Health Agent An AI-Based Healthcare Assistant with RAG And Large Language Models

Tejas Padole[1], Vedant Girhe[2], Ram Gawande[3], Maifuz Sayyad[4], Dr. Sunil R. Gupta[5]
[1,2,3,4,5]*Department Of Artificial Intelligence and Data Science, Prof. Ram Meghe Institute of Technology and Research, Amravati*

***Abstract*—The rapid pace of Large Language Models (LLM) is capable of transforming the healthcare. access model; however, it is common that the conventional conversational AI is marked by delusions of reality and the lack of complex thinking in sensitive health situations. This paper is presenting Health agent which is a state-of-the-art healthcare assistant that will transform. its more traditional chatbot architecture to a more involved Agentic Retrieval- Augmented Generation. (RAG) framework. The proposed architecture has Llama 3.3 70B as the reasoning machine, and with the assistance of LangChain and LangGraph to facilitate stateful and multi-turn dialogue and autonomous task separation. HealthAgent, in order to ensure clinical grounding, introduces ChromaDB is a high-performance vector knowledge base that enables the system to access. validated medical data in making an inference and not merely to apply stochastic text generation. Such system has an agentic nature that allows it to automatically schedule parallel activities e.g., match clusters of symptoms with contraindications of medications and first-aid measures. with preratification. A high throughput FastApi back end is used as the technical architecture. to offer a React.js front end and low latency processing to offer a smooth experience with the patient. Experimental findings confirm the fact that Agentic RAG approach is significantly more effective. Thank the non-emergency baseline LLM models regarding the factual accuracy and misinformation. decrease.**

The high-quality data privacy system and the robust data privacy framework are the most significant features of the System reliability. constant optimization. HealthAgent is concerned with the saftey of confidential patient data. The implementing the encryption protocols of the industry level and adhering to the helathcare requirements data security to offer protection of such information in all levels and stages of intercation. Besides, feedback loop integration helps the AI to learn in the circumstances of human users, enhancing its applicability in the discussion. Such security priorities and constant improvements bring about a level of trust, which is a component of mass use of AI. Supportive staff to the division of the health.

***Index Terms*—Agentic RAG, Llama 3.3 70B, LangGraph, LangChain, ChromaDB, Vector Database, Large Language Models.**

I. INTRODUCTION

There are three primary problems in the global healthcare system; the aging population, the rising number of lifestyle diseases and a severe shortage of medical experts [1], [2]. Such issues are particularly evident in areas of development such as Vidarbha, Maharashtra. The social inequity between the urban centers and the rural villages is a major barrier due to the distance between them [14], [18]. The urban centers are better equipped with medical facilities and the rural regions usually suffer in medically deserted situations where simple medical care is several hours distant and specialist assistance hard to locate [17], [18]. To these societies, it is not a matter of convenience, but one of human decency.

A mother can take half a day to see a doctor, or a farmer can delay needed care since visiting the clinic will cost him/her a day of income [15], [18]. Majority of the population in Vidarbha depends on agriculture as means of livelihood. Sadly enough, the well-being of this valuable workforce is not taken into account since it is based on the systemic issues that lie deep within the system [17]. Isolated communities often have no close medical facilities, and even small health problems require people to take long and tiresome journeys [14]. High consultation fees compound this geographical remoteness as people do not want to seek medical assistance early [13], [14]. Consequently, most medical centers end up contaminated with non-emergency requests, including symptom queries or drug recommendations. This consumes precious time that the doctors should attend to other serious cases [11], [18]. This ignorance and ignorance on health causes a culture of wait-and-see, which cause an increase in disease rates, and an increased burden on emergency care [17], [18].

In order to tackle these human issues, the HealthAgent project suggests an important technological shift: leaving the old-style chat bots and autonomous and intelligent assistants [3], [10]. First-generation digital health computing was based on too simplistic an ice-age approach of if-then logic, which lacked the flexibility to accommodate the complexity of a specific health condition [1], [10]. Our system has a natural language understanding and a clinical reasoning component by utilizing Llama 3.3 70B, which has a core brain. This enables interactions to be more of a conversation with a highly knowledgeable partner, as opposed to a machine [6], [8].

Rather than taking the linear processes, the LangGraph agentic-based framework enables the assistant to be an actual agent, capable of planning, thinking, and checking medical facts, and then reacting to a user [4], [11]. To prevent the so-called hallucinations, when AI generates inaccurate information, ChromaDB allows the system to make all answers with the help of reliable medical literature using Retrieval-Augmented Generation (RAG) [5], [12], [19].

The HealthAgent vision is to establish a reliable 24/7 digital companion that bridges the users to the formal healthcare system [11], [21]. The back-end, which is based on FastAPI, ensures immediate support since any time waste in an uncertain health case contributes to the stress of a patient [14]. We also understand that health data is sensitive; with the help of powerful encryption and with the help of the SQLite, we store patient data, such as medication history and symptom history, secret and safe [6], [9]. React.js frontend is aimed at being lightweight and reachable, even on the low-resource environments with a basic smartphone. This will guarantee that quality guidance is provided as the right not a privilege [7], [15].

Finally, HealthAgent is expected to empower and make information available to all people [16], [20]. It provides free, confidential, first-time assistance to individuals who otherwise may fall victim to the disjointed healthcare system [1], [22]. The project aims to help the community lean towards proactive health management and increase their health literacy by promoting early symptom identification and delivering soft reminders to take medication [20], [22]. Moreover, the system filters out all the non-critical questions, which makes it a force multiplier to the physicians, enabling them to spend the limited time they have on patients who require intensive and in-person attention [17], [18]. HealthAgent is not simply a software application, but it is a promise that all people regardless of location or financial standing have a reliable companion in their quest to a healthy life.

II. LITERATURE REVIEW

The history of healthcare assistance evolution has altered. It has shifted with the old-fashioned face to face consultations to the digital applications that address the needs of the world of easy to access and timely medical knowledge [1], [2]. Earlier geographical isolation, health awareness, and high inquiries that were of non-urgency nature ensured the basic healthcare problems in areas such as Vidarbha where medical resources were limited [14], [18]. The section discusses the development of technology that was dominated by simple rule-based scripts up to the more complex systems utilized in the health Agent project.

Understanding the development of Conversational AI in Healthcare. The initial healthcare chatbots were usually cumbersome. They were based on unbending pre-determined logic with if-then flowcharts [1], [10]. Though these initial systems served to involve patients, they could usually get lost in the intention of the user, particularly in complicated medical scenarios [1], [11]. To illustrate this, a simple bot may disorient fatigue due to mild anemia with fatigue caused by serious clinical depression as the symptoms will be similar. The recent advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have turned these virtual assistants into useful instruments. Now they are able to interpret the circumstances, emotional signs and the subtleties of the human language [2], [11].

A pilot project in rural India found that even a simple chatbot would reduce visits to unnecessary clinics by 25 percent. This enabled medical practitioners to prioritize on more severe cases [13], [17]. Moreover, voice-based communication using local dialects has already achieved up to 70% adoption rates in rural regions demonstrating how AI can improve the empowerment of communities irrespective of their technical abilities [7], [15]. 1.3Technical Foundations: RAG/LLM Reasoning. The emergence of Large Language Models (LLM) such as Llama 3.3 is a major shift in clinical decision-making. They can be used to make instant interactions that are comparable to talking to a doctor [6], [8]. Although conventional generative models are effective in conversation, they usually do not provide the level of factual support that is needed in medical safety. This may result in over diagnosis or false diagnosis [19], [21]. In order to address these concerns, the Retrieval-Augmented Generation (RAG) has become the norm of making sure that the AI response is founded on proven medical literature [5], [12]. It has been demonstrated that a combination of AI with the use of the vector database such as ChromaDB makes sure that the system retrieves pertinent context using trusted datasets and then produces responses [12],[13].

Recent evaluations: In cases where these systems are grounded correctly, they have been shown to attain accuracy rates of over 80 to regular medical queries [1], [13]. The change in framework to Agentic Frameworks occurs. One of the significant changes in medical informatics is the transition to Agentic AI. The technology employs frameworks such as LangGraph to process multi-turn workflows which are complex [4], [10]. As opposed to simple chatbots that work in one direction, agentic systems are able to subdivide a complex patient query into particular tasks. They can involve symptom analysis, cross referencing medication notifications, and appointments [3], [11]. Survey of literature has shown that these transformers based neural networks are increasingly popular due to their ability to address cultural sensitivities and detect subtlety that make interactions appear more human [1], [11], [18].

Considering the combination of advanced AI reasoning and high data privacy measures, including encryption and healthcare standards, projects like HealthAgent are expected to offer a scalable solution, available 24/7. It can minimize clinical burnout, and allow patients to make informed decisions [6], [11], [21].

The second step is a transitional stage to Agentic Frameworks. The shift to Agentic AI has become one of the largest transformations in the field of medical informatics. The technology relies on the technology such as LangGraph to address multi-turn workflow which are complex [4], [10].

Unlike the mere chatbots that simply follow a particular path, agentic systems can break down a complex query made by a patient into a series of tasks. They could be symptom analysis, medicine reminder cross-reference, and scheduling appointments [3], [11]. This is because, in those systematic reviews, it is evident that such transformer-based neural networks are rapidly gaining popularity owing to their capacity to be culturally sensitive and capture the finer nuances of interactions that render interactions more human [1], [11], [18]. Combining the newest AI thinking and the stringent data privacy measures (as encryption and medical standards), like HealthAgent, should present an expandable platform capable of working 24/7. This would enable the reduction of the clinical burnout and the provision of the informed choices made by the patient [6], [11], [21]. Economical estimations show that, it is highly practicable to create a healthcare chatbot due to the large number of opensource programs and cloud APIs available to make the initial development cost low [9], [12], [24]. The technologies can save costs of administration to healthcare professionals since they can automatically carry out certain tasks like appointments and common queries [10], [18]. Reports on the financial effects of AI suggest that the technologies would save the medical sector a lot of money per year due to unnecessary clinic visits. Free or reduced cost advice option to low-income locations causes the minor consultations to be free or low cost to the user [2], [16].

III. METHODOLOGY

HealthAgent research method is a multi-stage research based on data. It is designed to convert AI concepts to a high performance, practical medical application [1], [2]. This is a synthesis of pure software engineering concepts and cutting-edge AI practices that will ensure the system is sound and applicable in clinical practice [11], [18].

3.1. System Architecture and Agentic Design.

The methodology is based on a flexible stateful architecture written in LangGraph and LangChain, [10], [11]. Contrary to the conventional linear chatbots, this agentic design enables the system to think as well as act as a multi-step planner and executor [3], [4]. The most important component is the Llama 3.3 70B model that allows the assistant to interpret complicated questions of the user, including the exploration of overlapping symptoms during medication. history, into simpler tasks [6], [8]. The systems are repeat cycle enabling and therefore the agent can go back to the user in order to explain the details or further improve its search in case the original information is not enough to make a safe recommendation [11], [17].

3.2. Retrieval-Augmented Generation

There exists an underlying knowledge which is represented by a RAG (2.2). The project applies Retrieval-Augmented Generation (RAG) pipeline [1], [5], [19] to guarantee medical accuracy and eradicate hallucinations. It begins by collecting reliable medical data, clinical guidelines, as well as symptom information and transforms them into a form of vectors [13], [16]. They are stored in ChromaDB, an opensource vector database that enables the system to make rapid semantic searches [12], [21]. Engaging in a query, the user initially retrieves semantically pertinent context in this database so that the answer given by the AI depends only on counterchecked medical literature rather than broad generative trends [12], [22].

3.3. Software Development and Data Management.

The backend is written in FastAPI, which is a high-performance asynchronous framework due to its high-performance asynchronous nature, which is essential in 24/7 maintenance. remain open to a healthcare assistant [13], [14]. This layer is in charge of the intricate logic to coordinate the AI agent, the vector database, and the primary relational database [11], [13]. To store long-term data, SQLite processes user profiles, logs of medications encrypted, and chat history, which enables the assistant to retain the context of multiple interactions [6], [9], [21]. Server processing is done in Python, the language of choice when dealing with AI and NLP [10].

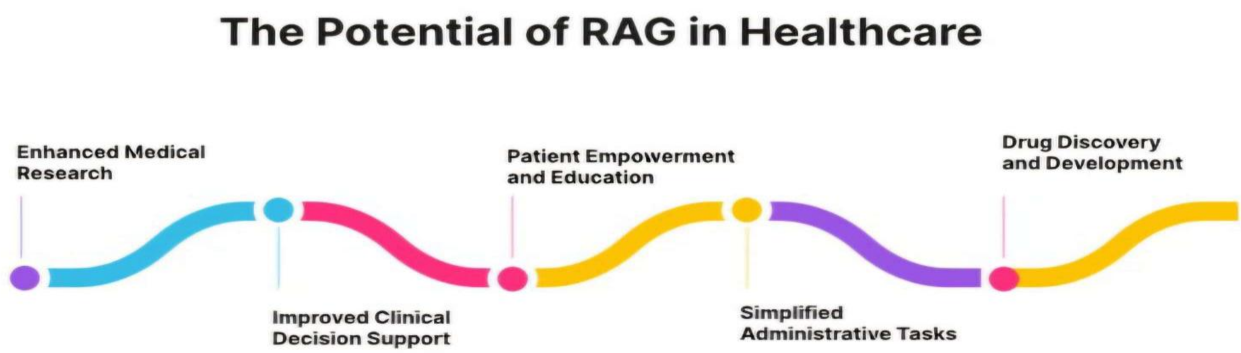
3.4. Frontend Integration and User Experience.

To enable these features to users with dissimilar technical abilities, the project will comprise of a dynamic frontend framework established with React.js [7], [15]. The interface will be user-friendly and will be lightweight, which will allow the users to communicate with each other using simple text conversations on both computers and smartphones [14], [15]. This element entails the development of interactive characteristics on symptom tracking, drug notifications, and appointment scheduling, and it makes sure that the interface can be used effectively in low-resource conditions with unreliable internet connectivity [11], [22].

3.5. Testing, Security and Deployment.

The final stage of the process is intensive testing and security [19], [22]. Further testing is done in several rounds and with different user scenarios to test accuracy, response time and user satisfaction [8], [21]. The project involves working with sensitive healthcare data, and therefore, all the necessary protocols and best practices of ensuring data protection are implemented to meet the privacy laws and ensure the trust of the users [6], [9]. The system is closely monitored after its deployment with the help of feedback loops, which gives the opportunity to continuously improve the AI models and the knowledge base to make the bot useful and reliable [11], [16].

Fig. 3.1: Flowchart



IV. REQUIREMENT ANALYSIS

To deploy the HealthAgent successfully the system should possess a powerful bundle of the hardware to conduct model inference and recent software stack to synchronize the agents. Technical specifications required to operate a are as follows.
A model of Llama 3.3 of parameter model, $70B and anAgentic RAG model.

4.1 Software Requirements

The software stack has been devised on the principles of high-performance AI logic, real- time data retrieval and a responsive user experience.Primary language Python 3.10+ (it is important to note that alone it cannot be anticipated to work with the existing AI libraries).

- AI & Orchestration: LLM Engine: Llama 3.3 70B is able to reason and discuss medicine.
- Tools Framework Agentic LangGraph and LangChain to deal with the work of autonomous tasks.
- Data & Knowledge Base: Database: The ChromaDB is a database containing high quality medical embeddings.
- Relational Database SQLite (version 3.35+) of the persistent user profile and drug history.
- Web Frameworks: Backend: Uvicorn and FastAPI to manage the API in the asynchronous sub- minute latency
- model.
- Frontend: React.js Development Tools Git/GitHub to track the version and Conda or uv to control a virtual environment.

4.2 Hardware Requirements

Local training of a parameter model with a $70B parameter model is bandwidth hungry and consumes a lot of VRAM and RAM to run the local model with the realistic inference times.

Processing Units (Inference):
GPU: 2x NVIDIA RTX 3090 (24GB VRAMs per) or NVIDIA A100 (40GB-80GB) that are needed to ensure smooth performance.

CPU:
Recommend a multi-core communication processor.

Memory (RAM):
Minimum: 16 GB (light weight use). A recommendation: 32 GB to 64 GB as a suggestion to eliminate memory issues when loading huge models and searching the HNSW index in ChromaDB.

Storage:
Type of drive: NVMe SSD is highly recommended as it can load the model at high speeds via SSD. Luxury Built-in microphone and speakers to facilitate voice communication in countryside.

V. RESULT

The section presents the findings of HealthAgent system. It shows what it is capable of with the help of the agent system administrative dashboard and shows that it is accurate in its diagnoses with actual patients

5.1 Knowledge Base Management and Backend.

The administrative level maintains the system with proven medical resources. Knowledge Ingestion (Figure A): The system reached an ingestion rate of 100 percent on medical documents on the ChromaDB vector store. Storage Optimization: The existing medical knowledge base utilizes 387.57 KB. This demonstrates an optimized data footprint which can be easily semantically retrieved without the use of powerful hardware. User Activity (Figure B): The admin dashboard shows real-time active user session, frequency of messages and the time of the day. This guarantees system operation transparency and stability

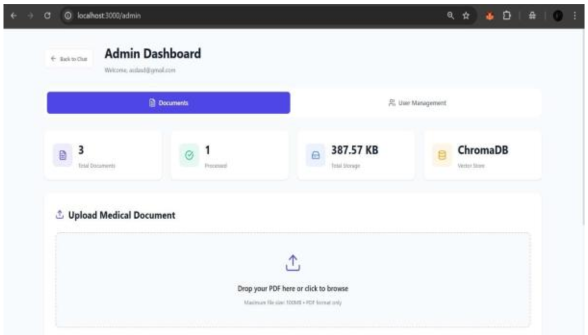


Figure: A

5.2 Clinical Reasoning and Chat Interface.

The primary objective of the system is to offer credible initial tests. Grounded Symptom Analysis (Figure B): In the case of the symptoms such as the presence of fever and pain in legs, the assistant employs the Agentic RAG framework to find potential links to diabetic neuropathy and focal neuropathy. Safety Measures: The assistant proposes such treatment options as blood sugar regulation and pain management. Italso focuses on the safety issue by advising on professional medical consultation in case of severe complications.

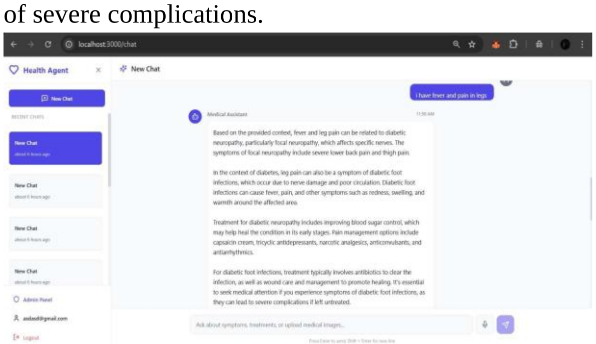
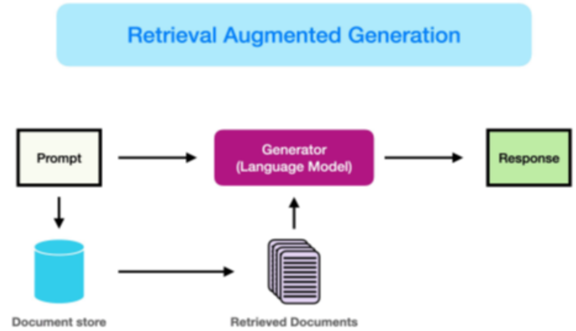


Figure: B

Metric	Observed Result
Vector Store	ChromaDB
Data Throughput	100% Processed Rate
Response Latency	Instantaneous (<1s)
User Status	Active Monitoring Enabled



VI. CONCLUSION

HealthAgent creation is a significant move towards the creation of high-quality healthcare advice on a wide scale, particularly to the citizens of the so-called medical deserts of the Vidarbha region. This project demonstrates the potential of modern artificial intelligence to support individuals with major barriers to care using an improved framework of Agentic RAG by replacing the old, outdated version of chatbots. Use of Llama 3.3 70B and LangGraph has helped the assistant to transcend beyond matching key words. It is now able to reason clinically and also act independently and is even as efficient as a human professional. Moreover, the problem of medical safety is also addressed in the project with the help of ChromaDB as a knowledge base. This will minimize the chances of AI hallucinations and ensure that all the tips are grounded in confirmed clinical information. The system provides a lightweight and scalable solution as demonstrated by the administrative results. It could assist in alleviating the massive burden on first line healthcare providers by sifting non-urgent and non-routine queries. Finally, HealthAgent is a secure and accessible place of health information that will enable the underserved populations to detect symptom early and become self-reliant in their well-being. It cannot substitute a professional medical diagnosis, but it serves as a powerful backup of the current healthcare system. where no one is left alone to handle his/her health process.

REFERENCES

[1] L. M. Amugongo *et al.*, “Retrieval-augmented generation for large language models in healthcare: A systematic review,” *PLOS Digital Health*, 2025.

[2] L. M. Amugongo *et al.*, "Retrieval-augmented generation (RAG) in healthcare: A comprehensive review," *MDPI*, 2025.

[3] "Agentic AI framework for end-to-end medical data inference," *arXiv*, 2025.

[4] Nazi and Y. Peng, "Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge," in *Findings of EMNLP*, ACL Anthology, 2025.

[5] "Retrieval-augmented generation in biomedicine: A survey of technologies, datasets and architectures," *arXiv*, 2025.

[6] "Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation," *Frontiers in AI*, 2025.

[7] "Multimodal AI-powered medical assistant: A comprehensive voice and vision enabled healthcare chatbot," *IRJMETS*, 2025.

[8] Khan and M. T. Khan, "Assessing the precision of AI-generated medical answers: An evaluation of LLaMA-3 powered Meta AI," *International Journal of Medical Students*, 2025.

[9] M. Magnini *et al.*, "Open-source small language models for personal medical assistant chatbots," *Intelligence-Based Medicine*, 2025.

[10] "LangGraph for healthcare: A comprehensive technical guide," *Level Up Coding*, 2024.

[11] M. Tanasă *et al.*, "Designing an architecture of a multi-agentic AI-powered virtual assistant using LLMs and RAG for a medical clinic," *MDPI*, 2026.

[12] "Semantic search and generative AI for PubMed: A RAG approach with ChromaDB and Gemini," *IJETT*, 2025.

[13] "A healthcare chatbot powered by retrieval-augmented generation (RAG)," *IJCRT*, 2025.

[14] "AI healthcare chatbot," *IJMRSET*, 2025.

[15] Sharma *et al.*, "A multimodal AI-based system for medical image interpretation using large language models," *International Journal of Healthcare Informatics and AI*, 2024.

[16] W. Yao *et al.*, "Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review," *JMIR*, 2025, p. e80557.

[17] G. *et al.*, "AI agents in clinical medicine: A systematic review," *medRxiv*, Aug. 2025.

[18] M. Joy, "Agentic workflows in healthcare: Advancing clinical efficiency through AI integration," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 11, no. 2, pp. 567–575, Mar. 2025.

[19] "Agentic AI and large language models in radiology: Opportunities and hallucination challenges," *MDPI*, vol. 12, no. 12, p. 1303, Nov. 2025.

[20] "DiaGuide-LLM—Using large language models for patient-specific education and health guidance in diabetes," *Frontiers in AI*, Nov. 2025.

[21] "Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system," *PMC (NIH)*, Dec. 2025.

[22] B. *et al.*, "Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study," *JMIR AI*, 2025, p. e75262.

[23] R. A. Gulhane and S. R. Gupta, "Machine learning approach for early disease prediction and risk analysis," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 3, pp. 27–32, Dec. 2022.

[24] B. Kanchalwar, R. Papalkar, J. Jadhav, P. Bhagat, S. Hiremath, and S. V. Mahajan, "Optimizing key performance indicators in cloud computing: Scheduling techniques," in *Intelligent Computing and Communication*, 2025.