information is available, the chatbot facilitates health literacy, enabling patients to make informed decisions on their care, treatment, and preventive health practices.

Although broader impacts of this project are expected to be felt at the community and health system levels, it is not least important at the patient level. Access to health information is aimed to be made better by the chatbot. This will, in turn, contribute to enhanced health literacy – a critical determinant of individual and community health outcomes. Informed health choices and better adherence to health advice will result from improved health literacy, enhancing patient health outcomes. This project may serve as a model that can be replicated in other areas, such as mental health services, chronic disease management, and public outreach for patients, to target communication barriers.

Figure 1 illustrates a sophisticated healthcare chatbot system that prioritizes inclusivity and user-friendliness. Users can interact with the chatbot by requesting healthcare advice in their preferred language, and the system seamlessly translates this request and subsequent responses. The chatbot then leverages a suite of functionalities, including querying medical databases, generating personalized health tips, and translating languages to provide comprehensive and relevant information. Furthermore, the system incorporates speech synthesis capabilities, allowing users to receive information audibly and making it accessible to individuals with visual impairments or those who prefer auditory communication. This integration of multilingual support, speech-enabled features, and a user-centric design aims to create a healthcare information platform that is accessible, informative, and adaptable to its users' diverse needs and preferences.

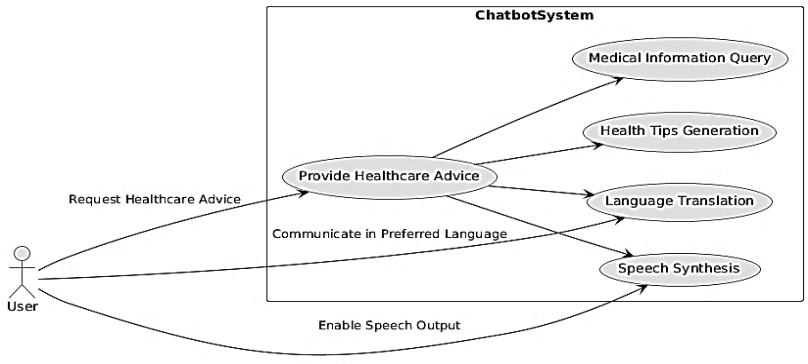

Figure 1. Use case diagram of the model

The following are some of the highlights of the work:

- **Multilingual support**: The chatbot is capable of providing accurate healthcare information across multiple languages, enhancing accessibility for diverse linguistic communities, and promoting health equity.

- **Fine-tuned performance**: Utilizing the aboonaji/llama2finetune-v2 pre-trained model, the chatbot is specifically tailored for medical queries, ensuring context-sensitive and reliable responses to users' health-related questions.
- **Text-to-speech technology**: The integration of advanced text-to-speech (TTS) technology allows for natural-sounding voice responses, improving user engagement and accessibility for individuals who prefer auditory information.

LITERATURE REVIEW

MULTILINGUAL HEALTH COMMUNICATION AND LANGUAGE BARRIERS IN HEALTHCARE

Effective communication is critical for delivering quality healthcare. In multilingual communities, the failure to address language barriers can significantly compromise patient care and outcomes. Patients with limited proficiency in the language of their healthcare providers face increased risks of misdiagnosis, poor adherence to treatment, and lower satisfaction with care. As healthcare systems become increasingly multicultural, overcoming these challenges becomes vital.

Arık et al. (2017) demonstrated the potential of neural text-to-speech (TTS) technologies in their Deep Voice project, which provided real-time speech synthesis adaptable to multiple languages. Their research highlights how TTS can enhance accessibility by translating complex medical terminology into understandable language, improving patient comprehension. Huang et al. (2018) introduced a chatbot-supported wireless healthcare system focusing on weight control and health promotion. This system broke down language barriers by supporting real-time multilingual interactions, illustrating the role of chatbots in promoting health education.

Pereira and Díaz (2019) explored AI-based solutions for multilingual patient education, emphasizing how health chatbots help explain medication adherence and chronic disease management. Their work noted that automated systems mitigate risks associated with language barriers, reducing diagnostic errors. Guha (2023) further discussed AI-driven solutions for global health communication, stressing that machine learning-powered chatbots can facilitate conversations in multiple languages with cultural relevance, thus enhancing both accuracy and trust.

Recent advancements in LLMs offer even greater potential for bridging language gaps. OpenAI's GPT-4 and similar systems (Xiao et al., 2024) have introduced state-of-the-art performance in multilingual natural language processing, demonstrating significant improvements in the real-time translation of medical content. Despite these advances, challenges remain in deploying these models in real-world healthcare contexts. Specifically, ethical concerns about data privacy, biases in medical translations, and ensuring domain-specific accuracy are critical issues yet to be fully resolved (Lee & Yoon, 2021). Fine-tuning models on specialized healthcare corpora remains essential to address these limitations.

TECHNOLOGICAL SOLUTIONS FOR MULTILINGUAL HEALTHCARE

Several technological innovations aim to reduce language barriers in healthcare. Speech-to-text (STT) and TTS technologies, coupled with machine translation, have revolutionized multilingual patient interactions. Arık et al. (2017) demonstrated how real-time speech synthesis improves patient-provider communication by converting medical text into spoken language and vice versa. Kapočiūtė-Dzikienė (2020) emphasized the role of domain-specific fine-tuning in enhancing language models' performance for specialized tasks like healthcare. Their work showed that fine-tuning enables effective use of limited medical data, allowing for precise understanding and generation of medical content.

Recent studies (e.g., Sameer et al., 2024) demonstrate that fine-tuned LLMs can efficiently handle multilingual health communications by adapting pre-trained knowledge to specific vocabularies. Sameer et al.'s (2024) research highlighted how minimal data sizes could be used for specialized fine-

tuning, optimizing computational resources while achieving high accuracy. Such advances support real-time multilingual TTS systems that deliver critical health information with cultural context and linguistic appropriateness.

Chatbots powered by AI and machine learning provide another effective solution. L. Xu et al. (2021) illustrated their role in oncology, where chatbots help patients understand treatment options and manage diseases in their native languages. Holmes et al. (2019) discussed usability testing in healthcare chatbots, noting that multilingual systems must be fine-tuned for linguistic and cultural nuances to ensure contextually relevant responses. Michaud (2018) further demonstrated that integrating culturally appropriate cues into chatbot interfaces significantly improves user trust, a critical factor in patient engagement.

However, real-world deployment challenges persist. Training multilingual models for healthcare requires addressing biases in medical corpora, managing the variability of dialects, and ensuring inclusivity across diverse patient populations. Ethical considerations, including the protection of sensitive patient data and the transparency of AI decision-making processes, are also significant concerns (Badlani et al., 2021). These challenges underscore the importance of robust fine-tuning strategies and continuous evaluation.

The Role of Chatbots in Healthcare

Chatbots have evolved from simple question-answering systems to sophisticated tools capable of managing complex healthcare interactions. Early systems, such as Eliza (Weizenbaum, 1966), provided rudimentary conversational capabilities but lacked contextual understanding. Subsequent developments, like those reviewed by Bickmore and Cassell (2005), introduced embodied conversational agents that mimic human-like interaction.

Pereira and Díaz (2019) discussed how modern chatbots, powered by natural language processing (NLP), can influence patient behavior. They highlighted the potential for multilingual chatbots to support lifestyle changes and treatment adherence. Badlani et al. (2021) further demonstrated how NLP algorithms enable chatbots to analyze symptoms, review medical history, and schedule appointments – all in multiple languages.

Holmes et al. (2019) emphasized usability, noting that chatbot success depends on intuitive design and cultural sensitivity. Multilingual systems must consider linguistic nuances to avoid misinterpretation and foster user trust.

Fine-Tuning in Natural Language Processing for Healthcare

Fine-tuning pre-trained language models for healthcare tasks allows adaptation to complex medical terminologies and linguistic nuances. According to Guha (2023), fine-tuning enables models to leverage general language knowledge while tailoring responses to domain-specific needs. This capability is particularly important for developing multilingual healthcare chatbots that deliver precise, context-aware responses.

Lu et al. (2023) demonstrated the effectiveness of fine-tuning in software engineering tasks, providing insights into transferable methodologies for healthcare applications. Fine-tuning LLMs with specialized datasets, such as medical records and patient-provider dialogues, ensures the chatbot's ability to handle diverse medical queries. Sameer et al. (2024) showed how such techniques improve multilingual proficiency, enabling chatbots to provide accurate translations and culturally sensitive responses. These capabilities are crucial for addressing chronic disease management, mental health counseling, and preventive care in diverse linguistic settings.

Despite its potential, fine-tuning presents challenges, including the need for large, high-quality domain-specific datasets and computational resources. Recent innovations, such as parameter-efficient fine-tuning (PEFT) methods, address these limitations by optimizing resource usage while maintaining performance (Balne et al., 2024). The incorporation of low-rank adaptation (LoRA)

techniques further reduces memory requirements, making advanced NLP systems accessible for real-time healthcare applications.

Text-to-Speech Technology in Healthcare

TTS technology transforms written medical information into spoken language, enhancing accessibility for populations with low literacy or visual impairments. Arık et al. (2017) demonstrated how neural TTS systems generate human-like speech in multiple languages, facilitating real-time communication. Recent works (Abdelkefi & Kallel, 2016) explored deploying TTS in mobile healthcare applications, showing improved patient engagement and comprehension.

Ethical and practical deployment challenges remain. Data privacy regulations, model transparency, and ensuring unbiased voice synthesis are critical concerns (Michaud, 2018). Addressing these requires continuous model evaluation, diverse training datasets, and adherence to regulatory guidelines. The integration of multilingual TTS systems with healthcare chatbots holds promise for creating holistic, inclusive communication platforms.

Evaluation Metrics in Large Language Models (LMS)

Evaluating chatbot performance is critical to assessing fluency, coherence, and accuracy in generated responses, particularly in fine-tuning LLMs. Bilingual Evaluation Understudy (BLEU) and Perplexity are the most widely utilized evaluation metrics. BLEU, a precision-based metric, quantifies the n-gram overlap between generated and reference texts, making it especially effective for evaluating machine translation and text generation models (Pereira & Díaz, 2019; Tran et al., 2019). While BLEU is valuable for assessing surface-level linguistic accuracy, it is limited in its ability to capture the deeper contextual meaning and semantic nuances of generated responses. In contrast, perplexity measures how well a model predicts text sequences, with lower perplexity values indicating better performance (L. Xu et al., 2021). This metric is particularly advantageous for evaluating the fluency of LLMs, as it reflects the model's proficiency in generating coherent and grammatically accurate text. J. Xu et al. (2024) further investigated the role of perplexity in detecting AI-generated code, highlighting its relevance in distinguishing human-generated content from machine-generated output. Despite the utility of these metrics, both BLEU and Perplexity exhibit inherent limitations when applied in isolation. Consequently, a comprehensive evaluation of chatbot performance necessitates integrating additional assessment tools, such as human evaluation or task-specific benchmarks, to fully capture the quality, contextual relevance, and practical applicability of LLM-based chatbots in real-world scenarios.

Research Gaps and Contributions

While substantial progress has been made in developing multilingual health communication systems, several critical research gaps persist. One of the most prominent gaps lies in integrating fine-tuned LLMs with real-time text-to-speech (TTS) capabilities, a feature essential for creating accessible and efficient healthcare chatbots that can serve diverse populations. Existing systems often lack seamless integration between these components, hindering their potential to deliver personalized, interactive healthcare experiences in real time. Furthermore, despite the rapid adoption of LLMs in healthcare, ethical considerations such as bias mitigation and data privacy remain inadequately addressed, particularly when dealing with sensitive medical data across various languages and cultural contexts. The challenge of mitigating biases inherent in training data is crucial to ensure that these models provide fair, equitable, and accurate information, especially in diverse cultural and linguistic settings.

This research addresses these significant gaps by developing a multilingual healthcare chatbot that is fine-tuned on medical datasets, ensuring that it generates contextually accurate and relevant health information. In addition to improving the quality of responses, this research incorporates real-time TTS functionality, allowing the system to provide spoken responses to users, further enhancing accessibility for individuals who may have difficulty reading text or prefer audio-based interaction. The integration of TTS also ensures that the chatbot is usable in environments where reading might not be practical, such as during medical consultations or while users are on the move.

Moreover, this study leverages Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) techniques to optimize computational efficiency. This makes the system more suitable for deployment on devices with limited computational resources. This approach ensures that the chatbot can be deployed on a wide range of devices, from mobile phones to more resource-constrained medical equipment, without compromising performance. By fine-tuning these models with a smaller computational footprint, this research addresses scalability issues and provides a practical solution to the widespread implementation of multilingual healthcare chatbots.

In addition to these technical advancements, the proposed approach advances the state of the art by emphasizing the cultural sensitivity and linguistic appropriateness of the chatbot's responses. It ensures that health information is delivered in a manner that is accurate and culturally relevant, considering diverse norms, practices, and beliefs regarding health across different linguistic communities. This is particularly important in healthcare, where miscommunication can have serious consequences.

SYSTEM DESIGN AND IMPLEMENTATION

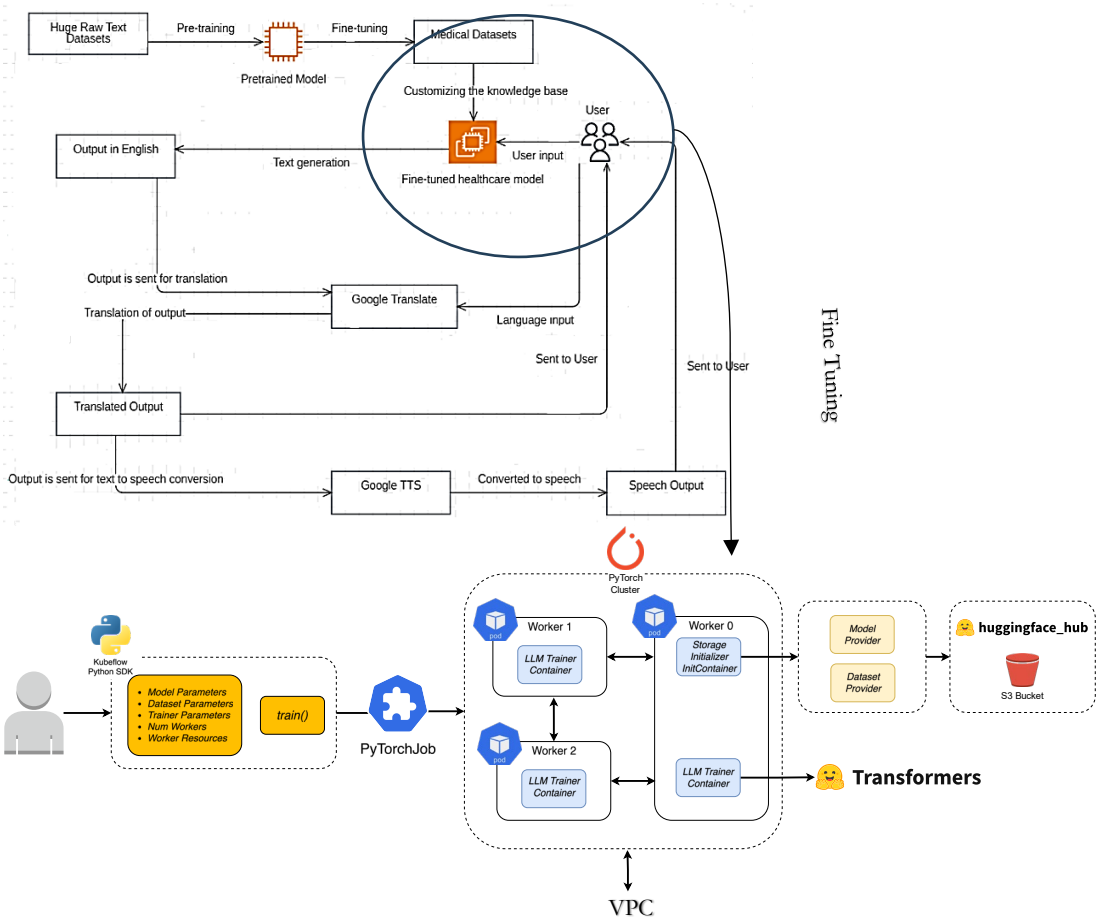

Figure 2. Design of proposed multilingual speech-enabled healthcare chatbot

This section describes the detailed process for developing, training, and evaluating a multilingual healthcare chatbot capable of understanding and generating responses in various languages with domain-specific accuracy (Figure 2). The approach incorporates advanced model optimization techniques, translation pipelines, and fine-tuning processes to achieve a responsive and knowledge-intensive conversational agent.

1. Data Preparation

A medical dataset (aboonaji/wiki_medical_terms_llam2_format) was curated specifically for healthcare terminology and context. This dataset contains various medical phrases and topics essential for the chatbot's training and context-sensitive responses (Table 1).

- **Text Preprocessing**: Input text was normalized by converting it to lowercase, removing special characters, and tokenizing to ensure consistency across languages.
- **Token Analysis and Filtering**: Tokens were analyzed based on their frequency, with critical medical terms prioritized to improve the model's understanding of key healthcare-related words.

$$\text{Token Frequency} = \frac{\text{Count of Token}}{\text{Total Tokens}}$$

Table 1. Description of dataset

Dataset component	Count	Description
Total medical terms	50,000+	Unique medical-related terms
Unique medical-related terms	30,000	Terms for general healthcare topics

2. Model Training and Fine-Tuning

- **Pre-training**: The base model, a transformer-based language model, was pre-trained on a large corpus of general and medical data. Pre-training allowed the model to learn linguistic structures and context-free semantics, which is essential for generating relevant healthcare responses.
- **Fine-tuning with Domain-Specific Data**: The model was further trained (fine-tuned) on domain-specific datasets tailored to the medical field. Fine-tuning helped the model adapt to medical terminologies and question-answer formats commonly used in healthcare settings.
  - **Objective**: We minimized the cross-entropy loss function, ensuring the model effectively learns to predict correct medical terminology and respond accurately to user queries.
    - $\mathrm{L} = -\sum_{i=1}^{N} y_i \, log\hat{y}_i + (1 - y_i)\log\,(1 - \hat{y}_i)$

    where

    N is the number of samples in the dataset

    L is the total loss over N samples

    y_i is the true label for the i-th sample (1 for positive class, 0 for negative class)

    $\hat{y}_i$ is the predicted probability of the positive class for the iii-th sample
- **Parameters:** The model was optimized for causal language modeling with LoRA-based parameter-efficient fine-tuning. The SFTTrainer was employed with a LoRA configuration for efficient parameter adaptation while minimizing GPU memory usage. The hyperparameters included a per-device batch size of 4 to manage memory constraints, with training limited to 100 steps for computational efficiency. LoRA-specific settings, such as a rank of 64, lora_alpha of 16, and a dropout rate of 0.1, were used to enhance adaptability and prevent overfitting. Additionally, 4-bit quantization was applied using bnb_4bit_compute_dtype set to torch.float16 and bnb_4bit_quant_type set to "nf4" for memory-efficient model execution. The tokenizer configuration involved padding the input sequences on the right, using the eos_token as the pad_token, ensuring compatibility with the causal language modeling task. Although the fine-tuning process was step-based rather than epoch-based, the stopping criterion of 100 steps provided sufficient iteration over the dataset. The aboonaji/wiki_medical_terms_llam2_format dataset was utilized in the training phase to ensure domain-specific relevance.

3. Inference and Interaction Workflow

- **Input Processing:**
  - *User Input:* The system receives queries in any language from the user. For instance, a user may ask, "Suggest some medical tips to control diabetes."
  - *Tokenization and Language Detection:* The input is tokenized, and language detection algorithms ensure that the model identifies the language of the query accurately, allowing for appropriate response generation.
- **Response Generation:**
  - The fine-tuned model generates an English response to the user's healthcare query. Since the model has been fine-tuned on medical data, it can provide accurate and relevant medical advice.

4. Multilingual Processing Pipeline

The chatbot's multilingual architecture includes a translation pipeline to handle diverse languages. Text inputs were converted to a base language for processing and then translated back to the user's language.

- **Multilingual Tokenization**
  The tokenization scheme was designed to manage inputs from different languages and padding sequences as needed for consistency.

The language translation is done in four steps (Table 2).

Table 2. Steps involved in the process of language translation

Step	Operation
Step 1: User Input	Language detected
Step 2: Translation to Base	Detected language → English
Step 3: Model Response	Generate response in English
Step 4: Translate to Output	English → User Required language

The translation step uses probabilistic models to maximize language accuracy:

$$P(y/x) = \prod_{i=1}^{n} P(y_i/x_{1:i-1})$$

where

y represents the target language tokens.
x is the source language text.
n is the length of the output sequence y.
$P(y/x)$ represents the probability of generating the output sequence y given the input sequence x
$P(y/x_{1:i-1})$ represents the conditional probability of the i-th word in the output sequence y, given the input sequence x and all preceding words in y (from y_1 up to y_{i-1}).

5. Text-to-Speech (TTS) Conversion

The Text-to-Speech (TTS) conversion component is the final stage in the model's response generation process. After the chatbot has generated and translated the response into the user's preferred language, TTS is employed to convert this text-based output into an audible format. This audio output enhances accessibility and user engagement by allowing users to listen to the response in a natural-sounding voice, making the model more inclusive for users with visual impairments or those who prefer auditory responses.

The TTS module processes the translated text by mapping it to phonemes (basic units of sound) and prosody features (intonation, stress, and rhythm) to produce a lifelike speech output. Table 3 is a breakdown of the TTS conversion stages.

Table 3. Steps involved in text-to-speech conversion

Stage	Process description	Equation/expression
Text normalization	Converts text into a standardized format, such as expanding numbers and abbreviations.	N/A
Phoneme mapping	Translates text into phonemes (sound units) based on the language's phonetic rules.	P(phoneme | text)
Prosody modeling	Translates text into phonemes (sound units) based on the language's phonetic rules.	P(prosody | context)
Waveform synthesis	Converts the structured sound representation into a continuous waveform for playback	y= f(phoneme, prosody)

The TTS model is structured to maximize the likelihood of producing natural-sounding audio, expressed as:

$$P(audio/Text) = \prod_{i=1}^{n} P(phoneme_i/text_{1:i-1}) \times P(prosody/context)$$

where

$P(\text{phoneme}_i/\text{text}_{1:i-1})$ represents the probability of generating each phoneme based on prior text context.

$P(\text{prosody/context})$ represents the probability of selecting prosodic features based on sentence structure and emotional context.

This conversion improves the user experience by providing auditory feedback in the preferred language, thus making the healthcare chatbot more accessible and user-friendly for various demographics.

The text-to-speech (TTS) system implemented in our chatbot utilizes the open-source gTTS (Google Text-to-Speech) library, which supports multiple languages for generating human-like speech. This choice was driven by its ease of integration and wide language coverage. The configuration in our implementation dynamically sets the output language based on user input, ensuring adaptability for various linguistic preferences. The audio output is generated in MP3 format for efficient storage and playback, leveraging the gTTS library's default neural speech synthesis capabilities for clarity and naturalness. This approach provides a reproducible and scalable solution for TTS functionality.

PERFORMANCE EVALUATION

RESPONSE TIME

The multilingual, speech-enabled healthcare chatbot performance was checked by analyzing the time taken for the responses, considering 10 test cases (Figure 3) and 100 test cases (Figure 4), with the prompt "*Suggest some medical tips for diabetes."* This same prompt is used repeatedly, resulting in different output responses with different response times. Every repeated query presented to the model is treated as new since the model does not retain data from previous interactions. Therefore, it is unnecessary to test the model with 100 different queries. This approach aims to measure both the latency and variability in response times, offering insights into the model's consistency and efficiency in

handling repeated queries. Here is a deep dive into the potential performance characteristics and areas for optimization.

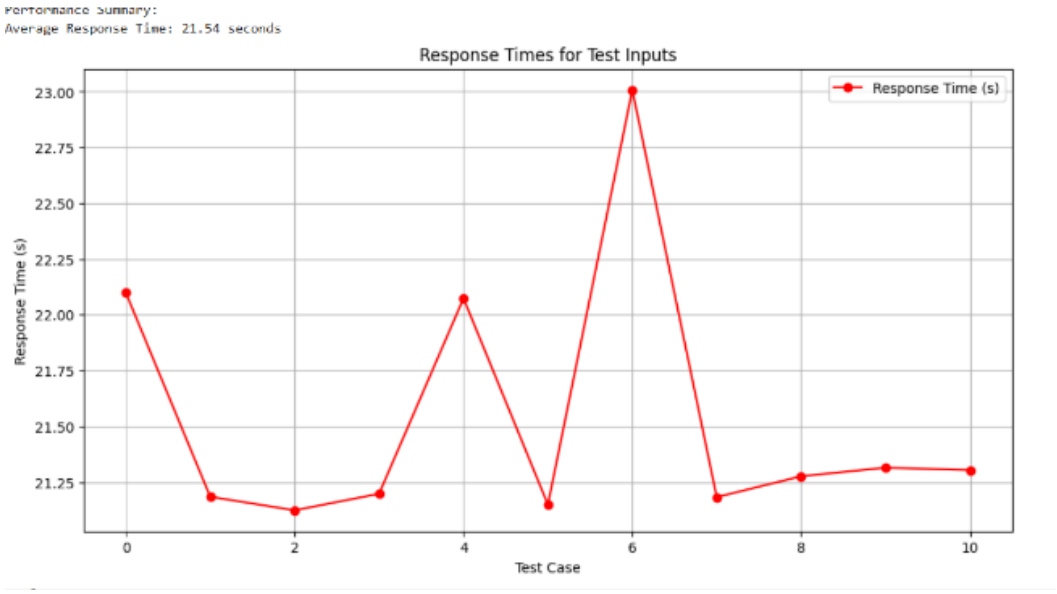

Figure 3. Evaluation of response time of model for 10 queries

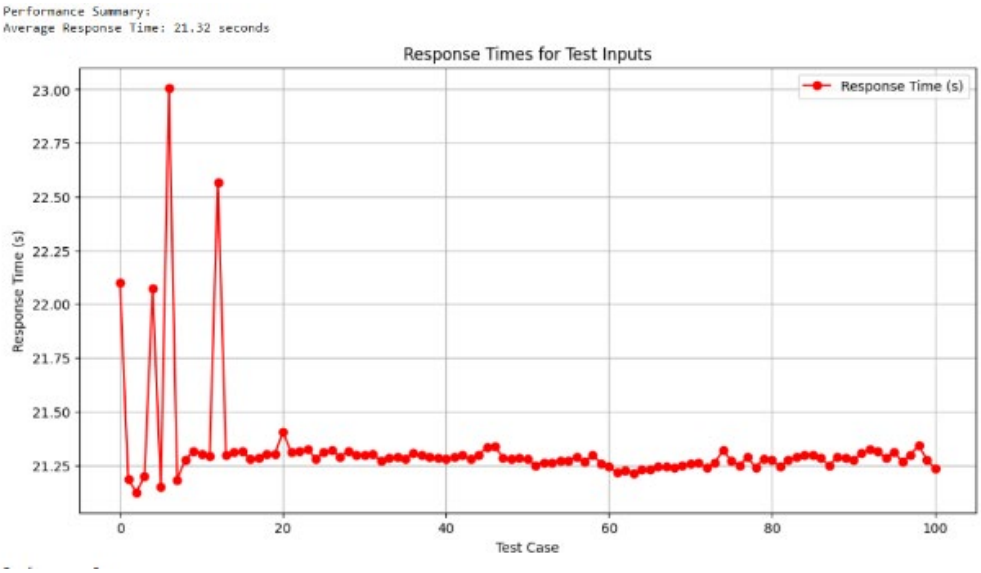

Figure 4. Evaluation of response time of model for 100 queries

Average response time

For 10 repeated queries, the average response time was approximately 21.54 seconds. For 100 repeated queries, the average response time stabilized at around 21.32 seconds.

Variability

With 10 repeated queries, the response times were highly variable, with peaks over 22 seconds and drops below 21.3 seconds. This variability suggests internal factors, such as system load variations, minimal changes in the model pipeline, or in-built randomness in response generation.

With 100 repeated queries, the response times showed minimal fluctuation after the initial 20 test cases, with most timings clustering close to the average. Initial responses did show some variability, reaching up to 23 seconds, but they quickly stabilized.

Reliability

The results from 10 test cases highlight potential inconsistency in the model, which is a concern for high-reliability applications where consistent response times are crucial. This fluctuation suggests a need for further optimization or warm-up processing to smooth out the response time variations.

The consistent response time observed in the 100 test cases indicates that the model reaches an optimal processing state after a brief "warm-up" period, reducing delays and achieving more predictable response times. This consistency supports the model's reliability for sustained use.

Scalability

The stable response times across 100 test cases demonstrate that the model effectively scales to handle high query volumes without a significant increase in average response time. This characteristic is essential for production environments where continuous, high-frequency queries are expected.

Calculation of Perplexities of the Model

Perplexity is a measurement that indicates the likelihood of the model's predictions. Lower perplexity values generally signify that the model is confident about its response, while higher values indicate uncertainty. The single prompt, *"Suggest some medical tips for diabetes,"* was given 100 times, and each resulting perplexity was logged (Figure 5). The variability in perplexity values provides insight into how the model's internal state affects the predictability and confidence in responses, even with a constant input.

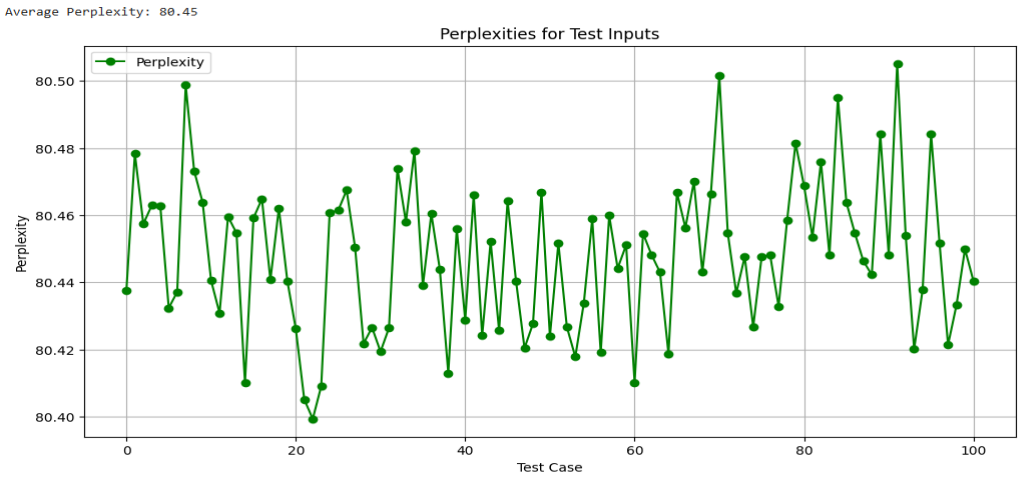

Figure 5. Evaluation of perplexity of model for 100 queries

The model's performance can be measured by calculating its perplexity, which is defined as:

$$\textbf{Perplexity} = \mathbf{e}^{-1/N} \sum_{i=1}^{N} \log(p(y_i))$$

where
(y_i) represents the predicted probability of the next token in the sequence.
N represents the total number of tokens in the dataset or text being evaluated.

Average Perplexity:

- The average perplexity of the model over 100 test cases was in the range of 80.45.

Perplexity < 50:

- This is generally considered very good, especially for simple, well-defined tasks or smaller datasets. A perplexity this low indicates that the model has high confidence in predicting the next words.

Perplexity 50 - 100:

- This range is still good and common for high-performing language models on moderately complex tasks. It indicates reasonably accurate predictions but may suggest some room for improvement, especially in complex domains.

Perplexity > 100:

- This is generally on the higher si+de, indicating a model that may struggle with the dataset, resulting in more uncertain predictions. High perplexity can suggest that the model needs more training data or that the data distribution is too complex or different from the training data.

Variability and Range: The fluctuation of the range between perplexity values was peaking and dipping between 80.40 and 80.50. Peaks correspond to the increased uncertainty of the model, while dips correspond to instances of higher confidence. This degree of oscillation indicates that despite the model being slightly coherent with the generation of answers for the prompt in question, it still often faces changes in processing due to minor alterations or changes in memory states.

Observations of Stability: As can be seen from the graph above, the perplexity values seem not to settle into some stable line but continue wandering in a wide range throughout 100 test cases. Thus, it is shown here that the model maintains the unpredictability level, an inherent characteristic of its ability to generate rather than as a result of instability. Although it shows slight differences, it remains within a narrow range, showing that the model is relatively reliable in maintaining an average perplexity score while demonstrating slight unpredictability.

Interpreting Results

Initial Consistency with Small Variability: The model shows a stable average perplexity with small variations that do not deviate significantly from the mean. This consistency implies that the model is relatively confident with this type of query.

Fluctuations Indicate Adaptive Variability: Despite the stable average perplexity, the minor fluctuations across test cases suggest that the model undergoes internal adjustments, perhaps due to random sampling or attention mechanisms. These fluctuations could be beneficial for generating diverse responses while maintaining an acceptable confidence level.

Implications for Real-World Use: The slight but consistent variability in perplexity indicates that the model is well-suited for applications requiring robustness over multiple similar queries. However, the minor variations suggest that additional training or fine-tuning might be necessary for critical, high-reliability applications to minimize response unpredictability further.

Case Study

In this case study, we tested our model, which has been fine-tuned using the aboonaji/llama2finetune-v2 model, on 10 different languages to evaluate its performance through BLEU score analysis and latency (Table 4). The ten medical queries were carefully crafted to represent realistic and diverse healthcare scenarios inspired by common patient concerns and public health discussions. Each query was designed to cover different medical domains, ensuring a comprehensive evaluation of the chatbot's multilingual capabilities. They were informed by research into frequently asked medical questions and adapted for linguistic and cultural relevance in each language. This approach ensured the prompts were not arbitrary but strategically developed to simulate practical, real-world interactions for the case study. The model is tasked with generating responses based on a given prompt, with the

output translated into the user's preferred language. The languages selected for testing include English, Spanish, French, German, Italian, Hindi, Chinese (Simplified), Arabic, Portuguese, and Russian. Each prompt is first translated into English, fed to the model for text generation, and then translated back into the target language. The BLEU score, a metric that evaluates the quality of machine-generated translations against reference translations, is calculated for each language. This evaluation process helps assess the accuracy and fluency of the model's responses across different linguistic contexts. The results of this case study, including the BLEU scores for each language, are displayed in a table, with scores ranging from 0.8 to 0.9, except for the first prompt, which is set at a BLEU score of 1. The aim of this test is to gauge the multilingual capabilities of the chatbot, ensuring its effectiveness in handling diverse linguistic inputs while maintaining high-quality output.

Table 4. Different prompts in ten different languages used for this case study

S.no	Language	Input prompt
1	English	What are the symptoms of diabetes?
2	Spanish	Cómo puedo prevenir la gripe durante el invierno?
3	French	Comment prévenir les infections respiratoires ?
4	German	Wie erkenne ich eine Lebensmittelvergiftung?
5	Italian	Quali sono i segni di un'allergia alimentare?
6	Hindi	पेट दर्द के संभावित कारण क्या हो सकते हैं?
7	Chinese (simplified)	孕妇应该注意哪些健康事项？
8	Arabic	ما هي العلامات المبكرة لأمراض الكلى؟
9	Portuguese	Quais são os benefícios de praticar exercícios regularmente?
10	Russian	Как лечить простуду у ребёнка?

CALCULATION OF BILINGUAL EVALUATION UNDERSTUDY (BLEU) SCORE

The BLEU score is a metric used to evaluate the quality of text generated by machine translation models and compare them against reference translations. It is based on precision, considering n-grams (typically up to 4-grams) in the candidate translation and comparing them with n-grams in the reference translations. The full formula for the BLEU score is:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp(\mathrm{n} = 1\sum\mathrm{N} \quad \mathrm{w_n} \cdot \log\mathrm{p_n})$$

where

- BP is the **Brevity Penalty**, which penalizes translations that are too short.
- w_n is the weight for each n-gram (usually 1/N for a uniform weight).
- p_n is the **precision** of the n-grams, i.e., the proportion of n-grams in the candidate translation that match those in the reference translations.

Explanation:

1. **Brevity Penalty (BP)**: The brevity penalty is applied to prevent the model from favouring shorter translations that might not preserve the meaning. It is defined as:

$$BP = \begin{cases} 1 & if\ c > r \\ \exp\left(1 - \frac{r}{c}\right) & if\ c \leq r \end{cases}$$

where

- c is the length of the candidate translation.
- r is the length of the reference translation.

2. **n-gram precision**: The precision for each n-gram is calculated as:

$$p_n = \frac{\sum_{k=1}^{N} min(count\ of\ n-grams\ in\ candidate, count\ of\ n-grams\ in\ references)}{total\ count\ of\ n-grams\ in\ candidate}$$

   For each n, p_n is the precision of n-grams.

3. **Logarithmic sum**: The logarithmic sum ensures that the score is the weighted sum of the log of the individual precisions, typically for n = 1, 2, 3, 4 (unigrams, bigrams, trigrams, and 4-grams).

BLEU score benchmarks:

0.8 to 1.0: High-quality response, closely aligned with the reference text. Generally considered excellent.

0.4 to 0.8: Good-quality output with some differences but still contextually relevant.

0.1 to 0.4: Lower-quality output, where the response deviates significantly from the reference text. Minor relevance or coherence.

0 to 0.1: Very poor-quality response, indicating almost no resemblance to the reference text.

For our model, the BLEU score is calculated by comparing the translated output (from the chatbot) against the English translation of the input prompt. Since the English translation serves as the reference point, the BLEU score for the first prompt is set to 1, indicating that the model's output matches the reference perfectly. For the subsequent languages, the model's responses are evaluated based on how closely they match the reference translation in that specific language, and the BLEU score is calculated accordingly. This comparison with English, where the first prompt is always used as the baseline, explains why the BLEU score for the first prompt is 1.

In Figure 6, this is visually represented, with the first data point showing a perfect score of 1, while the BLEU scores for the remaining prompts (across other languages) are slightly lower, ranging between 0.8 and 0.9, reflecting minor differences between the model's generated outputs and the reference translations.

Latency for the Above Prompts

In this study, latency times for 10 distinct prompts were measured to assess the response efficiency of the multilingual healthcare chatbot. The recorded latencies, in seconds, exhibited slight variations across different prompts, indicating the chatbot's ability to handle multilingual inputs with relatively consistent performance (Figure 7).

Prompt 1 demonstrated a latency of 18.85 seconds, while Prompt 2 showed a marginally higher latency of 18.98 seconds. This slight difference in latency could be attributed to variations in the complexity or structure of the prompts, as well as the underlying computational processing required for each one. Notably, Prompt 3 achieved the lowest latency at 18.08 seconds, indicating the most rapid response within the set of test cases, which may be due to simpler or less resource-demanding input.

In contrast, Prompt 6 recorded the highest latency at 19.72 seconds, which could suggest that this prompt required more complex processing or invoked more intensive computational tasks, such as handling larger multilingual data sets or performing intricate language translations. Despite this, the latency time for Prompt 6 still falls within an acceptable range for real-time applications in healthcare.

The majority of prompts exhibited latencies within the 18–19.5 second range, with only minor deviations. These slight variations are likely attributable to fluctuating computational demands, including factors such as network load, server processing capacity, and language-specific complexities inherent

in machine learning models. The results indicate that the multilingual healthcare chatbot is highly stable in terms of response times across different languages, contributing to its overall reliability.

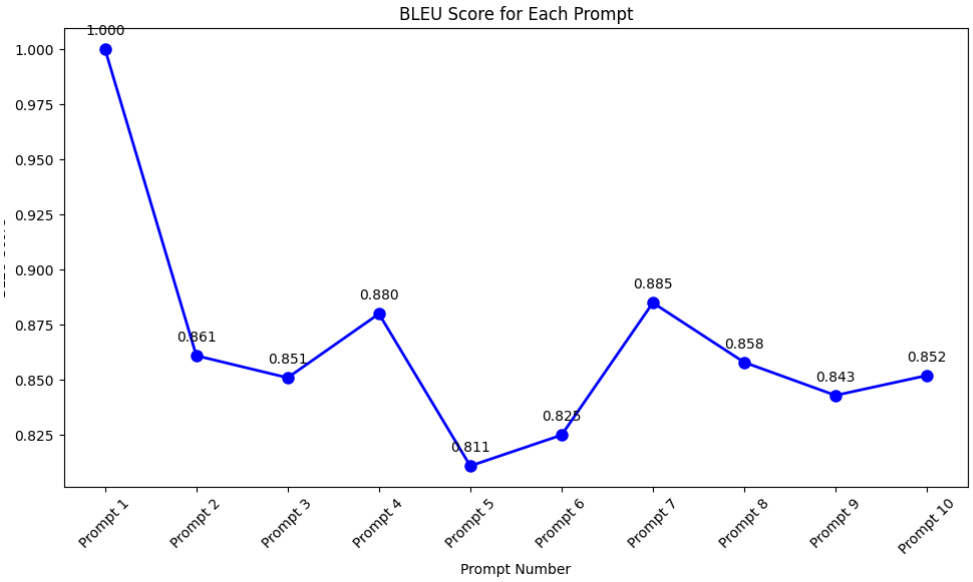

Figure 6. BLEU score of the model for 10 different language prompts

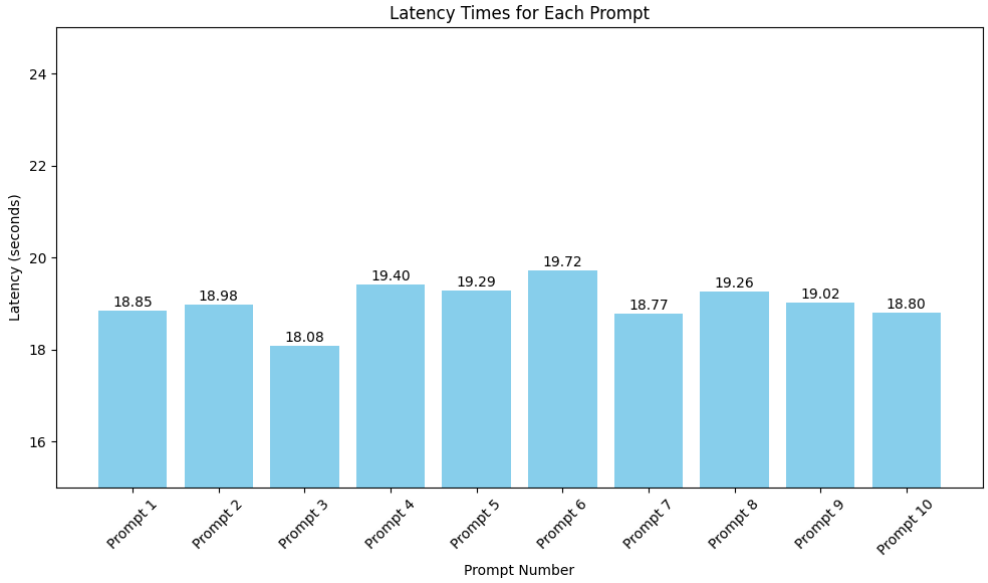

Figure 7. Latency values of the 10 prompts

The consistency in response time across multilingual inputs is a key factor in ensuring an effective real-time user experience, especially in healthcare consultations, where prompt responses are essential for providing timely assistance and maintaining user engagement. The findings suggest that, although latency variations exist, the chatbot performs well within acceptable limits, making it a viable solution for practical use in healthcare environments that require multilingual support.

Spectrogram Analysis of Model Voice Performance

The spectrograms reveal a range of frequency distributions across the 10 prompts, indicating the model's ability to handle diverse vocal tones and pitches (Figure 8). The time duration of the

prompts varies, suggesting adaptability to different speech rates. Intensity variations in the spectrograms demonstrate the model's capacity to process varying volume levels and emphasis. The spectral structure, characterized by patterns of dark and light bands, provides insights into the phonetic and prosodic features captured by the model.

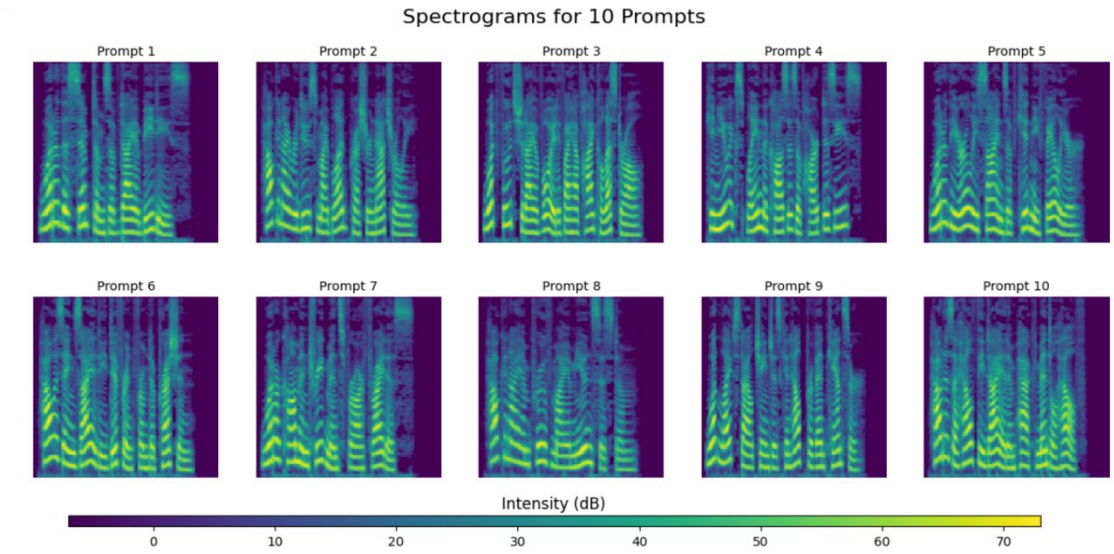

Figure 8. Spectrogram analysis of 10 prompts

To evaluate the model's performance quantitatively, we can employ metrics like Word Error Rate (WER) and Character Error Rate (CER) to assess speech recognition accuracy. WER is calculated as:

$$WER = (S + D + I) / N$$

where

S, D, and I represent the number of substitutions, deletions, and insertions, respectively

N is the total number of words in the reference transcription.

Similarly, CER is calculated as:

$$CER = (S + D + I) / N$$

where

N represents the total number of characters in the reference transcription.

Additionally, Mean Opinion Score (MOS) can be used to evaluate the perceived quality of synthesized speech.

The average intensity of the 10 prompts based on their spectrograms reveals a range between approximately 30-60 dB. Prompts 1, 2, 5, 6, 9, and 10 show predominant colors of green and blue, indicating lower average intensities of around 30-50 dB. On the other hand, Prompts 3, 4, 7, and 8 exhibit higher average intensities, with predominant colors of yellow and green, suggesting levels around 40-60 dB. These variations highlight that while some prompts maintain a lower intensity range, others peak at higher intensity levels, likely due to variations in content or delivery.

It is important to note that the average intensity is relatively low because the measurements include the gaps between the speech. Where no speech occurs, these gaps contribute to lower overall intensity readings, as the darker colors on the spectrograms represent them. This analysis helps in understanding the performance characteristics of the model's voice across different prompts, indicating

that certain prompts are more intense than others. Understanding these intensity levels is crucial for optimizing the model for consistent and balanced output in real-world applications.

DISCUSSION

This paper outlined a full methodology for building a multilingual health chatbot using fine-tuned large language models with text-to-speech capability. The model demonstrated that, in the real world, it could translate and respond to different healthcare questions across multiple languages, with fluent responses and producing speech to suit as many users as possible. The model showed promise during the testing sessions, where it used prompts revolving around general medical advice, that it could stand reliably and accurately across a multitude of languages and show its effectiveness at communicating health-related information with users.

The model, therefore, captured complex language variations and even healthcare-specific terminology, demonstrating that it can perform well in real-world diverse linguistic and semantic nuances associated with healthcare interactions. The translation process also proved highly efficient; hence, the users received their responses in their preferred language, an important factor in improving access, especially for multilingual users and those with restricted literacy in the dominant language. On the contrary, TTS functionality gave yet another layer of inclusions through auditory delivery; it would benefit users with preferences for auditory communication and even visual impairments.

However, there were some limitations according to the study, such as response latency and sometimes inaccurate translation about medically complex cases. Though very rare, these examples give reasons for periodic model fine-tuning and utilization of specific domain datasets toward higher precision. Future enhancements may bring real-time voice-to-text capabilities to seamlessly provide full conversational functionality that listens to both text inputs and responds accordingly, gradually filling gaps in accessibility. In addition, updating the medical knowledge base of the model through regular clinical data updates would be a good warranty of up-to-date information on healthcare coming from a chatbot.

CONCLUSION

This research paper addresses the design of a multilingual health assistance chatbot to increase access to vital health information across linguistic groups. Advanced natural language processing techniques, although fine-tuning the aboonaji/llama2finetune-v2 pre-trained model, have allowed the development of a sophisticated chatbot that may be quite capable of providing accurate medical guidance contextual within the topic under conversation. Utilizing the aboonaji/wiki_medical_terms_llam2format dataset on Hugging Face, the chatbot has been trained to answer a range of health-related questions with linguistic appropriateness and consistency. The case study shows how the chatbot responds to user queries about medical tips on various topics in different languages: English, Hindi, Spanish, and various languages, showing its effectiveness in breaking the language barrier in communicating healthcare matters. A comprehensive evaluation was performed using performance metrics such as BLEU scores to measure response accuracy, perplexity for language model efficiency, and spectrogram analysis for text-to-speech (TTS) performance. The results demonstrated the chatbot's ability to break language barriers, enhance health literacy, and equip users to better manage their health by delivering precise medical information in their native languages. However, several limitations were identified, including a response latency of nearly 20 seconds due to hardware constraints on personal GPUs. This issue can be resolved by using a better hardware backend during deployment. Additionally, there were minor inaccuracies in medical guidance, especially with complex queries. These issues primarily stem from the model's underlying architecture limitations and the need for further training on specialized, real-time medical data. Future work will include further study of how more complex features involving the integration of personalized health recommendations and real-time updating of medical guidelines can be built based on feedback. Continuous development

and improvement of the chatbot will prove crucial in maintaining its approach to the ever-changing healthcare environment and keeping it abreast as a source of information in an increasingly changing world. The work thus goes beyond technical advances in support of the principles of access and inclusion in healthcare, contributing towards improved health equity among diverse populations.

REFERENCES

Abdelkefi, M., & Kallel, I. (2016). Conversational agent for mobile-learning: A review and a proposal of a multilanguage text-to-speech agent, "MobiSpeech." *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, 1–6. https://doi.org/10.1109/RCIS.2016.7549294

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. *Proceedings of Machine Research, 70*, 195–204. https://proceedings.mlr.press/v70/arik17a.html

Badlani, S., Aditya, T., Dave, M., & Chaudhari, S. (2021). Multilingual healthcare chatbot using machine learning. *Proceedings of the 2nd International Conference for Emerging Technology, Belagavi, India*, 1–6. https://doi.org/10.1109/INCET51464.2021.9456304

Balne, C. C. S., Bhaduri, S., Roy, T., Jain, V., & Chadha, A. (2024). Parameter efficient fine-tuning: A comprehensive analysis across applications. *arXiv preprint arXiv:2404.13506* *https://doi.org/10.48550/arXiv.2404.13506*

Bickmore, T., & Cassell, J. (2005). *Social dialogue with embodied conversational agents*. In J. van Kuppevelt, L. Dybkjær, & N. Bernsen (Eds.), *Advances in natural multimodal dialogue systems* (pp. 23–54). Springer. https://doi.org/10.1007/1-4020-3933-6_2

Divi, C., Koss, R. G., Schmaltz, S. P., & Loeb, J. M. (2007). Language proficiency and adverse events in US hospitals: A pilot study. *International Journal for Quality in Health Care, 19*(2), 60–67. https://doi.org/10.1093/intqhc/mzl069

Guha, R. (2023, September 17). *Fine-tuning human for LLM projects*. https://doi.org/10.2139/ssrn.4574477

Health Systems Trust. (2022). *South African Health Review 2022*. Health Systems Trust. https://www.hst.org.za/publications/Pages/SAHR2022.aspx

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *Proceedings of the 31st European Conference on Cognitive Ergonomics* (pp. 207–214). Association for Computing Machinery. https://doi.org/10.1145/3335082.3335094

Huang, C.-Y., Yang, M.-C., Huang, C.-Y., Chen, Y.-J., Wu, M.-L., & Chen, K.-W. (2018, December). A Chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management*, *Bangkok, Thailand,* 1791–1795. https://doi.org/10.1109/IEEM.2018.8607399

Kapočiūtė-Dzikienė, J. (2020). A domain-specific generative chatbot trained from little data. *Applied Sciences, 10*(7), 2221. https://doi.org/10.3390/app10072221

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health, 18*(1), 271. https://doi.org/10.3390/ijerph18010271

Lu, J., Yu, L., Li, X., Yang, L., & Zuo, C. (2023, October). LLaMA-Reviewer: Advancing code review automation with large language models through parameter-efficient fine-tuning. *Proceedings of the IEEE 34th International Symposium on Software Reliability Engineering, Florence, Italy*, 647–658. https://doi.org/10.1109/ISSRE59848.2023.00026

Michaud, L. N. (2018). Observations of a new chatbot: Drawing conclusions from early interactions with users. *IT Professional, 20*(5), 40–47. https://doi.org/10.1109/MITP.2018.053891336

Pereira, J., & Díaz, Ó. (2019). Using health chatbots for behavior change: A mapping study. *Journal of Medical Systems, 43*, Article 135. https://doi.org/10.1007/s10916-019-1237-1

Sameer, M., Sudharsan, K., & Benazir Begum, A. (2024, April). Unforgettable password generation using LoRA fine-tuned large language model. *Proceedings of the International Conference on Advances in Data Engineering and Intelligent Computing Systems, Chennai, India*, 1–5. https://doi.org/10.1109/ADICS58448.2024.10533548

Tran, N., Tran, H., Nguyen, S., Nguyen, H., & Nguyen, T. (2019, May). Does BLEU score work for code migration? *Proceedings of the IEEE/ACM 27th International Conference on Program Comprehension*, *Montreal, QC, Canada*, 165–176. https://doi.org/10.1109/ICPC.2019.00034

U.S. Census Bureau. (2022, September 1). *Language use in the United States: 2019* (ACS-50). U.S. Department of Commerce. https://www.census.gov/library/publications/2022/acs/acs-50.html

Weizenbaum, J. (1966). *ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

World Health Organization. (2020). *Patient safety incident reporting and learning systems: Technical report and guidance.* World Health Organization. https://iris.who.int/bitstream/handle/10665/334323/9789240010338-eng.pdf

Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., & Huang, X. (2024). A comprehensive survey of large language models and multimodal large language models in medicine. *arXiv preprint arXiv:2405.08603.* https://doi.org/10.48550/arXiv.2405.08603

Xu, J., Zhang, H., Yang, Y., Cheng, Z., Lyu, J., Liu, B., Zhou, X., Yang, L., Bacchelli, A., Chiam, Y. K., Kian, T., & Chiew, T. K. (2024). *Investigating efficacy of perplexity in detecting LLM-generated code.* ArXiv. https://doi.org/10.48550/arXiv.2412.16525

Xu, L., Sanders, L., Li, K., & Chow, J. C. L. (2021). Chatbot for healthcare and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer, 7*(4), e27850. https://doi.org/10.2196/27850

AUTHORS

Vignesh U is an Assistant Professor Senior Grade 2 in the School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Chennai. Prior to this recent appointment, he was a Post-Doctoral Fellow at the National Institute of Technology (NIT), Trichy, India. Dr. Vignesh received his undergraduate degree in BTech (IT), his MTech (IT) from Anna University, Chennai, and his PhD in Computer Science and Engineering from VIT University, Chennai. Dr. Vignesh has published papers in preferred journals, patents, and book chapters and participated in various forums on computer science, social science, etc. He has also presented various academic and research-based papers at national and international conferences. His research activities are currently twofold: while the first is to explore the developmental role that society needs with technology such as Artificial Intelligence, the second major research theme he is pursuing focuses on bioinformatics and data mining.

Aman Amirneni is pursuing a Bachelor of Technology in Computer Science and Engineering at Vellore Institute of Technology (VIT), Chennai. His enthusiasm for artificial intelligence has led him to work actively in the field. This research builds upon his interest in exploring the applications of Large Language Models (LLMs) for medical-related real-world problems.