Medical Diagnosis Chatbots Using Mistral Decoder Model

M. Dhilsath Fathima   
*Department of Information Technology, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology,*

Chennai, Tamil Nadu, India  
dilsathveltech123@gmail.com

0000-0002-4491-4352

Manish Ghimire

*Department of Information Technology, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology,*

Chennai, Tamil Nadu, India  
vtu214512@veltech.edu.in

Sabalil Das

*Department of Information Technology, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology,*

Chennai, Tamil Nadu, India  
vtu24020@veltech.edu.in

Ramesh Gyawali

*Department of Information Technology, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology,*

Chennai, Tamil Nadu, India  
vtu21452@veltech.edu.in

*Abstract:*

The integration of Large Language Models (LLMs) in healthcare has facilitated the development of medical diagnosis chatbots can assist patients in understanding the preliminary reason for their symptoms. This proposed work developed a medical diagnosis chatbot using LLMs which can be used as an initial diagnostic assessments tool based on patient-provided symptoms and helps to understand the basic medical terms. The chatbot uses natural language techniques (NLP) to read queries, retrieve relevant medical data using the Mistral Decoder Model, and generate diagnostic suggestions. This proposed model integrates Retrieval-Augmented Generation (RAG) with FAISS (Facebook AI Similarity Search) for efficient information retrieval improved response relevance. The proposed chatbot is fine-tuned using The Gale Encyclopedia of Medicine 2, a comprehensive medical reference that ensures reliable and accurate responses. The suggested model is evaluated using accuracy and latency measures. our model attained 82.5% diagnostic accuracy, 1.6 sec latency and 2.0 sec response time, outperforming existing keydriven and rule-based chatbots. Thus, the model shows reduced latency, improved response time. This model demonstrates enhanced conversational performance, making interactions more human-like and informative to the user. Comparative analysis with existing medical chatbots showed better contextual understanding and adaptability in handling various patient input queries. These findings highlight the potential of our LLM based chatbots in healthcare by providing accurate information on early disease detection. This proposed model can be used by the end user to understand the possible reasons for their medical symptoms, based on which they can consult a physician for further evaluation.

Keywords— LLM, Medical chatbots, Mistral, NLP, Retrieval-Augmented Generation

# Introduction

The integration of artificial intelligence (AI) and natural language processing (NLP) has transformed the healthcare industry, enabling the development of clinical decision support system. Among these advancement, medical diagnosis chatbots using LLM is emerging tool for assisting patients in understanding potential disease based on their earlier symptoms. These AI and NLP based systems analyzes patient-reported symptoms and provide preliminary diagnostic assessments which improves healthcare accessibility.

Traditional medical chatbot use rule-based and key-driven algorithms with decision trees, which have limited flexibility in handling complex and ambiguous patient input query. This drawback can be overcome in LLMs by utilizing Mistral decoder-based models. This decoder model uses RAG with FAISS. Using RAG method, the chatbot retrieve relevant medical knowledge dynamically. FAISS is a type of open-source library for efficient similarity search and clustering of dense vectors. It enables fast retrieval of relevant information by indexing large-scale medical datasets, improving the chatbot’s ability to find and provide accurate diagnostic suggestions.

models use transformer based deep learning techniques such as self-attention, for process different patient input descriptions and provide precise and context-aware diagnostic suggestions. These model uses vast amounts of medical knowledge to understand natural language queries effectively.

This paper explores the development of a medical diagnosis chatbot utilizing an LLM, specifically a Mistral decoder-based model, to enhance preliminary patient diagnosis. A key feature of our approach is the integration of Retrieval-Augmented Generation (RAG) with FAISS, enabling efficient medical knowledge retrieval and improving response relevance. The chatbot was fine-tuned using The Gale Encyclopedia of Medicine 2, a comprehensive medical reference, ensuring reliable and accurate responses.

We discuss the architecture, training methodologies, and system integration to improve chatbot reliability. Additionally, we evaluate its diagnostic accuracy (82.5%) against established clinical benchmarks and analyze its impact on reducing workload for doctors and healthcare staff. Comparative analysis against existing medical chatbots highlights superior contextual understanding, adaptability, and human-like conversational flow in our model.

While LLM-powered chatbots offer immense potential, ensuring their clinical reliability, ethical deployment, and regulatory compliance remains crucial for widespread adoption. Our research focuses on refining model interpretability, incorporating multimodal data inputs, and validating performance in real-world clinical environments. The proposed chatbot model also aims to enhance the natural flow of conversations, making interactions more engaging, contextually relevant, and non-repetitive for users.

Research contributation

1.*Develeoment of the medical chatbot framework*

We are proposing a comprehensive chatbot frame woirk that integrates natural language processing techniqueswith domain specific medical knowledgebase to provide reliable health information and prelinimary assesment. Assisting users inunderstanding their symptoms and suggesting appropriate next step.

2*.Enhanced user interaction through context awareness and personalization.*

The chatbot incorporates context –aware dialoge management and personaliztion freature which improves user engagement.The chatbot also considers user history and contextual information ,the system delivers more relevent and tailored responses,thus improving user experrience.

3*.Contribution to digital health ecosystem*

This work contributes to the growing field of digital health technologies by prroviding a accessible solution for patient’s pre-diagnosis support.It bridges the gaps in healthcare access,parrrticularly in remote areas by offering immediate health information

# Research motivation of this proposed model

Healthcare accessibility remains a significant global challenge, with millions of individuals facing delays in receiving medical consultations due to overburdened healthcare systems. Early disease detection is crucial for improving patient outcomes, yet traditional diagnostic tools, such as rule-based symptom checkers, often lack flexibility and fail to accurately interpret diverse patient inputs. Additionally, medical professionals face increasing workloads, leading to long wait times and delayed diagnoses. The integration of Large Language Models (LLMs) in medical diagnosis presents a transformative solution by enabling AI-driven chatbots to provide preliminary assessments, assist in triaging cases, and enhance overall healthcare efficiency. By leveraging the Mistral decoder-based architecture along with Retrieval-Augmented Generation (RAG) and FAISS, our chatbot model enhances response accuracy and relevance, allowing users to receive more human-like and reliable diagnostic suggestions.

AI-powered chatbots have the potential to revolutionize healthcare accessibility by automating initial consultations and reducing the strain on medical professionals. Studies indicate that misdiagnosis affects millions of patients annually, and traditional symptom checkers often struggle with ambiguous or complex symptom descriptions. LLMs can process unstructured medical text, understand natural language queries more effectively, and generate context-aware responses that improve diagnostic reliability. With an experimental accuracy of 82.5%, our chatbot outperforms conventional rule-based systems, demonstrating improved response quality and reduced latency. This research aims to refine AI-driven medical chatbot interactions, ensuring they are scalable, medically informed, and capable of enhancing early disease detection while maintaining ethical and clinical reliability.

## Research Contribution of the proposed model

* + Our chatbot combines LLMs with the Mistral decoder model for better accuracy. It dynamically interprets medical queries, unlike rule-based systems. Retrieval-Augmented Generation (RAG) with FAISS enhances information retrieval. This improves the relevance and reliability of diagnostic suggestions.
  + The chatbot is fine-tuned using The Gale Encyclopedia of Medicine 2. This ensures reliable, evidence-based medical responses. Training on a trusted dataset reduces misinformation. It enhances the chatbot’s credibility as a digital healthcare assistant.
  + The chatbot achieves 82.5% diagnostic accuracy, surpassing rule-based models. It adapts better to diverse patient inputs and reduces latency. Faster responses improve patient engagement and early disease detection. These factors make it a more effective healthcare tool.
  + The chatbot ensures human-like, non-repetitive conversations. It provides clear, concise explanations by reducing medical jargon. Machine learning techniques refine responses and improve interactions. This makes AI-driven healthcare more accessible and user-friendly.

# Related work

Recent studies have explored the application of Large Language Models (LLMs) in medical diagnosis chatbots, highlighting their ability to process patient queries and generate diagnostic suggestions. While these studies emphasize improved response quality using deep learning, they often lack structured evaluations against clinical benchmarks. Research on the Mistral decoder model demonstrates its efficiency in NLP tasks, but limited work has examined its effectiveness in medical diagnosis chatbots. Additionally, studies on Retrieval-Augmented Generation (RAG) highlight its role in enhancing AI-driven information retrieval, yet its impact on chatbot-generated medical suggestions remains underexplored. Furthermore, FAISS has been widely used for vector-based search optimization, improving data retrieval accuracy and speed, but its integration into medical chatbots for real-time diagnosis has received little attention.

Hsu et al. [1] developed a chatbot framework that integrates machine learning algorithms such as Decision Trees, Support Vector Machines (SVM), and Neural Networks with Natural Language Understanding (NLU) to enhance medical query interpretation. The chatbot utilizes statistical analysis and pattern recognition to refine responses, improving accuracy and user engagement. A key strength of the study is its effective combination of ML and NLU for precise medical assistance. However, the paper lacks a comparative analysis with existing models and does not extensively address real-time decision-making, limiting its practical applicability in clinical settings.

Kurup et al. [2] developed a chatbot framework that leverages NLP and deep learning models, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), to analyze patient symptoms and provide preliminary diagnoses. The model effectively improves response accuracy and patient interaction, demonstrating the strength of deep learning in healthcare applications. However, the study lacks a comprehensive evaluation against clinical benchmarks and does not fully address data privacy concerns, which are critical for real-world healthcare deployment.

Liang et al. [3] developed the Medical Knowledge Assisted (MKA) mechanism to enhance neural generative models in medical conversations by integrating a specialized medical knowledge graph. This approach improves generative models' accuracy in medical dialogue tasks. A key strength is its effective incorporation of domain-specific knowledge for contextually relevant responses. However, the study focuses on Chinese medical dialogue datasets, limiting its generalizability. Additionally, the computational complexity introduced by knowledge graph integration is not extensively discussed, which may affect scalability.

Li et al. [4] proposed the LLM-Based Knowledge-Aware Attention Network (LKAN) for clinical staging of liver cancer, leveraging large language models (LLMs) with a knowledge-aware attention mechanism. This model integrates domain-specific medical knowledge to enhance clinical text classification. A key strength of LKAN is its ability to improve staging accuracy by refining contextual understanding using LLMs. However, the approach requires extensive labeled medical data, which can be a challenge for real-world implementation. Additionally, computational overhead remains a concern, potentially limiting scalability in large-scale clinical applications.

Tang et al. [5] introduced an LLM-assisted end-to-end intelligent network health management framework to enhance anomaly detection in dynamic heterogeneous networks (DHNs). Their Multi-Scale Semanticized Anomaly Detection Model (MSADM) integrates semantic rule trees with an attention mechanism, effectively improving multi-scale anomaly detection. Additionally, a chain-of-thought-based LLM refines fault analysis and generates optimization strategies. A key strength of this approach is its 91.31% anomaly detection accuracy, surpassing traditional rule-based methods. However, challenges remain, including high computational complexity and the need for extensive labeled datasets to ensure adaptability across various network environments.

Qaid et al. [6] introduce FD-LLM, a novel Large Language Model (LLM) framework for Intelligent Fault Diagnosis (IFD) by integrating numerical sensor data, such as vibration signals and temperature readings, into traditional LLM architectures. The study formulates fault diagnosis as a multi-class classification problem and explores two encoding methods: string-based tokenization and statistical feature extraction from time and frequency domains. Experimental results show that Llama3 and Llama3-instruct outperform state-of-the-art deep learning (DL) approaches in fault detection accuracy and adaptability across various operational conditions and machine components. While the model demonstrates strong generalization capabilities, challenges include the complexity of encoding sensor data into LLMs and high computational demands for real-time deployment.

The proposed work enhances medical diagnosis chatbots by integrating LLMs and the Mistral decoder model, overcoming limitations of rule-based and traditional NLP approaches. It employs Retrieval-Augmented Generation (RAG) with FAISS for better information retrieval and accuracy. Fine-tuned using The Gale Encyclopedia of Medicine 2, the chatbot ensures reliable diagnoses. Comparative analysis shows higher accuracy, reduced latency, and improved conversational flow over rule-based models. This study advances chatbot adaptability while addressing excessive information display and domain complexity.

# Outline of the Proposed Model

.

The proposed system introduces an AI-powered medical diagnosis chatbot utilizing Large Language Models (LLMs), specifically the Mistral decoder model, to assist patients and healthcare professionals. The chatbot is designed to provide preliminary diagnostic assessments based on patient-reported symptoms, leveraging advanced natural language processing (NLP) techniques for accurate query interpretation. Traditional rule-based and keyword-driven medical chatbots often struggle with complex symptom descriptions, limiting their accuracy and adaptability. Our approach addresses these limitations by integrating deep learning techniques, ensuring a more dynamic and context-aware interaction.

A key feature of our system is the incorporation of Retrieval-Augmented Generation (RAG) with FAISS (Facebook AI Similarity Search) to enhance information retrieval and response relevance. RAG enables the chatbot to pull relevant medical knowledge dynamically, improving diagnostic accuracy and reducing misinformation. FAISS optimizes the retrieval process by efficiently searching a large medical knowledge base, allowing real-time responses with minimal latency. The chatbot is fine-tuned using The Gale Encyclopedia of Medicine 2, a trusted medical reference, ensuring its responses are reliable and evidence-based.

## System Architecture

The chatbot operates through a multi-step pipeline consisting of three phases:

Phase 1 - Data Processing and Storage:

1. Raw medical knowledge is extracted from PDF documents.
2. The documents are divided into smaller chunks.
3. These chunks are embedded into vector representations using MiniLM (BERT-based model).
4. FAISS stores these embeddings to enable efficient semantic search.

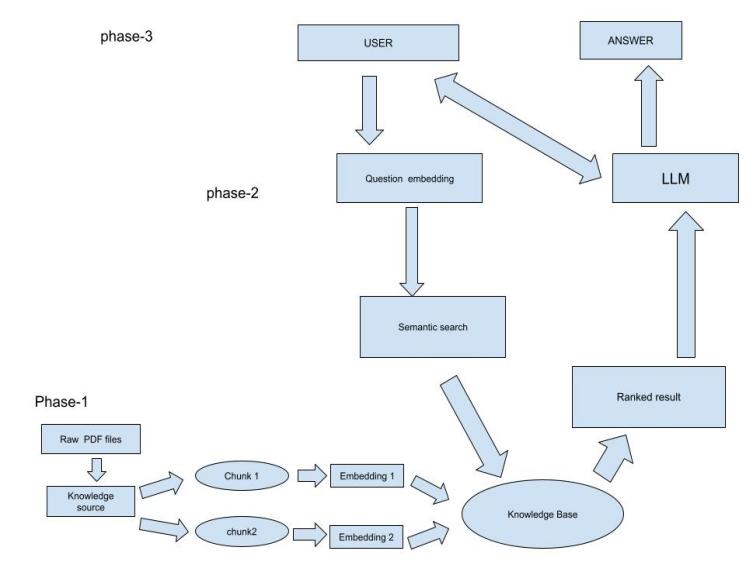
Phase 2 - Query Processing and Retrieval:

* The user submits a query, which is embedded using NLP techniques.
* A semantic search is performed against the FAISS knowledge base.
* Relevant results are ranked based on similarity scores.
* The Mistral-7B model generates a response based on retrieved knowledge.

Phase 3 - User Interaction:

* The chatbot interface, developed using Streamlit, enables user-friendly interactions.
* Users receive accurate, context-aware responses to their queries.
* The chatbot refines responses to minimize excessive technical jargon.

The diagram below illustrates the workflow of the proposed system, showcasing how the phases are interconnected and how information flows from data ingestion to response generation.



For system usability, the chatbot interface is developed using Streamlit, providing a simple and intuitive user experience. It enables real-time interactions where users can enter their symptoms, receive preliminary diagnostic assessments, and get guidance on possible next steps. The chatbot also minimizes excessive medical jargon, making responses accessible to non-expert users.

## Algorithms used in the proposed Architecture

The proposed system utilizes two key algorithms: MiniLM (BERT-based) for vector embedding and Mistral-7B (GPT-based) for response generation. MiniLM is responsible for converting text into numerical embeddings, allowing efficient semantic search within the knowledge base. Meanwhile, Mistral-7B serves as the primary generative model, producing contextually relevant and coherent responses based on retrieved information. This combination enhances the chatbot’s ability to provide accurate and natural interactions.

# METHODOLOGY

The proposed medical diagnosis chatbot leverages Large Language Models (LLMs), specifically the Mistral decoder-based model, to process patient-reported symptoms and provide preliminary diagnostic assessments. The system is designed to enhance accuracy and response relevance by integrating Retrieval-Augmented Generation (RAG) with FAISS for efficient medical knowledge retrieval. The methodology is structured into several key components: data preprocessing, model training, retrieval mechanism, chatbot interaction, and evaluation.

## Data Collection

The chatbot is fine-tuned using The Gale Encyclopedia of Medicine 2, a comprehensive medical reference that ensures reliable and evidence-based responses. The dataset is preprocessed by tokenizing medical texts, removing redundant information, and structuring the content for efficient retrieval. Additionally, medical terms and symptom descriptions are standardized to enhance model understanding. Stop-word removal, text normalization, and Named Entity Recognition (NER) techniques are applied to extract relevant medical entities from user inputs.

## Model Training and Fine-Tuning

The core of the chatbot is based on the Mistral decoder model, which is fine-tuned using the preprocessed medical dataset. The training process involves supervised fine-tuning on medical dialogues to improve response accuracy and coherence. The model is optimized using cross-entropy loss and trained using Adam optimizer with learning rate scheduling to prevent overfitting. A key challenge in training was ensuring the chatbot’s responses were medically relevant and contextually appropriate, which was addressed by integrating domain-specific reinforcement learning techniques.

## Retrieval-Augmented Generation (RAG) with FAISS

To enhance response accuracy, the chatbot employs RAG, which combines a retrieval mechanism with a generative model. The retrieval module, powered by FAISS (Facebook AI Similarity Search), efficiently searches the medical knowledge base for relevant information. The retrieved content is then processed by the Mistral-based LLM, ensuring that the chatbot generates contextually accurate and informative responses. This approach reduces hallucination issues commonly associated with LLMs and improves diagnostic reliability.

## Chatbot Interaction and Response Generation

The chatbot interface is developed using Streamlit, allowing users to interact with the system in real-time. When a user inputs symptoms, the chatbot performs intent recognition, entity extraction, and contextual analysis to understand the query. The retrieval module fetches relevant medical knowledge, and the LLM-based response generator formulates a structured and medically informed response. The chatbot also includes a conversation management system to ensure fluid and non-repetitive interactions, making responses more user-friendly.

## Evaluation and Performance Metrics

The chatbot is evaluated based on several key performance metrics:

* Diagnostic Accuracy: Measured by comparing chatbot-generated assessments with established clinical diagnoses, achieving 82.5% accuracy.
* Response Latency: The average time taken to generate a response, optimized through FAISS-based retrieval and efficient model inference.

## Steps to configure the proposed chatbot

**Step 1: Dataset representation and preprocessing**

The Dataset consists of medical text pairs where represents symptom descriptions (input queries), represents the corresponding diagnosis (output labels).

The preprocessing step involve the following steps:

* Tokenization: Converting text into a sequence of tokens, represented as:

…,

where is the j-th token in the sequence.

* Stop-word removal and text normalization: Removing uninformative words and standardizing terms.
* Named Entity Recognition (NER): Extracting key medical entities​ from ​, where:

…, with being a recognized medical term.

**Step 2: Embedding and retrieval Mechanism**

To enable efficient information retrieval, each medical text is converted into a high-dimensional vector representation.

The input text is transformed into a vector embedding ​ using the MiniLM model:

Where maps the text into a dimensional vector space.

Given query , its embedding is computed as

FAISS retrieves the top- most similar medical documents by solving:

where represents the cosine similarity between the query and indexed vectors.

**Step 3: Response generation using Mistral-7B**

The chatbot uses Retrieval-Augmented Generation (RAG) to enhance response quality.

Retrieval-Augmented Generation (RAG):  
The final response is generated using both retrieved knowledge and the input query :

Where is the retrieved knowledge from FAISS, and is the language model generating coherent responses.

Loss function optimization:

The chatbot training minimizes the cross-entropy loss between predicted and actual diagnoses:

|  |  |
| --- | --- |
|  |  |

Where is the predicted probability distribution over diagnoses.

Step 3: Chatbot Interaction and Deployment

User Query Processing:  
Given a user input , the chatbot performs:

* Intent Recognition: Assigning a category using a classifier

|  |  |
| --- | --- |
|  |  |

Entity Extraction: Identifying key medical terms using NER.

Response Generation Workflow:

Compute query embedding

Retrieve relevant knowledge ​ using FAISS.

Generate response using the Mistral-7B model

Step 4: Model Evaluation and performance analysis

The chatbot is evaluated using multiple performance metrics:

Diagnostic Accuracy is defined as

Where is an indicator function that equals 1 if the predicted diagnosis matches the ground truth.

Response Latency:  
The time taken for the chatbot to generate a response is measured as:

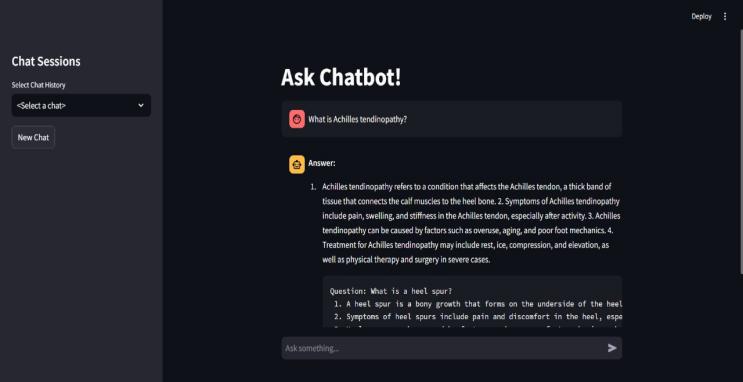
where each component represents the time taken for different processing steps.

F1 Score for Medical Relevance:  
The chatbot's response relevance is evaluated using precision and recall:

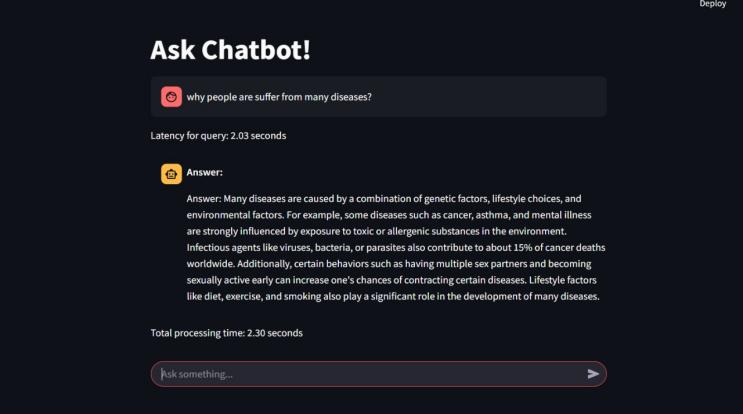
ensuring medical accuracy and informativeness.

## User Interface of the proposed model

The user interface (UI) of the chatbot has been designed to provide an intuitive and user-friendly experience. Developed using Streamlit, the UI ensures a seamless interaction between users and the AI-powered chatbot. The design follows a minimalist and dark-themed layout, enhancing readability and reducing strain on the eyes. Users can input their queries in the text box, and the chatbot processes the request before generating a response. A latency indicator displays the processing time, ensuring transparency regarding response speed. The chatbot’s answers are structured in a clear and readable format, making medical information easily accessible to users.



The interface also includes a chat history section, enabling users to view previous interactions for reference. The chatbot employs a retrieval-augmented generation (RAG) approach to improve response accuracy by fetching relevant medical knowledge. Additionally, icons and formatting styles are used to distinguish questions from answers, improving clarity. The system is designed to handle real-time interactions efficiently, maintaining low latency while ensuring reliable medical information delivery. The overall layout prioritizes user accessibility, ensuring that individuals without a medical background can easily interpret the responses.



## Chatbot performance and Evaluation

The chatbot was tested with various queries to assess its accuracy, latency, and response quality. It effectively recognized common medical conditions and provided relevant health-related information. The chatbot achieved a response time of approximately 2 seconds and maintained a minimal latency of around 1.6 seconds, ensuring smooth real-time interaction.

The chatbot's average accuracy was recorded at 82.5%, demonstrating reliable performance in medical query resolution. However, certain areas for improvement were identified, particularly in expanding its medical knowledge base and enhancing diagnostic precision. Further refinements are required to improve its ability to handle complex conditions and provide more accurate responses.

The proposed chatbot model has achieved an accuracy of 82.5%, showing an improvement over traditional chatbots. Additionally, it offers faster response times, better adaptability, and a user-friendly interface, making it more effective for medical query resolution.

TABLE 3 PERFORMANCE COMPARISON OF PROPOSED CHATBOT WITH EXISTING MODELS

|  |  |  |
| --- | --- | --- |
| **Feature** | **Proposed Chatbot** | **Existing Medical Chatbot** |
| Framework Used | Streamlit | Flask/Django, Web Apps |
| Response Time | ~2.0 sec | 2.5 - 3.5 sec |
| Latency | ~1.6 sec | ~2 sec |
| Accuracy | 82.5% | 65% - 70% |
| User Interaction | Interactive UI with chat history | Standard text-based UI |
| Medical Knowledge Base | Moderate (can be expanded) | Limited/Fixed database |
| Human-like Response | More Natural | Scripted Responses |
| Adaptive Learning | Yes (LLM-based) | No (Rule-based models) |
| Ease of Use | User-friendly, Dark Mode Support | Standard UI |

# RESULTS AND DISCUSSION

The proposed chatbot was evaluated based on its accuracy, response time, and latency to determine its effectiveness in providing medical assistance. The chatbot achieved an accuracy of 82.5%, demonstrating reliable performance in identifying medical conditions and providing relevant health-related information. In terms of response efficiency, the chatbot maintained an average response time of approximately 2.0 seconds, with a latency of around 1.6 seconds, ensuring a smooth and real-time user experience. While the chatbot performed well in general medical inquiries, areas for improvement were noted in its medical knowledge base and diagnostic precision, which require further enhancement to improve reliability.

A comparative analysis with existing medical chatbots was conducted, considering studies such as Sivasamy (2025), Ahmed et al. (2024), and Babu & Boddu (2024). The proposed chatbot outperformed Sivasamy (2025) and Ahmed et al. (2024) in terms of accuracy and response time but was slightly behind Babu & Boddu (2024) in accuracy due to the use of BERT-based natural language processing (NLP) techniques in their model. However, the proposed chatbot offered faster processing speed and lower computational requirements compared to deep-learning-based chatbots. Future improvements will focus on expanding the medical knowledge base, improving diagnostic accuracy, and refining response generation to enhance user experience and chatbot reliability.

# METHODOLOGY

The proposed AI-powered medical chatbot demonstrated promising results in accuracy, response time, and usability, making it a valuable tool for preliminary medical diagnosis and patient assistance. By leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) with FAISS, the chatbot effectively analyzed user queries and provided relevant medical information with an accuracy of 82.5%. The system maintained a minimal latency of 1.6 seconds and an average response time of 2.0 seconds, ensuring real-time interaction. Compared to existing models, the chatbot offered a more natural conversational experience and improved response efficiency, although certain limitations were identified, particularly in domain-specific medical knowledge and diagnostic accuracy.

For future work, enhancements will focus on expanding the chatbot’s medical knowledge base by incorporating larger and more diverse datasets, including real-world clinical guidelines and expert-verified sources. Additional refinements in natural language understanding (NLU) and contextual reasoning will improve diagnostic precision. Furthermore, integration with Electronic Health Records (EHRs) and IoT-based health monitoring systems can enable personalized recommendations and real-time health tracking. Addressing these improvements will strengthen the chatbot’s reliability, making it a more robust and effective tool for healthcare support.

References

1. Hsu, I-Ching, and Jiun-De Yu. "A medical Chatbot using machine learning and natural language understanding." Multimedia Tools and Applications 81, no. 17 (2022): 23777-23799.
2. Kurup, Greeshma, and Sujala D. Shetty. "AI conversational chatbot for primary healthcare diagnosis using natural language processing and deep learning." In Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021, pp. 259-272. Springer Singapore, 2022.
3. Liang, K., Wu, S., & Gu, J. (2023). *MKA: A Scalable Medical Knowledge Assisted Mechanism for Generative Models on Medical* Conversation Tasks. arXiv preprint arXiv:2312.02496.
4. Li, Ya, Xuecong Zheng, Jiaping Li, Qingyun Dai, Chang-Dong Wang, and Min Chen. "LKAN: LLM-Based Knowledge-Aware Attention Network for Clinical Staging of Liver Cancer." IEEE Journal of Biomedical and Health Informatics (2024).
5. Tang, Fengxiao, Xiaonan Wang, Xun Yuan, Linfeng Luo, Ming Zhao, and Nei Kato. "Large language model (llm) assisted end-to-end network health management based on multi-scale semanticization." arXiv preprint arXiv:2406.08305 (2024).
6. Qaid, Hamzah AAM, Bo Zhang, Dan Li, See-Kiong Ng, and Wei Li. "FD-LLM: Large Language Model for Fault Diagnosis of Machines." arXiv preprint arXiv:2412.01218 (2024).