Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Enabled Public Health Agent for SupportingFrontline Healthcare Workers in Primary CareSettings: Design, Evaluation, and Decision Support

Sayed Mohammed Zeeshan

msayed@gitam.edu

GITAM University

Research Article

Keywords: Artificial Intelligence, Retrieval-Augmented Generation, Primary Healthcare, Community Health Centres, Nursing Informatics, Healthcare Assistants, Clinical Decision Support Systems, Public Health Informatics, Digital Health, India

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10120163/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-Enabled Public Health Agent for Supporting Frontline Healthcare Workers in Primary Care Settings: Design, Evaluation, and Decision Support

Sayed Mohammed Zeeshan

Department of Mathematics & Statistics
GITAM (Deemed to be University), Visakhapatnam Campus
Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh 530045, India

Corresponding Author
Sayed Mohammed Zeeshan
Department of Mathematics & Statistics
GITAM (Deemed to be University), Visakhapatnam Campus
Visakhapatnam, Andhra Pradesh 530045, India
Email: msayed@gitam.edu

Abstract

Background: Primary healthcare centres, community health centres, dispensaries, and nursing facilities play a vital role in healthcare delivery. However, frontline healthcare workers, nursing staff, healthcare assistants, and community health personnel often face increasing workloads related to patient interaction, symptom collection, referral coordination, documentation, and operational support. Artificial Intelligence (AI) offers opportunities to augment these healthcare workflows by providing rapid access to relevant healthcare information and decision-support resources.

Materials and Methods: This study developed and evaluated a lightweight Retrieval-Augmented Generation (RAG)-based Public Health Agent designed to support operational healthcare workflows. The framework integrates domain-aware routing, semantic retrieval, structured healthcare repositories, deterministic physician and outpatient department (OPD) lookup, and rule-based triage mechanisms. Five healthcare repositories were incorporated, including Public Health, Doctor Registry, Patient Management, OPD Scheduling, and Case Triage datasets. The system was evaluated using a benchmark of 200 healthcare queries spanning public-health information, physician referral, patient management, appointment scheduling, and triage scenarios. Performance was compared against a retrieval-only baseline using retrieval effective-

ness, response quality, groundedness, routing accuracy, triage accuracy, latency, and memory-consumption metrics. Sensitivity analysis was performed for retrieval depths ((k=1,3,5)), and bootstrap confidence intervals were estimated for major evaluation metrics.

Results: The proposed lightweight RAG framework consistently outperformed the retrieval-only baseline across all evaluated retrieval depths. At the best-performing configuration ((k=5)), Recall@k improved from 0.677 to 0.739, Mean Reciprocal Rank (MRR) increased from 0.643 to 0.712, and nDCG@k improved from 0.630 to 0.700. Groundedness increased from 0.825 to 0.846, while the unsupported sentence rate decreased from 0.175 to 0.154. Domain-routing accuracy reached 87

Conclusions: The proposed lightweight RAG-based Public Health Agent demonstrated improved retrieval effectiveness and response grounding compared with a retrieval-only baseline while maintaining minimal computational requirements. The findings support the technical feasibility of lightweight retrieval-based healthcare information-support systems for resource-constrained environments. However, the evaluation was conducted using expert-curated synthetic healthcare datasets and internally developed benchmark queries. Further clinician-centered evaluation, external validation, and real-world deployment studies are required before conclusions regarding clinical effectiveness can be drawn.

Keywords: Artificial Intelligence; Retrieval-Augmented Generation; Primary Healthcare; Community Health Centres; Nursing Informatics; Healthcare Assistants; Clinical Decision Support Systems; Public Health Informatics; Digital Health; India.

1 Introduction

Primary healthcare centres, community health centres, dispensaries, nursing centres, and rural health clinics form the foundation of healthcare delivery in India. Over the past several decades, substantial investments have been made in expanding healthcare infrastructure, increasing healthcare coverage, and strengthening public health services. Despite these advancements, many healthcare facilities continue to experience shortages of trained support personnel, including nursing assistants, community healthcare workers, health educators, patient coordinators, and healthcare support staff. Although physicians may be available at many facilities, the growing number of patients and increasing administrative responsibilities often place significant pressure on healthcare teams, creating challenges for efficient patient management and timely healthcare delivery.

Frontline healthcare workers play an essential role in supporting physicians and ensuring the smooth operation of healthcare services. Personnel such as Accredited Social Health Activists (ASHAs), Auxiliary Nurse Midwives (ANMs), Community Health Officers (CHOs), nursing

staff, health assistants, and other healthcare support workers are frequently responsible for collecting patient information, recording symptoms, conducting preliminary assessments, providing health education, maintaining records, supporting public health programs, and coordinating patient care. Through nursing education, paramedical training, government-sponsored health programs, and practical field experience, these workers acquire valuable healthcare knowledge and serve as an indispensable bridge between patients and physicians.

However, healthcare support staff often operate under demanding conditions characterized by high patient volumes, limited time for patient interaction, increasing documentation requirements, and restricted access to decision-support resources. In many settings, healthcare workers must rapidly assess patient complaints, identify potential risk factors, retrieve relevant health information, and communicate effectively with physicians while maintaining accuracy and consistency. These challenges highlight the growing need for intelligent tools that can assist healthcare workers in symptom assessment, information retrieval, patient triage, and healthcare decision support without disrupting existing clinical workflows.

Recent advances in Artificial Intelligence (AI) have created new opportunities for enhancing healthcare delivery and supporting healthcare professionals. AI technologies are increasingly being applied to disease prediction, clinical decision support, medical information retrieval, telemedicine, healthcare management, and public health surveillance. Previous studies have demonstrated that AI can improve operational efficiency, reduce workload, enhance diagnostic support, and strengthen healthcare accessibility. Nevertheless, many existing AI solutions require extensive computational resources, large clinical datasets, continuous internet connectivity, or specialized technical expertise, limiting their practicality in primary healthcare settings and resource-constrained environments.

The emergence of Large Language Models (LLMs) has further expanded the potential of AI-assisted healthcare support through natural-language interaction and knowledge-based reasoning. Healthcare workers can interact with these systems using simple language to obtain information, assess symptoms, and receive guidance. However, direct use of LLMs in healthcare raises important concerns regarding reliability, transparency, explainability, and factual accuracy. Generative AI systems may occasionally produce plausible but incorrect responses, commonly referred to as hallucinations, which can reduce trust and potentially compromise healthcare decision-making.

Retrieval-Augmented Generation (RAG) has emerged as a promising solution to address these challenges. By combining information retrieval techniques with generative language models, RAG systems generate responses grounded in curated knowledge repositories rather than relying solely on the internal knowledge of the model. This approach improves factual consistency, transparency, interpretability, and user confidence, making it particularly suitable for healthcare applications where evidence-based recommendations are essential.

Motivated by the need to strengthen healthcare support services within primary healthcare environments, this study presents an AI-enabled Public Health Agent designed to assist frontline healthcare workers, nursing staff, and healthcare support personnel. The proposed system employs a lightweight Retrieval-Augmented Generation framework that combines keyword-based document retrieval with large language model reasoning over expert-developed public-health fact sheets. The system is intended to support symptom assessment, health information retrieval, patient triage assistance, and healthcare decision support while maintaining simplicity, transparency, and computational efficiency.

The primary contributions of this study are fourfold. First, we develop a lightweight and interpretable AI-assisted information-support framework specifically designed for healthcare support staff operating in primary healthcare settings. Second, we construct and publicly release a collection of expert-curated synthetic healthcare repositories covering public-health information, physician referral, patient management, OPD scheduling, and case-triage workflows. Third, we integrate domain-aware routing, semantic retrieval, structured healthcare repositories, deterministic physician and OPD lookup, and rule-based triage mechanisms to provide evidence-supported responses across multiple healthcare domains. Finally, we evaluate the proposed framework using a benchmark of 200 healthcare queries and compare its performance against a retrieval-only baseline using retrieval effectiveness, response quality, groundedness, routing accuracy, triage accuracy, latency, and computational-efficiency metrics. Additional sensitivity analysis, bootstrap confidence interval estimation, and error analysis are performed to assess robustness and identify system limitations. The results demonstrate that the proposed lightweight RAG framework consistently improves retrieval effectiveness and response grounding compared with direct retrieval while maintaining minimal computational overhead. These findings support the technical feasibility of lightweight retrieval-based healthcare information-support systems for resource-constrained environments and provide a foundation for future clinician-centered evaluation and real-world deployment studies.

2 Literature Review

Recent studies have highlighted the growing role of AI in strengthening primary healthcare systems through applications such as risk prediction, disease surveillance, telemedicine, personalized care, clinical decision support, and healthcare management. Joshi and Jasani [4] emphasized the potential of AI to improve the efficiency, accessibility, and responsiveness of primary healthcare services. Similarly, Pathak et al. [7] demonstrated that AI technologies can enhance diagnostic accuracy, optimize triage processes, automate routine administrative tasks, and expand access to specialist-level healthcare services in underserved regions.

One of the most promising applications of AI in healthcare is the development of Clinical De-

cision Support Systems (CDSSs). Nayak et al. [6] discussed how AI-powered CDSS platforms leverage machine learning and advanced data analytics to support diagnostic interpretation, treatment planning, and personalized healthcare delivery. By identifying hidden relationships within complex medical datasets, such systems can assist healthcare professionals in making more informed and evidence-based decisions. These capabilities are particularly important in healthcare settings where specialist expertise and advanced diagnostic resources are limited.

While AI offers substantial opportunities for improving healthcare delivery, successful implementation depends heavily on healthcare workers' trust and acceptance. Tun et al. [9] conducted a systematic review of AI-based Clinical Decision Support Systems and identified transparency, explainability, clinical reliability, usability, and user training as critical determinants of trust among healthcare workers. The review highlighted that healthcare professionals are more likely to adopt AI systems when recommendations are interpretable and can be verified through trusted evidence sources. These findings reinforce the need for AI systems that augment rather than replace human decision-making.

Beyond supporting clinical decisions, AI has also been recognized as a mechanism for improving healthcare equity and accessibility. Reddy [8] proposed a framework illustrating how AI can democratize healthcare access by reducing geographical barriers, optimizing resource utilization, improving preventive care, and enhancing healthcare affordability. Such capabilities are particularly relevant for low- and middle-income countries where healthcare resources are often unevenly distributed and rural communities face persistent access challenges.

The potential of AI to support healthcare workers themselves has received increasing attention in recent years. Virgillito and Ledda [10] reported that AI-enabled technologies can improve workflow efficiency, reduce stress and burnout, support workplace health promotion, and enhance healthcare worker engagement. Their findings suggest that AI can contribute not only to improved patient outcomes but also to strengthening workforce sustainability and organizational performance within healthcare systems.

The COVID-19 pandemic highlighted the importance of digital technologies and intelligent systems in public health preparedness. Abdelouahed et al. [1] examined the integration of AI into public health education, healthcare training, and epidemic preparedness. Their work demonstrated how AI-driven epidemiological modeling, outbreak prediction, and healthcare workforce training can strengthen healthcare resilience and improve readiness for future public health emergencies.

AI applications have also demonstrated value in complex healthcare operational environments. Alquayt et al. [2] investigated AI-driven healthcare innovations implemented during the Hajj pilgrimage and reported improvements in patient care, healthcare coordination, resource management, and clinical service delivery. These findings illustrate the flexibility of AI systems in supporting healthcare delivery across diverse and large-scale operational settings.

As AI becomes increasingly integrated into healthcare systems, ethical governance and stakeholder participation have become important areas of discussion. Cleofas [3] proposed a human-rights-based framework for healthcare AI governance and emphasized the critical role of nursing academia in ensuring equitable, safe, and ethical deployment of AI technologies. The study highlighted the importance of transparency, accountability, human rights considerations, and stakeholder engagement in shaping responsible healthcare AI ecosystems.

Despite the growing promise of AI, many existing healthcare solutions remain dependent on cloud infrastructure and computationally intensive architectures that may not be practical in low-resource environments. To address these limitations, Latif et al. [5] proposed a lightweight Edge-AI framework capable of providing real-time diagnostic support on low-power devices without continuous internet connectivity. Their results demonstrated reduced latency, improved reliability, enhanced privacy, and greater usability for non-technical users, highlighting the feasibility of deploying AI solutions in remote and underserved communities.

Collectively, the existing literature demonstrates that AI has substantial potential to improve healthcare accessibility, diagnostic quality, workforce effectiveness, public health preparedness, healthcare governance, and healthcare equity [4, 7, 6, 9, 8, 10, 1, 2, 3, 5]. However, most existing studies focus primarily on physicians, hospitals, specialist healthcare services, or large healthcare systems. Comparatively limited attention has been given to empowering grassroots healthcare workers such as Accredited Social Health Activists (ASHAs), Auxiliary Nurse Midwives (ANMs), and Community Health Officers (CHOs), who form the backbone of healthcare delivery in rural India. Motivated by this gap, the present study develops a lightweight Retrieval-Augmented Generation (RAG)-based Public Health Agent designed specifically to support frontline healthcare workers through evidence-grounded symptom-based decision support in low-resource settings.

3 Materials and Methods

3.1 System Architecture

The proposed Public Health Agent was developed as a lightweight Retrieval-Augmented Generation (RAG) framework designed to support frontline healthcare workers, nursing staff, healthcare assistants, and community health personnel operating in primary healthcare settings. The system combines domain-aware routing, semantic retrieval, structured healthcare repositories, deterministic physician and outpatient department (OPD) lookup, and rule-based triage mechanisms to provide evidence-supported responses for symptom assessment, healthcare information retrieval, patient triage support, referral assistance, and operational healthcare queries.

The framework consists of five domain-specific knowledge repositories:

1. Public Health Repository
2. Doctor Registry Repository
3. Patient Management Repository
4. OPD Scheduling and Room Management Repository
5. Case Triage and Escalation Repository

Each repository is stored as a structured JSON dataset containing expert-curated healthcare information relevant to operational healthcare workflows. Incoming user queries are first classified and routed to the most appropriate healthcare domain. Semantic retrieval is then performed to identify the most relevant records within the selected repository. For physician referral and OPD-related queries, deterministic lookup procedures are applied to retrieve structured information regarding healthcare providers, specialties, consultation schedules, and room assignments. For triage-related queries, rule-based severity assessment and escalation logic are used to generate management recommendations.

The retrieved records are subsequently aggregated into a context representation from which an evidence-supported response is generated. By combining semantic retrieval with structured healthcare workflows, the framework aims to improve retrieval quality, response grounding, and operational usability while maintaining low computational overhead suitable for deployment in resource-constrained healthcare environments.

3.2 Datasets

The proposed Public Health Agent was developed using five structured healthcare repositories represented in JavaScript Object Notation (JSON) format. These repositories collectively support disease-information retrieval, patient management, physician referral, outpatient scheduling, and clinical triage operations. The datasets were curated to emulate information commonly encountered in primary-care and public-health environments while remaining lightweight enough for deployment in low-resource settings.

The Public Health Repository contains 101 disease-specific fact records covering communicable diseases, vector-borne infections, vaccine-preventable diseases, respiratory illnesses, sexually transmitted infections, parasitic diseases, and other public-health conditions. Each record includes disease descriptions, symptoms, transmission routes, prevention measures, and treatment recommendations.

The Doctor Registry Repository contains information for 10 physicians representing 10 medical specialties, including General Medicine, Cardiology, Pediatrics, Orthopedics, Dermatology,

Neurology, Gynecology, ENT, Psychiatry, and Ophthalmology. Physician availability, consultation schedules, and OPD locations are maintained to support referral decisions.

The OPD Scheduling and Room Management Repository contains 10 consultation rooms distributed across multiple specialty blocks. Information includes room availability, specialty assignments, physician occupancy, and next available appointment slots.

The Case Triage and Escalation Repository contains more than 86 triage records generated from simulated patient encounters. Each record stores symptom information, severity scores, agent confidence scores, management decisions, escalation routes, and physician assignments. These records are used to evaluate escalation logic and referral workflows.

The repositories were manually curated from standard public-health knowledge sources and healthcare workflow requirements. Records were reviewed for consistency, completeness, and operational relevance prior to integration into the retrieval system.

The repositories were manually curated from standard public-health knowledge sources and healthcare workflow requirements. The datasets are publicly available through Zenodo [11] and include expert-curated disease information together with synthetic healthcare workflow records developed with input from healthcare professionals and administrators. Records were reviewed for consistency, completeness, and operational relevance prior to integration into the retrieval system.

Inclusion criteria:

- Diseases commonly encountered in primary-care and community-health settings.
- Physician specialties frequently required for referral workflows.
- OPD scheduling information necessary for routing and appointment allocation.
- Triage scenarios representing both low-risk and high-risk patient presentations.

Exclusion criteria:

- Duplicate records.
- Incomplete entries lacking essential fields.
- Conditions outside the scope of frontline public-health decision support.
- Records with insufficient information for retrieval or escalation decisions.

Table 1 summarizes the healthcare repositories used in the proposed framework.

Table 1: Summary of datasets used in the Public Health Agent.

Repository	Records	Purpose
Public Health Repository	101	Disease facts and health guidance
Doctor Registry Repository	10	Physician information and specialties
OPD Scheduling Repository	10	Room allocation and appointment routing
Case Triage Repository	86+	Escalation and referral decisions
Total Structured Records	207+	Integrated healthcare knowledge base

3.3 Retrieval-Augmented Generation Framework

The proposed framework follows a lightweight Retrieval-Augmented Generation (RAG) architecture that combines domain-aware routing, semantic retrieval, structured healthcare repositories, deterministic physician and OPD lookup, and rule-based triage mechanisms. Given a user query (q), the routing module first identifies the most appropriate healthcare domain. Semantic retrieval is then performed within the selected repository to identify the top-(k) most relevant records. Sensitivity analysis was conducted using (k=1), (k=3), and (k=5), with (k=5) providing the best overall retrieval performance.

For physician referral and OPD-related queries, deterministic lookup procedures are used to retrieve structured information regarding specialists, consultation schedules, and room assignments. For triage-related queries, rule-based severity assessment and escalation logic are applied to support operational decision-making. Responses are generated exclusively from retrieved records and structured repository content, thereby improving transparency, traceability, and evidence support while reducing the likelihood of unsupported statements.

The retrieval process, ranking formulation, and workflow algorithm are described mathematically and algorithmically in the subsequent methodology subsections to facilitate reproducibility and independent implementation.

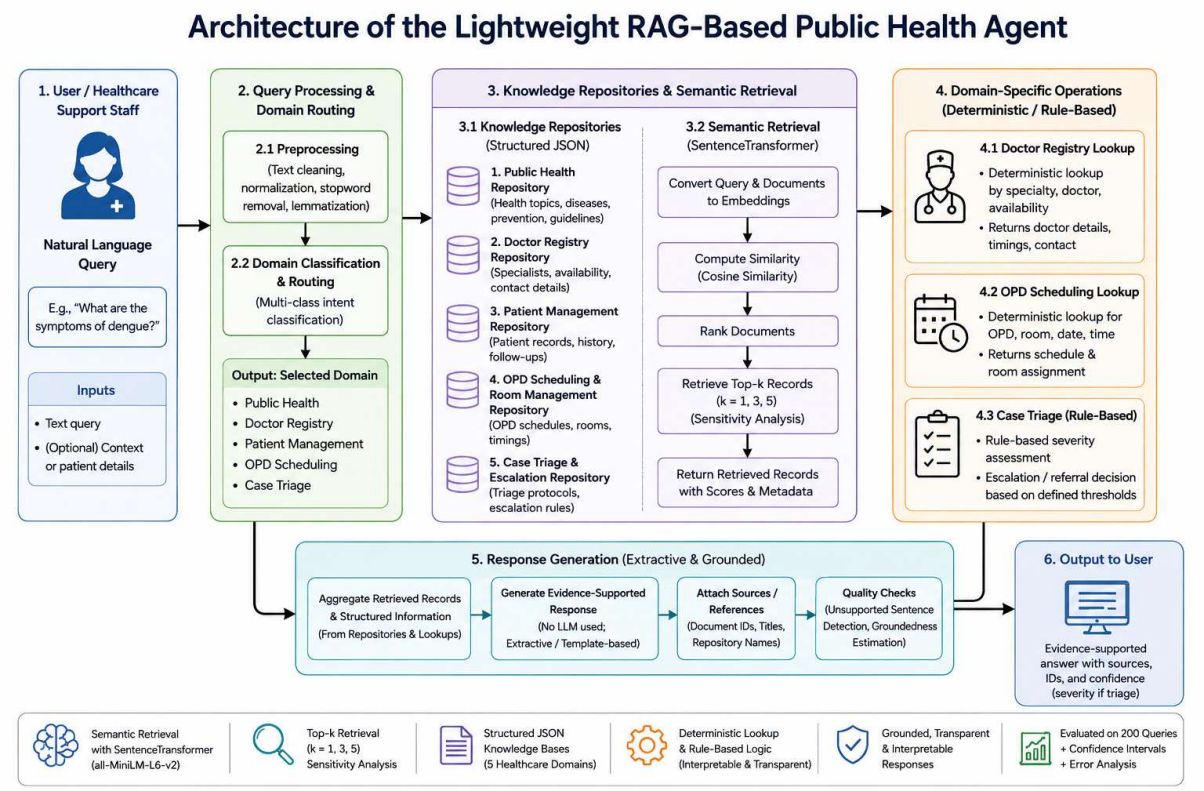


Figure 1: Architecture of the proposed lightweight RAG-based Public Health Agent.

Figure 1 illustrates the overall workflow of the proposed framework. User queries are first routed to one of five healthcare repositories, followed by semantic retrieval, deterministic healthcare operations, and rule-based triage when required. Retrieved information is subsequently used to generate evidence-supported responses while maintaining low computational overhead suitable for resource-constrained healthcare environments.

3.4 Retrieval Algorithm

Let q denote a user query and let $D_c = \{d_1, d_2, \ldots, d_n\}$ denote the collection of records belonging to the selected healthcare domain c. Following domain routing, semantic retrieval is performed within the corresponding repository using sentence embeddings generated by the all-MiniLM-L6-v2 SentenceTransformer model.

Let

$$\mathbf{e}_q = f(q)$$

denote the embedding representation of query q, where $f(\cdot)$ represents the sentence-transformer encoder. Similarly, each repository record d_i is converted into an embedding vector

$$\mathbf{e}_i = f(d_i).$$

The semantic similarity between query q and record d_i is computed using cosine similarity:

$$S(q, d_i) = \frac{\mathbf{e}_q^T \mathbf{e}_i}{\|\mathbf{e}_q\| \, \|\mathbf{e}_i\|}.$$

All repository records are ranked according to $S(q, d_i)$, and the top-k records are selected:

$$R_k(q) = \mathrm{TopK}_{d_i \in D_c} \, S(q, d_i),$$

where $k \in \{1, 3, 5\}$. Sensitivity analysis was performed for multiple values of k, and $k = 5$ produced the best overall retrieval performance in the benchmark evaluation.

For physician-referral and OPD-related queries, deterministic lookup procedures are additionally applied to retrieve structured information regarding specialties, physicians, room assignments, schedules, and availability. For triage-related queries, retrieved information is combined with rule-based severity assessment and escalation logic. Responses are generated directly from retrieved records and structured repository fields to maintain transparency, traceability, and factual support.

Algorithm 1 Lightweight healthcare retrieval and response generation

Require: User query q, healthcare repositories $\mathcal{D}$, retrieval depth k

Ensure: Evidence-supported response A

Route query q to healthcare domain c
Select repository $D_c \in \mathcal{D}$
Encode query as $\mathbf{e}_q \leftarrow f(q)$
for each record $d_i \in D_c$ **do**
    Encode record as $\mathbf{e}_i \leftarrow f(d_i)$
    Compute $S(q, d_i) \leftarrow \cos(\mathbf{e}_q, \mathbf{e}_i)$
end for
Rank records in descending order of $S(q, d_i)$
Retrieve top-k records $R_k(q)$
if query belongs to Doctor Registry **then**
    Apply deterministic physician and specialty lookup
else if query belongs to OPD Scheduling **then**
    Apply deterministic room and availability lookup
else if query belongs to Case Triage **then**
    Apply severity and escalation rules
end if
Aggregate retrieved records and structured information
Generate evidence-supported response A
return A

3.5 Evaluation Datasets

The evaluation was conducted using five operational healthcare repositories:

- Public Health Fact Sheets
- Doctor Registry Dataset
- Patient Management Dataset
- OPD Scheduling and Room Management Dataset
- Case Triage and Escalation Dataset

A benchmark consisting of 200 healthcare queries was developed to evaluate retrieval quality, response generation performance, response grounding, routing accuracy, triage performance, and computational efficiency. The benchmark included public-health information requests, physician referral queries, patient-management tasks, OPD scheduling requests, and triage-related scenarios designed to reflect conversational queries commonly encountered by healthcare workers and laypersons. Retrieval relevance annotations were available for 142 benchmark instances and were used for retrieval-performance evaluation.

In addition, a retrieval-only baseline was implemented to assess the contribution of the proposed lightweight RAG framework. Comparative evaluation was performed between the retrieval-only baseline and the proposed system using identical benchmark queries.

3.6 Evaluation Metrics

Retrieval effectiveness was evaluated using Precision@k, Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG@k). Response quality was assessed using ROUGE-L, BLEU, and Token-Level F1 scores.

To assess factual reliability, groundedness and unsupported sentence rate were computed. Operational performance was evaluated using domain-routing accuracy and triage-decision accuracy. Computational efficiency was assessed using average response latency, 95th-percentile latency, and peak memory consumption.

A retrieval-only baseline was implemented to quantify the contribution of domain routing, structured lookup, and triage mechanisms beyond direct retrieval alone. Sensitivity analysis was performed using (k=1), (k=3), and (k=5) retrieved records to evaluate the effect of retrieval depth on system performance. To quantify uncertainty in the reported results, 95% bootstrap confidence intervals were estimated for major retrieval, grounding, routing, and triage metrics. In addition, error analysis was conducted to identify routing errors, triage-decision errors, and

retrieval failures, thereby providing insight into the strengths and limitations of the proposed framework.

4 Results

4.1 Expanded Benchmark and Baseline Evaluation

The system was evaluated using an expanded benchmark of 200 healthcare queries spanning public-health information, physician referral, patient management, OPD scheduling, and case-triage scenarios. Retrieval metrics were computed on 142 benchmark instances for which relevant document identifiers were available. The proposed lightweight RAG method was compared with a retrieval-only baseline to assess the contribution of domain routing, deterministic doctor and OPD lookup, rule-based triage, and extractive grounded response generation.

Table 2 compares the retrieval-only baseline with the proposed lightweight RAG framework.

Table 2: Baseline comparison between retrieval-only and proposed lightweight RAG methods.

Metric	Retrieval Only	Proposed Lightweight RAG
Precision@3	0.366	0.406
Recall@3	0.624	0.674
MRR	0.634	0.701
nDCG@3	0.615	0.681
ROUGE-L	0.115	0.115
BLEU	0.098	0.098
Token F1	0.124	0.124
Groundedness	0.825	0.846
Unsupported Sentence Rate	0.175	0.154
Category Accuracy	0.870	0.870
Triage Decision Accuracy	0.800	0.800
Latency Mean (ms)	0.781	0.944
Latency P95 (ms)	1.635	2.054
Peak Memory per Query (KiB)	8.10	8.71

The proposed lightweight RAG system improved retrieval effectiveness over the retrieval-only baseline. Precision@3 increased from 0.366 to 0.406, Recall@3 increased from 0.624 to 0.674, MRR improved from 0.634 to 0.701, and nDCG@3 increased from 0.615 to 0.681. These improvements indicate that the proposed method ranked relevant healthcare records more effectively than direct retrieval alone.

The groundedness score increased from 0.825 to 0.846, while the unsupported sentence rate decreased from 0.175 to 0.154. This suggests that the proposed lightweight RAG method produced responses that were more consistently supported by retrieved evidence. ROUGE-L, BLEU, and Token F1 remained unchanged because both evaluated methods used extractive response generation rather than free-form neural text generation.

4.2 Sensitivity Analysis for Retrieval Depth

To examine the effect of retrieval depth, sensitivity analysis was performed using $k = 1$, $k = 3$, and $k = 5$. Table 3 shows that increasing k improved Recall@k for both methods, although Precision@k decreased as more documents were retrieved.

Table 3 presents the sensitivity analysis for different retrieval depths.

Table 3: Sensitivity analysis for different values of k.

Method	k	Precision@k	Recall@k	MRR	nDCG@k
Retrieval Only	1	0.585	0.407	0.585	0.585
Proposed Lightweight RAG	1	0.648	0.445	0.648	0.648
Retrieval Only	3	0.366	0.624	0.634	0.615
Proposed Lightweight RAG	3	0.406	0.674	0.701	0.681
Retrieval Only	5	0.244	0.677	0.643	0.630
Proposed Lightweight RAG	5	0.280	0.739	0.712	0.700

The proposed lightweight RAG system outperformed the retrieval-only baseline at all evaluated values of k. At $k = 5$, Recall@k increased from 0.677 to 0.739, MRR improved from 0.643 to 0.712, and nDCG@k increased from 0.630 to 0.700. These results indicate that the proposed framework benefits from a larger retrieval context while maintaining improved ranking quality compared with direct retrieval alone.

4.3 Confidence Interval Analysis

To quantify uncertainty in evaluation performance, non-parametric bootstrap resampling with 1000 iterations was used to estimate 95% confidence intervals for major metrics. Table 4 summarizes the bootstrap estimates.

Table 4 reports the bootstrap 95% confidence intervals.

Table 4: Bootstrap 95% confidence intervals for major evaluation metrics.

Metric	Method	Mean	95% CI
Recall@5	Retrieval Only	0.677	[0.601, 0.746]
Recall@5	Proposed RAG	0.739	[0.671, 0.798]
MRR	Retrieval Only	0.643	[0.569, 0.714]
MRR	Proposed RAG	0.712	[0.642, 0.777]
nDCG@5	Retrieval Only	0.630	[0.559, 0.700]
nDCG@5	Proposed RAG	0.700	[0.634, 0.763]
Groundedness	Retrieval Only	0.825	[0.770, 0.875]
Groundedness	Proposed RAG	0.846	[0.799, 0.889]
Category Accuracy	Retrieval Only	0.870	[0.820, 0.915]
Category Accuracy	Proposed RAG	0.870	[0.820, 0.915]
Triage Accuracy	Retrieval Only	0.800	[0.640, 0.960]
Triage Accuracy	Proposed RAG	0.800	[0.640, 0.960]

The confidence intervals show that the proposed lightweight RAG system consistently achieved higher retrieval-oriented performance than the retrieval-only baseline. However, intervals overlap for some metrics, indicating that the improvements should be interpreted as evidence of technical improvement rather than definitive proof of superiority across all possible query distributions.

4.4 Domain Routing Performance

The domain-routing mechanism achieved a category accuracy of 0.870 across the 200-query benchmark. This result indicates that most user queries were assigned to the correct healthcare repository. However, the result also shows that routing was not perfect, particularly for conversational layperson-style queries containing overlapping terms across public-health, doctor-registry, OPD, and triage domains.

4.5 Case Triage Performance

Triage decision accuracy was 0.800 on the 25 benchmark queries containing escalation labels. The system correctly handled most escalation and management decisions, but some errors remained. These errors indicate that further refinement of triage rules and severity thresholds is needed before clinical deployment.

4.6 Error Analysis

Error analysis was conducted for routing errors, triage decision errors, and retrieval failures. Table 5 summarizes the observed error types.

Table 5 summarizes the routing, triage, and retrieval errors.

Table 5: Summary of error analysis for the proposed lightweight RAG system.

Error Type	Count
Routing errors	26
Triage decision errors	5
Retrieval failures	25

Among 200 benchmark queries, 26 routing errors were observed, corresponding to a routing error rate of 13%. Most routing errors occurred for conversational public-health queries that resembled triage requests, such as fever, rash, breathing difficulty, dehydration, or urgent-care phrasing. Five triage decision errors were observed among 25 triage-labeled cases. These errors were associated with seizure-related queries in children, where the rule-based triage mechanism assigned a severity score below the escalation threshold. In addition, 25 retrieval failures were observed among relevance-annotated queries, representing a retrieval failure rate of approximately 17.6%. Retrieval failures commonly involved disease-specific treatment or transmission queries where the retrieved records did not match the manually annotated relevant document identifiers.

4.7 Efficiency Evaluation

Both methods demonstrated low computational overhead. The retrieval-only baseline achieved a mean latency of 0.781 ms, while the proposed lightweight RAG method achieved a mean latency of 0.944 ms. The 95th-percentile latency was 1.635 ms for retrieval only and 2.054 ms for the proposed method. Average memory consumption remained below 9 KiB per query for both methods.

5 Discussion

The present study developed and evaluated a lightweight Retrieval-Augmented Generation (RAG)-based Public Health Agent designed to support frontline healthcare workers, nursing staff, healthcare assistants, and community health personnel operating within primary healthcare environments. Unlike many healthcare AI systems that primarily target physicians or

specialist clinicians, the proposed framework focuses on supporting healthcare workers who frequently serve as intermediaries between patients and healthcare providers. Such personnel are often responsible for patient registration, symptom collection, health education, triage assistance, referral coordination, and communication with physicians.

Evaluation on a benchmark of 200 healthcare queries demonstrated that the proposed lightweight RAG framework consistently outperformed a retrieval-only baseline. At the best-performing retrieval depth ((k=5)), Precision@k increased from 0.244 to 0.280, Recall@k improved from 0.677 to 0.739, MRR increased from 0.643 to 0.712, and nDCG@k improved from 0.630 to 0.700. These findings indicate that the integration of domain routing, semantic retrieval, deterministic physician and OPD lookup mechanisms, and rule-based triage contributed to improved retrieval quality and ranking performance. The results suggest that combining healthcare-specific operational modules can provide measurable benefits beyond direct retrieval alone.

Sensitivity analysis further demonstrated that retrieval performance varied with retrieval depth. Increasing (k) improved Recall@k for both methods, while Precision@k decreased as additional documents were retrieved. Nevertheless, the proposed lightweight RAG framework consistently outperformed the retrieval-only baseline across all tested values of (k). This finding suggests that the framework remains robust under different retrieval configurations and that the observed improvements are not limited to a single parameter setting.

An important finding of the study concerns response grounding. The proposed lightweight RAG framework achieved a groundedness score of 0.846 compared with 0.825 for the retrieval-only baseline, while reducing the unsupported sentence rate from 0.175 to 0.154. Although groundedness was not perfect, the results demonstrate that the majority of generated responses remained supported by retrieved evidence. In healthcare applications, reducing unsupported statements is particularly important because inaccurate or unsupported information may negatively influence operational decision-making. The observed improvement in groundedness suggests that domain-aware retrieval and structured healthcare workflows can improve response reliability without requiring computationally intensive large language model pipelines.

Bootstrap confidence interval analysis provided additional insight into result stability. The estimated 95% confidence intervals for Recall@k, MRR, nDCG@k, groundedness, routing accuracy, and triage accuracy remained relatively narrow, suggesting reasonable consistency of performance across the benchmark. Although confidence intervals for some metrics overlapped between the retrieval-only and proposed approaches, the proposed framework consistently achieved higher point estimates across retrieval-oriented measures. These findings provide additional support for the robustness of the observed performance improvements.

The operational evaluation demonstrated encouraging, although not perfect, performance. Domain-routing accuracy reached 87%, indicating that most benchmark queries were correctly assigned to their respective healthcare repositories. Routing errors primarily occurred for conversational

layperson queries containing overlapping terminology across public-health, referral, and triage domains. Similarly, triage decision accuracy reached 80% on benchmark cases containing escalation labels. Error analysis revealed that all five triage errors involved seizure-related pediatric scenarios where the rule-based severity score remained below the escalation threshold. These findings suggest that future versions of the framework should incorporate more refined severity assessment rules and clinical escalation criteria.

Error analysis further identified 25 retrieval failures among relevance-annotated benchmark queries. Most retrieval failures involved disease-specific treatment, transmission, or symptom queries where the retrieved records differed from manually annotated relevant records. While such failures reduced retrieval effectiveness, they also provide useful guidance for future refinement of retrieval strategies, repository design, and semantic matching approaches.

As illustrated in Figure 1, the framework integrates domain-aware routing, semantic retrieval, structured healthcare repositories, deterministic lookup modules, and rule-based triage within a lightweight operational pipeline. A notable advantage of the proposed framework is its computational efficiency. Average response latency remained below 2 ms even at (k=5), while memory consumption remained below 13 KiB per query. These findings suggest that the framework can operate on modest computing infrastructure without requiring specialized hardware, cloud computing resources, or continuous access to commercial AI services. Such characteristics may be particularly valuable in low-resource healthcare settings where reliable internet connectivity and computational resources are limited.

From a healthcare operations perspective, the proposed framework should be regarded as an assistive information-support system rather than a diagnostic or clinical decision-making system. The objective is not to replace physicians, nurses, or other healthcare professionals. Instead, the system is intended to facilitate access to healthcare information, assist with routing and referral workflows, support symptom-based information retrieval, and provide operational guidance within predefined healthcare processes. Final clinical decisions must remain the responsibility of qualified healthcare professionals.

Several limitations should be acknowledged. First, the evaluation was conducted using expert-curated synthetic healthcare datasets and benchmark queries rather than real-world clinical interactions. Consequently, the reported results may not fully reflect performance under operational healthcare conditions. Second, although the benchmark was expanded to 200 queries, the evaluation remains limited to internally developed datasets. Third, the retrieval component relies primarily on lightweight semantic matching and structured lookup mechanisms and may not capture all complex relationships between symptoms, diseases, and healthcare workflows. Fourth, no clinician-centered evaluation was performed; therefore, the practical usefulness of the system for healthcare workers remains to be validated through user studies.

Future work should focus on external validation using healthcare professionals and real-world

operational scenarios. Additional investigations involving physicians, nurses, community health workers, and public-health personnel could provide valuable insights regarding usability, trust, workflow integration, and clinical relevance. Further enhancements may include knowledge-graph integration, multilingual support, voice-based interaction, mobile deployment, and integration with electronic health record systems. Such developments may improve retrieval quality, usability, and applicability across diverse healthcare environments.

Overall, the study demonstrates the technical feasibility of a lightweight healthcare-oriented RAG framework and highlights the potential of retrieval-based information-support systems for operational healthcare workflows. However, further validation is required before conclusions regarding real-world clinical effectiveness can be drawn.

6 Conclusion

This study presented a lightweight Retrieval-Augmented Generation (RAG)-based Public Health Agent designed to support frontline healthcare workers, nursing personnel, healthcare assistants, and community health staff operating in primary healthcare environments. The framework integrates structured healthcare repositories, domain-aware routing, semantic retrieval, deterministic healthcare operations, and rule-based triage workflows to provide evidence-supported responses across multiple healthcare domains, including public health information, physician referral, patient management, OPD scheduling, and case-triage support.

Evaluation using a benchmark of 200 healthcare queries demonstrated that the proposed lightweight RAG framework consistently outperformed a retrieval-only baseline across all evaluated retrieval depths. At the best-performing configuration ((k=5)), the framework achieved Recall@k of 0.739, MRR of 0.712, nDCG@k of 0.700, and groundedness of 0.846. Sensitivity analysis showed that performance improvements remained consistent across (k=1), (k=3), and (k=5), while bootstrap confidence interval analysis indicated that the observed results were reasonably stable across benchmark samples. Error analysis further identified routing ambiguities, triage-threshold limitations, and retrieval mismatches that provide valuable directions for future system refinement.

The framework maintained low computational overhead, with average response latency below 2 ms and memory consumption below 13 KiB per query. These characteristics suggest that the proposed architecture may be suitable for deployment in resource-constrained healthcare environments where computational resources, internet connectivity, or access to commercial AI services may be limited.

The findings demonstrate the technical feasibility of using a lightweight healthcare-oriented RAG architecture to support operational healthcare information workflows. However, the current evaluation was conducted using expert-curated synthetic healthcare datasets and internally

developed benchmark queries. Consequently, the reported results should be interpreted as evidence of technical feasibility and system performance under controlled evaluation conditions rather than proof of real-world clinical effectiveness.

The proposed framework is intended as an assistive information-support tool and not as a diagnostic or autonomous clinical decision-making system. Final healthcare decisions must remain the responsibility of qualified healthcare professionals. Future work should focus on external validation, clinician-centered evaluation, real-world deployment studies, multilingual support, knowledge-graph integration, enhanced semantic retrieval techniques, and integration with electronic health record systems. Such investigations will be necessary to determine the practical utility, reliability, safety, and adoption potential of the framework in operational healthcare environments.

6.1 Data Availability

The datasets developed for this study are publicly available through the Zenodo repository [11] and may be accessed for research, educational, and benchmarking purposes. The repository contains the Public Health, Doctor Registry, Patient Management, OPD Scheduling, and Case Triage datasets used in the experiments reported in this paper. All patient-related, scheduling, and triage records are synthetic and were created exclusively for methodological evaluation and research. No personally identifiable information (PII) or protected health information (PHI) is contained within the released datasets.

Clinical Trial Number

Not applicable.

Funding

The author received no specific funding for this work.

Ethics and Consent to Participate

Not applicable.

References

[1] Mustapha Abdelouahed, Dana Yateem, Chadi Amzil, Ilyas Aribi, El Hassan Abdelwahed, and Salim Fredericks. Integrating artificial intelligence into public health education and healthcare: Insights from the covid-19 and monkeypox crises for future pandemic readiness. *Frontiers in Education*, 10, 2025.

[2] Ala'a Alquayt, Ohoud Aljuhani, Abdullah F. Alharthi, Rahaf Alqahtani, Anas Khan, Ahmed Al-jedai, Abdulqader Almoeen, Mohammed Alshennawi, Hisham A. Badreldin, Abdulrhman Aljouie, Lubna A. Alnasser, Abdulmajeed M. Alshehri, Mohammed Y. Alzahrani, Haifa A. Alhaidal, Raghad Alhajaji, Salman Alotaibi, Esraa Z. Redhwan, Fahad Alharthi, Badr G. Alghamdi, and Khalid Al Sulaiman. Ai-driven healthcare innovations for enhancing clinical services during mass gatherings (hajj): Task force insights and future directions. *BMC Health Services Research*, 25:876, 2025.

[3] Jerome Visperas Cleofas. Towards a framework for human rights-based stakeholdership of nursing academia in the healthcare artificial intelligence governance ecosystem: A discussion paper. *Nurse Education in Practice*, 92:104737, 2026.

[4] Krupal Joshi and Krishna Jasani. Role of artificial intelligence in strengthening the primary healthcare system. In *Multimodal AI in Healthcare: Fusion Intelligence for Advanced Health Diagnostics*, pages 387–413. Springer Nature Singapore, Singapore, 2026.

[5] Rana Muhammad Amir Latif, Zakawat Liaqat, Farah H. Zawaideh, Marwan Abu-Zanona, Abdulmalik Ahmad Al Jabr, and Muhammad Adrees. Empowering remote healthcare through edge-ai: A lightweight framework for low-resource environments. In *2026 7th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–5, 2026.

[6] Ankita Nayak, Ipseeta Satpathy, and Alex Khang. Employing ai-powered clinical decision support to revolutionize healthcare: Integrating data analytics with informed diagnostics. In Alex Khang, editor, *Revolutionizing Digital Healthcare Through Artificial Intelligence and Automation*, pages 433–446. Academic Press, 2026.

[7] Vineet Kumar Pathak, Ajeet Singh Bhadoria, Shubham Gupta, Mahek Rai, and Priyanka Parhad. Artificial intelligence in primary health care. In *Multimodal AI in Healthcare: Fusion Intelligence for Advanced Health Diagnostics*, pages 363–386. Springer Nature Singapore, Singapore, 2026.

[8] Sandeep Reddy. Foundation for artificial intelligence-driven democratization of healthcare access. *Learning Health Systems*, 10(3):e70093, 2026.

[9] Hein Minn Tun, Hanif Abdul Rahman, Lin Naing, and Owais Ahmed Malik. Trust in artificial intelligence-based clinical decision support systems among health care workers: Systematic review. *Journal of Medical Internet Research*, 27:e69678, 2025.

[10] Daniele Virgillito and Caterina Ledda. Personalized ai for workplace health promotion: Performance management and healthcare worker engagement through digital analytics. *Frontiers in Public Health*, 13, 2026.

[11] Sayed Mohammed Zeeshan. Expert-curated synthetic healthcare datasets for retrieval-augmented public health and clinical decision support systems, 2026.