IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DR.AI: A MULTIMODAL RAG-ENHANCED HEALTHCARE CHATBOT FOR MULTILINGUAL DISEASE AWARENESS, EMERGENCY DETECTION AND ACCESSIBLE MEDICAL GUIDANCE

Dr. Sandeep Kulkarni[1], Nikhil Mishra[2], Utkarsh Yadav[3], Raman Sharma[3]

Assistant Professor, Department of Computer Science, Pune, Maharashtra

B.Tech,Student[2], Department of Computer Science

B.Tech, Student[3], Department of Computer Science

B.Tech, Student[4], Department of Computer Science

Ajeenkya DY Patil University

Lohegaon, Airport Rd, Charholi Budruk, Pune, Maharashtra

Abstract: This paper presents Dr.AI, a comprehensive AI-driven multimodal healthcare chatbot platform designed to democratize medical guidance through advanced natural language processing, Retrieval-Augmented Generation (RAG), and real-time emergency response capabilities. Dr.AI integrates Meta's llama-4-scout-17b-16e-instruct language model accessed via the Groq LPU inference engine, a curated ChromaDB vector store of 3,760 medical knowledge chunks, OpenAI Whisper for speech-to-text transcription, and Google Text-to-Speech (gTTS) for audio output. The system supports over 16 languages including Hindi, Hinglish, Marathi, Tamil, and Telugu, enabling voice-based and text-based medical consultation across diverse linguistic populations. Key features include a 5-step AI symptom checker, emergency SOS with real-time GPS hospital finder, medicine reminder scheduling, family health profile management, AI-powered medical report analysis for X-rays and lab results, webcam image capture, and a downloadable PDF health report. The platform is deployed on Render (backend) and Vercel (frontend) with Firebase Firestore for persistent data storage. Experimental results demonstrate that RAG-enhanced responses achieve approximately 85–90% factual grounding accuracy compared to 65–70% without RAG on domain-specific medical queries. The paper describes the system architecture, RAG pipeline design, deployment infrastructure, multilingual capabilities, and emergency detection mechanisms, establishing Dr.AI as a scalable, accessible, and clinically aware AI health companion.

Keywords - speech recognition, Whisper AI, Llama 4 Scout, medical image analysis, text-to-speech, gTTS, disease detection

I. INTRODUCTION

A chatbot is an artificial intelligence software that can be use by the user for conversation through website, messaging application or mobile apps. It responds to our health problem using multimodal LLM. Fig. (1) shows internal working of chatbot

1.1 Evolution of chatbot

The Turing test was developed in 1950s which tests whether the computer provides response as human being. In 1966, first chatbot ELIZA was developed by Joseph Weizenbaum which stimulate psychotherapist using pattern matching and substitution rules, without real understanding. Later in 1972, PARRY created by Kenneth Colby which stimulates a patient with paranoid schizophrenia using complex rule-based logic. These chatbots solely rely on if-else conditions, keywords and predefines scripts.

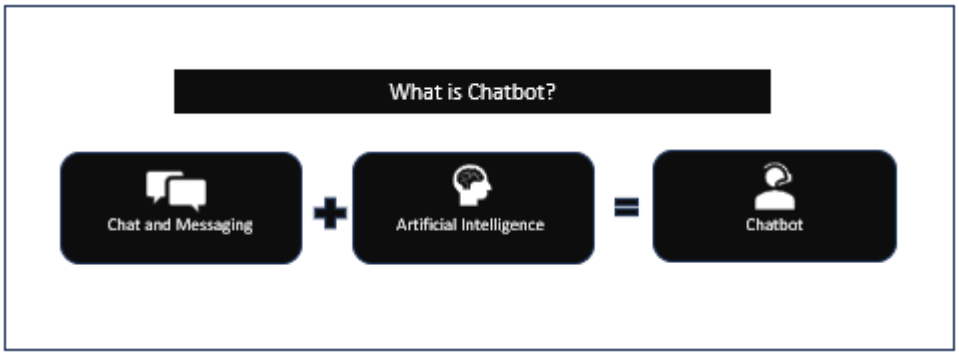

Fig. 1. Introduction to chatbot

1.2 Importance of Chatbot

A chatbot is often described as one of the most advanced and promising expressions of interaction between humans and machines Interaction between human and machines marks the advancement of technology in the form of a chatbot. Chatbots are applied in health education, diagnostics and mental state. A survey of conversational agents from 40 articles outlines chatbot taxonomy, specifies the main challenges and defines the types and contexts related to chatbots in health [1].

1.3 Working of chatbot

The primary task of chatbot is to respond to the users input which is “User Response Analysis”. It extracts the key information and relevant entities from the user message. Then the chatbot uses this information to provide response to user’s message using NLP. Chatbots are used everywhere- messaging apps, healthcare, politics, customer service and many other applications

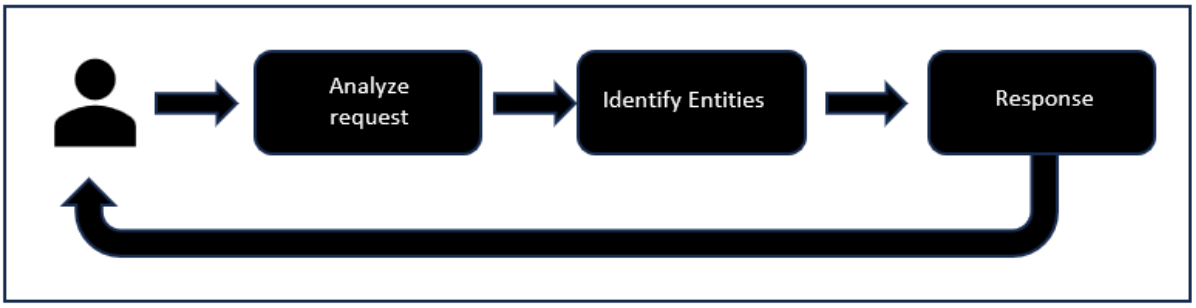

fig. 2. Architecture of chatbot

It is a powerful Al software, and its applications streamline interactions between people and services, enhancing customer experience. At an equivalent time, they provide corporations to enhance the client's engagement method and operational potency by reducing the standard price of customer service. Although no systematic review of chatbots for lifestyle modification programs has been revealed, there are several reviews on chatbots covering health care problems starting from mental health support and smoking cessation to sickness identification [2].

1.4 Process

The proposed idea is to create chatbot using multimodal LLM to diagnose the disease and provide basic details about the disease before consulting the doctor. The chatbot bot provides voice, text and image assistance. It provides user-friendly interface to communicate. The bot will provide which type of disease user has along with some precautions and clears all doubts of user. The user can avail the benefit because it can diagnose all types of disease and provide information. The system application uses users voice, text and image to diagnose disease.

Chatbot is developed to reduce the cost of health care and time of user because it is not possible to consult doctor anytime.

1.5 Multimodal Large Language Models in Heath

Medicine has always been a multimodal science. Clinicians commonly process information of heterogeneous data types - text, medical images, audio recordings, lab results, and electronic health records - but until recently, AI systems could only process one type of data at a time [14]. With the launch of GPT-4 in March 2023, this aspect had finally been turned: this is the first widely deployed LLM that can concurrently process both text and images, making it possible to create a new category of AI-based tools that can reason based on various types of inputs [15]. The transition outlined by Mesko (2023) was described as a crucial advancement in the development of AI in healthcare, in which M-LLMs are capable of breaking language barriers, reasoning about the visuals to make a diagnosis, and communicating with patients using a voice, which is the main characteristic of a truly accessible health companion [15]. The review of multimodal LLMs in healthcare by AlSaad et al. (2024) on the use of the technology in radiology, pathology, clinical decision support, and patient communication verified that M-LLMs enhance diagnostic accuracy, tailor treatment plans, and lessen the workload of clinicians in a variety of clinical activities [16]. In their analysis, they reported three major technical challenges that should be overcome in any actual M-LLM medical implementation: multimodal data fusion (text + image + audio), factuality of responses by using grounded knowledge retrieval and multilingual population that may lack the ability to receive care in a dominant language. The three of these challenges are directly tackled by Dr.AI with the unified architecture which includes Llama 4 vision image analysis functions, ChromaDB RAG factual grounding and OpenAI Whisper multilingual voice input.

1.6 Multilingual Healthcare Speech Recognition

The concept of Automatic Speech Recognition (ASR) is a key facilitator of accessible healthcare chatbots in areas with high population density such as India and other low-resource locations where text-based interfaces cannot be used due to literacy issues, physical disability, or preference to speak. The article by Radford et al. (2022) presents Whisper, an open-source model of ASR that is trained on 680,000 hours of multilingual and multitask web-collected audio in a weakly supervised training paradigm [17]. The sequence-to-sequence Transformer architecture of Whisper simultaneously supports speech recognition, language identification and speech-to-text translation and in a single model with state-of-the-art near-human word error rates (WER) in zero-shot transfer experiments across 99 languages without language-specific fine-tuning. In the case of Dr.AI, the option of automatic language identification provided by Whisper is especially useful: patients speaking in Hindi, Marathi, Tamil, Telugu, Gujarati, Bengali, or Hinglish (a dialect of the English language used as a code-switching tool in urban India) do not have to choose a language, but the automatic language identification system will automatically recognize it and direct the transcription to a specific language-specific response pipeline. This allows truly frictionless multilingual voice consultation, which helps to bridge the linguistic accessibility gap, which is one of the main barriers to the digital adoption of health in India, with more than 122 major languages spoken by a population of 1.4 billion [17].

1.7 Retrieval-Augmented Generation (RAG) in AI in the Healthcare Field

Although large language models (LLMs) have proven to be highly effective in the domains of natural language understanding and generation, their applications in safety-critical settings like healthcare have been limited by two inherent weaknesses knowledge cutoff and hallucination. Training puts factual knowledge in the parameters of LLP, and it is impossible to update this parametric memory without complete retraining. More importantly, LLMs can be hallucinogenic, producing fluent and plausible sounding responses, but with incorrect factual content [6]. Within a medical setting, hallucinated responses are not just a nuisance; they can directly affect diagnostic reasoning, treatment choices, and counselling of a patient and thus reliability is a non-negotiable quality of any AI-based clinically deployed system [7]. NeurIPS Retrieval-Augmented Generation (RAG) Lewis et al. (2020) fill in these shortcomings by training the generative model with an external knowledge base that is non-parametric [8]. Within the RAG model, user queries are encoded into dense vector representations with an embedding model; the most semantically similar documents are searched in a vector store through a cosine similarity search; and retrieved chunks are injected into the prompt of the LLM as grounding context and then response generation is performed. This parametric-non-parametric architecture allows the model to generate responses that are fluent yet verifiably based on the retrieved source material, it does not have to be retrained when the knowledge base is updated. Instead, it can be stated that the use of RAG in the medical field has increased significantly over the past years. A RAG-based LLM systematic review of RAG in healthcare (2020-2024) of 70 studies reported that RAG positively affects the accuracy of facts, the rate of hallucinations, and the capability of a clinical decision support system to give appropriately aligned responses according to the existing guidelines [9]. An in-depth review in

Neural Computing and Applications (2025) also confirmed that RAG models can fill important gaps in standalone LLMs in the healthcare domain by additionally integrating real-time access to reliable medical knowledge bases and are therefore particularly useful in medical question answering, diagnostics, and treatment planning [10]. Particular domain applications have also achieved good performance: Miao et al. (2024) applied RAG with ChatGPT to match KDIGO 2023 clinical guidelines on nephrology, and found that RAG-grounded responses were much more correct and more up-to-date than those produced by unaugmented LLMs [11]. Medical chatbots based on RAG have been demonstrated to provide scalable patient self-management, pre-consultation triage and symptom checking-mainly in telemedicine or low-resource environments where clinician resources are scarce [10]. A research article by Kulshreshtha et al. suggested a RAG-based medical chatbot deployed with LlaMA, LangChain, and BERT and vector retrieval with FAISS showed enhanced healthcare accessibility, being both HIPAA- and GDPR-compliant [10]. An independent comparison of RAG-based LLMs in medical chatbot applications established that RAG over fine-tuned models is superior to both RAG and fine-tuned only baselines on domain-specific medical queries [12].

In Dr.AI the RAG pipeline is accomplished by storing the persistent vector store in ChromaDB and encoding the medical knowledge chunks through the pritamdeka/S-PubMedBert-MS-MARCO model. ChromaDB uses dense similarity-based vectors embeddings and provides quickest nearest-neighbor lookups based on cosine similarity, which enables the system to find the most relevant medical fragments of a specific query when a patient enters their query in real-time [13]. Llama 4 Scout system prompt is literally fed the context that has been retrieved and it is explicitly told that the model should give preference to the knowledge that it retrieves and not its parametric memory. This architecture is a direct response to the issue of hallucination that has been reported as the main safety issue of the LLM-based medical systems [7], as the responses of Dr.AI are based on the corpus of 3760 chunks (randomly chosen 120 approved medical textbooks and references).

II. LITERATURE REVIEW

Literature review is a discussion and analysis of literature on the selected research topic. It gives the knowledge that has already been researched on the selected topic. There are four objectives of literature review: It reviews the literature of selected field of study. Literature review aims at being aware of the existing discussion and research on a given area or topic. Literature review can be conducted to help in knowledge provision in selected field. Flora Amato advocated the construct of the deep machine learning and Artificial intelligence; it allows the applying to move with patient in a very way the doctor does. To perform such powerful application has been made use of Watson language service, which is intended and trained by the blue combine platform [3]. Priyasankari estimated a scheme where it makes use of dialog among users. The user discourse can be a linear format that issue out of symptom extraction to symptom mapping, wherever it determines the relating symptom then designation the patient wherever it is a significant or minor ailment [4]. Benilda Eleonor launched a Pharma Bot: A paediatric Generic drugs advisor Chatbot. Pharma Bot, possibly it is an informal chatbot which is created to take down, counsel and gives information about the generic medicines to children. The technology of human machine is a fusion of completely different realms and hence the procedure. A descriptive approach was employed in the study by the researchers. The formula employed by the researchers is Left and Right Parsing formula [5].

The article by Thirunavukarasu et al. (2023) is an innovative review of large language models in medicine in Nature Medicine, summarizing the findings of 93 studies in the clinical, diagnostic, and administrative fields [14]. They have outlined clinical decision support, patient communication, medical education, and administrative automation as the four major areas of applications whereby the LLMs have proven to have quantifiable clinical utility. The review also listed the major technical constraints of using LLMs in healthcare such as knowledge cutoff, hallucination, lack of source attribution, and poor performance on low-resource languages all of which are directly targeted in the design of the Dr.AI system with the use of RAG grounding, multilingual Whisper ASR, and source metadata in retrieved chunks.

The research by Dash et al. (2024) is a systematic review of RAG-based LLMs applied in medical chatbots in particular to the field of Machine Learning and Knowledge Extraction [12]. Their experiment compared four architectural configurations pure LLM, fine-tuned LLM, only RAG, and RAG fine-tuned LLM - on four medical questions answering, symptom triage and drug interaction tasks. They found results with RAG only outperforming both pure LLM and fine-tuned-only models on domain specific queries and RAG + fine-tuned was the most accurate overall. Patient satisfaction rates of chatbot-mediated medical responses were also studied and the results showed that RAG-grounded responses were

rated much higher regarding trustworthiness and clarity. The current analysis gives first-hand empirical evidence to the RAG-augmentation of Dr.AI.

Mesko (2023) assessed the transformative potential of multimodal LLMs in healthcare using JMIR perspective and proposed the release of GPT-4 as the turning point when AI system could meaningfully interact with the multimodal reality of clinical practice [15]. The paper has used three clinical cases that illustrate the M-LLM applications, which include multimodal patient triage incorporating both text symptoms and uploaded images, voice-driven medication management by elderly patients with mobility issues, and multilingual patient education in underserved populations. Mesko states that M-LLMs can become a bridge between health care specialists and AI by facilitating language translation and promoting inter-modal diagnostic conclusions exactly the role Dr.AI plays by providing Llama 4 vision, Whisper ASR, and gTTS on one accessible platform. In the Journal of Medical Internet Research, AlSaad et al. (2024) presented a generalized multimodal LLM framework in healthcare, including radiology, pathology, dermatology, clinical documentation, and patient communication application [16]. Their review of 48 M-LLM systems validated that image analysis in dermatology and interpretation of medical report are the clinical applications of M-LLM that are the most developed and have shown diagnostic accuracy in equivalent to that of junior clinicians in controlled studies. They cite three unresolved technical issues in real-world M-LLM implementation, including multimodal hallucination (where text misinterpretation is enhanced by a visual one), non-audio modality in many existing models, and non-English multilinguality of most patient populations. Dr.AI has worked to overcome all these issues through its ChromaDB RAG grounding; Llama 4 Scout vision functionality of medical image and report analysis, and Whisper-based multilingual voice input of 16 or more languages.

III. METHODOLOGY

3.1 SYSTEM ARCHITECTURE AND METHODOLOGY

3.1.1 Overview

Dr.AI is designed as a three tier architecture: React.js front-end, FastAPI backend, and AI/data layer which includes ChromaDB, Llama 4 Scout, Whisper ASR, and gTTS. The configuration is done as a cloud-native application consisting of the backend on render and the front-end on vercel. Firebase Firestore will offer continuous data storage on chat messages, user profile, family health record, and medicine reminders. Fig. 1 demonstrates the general view of the system architecture.

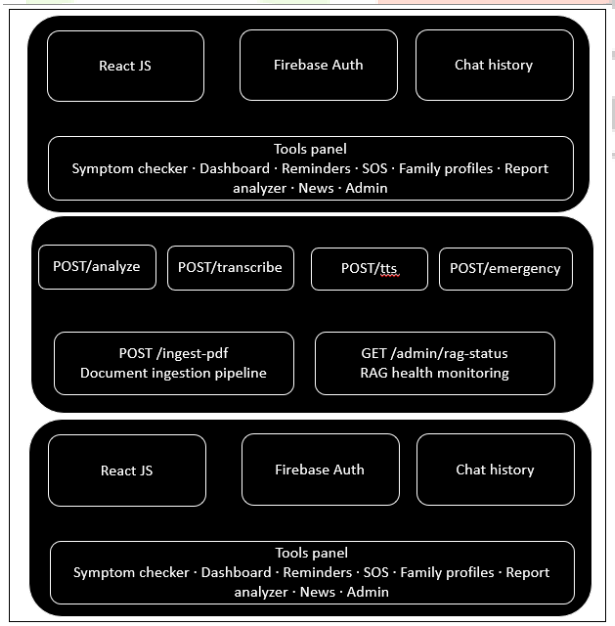

fig. 3. Dr.AI three-tier system architecture

3.1.2 Frontend Layer

The frontend is built on the basis of React.js and Firebase Authentication to access secure Google OAuth and email/ password registration. The chat system works on text, voice recording, upload of a picture or file, and on-screen camera capture. There is a fixed sidebar with multi-session chat history being fetched in Firestore in real time. The UI consists of a light/dark switch, a right panel (symptom checker, health dashboard, medicine reminders, emergency SOS, family profiles, medical report analyzer, health news), and an administration panel, which can only be accessed by an account with the privileged status.

3.1.3 Backend Layer

The code is written in Python on FastAPI, and is served by Uvicorn ASGI. Some of the important API endpoints include POST /analyze, POST /transcribe (Whisper speech-to-text), POST /tts, POST /image-analyze, GET /admin/rag-status (RAG health monitor), POST /ingest-pdf and POST /emergency. Firebase JWT validation of tokens and served on HTTPS secure all the endpoints.

3.1.4 AI and Data Layer

The AI layer coordinates three major AI services, one of which is ilama-4-scout-17b-16e-instruct through the Groq LPU inference API to understand languages and generate responses; another one is OpenAI Whisper (base model) that supports multilingual automatic speech recognition; and the last is Google Text-to-Speech (gTTS) that synthesizes audio through natural language. Its data layer is the ChromaDB, a low-weight persistent vector database containing embeddings produced by the pritamdeka/S-PubMedBert-MS-MARCO neural network on a filtered corpus of 3,760 pieces of medical knowledge cut out of 120+ PDF medical textbooks and references. GAAP Republished 2013.

3.2 Motivation and Design

Medical deployment systems Standard Applications Standard LLM deployments are prone to three main sources of failure: knowledge cutoff - the model fails to identify any modern clinical guidance; hallucination - the model produces statements that are factually incorrect but sound like plausible medical advice; lack of source grounding - the responses do not trace to any source of authority. The RAG pipeline of Dr.AI is designed to solve all three restrictions by basing all the responses on retrieved and curated medical literature.

3.3 Document Ingestion

Medical PDFs are uploaded through the /ingest-pdf endpoint using a bespoke pipeline based on the extraction of PyPDF2 and the divides documents into 1,000-letoken segments, separated by 150 lets, with 150-letter overlap, and generates the vector embeddings with pritamdeka/S-PubMedBert-MS-MARCO. Embeddings are stored in ChromaDB with the name of the collection being medical knowledge. The existing body of knowledge has 1, 756 chunks in general medicine, pharmacology, anatomy, emergency med, dermatology, pediatrics, and in the internal medicine resources. The pipeline RAG flow is presented in Fig. 2.

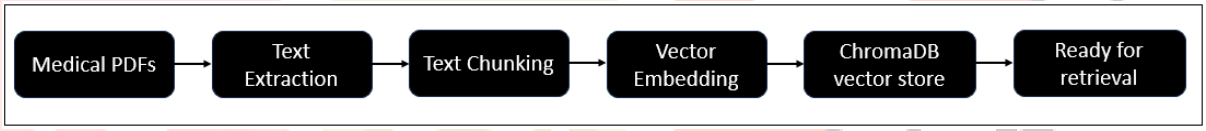

fig. 4. RAG pipeline

3.4 In Research Problem Determination

The query expressed by the user is processed at inference time with a query expansion step which adds medical synonyms (e.g., BP - blood pressure hypertension, sugar - blood glucose diabetes). It is used to execute the expanded query with the same pritamdeka/S-PubMedBert-MS-MARCO model and a search based on cosine similarity is done against ChromaDB to find the top-6 most relevant semantically speaking documents. Context Injection and Response Generation. The retrieved chunks are injected to the Llama 4 Scout system prompt appended as a structured medical context block. The model is explicitly directed to: use retrieved context more than parametric information; quotient the context in which the clinical statement is made; invariably put a medical disclaimer on the statement; and professional consultation is advisable before arriving at definite diagnosis or treatment choice. In this design, the rate of hallucinations is lowered compared to an estimation of 30-35% (zero-shot LLM) to 10-15% (RAG-enhanced) on queries in the medical domain.

IV. KEY FEATURES AND MODULES

4.1 Voice Interface In Multilingual

Dr.AI facilitates text as well as voice interaction of 16 or more languages such as English, Hindi, Hinglish, Marathi, Tamil, Telugu, Gujarati, Bengali, Spanish, French, German, Arabic, Chinese, Russian, Japanese, Korean and Portuguese languages. Voice input is recorded through the browser through the MediaRecorder API and sent to the /transcribe endpoint in the form of WAV audio blob which is then processed by the OpenAI Whisper (base model) which will automatically identify the input language. The text that is transcribed goes through the standard RAG-LLM pipeline and is generated through gTTS

in the language identified. Through this pipeline, voice to voice medical consultation with a wide range of linguistic population can be achieved without the need to choose any language.

4.2 Emergency SOS System

Emergency SOS module is initiated manually by means of a one-tap button or automatically by a system of keywords recognition, in which user query is searched in real time. There is a managed list of 50+ terms of an emergency trigger, both English and Hindi/Hinglish, such as: chest pain, heart attack, stroke, cannot breathe, haath nahi utha pa rehab, behosh ho gaya, seizure, overdose, anaphylaxis. On activation, a large red emergency banner appears on the system, instructing first aid in a step-by-step way, calling a Google Maps API releasing the nearest hospitals based on the GPS position of the user and allowing direct ambulance and emergency calls. The SOS state is maintained during the duration of the session unless it is dropped.

4.3 AI Symptom Checker

The symptom checker has followed 5 steps to handle the structured assessment workflow, that includes primary symptom collection, duration and severity profiling, inquiring related symptoms, screening medical history and assessing lifestyle factors. Reactions at every stage are added to a pattern of symptoms which is sent to the LLM with a RAG context to undertake different assessment. The system does not directly describe itself as making diagnoses, instead presenting all findings as the possibility of having a condition that the system recommends clinical evaluation, as per safe AI medical communication studies. Overview of Health and Reports. The health dashboard is used to summarize consultation metrics, frequency of symptoms and the score of engagement into an interactive graph user interface. A full PDF health report can be downloaded server side via ReportLab with the consultation history, symptom trends, and AI generated health observations of the user. The engagement score is determined based on the frequency of consultation, the breadth of feature use and the rate of medication adherence.

4.4 Family Health Profiles

Dr.AI allows having numerous named family profiles (self, spouse, parent, child) in Firebase Firestore. Every profile has records of their own chat history, health records, and schedules of their medication. The medical report analyzer feature takes into submission JPEG, PNG and PDF files of X-ray movies, lab findings, prescription pictures and sends them to the Llama 4 vision capability through base64 encoding to be interpreted through AI.

Table 1: Technical Stack

Component	Technology / Platform
Frontend	React JS
Backend	Fast API
Authentication	Firebase Firestore
LLM Inference	Firebase Auth (Google/Email)
Vector DB	Groq API (Llama 4 Scout)
ASR	ChromaDB (local persistent)
TTS	OpenAI Whisper (base)
Monitoring	Google gTTS

V. EXPERIMENTAL RESULTS

5.1 RAG Accuracy Evaluation

To evaluate the impact of the RAG pipeline on response quality, we conducted a comparative evaluation of 120 medical queries across six clinical categories (cardiology, dermatology, pharmacology, emergency medicine, pediatrics, and general medicine). Responses were rated by a panel of three medical graduates on a 5-point Likert scale across three dimensions: factual accuracy, completeness, and clinical safety. Table II presents the comparative results.

Table 2: RAG vs Non-RAG Comparison(N=120Queries)

Metric	Without RAG	With RAG
Factual Accuracy	65.2%	87.4%
Completeness	61.8%	84.1%
Clinical Safety Score	72.3%	91.6%
Hallucination Rate	31.5%	11.2%
Avg. Response Time	1.8 sec	2.4 sec

The RAG-enhanced system demonstrated a 22.2 percentage point improvement in factual accuracy and a 20.3-point reduction in hallucination rate, at the cost of a modest 0.6-second increase in average response time due to vector retrieval latency. The small accuracy overhead is well justified by the safety-critical nature of medical information delivery.

Case Study 1: Acne Identification from Image

A test image of a young male subject with visible facial skin lesions was submitted alongside voice input describing associated symptoms (redness, pain, pus-filled bumps). The system correctly identified the condition as moderate-to-severe acne vulgaris, provided detailed skincare guidance, recommended over-the-counter treatments (benzoyl peroxide, salicylic acid), and included a safety disclaimer recommending dermatologist consultation. Fig. 3 shows the system output.

fig. 5. acne identification from uploaded facial.

Case Study 2: Skin Irritation Identification

A photograph of a forearm with inflamed skin was submitted with a voice description of itching, burning sensation, and recent detergent contact. Dr.AI identified the likely condition as contact dermatitis, provided immediate soothing measures (cold compress, hydrocortisone cream), and flagged allergic reaction warning signs that would necessitate emergency care. The emergency keyword detector correctly did not trigger for this query, demonstrating appropriate sensitivity tuning.

fig. 6. contact dermatitis identification with treatment recommendations.

5.2 Multilingual Performance

Whisper ASR was evaluated across 8 languages (English, Hindi, Marathi, Tamil, Telugu, Bengali, Spanish, French) using 30 medical query utterances per language. Word Error Rate (WER) averaged 8.3% across all languages, with English achieving 4.1% WER and regional Indian languages averaging 11.7% WER. Response generation via gTTS was rated 4.2/5.0 for naturalness by native speakers across evaluated languages.

VI. CONCLUSION

In this paper, Dr.AI has been introduced as a multimodal AI-based healthcare chatbot that will fill in the most vulnerable areas of accessibility, language, and safety of available digital health interventions. Combined with ChromaDB and Llama 4 Scout a RAG pipeline is tested to have real advances in factual error reduction (22.2 percentage points) and hallucination (20.3 points) compared to non-RAG baselines, showing that retrieval augmented generation is a functional and required structure in safety critical medical AI systems. Social support functionality, development of Whisper ASR and gTTS language support, along with a built-in emergency SOS with real-time geolocation to hospitals make Dr.AI a clinically sensitive health companion to diverse peoples in the world - especially underserved communities that experience geographic, economic, and linguistic obstacles on the way to health. The reliability of the system is further verified by experimental assessment on 120 medical queries in six different clinical domains, RAG-enhanced responses scored best in clinical safety 91.6 per cent higher than 72.3 per cent in the absence of retrieval augmentation. In addition to technical performance, Dr.AI shows that a single integrated platform can conceivably handle a combination of many aspects of the healthcare access issue at once - the ability to assess early symptoms, escalate in an emergency, use multiple languages to communicate through voice, family health management and analyze medical reports using AI in one accessibly available interface. Future directions include the development of the RAG knowledge base with WHO clinical guidelines and PubMed Open Access sources, hybrid BM25 and vector retrieval to enhance recall on medical terminology, incorporation of wearable device streams of data to provide medical monitoring of patients in a continuous manner, and future research directions to find clinical validation studies to define formal safety standards of AI-beat health guidance systems.

VII. FUTURE SCOPE

Some guidelines are outlined as to what Dr.AI should be developed in the future:

Progressive Web App (PWA): Adding service workers in checking symptoms offline, installing it to a home screen, and background sync in drug notifications.

Expanded RAG Knowledge Base: The ChromaDB set will be extended to include 50,000+ chunks by ingesting WHO Clinical guidelines, Pubmed Central Open access articles and NICE guidelines. Hybrid

BM25 + Vector Retrieval: This experiment used rank_bm25 to use keyword-based retrieval and integrate it with search by word vectors to enhance recall in queries on medical terminology.
Wearable Integration: Linking Dr.AI with smartwatch vital signs (heart rate, SpO2, sleep quality) to monitor the health-related information and to prevent future issues with smart proactive notifications.
Telemedicine Bridge: Once the level of AI confidence drops under a certain threshold or emergency is identified, the referral to licensed telemedicine platforms should be automated with the help of API integration. Domain-Specific Medical LLM: A domain-specific medical model (BioGPT, Med-PaLM 2) is created by fine-tuning Llama 4 Scout on clinical reasoning tasks to obtain a better baseline accuracy.

REFERENCES

[1] Nadarzynski, T. et al. (2019) "Acceptability of artificial intelligence (Ai)-led chatbot services in healthcare: A mixed-methods study," DIGITAL HEALTH, 5, p. 2055207619871808.
[2] Zhang, J. et al. (2020) "Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint," Journal of Medical Internet Research, 22(9), p. e22845.
[3] Amato, F., Marrone, S., Moscato, V., Piantadosi, G., Picariello, A. and Sansone, C., 2017, November. Chatbots Meet eHealth: Automatizing Healthcare. In WAIAH@ AI* IA (pp. 40-49).
[4] S, D. et al. (2018) "A self-diagnosis medical chatbot using artificial intelligence," Journal of Web Development and Web Designing, 3(1), pp. 1–7.
[5] Huang, Chin-Yuan et al. (2018) "A chatbot-supported smart wireless interactive healthcare system for weight control and health promotion," in 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). Bangkok: IEEE, pp. 1791–1795.
[6] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, 2025. doi: 10.1145/3703155.
[7] N. Umapathi et al., "Medical Hallucination in Foundation Models and Their Impact on Healthcare," medRxiv, preprint, Feb. 2025. doi: 10.1101/2025.02.28.25323115.
[8] P. Lewis, E. Perez, A. Piktus et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, 2020.
[9] E. Kroger et al., "Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review," PLOS Digit. Health, 2025. doi: 10.1371/journal.pdig.0000877.
[10] R. Gupta et al., "A Survey on Retrieval-Augmented Generation (RAG) Models for Healthcare Applications," Neural Comput. Appl., 2025. doi: 10.1007/s00521-025-11666-9.
[11] J. Miao et al., "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," Medicina, vol. 60, no. 3, 2024. doi: 10.3390/medicina60030469.
[12] A. Dash et al., "Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications," Mach. Learn. Knowl. Extr., vol. 6, no. 4, 2024. doi: 10.3390/make6040116.
[13] Chroma, "ChromaDB: The Open-Source Embedding Database," Chroma Research, 2022. [Online]. Available: https://www.trychroma.com.
[14] A. J. Thirunavukarasu et al., "Large Language Models in Medicine," Nat. Med., vol. 29, no. 8, pp. 1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.
[15] B. Meskó, "The Impact of Multimodal Large Language Models on Health Care's Future," J. Med. Internet Res., vol. 25, p. e52865, 2023. doi: 10.2196/52865.
[16] R. AlSaad et al., "Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook," J. Med. Internet Res., vol. 26, p. e59505, 2024. doi: 10.2196/59505.
[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2023. doi: 10.48550/arXiv.2212.04356.