iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71640

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

AI-Driven Bilingual Voice Chatbot

S Manimala[1], Shashank Uppunda[2], Manoj Kumar N[3], B Charan Sai[4], Bhuvan R[5]
Department of CS&E, JSS Science and Technology University, JSS Technical Institutions Campus, Mysuru, 570006, Karnataka, India

Abstract: ***This project proposes an AI-driven, bilingual voice- enabled health chatbot that aims to enhance healthcare access in Karnataka's rural and semi-urban zones. The system was initially developed with a fine-tuned GPT-2 model; however, it generated unpredictable and sometimes unrelated answers, particularly for advanced or bilingual inputs. To address these constraints, the model was substituted with Mistral LLM through integration on LangChain, allowing Retrieval-Augmented Generation (RAG) to generate more precise and context-specific responses from a carefully curated health QA dataset. Supporting both Kannada and English languages, voice interaction in real time, and a friendly Gradio interface, the chatbot offers inclusive, voice-to- voice health support customized for those with low literacy and restricted digital exposure.***
Index Terms: GPT-2, Mixtral LLM, LangChain, Voice Chat- bot, Kannada NLP, Healthcare AI, Bilingual Interface, Gradio, Retrieval-Augmented Generation.

I. INTRODUCTION

While healthcare has been rapidly digitized, rural and semi-rural populations in Karnataka still face considerable impediments in getting access to medical advice because of language, literacy, and technology differences. While most health chatbots support English-literate users and are not available in Kannada language or voice interaction modes, disadvantaged groups like women, low-literacy users, and the elderly remain underserved. Classic rule-based systems compound this divide with stiff, impersonal exchanges that are insensitive to context and culture. To counter these limitations, we created an AI-based biilingual health chatbot supporting voice-first conversation in both English and Kannada. Early builds with a highly tuned GPT-2 model exposed shortcomings such as variable responses, hallucinated responses, and poor management of bilingual requests. Switching to Mix- tral (Mixtral-8x7B-Instruct) combined with LangChain and Retrieval-Augmented Generation (RAG) addressed these prob- lems by anchoring responses in a handcurated health dataset with FAISS vector search. Speech-to-speech translation in real time, speech recognition through the Google Speech API, and speech synthesis through gTTS allow seamless voice-to-voice communication, while a user-friendly Gradio interface is made available to meet the needs of users with minimal digital liter- acy. This strategy fills healthcare accessibility gaps by placing emphasis on support in Kannada language, guidance sensitive to culture, and a voice-native approach, enhancing access to trusted medical information for underprivileged communities in their native language.

II. LITERATURE REVIEW AND GAP IN THE LITERATURE

The rapid development of artificial intelligence and nat- ural language processing has had a profound effect on the healthcare industry, facilitating the creation of ever more advanced chatbot systems and automated question answering (QA) systems. The technologies hold the promise of increasing patient engagement, simplifying triage, and broadening access to health information, particularly in resource-limited environ- ments. Despite these developments, there are ongoing chal- lenges: most systems are limited in their linguistic and cultural versatility, support regional or low-resource languages infre- quently, and typically have weak voice interaction mechanisms or handling of the subtle, context-heavy questions common in actual healthcare applications. In addition, transparency, answer reliability, and source identification remain issues that erode user trust and impede clinical uptake of QA systems. These gaps call for health assistants based on state-of-the-art language models with bilingualism, inclusivity, transparency, and accessibility-driven research directions and innovations as detailed in subsequent sections.

A. Labrak et al. (2024) biomistral: a set of open-source pretrained large language models for medical domains

Labrak et al. introduced BioMistral [8], a set of open-source large language models (LLMs) specifically created for biomed- ical and healthcare purposes. They are founded on the Mistral architecture and additionally pre-trained on vast biomedical corpora, like PubMed Central. BioMistral is particularly strong in medical question answering, text classification, and medical language understanding tasks, and is built to be fine-tuned simply for a range of downstream medical NLP applications. The model performs well on medical benchmarks and is made available to researchers and practitioners alike, rendering it an invaluable asset to the development of medical AI.

Pros:

- Open-source and available, encouraging transparency and shared research.
- High medical question-answering and text- comprehension performance.
- Expertise in medical terminol- ogy, which makes it extremely suitable for professional health- care settings.
- Can be modified for other medical NLP tasks.

Cons:

- Limited Multilingual/Regional Language Support: Though strong in English, BioMistral is weak in supporting regional or low-resource languages such as Kannada. This restricts its usage in multilingual settings, particularly where English skills are low.
- No Voice Interaction: BioMistral is tailored for text-based tasks and lacks voice-based user interaction, which is highly important for accessibility among low-literacy or visually impaired users.

B. Mokmin Ibrahim (2021) The Evaluation Of Chatbot As A Tool For Health Literacy Edu- Cation Among Undergraduate Students

The paper [3] evaluates the impact of a health chatbot in enhancing health literacy among students at the undergrad- uate level. The chatbot is tested for its capacity to interact with users, provide health information, and improve learning outcomes. The paper proves that chatbots can be used as efficient means of health education, especially among youth and educated cohorts. Pros: • Demonstrates that chatbots can considerably improve health literacy and interaction. • High- lights the health education and awareness campaign potential of conversational AI. • Generates feedback on user experience and satisfaction with chatbot technology. Cons: • Population Limitation: The research sample consists of only undergrad- uates and does not take into account rural, low-literacy, or bilingual users who can have other needs and difficulties. • English-Centric, Text-Based: The chatbot is English-only and text-based, excluding users who need voice support or regional language option.

C. Vignesh Amirn (2025) Breaking Language Barriers In Healthcare: A Voice Activated Multilingual Health Assistant

A voice-activated health assistant is proposed in the paper [11] to enhance the accessibility of healthcare for non-English speakers. It is multilingual and uses voice interaction to make health information and services more accessible, especially to those with limited English proficiency. Pros: • explicitly addresses the requirement for both voice and multilingual functionality in healthcare assistants. • Supports better access for non-English speaking users. • Emphasizes the need for voice interfaces for inclusion and user experience. Cons: • Limited Regional Coverage: Although the system is multilin- gual, it does not strongly support certain regional languages like Kannada that are necessary for real inclusivity in regions like South India. • Lack of Source Attribution and Trans- parency: The system does not prioritize providing source- backed answers or transparent information, which are essential for building trust in healthcare applications.

D. Singhania ET AL. (2024) — Medibuddy: A Healthcare Chatbot using AI

Medibuddy [16] is a chatbot based on artificial intelligence intended to give users medical information and assistance. The system uses NLP to respond to an array of health-related questions and direct users to relevant resources. It is used in real-life environments, showing hands-on applicability and user interaction. Pros: • Real-life implementation with user feedback and actual deployment. • Able to respond to all types of health-related questions and situations. • Serves to show the viability of AI chatbots in healthcare. Cons: • English-Only, Text-Based: Medibuddy primarily supports English and text interaction, limiting accessibility for regional language speak- ers and those who require voice input/output. • No Source Attribution: The chatbot does not provide explicit references or transparency in its responses, which can undermine user trust and the reliability of the information provided.

E. Prashanth ET AL. (2023) AI Enabled Chatbot for COVID'19

This article [5] reports on a COVID-19 information dissemi- nation and triage AI-powered chatbot. The chatbot is capable of giving quick answers to public health questions, symptom checking, and helping the public navigate through the pan- demic. Pros: • Illustrates the potential of chatbots in public health crises to speed up information dissemination. • Can perform elementary triage and symptom checking, enhancing healthcare processes in times of crisis. • Indicates versatility of chatbot technology to meet pressing and changing healthcare demands. Cons: • Limited Voice and Language Support: The bot is not designed to be optimized for local languages or voice, which are crucial for universal communication among diverse groups. • Understanding in Context: The system can be challenged with sophisticated, context-rich questions typical of real-life healthcare, which restricts its capability for

handling difficult or individualized questions.

F. How Our Project overcomes these LIM

ITATIONS Although the chosen literature shows impres- sive strides in the advancement and deployment of health chatbots and medical language models, persistent gaps ex- ist—particularly in multilingual and regional language cover- age, voice-based accessibility, and transparency. The majority of current solutions are English-dominant and text-based, cutting out users who use regional languages such as Kannada or those who need voice interaction because of literacy or visual impairments. In addition, the absence of explicit source citation in answers reduces user confidence and the accuracy of health information. Our project is uniquely positioned to overcome such limitations. With a bilingual (Kannada- English), voice-based, retrieval-augmented health assistant, we make it possible for users with varying linguistic back- grounds—rural or low-resource settings included—to access accurate healthcare information. The assistant is both text and voice input/output, thus accessible to users of different literacy levels and physical capabilities. Additionally, through the provision of clear source-backed responses, our system facilitates greater transparency and trust, and more directly addresses transparency and reliability concerns mentioned previously. Overall, our project provides a more inclusive, accessible, and more trustworthy healthcare chatbot solution that is suited to the identified real-world needs and constraints in the existing literature.

III. METHODOLOGY

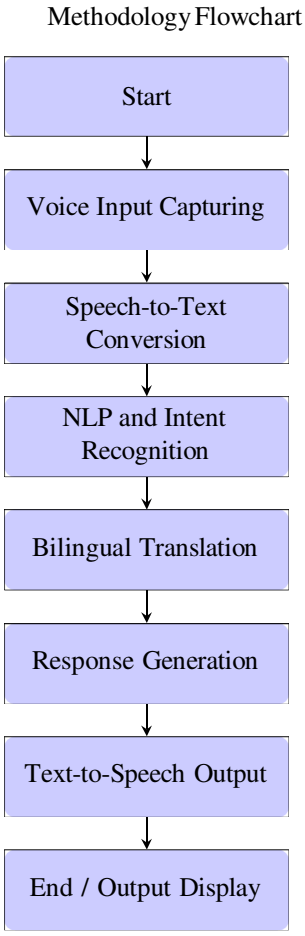


Fig. 3.1 . Ai-Driven health Companion Methodology Flowchart

The methodology for the research conducted is as shown in the [Fig.2.1]

A. Dataset Preparation

In summary, the dataset preparation phase of the AI- Driven Health Companion project set a unique baseline for producing accurate, context-based, and bilingual responses to user inquiries in a healthcare environment. The project started with the collection of consistent and reputable medical documents, in PDF format. These included frequently asked questions, symptom-treatment documents, and general health information, which we identified heavely and vetted in ad- vance, included in the FAISS knowledge database to categorize and identify level consumer health knowledge. These PDF files were extracted using automated PDF extraction libraries, PyPDFLoader and DirectoryLoader, which will help scale ingestion across defined data/ directory . The raw text extracted from the PDFs has no data structure on its own and so it needed to be broken into smaller text segments, that potentially overlap using RecursiveCharacterTextSplitter to maintain a stream of awareness across text segments; this was particularly critical for medically linked topics like symptoms, causes, and treatments. Each identified text segment was then con-verted to a semantic vector embedding, using the sentence- transformers/all-MiniLM-L6-v2 model, which is efficient and preserves relatively consistent retentivity of natural language. The generated embeddings were all stored in a FAISS vector database for identifiable recall based on user prompt similarity. This vector knowledge store was also the heart of the system's Retrieval-Augmented Generation (RAG) pipeline and this cre- ates accurate data organized to inform the Mistral Language model. An important focus was also placed on bilingual, and in this case, importantly building a knowledge back that deals with english and kannada context.

1) Medical Relevance and Accuracy

Selection of content that is medically accurate and domain- specific is the basis of any trustworthy healthcare AI system. As for this project, it was chosen that the verified health documents and QA-based PDFs should be opted for by the model to fetch trustworthy and clinically relevant information. This is a must in healthcare settings where even the slightest deviation might mean life or death. All sources must be from verified publications or medically approved guidelines. This means that the chatbot's information is factually grounded, offering a safe environment for the user to interact with.

2) Language Compatibility (Bilingual Support)

Because it is providing services that support both English and Kannada speakers, the dataset needs to support both languages in queries and responses. Specifically, the potential for curating bilingual medical content or converting existing English content into Kannada should utilize credible trans- lation engines/tools for the English –¿ Kannada process. It is recognized that translations in a medical context should preserve the term and context and that literal translations may result in misleading understanding. Proper alignment will result in the AI being able to respond naturally and accurately in both languages. Proper alignment will also enhance access for individuals in rural or regional areas.

3) Context Preservation in Chunking

In organizing documents into smaller data chunks for em- bedding and retrieval, text chunking arises. However, chunking text carelessly breaks semantic coherence; as in the case of health information, where it is important to keep symptoms and treatment options logically grouped. To get around this situation, recursive character-based chunking with overlaps is used to maintain context. Overlapping allows adoption of information from one chunk to another, reducing any loss of meaning. The system will provide complete and meaningful answers to queries.

4) Embedding Quality and Model Suitability

To embed text is to transform it into a numerical vector that can be digested by AI. Using an appropriate embedding model like sentence-transformers/all-MiniLM-L6-v2 will provide se- mantically meaningful representations of medical content in an efficient manner. The chosen embedding model also needs to strike a good fit with respect to performance and computational limits, especially in timing contexts. If the embedding model is not using quality techniques for embedding, subtle diagnostic differences are often missed or worsened in the medical phrasing or concept. Therefore, a reliable combination of quality embedding model and thorough pre-processing will improve the reliability of query matching.

5) Retrieval-Readiness and Clean Preprocessing

It is necessary to be cleaned and pre-processed in a partic- ular way. Clean data includes irrelevant navigational aspects such as page numbers, headers and footers, embellishments, and general formatting that may impact the quality of the embedding.

Clean data will also improve the user's semantic matching experience and ensure that the user does not receive junk content, or content that doesn't hold health and relevance. Just as it is essential to have clean data, there is a certain factor sanitizing and structuring will enable the users to extract from the FAISS vector store efficiently and optimally. Not only does preprocessing and cleaning productivity aspects play a role in retrieving relevant and relevant for their query; it can directly impact users' overall experience and reliability of the content.

B. System Architecture

The system is structured into separate, interacting modules, each having a particular role in the information processing chain. This layered, modular architecture facilitates maintain- ability, scalability, and future upgrades:

1) User Interface (using Gradio)

Acts as the main point of engagement, enabling users to enter queries through voice or text and obtain answers in both modes. The platform is intuitive and easy to use, with a user- friendly interface that caters to users who are at different levels of digital literacy.

2) Speech Recognition Module

Translates user input through speech into text. It caters to both English and Kannada, using strong algorithms to recog- nize variable accents and moderate ambient noise, producing accurate transcriptions.

3) 3 Language Detection Translation Module

Detects automatically the input text language. If a query is in Kannada, it gets translated into English for downstream processing. The translation system is semantic accuracyopti- mized, especially for health-related terminology.

4) Query Embedding FAISS Retrieval Module

Converts the (potentially translated) question into a vector representation with a sentence transformer. FAISS is then utilized to query a carefully curated medical QA database for the most contextually applicable information to guarantee that the system's answers are rooted in correct medical facts.

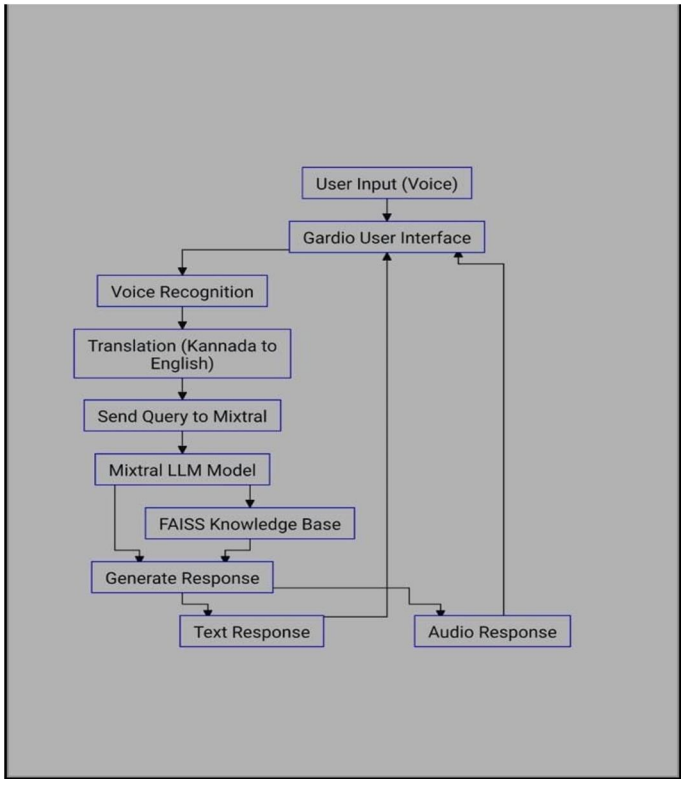


Fig. 3.2.1 . System Architecture

5) Mixtral LLM (via LangChain)

The Mixtral-8x7B-Instruct model, coordinated by LangChain, takes in the user's question and the context retrieved. It creates an accurate, context-sensitive answer appropriate for the user's query, tapping both the language model's reasoning and the knowledge base's factual material.

6) Back-Translation (if necessary)

If the user originally used Kannada, the English response obtained is translated back into Kannada so that the answer is returned in the user's chosen language.

7) Text-to-Speech (TTS) Module

Translates the final answer into natural speech with gTTS, allowing the information to be accessible to low literacy or visually impaired users. The audio and text answers are both displayed through the interface.

8) Knowledge Base (FAISS + Medical QA Dataset)

Stores pre-curated medical questions and answers in a vectorized manner, allowing for fast and accurate retrieval of useful information for a given user query.

C. Model Training

1) Training Setup

Data set was fetched with the help of Tensorflow and keras libraries. The training of the model was done in Google Colab, where a Tesla T4 GPU provided ample processing for the fine-tuning. The fine-tuning was carried out using the Hugging Face Transformers with PyTorch as the backend. The environment was set up with the necessary dependencies such as transformers, datasets, and tokenizers. A pre-trained GPT-2 was used as the base for transfer greatly reducing the training size or compute requirement.

2) Training process

The dataset was preprocessed and tokenized into prompt- response format using GPT-2's tokenizer. The model was fine- tuned using Adam optimizer with a learning rate of 5e-5, batch size of 4, and trained for 3 to 5 epochs. Cross-entropy loss was minimized to optimize response accuracy. Throughout the training, metrics such as loss and perplexity were monitored, and checkpoints were saved regularly to prevent data loss and evaluate performance at each stage.

D. Equations

1) Cosine Similarity

Cosine Similarity measures the semantic closeness between the chatbot's output and the reference answer, with values approaching 1 indicating a high degree of similarity.

$$\cos(\vartheta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \qquad (1)$$

where:

- **A** and **B** are the embedding vectors of the two text sequences (predicted and actual response).
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the vectors.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ and are the magnitudes (Euclidean norms) of the vectors.

2) F1 Score

F1 Score measures the balance between precision and recall in the chatbot's response compared to the reference answer. Values closer to 1 indicate a better overlap and more accurate retrieval of relevant information.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

where:

- $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
- $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$

3) BERT Score

BERT Score evaluates the semantic similarity between the chatbot's generated answer and the reference answer using deep contextual embeddings from the BERT model. Values closer to 1 indicate stronger semantic alignment, even if the exact wording is different.

$$\text{BERTScore}_{F_1} = 2 \times \frac{\text{Precision}_{BERT} \times \text{Recall}_{BERT}}{\text{Precision}_{BERT} + \text{Recall}_{BERT}} \quad (3)$$

where:

- Precision_{BERT} is the average similarity of predicted tokens to reference tokens.
- Recall_{BERT} is the average similarity of reference tokens to predicted tokens.
- These similarities are computed using cosine similarity between BERT embeddings of each token.

E. Model Evaluation

Post training, the model was evaluated against the test dataset.

1) Language Model Accuracy and Transition

Initially, the project utilized a fine-tuned GPT-2 model for answering medical questions. However, due to context retention issues, hallucination, and inaccuracies, the GPT-2 model was replaced with the Mixtral-8x7B-Instruct model. The system dramatically improved its accuracy and context- aware responses using the FAISS-based medical knowledge base with Hugging Face and LangChain's RetrievalQA.

2) Bilingual Response Accuracy

The pipeline was evaluated in two languages, English and Kannada. The responses were checked for accuracy and equiv- alency in translations. In all instances, the system recognized the voice input correctly and then translated the voice input to English. The English text was then processed through the QA chain which resulted in another text output. The output text was then translated back into Kannada when needed. The quality evaluations indicated the translated texts in Kannada preserved the intent of the medical purpose and accuracy, with very few instances of semantic error.

3) Results Analysis

In order to earnestly evaluate the performance of our health chatbot, we utilized three quantitative measures: Cosine Sim- ilarity, F1 Score, and BERT Score. These measures assess the extent of semantic and lexical correspondence between the chatbot's generated responses and the predetermined ground truth answers. In the subsequent sections, we will discuss each evaluation measure and present the corresponding results derived from our project. This method generates a complete picture of the accuracy of the chatbot, semantic usefulness, and overall performance in the provision of quality healthcare responses.

TABLE 3.1
TRAINING METRICS OVER EPOCHS

Epoch	Loss	Grad Norm	Learning Rate
1.54	2.8114	7.67	1.13e-5

1.55	2.8314	4.00	1.11e-5
1.56	2.8309	3.92	1.09e-5
1.57	2.8359	4.28	1.06e-5
1.58	2.7932	4.36	1.04e-5
1.59	2.7791	4.48	1.02e-5
1.60	2.7952	5.16	1.00e-5
1.61	2.8252	5.01	9.77e-6
1.62	2.7822	4.82	9.54e-6
1.63	2.7952	3.41	9.31e-6

TABLE 3.2
COSINE SIMILARITY AND CORRECTNESS

Cosine Similarity	Correct
0.7113	TRUE
0.8075	TRUE
0.7016	TRUE
0.7172	TRUE
0.7291	TRUE

The table 3.2 presents Cosine Similarity values for five test cases, each paired with a correctness label. In our evaluation, all Cosine Similarity values exceed 0.70, with the highest reaching 0.8075. Importantly, every response in this set is marked as ”TRUE” for correctness. This consistent pattern demonstrates that when Cosine Similarity surpasses the 0.70 threshold, the chatbot reliably produces factually accurate and contextually relevant answers. This metric thus serves as an effective initial filter for response validation in our system.

TABLE 3.3
FULL EVALUATION METRICS

Cosine Similarity	Correct	F1 Score	BERT Score
0.7113	True	0.4396	0.1424
0.7421	True	0.4925	0.2815
0.7093	True	0.4038	0.1075
0.7172	True	0.4235	0.0090
0.7204	True	0.4274	0.0981
0.6615	False	0.5321	0.1990

The table 3.3 expands the evaluation to include the F1 Score, which balances precision and recall to reflect how much of the reference answer is captured by the chatbot’s output. In our results, F1 Scores range from 0.4038 to 0.5321. Higher F1 Scores (above 0.49) suggest that the chatbot is capturing most of the relevant information, while lower scores indicate that some essential details may be missing or mismatched. The table shows that even when F1 Scores are relatively high, a response may still be marked as incorrect (as seen with the score of 0.5321), highlighting the importance of using F1 Score in conjunction with other metrics and manual review.

F. Model Saving and Deployment

1) Saving the Model

With this project, the fine-tuned GPT-2 model was saved to a dedicated host folder using the method of the Hugging Face. This actually saves the model weights, tokenizer, and config files so that any time, they can be brought up again for use. Save the model to the host folder so that while loading back it can be called without training it again, thus making it always at hand for integration into the Flask web app to deliver real-time responses.

2) *Loading the Model in the Web Application*

In the web-based application, the AI model runs using the HuggingFaceEndpoint class from the langchain huggingface library. The application securely connects to the Mixtral- 8x7B-Instruct model hosted on Hugging Face by layout the HUGGINGFACEHUB API TOKEN, which has been stored in the environment variable. Additionally, it configures the model with parameters, such as temperature and max token length, to control the response length and style. This demonstrates the application's design is done in such a way that every user query can be accommodated through a powerful, instruction- tuned language model. The actual loading of the model is done at startup, getting the system ready to engage users with the Gradio interface in real-time.

IV. COMPARATIVE ANALYSIS

A. *GPT-2 Implementation and Results*

In the first phase of the development process, we used the GPT-2 model for the implementation of the healthcare chatbot. The GPT-2 model is a widely used transformer-based model that is well recognized for its fluency in producing human-like text. When tested within the framework of medical question answering, however, some severe limitations were revealed.

1) Lack of Domain Accuracy

GPT-2 consistently tended to produce responses that were either too generic or factually incorrect. When asked about particular medical symptoms, treatments, or recommendations, the model typically fell back to general statements or, in a few instances, made-up information that was not based on any medical source. Such a lack of domain-specific accuracy is a highly reported limitation of general-purpose language models since they are not intrinsically endowed with current or domain-specific knowledge necessary for health care use. Consequently, users were unable to depend on the chatbot for accurate or clinically useful advice, which is crucial in health situations.

2) Irrelevance and Hallucination

By way of illustration of a recurrent issue, GPT-2 was found to generate false facts or sometimes even facts that were irrelevant but that glided through a reputable- sounding discourse pattern, a phenomenon called hal- lucination. In medicine, this is quite dangerous while false or inaccurate information could be saved by a user acting on faulty medical guidance. Despite attempts to fine-tune GPT-2 with a set of hand-curated medical information, the model failed to produce accurate and relevant responses consistently, brushing aside much of the nuance and complexity that medical queries entail.

3) No Strong Source Attribution

No strong knowledge base or source document was considered in the generation of any of its answers."While we took standardized medical datasets as the impetus for GPT-2's output, the model could not reliably anchor its responses in the given data. Consequently, answers tended to be detached from identifiable ties to credible sources, making it tricky for users to ascertain the accuracy or source of the information. This lack of transparency is a serious limitation in healthcare environments, where source-supported accuracy is critical to user confidence and safety. The failure to enable attribution of answers to authoritative sources further undermined confidence in the system, pointing to a major obstacle to responsible use of general-purpose language models such as GPT-2 in medical contexts.

4) User Experience

In general, the GPT-2-driven chatbot failed to instill user trust. Repeated errors, vagueness, and the impossibility of supporting multiple languages resulted in a less-than- ideal user experience. Users considered the system un- trustworthy and unreliable, which is in line with wider literature findings citing usability and accessibility prob- lems as primary impediments to successful uptake of healthcare chatbots.

Considering these limitations, GPT-2 was found unsuitable for release in a healthcare environment where accuracy, de- pendability, and bilingual support are not negotiable needs.

B. *Benefits of Mistral + LangChain*

To overcome the above-found limitations with GPT-2, we took up an even more sophisticated architecture based on the Mixtral-8x7B large language model (a variant of Mistral) combined with the LangChain framework and a retrieval- augmented generation (RAG) pipeline. This shift saw substan- tial gains across all essential dimensions.

1) Significant Accuracy Improvements

Mixtral-8x7B, as a bigger and instruction-fine-tuned model, showed significantly higher proficiency in medical questions. With LangChain's retrieval features added, the system provided answers that were accurate and contextually appropriate all the time, based on actual medical information.The improvement was most evident on advanced questions that required synthesis of in- formation from over one source or esoteric knowledge of medical nomenclature, addressing the root issue of domain accuracy within GPT-2.

2) Fact-Based, Reliable Answers

The addition of a RAG pipeline allowed the model to look up a compiled medical knowledge base for the applicable context before generating an answer. This reduced hallu- cinations by a large amount and ensured that responses were supported by actual source documents, thus making the chatbot more factually sound. Users can now be presented with answers that not only provide the response but also are verifiable against trusted sources, a core strength for healthcare use cases where disinformation can be deadly.

3) Transparency and Trust

With each response it produced, the system gave the original source documents or passages used. In addition to heightening user trust, this openness also enabled independent checks on the information presented. The transparent identification of sources constituted a major improvement over the GPT-2 deployment, putting the assistant in line with best practice in responsible AI for health and advancing the ethical mandate of transparency in the provision of medical information.

4) Improved User Experience

The enhanced chatbot had both text and voice in- put/output, gave precise and concise answers, and pre- sented a friendly interface. The feature of text-to-speech and source document display made the system more accessible to users with different literacy levels and requirements. These features made the system more in- clusive and pleasant to use, leading to greater engagement and trust across different user groups.

Integrating Mixtral-8x7B with LangChain and retrieval- augmented generation substantially enhanced the accuracy, dependability, and bilingual nature of the chatbot. By basing responses on a carefully curated medical knowledge base and offering source references, the system minimized hallucina- tions and maximized user trust. The improved support for both English and Kannada, as well as voice and text interfaces, increased the assistant's accessibility to more people. As a whole, this design presents a resilient, open, and user-friendly solution optimally appropriate for real-world healthcare uses.

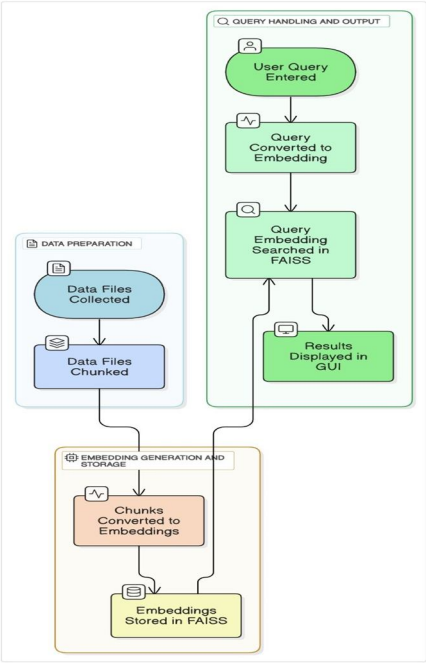


Fig. 4.2.1 . Design Overview For Lanchain

V. WEBSITE DEVELOPMENT

The shown frontend is implemented with the use of the Gradio framework, which makes web application inter- faces easy to interact with. This particular interface belongs to an AI-powered bilingual medical question-answer sys- tem—implemented as a health buddy chatbot.

A. Website Framework and Technologies

The AI-Driven Health Companion uses Gradio framework to build a fancy and interactive frontend, allowing for audio answers, language selection, and voice input. Python and Flask are used for backend development to manage user queries and model integration. MongoDB stores data related to user interactions. The core AI technologies used are Google's Cloud APIs (speech recognition and synthesis), translation APIs for supporting multiple languages, and GPT-2 through Hugging Face for response generation. spaCy is used by the system for natural language processing tasks.

B. Website Features

1) Voice Input Button (Microphone Integration)

This is usually a big prominent microphone button al- lowing users to speak about their symptoms or health- related question. The voice is recognized with voice input, converting it into text under the aids of speech recognition APIs, making the system entirely hands-free and thus, usable even by anyone who cannot or prefers not to type.

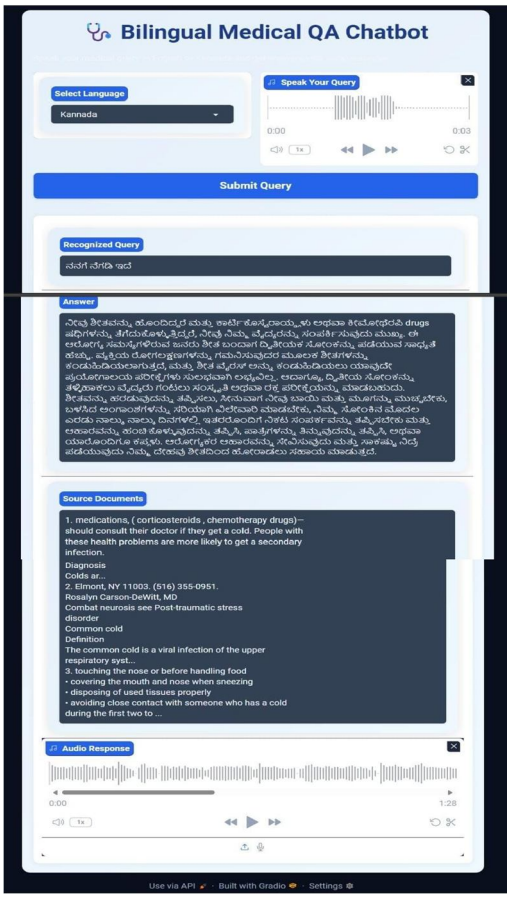


Fig. 5.1 . Website Screenshots

2) Language Selection Dropdown

Just in case a user wants their preferred language to be registered, there is an available dropdown on the frontend, thereby supporting bilingually. It is to ensure users do not face trouble communicating with the system due to linguistic barriers, hence making it inclusive and able to adapt to local contexts.

3) Chat Display Area (User Query and Response View)

This area displays the recognized user input while show- ing the translations in both original and translated ver- sions, along with AI-generated answers for healthcare- related queries. The clean, scrollable interface should allow users to see the whole conversation to follow through in clear steps, enabling them to look back at the previous suggestions and understand the health ideas offered easily.

4) Audio Playback of AI Responses

An audio player or playback button will enable the system's users to listen to the spoken responses. This becomes very useful for the elderly and visually impaired clients. The frontend integrates the text-to-speech output and offers controls to pause or replay the sound.

5) Responsive and Minimal UI Design

Frontend is clean, responsive, and easy to use, developed with framework such as Gradio. It facilitates seamless operation on multiple screen sizes and devices, with sim- ple navigation, legible fonts, and the necessary controls made easily accessible.

C. Website Enhancements

1) Improved Bilingual Support and Accuracy

Ensuring more precise translation and broadening the num- ber of regional languages supported could lend to better user experience. This entails, yet is not limited to, the integrated use of translation APIs to more accurately capture medical terminology, whereas the input side keeps clarity and context on its own or on the output side. Accurate translation, espe- cially for rural, non-English-speaking users, is imperative for any institution trying to offer a pronounceable, safe medical opinion.

VI. FUTURE SCOPE AND CONCLUSION

The development of the AI-Driven Health Companion has been a transformational process that helped us develop the technical expertise and user-centered design skills necessary to develop inclusive digital health solutions. Initially, we built our system using a fine-tuned GPT-2 model. However, we re- alized that the model had limitations in accuracy of responses and persisted contextual reliability. Therefore, we chose to switch to the Mixtral-8x7B-Instruct model with LangChain and a FAISS knowledge matrix. This enhancement enabled contextual, fact-based responses based on a curated medical QA dataset, which improved the capacity for delivery. In the process of the project, we had engaged exposure to speech recognition, translation, and text to speech technologies as we adapted Gradio to develop a multilingual and accessible user interface. There were challenges such as latency, inference, and multilingual integration but were able to put together a system that was both a user interface that demonstrated technical promise. In the future we aim to improve speed through an asynchronous processing system, keep our interface user-driven design, and increase our dataset with more diverse and domain-specific medical data. Additional goals include adding image-based QA using computer vision approaches, multi-turn conversational possibilities, clearer audio for diverse accents, secure data storage for user personalization, and full as a deployed interface.A variety of tools are introduced to extend the program's scalability, inclusiveness, and utility to a broader array of uses, notably within under-served zones.

REFERENCES

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners", Sensors, 2019. DOI:CorpusID:160025533

[2] Abacha, A. B., Mrabet, Y., Demner-Fushman, D. "A question- entailment approach to question answering", BMC Bioinformatics, 20(1), 1–11.Sensors, 2019. DOI:10.1186/s12859-019-3119-4

[3] Mokmin, N. A. M., Ibrahim, N. A. (2021). "The evaluation of chatbot as a tool for health literacy education among undergraduate students", Education and information technologies, 26(5), 6033–6049. Sensors, 2021. DOI:10.1007/s10639-021-10542-y

[4] Denecke, K., May, R., Rivera-Romero, O. (2024). "Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcom- ings and Risks", Journal of medical systems, 48(1), 23. Sensors, 2024. DOI:10.1007/s10916-024-02043-5

[5] Prashanth, Mallellu Reddy, P Swapna, Mudrakola. (2023). "AI En- abled Chat Bot for COVID'19", Sensors, 2023. DOI:10.1007/978-3-031- 27524-1$_6$8

[6] Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouche B. (2024). Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. Sensors, 2024. DOI:preprint/56930

[7] Ahan Bhatt, Nandan Vaghela. (2024). Med-Bot: An AI-Powered As- sistant to Provide Accurate and Reliable Medical Information, Sensors, 2024. DOI:10.48550/arXiv.2411.09648

[8] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gour- raud, Mickael Rouvier, Richard Dufour. (2024). BioMistral: A Collec- tion of Open-Source Pretrained Large Language Models for Medical Domains, Sensors, 2024. DOI:10.48550/arXiv.2402.10373

[9] Vishal Vinod, Susmit Agrawal, Vipul Gaurav, Pallavi R, Savita Choud- hary. (2021). Multilingual Medical Question Answering and Informa- tion Retrieval for Rural Health Intelligence Access, Sensors, 2021. DOI:10.48550/arXiv.2106.01251

[10] Qiming Bao, Lin Ni, Jiamou Liu. (2020). HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi- Directional Attention, Sensors, 2020. DOI:10.1145/3373017.3373049

[11] Vignesh U, Aman Amirneni. (2025). Breaking Language Barriers in Healthcare: A Voice Activated Multilingual Health Assistant, Sensors, 2025. DOI:10.28945/5455

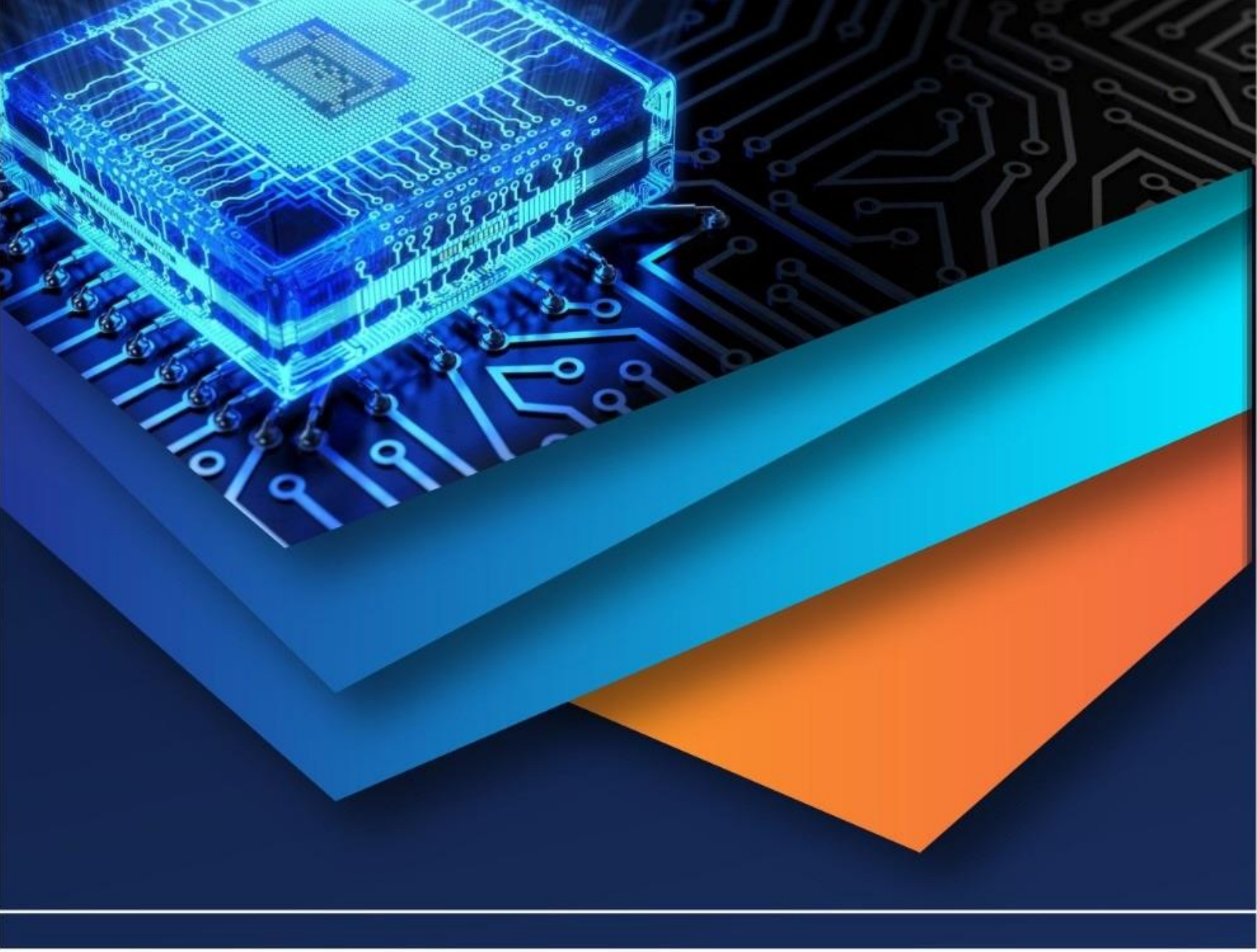
[12] Vince Bartle, Janice Lyu, Freesoul El Shabazz-Thompson, Yunmin Oh, Angela Anqi Chen, Yu-Jan Chang, Kenneth Holstein, Nicola Dell. (2022). "A Second Voice": Investigating Opportunities and Challenges for Interactive Voice Assistants to Support Home Health Aides, Sensors, 2022. DOI:10.1145/3491102.3517683

[13] Leibo Liu,Oscar Perez-Concha,Anthony Nguyen,Vicki Bennett,Louisa Jorm. (2022). De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models, Sensors, 2022. DOI:10.1016/j.jbi.2022.104215

[14] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, Dakuo Wang. (2024). Talk2Care: An LLM-based Voice Assistant for Communica- tion between Healthcare Providers and Older Adults, Sensors, 2024. DOI:10.1145/3659625

[15] Kowatsch, Tobiascc Nißen, Marcia Katharinacc Shih, Chen-Hsuan I. (2017). Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Con- trolled Trial on Childhood Obesity,, Sensors, 2017. DOI:10.3929/ethz- b-000218776

[16] Singhania, R., Badagan, S., Reddy, D., Sai Teja, K. T., Jett, C. (2024). Medibuddy – A healthcare chatbot using AI. International Journal of Soft Computing and Engineering (IJSCE), 14(3), Article G9902. Sensors, 2024. DOI:10.35940/ijsce.G9902.14030724



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY