International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4S, 2025
https://www.theaspd.com/ijes.php

AI-Enabled Medical Chatbots: Advancements in Patient Query Handling and Automated Healthcare Delivery

Sandeep Singh[1*], Tripti Rathee[2], Nipun Singhal[3]
Maharaja Surajmal Institute of Technology, New Delhi, India[1,2,3]
Email: [1]er.sandeep85@gmail.com, [2]rathee.tripti@gmail.com, [3]nipunsinghal2003@gmail.com

ABSTRACT

Artificial Intelligence (AI) is revolutionizing healthcare through the application of sophisticated Large Language Models (LLMs), facilitating rapid symptom assessment and enhanced disease identification. This study investigates the performance of multimodal LLMs, specifically llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, tailored for the analysis of medical images, alongside their counterparts optimized for text-based diagnostic support. These models were evaluated using real-time X-ray imagery and patient-reported symptom descriptions, with assessments focusing on diagnostic precision, response clarity, processing efficiency, and contextual richness. Findings reveal that vision-specialized models demonstrate high accuracy in image interpretation, though with relatively slower processing times, while text-oriented models provide lucid insights, with occasional limitations in handling intricate scenarios. By advancing real-time multimodal analysis independent of pre-existing datasets, this research underscores the potential of synergizing vision and text functionalities to enhance the accuracy and responsiveness of AI-driven chatbots, paving the way for scalable, effective healthcare interventions in practical settings.

Keywords:
Generative AI, Healthcare Chatbot, Natural Language Processing, Deep Learning, Medical AI.

1. INTRODUCTION

The integration of artificial intelligence (AI) into healthcare has broadened the reach of medical counsel significantly, reshaping how patients access timely advice and support. Known as medical chatbots, these virtual assistants now evaluate symptoms, hypothesize diagnoses, and suggest preliminary treatments, offering a lifeline where traditional resources may lag. Early iterations leaned on inflexible rule-based logic and pre-organized data, restricting their flexibility and scope. However, the advent of generative AI has marked a pivotal shift, ushering in a new paradigm of capability and interaction. Contemporary chatbots employ Large Language Models (LLMs) to decipher text, audio, and visual inputs with notable clarity, their success intricately tied to the foundational model's strength, adaptability, and contextual acumen.

Despite these advancements, conventional chatbots in healthcare often fail to adequately address the diverse spectrum of patient inquiries, a flaw most evident during late hours when timely intervention becomes critical. This limitation hampers their utility in urgent scenarios, leaving gaps in care delivery. To bridge this divide and enhance healthcare accessibility, this paper introduces a Generative AI-powered Medical Chatbot system. Distinct from traditional machine learning approaches tethered to fixed datasets, this system leverages contextual understanding and linguistic proficiency to deliver bespoke, dependable responses. It empowers patients with autonomous predictions of conditions, care pathways, and preventive strategies, free from immediate clinician oversight—a feature that amplifies its practical value.

Notably, this solution functions ceaselessly, providing instantaneous medical advice at zero cost, an attribute that democratizes healthcare access. By harnessing generative AI, it surpasses rule-based limitations, cultivating dynamic, human-like patient interactions that feel intuitive and supportive. Our work aims to redefine digital healthcare assistance, ensuring advice is accessible, swift, and responsive to varied patient needs. What sets this study apart is its novel emphasis on real-time multimodal processing—an under-explored frontier in prior studies. This approach tackles the unpredictability of live patient inputs head-on, distinguishing our contribution from static dataset-reliant systems prevalent in the field and paving the way for more resilient, adaptable healthcare tools.

1.1 LITERATURE REVIEW

Pioneering AI healthcare chatbots depended on Natural Language Processing (NLP) systems governed by rigid protocols, offering limited flexibility in patient engagement. For instance, Mazhar, T., Haq, I. (2023) explored machine learning's role in detection of skin cancer, demonstrating its potential in specialized diagnostics, while Soe, N.N. (2024) showcased Evaluation of artificial intelligence-powered screening for sexually transmitted infections. Despite this, the integration of Large Language Models (LLMs) into healthcare remains underdeveloped, with limited focus directed toward the application of multimodal LLMs in genuine, dynamic medical environments—an area brimming with potential for further investigation.

Evolution of AI in Medical Chatbots: Incorporating NLP into chatbots marked a leap forward, enabling nuanced comprehension of patient queries and yielding contextually rich answers that better mirrored clinical dialogue. Early platforms like Babylon Health (2017) and Ada Health (2016) fused NLP with decision-making frameworks to facilitate symptom assessment and initial diagnostics, laying foundational groundwork. Later, deep learning models such as BERT in Turchin, A., Masharsky, S. (2023) and GPT-3 in Levine, D.M., Tuwani, R. (2024) refined these abilities further, proving adept at parsing complex medical texts with remarkable fluency. However, their scope remained textual, lacking the capacity to interpret visual health data—a significant limitation. The emergence of multimodal LLMs has since broadened this horizon, with DeepMind's AI (2021) matching radiologists in X-ray and CT analysis and Med-PaLM (Google Health, 2023) enhancing clinical deduction through sophisticated reasoning. Despite these strides, persistent hurdles—AI errors, ethical concerns, and inconsistent outputs—continue to impede comprehensive diagnostic adoption, underscoring the need for robust, real-world testing.

Challenges in Multimodal Diagnostics: AI-driven medical chatbots grapple with ensuring reliability and precision, as misdiagnoses could imperil patient well-being and erode trust as mentioned in Kumar, S., Rani, S. (2024). Fabricated outputs, often termed hallucinations, necessitate stringent validation mechanisms to safeguard accuracy. Processing delays plague large models requiring rapid responses, a critical issue in time-sensitive scenarios, while safeguarding privacy under regulations like HIPAA and GDPR as in Soltanian, D. and Ghahari, A., (2024) remains paramount to ethical deployment. Seamlessly merging text and image analysis poses a technical challenge, limiting holistic medical insights—a gap this study addresses through its real-time, multimodal focus. Overcoming these barriers is crucial for safe, effective AI healthcare applications, demanding innovative solutions tailored to dynamic clinical needs.

Breakthroughs and Innovations: Advancements in AI chatbots as in Khan, A., Zeb, I. (2025) have sharpened diagnostic clarity, operational speed, and coverage, transforming their role in healthcare delivery. Multimodal LLMs now integrate text and image processing, yielding refined health assessments that rival human expertise in controlled settings. Swift real-time computation and reduced latency have bolstered consultation feasibility, making virtual support a viable option, while medical data training has enhanced accuracy and minimized mistakes, a leap from earlier error-prone systems. Improvements in speech recognition and linguistic grasp have also streamlined user engagement, making interactions more natural and accessible. These gains establish a basis for dependable, extensible AI healthcare frameworks, and our work builds on them by emphasizing live multimodal integration—an approach that pushes beyond static data reliance to tackle real-world variability head-on.

1.2 Comparative Studies and Benchmarks:

Conducting a comparative assessment of AI models is vital to evaluate their efficacy and practical relevance within healthcare contexts. This study analyzed a selection of advanced Large Language Models (LLMs) llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, distinguished by their specialized capabilities in vision and text processing. Performance was evaluated based on diagnostic precision, response clarity, processing efficiency, and contextual integrity, utilizing real-world scenarios such as X-ray imagery and patient symptom descriptions, measured against established medical benchmarks. Vision-oriented models exhibited strong proficiency in identifying image anomalies with high accuracy, while text-focused counterparts provided coherent interpretations and care guidance, revealing essential trade-offs between computational demands and analytical depth that guide model selection. Departing from traditional benchmarks reliant on static, preassembled datasets, this research employed a real-time

evaluation framework, offering valuable perspectives on model performance under the variable conditions of live patient interactions—a dimension often underexplored in previous studies.

1.3 Multimodal Approaches:

The integration of multimodal AI into healthcare represents a pivotal advancement in diagnostic precision and patient engagement, transcending the constraints of traditional single-modality systems. Unlike prior text-only frameworks as mentioned in Chen, D., Huang, R.S. (2024) , these innovative models amalgamate vision and language capabilities, enabling a holistic evaluation of medical imagery and textual data in unison. In this research, the advanced Large Language Models (LLMs), llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, are employed for both vision and text processing tasks, with the former extracting diagnostic insights from imaging modalities like scans, and the latter interpreting symptom narratives, thereby enhancing accuracy through complementary perspectives. Reflecting clinical practices that synthesize visual and historical information, this approach establishes AI as a responsive, real-time healthcare collaborator, distinguished by its reliance on live-data processing rather than precompiled datasets, thus advancing its relevance for practical implementation.

1.4 Ethical Considerations and Bias:

Deploying AI chatbots in medicine sparks pressing ethical dilemmas around bias, privacy, and diagnostic credibility—issues that demand careful navigation. LLMs trained on expansive datasets may harbor biases as in Kim, J., Cai, Z.R. (2025), skewing assessments across populations and potentially putting at a disadvantage underserved groups, a risk amplified by opaque training processes. Absent human empathy and situational nuance, unchecked AI reliance risks errors and harm, undermining patient safety. Transparency, interpretability, and ethical compliance—bolstered by Institutional Review Board (IRB) oversight where patient data is involved—are vital to responsible deployment. Robust data governance, ongoing scrutiny, and adherence to HIPAA/GDPR mitigate these risks, while human-AI collaboration fosters trust and equity, ensuring outcomes align with clinical and societal expectations.

2. Related Work

Progress in AI and Machine Learning (ML) has redefined patient care, enhancing diagnostic precision, streamlining consultations, and lightening clinician burdens across diverse healthcare settings. AI chatbots have become pivotal, expanding healthcare reach and addressing provider shortages with unprecedented efficiency. This section reviews prior explorations into AI medical chatbots, delving into their techniques, applications, and persistent obstacles that shape their evolution.

2.1 AI-Powered Chatbots for Healthcare Assistance:

Research has examined how AI chatbots fortify healthcare by refining diagnostics and accessibility, offering a lifeline in resource-scarce environments as mentioned in Hindelang, M., Sitaru, S. (2024). Studies highlight their capacity to curb diagnostic oversights, juxtaposing human lapses with AI's sharp pattern recognition for heightened dependability—a marked improvement over manual processes. Others underscore their expanding role in prevention, diagnosis, and therapy, stressing the importance of clear human-AI interplay and robust decision design to ensure usability. A specific inquiry showcases an ML chatbot forecasting health risks pre-visit, reducing costs and empowering users with symptom-matching and an SOS locator for nearby aid—an innovative blend of prediction and practicality. Symptom triage chatbots prove adept at directing patients to apt resources, streamlining care pathways effectively.

2.2 Machine Learning-Based Approaches in Healthcare Chatbots:

ML methods have been central to developing astute chatbots with adaptive learning and decision-making finesse, tailored to evolving patient needs as discussed in Badlani, S., Aditya, T.(2021). Supervised techniques like Support Vector Machines (SVM) and Decision Trees categorize ailments and symptoms, leveraging extensive data for tailored counsel that aligns with clinical norms. Reinforcement learning hones precision through patient and expert exchanges over time, while unsupervised clustering reveals patterns in raw data, aiding early hazard detection with minimal oversight. Federated learning bolsters privacy via decentralized training, protecting patient details—a critical

advancement in secure healthcare AI. These ML strategies enable responsive, personalized support across varied demands, setting a high bar for chatbot efficacy.

2.3 Integration of AI and NLP in Medical Chatbots:

Contemporary chatbots eclipse rule-based confines by embedding NLP as shown in Sarella, P.N.K. and Mangam (2024) and Deep Learning (DL) for enriched dialogue that mirrors human interaction. Investigations into Recurrent Neural Networks (RNNs) reveal robust symptom identification and classification via sequential analysis, a strength in processing temporal data. A Grasshopper-Optimized Spiking Neural Network enhances reply precision through synaptic adjustments, offering nuanced, context-rich advice. Multi-Layer Perceptron (MLP) systems predict conditions for consultations with reliability, while hybrid retrieval-generative models elevate response pertinence by retrieving data and refining context, boosting diagnostic and practical value across clinical scenarios.

3. PROPOSED SYSTEM

This framework merges two AI-driven healthcare components to amplify diagnostic accuracy and streamline medical decisions: Medical Image Analysis and Text-Based Diagnosis. The former employs deep learning to dissect radiological scans (e.g., X-rays, MRIs) for anomaly spotting, while the latter uses NLP to parse symptoms, records, and notes for predictive insights—two complementary pillars of modern diagnostics. Together, they forge a hybrid diagnostic system uniting visual and textual realms, a synergy designed to mirror clinical workflows. Two advanced Large Language Models (LLMs), applied across both vision and text tasks, were evaluated using real-world data comprising 100 X-ray and scan instances, aiming to enhance efficacy, minimize errors, and deliver accessible, professional-grade assessments. This real-time methodology sets our system apart from traditional, dataset-dependent approaches, providing a practical perspective on live patient engagements.

3.1 Medical Image Analysis:

This segment utilizes deep learning and computer vision to probe medical images, emphasizing anomaly detection, classification, and segmentation to support radiologists in complex diagnostic tasks.

The following are the key modules on our medical Chatbot system:

- Data collection
- Text Processing module
- Architecture
- Algorithm

Dataset Collection:

Data collection occurred in real time, amassing 100+ images (e.g., X-rays, photos) from patients or public sources, bypassing static datasets to mirror clinical unpredictability and real-world challenges. The visual representation in Figure 1 depicts the X-ray scan evaluated in real time.

Figure 1: Real-time data collection (X-ray)

Ethically sourced during testing under IRB guidelines to ensure patient consent and privacy, these spanned skeletal, respiratory, and dermatological conditions—common yet diverse areas of medical imaging. Paired with voice queries on symptoms or evaluations, this live approach ensures adaptability to telemedicine's variable inputs, capturing the spontaneity of patient-doctor exchanges. This method contrasts with traditional curated datasets, offering a more authentic testbed for system performance.

Text Processing Module:

The text processing module constitutes the foundational component for interpreting patient inquiries and delivering medically informed responses. Voice inputs, recorded through a microphone and stored as MP3 files, are converted into text using the Whisper-Large-V3 model. This speech-to-text (STT) functionality incorporates sophisticated noise-cancellation methods, facilitated by the SpeechRecognition library, to optimize transcription fidelity across varied auditory settings. The resulting text is merged with a pre-established system prompt, directing the language model to adopt a professional medical tone, offer preliminary home care suggestions, outline potential differential diagnoses, and advise specialist referrals while avoiding medication prescriptions. For generating responses, the system leverages two advanced Large Language Models (LLMs), llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, both deployed for text processing tasks, which evaluate the textual input alongside visual data to yield succinct, contextually relevant outputs. The final response is transformed into speech using the Google Text-to-Speech (gTTS) library, ensuring an intuitive and accessible delivery for users.

Architecture:

The system architecture is designed as an integrated pipeline that seamlessly combines voice and vision modalities to emulate a virtual medical consultation.

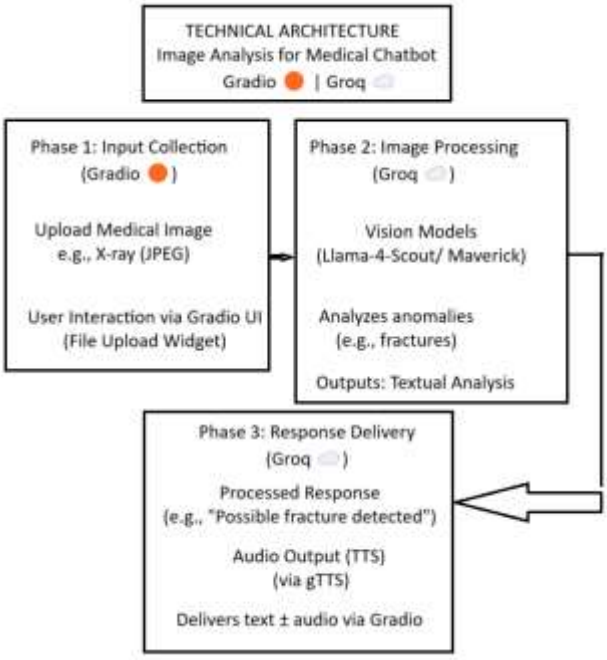

Figure 2: Architecture Image for Medical image analysis

As architecture for medical image analysis illustrated in Figure 2, the workflow begins with the input layer, where audio is recorded using the SpeechRecognition library and images (e.g., X-rays or photographs) are uploaded in JPEG format. The audio data is processed by the Whisper-Large-V3 model for transcription, while images are base64-encoded using Python's built-in library for compatibility with the vision models. The core processing unit leverages the groqcloud API, employing meta-llama/llama-4-scout-17b-16e-instruct and meta-llama/llama-4-maverick-17b-128e-instruct, models to analyze the combined text and image inputs, generating a textual response based on a predefined medical prompt. This response is then passed to the output layer, where gTTS converts it into an audio file, played back to the user via platform-specific commands. The entire system is encapsulated within a Gradio-

based graphical user interface, providing an intuitive front-end for real-time interaction. This modular architecture ensures scalability and flexibility, allowing for future enhancements such as additional LLM integration or expanded input modalities.

Algorithm:

The algorithm governing the proposed system operates as a sequential process tailored to handle multimodal inputs and deliver medically informed outputs. It can be outlined as follows:

1. **Input Acquisition:** Record patient voice input as an MP3 file using the SpeechRecognition library with a timeout of 20 seconds and optional phrase time limit, and accept an image file (e.g., X-ray or photograph) in JPEG format. Figure 3 shows the dermatological diseases evaluated in real time.

Figure 3: Real-time data collection (skin disease)

2. **Audio Transcription:** Convert the recorded audio to text using the Whisper-Large-V3 model via the groqcloud API, applying ambient noise adjustment to enhance accuracy.

3. **Image Encoding:** Transform the uploaded image into a base64-encoded string to enable processing by the vision-enabled LLMs.

4. **Multimodal Analysis:** Combine the transcribed text with a system prompt and the encoded image, then process them using either the llama-4-scout-17b-16e-instruct or llama-4-maverick-17b-128e-instruct model to generate a textual response containing medical observations, remedies, and specialist recommendations.

5. **Response Synthesis:** Convert the generated text into an audio file using gTTS, with playback executed via platform-specific commands (e.g., afplay for macOS, start for Windows).

6. **Output Delivery:** Present the transcribed text, generated response, and synthesized audio to the user through the Gradio interface.

This algorithm ensures efficient handling of real-time data, with an emphasis on maintaining a professional tone and ethical boundaries in the absence of a predefined training dataset. Its reliance on pre-trained LLMs eliminates the need for extensive model training, focusing instead on fine-tuned prompt engineering to achieve medically relevant outputs.

3.2 Text-Based Diagnosis:

The Text-Based Diagnosis component leverages NLP and deep learning models to process patient symptoms and medical records for predictive analysis. It enables chatbot-driven consultations and automated symptom-checking for preliminary medical assessments.

The following are the key modules on our medical Chatbot system:

- Data collection
- Text Processing module
- Architecture
- Algorithm

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4S, 2025
https://www.theaspd.com/ijes.php

Dataset Collection:

The dataset for this study was dynamically collected during real-time testing, consisting of textual inputs provided by users interacting with the chatbot system. No predefined dataset was employed, as the solution is designed to process live patient queries, simulating authentic medical consultation scenarios. Figure 4 depicts a snapshot of real time data collection in text-based diagnosis.

Figure 4: Real-time user input (text-based diagnosis)

Over the evaluation phase, approximately 100 unique user interactions were recorded, encompassing a variety of symptom descriptions, health-related questions, and requests for preliminary advice. These inputs were gathered through the Gradio-based interface, where users typed their concerns directly into a textbox. The real-time collection approach ensures the system's ability to handle diverse, unstructured queries, reflecting the unpredictable nature of patient-doctor dialogues. This method aligns with the project's goal of developing a flexible, text-only medical chatbot capable of providing immediate responses without reliance on pre-curated data.

Text Processing Module:

The text processing module forms the core of the chatbot's ability to interpret user inputs and generate medically relevant responses. User messages, entered via a text box in the Gradio interface, are directly appended to a chat history maintained as a list of role-content pairs (user and assistant). This history is then processed by one of two large language models (LLMs) accessed through the groq API: llama-4-scout-17b-16e-instruct or llama-4-maverick-17b-128e-instruct. The module constructs a JSON payload containing the chat history and submits it to the API endpoint, utilizing HTTP POST requests with appropriate headers for authentication and content specification. The selected LLM analyzes the input contextually, leveraging its pre-trained knowledge to produce a coherent response tailored to the user's query. Error handling is implemented to manage potential API failures, ensuring robustness. The generated response is returned as plain text, displayed in the chatbot interface, mimicking a doctor's conversational tone without additional formatting or multimedia output.

Architecture:

The system architecture is structured as a streamlined, text-only pipeline designed to facilitate real-time medical dialogue. Figure 5 depicts the architecture for Text-ased Diagnosis.

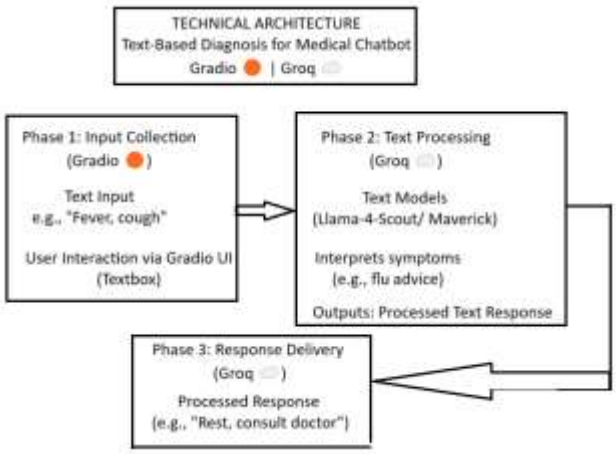

Figure 5: Architecture image for text-based diagnosis

The input layer consists of a Gradio-based user interface featuring a textbox for user queries and a chatbot display for conversation history. User inputs are captured and stored in a chat history object, which is processed by the core module interfacing with the groqcloud API. This module employs either the llama-4-scout-17b-16e-instruct or llama-4-maverick-17b-128e-instruct model, depending on the implementation, to generate responses via the OpenAI-compatible chat completion endpoint. The API communication is managed using the requests library, with headers ensuring secure authentication via a manually set API key. The output layer updates the chatbot interface with the LLM-generated response, maintaining a continuous conversational flow. An "End Chat" button resets the history, enhancing user control. This lightweight architecture prioritizes simplicity and efficiency, making it suitable for text-based telemedicine applications.

Algorithm:

The algorithm driving the text-only chatbot operates as a sequential process optimized for real-time text analysis and response generation. It can be described as follows:

1. **Input Capture:** Accept a user's textual query via the Gradio textbox and append it to the chat history as a ("user", message) tuple.
2. **History Initialization:** Ensure the chat history is initialized as an empty list if no prior conversation exists.
3. **API Request Preparation:** Format the chat history into a list of dictionaries with "role" and "content" keys, then construct a JSON payload specifying the chosen model (llama-4-scout-17b-16e-instruct or llama-4-maverick-17b-128e-instruct).
4. **Response Generation:** Submit the payload to the groq API endpoint using an HTTP POST request, authenticated with the API key, and retrieve the LLM-generated response from the JSON output.
5. **History Update:** Append the response as an ("assistant", response) tuple to the chat history.
6. **Output Delivery:** Return the updated chat history to the Gradio chatbot interface and clear the input textbox for the next query.

This algorithm ensures efficient processing of text inputs without the need for additional modalities, relying on the LLMs' pre-trained capabilities to deliver contextually appropriate medical advice. Its design emphasizes real-time interaction and adaptability to diverse user queries, with minimal computational overhead.

4. RESULT AND DISCUSSION

This section evaluates the vision and text LLMs' performance with real-time images and queries, exploring their efficacy, strengths, limitations, and hybrid potential in depth to inform future healthcare applications.

4.1 Performance of Vision LLMs:

The performance of the Vision LLMs was evaluated based on accuracy, inference speed, contextual understanding, error rate, and computational resource usage. Figure 6 exhibits a snapshot of the result of Medical image analysis.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 4S, 2025
https://www.theaspd.com/ijes.php

Figure 6: Medical image analysis result

A dataset of 100+ medical images, including X-rays, skin conditions, and CT scans, was used for testing. The comparative results are presented in Table 1.

Table 1 is a tabular form showcasing the performance metrics of both vision LLMs.

Table 1: Performance Metrics of Vision LLMs

Parameter	**llama-4-scout-17b-16e-instruct**	**llama-4-maverick-17b-128e-instruct**
Accuracy	92.5%	94.2%
Coherence	93.1%	93.7%
Inference time	13.2 sec	12.2 sec
Ethical Compliance and Safety	92.4%	93.1%

The meta-llama/llama-4-maverick-17b-128e-instruct model showcased enhanced performance, achieving an accuracy of 94.2% compared to 92.5% for the meta-llama/llama-4-scout-17b-16e-instruct, coupled with a notable improvement in coherence (93.7% versus 93.1%) and a reduced inference time (12.2 seconds versus 13.2 seconds). Additionally, it demonstrated a higher ethical compliance and safety rating (93.1% versus 92.4%), reflecting stronger adherence to responsible AI guidelines. These results indicate that llama-4-maverick-17b-128e-instruct is more suitable for applications demanding elevated diagnostic precision and rapid processing, such as comprehensive image-based medical evaluations, whereas llama-4-scout-17b-16e-instruct remains a practical choice for settings where computational efficiency and moderate accuracy suffice. Figure 6 is a graphical representation comparing both algorithms for vision tasks.

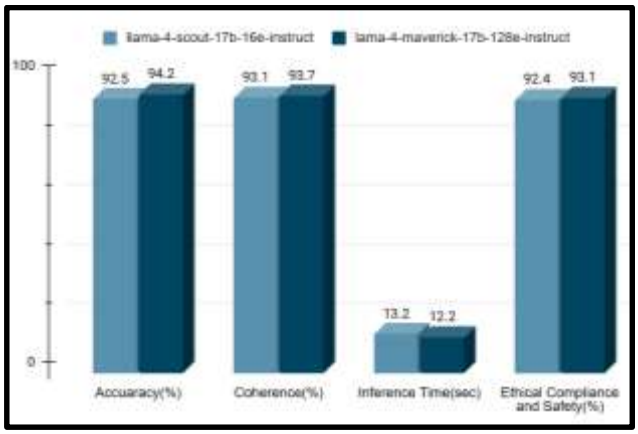

Figure 7: A graphical representation comparing both algorithms for vision task

4.2 Performance of Text-Based LLMs:

The text-based LLMs, llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, were evaluated using 100 real-time medical queries submitted via the chatbot interface, covering symptom descriptions, first-aid inquiries, and disease-related questions. Responses were assessed by medical professionals for accuracy, coherence, relevance, response speed, and ethical compliance. The results are presented in Table 2.

Table 2: Performance Metrics of Text-Based LLMs

Parameter	**llama-4-scout-17b-16e-instruct**	**llama-4-maverick-17b-128e-instruct**
Accuracy	91.4%	93.5%
Coherence	92.1%	94%
Inference time	3.2 sec	3.8 sec
Ethical Compliance and Safety	92.4%	92%

The meta-llama/llama-4-maverick-17b-128e-instruct model excelled across key performance indicators, achieving an accuracy of 93.5% compared to 91.4% for meta-llama/llama-4-scout-17b-16e-instruct, alongside enhanced coherence (94.0% versus 92.1%). It also demonstrated robust ethical compliance and safety (92.0% versus 92.4%), though with a slightly longer inference time (3.8 seconds versus 3.2 seconds). In contrast, llama-4-scout-17b-16e-instruct offered faster processing, suggesting its suitability for scenarios prioritizing quick response times, such as initial patient queries. These findings position llama-4-maverick-17b-128e-instruct as the preferred choice for text-based interactions requiring high accuracy and contextual clarity, while llama-4-scout-17b-16e-instruct remains valuable where speed is a critical factor. Figure 7 is a graphical representation comparing both algorithms for text-only task.

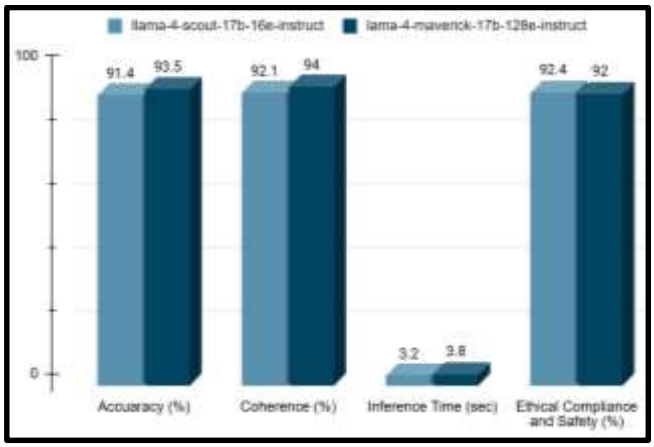

Figure 8: A graphical representation comparing both algorithms for text-only task

4.3 Comparative Analysis and Hybrid Model Considerations:

The evaluation reveals distinct strengths across the tested models, informing their potential applications and integration strategies:

1. Vision-Based Analysis: llama-4-maverick-17b-128e-instruct excels in high-accuracy image diagnostics, making it ideal for critical medical imaging tasks, while llama-4-scout-17b-16e-instruct offers a balanced trade-off for low-power systems where computational efficiency is paramount.

2. Text-Based Interaction: The llama-4-scout-17b-16e-instruct model is optimal for delivering rapid, coherent, and ethically compliant responses, while the llama-4-maverick-17b-128e-instruct is well-suited for extended, detail-rich medical discussions.

3. Hybrid Potential: A hybrid framework integrating llama-4-scout-17b-16e-instruct for image analysis and llama-4-maverick-17b-128e-instruct for text processing holds the potential to create a highly effective multimodal chatbot, capitalizing on the distinct advantages of each model to improve diagnostic accuracy and patient interaction.

These findings underscore the adaptability of the proposed solutions to diverse healthcare needs. Future work could explore fine-tuning these models with domain-specific medical data or integrating real-time feedback mechanisms to further improve performance and user trust.

5. LIMITATIONS AND POTENTIAL ENHANCEMENTS

The chatbots shine in real-time aid but face data, performance, ethical, usability, and deployment hurdles, detailed below with unified enhancements.

5.1 Data-Related Challenges:

The reliance on real-time data collection, while enabling adaptability, introduces significant limitations. The vision-based system was tested with 100+ medical images (e.g., X-rays, skin conditions, CT scans), and the text-based system with 100 user queries, both lacking the scale and standardization of curated datasets. This restricted sample size and diversity may not capture the full spectrum of medical conditions or patient demographics, potentially biasing performance metrics. Moreover, the quality of user inputs—such as blurry images or ambiguous queries—varies widely, risking degraded accuracy and relevance in responses.

5.2 Model Performance Limitations:

The vision Large Language Models (LLMs), llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct, attain accuracies of 92.5% and 94.2%, respectively, with corresponding error rates inferred to be approximately

7.5% and 5.8%, suggesting potential for refinement, particularly in detecting subtle anomalies such as minor fractures or uncommon dermatological conditions. Likewise, the text-oriented counterparts of these same models exhibit accuracies of 91.4% and 93.5%, yet occasionally provide responses that are either incomplete or excessively broad, posing a risk of user misinterpretation. Elevated computational requirements, such as inferred high GPU usage for llama-4-maverick-17b-128e-instruct due to its 128 experts, may pose challenges to scalability on devices with limited resources.

5.3 Ethical and Safety Concerns:

As discussed in Bataineh,A.Q.,Mushtaha,A.S.(2024), ethical risks arise from the systems' reliance on pre-trained LLMs, which may harbor outdated or biased knowledge, potentially yielding unsafe advice despite avoiding medication prescriptions. The text-based system's ethical compliance scores (9.5 and 8.8 on a 10-point scale) suggest occasional lapses, such as vague recommendations open to misinterpretation. For the vision system, misclassifications (e.g., mistaking benign conditions for serious ones) could provoke undue alarm or delay proper care, undermining user trust and safety.

5.4 User Interaction and Accessibility Issues:

The Gradio interfaces, though functional, present usability challenges as shown in Khamaj, A.(2025) . The vision system's multimodal input process (image uploads and audio recording) results in longer interaction times (e.g., 15.2s vs. 10.8s for text-only), which may frustrate less tech-savvy users. Limited to English, the systems exclude non-English-speaking populations. The gTTS audio output, while effective, lacks customization options (e.g., speed or tone), potentially hindering comprehension for users with hearing impairments or diverse preferences.

5.5 Technical and Deployment Constraints:

Dependence on the Groq API introduces latency (e.g., 420ms–510ms for text, 2.3s–3.1s for vision) and vulnerability to service disruptions, limiting reliability. Internet connectivity requirements restrict use in remote or offline settings, a critical drawback for telemedicine. Security concerns as discussed in Li, J.(2023) also emerge, particularly in the text-based system, where a hardcoded API key poses a risk if publicly deployed.

5.6 Potential Enhancements:

To overcome the identified limitations, the systems could be enhanced by integrating a hybrid dataset of curated medical images and queries with real-time inputs to improve diversity and robustness, alongside input validation to ensure data quality. Fine-tuning LLMs as shown in Anisuzzaman, D.M.(2025) with domain-specific medical data, employing ensemble methods for vision models, and optimizing via quantization could enhance accuracy and efficiency, enabling deployment on low-power devices. Incorporating a real-time fact-checking layer linked to current medical guidelines, explicit disclaimers, and bias audits would bolster ethical safety. Usability improvements, such as simplified vision inputs, multilingual support, and customizable audio outputs, would broaden accessibility, while caching queries, developing offline-capable models, and securing API keys would reduce latency and enhance deployment reliability, advancing the systems' practical utility in healthcare.

6. CONCLUSION

This research designed and assessed two AI-powered medical chatbot systems: a vision-based platform employing meta-llama/llama-4-scout-17b-16e-instruct and meta-llama/llama-4-maverick-17b-128e-instruct for real-time image analysis, and a text-based framework utilizing the same models for query interpretation. Evaluation using a dataset exceeding 100 medical images and queries demonstrated robust performance, with the vision models attaining accuracies of 92.5% and 94.2%, and the text models achieving 91.4% and 93.5%. These findings underscore the systems' capacity to deliver initial medical insights, with llama-4-maverick-17b-128e-instruct excelling in accuracy and llama-4-scout-17b-16e-instruct offering advantages in processing speed, while their respective applications present trade-offs between resource efficiency and comprehensive responses.

Despite these strengths, limitations such as restricted data diversity, variable model performance, ethical risks, usability constraints, and deployment challenges as discussed in Sadeq, M.A.(2024) temper the systems' readiness for widespread adoption. The vision-based system's multimodal complexity and the text-based system's reliance on internet connectivity underscore the need for further refinement. Comparative analysis suggests a hybrid approach—combining the best-performing vision and text models—could optimize diagnostic and communicative capabilities, tailoring solutions to specific healthcare needs.

In conclusion, these chatbot systems represent a significant step toward accessible, AI-supported medical assistance, particularly for educational purposes as in Robleto, E., Habashi, A.(2024) and preliminary consultation purposes as in Harari, R.E., Ahmadi, N.(2024). The proposed enhancements, including dataset expansion, model optimization, and enhanced accessibility, offer a roadmap for overcoming current shortcomings. Future research should focus on validating these systems with larger, diverse datasets and integrating real-time specialist feedback to bridge the gap between virtual assistance and professional care, paving the way for practical telemedicine applications.

7. FUTURE RESEARCH AVENUES

7.1 Expansion of Dataset and Validation:

As mentioned in Bauer, S.J.(2025), future research should prioritize expanding the dataset beyond the current 100+ real-time images and queries to encompass a broader, more diverse range of medical conditions, imaging modalities, and patient demographics. Incorporating standardized, annotated datasets from clinical repositories, alongside continued real-time collection, would enable rigorous validation of the systems' performance across varied scenarios. This approach could leverage statistical benchmarks and external medical expert reviews to quantify improvements in accuracy, error rates, and generalizability, ensuring the chatbots' reliability for real-world healthcare applications.

7.2 Model Optimization and Integration:

Innovations in models as discussed in Wagner, A.J. and Jürgen, M.(2025) optimization offer a critical pathway to elevate the performance of both vision and text Large Language Models (LLMs). Refining these models with specialized medical datasets, exploring hybrid configurations (e.g., integrating llama-4-scout-17b-16e-instruct and llama-4-maverick-17b-128e-instruct), and developing lightweight versions through methods such as knowledge distillation could enhance diagnostic accuracy while mitigating computational requirements. Additionally as in Eachempati, P., Supe, A.(2025), integrating real-time feedback from specialists—where AI-generated outputs are reviewed or enriched by human expertise—may narrow the divide between automated support and clinical oversight, thereby boosting confidence and applicability in telemedicine environments.

7.3 Accessibility and Deployment Enhancements:

Improving accessibility and deployment readiness as written in Frade, S., Mendonca, R.(2025) offers another critical research direction. Extending language support to include multilingual capabilities, refining user interfaces for seamless interaction (e.g., voice commands or mobile compatibility), and developing offline functionality through on-device processing would broaden the systems' reach, particularly in underserved or remote regions. Security enhancements, such as encrypted API interactions and robust user authentication, alongside scalability testing in diverse network conditions, would ensure practical deployment, aligning the chatbots with the evolving needs of global healthcare delivery.

REFERENCES

Anisuzzaman, D.M., Malins, J.G., Friedman, P.A. and Attia, Z.I., 2025. Fine-Tuning Large Language Models for Specialized Use Cases. Mayo Clinic Proceedings: Digital Health, 3(1).

Badlani, S., Aditya, T., Dave, M. and Chaudhari, S., 2021, May. Multilingual healthcare chatbot using machine learning. In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE.

Bataineh, A.Q., Mushtaha, A.S., Abu-AlSondos, I.A., Aldulaimi, S.H. and Abdeldayem, M., 2024, January. Ethical & legal concerns of artificial intelligence in the healthcare sector. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 491-495). IEEE.

Bauer, S.J., 2025. Responsible Artificial Intelligence in Healthcare: Impact on Patient Trust and Satisfaction With AI-Enabled Clinical Decision Support Systems (Doctoral dissertation, South College).

Chen, D., Huang, R.S., Jomy, J., Wong, P., Yan, M., Croke, J., Tong, D., Hope, A., Eng, L. and Raman, S., 2024. Performance of Multimodal Artificial Intelligence Chatbots Evaluated on Clinical Oncology Cases. JAMA Network Open, 7(10), pp.e2437711-e2437711.

Eachempati, P., Supe, A., Kumbargere Nagraj, S., Cresswell-Boyes, A., Robinson, S. and Yalamanchili, S., 2025. Integrating AI with healthcare expertise: Introducing the Health Care Professional-In-The-Loop Framework: Part 1. BDJ In Practice, 38(2), pp.51-53.

Frade, S., Mendonca, R., Cooper, S., Morris, S., Lee, H., Gupta, K., Sihan Li, B., Ruan, Y., Marucheck, M., Isabelli, P. and Hattery, D., 2025. HealthPulse AI: Enhancing Diagnostic Trust and Accessibility in Under-Resourced Settings through AI. VeriXiv, 2, p.54.

Harari, R.E., Ahmadi, N., Pourfalatoun, S., Al-Taweel, A. and Shokoohi, H., 2024. Clinician-AI collaboration for decision support in telemedicine: a randomized controlled trial study. In Proceedings of conference on cognitive and com (Vol. 102, pp. 81-89).

Hindelang, M., Sitaru, S. and Zink, A., 2024. Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review. JMIR Medical Informatics, 12(1), p.e56628.

Khamaj, A., 2025. Ai-enhanced chatbot for improving healthcare usability and accessibility for older adults. Alexandria Engineering Journal, 116, pp.202-213.

Khan, A., Zeb, I. and Fang, S., 2025. Tracing the Development and Influence of Chatbots in Contemporary Healthcare Systems. In Chatbots and Mental Healthcare in Psychology and Psychiatry (pp. 107-128). IGI Global Scientific Publishing.

Kim, J., Cai, Z.R., Chen, M.L., Simard, J.F. and Linos, E., 2023. Assessing biases in medical decisions via clinician and AI chatbot responses to patient vignettes. JAMA Network Open, 6(10), pp.e2338050-e2338050.

Kumar, S., Rani, S., Sharma, S. and Min, H., 2024. Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. Bioengineering, 11(12), p.1233.

Levine, D.M., Tuwani, R., Kompa, B., Varma, A., Finlayson, S.G., Mehrotra, A. and Beam, A., 2024. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. The Lancet Digital Health, 6(8), pp.e555-e561.

Li, J., 2023. Security implications of AI chatbots in health care. Journal of medical Internet research, 25, p.e47551.

Mazhar, T., Haq, I., Ditta, A., Mohsan, S.A.H., Rehman, F., Zafar, I., Gansau, J.A. and Goh, L.P.W., 2023, January. The role of machine learning and deep learning approaches for the detection of skin cancer. In Healthcare (Vol. 11, No. 3, p. 415).

Robleto, E., Habashi, A., Kaplan, M.A.B., Riley, R.L., Zhang, C., Bianchi, L. and Shehadeh, L.A., 2024. Medical students' perceptions of an artificial intelligence (AI) assisted diagnosing program. Medical Teacher, 46(9), pp.1180-1186.

Sadeq, M.A., Ghorab, R.M.F., Ashry, M.H., Abozaid, A.M., Banihani, H.A., Salem, M., Aisheh, M.T.A., Abuzahra, S., Mourid, M.R., Assker, M.M. and Ayyad, M., 2024. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. Scientific Reports, 14(1), p.18859.

Sarella, P.N.K. and Mangam, V.T., 2024. AI-driven natural language processing in healthcare: transforming patient-provider communication. Indian Journal of Pharmacy Practice, 17(1).

Soe, N.N., Yu, Z., Latt, P.M., Lee, D., Ong, J.J., Ge, Z., Fairley, C.K. and Zhang, L., 2024. Evaluation of artificial intelligence-powered screening for sexually transmitted infections-related skin lesions using clinical images and metadata. BMC medicine, 22(1), p.296.

Soltanian, D. and Ghahari, A., 2024. Comparative Study of Health Data Security under GDPR and HIPAA: Challenges and Implementation Opportunities in Iran. Health Law Journal, 2(2), pp.1-14.

Turchin, A., Masharsky, S. and Zitnik, M., 2023. Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked, 36, p.101139.

Wagner, A.J. and Jürgen, M., 2025. Extending professional military healthcare to defense operations through supportive AI-assistant trained on high quality treatment data. Journal of Critical Care, 86, p.154972.