Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

HealthEdAI: An End-to-End Agentic Health Education Platform

Combining LLaMA-3 Inference, Retrieval-Augmented Generation, and Multi-Role Clinical Workflows

Mrs. J. Hima Bindu[1], C. Yetnesh Reddy[2], Gummalla Pavana Lakshmi Narasimha[3]
[1]Assistant Professor, Department of Information Technology, Mahatma Gandhi Institute of Technology
[23]UG Students, Department of Information Technology, Mahatma Gandhi Institute of Technology

Abstract - The issue of medical expertise versus patient comprehension persists as an ongoing challenge in the discipline of medicine. In many places, patients are provided with medical information concerning their ailments that might not be easily comprehensible to them. There are clear consequences to this situation, including instances where patients fail to take their prescribed medications, medical delays, and avoidable deterioration of medical conditions. The current digital health interventions that exist in our environment have not been successful in solving this challenge. This could be due to the supply of generic medical information that does not take into consideration different health literacies among patients.

The current paper introduces HealthEdAI, which is a production-ready multi-task patient education framework based on the integration of three closely related AI sub-systems: inferencing of the pretrained LLaMA-3.1-8B model using Groq free tier API, crawler for the RAG system, and a ReAct agent empowered by LangChain and utilizing three auxiliary utilities. The relevance checking is one of the key innovations introduced in this study. Rather than blindly returning any retrieved document from the RAG system to the language model, the ReAct agent performs a preliminary verification if it covers the specified disease at all; otherwise, parametric knowledge is used. Thanks to the above, response generation quality was boosted up to 87% for rare diseases compared to 12%.

longside the question-answering facility, the system offers many clinically-oriented services for urgent symptom assessment, drug interaction detection, laboratory reports, physical activity indicators, appointment scheduling, reminders regarding medication dosage, multilingual response translation, and immediate location-based identification of nearest medical professionals through Practo. In total, there are three different user types that are pre-defined upon registration: patient, doctor, and administrator. The evaluation metrics calculated for 120 health queries include FleschKincaid Grade Level of 6.2, ROUGE-1 F1 of 0.74, BLEU-4 of 0.61, and 100% safety rating, outperforming all three baseline models.

Keywords: patient health education, large language models, retrieval-augmented generation, agentic AI, health literacy, LangChain, clinical decision support

I. INTRODUCTION

One of the most important issues that affect the field of medicine but that remain unresolved until today is communication in healthcare. Based on recent research findings, it is estimated that a patient can forget about 40 to 80 percent of the information provided by their physician within one minute of leaving the office visit. Additionally, it is also observed that half of the information that is actually remembered by the patient is wrong.

However, digital health technologies presented a new alternative. Unfortunately, there were two prevalent versions of implementing these technologies, and both of them were insufficiently focused on patients. First, medical websites such as WebMD, Mayo Clinic, and NHS Inform contain highly precise information; nonetheless, this information is not personalized at all. This means that a patient suffering from lupus will receive the same article on proper nutrition as a patient with Type 2 diabetes, since the system would be unable to determine what exact question had been asked by the user. Second, there exist artificial intelligence chatbots and language models that create a possibility for interactive communication but lack any scientific information and only deflect attention from it.

All of these issues are already addressed by HealthEdAI. The information available on the website about health issues is scientifically proven and provides direct answers for commonly occurring health issues, which also includes instructions for taking OTC drugs, including the proper dose as recommended by WHO, NIH, and MedlinePlus. All of these answers have been customized for a particular patient, based on their illness, education level, geographic location, age, and any other medication that the patient might be consuming.

A. Problem Statement

The problems could be defined as follows: (1) Lack of Personalization, where all users get the same material

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

regardless of their illnesses, literacy level, geographical location, and medication they are taking; (2) Deflection, where the AI will never reply to any query relating to a health matter if it could be done using the WHO/NIH/MedlinePlus protocols; thus, it is completely useless for those who search for information concerning their disease; and (3) Lack of Integration of RAG Systems into the Work Process of Healthcare Providers, where there are no systems allowing to incorporate all above-described functions.

Apart from the problem described above, there is one more difficulty associated with technology that makes the situation worse. When an RAG system finds literature related to Addison's disease, the documents retrieved concern generic fatigue management since the corpus does not contain any material on Addison's disease. Thus, the patient cannot detect what causes such a response on the part of the algorithm, and he or she gets irrelevant materials that have nothing in common with the person's problems.

B. Existing System

The current state of patient education strategies can be categorized into three categories, each carrying unique flaws. The patient portal such as WebMD, Mayo Clinic, or NHS Inform supplies general information using complex language without customizing the information based on the patient's level of literacy, age, illness, culture, or medication. Customization and updating of the information provided is not possible since personalization is not available, and no extra information on the patient can be included. Information provided by means of brochures and documents may be out-of-date because of frequent modifications made to the guidelines; furthermore, patients cannot ask questions about the information provided. An artificial intelligence healthcare chatbot is truly a remarkable breakthrough in medical practice; however, it gives answers predetermined to prevent any harm to the program.

Disadvantages of Existing Systems:

- Non-personalised, one-size-fits-all content regardless of patient diagnosis or literacy
- Blanket deflection — refuses to answer routine queries with 'consult a doctor'
- No medicine guidance — will not suggest OTC medicines from standard guidelines
- Static knowledge bases that do not reflect current WHO/CDC/NIH guidelines
- No integrated workflow — doctor discovery, appointments, reminders handled separately
- Generic responses for rare disease patients due to RAG corpus coverage gaps
- No role separation — no doctor-facing or admin-facing data interfaces

PROPOSED SYSTEM

A. Architecture of Proposed System

The system utilizes a nine-module Python backend with more than 25 Flask REST APIs, which it consumes through the React 18 frontend with thirteen pages and role-based routing. The three artificial intelligence layers used—retrieval, agentic reasoning, and generation—are performed in sequence for each patient query.

The live web scraping is carried out by the Retrieval layer to scrape the currently existing texts from WHO, NIH, CDC, MedlinePlus, and Mayo Clinic websites per each query provided by the user. The live scraped texts are segmented into text blocks with 400 words each with an overlap of 80 words between two neighboring text blocks to ensure sentence context. Afterward, the text blocks are encoded into vectors using all-MiniLM-L6-v2 sentence transformer model which encodes texts into 384 dimensional vectors.

In the case of the Agentic Reasoning Layer, we use a LangChain AgentExecutor with a custom ReAct (Reasoning + Acting) prompt. The three functions that are used to operate the tools are created for each request individually by means of Python closures and are stateless, meaning that the same function may be executed by different users simultaneously without influencing each other during their operation. To start with, Tool 1 retrieves the patient's whole profile prior to taking any decision about searching the needed information, relying on clinical data concerning the patient in particular. The second tool searches a FAISS vector by a query that relates to the specified condition but not the patient's request in order to get as accurate results of the procedure as possible. The third tool filters the selected documents according to two criteria: the average cosine similarity between the documents should be 0.35 or above and the disease of the patient is mentioned as one of the keywords in the document. If those criteria are satisfied, the system will function in the RAG mode, otherwise it will be operating in the knowledge mode.

The first inference service that we used is LLaMA3.1-8B-Instant from Groq's free API service, while the second inference service is offered by Huggingface router, and the third inference service is google/flan-t5-base from the locally installed source in case there is no internet connectivity. The system prompt needs the model to respond to the queries related to health topics with recommendations for appropriate OTC medications based on purpose, drug names, and dosages for adult patients extracted from WHO, NIH, and MedlinePlus sources, along with one disclaimer for each response. The system prompt forbids starting responses for queries related to public health topics with advice to see a doctor.

The users' roles in the overall data access process have been defined into three types. The patients will have access to all the features of the clinical software package. The physicians can access a condensed form of the data of those patients that have linked up with them, which include the features like

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

diagnosis, medications, and health data generated by the AI. The information on all the patients in the system and their interaction with the AI is visible to the administrators.

Advantages of Proposed System:

- Direct, evidence-grounded answers with OTC medicine suggestions — no deflection
- Relevance-gating raises rare/custom disease accuracy from 12% to 87%
- Live web crawl ensures responses reflect current WHO/CDC/NIH guidelines
- Three-role clinical workflow with appropriate data boundaries per role
- End-to-end health companion: doctor finder, symptom triage, health tracking, reminders
- 100% safety validation with FKGL/BLEU/ROUGE metrics on every response

II. LITERATURE SURVEY

Retrieval-Augmented Generation was first proposed by Lewis et al. [1]. It was found that incorporating retrieval for generation can aid in minimizing the rate of hallucinations and help ensure facts and truths during tasks needing detailed knowledge. In keeping up with this framework, HealthEdAI replaces datasets indexed beforehand with webcrawling to get responses that conform to current medical guidelines.

The Speculative RAG algorithm was introduced by Wang et al. [2], where a small drafter model generates candidate answers that are validated by a large verifier model. Although this enhances the latency efficiency when applied to popular natural language processing tasks, there is no analysis conducted to see whether this is accurate from a healthcare perspective, and whether the content produced takes into account the patient's reading level, something that HealthEdAI addresses.

The ReAct framework was developed by Yao et al. [3], demonstrating that reasoning intermixed with chain-of-thought and tool use from the environment leads to markedly improved multi-hop behavior than when using prompts only. The HealthEdAI system implements the ReAct framework via LangChain AgentExecutor and adapts the reasoning chain of three steps to fit patient education needs.

In this regard, according to He et al. [4], there is a systematic review of RAG techniques that can be used in large language models in healthcare, which states that the increase in factual accuracy was observed; however, the unchanged problem in benchmarking patient-centric evaluation persists. Thus, in the current case, it is important to emphasize the absence of methods involving the application of both skills in fact-finding and the adaptation of facts to the level of readability of laypeople.

The sentence transformers framework was originally introduced by Reimers and Gurevych [5], where the authors showed that the siamese BERT architecture is able to create dense embeddings for capturing the semantic features that can be used for similarity search. In this case, HealthEdAI uses allMiniLM-L6-v2 from the sentence transformers library for creating 384 dimensional vectors used for FAISS search.

The paper published by Johnson et al. [6] that demonstrated the ability to perform vector searches using billions of vectors in GPU using FAISS contributed to making FAISS the most suitable solution for retrieval operations. HealthEdAI leverages the use of FAISS in CPU-based flat inner product search, not requiring any GPU component.

Gan et al. [7] carried out an analysis of existing RAG performance evaluation metrics, identifying inconsistencies in the measurement of factual grounding and response quality. The requirement for metrics that will integrate readability, factual correctness, and comprehension accuracy is fulfilled by four separate per-response metrics adopted by HealthEdAI: FKGL and SMOG readability scores, BLEU-4 score for precision of n-grams compared to the retrieved text, and ROUGE-1 F1 score for wordlevel recall.

The study conducted by Wang et al. [8] highlighted the issues concerning RAG in the context of medical and nursing education literature, emphasizing the need for standardization with respect to safety and readability criteria that favor patients. The validation process involved in the case of HealthEdAI takes into account the issue of standardization of safety criteria highlighted above by applying the FKGL criterion.

As per Sheeran & Kass-Hout [9], the following three basic issues that the healthcare sector currently encounters can be overcome by implementing an agentic AI model: personalization, automation of the process, and preventive measures. In spite of this being a theoretical paper without any empirical studies, it illustrates the rationale for developing the HealthEdAI agentic framework. The three operations executed by the LangChain library, i.e., get_patient_context, retrieve_medical_docs, and check_rag_relevance, exemplify this answer mechanism.

Data privacy and clarity of ethics have been identified as two primary challenges of the application of RAG in the healthcare domain according to research conducted by Amugongo et al. [10]. In addressing such challenges, HealthEdAI has adopted the strategy of limiting the availability of data based on the roles that individuals play in an organization. Specifically, doctors can access only health summaries and sources of all responses.

Quantitative analysis of the relationship between health literacy and drug adherence was conducted by Valero-Chillerón et al. [11] to show that health illiteracy had quite a big impact on non-compliance and probability of admission. The recommendation to conduct interventions using literacy-based AI educational tools is supported by the practical reasoning behind HealthEdAI literacy-adaptive content generation at the FKGL reading level of preference for each person.

Hallucination and obsolescence of knowledge have been determined to be the two most enduring failure points of

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

standalone language models by Gao et al. [12]. The healthEdAI approach of live crawling explicitly tackles the obsolescence issue, whereas the relevance gating process is effective against the retrieval-induced hallucination phenomenon wherein the model produces plausible but irrelevant answers upon retrieval failures.

Bionghi et al. [13] reviewed the implementation challenges of GenAI in healthcare from an implementation-science perspective, emphasising governance, evaluation, and integration with clinical workflows. The study noted a gap between highlevel strategic frameworks and practical, patientfacing implementations. HealthEdAI bridges this gap as a fully deployed, end-to-end system with logging, metric scoring, and role-based access — demonstrating practical implementation of the governance principles Bionghi et al. advocate.

Given that the release of the LLaMA-3 foundation model [14] from Meta AI is in an open-access framework, one can realize that it is possible to obtain quality language generation despite the limited 8 billion parameter size with no cost involved. Therefore, the use of the LLaMA-3.1-8BInstant model provided by HealthEdAI on Groq API is feasible in the context of clinical NLP models.

In particular, the findings of the research conducted by Rojewska et al. [15] about medication nonadherence interventions showed that patient understanding and access to information constitute factors that are easy to change compared to other factors. Specifically, personalized education that considers patients' literacy was always more effective compared to disseminating general information, forming the basis of HealthEdAI's approach.

III. RESULTS AND EVALUATION

We evaluated HealthEdAI on 120 health-related questions, covering both common and rare diseases, medications, and lifestyle-related topics with the outputs constrained to FKGL ≤ 6.0. As baselines, we consider three approaches: RAG-only, agentic, and direct LLM use without retrieval.

HealthEdAI showed an FKGL score of 6.2 and SMOG of 7.1 with guaranteed readability and grounding (BLEU-4 0.61, ROUGE-1 0.74). The primary distinction between the two comes from the improvement of the disease-specific question accuracy from 12% (RAG-only) to 87%.

Surprisingly, the gain does not result from the improved ability to retrieve the sources; it stems from failure handling. Indeed, in the case when a relevant source could not be retrieved, RAG-only approaches generate topical but inaccurate results. In contrast, the HealthEdAI model detects the inconsistency and creates an answer relying solely on the language model's knowledge while maintaining relevance.

Finally, the direct use of a language model yields a 61% accuracy but provides less grounded outputs (BLEU-4 0.38, ROUGE-1 0.51) along with the 6% safety violation rate.

In conclusion, the results imply that the ability to handle retrieval failures outweighs its success in improving disease-specific answers' accuracy.

IV. CONCLUSION

HealthEdAI is the full-fledged GenAI-powered solution for patient education that addresses the three key areas where the current health information systems fail: lack of personalization, defensive deflection, and lack of integration with clinical workflow.

The main technical contributions include: (1) a new relevance-gating function in a LangChain ReAct agent that improves the accuracy of rare and personalized disease responses by 87% through the detection and correction of RAG errors; (2) a new architecture for LLM system prompts with provision of direct evidence-driven answers without resorting to defensive deflection and providing suggestions on over-the-counter medications based on WHO, NIH, and MedlinePlus guidelines; (3) a 3-role clinical platform with automatic role detection, live search for doctors using Practo and a set of 10 clinical features for patients' educational needs.

The validation through the entire evaluation model establishes the success of the platform: FKGL 6.2 (can be understood by someone who has basic literacy skills), BLEU-4 0.61 and ROUGE-1 0.74 (has excellent facts based on the information collected), and a 100% safety check in all test questions. This platform works completely through free-tier cloud computing services, allowing its usage in resource-limited healthcare environments without the need for GPUs or any API token fees.

The next steps will involve three aspects: learning the threshold value 0.35 based on the condition, integration with India's National Medical Register for the verified government doctors' database, and multilingualism expansion from machine-translated LLMs to native Indic language models (IndicTrans2, Airavata).

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[2] Z. Wang, K. Zhao, and J. Lee, "Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting," Google Research Technical Report, arXiv:2407.08223, 2024.

[3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. of the 11th International Conference on Learning Representations (ICLR), 2023.

[4] Z. He, J. Zhao, and L. Chen, "RetrievalAugmented Generation for Large Language Models in Healthcare: A Systematic Review," PLOS Digital Health, vol. 3, no. 4, pp. 1–20, 2025. doi: 10.1371/journal.pdig.0000877

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERTNetworks," in Proceedings of the 2019 Conference on Empirical Methods in Natural

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Language Processing (EMNLP), pp. 3982– 3992, 2019. doi: 10.18653/v1/D19-1410

[6] J. Johnson, M. Douze, and H. Jégou, "BillionScale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021. doi: 10.1109/TBDATA.2019.2921572

[7] A. Gan, Y. Song, and J. Liu, "Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey," arXiv:2504.14891, 2025.

[8] Z. Wang, et al., "Improving Large Language Model Applications in the Medical and Nursing Domains With Retrieval-Augmented Generation: Scoping Review," JMIR Medical Education, vol. 11, p. e80557, 2025. doi: 10.2196/80557

[9] D. Sheeran and T. Kass-Hout, "How Agentic AI Systems Can Solve the Three Most Pressing Problems in Healthcare Today," GE HealthCare Insights, 2024. Available: https://www.gehealthcare.com/insights

[10] L. M. Amugongo, J. Kachoka, and T. Mwaka, "Bridging AI and Healthcare: A Scoping Review of Retrieval-Augmented Generation — Ethics, Bias, Transparency," medRxiv, 2025. doi: 10.1101/2025.04.01.25325033

[11] M. J. Valero-Chillerón, et al., "Health Literacy and Medication Adherence in Polypharmacy: A Systematic Review and Meta-Analysis," PMC / PLOS ONE, 2025. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12360272/

[12] Y. Gao, T. Li, and H. Zhang, "RetrievalAugmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023. Available: https://arxiv.org/abs/2312.10997

[13] N. Bionghi, S. Norris, and R. McMahon, "Leveraging Artificial Intelligence to Advance Implementation Science," Implementation Science (BMC), vol. 19, p. 17, 2024. doi: 10.1186/s13012-024-01346-y

[14] Meta AI, "Llama 3: Open Foundation and FineTuned Chat Models," Technical Report, Meta Platforms Inc., 2024. Available: https://llama.meta.com

[15] A. Rojewska, P. Tomaszewski, and M. Wysocki, "Enhancing Therapy Adherence: Impact on Clinical Outcomes, Health Expenditures, and Quality of Life," Medicina (MDPI), vol. 61, no. 1, pp. 1–22, 2025. doi: 10.3390/medicina61010001

(This work is licensed under a Creative Commons Attribution 4.0 International License.)