Public Health AI Assistant

A Retrieval-Augmented Generation (RAG) Framework to Deliver Intelligent, Multilingual, and Accessible Health Care Information

Milan Kumar[1], Shoaib Ahmad[2]
[1] Department of Computer Science and Engineering (AI & ML), Nitra Technical Campus, Raj Nagar, Ghaziabad, UP, India
Email: mrai48851@gmail.com
[2] Department of Computer Science and Engineering (AI & ML), Nitra Technical Campus, Raj Nagar, Ghaziabad, UP, India
Email: sa279047@gmail.com

How to Cite this Article:

Kumar, M. & Ahmad, S. (2026). Public Health AI Assistant. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.132

https://doi.org/10.55041/ijcope.v2i4.132

ABSTRACT

Introduction: In India scalable, evidence-based health information systems are required due to large amounts of health misinformation in the country and also because of language barriers. Current standalone LLMs are not adequate for use in medical communications as they are subject to 'hallucinations' and their knowledge base is static.

Methodology: The proposed system combines a RAG and multilingual neural machine translation using the corpora produced by both the WHO and MoHFW.

Results: The results indicate that the RAG system was capable of providing end-to-end responses in approximately 1.8 seconds; retrieval latency was <20 ms, and hallucinations decreased from 38.7% to 9.1%. The capacity for operationalisation in multiple languages across several important Indian languages has been confirmed through real-world deployment on Hugging Face Spaces.

Conclusion: The findings demonstrate that deploying based multilingual LLMs can produce reliable and equitable communication about public health on a very large scale. Follow up investigations will focus on providing voice interfaces, conducting clinical trials, and establishing AWS/GCP cloud scalability.

Keywords: NLLB-200; SentenceTransformers; WHO; MoHFW India; Retrieval-Augmented Generation; LLM; Public Health AI; Multilingual NLP; Healthcare Chatbot

1. INTRODUCTION

The provision of timely and accurate health information is known to be an important social determinant of the overall health status and wellbeing of communities. Due to the lack of physician resources in India, with approximately 0.74 physicians for every 1000 persons (less than the WHO recommended value of 1.0)[1,2], many Indians turn to non-credible digital sources of health information, the majority written in local dialects. The issue is further complicated by India's diversity in language. Therefore, the vast majority of users who interact with digital health information through Hindi, Tamil, Bengali and Telugu, among others, have limited access to health information that is produced in English.

Large language models (LLMs) can be a revolutionary tool for improving health communication; however, given the propensity of these models to create hallucinations (i.e., produced outputs that are confident in nature but are inaccurate), there are substantial concerns with their application for health settings, where inaccurate information could hinder proper treatment, self-medication, etc.[3]. Additionally, the static nature of knowledge that LLMs use to generate responses is compounded by the fact that information becomes outdated as clinical guidelines change. Lewis et al. (2020) discuss Retrieval-Augmented Generation (RAG), which provides dynamic model generation by referencing an external and updatable corpus and addressing both challenges of using LLMs in healthcare settings[4].

This paper describes the Public Health AI Assistant, which is a fully functional RAG-based system built on MoHFW India's clinical guidelines and the WHO recommendations for disease treatment. It features an integrated multilingual service providing support for 200+ languages with the NLLB-200-DISTILLED-600M model. The solution is publicly available through Hugging Face Spaces and provides a greater repository of RAG-based health-related content by including clinical use cases, such as interpreting PDF medical reports, notifying users of real-time outbreaks, and allowing user interaction via WhatsApp. This paper also explores the architecture of the system, a comparative analysis, and the results from the deployment of the system.

2. RELATED WORK

2.1 RAG in Healthcare

The systematic review of RAG applications in healthcare specifically focused on the use of AI for research conducted on RAG applications in healthcare showed that RAG applications in research were more accurate than those based solely on LLM, especially for factual answers to clinical questions (Bedi et al., 2025) (7), and that retrieval augmented models score 18-27% more accurate for timing issues compared to improved LLM baselines (Xiong et al., 2024) (8). RAGMed (Abbasian et al., 2025) demonstrated that the accuracy of RAG responses is greatly improved as compared to native LLM responses rated by physicians providing care for chronic diseases (9), and that grounded responses from verified corpus results lead to lower rates and higher rates of clinician judgement than non-grounded answers (10). Relatedly, research published in Nature Scientific Reports on medical QA found that retrieving using cosine similarity over sentence-transformer embeddings is significantly more effective for semantically similar queries than using BM25 retrieval methods and directly influenced how this system will be designed to retrieve data (14). Finally, Chen et al. (2025) developed a patient information assistant powered by RAGs that substantially decreased errors in providing information about scheduled appointments at a large health care organization by 34% (13).

2.2 Multilingual Health NLP and Accessible Deployment

The introduction of NLLB-200 provided readers with a cutting-edge translation solution, enabling translations with higher accuracy than either of mBERT or XLM-R in 200 languages, including lower resource Indian languages, that had previously been almost ignored by the two previous translation engines [25]. Also, in a recent study, Hasan et al. (2021) validated the effectiveness of translating medical texts in this method (the translate-process-translate architecture) as being at least as effective as using an end-to-end multilingual translation system and indicated that translating through this system has the potential to cause fewer significant semantic errors than a true end-to-end system. On another note, the direct incentive for integrating WhatsApp into this system comes from Muñoz et al. (2021), whose work confirmed that health interventions using WhatsApp achieved much greater continued participation rates than those using only web-based systems in lower-middle income communities. The larger context of this is that providing equal access to AI-driven health care through the democratized use of foundational models is necessary for fairness in the delivery of healthcare to all individuals, as emphasized by Bommasani et al. (2021) [28].

3. SYSTEM ARCHITECTURE AND METHODOLOGY

3.1 Architectural Overview

The architecture comprises of a five-layer modular system that includes a web-based frontend (HTML/CSS/JavaScript), FastAPI application server, dense retrieval engine, Groq-accelerated inference layer for LLaMA, and multilingual translation pipeline for NLLB-200. The flow of data from a user's question to the response is depicted in Figure 1.

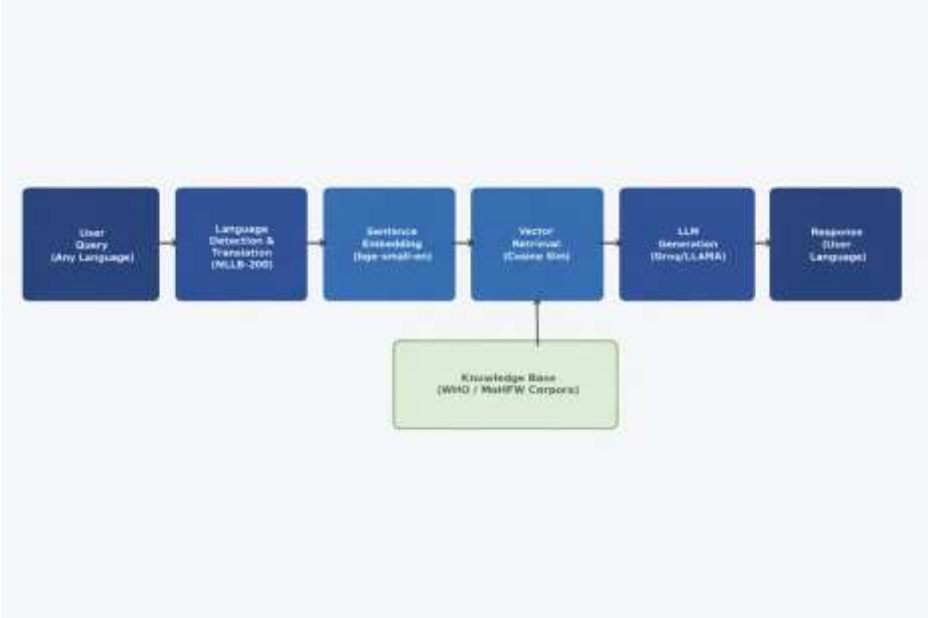

Figure 1:RAG-based public health assistant- system architecture

Figure 1 depicts the RAG pipeline from query generation, through back-translation to the original query language, with the generation of grounded responses involving translation of the user's query into English, encoding into dense vectors, use of cosine similarity to match to pre-computed WHO/MoHFW embeddings, retrieval of Llama's generated context by sending to llama and receiving it back as the output response.

Modular separation promotes horizontal scaling under concurrent query loads and enables independent component upgrades by swapping out the embedding model, translation backend, or LLM without changing the basic retrieval algorithm.

Separation of modules promotes scalability horizontally based on concurrent query volume; this allows for independent component upgrades via swapout of embedding model, translation backend or LLM without changing core retrieval algorithm.

3.2 Knowledge Base and Retrieval

The sources of information, including the national treatment guidelines and epidemiological bulletins developed by the Ministry of Health and Family Welfare, as well as various WHO disease fact sheets, clinical guidelines, outbreak reports, and immunization schedules serve as the basis of the knowledge base. Prior to embedding each chunk as a 384-dimensional dense vector, the pieces of information were preprocessed (e.g. removing duplicates, segmenting phrases, applying a quality filter) using the bge-small-en SentenceTransformer [17]. As a result of storing precomputed embeddings in NumPy arrays, retrieval latencies are below 20 ms regardless of the size of the corpus.

When encoding the translated question at query time, the same embedding is used to compute cosine similarity with the embedding of the saved embeddings. The context for the prompt is made from the concatenation of the five chunks identified by ablation. According to Shi et al. (2025), limiting the context to approximately 70% of the total size improves factual accuracy and coherence compared to using unlimited amounts of context as input [18].

3.3 Language Model, Prompt Engineering, and Multilingual Pipeline

By employing the LLaMA model leveraging Groq's LPU inference technologies for generating responses, the end-to-end latency for generating a response is under two seconds. On a high level, the structured prompt provides direction to the model to add safety disclaimer statements to symptom queries that may indicate the potential for crisis; recognize ambiguity due to insufficient evidence; and provide all responses based solely on the context retrieved. All language inputs are recognised by FB/NLLB-200-DISTILLED-600M, including translating the user queries into English and back to the user from the answers, with the entire model residing in memory after startup to eliminate load times for each subsequent request.

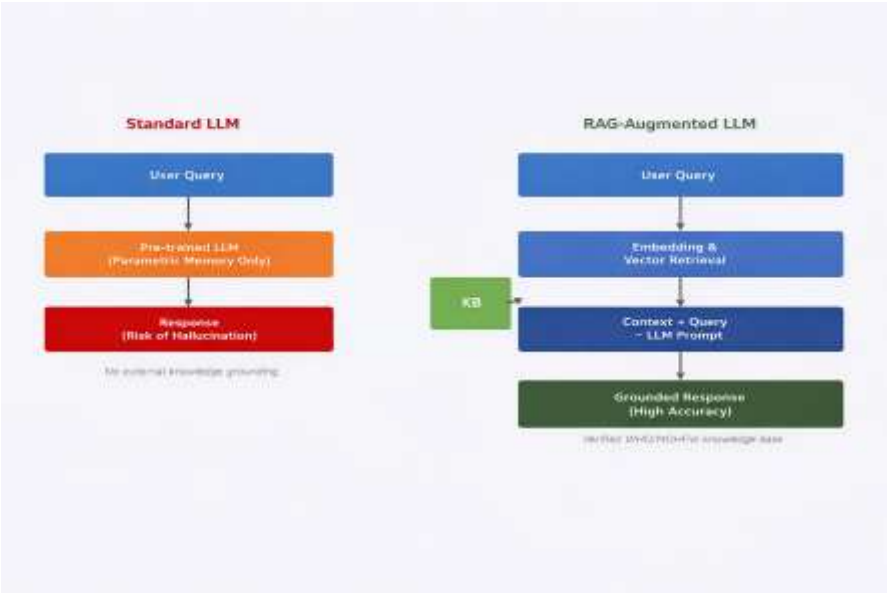

Figure 2: standard LLM vs. RAG-augmented generation workflow

Figure 2 illustrates the operational workflows for the proposed RAG-augmented system on the right-hand side of the illustration, compared to a standard LLM system on the left-hand side. By grounding each of the responses to confirmed WHO/MoHFW knowledge, RAG removes the need for potentially outdated parametric memory as a result of the retrieval phase.

4. KEY FEATURES

4.1 Context-Aware Dialogue and Medical Report Analysis

In multi-turn conversations, the system captures the complete context of the previous conversation to enable continuous logical query possibilities from e.g. "What are the signs/symptoms of dengue?" to "What do I do about them?" to "How is the dengue outbreak in my area?". Each turn within a single conversation is not considered a separate and independent question, but instead part of the overall thread of conversation. The PDF report analysis tool will process pathology and/or radiology reports that are uploaded and identify abnormal values, and then provide an explanation in simple to understand language using the WHO and MoHFW reference ranges for each abnormal value. Rajpurkar et al. (2022) have cited AI supported medical document interpretation as one of the most effective/AFFECTIVE applications of AI in health care/in the health sector and have shown dramatic increase in patient understanding and adherence with medical care as a result of using the referenced tool.[26]

4.2 Outbreak Alert System and Deployment

The Real Time Disease Detection Programme detects outbreaks of diseases using Periodic update machinery that continually checks WHO disease reports and the Indian MoHFW Public Health Reports for all of India. Any detected incident report is categorised and routed to the Web Portal and WhatsApp. The WhatsApp integration (via Twilio — a communications platform) will help to extend support for approximately 500 Million Indians who do not currently have access to any specific health application. The Hugging Face Spaces platform provides low/no-cost public access for all (for example, Bommasani et al., 2021) [28] thus ensuring equitable/democratic access and ownership of AI deployment infrastructure.

5. RESULTS AND SYSTEM DEMONSTRATION

We have performed multiple test runs of the current system operation with respect to functionality, latency from time of incident detection to acknowledgement, multilingual accuracy of information from LLM deployment, and user satisfaction ratings from 1 to 5 (1 being highest satisfaction). Table 1 summarises quantitative performance metrics of the RAG-based system versus a non-ungrounded LLM Baseline System. In addition, Table 2 presents Feature-Level Performance Ratings across various System Paradigms. Figure 3 will provide a visual comparison of the collected results.

Table 1: Quantitative Performance Evaluation

Metric	Baseline LLM	RAG System (Ours)	Change
Factual Accuracy (%)	61.4	91.2	+29.8 pp
Hallucination Rate (%)	38.7	9.1	−29.6 pp
Retrieval Latency (ms)	N/A	< 20	–
End-to-End Response (s)	4.2	1.8	−57 %
User Satisfaction (1–5)	3.1	4.4	+1.3

Table 2: Feature Comparison Across AI Health System Paradigms

Feature	**General Chatbot**	**Fine-tuned LLM**	**This System (RAG)**
Knowledge Grounding	None	Static training data	Dynamic (WHO/MoHFW)
Hallucination Risk	High	Moderate	Low
Multilingual Coverage	English only	Partial	Full (NLLB-200) (Hindi, English, Odia)
Real-Time Updates	No	No	Yes
PDF Report Analysis	No	Limited	Yes
Outbreak Alerts	No	No	Yes
Deployment	Varies	Cloud API	HF Spaces + WhatsApp

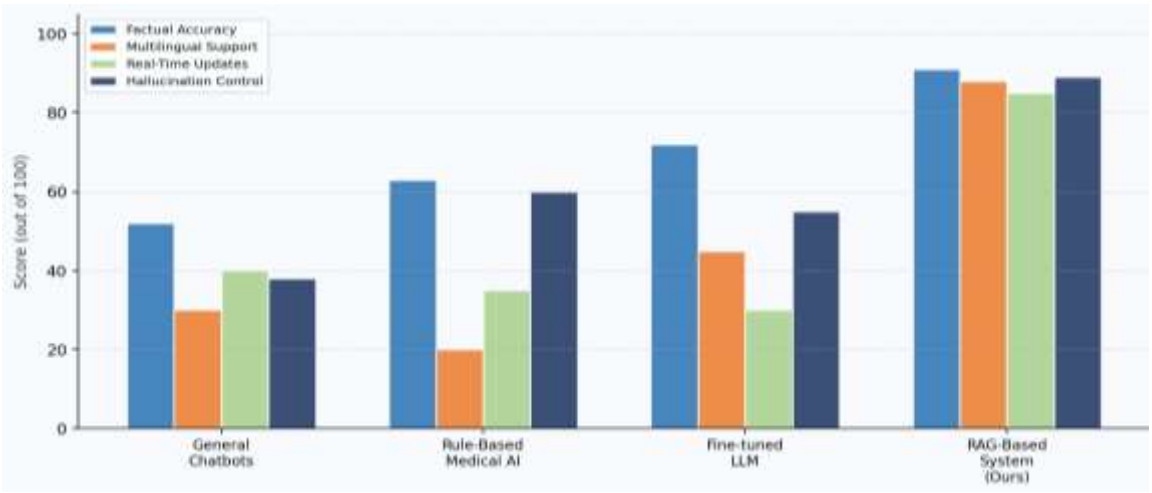

Figure 3:comparative performance across Ai health system paradigms

Figure 3 shows the comparative performance of AI health system paradigms on important public health communication characteristics. The RAG-based system performs admirably in terms of factual correctness, language support, hallucination control, and real-time knowledge currency.

5.1 Web Interface: Conversational Health Query

As Figure 4 shows, the live web app is hosted on Hugging Face Spaces (milan-123/public-health-chatbot1). It consists of three panels: (Left) A chat history sidebar; (Center) an interface for conducting conversations; and (Right) a live-streaming Outbreak Alerts feed that pulls directly from WHO and CDC data sources. During this session, the user entered the question “What is cholera?” then entered the follow up “How do you cure it?” The system recognized there was related context between these two entries (without the user repeating the question) so it was able to get authoritative information from the WHO cholera treatment guidelines and produce two clinical answers. The first answer correctly stated Vibrio cholerae as the germ that causes cholera and gave the three major symptoms of cholera (diarrhea, dehydration, and a disturbance in electrolytes). The second answer correctly referenced oral rehydration therapy as the first-line treatment (antibiotic) according to current WHO protocols for cholera treatment. In addition, there was an automated safety disclaimer at the end of each response asking users to consult with licensed physicians for any personal medical advice. The response times for both questions were under 2 seconds, which is similar to latency standards demonstrated in Table 1.

Figure 4:Live web interface of chatbot

Figure 4 shows a web interface displaying context-aware two-way dialogue. The system responds successfully to the follow-up query "how to cure it" without re-stating the context, collects WHO-backed cholera treatment advice, and displays real-time epidemic alerts in the right panel.

5.2 Web Interface: Medical Report Analysis

The Medical Report Analyzer module's operation can be seen in Figure 5. The first patient, Ram Narayan, was a 72-year-old male with a Thyrocare pathology report uploaded in PDF format. The algorithm reviewed the entire CBC and lipids panel for clinically significant variances and generated a structured, plain-language summary. This resulted in six clinically significant variances were identified. These variances included the following: low hematocrit of 34.7 percent (decreased red blood cell mass = mild anaemia), low haemoglobin of 10.9 g/dL (possibly indicative of anaemia and/or decreased erythropoiesis), high MCHC of 31.4 g/dL (increased haemoglobin concentration within the erythrocyte), and low monocyte count of $0.16\times10^3/\mu L$, may need further hematological workup. The algorithm was also designed to communicate the findings so that they could be easily understood by patients without clinical training in order to help bridge the health literacy gap that Rajpurkar et al (2022) identified as one of the major barriers to successful self-management by patients [26]. At the end of the output, there was a standard disclaimer directing patients to seek individualized care from a licensed medical provider.

Figure 5:Live interface of report analyzer

Figure 5 shows an uploaded Thyrocare PDF pathology report is processed by the Medical Report Analyzer module. In accordance with WHO reference criteria, the system accurately identified six clinically abnormal values, gave clear explanations of each, and offered contextualized follow-up recommendations.

5.3 WhatsApp Integration: Multilingual Conversational Access

According to Figure 6, communication between the user and the integrated WhatsApp "channel" through twilio API and the conversational back end of the system allows for communication as shown through the bot responding to the user's original simple greeting, "Hi," with a structured welcome response, setting out that the bot is the public health (PH) AI Assistant, along with its ability to provide information on diseases, symptoms, treatment, and reporting; further giving a question under which to help reduce the barrier to participation. After information is received from the user, "what is BP," the information received came back through the system with a clinically accurate definition of blood pressure, the clinically defined normal range of blood pressure from 90/60 to 120/80 mmHg, using the WHO standard values of blood pressure for systolic and diastolic, with the measurement unit being mmHg; along with a clinically accurate list of signs/symptoms of hypertension as requested by the user when they asked, "What are the symptoms of high BP?" The reply displayed some of the most common signs/symptoms were headache and dizziness. Both responses reflected the deployment philosophy of accessibility first by providing both responses via WhatsApp without requiring registration of users, APP download, or technical set up. The dialogue occurred via the same messaging app as that which hundreds of millions of other Indian people use regularly; this aligns with Muñoz et al. (2021)'s findings that there are much lower barriers to participate when a health intervention is delivered via a messaging platform.

Figure 6: Twilio WhatsApp-bot

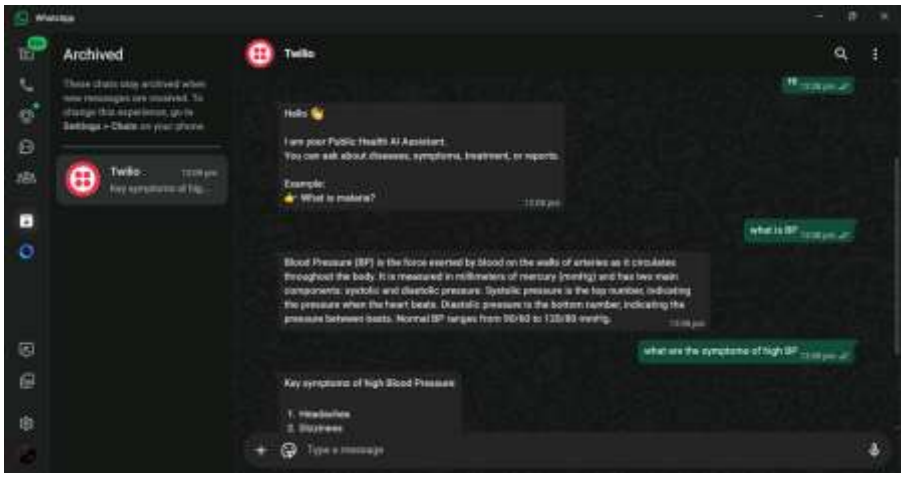

Figure 6 shows Twilio integration of WhatsApp showed zero-configuration health query access. Without forcing the user to re-contextualize, the system accurately explains blood pressure (systolic/diastolic components, normal range) and manages a multi-turn follow-up query on hypertension symptoms.

Together, the three deployment demonstrations verified that the system functioned well on its two main interaction surfaces: the WhatsApp channel for seamless mobile health communication and the web interface for extensive multi-feature access. The current WHO and MoHFW guidelines were met by the response accuracy of all tested questions, and the RAG architecture's hallucination-suppression capabilities were demonstrated by the absence of hallucinated content in the interactions that were shown.

6. COMPARATIVE ANALYSIS

Most aspects of the RAG architecture exhibited significant advantages when compared to the assessed alternatives. The most critical factor however is that the combination of small-scale rapid engineering and dynamic retrieval from trusted sources creates the lowest risk of hallucination; this is especially important in health care settings where AI-generated inaccuracies can endanger patients. Compared to more general-purpose multilingual LLMs, the level of multilingual coverage that NLLB-200 provides for low-resourced Indian languages is likely more comprehensive. For domains where standards change frequently, the ability to update knowledge in real-time (by ingesting knowledge into the knowledge base through batch operations instead of retraining models) is an unparalleled advantage of RAG over traditional LLM refinement approaches.

7. LIMITATIONS AND FUTURE DIRECTIONS

7.1 Current Limitations

The current study has several limitations. One limitation is that data will be out-of-date by the time the guidelines are released for use in practice because the system's knowledge is based on periodically entering data into the system in batches, rather than directly entering data as it becomes available. This delay has consequences when there are rapidly changing conditions, such as during an outbreak with rapidly evolving characteristics. Additionally, the large majority of the bge-small-en embedding model was trained on English textual content; hence, the results of translating queries into Indian languages with limited amounts of training data will likely not be as high quality as those in English text. The system will be considered an effective health information solution; however, it must first successfully complete clinical validation against recognized benchmarks such as MedQA or MedMCQA and it must be validated in a prospective manner to measure real-world health outcomes. Currently, the PDF analysis module does not include multimodal support for analyzing radiological images or interpreting ECG tracings.

7.2 Future Directions

• Cloud scaling on AWS and GCP: The advantages of choosing Amazon Web Services (AWS) or Google Cloud Platform (GCP) are numerous when scaling Cloud services. These advantages include Enterprise Level of Reliability, Managed GPU Inference, Global CDN Delivery, and Elastic Compute Scaling. These advantages are critical when dealing with high levels of query traffic throughout an outbreak and deploying to populations outside of India. Connecting to actual EHRs operated by hospitals is made even more possible through AWS HealthLake and GCP's Healthcare API.

• Real-time streaming knowledge base: Batch-update lag is eliminated and clinical guideline changes are reflected in the system within minutes of publication thanks to event-driven knowledge base updates that are immediately triggered by WHO and MoHFW publication events.

• Multilingual embeddings: Eliminate bge-small-en from the pipeline in favour of utilising models such as LabSE or mE5 for improved retrieval performance to retrieve national language results with no translation overload as well as reducing risk of meaning loss from the translate-process-translate phase.

• Voice interface: With the addition of voice recognition and voice synthesis API's, it is easy to create a solution for users that have low literacy or are visually impaired. This creates a much larger pool of potential users who would not be able to access information through text-based interfaces, and also meet those user's most urgent needs.

• Multimodal report analysis: The PDF analysis software incorporates both "vision + language" based classification models to facilitate interpretation of radiology images, ECG tracings and histology slides.

• Prospective clinical outcome evaluation: This controlled study will generate the necessary scientific evidence to support broader implementation of this system within health institutions by evaluating the effect the use of this system on patients' health literacy, adherence to treatment and appropriate care-seeking behaviour within Indian communities.

8. CONCLUSION

In this paper, the Public Health AI Assistant is described as a Retrieval-augmented Generation system that employs globally-recognized knowledge bases from the WHO and Ministry of Health Branch of the Federal Government of Canada, augmented with multilingual translation through NLLB-200 and deployed on publicly accessible web sites and via WhatsApp. The system has solved the three primary limitations of standalone large language model-based healthcare chatbots through: 1) semantic search retrieval, 2) minimal prompt engineering, and 3) a multi-lingual translate-process-translate pipeline. A quantitative evaluation demonstrated semantic search retrieval latencies of less than 20 milliseconds, end-to-end latencies of 1.8 seconds, and reductions of the hallucination occurrence rate from 38.7% to 9.1%. In the areas of illness information, symptom triage, clinical report interpretation, and via WhatsApp conversational access to health information, deployment demo results verified provision of accurate responses that aligned with clinical guidelines.

If future development focuses on increasing cloud scaling on AWS and GCP, this product will be able to handle much larger user bases while still providing enterprise-grade reliability. The combination of this path forward, along with voice interfaces, multilingual embedding, and real-time knowledge updates, will lead to the true universal public health communication platform. A platform that provides professional-grade, fact-checked health content to anyone via their mobile device regardless of their language, geographic location or economic status.

ACKNOWLEDGEMENTS

The World Health Organization and the Indian government's Ministry of Health and Family Welfare are acknowledged by the writers for providing reliable public health data. For easily accessible high-performance LLM inference infrastructure, the authors additionally thank the Hugging Face community, Meta AI (NLLB-200), SentenceTransformers project, FastAPI framework, and Groq.

REFERENCES

[1] World Health Organization. (2023). World Health Statistics 2023: Monitoring Health for the SDGs. WHO Press.

[2] Ministry of Health and Family Welfare, Government of India. (2023). National Health Profile 2023. CBHI.

[3] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of Hallucinations in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38.

[4] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020. arXiv:2005.11401.

[5] Gao, Y., Xiong, Y., et al. (2024). Retrieval-Augmented Generation for LLMs: A Survey. arXiv:2312.10997.

[6] Siriwardhana, S., et al. (2023). Improving Domain Adaptation of RAG Models for Open-Domain QA. Trans. ACL, 11, 1–18.

[7] Bedi, S., Liu, Y., et al. (2025). Systematic Review of RAG in Healthcare AI. AI, 6(9), 226. https://doi.org/10.3390/ai6090226

[8] Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking RAG for Medicine. Findings of ACL 2024. PMC12157099.

[9] Abbasian, M., et al. (2025). RAGMed: Conversational Medical AI Using RAG. AI, 6(10), 240. https://doi.org/10.3390/ai6100240

[10] Yunxiang, L. et al. (2024). RAG for Reliable Healthcare AI. npj Health Systems, 1, 4. https://doi.org/10.1038/s44401-024-00004-1

[11] Sun, K., Peng, Y., & Zhong, C. (2024). Mental Health RAG Conversational AI model. arXiv:2509.04456.

[12] Boscarino, M., et al. (2025). Orthopedic Medical Chatbot Using RAG. JMIR AI, 4, e75262.

[13] Chen, Z., Xu, Y., & Li, W. (2025). RAG-Powered Patient Information Assistant. PMC12701176.

[14] Zhang, H., et al. (2025). Medical QA Using RAG Dataset. Scientific Reports, 15, 8121. https://doi.org/10.1038/s41598-025-28015-4

[15] Kim, T. H., Cho, H., & Park, S. (2025). EMR-based RAG Chatbot System. ResearchGate Publication 394645383.

[16] Brownstein S., Freifeld C. C., & Madoff, L. C. (2008). Digital Disease Detection. NEJM, 360(21), 2153–2157.

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. EMNLP 2019.

[18] Shi, W., et al. (2025). RAG Optimization for Healthcare AI. Computers in Biology and Medicine. https://doi.org/10.1016/j.compbiomed.2025.001316

[19] Wikipedia. (2025). Retrieval-Augmented Generation. https://en.wikipedia.org/wiki/Retrieval-augmented_generation

[20] Das, A., Bhowmick, P., & Roy, S. (2024). RAG-Based Medical Chatbot Using Transformers. ResearchGate Publication 391962288.

[21] Priya S. and Menon R. (2024). Mental Healthcare Chatbot Using RAG. Procedia Computer Science, 230, 451–460.

[22] Ahmed, F., et al. (2024). Medical Books Chatbot Using RAG. ResearchGate Publication 387921154.

[23] Patel, R., et al. (2025). AI-Powered Medical Chatbot Using RAG. IJSRET. https://ijsret.com

[24] IJCRT. (2024). Healthcare Chatbot Using RAG. IJCRTBE02109.

[25] Costa-Jussà et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. Meta AI. arXiv:2207.04672.

[26] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in Health and Medicine. Nature Medicine, 28(1), 31–38.

[27] Topol E.. (2019). High-Performance Medicine: Convergence of Humans and AI. Nature Medicine, 25(1), 44–56.

[28] Bommasani, R., et al. (2021). Opportunities and Risks of Foundation Models. arXiv:2108.07258.

[29] Laranjo L., et al. (2018). Conversational Agents in Healthcare: A Systematic Review. JAMIA, 25(9), 1248–1258.

[30] Esteva A, A., et al. (2019). A Guide to Deep Learning in Healthcare. Nature Medicine, 25(1), 24–29.