Medical Diagnosis Chatbots Using Mistral Decoder Model

M. Dhilsath Fathima[1], Sabalil Das[2], Ramesh Gyawali[3] and Manish Ghimire[4]
{dilsathveltech123@gmail.com[1], vtu24020@veltech.edu.in[2], vtu21452@veltech.edu.in[3], vtu214512@veltech.edu.in[4]}

Department of Information Technology, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[1, 2, 3, 4]

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 key driven 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.

1 Introduction

The change on the horizon in healthcare industry is the way AI (Artificial Intelligence) and NLP (Natural Language Processing) have been integrated to develop clinical decision support systems. One such medical diagnosis chatbot which is coming to existence using LLMs (Large Language Models) and can help out with the patients in predicting diseases based on symptoms he/she had faced earlier (Hsu et al., 2022) [1]. These AI/NLP–based systems are used to analyze symptoms reported by patients, leading to early diagnostic outcomes thereby enhancing healthcare penetration (Kurup & Shetty 2022) [2].

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

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357844

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.

1.1 Research contribution

1.1.1 Development of the medical chatbot framework

We are proposing a comprehensive chatbot frame work that integrates natural language processing techniques with domain specific medical knowledgebase to provide reliable health information and preliminary assessment. Assisting users in understanding their symptoms and suggesting appropriate next step.

1.1.2 Enhanced user interaction through context awareness and personalization

The chatbot incorporates context aware dialog management and personalization feature which improves user engagement. The chatbot also considers user history and contextual information, the system delivers more relevant and tailored responses, thus improving user experience.

1.1.3 Contribution to digital health ecosystem

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

2 Research Motivation of this Proposed Model

Globally healthcare is still a big challenge, with millions waiting for too long for their turn to see a doctor due to overburdened health systems. The early detection of diseases can mean the difference between life and death, but diagnostic algorithms (whether for reading a piece of data or interpreting hundreds of millions from a patient) are rigid in operation (Hsu et al., 2022) [1]. Further the population demands more health care services making doctors work harder, cause increased waiting time and probable delay in diagnosis (Kurup & Shetty, 2022) [2]. Large Language Models (LLMs) will be a game changer in the medical diagnosis space by deploying AI chatbots to conduct preliminary assessments, for triaging situations more efficiently and effectively (Liang et al. 2023) [3]. With the chatbot model and the Mistral decoder-based architecture, it allows more accurate and relevant responses through Retrieval-Augmented Generation (RAG) as well as using FAISS to offer diagnostic suggestions that are human-like and believable.

After all, AI-driven chatbots present an opportunity for streamlined healthcare access that, if implemented at scale, could quite literally reshape how patients access care and operate as a filter through which to direct patients to the most appropriate healthcare professional. Misdiagnosis is believed to impact hundreds of thousands of patients each year with some estimates estimating the number as over a million patients and traditional symptom checkers often cannot provide appropriate diagnoses using highly ambiguous or complicated symptom descriptions. LLMs can convert unstructured medical text to structured data [4] and may help in better understanding of natural language queries [5], ultimately resulting in diagnosing more reliable because their answer generation is conditioned with context. Our chatbot had an experimental accuracy of 82.5%, with performance significantly better than conventional rule-based systems, of higher response quality and lower latency [6]. The goal of this study is to algorithmically create scalable, medically accurate and clinically meaningful health chatting bot responses that are able to promote early disease detection in an ethical and objective way.

2.1 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.

3 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] invented a chatbot framework which merges support from machine learning Decision Trees with SVM and Neural Networks and Natural Language Understanding (NLU) for refining medical query understanding. The chatbot performs statistical analysis and pattern recognition functions that enable better responses for increased accuracy and user interaction. The research demonstrates its strength through an optimal integration of ML and NLU which delivers precise medical assistance. The research fails to conduct a model comparison with existing systems and does not provide thorough information about real-time decision capabilities which impairs its usefulness in medical environments.

A chatbot framework created by Kurup et al. [2] combines NLP with deep learning models specifically RNN and LSTM to evaluate patient symptoms which leads to preliminary diagnostic results. The deep learning framework enhances patient responses along with interactions while proving its effectiveness in healthcare settings. The research fails to demonstrate clinical performance benchmarks and does not resolve essential data privacy issues that would be necessary for healthcare system deployment in real-world settings

Liang et al. [3] created the Medical Knowledge Assisted (MKA) mechanism which uses specialized medical knowledge graphs to improve neural generative models during medical conversations. The implementation of this method leads to better accuracy performance from generative models when processing medical dialogues. The system proves successful at implementing medical knowledge systems for producing context-appropriate responses. The research analyzes Chinese medical dialogue datasets which reduces its applicability to other medical settings. The study fails to thoroughly address the computational complexity challenges that arise from integrating knowledge graphs even though this could impact scalability in the system.

Li et al. [4] developed the LLM-Based Knowledge-Aware Attention Network (LKAN) which applies large language models (LLMs) with a knowledge-aware attention mechanism for clinical staging of liver cancer. Specialized medical data enhances the performance of the model when categorizing clinical texts. The main advantage of LKAN emerges from its LLM-based contextual refinement abilities which enhance staging accuracy. The method's deployment encounters difficulties because it needs large amounts of medical data that need to be properly

labeled. The system encounters obstacles with its resource consumption needs which limit its potential for growth across extensive medical facilities.

The framework presented by Tang et al. [5] employs LLM assistance to create an end-to-end intelligent network health management system for dynamic heterogeneous networks (DHNs) anomaly detection enhancement. MSAD an anomaly detection model that combines semantic rule trees with an attention mechanism to enhance multi-scale anomaly detection capabilities. The chain-of-thought-based LLM enhances both fault analysis and produces optimization strategies. The anomaly detection accuracy of 91.31% stands as a major benefit of this approach because it exceeds conventional rule-based detection methods. The system encounters two major challenges because it demands costly calculations in addition to requiring numerous labeled examples to operate effectively across different network conditions.

The research by Qaid et al [6] presents FD-LLM as a new Large Language Model (LLM) framework which combines numerical sensor information with standard LLM structures for Intelligent Fault Diagnosis systems. The research transforms fault diagnosis into a multi-class classification task while evaluating two encoding approaches that involve string-based tokenization and statistical feature analysis from time and frequency domains. The experimental findings demonstrate that Llama3 and Llama3-instruct yield superior fault detection results in comparison to current deep learning (DL) approaches regardless of operational conditions and machinery components. Real-time implementation of the model encounters two primary challenges from needing strong computational resources along with complex data processing requirements for LLM structure creation.

This research develops a medical diagnosis chatbot system through the integration of LLMs and Mistral decoder model to surpass traditional rule-based and NLP diagnostic methods. The system implements Retrieval-Augmented Generation (RAG) together with FAISS to enhance both information retrieval and accuracy levels. The chatbot depends on The Gale Encyclopaedia of Medicine 2 for its fine-tuning process to generate dependable medical diagnoses. According to analysis LLM-based models produce more precise outputs at faster rates with enhanced natural dialogue capabilities than traditional rule-based systems. The research improves chatbot flexibility through solutions for both information display issues and domain complexity problems.

4 Outline of the Proposed Model

The AI-powered medical diagnosis chatbot system implements Mistral decoder model within its Large Language Models (LLMs) for supporting healthcare providers and patients. The system utilizes NLP technology to process patient symptoms which enables it to generate diagnostic assessments through its designed approach. The utilization of rules and keywords leads medical chatbots to provide less accurate and flexible services during complex symptom description. Deep learning methods integrated within our solution improve context-based interactions which resolve the mentioned restrictions.

Our system includes Retrieval-Augmented Generation (RAG) together with FAISS (Facebook AI Similarity Search) which improves both information retrieval efficiency and response relevance. The RAG system allows the chatbot to dynamically retrieve medical knowledge which improves diagnosis precision and minimizes inaccurate information. The medical knowledge base retrieval process receives optimization through FAISS which enables quick and almost instantaneous responses with low latency values. The chatbot receives its fine-tuning through The Gale

Encyclopaedia of Medicine 2 which serves as a trusted medical reference for delivering evidence-based and reliable responses.

4.1 System Architecture

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

Phase 1 - Data Processing and Storage:

1. The extraction process starts with medical knowledge retrieval from PDF documents.
2. The documents receive segmentation into smaller parts.
3. MiniLM (BERT-based model) embeds the chunks into vector representations for storage.
4. The FAISS system stores the embedded data for facilitating quick semantic search operations.

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 produces its response by accessing stored knowledge.

Phase 3 - User Interaction:

- The Streamlit-based interface of the chatbot system provides users with an easy-to-use interface for interaction.
- Users obtain precise answers that consider their query context through the system.
- The chatbot refines responses to minimize excessive technical jargon.

The workflow of the proposed system is shown in the fig 1 diagram below which illustrates how phases are interconnected and how information flows from intake of data to creation of instance response.

The chatbot interface is developed using Streamlit for system usability to have a simple and intuitive user experience. Real time interactions where symptoms are entered, an initial diagnostic assessment is run and users are guided on the next steps. It also reduces the amount of excessive medical jargon in the responses, so that the responses are accessible to non-expert users.

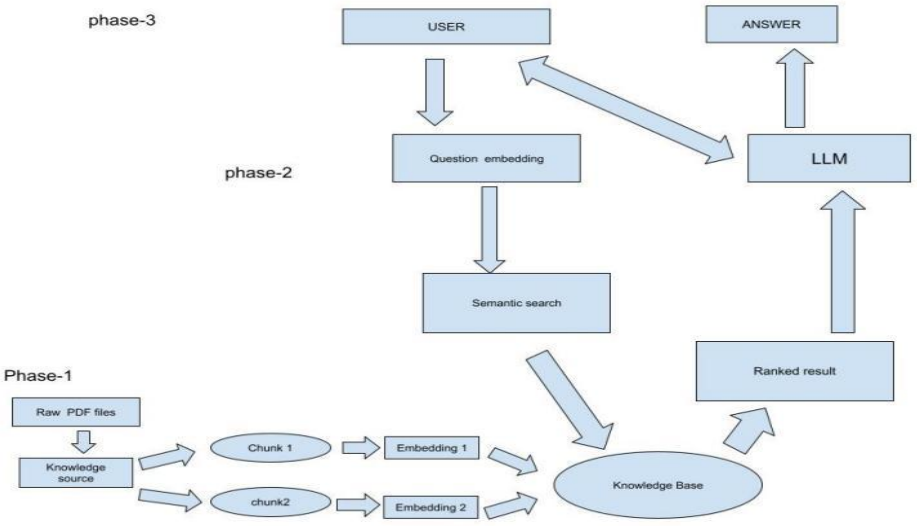

Fig. 1. The workflow of the proposed system.

4.2 Algorithms used in the proposed Architecture

This work proposes two key algorithms, MiniLM (a BERT-based vector embedding), and Mistral-7B (a GPT-based response generation), and shows that they can achieve good performances on tasks that combine whether questions are answered by Yes or No. The text is converted into numerical embeddings by MiniLM, which can then perform efficient semantic search within the knowledge base. On the other hand, Mistral-7B is the main generative model which produces contextually relevant and coherent responses from retrieved information. This combination allows the chatbot to have more accurate and natural interactions.

5 Methodology

The medical diagnosis chatbot utilizes Mistral decoder-based models from the Large Language Models family to analyze symptoms submitted by patients for initial diagnostic evaluation. The system combines Retrieval-Augmented Generation (RAG) and FAISS to enhance accuracy and response relevance through efficient medical knowledge retrieval. The system follows a methodology divided into four main sections including data preprocessing and model training followed by retrieval mechanism and chatbot interaction and evaluation.

5.1 Data Collection

The chatbot utilizes The Gale Encyclopedia of Medicine 2 as its reference material for delivering evidence-based and reliable responses. The medical information goes through preprocessing steps that include tokenization of texts along with redundant information removal and content organization for efficient retrieval. Standardization of medical terminology together with symptom descriptions exists to improve the model's comprehension abilities. The

system implements Stop-word removal together with text normalization and Named Entity Recognition (NER) methods to extract necessary medical entities from user-submitted content.

5.2 Model Training and Fine-Tuning

The Mistral decoder model serves as the foundation of the chatbot system while using the preprocessed medical dataset for its fine-tuning process. The training process executes supervised fine-tuning on medical dialogues to enhance both response accuracy and coherence. The model receives optimization through cross-entropy loss while Adam optimizer with learning rate scheduling ensures it avoids overfitting. The main training difficulty involved making sure the chatbot delivered medically valid responses within proper context and this issue was solved through reinforcement learning methods tailored for medical domains.

5.3 Retrieval-Augmented Generation (RAG) with FAISS

Response accuracy of the chatbot increases through the implementation of RAG that unites retrieval capabilities with generative model functions. FAISS (Facebook AI Similarity Search) drives the retrieval module which enables efficient searching of medical knowledge bases for relevant information. The LLM based on Mistral processes retrieved content to produce contextually sound and informative responses. The implementation of this method helps LLMs avoid hallucinations while enhancing their diagnostic accuracy.

5.4 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 symptom, 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.

5.5 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.

5.6 Steps to configure the proposed chatbot

Step 1: Dataset representation and preprocessing

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

The preprocessing step involve the following steps:

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

$$T(X_i) = \{t_1, t_2, \ldots., t_n\} \quad (1)$$

where $t_j$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 E_i from X_i, where:

$$E_i = \{e_1, e_2, \dots, e_n\} \quad (2)$$

with e_j 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 X_i is transformed into a vector embedding V_i using the MiniLM model:

$$V_i = f_{MiniLM}(X_i) \in R^d \quad (3)$$

Where f_{MiniLM} maps the text into a d dimensional vector space.

Given query Q, its embedding V_Q is computed as

$$V_Q = f_{MiniLM}(Q) \in R^d \quad (4)$$

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

$$argmax \cos(V_Q, V_i) \quad (5)$$

where $\cos(V_Q, V_i)$ 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 $R$is generated using both retrieved knowledge D_k, and the input query Q:

$$R = f_{Mistral-7B}(Q, D_k) \quad (6)$$

Where D_k is the retrieved knowledge from FAISS, and $f_{Mistral-7B}$ is the language model generating coherent responses?

Loss function optimization:

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

$$R(Y, \overline{Y}) = \sum_{i=1}^{N} Y_i \log \overline{Yi} \quad (7)$$

Where $\overline{Yi}$ is the predicted probability distribution over diagnoses?

Step 3: Chatbot Interaction and Deployment

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

- Intent Recognition: Assigning a category C using a classifier

$$C = argmaxP(C|Q) \quad (8)$$

Entity Extraction: Identifying key medical terms using NER.

Response Generation Workflow:

Compute query embedding V_Q

Retrieve relevant knowledge D_k using FAISS.

Generate response R 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

$$Accuracy = \frac{\sum_{i=1}^{N} 1\,(Y_i = \overline{Yi}}{N} \quad (9)$$

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

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

$$T = T_{embedding} + T_{retrievel} + T_{generation} \quad (10)$$

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:

$$F1 = \frac{2*precision*Recall}{Precision+Recall} \quad (11)$$

ensuring medical accuracy and informativeness.

5.7 User Interface of the proposed model

The chatbot interface uses a design that lets users interact easily with the system. The Streamlit development platform enabled the creation of a user interface that achieves smooth communication between AI-powered chatbot users. A dark-themed minimalistic design approach makes the interface both reading-friendly and eye-comfortable. The text box allows users to submit their queries which the chatbot processes to deliver a response. Users can see the processing time through an indicator which shows the system response speed. Users can easily access medical information through the chatbot because its responses use a structured presentation format that is easy to read. Fig 2 shows Streamlit-Based Medical Chatbot Interface Architecture.

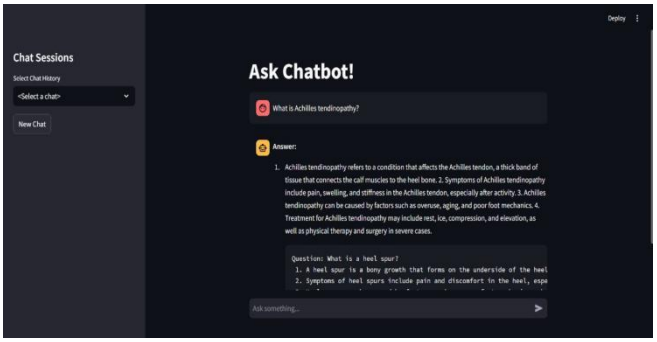

Fig. 2. Streamlit-Based Medical Chatbot Interface Architecture.

Users can access previous interactions through the chat history section that is part of the interface. The RAG approach incorporated by the chatbot enables it to retrieve medical knowledge for enhanced response accuracy. The interface uses icons together with formatting styles to separate questions from answers which enhances readability. The system operates with real-time handling capabilities to deliver reliable information at optimal speed. The layout design puts user accessibility first to enable non-medical users to understand the responses without difficulty. Fig 3 shows User Interface Design of a Streamlit-Based Medical Chatbot with RAG Integration.

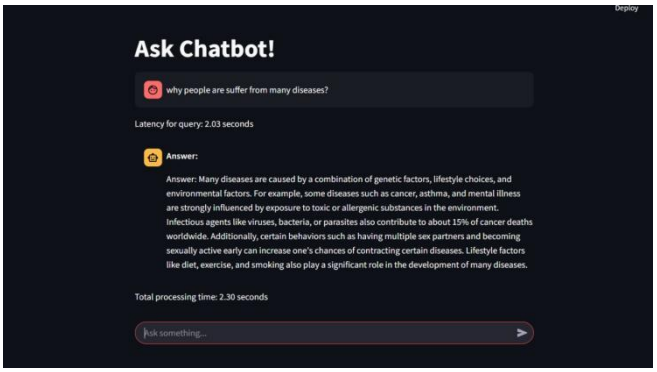

Fig. 3. User Interface Design of a Streamlit-Based Medical Chatbot with RAG Integration.

5.8 Chatbot performance and Evaluation

The system underwent testing through multiple queries to assess its accuracy levels and response time performance as well as information quality standards. Standard medical problems were successfully identified by the system which then provided appropriate health-related information. All requests processed by the bot took no longer than 2 seconds before providing immediate service that maintained delays under 1.6 seconds.

The medical queries processed by the chatbot achieved an 82.5% success rate in resolution. The system demonstrates consistent performance in medical question answering tasks. The system needed better medical information for processing and improved diagnostic capabilities to function properly. The system requires additional optimization changes to achieve better diagnostic outcomes in advance patient situations. Table 1 shows Performance Comparison of Proposed Chatbot with Existing Models.

The created chatbot system outperforms traditional models because it reaches 82.5% accuracy in its operational tasks. Users benefit from an advanced interface and quick responses and adaptable features that improve medical query responses.

Table 1. 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

6 Results and Discussion

Our investigation includes a comparison of the accuracy of the proposed chatbot system in delivering medical assistance as well as response and waiting times. The test success demonstrates 82.5% meaning the system can give healthcare information and diagnosis. The response time for the system was 2.0 seconds on average, and a delay of 1.6 seconds for the response gave both medical users and chatbot interactants as if they had chatted online naturally

with each other. The medical questions part of the chatbot was functioning well however its medical data and diagnostic tools require further work to make them more reliable in updated releases.

The suggested chatbot system processed tasks faster and needed less computing power when compared to deep-learning-based chatbots. Future developers need to update medical records and disease detection systems plus chatbot response logic to build higher-quality user-friendly systems.

7 Conclusion

The AI healthcare solution demonstrated excellent results across several evaluation metrics thus making it an important tool for both medical diagnosis initiation and patient care assistance. Large Language Models and Retrieval-Augmented Generation with FAISS enabled efficient processing time which resulted in a user accuracy rate of 82.5%. The system delivered its fastest response within 1.6 seconds yet generated typical responses at 2.0 seconds thus allowing quick patient interaction. The new chatbot system improved user interaction through natural dialogue although it faced limitations regarding medical scope and diagnosis problems.

Future development efforts will focus on expanding the medical database of the chatbot through extensive medical guidelines verified by medical professionals. Medical diagnosis quality will advance as NLU and contextual reasoning systems develop better capabilities. Through the integration of Patient Electronic Health Records with Internet of Things-based health monitoring systems the chatbot maintains constant communication of individual patient care guidelines and health measurement updates. The healthcare support tool will become more reliable after receiving enhancements that convert it into an effective system.

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).