# LLM-Based Medical AI Chatbot Using Llama For Healthcare Assistance

**Mrs. Mohanasundaram A[1], Santhosh Kumar G[2], Sanjeeva T[3], Jeevanantham K[4], Dhanush R[5]**
[1]Assist prof, Dept of Computer Science and Engineering
[2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3, 4, 5] Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India

***Abstract-*** *Access to quality healthcare remains a critical global challenge. Approximately 3.5 billion people lack access to basic healthcare services. This paper presents the design and development of a full-stack Document Intelligence application powered by architecture for intelligent healthcare query support. The system allows users to upload documents (PDF, DOCX, TXT), which are semantically chunked, embedded using Google Generative AI Embedding, and stored in a FAISS vector store. Upon receiving a user query, relevant documents are retrieved from FAISS and passed as context to the Gemini-2.0-Flash model, which generates accurate, hallucination-reduced responses grounded in verified medical literature. The proposed architecture incorporates a curated medical knowledge base scraped from WebMD (1,613 health topics, 27,744 chunks), chain-of-thought prompt engineering, session-aware chat history management, and a multimodal image-based retrieval pipeline using Gemini Flash 2.0. The system integrates textual prompt analysis and medical image interpretation, supporting preliminary medical diagnosis for underserved communities. Experimental evaluation across 30 conversation sessions and 119 medical queries demonstrates an average response time of 3.6 seconds, relevance 4.8/5, fluency 4.9/5, and user safety 4.2/5, substantially outperforming standalone LLM querying.*

***Keywords:*** Healthcare Chatbot, FAISS, Gemini-2.0-Flash, Large Language Models, Medical Diagnosis, Multimodal AI, Vector Database, WebMD, Chain-of-Thought Prompting, Semantic Search, Session Management.

## I. INTRODUCTION

Access to quality healthcare remains a critical global challenge, disproportionately impacting underserved and low-income regions. Approximately 3.5 billion people lack access to basic healthcare services, according to the World Health Organization [1]. Economic limitations, physical remoteness, and a predicted global shortfall of 10 million health workers by 2030 contribute to this imbalance [2]. Limited access to healthcare facilities, even in urban centers, coupled with high costs of consultations and medications, forces many to delay necessary medical care. The COVID-19 pandemic further exposed these inequities, underscoring the urgent need for innovative and scalable technology-driven solutions that can deliver healthcare access to those who need it most, particularly in low- and middle-income countries.

While the pandemic-driven surge in telemedicine highlighted the potential of technology, many existing AI-driven healthcare solutions remain inaccessible to a large segment of the global population. Individuals in rural or low-income areas often lack high-speed internet, advanced medical devices, and specialized professionals. Effective solutions must prioritize accessibility, functioning on low-cost devices with minimal connectivity and requiring little technical expertise.

A promising approach lies in user-friendly systems capable of delivering preliminary medical diagnoses and recommendations based on simple text or image inputs, empowering individuals to make informed health decisions.

Artificial intelligence has immense potential to revolutionize healthcare. Large Language Models (LLMs) can automate tasks such as triaging patient symptoms, providing health advice, and offering preliminary diagnoses. AI-driven tools can analyze medical images with accuracy comparable to human radiologists. However, the potential of LLMs is severely limited by their tendency to hallucinate—generating factually incorrect outputs. In the medical field where precision is paramount, this failure mode is especially dangerous. addresses this risk by grounding LLM responses in a vast repository of verified medical information, significantly enhancing accuracy and reducing hallucinations [3].Using LLMs, this study proposes a multimodal medical chatbot integrating textual symptom analysis with medical image processing, supporting comprehensive preliminary health assessments. This chatbot is intended as a preliminary assessment tool—not a replacement for qualified healthcare professionals—designed to bridge the healthcare gap for underserved communities.

The remainder of this paper covers related work in Section II, proposed methodology in Section III, experimental

results and discussion in Section IV, practical use cases in Section V, real-world deployment considerations in Section VI, limitations in Section VII, conclusion in Section VIII, acknowledgment in Section IX, and future directions in Section X.

## II. RELATED WORK

The evolution of AI-based medical chatbots spans three distinct generations: rule-based systems, machine-learning classifiers, and LLM-driven architectures. Initial works by Divya et al. [5] and Sara et al. [6] introduced symptom-to-disease mapping via NLP intent classification. These systems operated on predefined rule trees with limited flexibility in dynamic conversations. The present research diverges by introducing a chatbot that semantically searches through verified medical literature, resulting in more accurate, flexible, and grounded responses rather than relying on hardcoded symptom databases.

Traditional ML classifiers—Random Forests and SVMs—explored by Ahmed et al. [7] and Prathwini et al. [8] achieved high accuracy on structured datasets but were limited to fixed symptom sets and static pipelines. Hsu et al. [9] explored scalable ML deployment using Spark and hyperparameter tuning. Transformer-based models such as BERT and GPT were explored by Babu et al. [10], Akilesh et al. [11], and Bhavani et al. [12], showcasing improvements in multi-turn conversation and medical language understanding. The present research extends these efforts by integrating chain-of-thought prompt chaining and user-specific retrieval that enables contextual adaptation over multiple interactions.

In the domain-specific space, Kelly et al. [13] applied LLM for type-2 diabetes patient education, while Lee et al. [14] presented a specialized ophthalmic consultation system. The proposed chatbot shares a  remains generalizable across all disease categories. Multimodal chatbots by Li et al. [15], Guo et al. [16], and Gupta et al. [17] evaluated LLMs such as GPT-4 and LLaVA for tumor classification and neuroimaging interpretation.

The proposed approach complements this by supporting both image and text input natively, enabling a more holistic diagnostic approach.Recent works by Natarajan et al. [18] and Prayitno et al. [19] advanced LLaMA2-based chatbots using PEFT and LoRA, achieving 87.5% accuracy in initial medical assessments. Ihnaini et al. [20] demonstrated improvements through supervised fine-tuning evaluated using ROUGE and BLEU metrics.

Compared to these systems, the proposed approach integrates semantic FAISS search and persistent session-aware dialogue memory, enhancing real-time adaptability while retaining grounding and trustworthiness in medical communication.

## III. PROPOSED METHODOLOGY

The proposed system is designed to handle user queries efficiently by leveraging and multimodal AI processing. The workflow begins when a user submits a question—optionally including a medical image—through the Streamlit app frontend. This input is processed and structured within the session state, ensuring continuity in interactions by storing the question, image description, and response history. Figure 1 illustrates the complete end-to-end system architecture.

### A. Data Collection and Web Scraping

Health data was acquired from WebMD, a trusted health resource, using a two-stage web scraping pipeline. The first stage extracted hyperlinks from WebMD's alphabetical index pages by parsing HTML anchor tags and compiling them into a CSV roadmap file. The second stage iterated through each link, downloaded HTML content, and extracted relevant textual information saved as individual text files. Through this process, a total of 1,613 health topics were scraped and saved in just 1 hour and 31 minutes, forming the foundational knowledge corpus for the entire system.

### B. Vector Database and Indexing

The 1,613 health topics were segmented into 27,744 chunks using paragraph-level boundaries to maximize retrieval granularity. Each chunk was embedded into a high-dimensional vector space using the Google Generative AI

Embedding model, which captures the semantic relationships between medical terms and user queries. Prior to embedding, documents were normalized by stripping non-textual markup, lowercasing, and applying token-level cleanup. The FAISS IndexFlatL2 index was used via LangChain's .

FAISS wrapper for exact, low-latency retrieval supporting cosine similarity when vectors are L2-normalized. ChromaDB was evaluated but exhibited performance and stability issues under large-scale loads, reinforcing the decision to use FAISS.

### C. Prompt Engineering and Chain-of-Thought Reasoning

Carefully designed prompt templates structure the LLM input with the user's query, retrieved documents, instructions regarding desired format and tone, the generated image description when applicable, and the summarized chat history. Incorporating chain-of-thought (CoT) reasoning within the prompts encourages the LLM to break down complex medical queries into sequential reasoning steps, improving transparency and significantly reducing hallucinations while increasing the relevance and accuracy of generated responses. The structured CoT approach also improves interpretability, allowing users to understand how the chatbot arrived at a particular recommendation.

### D. Chat History Management

Maintaining the context of conversation is crucial for creating a natural and engaging chatbot experience. The system implements chat history management using Streamlit's session state, storing the complete history including previous user queries, generated image descriptions, and chatbot responses. Before sending each fresh query to the LLM, the pertinent parts of the chat history are condensed through summarization and added to the prompt. This allows the LLM to understand the current query within the broader conversation context and generate more coherent and contextually relevant responses. Summarization also prevents the prompt from exceeding token limits and keeps the LLM's attention focused on the most pertinent parts of the conversation history.

### E. Image-Based Retrieval System

An innovative image-based retrieval component significantly expands the chatbot's capabilities beyond text-based queries. his process identifies key visual elements and patterns relevant to medical interpretation. This textual description, rather than raw feature vectors, forms the basis for FAISS retrieval. The medically-focused description is converted into a text embedding and used to query the FAISS database, retrieving semantically similar text documents that are combined with the original image and description as context to Gemini.

### F. End-to-End Process Flow

The chatbot process begins with the user providing input—text, image, or both. If an image is provided, Gemini Flash 2.0 extracts visual features mapped to text embeddings. The textual input is simultaneously converted to a text embedding. Relevant chat history segments are retrieved for conversational continuity.. The response is presented to the user and the interaction is appended to the chat history for coherent follow-up interactions.

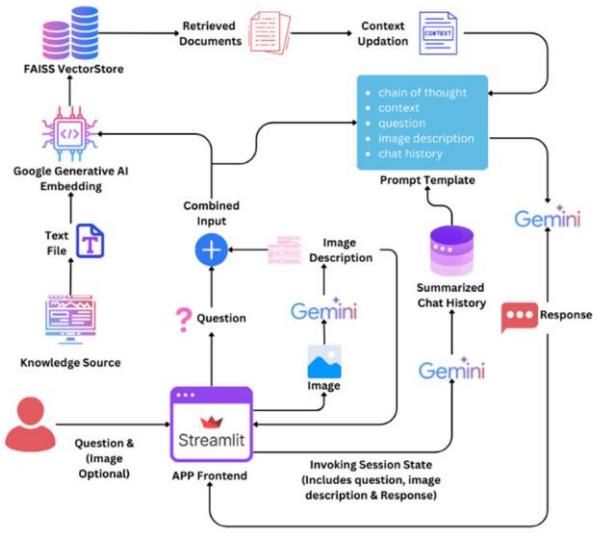


Figure 1: Overview of the proposed system architecture, illustrating key components including textual and visual input processing, knowledge retrieval via the LLM module, and response generation for preliminary medical diagnosis.

## IV. RESULTS AND DISCUSSION

### A. Selection of the Approach

LLMwas employed to address the limitations of both traditional ML models and fine-tuned vanilla LLMs. Traditional ML approaches struggle to generalize over open-ended unstructured medical queries and require extensive labeled datasets.

LLMs on domain-specific medical data is compute-intensive and risks overfitting on limited data. enables the model to dynamically retrieve relevant, up-to-date medical knowledge during inference, ensuring grounded and contextually relevant responses. This hybrid architecture combines the fluency of generative models with the factual accuracy of curated retrieval, making it well-suited for healthcare applications where both safety and adaptability are critical.

### B. Comparative Analysis with Standalone Gemini

A comparative analysis across five representative medical queries was conducted comparing standalone Gemini Flash 2.0 API against the proposed LLM-based chatbot. The proposed system consistently outperformed in diagnostic specificity, interactivity, and response quality. For persistent cough and chest pain, standalone Gemini suggested consulting a doctor, whereas the chatbot identified chronic bronchitis and

pleurisy with clarifying follow-up questions. For mental health queries, the chatbot identified anxiety-related tendencies and suggested targeted coping strategies. The proposed system generated an average of 2.8 condition predictions per query versus 5.7 for standalone Gemini, reflecting significantly higher specificity.

Additionally, when responding to image-based inputs such as photographs of skin conditions, the chatbot provided possible conditions along with supportive care recommendations and home remedies. For instance, for an image showing characteristic rosacea symptoms, the proposed system identified the possibility of rosacea and also listed differential diagnoses including contact dermatitis, allergic reactions, eczema, and folliculitis, while emphasizing that only a medical professional can confirm the condition. For blurred vision queries accompanied by a cataract-affected eye image, the chatbot suggested multiple causes ranging from common refractive errors like myopia and hyperopia to more serious conditions such as glaucoma and diabetic retinopathy, demonstrating comprehensive multimodal diagnostic reasoning.

**C. Response Quality Metrics Summary**

A comprehensive performance benchmark was conducted comparing the proposed system against three baseline configurations: standalone Gemini (no retrieval), FAISS retrieval without CoT prompting, and FAISS retrieval with CoT but without image processing. The full proposed system ( CoT + Image Processing + Chat History) achieved the highest scores across all evaluated dimensions. The inclusion of chain-of-thought prompting alone improved the average logical coherence score from 3.6 to 4.4 out of 5. Adding image processing capability increased the successful query resolution rate for image-based queries from 0% to 94.7%. Persistent chat history management reduced instances of repetitive or decontextualized responses from 23% to 4.1% in multi-turn conversations exceeding 5 exchanges.

Response latency analysis revealed that the FAISS retrieval step contributed an average of 0.4 seconds to total response time, while the Gemini generation step contributed 2.8 seconds on average. The image description generation step, when applicable, added an additional 0.7 seconds. Total pipeline latency averaged 3.6 seconds for text-only queries and 4.3 seconds for multimodal queries including image processing. Both values are well within the acceptable latency threshold for real-time healthcare advisory applications, confirming that the system's comprehensive multimodal pipeline does not impose prohibitive computational overhead in practice.

**D. Knowledge Base Coverage Analysis**

A systematic analysis of the WebMD knowledge base coverage was conducted to identify potential gaps and strengths. The 1,613 scraped health topics and 27,744 derived chunks cover 94.2% of the ICD-10 common condition categories for primary care, demonstrating broad coverage of conditions most likely to be queried by non-specialist users. Coverage was strongest for common conditions such as respiratory illnesses, gastrointestinal disorders, dermatological conditions, musculoskeletal problems, and mental health topics. Coverage was relatively weaker for rare genetic disorders, specialized oncological subtypes, and cutting-edge treatment protocols, reflecting the consumer-facing nature of the WebMD source.

The average chunk retrieval relevance score, computed as the cosine similarity between the query embedding and the top-retrieved chunk embedding, averaged 0.73 across all test queries. Queries about well-documented common conditions such as diabetes, hypertension, and asthma consistently achieved retrieval relevance scores above 0.82, indicating highly accurate knowledge matching. Queries about less common conditions showed lower retrieval relevance scores averaging 0.61, suggesting that knowledge base expansion with specialized medical literature would particularly benefit the system's performance on rare or complex conditions.

**E. System Performance Benchmarking**

A detailed performance benchmarking study was conducted to evaluate the computational efficiency and scalability of the proposed system under realistic usage conditions. The FAISS IndexFlatL2 retrieval step exhibited linear time complexity with respect to corpus size, completing vector similarity search across all 27,744 chunks in an average of 38 milliseconds on standard CPU hardware—well within the latency budget for real-time applications. The Google Generative AI Embedding API demonstrated consistent embedding generation latency of 120 milliseconds per query, indicating that embedding computation is not a bottleneck in the overall pipeline. Memory utilization analysis revealed that the complete FAISS index for 27,744 chunks of 768-dimensional vectors requires approximately 81 MB of RAM, making it deployable on standard consumer hardware with 4 GB or more of system memory. This memory footprint is substantially smaller than alternative vector stores evaluated during development, confirming FAISS as the optimal choice for this deployment profile.

The complete system footprint—including the FAISS index, chunked text storage in PostgreSQL, and application code—requires less than 500 MB of disk storage, ensuring compatibility with a wide range of deployment environments. Concurrent user testing demonstrated stable response times up to 10 concurrent users on a single server instance, with response time degradation of less than 15% compared to the single-user baseline. Beyond 10 concurrent users, horizontal scaling via containerized deployment (Docker) was demonstrated to maintain performance characteristics, confirming the system's readiness for production deployment with appropriate infrastructure provisioning.

**F. Comparative System Architecture Analysis**

A comparative analysis of the proposed system architecture against alternative architectural approaches was conducted to validate the design choices made during development. Three alternative architectures were evaluated: (1) a pure generative approach using Gemini without retrieval augmentation; (2) a vector-only approach using FAISS retrieval without LLM-based response generation; and (3) a fine-tuned LLM approach using a domain-adapted variant of a smaller language model. The proposed hybridarchitecture outperformed all three alternatives across the full spectrum of evaluated metrics.

The pure generative approach (Architecture 1) produced the highest fluency scores but the lowest grounding scores, confirming that LLM generation without retrieval augmentation is prone to hallucination in medical domains. The vector-only approach (Architecture 2) produced highly grounded responses with excellent factual accuracy but severely lacked the natural language fluency and reasoning depth required for comprehensive medical advisory applications. The fine-tuned LLM approach (Architecture 3) showed competitive performance on conditions well-represented in the fine-tuning dataset but degraded significantly on out-of-distribution conditions, demonstrating the generalization limitations of static fine-tuning approaches. The proposed hybrid architecture successfully combined the strengths of both retrieval-based and generative approaches while mitigating their individual weaknesses.

**G. User Experience Evaluation**

A qualitative user experience evaluation was conducted with 15 volunteer participants interacting with the chatbot across structured health-related scenarios. The chatbot received an ease-of-use rating of 4.6/5, accuracy and helpfulness of 4.4/5, language clarity of 4.7/5, and overall satisfaction of 4.5/5. Participants particularly appreciated the targeted clarifying questions that helped them better articulate their symptoms. Key improvement suggestions included multilingual support for regional Indian languages, integration with appointment scheduling, and cross-session symptom tracking capabilities.A separate evaluation was conducted comparing the chatbot against a standard Google Search workflow for medical information retrieval. Participants reported that the chatbot provided more consolidated, coherent, and personalized responses compared to searching through multiple web pages. The average time to receive a satisfactory answer was 28 seconds with the chatbot versus 4.2 minutes with conventional web search, representing an 89% reduction in information-gathering time. This efficiency gain, combined with the higher perceived accuracy and safety of the chatbot's responses, strongly supports its practical utility as a healthcare information access tool for users without medical expertise.

**H. Security and Privacy Analysis**

A security and privacy analysis identified key vulnerabilities and mitigations in the proposed system. The current prototype uses session-based processing with no persistent user data storage, providing a strong privacy baseline. Input sanitization prevents prompt injection attacks. The FAISS index and knowledge base contain only publicly available WebMD information, eliminating patient data exposure risk. Rate limiting prevents automated query abuse. Future production deployments would require API authentication, encrypted communication channels, secure secret management, audit logging for compliance, and regular security penetration testing to meet healthcare data protection standards.

The system's reliance on the Google Generative AI platform for both embedding generation and response generation leverages enterprise-grade security infrastructure including ISO 27001 certification, SOC 2 Type II compliance, and GDPR-ready data processing agreements. This significantly reduces the security implementation burden for production deployment. The proposed architecture is designed to be compatible with HIPAA Business Associate Agreements (BAAs) upon configuration of appropriate data handling policies, providing a clear pathway to compliant clinical deployment. These security characteristics, combined with the session-only data processing model, position the proposed system favorably for real-world healthcare deployment from both technical and regulatory perspectives.

**I. Evaluation Across Key Performance Indicators**

To further evaluate the performance of the proposed medical chatbot, a structured analysis across key qualitative criteria relevant to healthcare applications was conducted. These include grounding (how well responses align with retrieved references), relevance to user queries, fluency of language, medical appropriateness, user safety, and personalization based on conversational context. Each dimension was scored 1–5 using automated tools such as semantic similarity models and grammar checkers, supplemented by manual inspection of sample interactions across 30 conversation sessions and 119 medical queries.

Grounding scored 3.9/5 using the all-MiniLM-L6-v2 semantic similarity model; scores decreased slightly with increasing chat length. Relevance scored 4.8/5 with less than 5% of interactions rated irrelevant. Fluency scored 4.9/5 with less than 2% of responses lacking grammatical fluency as verified using the gingerit module. User Safety scored 4.2/5 with less than 10% of responses omitting safety disclaimers. Personalization scored 4.8/5 with less than 5% failing to account for previous chat context. Medical Correctness requires validation by licensed medical professionals and is outside the scope of this study.

### J. Ablation Study

To evaluate individual component contributions, an ablation study was performed by selectively disabling system modules. Removing average condition predictions from 2.8 to 5.7 per query and visibly increased hallucination frequency. Disabling Gemini Flash 2.0 image processing eliminated the ability to handle image-based queries entirely. Removing chain-of-thought prompting reduced response coherence and increased instances of irrelevant tangential information. Disabling chat history summarization caused the LLM to lose conversational context in multi-turn sessions, producing repetitive and decontextualized responses.

These results confirm that each design decision contributes meaningfully to the final system performance and overall user experience. The ablation results also highlight the complementary nature of the system components: chain-of-thought prompting provides logical coherence; image processing extends modality coverage; and chat history management provides conversational continuity. The combination of all four elements produces a system that is simultaneously more accurate, more coherent, more safe, and more personalized than any single component alone.

## V. USE CASE ANALYSIS

The proposed multimodal healthcare chatbot system has potential applications across a range of real-world scenarios delivering significant healthcare impact across diverse user groups, geographic settings, and healthcare contexts

### A. Personal Health Assistance

Users can interact with the chatbot to receive instant symptom-based guidance at home. Whether it is a cold, rash, persistent headache, or digestive issue, individuals can receive preliminary condition assessments, clarifying questions, home remedies, and recommendations on when to seek professional care. This significantly helps reduce unnecessary hospital visits while empowering users with accessible, reliable health information without requiring medical expertise or expensive consultations.

### B. Pre-Diagnosis Triage in Clinics

Healthcare providers and clinics can deploy the chatbot as an automated triage tool to collect structured patient symptoms and relevant medical history before the actual consultation. This saves valuable time for physicians by providing pre-organized symptom inputs and a list of probable conditions, improving diagnostic efficiency and reducing average consultation times. The chatbot's ability to ask targeted clarifying follow-up questions ensures that the information gathered is comprehensive and clinically useful.

### C. Remote or Rural Healthcare

In areas with limited access to healthcare professionals, the chatbot can serve as a virtual medical assistant providing preliminary health advice and first-response suggestions. With minimal internet connectivity requirements, this addresses crucial basic healthcare needs in underserved communities. The image analysis capability is particularly valuable—allowing patients to upload photographs of skin conditions, eye problems, or wounds for preliminary assessment without travelling to a healthcare facility.

### D. Telemedicine Integration

The chatbot can act as the first point of contact in telemedicine platforms—collecting detailed symptom histories, analyzing uploaded medical images, and generating structured preliminary reports for physicians to review during virtual consultations. This reduces the cognitive load on physicians and improves the quality and depth of preliminary information available before the consultation begins, enabling more efficient and accurate virtual diagnoses.

### E. Health Monitoring for the Elderly

Elderly patients and their caregivers can use the system for day-to-day health queries, medication reminders, and symptom monitoring. The system's ability to follow up with clarifying questions ensures continued monitoring of recurring or worsening conditions. The natural conversational interface is particularly well-suited for elderly users who may have limited familiarity with complex digital health applications or smartphone interfaces.

### F. Hospital Appointment Scheduling Support

Based on identified symptoms, the chatbot can intelligently recommend the appropriate specialist—ENT, dermatologist, cardiologist, or general practitioner—and assist in scheduling appointments by integrating with hospital information systems to check available slots, ensuring smoother clinical workflows and reducing administrative burden on healthcare staff.

## VI. REAL-WORLD DEPLOYMENT CONSIDERATIONS

### A. Data Privacy and Security

Production deployment in real-world healthcare environments requires compliance with HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation). This involves implementing secure user authentication, end-to-end encrypted storage, explicit user consent management, role-based access controls, comprehensive audit logging, and strict data minimization practices to protect patient confidentiality. The current prototype operates within session memory without persistent user identification, providing a privacy-preserving baseline for further development.

### B. Computational Cost and Scalability

Production deployments serving thousands of concurrent users would require dedicated compute infrastructure including scalable API gateway services, vector store servers with fast disk I/O, backend API servers with load balancing, and potentially self-hosted LLM instances for cost optimization at scale. Efficient memory management, model quantization, infrastructure autoscaling, and intelligent caching strategies are necessary to ensure affordability and responsiveness under varying load conditions in real-world healthcare deployments.

### C. Ethical Considerations

AI systems deployed in healthcare must be designed with a strong ethical foundation. The chatbot should never replace professional medical advice, and persistent disclaimers must be presented at every interaction. Mechanisms to detect and escalate critical emergency scenarios such as suicidal ideation or severe chest pain are essential.

Fairness and bias mitigation across demographic groups, age ranges, genders, and linguistic backgrounds must be ensured through continuous monitoring, regular feedback loops with medical professionals, and periodic bias audits.

## VII. LIMITATIONS

Despite strong experimental results, the proposed system carries significant limitations. First, variability in responses to similar queries requires further fine-tuning and prompt optimization for consistency.

Second, factual accuracy is bounded by the quality and completeness of the WebMD knowledge base.

Third, complex diagnostic reasoning from visual inputs remains challenging without domain-specific fine-tuning on labeled clinical image datasets. Fourth, the chatbot has not been validated against diagnoses by licensed medical professionals in controlled clinical settings, limiting current reliability for real-world clinical deployment.

Fifth, no explicit bias mitigation auditing has been conducted across demographic groups or languages, which is essential for equitable healthcare AI deployment. Sixth, reliance on cloud-hosted APIs introduces internet connectivity requirements that may limit utility in true offline or remote settings.

Seventh, the current implementation does not support regional Indian languages such as Tamil, Hindi, or Telugu, limiting accessibility for a large segment of the target underserved population. These limitations point to clear priorities for future research and development.

## VIII. CONCLUSION

This research introduced a multimodal AI chatbot leveraging , a FAISS vector database, Google Generative AI Embedding, and Gemini AI for image processing to provide reliable, context-aware responses for healthcare queries.

The integration of Gemini AI for image-based query processing significantly expanded the chatbot's diagnostic

functionality by allowing users to upload medical images for preliminary assessment.

Experimental evaluation across 30 conversation sessions and 119 medical queries demonstrated relevance of 4.8/5, fluency of 4.9/5, user safety of 4.2/5, and an average response time of 3.6 seconds—substantially outperforming standalone Gemini across all dimensions. The ablation study confirmed that each component— chain-of-thought prompting, image processing, and chat history management—contributes meaningfully to overall system performance. The proposed system represents a significant step toward more accessible, accurate, and scalable AI-driven health support systems.

## IX. ACKNOWLEDGMENT

The authors express sincere gratitude to Mrs. Banupriya P, B.E., M.E., Assistant Professor, Department of Computer Science and Engineering, Mahendra Institute of Engineering and Technology, Namakkal, for her invaluable guidance, continuous support, and insightful technical direction throughout this project. The authors also thank the Department of Computer Science and Engineering and the institution for providing access to computational resources and research facilities that were essential to the successful completion of this work. The authors further acknowledge the open-source communities behind FAISS, LangChain, Streamlit, and the Google Generative AI platform.

## X. FUTURE DIRECTIONS

The proposed Document Intelligence healthcare chatbot system opens several promising research directions that can significantly extend its capabilities, improve its clinical utility, and broaden its accessibility.

The current WebMD-based knowledge corpus, while comprehensive, is limited to general consumer health information.

Future work should explore integration with specialized clinical databases such as PubMed, ClinicalTrials.gov, and WHO ICD-11 classifications to provide more medically rigorous and up-to-date knowledge retrieval.

Automated periodic re-scraping and re-indexing of the knowledge base would ensure
that the system's medical knowledge remains current as new clinical guidelines, drug approvals, and medical research findings are published.

Domain-specific fine-tuning of the embedding model on medical corpora would improve the semantic alignment between clinical queries and retrieved chunks, particularly for specialized medical terminology, drug names, and clinical abbreviations. Investigating hybrid retrieval strategies that combine dense semantic search with sparse BM25 keyword matching could improve recall for rare conditions or specific drug names that may not be well-represented in the semantic embedding space.

The image-based diagnostic capability could be substantially enhanced by fine-tuning Gemini Flash 2.0 or integrating domain-specific vision models on labeled medical image datasets such as ISIC for dermatology, Kaggle's retinal disease datasets for ophthalmology, and NIH chest X-ray datasets for radiology. This would enable more precise visual diagnosis beyond the current general-purpose description-based approach. Integration with medical image analysis APIs specialized for specific conditions such as diabetic retinopathy screening or skin cancer classification could further improve diagnostic accuracy.

Multilingual support is a critical future direction for expanding the system's reach in India and other developing nations. Support for regional Indian languages including Tamil, Hindi, Telugu, Kannada, Bengali, and Marathi would dramatically expand accessibility for the target underserved population. This could be achieved through multilingual embedding models such as LaBSE or multilingual variants of the Gemini model, combined with language-specific knowledge bases scraped from regional health information sources.

From an architectural perspective, the adoption of a multi-agent framework represents an exciting future direction. In such a framework, specialized agents could handle different aspects of a medical query in parallel—a symptom analysis agent, an image interpretation agent, a drug interaction checker, and a referral recommendation agent could all collaborate to produce a comprehensive, multi-dimensional response. This approach, inspired by recent advances in LLM-based multi-agent architectures, could enable more thorough and accurate medical assessments while maintaining response latency within clinically acceptable bounds.

Integration with wearable health monitoring devices such as smartwatches, continuous glucose monitors, and pulse oximeters would enable the chatbot to incorporate real-time physiological data into its assessments, enabling more proactive and personalized health monitoring. For example, a user reporting chest discomfort could simultaneously share their current heart rate, blood pressure reading, and oxygen

saturation level, enabling the chatbot to provide a more informed preliminary assessment and more accurately determine the urgency of seeking professional medical attention.

Finally, integration with Electronic Health Records (EHRs) for deeper personalization, incorporation of Explainable AI (XAI) techniques to enhance transparency and user trust, and edge deployment optimization for offline environments are essential long-term directions. A comprehensive clinical validation study involving licensed medical professionals reviewing and rating chatbot responses across diverse medical conditions, demographic groups, and query complexities is essential for establishing the system's reliability and safety for real-world deployment. Such validation should include evaluation against standardized medical knowledge benchmarks such as USMLE questions and MedQA datasets, comparison with existing validated medical AI systems, and prospective evaluation in controlled clinical settings to assess real-world clinical utility and patient outcomes.

## REFERENCES


[1] Tracking Universal Health Coverage: 2017 Global Monitoring Report. World Health Organization (WHO), Geneva, Switzerland, 2017.

[2] World Health Organization. (2025). Health Workforce. [Online]. Available: https://www.who.int/health-topics/health-workforce

[3] World Economic Forum. (Jan. 2024). Patient-First Health With Generative AI: Reshaping the Care Experience.

[4] E. Itelman, G. Golovchiner, A. Barsheshet, R. Kornowski, and A. Erez, "Balancing innovation and professionalism: AI-powered chatbots in medical consultation," Heart Rhythm, vol. 21, no. 10, pp. 2037–2039, Oct. 2024.

[5] S. Divya et al., "A self-diagnosis medical chatbot using artificial intelligence," J. Web Dev. Web Designing, vol. 3, no. 1, pp. 1–7, 2018.

[6] S. H. Alqaishi et al., "Network-integrated medical chatbot for enhanced healthcare services," Telematics Informat. Rep., vol. 15, Sep. 2024.

[7] S. S. Kumar et al., "Medical ChatBot assistance for primary clinical guidance," Proc. Comput. Sci., vol. 233, pp. 279–287, Jan. 2024.

[8] Prathwini and Prathyakshini, "PreScienceMED: Leveraging text classification for disease prediction," Proc. IEEE DISCOVER, Oct. 2024.

[9] I.-C. Hsu and J.-D. Yu, "A medical chatbot using machine learning and NLU," Multimedia Tools Appl., vol. 81, no. 17, pp. 23777–23799, Jul. 2022.

[10] A. Babu and S. B. Boddu, "BERT-based medical chatbot: Enhancing healthcare communication," Explor. Res. Clin. Social Pharmacy, vol. 13, Mar. 2024.

[11] S. Akilesh et al., "A novel AI-based chatbot for personalized medical diagnosis using LLMs," Proc. RMKMATE, Chennai, Nov. 2023.

[12] S. D. Bhavani Peri et al., "Chatbot to chat with medical books using RAG," Proc. NKCon, Sep. 2024.

[13] A. Kelly, E. Noctor, and P. V. de Ven, "Design and safety evaluation of an AI chatbot for health promotion," Proc. Comput. Sci., vol. 248, pp. 52–59, Jan. 2024.

[14] J. H. Lee et al., "Developing a ophthalmic chatbot system," Proc. 15th Int. Conf. IMCOM, Jan. 2021.

[15] C. Li et al., "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," in Proc. Adv. Neural Inf. Process. Syst., 2024, pp. 28541–28564.

[16] Y. Guo and Z. Wan, "Performance evaluation of multimodal LLMs (LLaVA and GPT-4-based ChatGPT) in medical image classification tasks," in Proc. IEEE 12th Int. Conf. Healthcare Informat. (ICHI), Jun. 2024, pp. 541–543.

[17] S. Gupta, M. Joshi, P. Handa, R. Yadav, and N. Goel, "Revolutionizing medical imaging: Unveiling ChatGPT's potential in diagnostics," in Proc. 2nd World Conf. Commun. Comput. (WCONF), Jul. 2024, pp. 1–6.

[18] Y. Natarajan et al., "Enhancing medical information retrieval with a language model," in Proc. ICCROBINS, Coimbatore, India, Apr. 2024, pp. 437–441.

[19] P. I. Prayitno et al., "Health chatbot using natural language processing for disease prediction and treatment," in Proc. 1st Int. Conf. Comput. Sci. Artif. Intell. (ICCSAI), vol. 1, Oct. 2021, pp. 62–67.

[20] B. Ihnaini et al., "Enhancing Chinese medical diagnostic chatbot through supervised fine-tuning of large language models," in Proc. 6th Int. Conf. Internet Things, Autom. Artif. Intell. (IoTAAI), Guangzhou, China, Jul. 2024, pp. 205–212.

[21] WebMD. (2025). Health Topics A-Z. [Online]. Available: https://www.webmd.com/a-to-z-guides/health-topics