World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

# VERA: A WHATSAPP-BASED RETRIEVAL-AUGMENTED GENERATION SYSTEM FOR AUTOMATED HEALTH MISINFORMATION FACT-CHECKING IN LOW- AND MIDDLE-INCOME COUNTRIES

Brighton Mukundwi [1,*] and Delvin Tadiwa Vengesai [2]

[1] Department of Data Analytics and Visualisation, Faculty of Computer Science, Yeshiva University, NY, USA.
[2] Department of Biological Sciences and Ecology, Faculty of Science, University of Zimbabwe, Harare, Zimbabwe.
* Corresponding Author

ORCID Details
D elvin Tadiwa Vengesai: https://orcid.org/0009-0001-4948-5729
Brighton Mukundwi: https://orcid.org/0009-0003-8516-9656

World Journal of Advanced Research and Reviews, 2026, 31(03), 1325–1335

Publication history: Received on 11 August 2026; revised on 17 September 2026; accepted on 19 September 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.3.2435



## ABSTRACT

Health misinformation is a growing public health emergency, with disproportionate impact in low- and middle-income countries (LMICs), where fragile information ecosystems increase susceptibility to infodemic harm. This paper presents VERA (Verified Evidence Retrieval Assistant), a WhatsApp-based fact-checking system that combines Retrieval-Augmented Generation (RAG) with an open-source large language model (LLM) to counter health misinformation in resource-constrained settings. VERA accepts user-submitted health claims, forwarded messages, or article links through WhatsApp, retrieves evidence from PubMed, Semantic Scholar, and WHO/CDC sources, and returns a structured verdict, supported, refuted, nuanced, or evidence-insufficient, with citations and a plain-language explanation. Llama 3 8B serves as the reasoning backbone, chosen for its competitive biomedical performance under RAG, zero licensing cost, and deployment feasibility on modest cloud infrastructure. VERA is a theoretical architecture and hasn't been deployed or empirically validated therefore, the contribution of this paper is the design itself, together with a rigorous evaluation protocol, precedent from comparable deployed systems, and worked examples demonstrating how the pipeline would process representative claim types. We detail the complete technical architecture, the rationale behind each design choice, a multi-dimensional evaluation framework spanning technical accuracy, usability, and public health impact, and the governance safeguards required before real-world deployment. VERA demonstrates that scientifically rigorous, evidence-grounded fact-checking can be delivered through infrastructure LMIC populations already use and can afford.

**Keywords:** Health Misinformation; Infodemic; Retrieval-Augmented Generation; Large Language Models; Whatsapp; Low- And Middle-Income Countries.

# 1 INTRODUCTION

The World Health Organization (WHO) coined the term "infodemic" to describe the oversupply of accurate and misleading information during epidemics, which can undermine public health responses [1]. Its effects are measurable: false claims during COVID-19 contributed to methanol poisoning deaths in Iran, while disinformation reduced mask adoption, vaccination uptake, and care-seeking elsewhere [2] [3]. A review identified 2,311 reports of COVID-19 rumours and conspiracy theories across 87 countries, 82% of which were demonstrably false [2]. LMICs face greater vulnerability due to lower digital health literacy, reliance on unofficial information networks, limited access to reliable sources, and institutional mistrust [4]. Misinformation is also more prevalent in lower-income settings, where people are more exposed to it and less likely to receive corrections [5]. WhatsApp presents both a challenge and an opportunity for public health in LMICs. It is widely used across Sub-Saharan Africa, South Asia, and Latin America [6] [7], while also serving as a major channel for health disinformation [8]. WhatsApp-based interventions avoid barriers such as app installation, registration, and web browsing. The platform has also been repeatedly zero-rated by mobile carriers across Africa [9] [10], with adoption in South Africa linked to reduced connectivity costs for lower-income users [11]. For the poorest Africans, mobile data can consume nearly a third of monthly income [12], making a WhatsApp-based system potentially more accessible than browsing multiple fact-checking websites. Existing automated fact-checkers, dedicated websites, warning labels, and platform classifiers have limited reach in LMICs, where digital health literacy and awareness of fact-checking services remain low. No deployed system currently provides interactive, real-time, evidence-based health claim verification directly through a messaging app. This paper proposes VERA (Verified Evidence Retrieval Assistant), a complete technical architecture for a WhatsApp-based health fact-checking chatbot.

This paper's contributions are:

- A deployable, modular architecture combining a RAG pipeline for health fact-checking with the WhatsApp Business API;
- Evidence-based selection of an open-source LLM backbone suited to LMIC resource constraints;
- A rigorous, multi-dimensional evaluation protocol, spanning automated RAG-specific metrics, adversarial robustness testing, and a phased human-subjects evaluation roadmap, precise enough for other research teams to build and test against;
- A governance and safety framework addressing consent, human escalation, data protection, and harm from incorrect verdicts, adapted to the regulatory and infrastructural realities of LMICs;
- Grounding of the design in precedent, through comparative analysis of deployed WhatsApp and chatbot health-communication systems.

## 1.1 AI-Based Health Misinformation Detection

Deep neural architectures have superseded rule-based classifiers in automated health misinformation detection. On labelled COVID-19 misinformation datasets, hybrid CNN-LSTM models enhanced with psychological elaboration cues have achieved accuracy above 97% [13], while ensemble machine learning techniques show 79-95% accuracy across major social media platforms [14]. These models are intrinsically reactive. They categorize posts that already exist but cannot respond to natural-language enquiries or explain their decisions, qualities that matter for public engagement. LLM-based methods have pushed the field forward recently. TrumorGPT outperforms several healthcare benchmarks using graph-based RAG over continuously updated semantic health knowledge graphs to separate confirmed claims from rumours [15]. Moreover, a tripartite collective intelligence system combining deep neural networks, LLMs, and crowdsourced human judgment further improved detection of subtle misinformation [16]. A 2025 scoping review of 63 studies on AI-based infodemic interventions found that most concentrated on monitoring and surveillance, with fewer than 10% deploying interactive user-facing tools and minimal LMIC-targeted deployment [17]. VERA aims to directly close this gap.

## 1.2 Retrieval-Augmented Generation for Medical Knowledge

In high-stakes scenarios, RAG minimises hallucination and knowledge staleness, two critical failure modes of standalone LLMs [18]. The MedRAG benchmark, comparing 41 combinations of corpora, retrievers, and backbone LLMs across 7,663 medical questions, exhibited accuracy increases of up to 18% over chain-of-thought prompting, elevating GPT-3.5 and open-source Mixtral to GPT-4-level performance [19]. Self-BioRAG, a biomedical RAG framework with self-reflective generation, outscored leading open-source models by 7.2 percentage points on medical QA benchmarks [20]. A thorough evaluation of RAG in healthcare across 67 studies supports RAG as the leading way for grounding medical LLM outcomes, with text-based RAG accounting for 54% of approaches [18]. Among open-source models, Llama 3-based variations have the strongest medical performance. MMed-Llama 3 (8B parameters), trained on a 25.5-billion-token multilingual medical corpus spanning six languages, compares with GPT-4 on conventional medical benchmarks [21]. Llama 3 70B was the best-performing open-source model in diagnostic benchmarking across 1,933 radiology cases, closely trailing GPT-4o; smaller 8B variations provided competitive accuracy at much lower compute cost [22]. Under

RAG conditions specifically, the performance gap between open-source and proprietary models narrows dramatically, making open-source models a feasible solution for resource-constrained deployments [19].

### 1.3 WhatsApp and mHealth in LMICs

WhatsApp has demonstrated practical value in research and clinical applications across LMICs. A scoping review identified eleven LMIC-based health research studies using WhatsApp for participant engagement, data collection, and behaviour-change interventions [6]. An IEEE analysis highlighted its use for clinical coordination, patient follow-up, and communication among community health workers [7]. Practitioner perspectives on vaccine promotion further emphasize WhatsApp's affordability, multimedia capabilities, and group messaging, noting that most of its roughly two billion active users are in LMICs [23]. In Africa, health chatbots have supported education and counselling, although evidence of effectiveness remains limited and concerns around trust, privacy, and evaluation persist [24]. Deployed examples include Malawi's Ministry of Health WhatsApp chatbot, which recorded over 347,000 accesses during COVID-19 and later added vaccine-misinformation tracking [25]; South Africa's GREAT4Diabetes chatbot, which provided multilingual diabetes education to over 8,000 users during pandemic-related care disruptions [26]; and a randomized WhatsApp messaging intervention in Zimbabwe that increased COVID-19 knowledge and reduced non-adherence to public health guidance among 27,000 subscribers [27]. Despite this evidence, no documented architecture has integrated WhatsApp with an LLM-RAG fact-checking pipeline for infodemic management. Section 3.1 examines these systems in greater depth as precedents for VERA's design.

## 2 MATERIALS AND METHODS

### 2.1 System Architecture

VERA's architecture (Figure 1) comprises seven functional modules across three tiers: a messaging interface layer, an orchestration and retrieval layer, and an inference layer. The design prioritizes modularity, stateless computation, and cloud deployability without requiring on-premises GPU hardware.

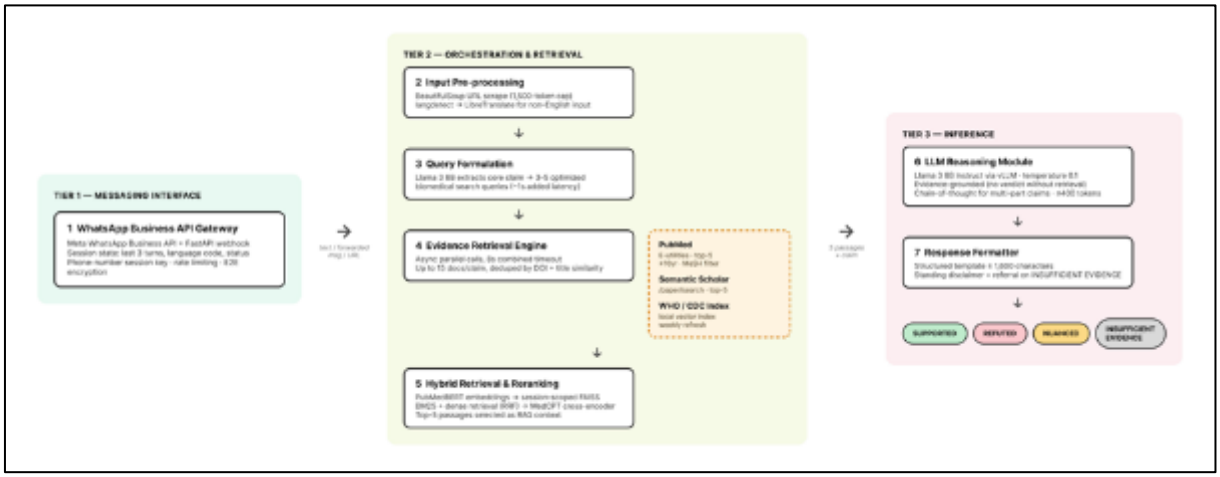


**Figure 1: VERA's three-tier architecture: the messaging interface, orchestration and retrieval, and inference layers, annotated with the technology used at each module, the three external evidence sources queried in parallel, and the four possible verdict outputs.**

#### 2.1.1 WhatsApp Business API Gateway

Users interact with VERA exclusively through WhatsApp. VERA operates as a verified WhatsApp Business account via Meta's WhatsApp Business API. Inbound messages, plain-text claims, forwarded messages, or article URLs, trigger webhook events to a Python/FastAPI server. The gateway handles message parsing, session management keyed to phone number, and rate-limiting. WhatsApp's end-to-end encryption protects submitted claims in transit. Each session maintains a thin state object: the last three conversational turns for context continuity, an auto-detected language code, and a processing status flag. This lightweight design keeps memory overhead negligible and enables horizontal scaling through container orchestration, for example AWS ECS or Google Cloud Run.

#### 2.1.2 Input Pre-processing Module

Input arrives in one of three forms: a free-text health claim, a forwarded WhatsApp message, or a URL to an external article. For URL inputs, a web scraper (BeautifulSoup) extracts article body text, strips navigation and boilerplate, and truncates output to 1,500 tokens. A language detection module (langdetect) identifies non-English input and routes it

to LibreTranslate, an open-source translation API, before further processing. This step matters for LMIC deployment, where English is rarely the primary language of health misinformation.

### 2.1.3 Query Formulation Module

Health claims are rarely phrased using scholarly vocabulary consistent with bibliographic databases. A lightweight prompt instructs Llama 3 8B to extract the core factual assertion from the pre-processed input and generate three to five optimized search queries using medical terminology. For instance, the assertion "turmeric cures cancer" is expanded into queries such as "curcumin anti-tumour efficacy systematic review" and "turmeric phytochemical cancer clinical trial." This query-extension step measurably increases retrieval recall on biomedical databases, at a cost of roughly one second of added latency.

### 2.1.4 Evidence Retrieval Engine

VERA three source categories in parallel using asynchronous HTTP calls, with a combined timeout of 8 queries seconds:

- PubMed (NCBI E-utilities API): the primary biomedical literature source, providing access to over 37 million citations. VERA uses the ESearch and EFetch endpoints to retrieve the top-5 most relevant abstracts per query, filtered to publications within the last 10 years. MeSH terms enable semantic filtering and improve precision.
- Semantic Scholar API: supplements PubMed with broader coverage of interdisciplinary health literature and open-access preprints. The /paper/search endpoint returns the top-5 results per query with abstract availability verified before inclusion.
- WHO/CDC Health Topic Index: a curated, locally maintained vector index of WHO advisory pages and CDC health topic summaries, updated weekly through scheduled scraping. This source matters most for infectious disease, vaccine, and nutrition claims, where WHO position statements constitute authoritative consensus. Local or regional public health authority pages can be appended to this index to add context for specific LMIC settings.

Retrieved documents, up to 15 per claim, are deduplicated by DOI and title-level similarity before the retrieval stage.

### 2.1.5 Hybrid Retrieval and Reranking

Retrieved abstracts are encoded using PubMedBERT-base, a domain-adapted sentence embedding model, and stored in a session-scoped FAISS in-memory index. VERA employs hybrid retrieval: BM25 sparse retrieval (keyword precision) is combined with dense vector retrieval (semantic similarity) through reciprocal rank fusion. The top-5 most relevant passages form the RAG context window. A cross-encoder re-ranker (MedCPT-Cross-Encoder) performs a final relevance pass to ensure the highest-quality evidence is prioritized before LLM synthesis, directly addressing the "lost-in-the-middle" problem documented in medical RAG benchmarking [19].

### 2.1.6 LLM Reasoning Module

The five retrieved passages are injected into a structured system prompt alongside the original claim. VERA uses Llama 3 8B Instruct as the generation backbone, served via vLLM on a single cloud GPU instance. The prompt instructs the model to: (1) assess each retrieved source's relevance and position relative to the claim; (2) synthesize a verdict from four categories, SUPPORTED (evidence corroborates the claim), REFUTED (evidence contradicts the claim), INSUFFICIENT EVIDENCE (no peer-reviewed evidence located), or NUANCED (claim contains partially accurate elements); (3) produce a one-paragraph plain-language explanation; and (4) list cited sources as DOIs or URLs. The system prompt explicitly prohibits verdict generation without retrieved evidence, enforcing grounded responses and preventing hallucination. Temperature is set to 0.1 for factual consistency. Chain-of-thought prompting is applied for complex multi-part claims [18]. A hard output length limit of 400 tokens keeps responses within WhatsApp's practical display constraints.

### 2.1.7 Response Formatter

LLM output is post-processed into WhatsApp-compatible plain text following a structured template:

- CLAIM: [extracted claim]
- VERDICT: [SUPPORTED / REFUTED / INSUFFICIENT EVIDENCE / NUANCED]
- EXPLANATION: [plain-language summary]
- SOURCES: [1] title, URL...

Response length is limited to 1,600 characters, with a continuation option for detailed outputs. A standing disclaimer advises users to consult local health authorities for urgent guidance, and every INSUFFICIENT EVIDENCE verdict includes referral language directing users to their national health ministry or WHO country office.

### 2.2 LLM Selection and Deployment Rationale

The selection of Llama 3 8B Instruct was determined against three criteria: biomedical task performance, deployment cost, and multilingual capability. Each criterion was evaluated against the published evidence base.

#### 2.2.1 Performance

Under RAG conditions, VERA's operational mode, smaller open-source models achieve their greatest relative gains. MedRAG raised both GPT-3.5 and open-source Mixtral to GPT-4-level performance on medical QA, showing that retrieval quality dominates model size as a performance determinant [19]. Self-BioRAG with a 7B backbone outperformed larger unaugmented models on biomedical QA [20]. In open diagnostic benchmarking across 1,933 clinical cases, Llama 3 70B closely trailed GPT-4o, while 8B variants showed competitive accuracy at a fraction of the compute cost [22]. Notably, Meditron-70B, a widely cited medical open-source model, showed surprisingly poor performance in recent multi-model evaluations, 52.9% accuracy in specialized neurological reasoning, suggesting domain-specific fine-tuning alone does not guarantee superiority over strong general-purpose models under RAG [28].

#### 2.2.2 Cost and Deployability

Llama 3 8B can be served via vLLM on a single NVIDIA A10G GPU, with throughput sufficient for hundreds of concurrent users. For minimal-resource deployments, 4-bit quantized GGUF variants run on CPU-only infrastructure, trading roughly 5% accuracy for zero GPU cost, a viable configuration for NGO or ministry deployments without cloud GPU budgets. Section 3.1.1 situates this cost model against the reported six-month operating cost of a comparable deployed WhatsApp health chatbot.

#### 2.2.3 Multilingual Capability

Llama 3 8B supports multiple languages natively, and MMed-Llama 3 8B, a medically fine-tuned derivative, achieves GPT-4-comparable performance while supporting six core languages, a property that matters directly for LMIC deployments where health misinformation circulates in diverse local languages [21]. Operators may substitute MMed-Llama 3 8B for language-heavy deployments with minimal architectural change.

### 2.3 Connectivity and Data-Cost Resilience

WhatsApp does not function without a data connection, and VERA does not claim otherwise. What VERA changes is the connection a user needs and what that connection costs, not whether one is required. Two distinct connectivity dependencies are relevant here, and this section treats them separately.

#### 2.3.1 Why WhatsApp lowers the barrier

Mobile carriers across Africa have repeatedly zero-rated WhatsApp, exempting its traffic from standard data charges [9] [10]. In South Africa, this increased WhatsApp usage regardless of connection type [9], while its low per-message cost has been credited with reducing the cost barrier to sustained connectivity [11]. Even without zero-rating, text-based WhatsApp exchanges use far less data than loading and navigating multiple fact-checking websites. This is significant because mobile data can consume nearly a third of monthly income for the poorest Africans [12], while connectivity remains a major barrier to digital health interventions in LMICs [29]. VERA preserves this advantage through plain-text inputs and outputs, with responses capped at 1,600 characters to keep data use small and predictable. It also benefits from WhatsApp's store-and-forward delivery, allowing messages sent without a signal to queue and deliver when connectivity returns without additional engineering by VERA.

#### 2.3.2 Resilience on VERA's own side

VERA's backend connectivity is separate from the users'. Its FastAPI backend must reach PubMed, Semantic Scholar, and WHO/CDC endpoints to retrieve evidence, and this connection can fail due to hosting-provider network issues, upstream outages, or rate limits. Two measures address this without claiming full offline operation. First, the WHO/CDC Health Topic Index is a locally maintained, weekly-refreshed vector index, so high-consensus claim categories, including vaccines and common infectious diseases, do not require live retrieval at query time. Extending this caching approach to high-frequency PubMed and Semantic Scholar queries, following store-and-forward and local-first patterns used in offline-first mHealth systems [30][31][32], would further reduce reliance on live upstream calls. Second, if a PubMed or Semantic Scholar request times out, VERA fails transparently. It returns an INSUFFICIENT EVIDENCE verdict stating that retrieval could not be completed, rather than hanging or generating a verdict from the WHO/CDC cache alone. The original query is then queued for retry when backend connectivity is restored.

### 2.4 Proposed Evaluation Framework

Evaluation of VERA must address three dimensions: technical accuracy, LMIC usability, and public health impact. Given the conceptual stage of this work, we propose a phased evaluation roadmap, detailed enough that an independent team could build VERA from this specification and reproduce the tests below without further design decisions.

#### 2.4.1 Technical Accuracy and RAG-Specific Rigor

Automated evaluation uses established health-claim benchmarks. The HealthFC dataset, containing 750 claims with expert evidence assessments and verdict labels, provides ground truth for macro-F1 scoring across verdict categories [33]. Supplementary evaluation on LIAR-PLUS [34] extends coverage to non-biomedical health-adjacent claims. Primary metrics are macro-F1 across the four verdict classes (target: >0.75), citation precision, the proportion of returned DOIs linked to evidence relevant to the claim (target: >0.90), hallucination rate, the proportion of model claims unsupported by retrieved passages (target: <5%), and end-to-end latency, with a 15-second target at the 90th percentile. However, end-to-end accuracy alone is insufficient for evaluating a RAG system, so three additional measures are included. First, component-level metrics assess retrieval and generation separately. Faithfulness measures whether explanations are supported by retrieved passages, while context precision and recall are computed using the RAGAs framework [35]. Second, adversarial and noise-robustness testing, following medical RAG benchmarks [36][37], introduces misleading evidence, contradictory sources, and claims with no retrievable evidence to test whether VERA correctly produces INSUFFICIENT EVIDENCE and NUANCED verdicts. Systematic RAG evaluation reviews likewise emphasize component-level and adversarial testing alongside end-to-end accuracy [38] [39]. Third, retrieval and generation are logged and scored at each pipeline stage (Section 2.1), allowing failures to be attributed to query formulation, retrieval, reranking, or generation rather than only to the final output.

#### 2.4.2 LMIC Usability and Comprehension

A structured mixed-methods pilot study with 120 to 150 participants across two LMICs, stratified by age, literacy, and prior health knowledge, would evaluate: verdict comprehension using a structured post-interaction questionnaire; trust calibration, whether users appropriately discount INSUFFICIENT EVIDENCE verdicts rather than treating absence of evidence as disproof; task completion rate across the three input modalities, free text, forwarded message, and URL; and willingness to share VERA verdicts with family and community members as a proxy for network-level impact. Qualitative focus groups would explore barriers to adoption and the language appropriateness of explanations.

#### 2.4.3 Public Health Impact

A cluster-randomized controlled feasibility trial in a subsequent phase would assess change in self-reported misinformation belief scores, using validated scales from prior infodemic research, before and after VERA access; intent to comply with evidence-based public health recommendations; and referral rate to health authorities following INSUFFICIENT EVIDENCE verdicts, an indicator of whether VERA appropriately redirects users when its evidence base is insufficient.

## 3 RESULTS AND DISCUSSION

### 3.1 Case Studies

Since VERA has not been deployed, this section places the design against systems that have been deployed and evaluated to establish architectural feasibility. It does not substitute for the proposed empirical evaluation procedures. Five deployed systems help define VERA's design space and demonstrate its feasibility.

Firstly, Malawi's Ministry of Health operated a national COVID-19 WhatsApp chatbot from 2020 to 2023, recording over 347,000 accesses and later adding vaccine-misinformation tracking that captured 198 user-reported rumours [25]. This demonstrates that government-operated WhatsApp health services can reach national scale in an LMIC and that users will report misinformation through the same channel they use to seek information. However, the chatbot experienced downtime on 38% of days, particularly during a major feature expansion, highlighting the backend reliability challenges addressed in Section 2.3.

Auntie Meiyu, a fact-checking chatbot embedded in LINE, is the closest functional analogue to VERA [40]. Interviews with 27 adopters found that users used it to protect family members from misinformation in private group chats, but reported frustration with detection errors and a "robotic" conversational style. These findings support VERA's plain-language explanations, local-health-authority disclaimer, and inclusion of tone and error framing in its LMIC usability evaluation.

Third, South Africa's GREAT4Diabetes WhatsApp chatbot delivered diabetes education through voice messages in English, Afrikaans, and isiXhosa to over 8,000 users during pandemic-related disruptions to in-person care [26]. It demonstrates multilingual WhatsApp delivery at scale and reports approximately US$30,600 in operating costs over six months, providing a real-world benchmark for VERA's cost model.

A randomized WhatsApp intervention in Zimbabwe reached 27,000 subscribers through a civil-society partner and produced a 0.26 standard-deviation increase in COVID-19 knowledge and a 30-percentage-point reduction in non-adherence to lockdown guidance [27]. This provides direct evidence that WhatsApp-based corrective health messaging can influence behaviour in an LMIC population.

Finally, a proposed architecture for identifying previously fact-checked content within WhatsApp while preserving end-to-end encryption addresses a different use case: passive detection of known misinformation rather than active, user-submitted verification [41]. It nevertheless confirms that privacy constraints are an established design consideration, supporting VERA's approach of accessing message content only when users explicitly submit it for verification.

### 3.2 Ethical, Governance, and Safety Considerations

Ethical deployment of a health fact-checking system carries risks distinct from most consumer chatbot applications. An incorrect verdict can be acted on as health guidance, with potentially fatal downstream consequences. This section sets out the governance architecture VERA requires before any real-world deployment.

#### 3.2.1 Transparency and consent

Every VERA response discloses that the verdict was generated by an AI system, not a clinician or fact-checking professional, consistent with recommendations that AI-generated health content be clearly labelled to avoid misplaced authority [42]. Labelling alone is a weak safeguard: experimental evidence shows that users can come to over-rely on AI-generated advice even when told a system is AI-based, particularly once they have built up habitual trust in it [43][44]. VERA's standing disclaimer and mandatory referral language on every INSUFFICIENT EVIDENCE verdict are designed against this specific failure mode.

#### 3.2.2 Human oversight and escalation

Fully automated moderation of health content is not adequate on its own; ethical frameworks for AI-driven health content moderation consistently call for hybrid human-AI review rather than AI-only judgment, particularly for ambiguous or high-stakes cases [45]. VERA is designed with an escalation path: claims returning NUANCED or INSUFFICIENT EVIDENCE verdicts at a rate above a defined threshold for a given topic are flagged for review by a clinical advisory board, and a user-facing feedback mechanism lets any user report a verdict they believe is wrong, routing it to the same review process.

#### 3.2.3 Data protection under a fragmented regulatory landscape

Data protection law across the African countries VERA targets is uneven. A scoping review across twelve African countries found no country-specific AI regulation and highly fragmented data-protection coverage [46], and legal analyses in specific jurisdictions confirm this fragmentation extends to health data specifically [47]. Rather than designing to the least protective applicable law, VERA's architecture follows the more specific pattern of "governance-by-design" proposed for digital health in Sub-Saharan Africa, embedding data-minimization and purpose-limitation rules directly into the system's infrastructure rather than relying on policy compliance alone [48]. Concretely, VERA's session state retains only the last three conversational turns, phone numbers are not linked to any claim data beyond the active session, and no submitted claim content is retained after a verdict is delivered unless a user explicitly opts into flagging it for review.

#### 3.2.4 Harm from false negatives

A claim wrongly marked SUPPORTED is a more dangerous failure than one wrongly marked INSUFFICIENT EVIDENCE, because the former is more likely to be acted on and shared. VERA's temperature setting, evidence-grounding constraint, and NUANCED category are conservative by design for this reason: the system is built to under-claim rather than over-claim confidence, and the adversarial testing proposed in Section 2.4.1 is specifically designed to surface cases where this bias fails.

#### 3.2.5 Community and clinical accountability

A clinical advisory board conducting periodic quality audits, transparency disclosures on AI involvement, and compliance with applicable data protection legislation are treated in this design as non-negotiable preconditions for scaling past a pilot, consistent with governance frameworks for responsible AI in healthcare more broadly [42].

### *Limitations and Future Work*

Several limitations remain despite the additions above. Firstly, this paper's primary limitation is that VERA has not been built end-to-end or tested against real users or real claims. Section 2.4 specifies exactly what that testing would look like, but until it is carried out, every performance target in this paper is a design target, not a measured result. Secondly, PubMed and Semantic Scholar are overwhelmingly English-language scholarly sources. Automated translation partially mitigates this, but LMIC languages with limited biomedical literature, Chichewa, Amharic, and Tigrinya among them, will produce frequent INSUFFICIENT EVIDENCE verdicts that may read to users as "no evidence exists" rather than "no evidence was found in the sources searched." Future work should pair VERA with community-health-worker-curated vernacular knowledge repositories and regional health ministry materials, rather than treating automated translation alone as sufficient. Moreover, VERA verifies claims; it does not address the social mechanisms that make misinformation compelling in the first place: emotional resonance, community identity, and trusted messengers. Integration with behavior-change communication frameworks and community health worker networks, rather than claim verification alone, is the more direct path to population-level impact, and is left for future work.

## 4 CONCLUSION

VERA demonstrates that accessible, evidence-grounded health fact-checking is architecturally viable through WhatsApp in LMIC contexts. Section 3.1 shows that comparable systems have already reached hundreds of thousands of users at national scale, and Section 2.3 shows why WhatsApp specifically lowers the cost and connectivity barrier rather than requiring users to browse the open web. By combining WhatsApp's reach, the hallucination-mitigating structure of RAG, and the zero-licensing cost of Llama 3 8B, VERA offers a technically grounded blueprint for infodemic management at the intersection of AI, digital health, and global equity. It requires no new app, no registration step, and no digital literacy beyond sending a WhatsApp message, deliberately matching the communities most affected by health misinformation to the tool they already use. The evaluation protocol in Section 2.4 and the governance framework in Section 3.2 are what would need to be executed, in that order, to move this design from architecture to deployed system; that empirical work, not further architectural refinement, is the necessary next step.

## STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

### *Acknowledgments*

The authors declare that this work received no external funding. The proposed VERA architecture relies entirely on open and publicly accessible resources, and the authors gratefully acknowledge the infrastructure that makes this work reproducible without proprietary dependencies.

### *Disclosure of conflict of interest*

The authors declare no conflict of interest.

### *Statement of ethical approval*

This work presents a proposed system architecture and does not report empirical research involving human participants, animal subjects, or the collection or analysis of personal data. No ethical approval was required for the work described in this paper. The phased evaluation roadmap proposed in Section 2.4, including the LMIC usability pilot and the cluster-randomized feasibility trial, would require institutional ethical review and informed consent procedures in each participating country before implementation; these approvals are a precondition for that future work, not something obtained for this paper.

### *Statement of informed consent*

This paper presents a proposed system architecture and does not involve human participants, case studies, surveys, or interviews. No individual data was collected, and informed consent was therefore not applicable to this work. The phased evaluation roadmap described in Section 2.4 would require informed consent from participants once carried out, but that requirement applies to future work, not to this paper.

## REFERENCES

[1] World Health Organization. Infodemic. Geneva: WHO; 2020. Available from: https://www.who.int/health-topics/infodemic

[2] Islam MS, Sarkar T, Khan SH, Kamal AH, Hasan SM, Kabir A, Yeasmin D, Islam MA, Chowdhury KI, Anwar KS, Chughtai AA. COVID-19–related infodemic and its impact on public health: A global social media analysis. The American journal of tropical medicine and hygiene. 2020 Aug 10;103(4):1621.

[3] Tasnim S, Hossain MM, Mazumder H. Impact of rumors and misinformation on COVID-19 in social media. Journal of preventive medicine and public health. 2020 Apr 2;53(3):171.

[4] Naeem SB, Bhatti R, Khan A. An exploration of how fake news is taking over social media and putting public health at risk. Health Information & Libraries Journal. 2021 Jun;38(2):143-9.5. McCool J, Ahmad S, Mukhtar R, Mason-Jones S. Mobile health (mHealth) in low- and middle-income countries. Annu Rev Public Health. 2021;42:297-315.

[5] Manji K, Hanefeld J, Vearey J, Walls H, De Gruchy T. Using WhatsApp messenger for health systems research: a scoping review of available literature. Health policy and planning. 2021 Jun 1;36(5):774-89.

[6] Weaver NS, Roy A, Martinez S, Gomanie NN, Mehta K. How WhatsApp is transforming healthcare services and empowering health workers in low-and middle-income countries. In2022 IEEE global humanitarian technology conference (GHTC) 2022 Sep 8 (pp. 234-241). IEEE.

[7] Gagnon-Dufresne MC, Azevedo Dantas M, Abreu Silva K, Souza dos Anjos J, Pessoa Carneiro Barbosa D, Porto Rosa R, de Luca W, Zahreddine M, Caprara A, Ridde V, Zinszer K. Social media and the influence of fake news on global health interventions: Implications for a study on dengue in Brazil. International Journal of Environmental Research and Public Health. 2023 Mar 28;20(7):5299.

[8] Chen A, Feamster N, Calandro E. Exploring the walled garden theory: An empirical framework to assess pricing effects on mobile data usage. Telecommunications Policy. 2017 Aug 1;41(7-8):587-99.10. Stork C, Esselaar S, Chair C. OTT - threat or opportunity for African telcos? Telecomm Policy. 2017.

[9] Shambare R. The adoption of WhatsApp: breaking the vicious cycle of technological poverty in South Africa. Journal of economics and behavioral studies. 2014;6(7):542-50.

[10] Begazo T, Blimpo M, Dutz M. Digital Africa: Technological transformation for jobs. World Bank Publications; 2023 Mar 13.

[11] Sikosana M, Maudsley-Barton S, Ajao O. Advanced health misinformation detection through hybrid CNN-LSTM models informed by the elaboration likelihood model (ELM). In2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) 2025 Aug 7 (pp. 1-11). IEEE.

[12] Naeem J, Gül ÖM, Parlak IB, Karpouzis K, Kadry SN, Salman YB. Detection of Misinformation Related to Pandemic Diseases Using Machine Learning. InInternational Conference on Robotics and Networks 2023 Dec 15 (pp. 147-159). Cham: Springer Nature Switzerland.

[13] Hang CN, Yu PD, Tan CW. TrumorGPT: Graph-based retrieval-augmented large language model for fact-checking. IEEE Transactions on Artificial Intelligence. 2025 May 5;6(11):3148-62.

[14] Zhang Y, Zong R, Shang L, Yue Z, Zeng H, Liu Y, Wang D. Tripartite intelligence: Synergizing deep neural network, large language model, and human intelligence for public health misinformation detection (archival full paper). InProceedings of the ACM Collective Intelligence Conference 2024 Jun 27 (pp. 63-75).

[15] Cianciulli A, Santoro E, Manente R, Pacifico A, Quagliarella S, Bruno N, Schettino V, Boccia G. Artificial intelligence and digital technologies against health misinformation: A scoping review of public health responses. InHealthcare 2025 Oct 18 (Vol. 13, No. 20, p. 2623). MDPI.

[16] Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health. 2025 Jun 11;4(6):e0000877.

[17] Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. InFindings of the Association for Computational Linguistics: ACL 2024 2024 Aug (pp. 6233-6251).

[18] Jeong M, Sohn J, Sung M, Kang J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics. 2024 Jul;40(Supplement_1):i119-29..

[19] Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, Wang Y, Xie W. Towards building multilingual language model for medicine. Nature Communications. 2024 Sep 27;15(1):8384..

[20] Kim SH, Schramm S, Adams LC, Braren R, Bressem KK, Keicher M, Platzek PS, Paprottka KJ, Zimmer C, Hedderich DM, Wiestler B. Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. NPJ digital medicine. 2025 Feb 12;8(1):97.

[21] Jamison A, Rath S, Salomon-Ballada S, Diamond-Smith N, Johnston JS, Barrera A, Rimal R, Ganjoo R. Using WhatsApp for vaccine promotion in low-and middle-income countries: practitioner-informed perspectives. BMJ global health. 2025 Oct;10(10):e018983.

[22] Phiri M, Munoriyarwa A. Health chatbots in Africa: scoping review. Journal of medical Internet research. 2023 Jun 14;25:e35573.

[23] Ndemera I, Hsieh YT, Chirwa K, Ngwira S, Denny A, Wu TS, Lee HY, Chirambo GB, Yosefe SL, Chilima B, Kagoli M. Describing a National Chatbot Deployed by the Ministry of Health in Malawi During the COVID-19 Pandemic: Retrospective Data Analysis. Journal of Medical Internet Research. 2026 Jul 16;28:e80960.

[24] Mash R, Schouw D, Fischer AE. Evaluating the implementation of the GREAT4Diabetes WhatsApp chatbot to educate people with type 2 diabetes during the COVID-19 pandemic: convergent mixed methods study. JMIR diabetes. 2022 Jun 24;7(2):e37882.

[25] Bowles J, Larreguy H, Liu S. Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe. PloS one. 2020 Oct 14;15(10):e0240005.

[26] Sorka M, Gorenshtein A, Aran D, Shelly S. A multi-agent approach to neurological clinical reasoning. PLOS Digital Health. 2025 Dec 4;4(12):e0001106.

[27] Kaboré SS, Ngangue P, Soubeiga D, Barro A, Pilabré AH, Bationo N, Pafadnam Y, Drabo KM, Hien H, Savadogo GB. Barriers and facilitators for the sustainability of digital health interventions in low and middle-income countries: a systematic review. Frontiers in digital health. 2022 Nov 28;4:1014375.

[28] Were MC, Savai S, Mokaya B, Mbugua S, Ribeka N, Cholli P, Yeung A. mUzima mobile electronic health record (EHR) system: development and implementation at scale. Journal of medical Internet research. 2021 Dec 14;23(12):e26381.

[29] Burka D, Gupta R, Moran AE, Cohn J, Choudhury SR, Cheadle T, Mullick R, Frieden TR. Keep it simple: designing a user-centred digital information system to support chronic disease management in low/middle-income countries. BMJ Health & Care Informatics. 2023 Jan 13;30(1):e100641.

[30] Istambul MR, Parlindungan P, Wijaya JH, Zezarina R, Fachrizsal D. SiKurang: Development and Field Evaluation of an Offline-First mHealth System for Community-Based Stunting Risk Detection and Counselling in Low-Connectivity Primary Care Settings. Glosains: Jurnal Sains Global Indonesia. 2026 Apr 7;7(2):286-305.

[31] Vladika J, Schneider P, Matthes F. HealthFC: Verifying health claims with evidence-based medical fact-checking. InProceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) 2024 May (pp. 8095-8107).

[32] Wang WY. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. InProceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2017 Jul (pp. 422-426).

[33] Es S, James J, Anke LE, Schockaert S. Ragas: Automated evaluation of retrieval augmented generation. InProceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations 2024 Mar (pp. 150-158).

[34] Ngo NT, Van Nguyen C, Dernoncourt F, Nguyen TH. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. arXiv preprint arXiv:2411.09213. 2024 Nov 14.

[35] Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. InProceedings of the AAAI conference on artificial intelligence 2024 Mar 24 (Vol. 38, No. 16, pp. 17754-17762).

[36] Brown A, Roman M, Devereux B. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. arXiv preprint arXiv:2508.06401. 2025 Aug 8.

[37] Friel R, Belyi M, Sanyal A. Ragbench: Explainable benchmark for retrieval-augmented generation systems. arXiv preprint arXiv:2407.11005. 2024 Jun 25.

[38] Lee TH, Fussell SR. Countering misinformation in private messaging groups: Insights from a fact-checking chatbot. Proceedings of the ACM on human-computer interaction. 2025 Jan 10;9(1):1-30.

[39] Reis J, Melo PD, Garimella K, Benevenuto F. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation?. arXiv preprint arXiv:2006.02471. 2020 Jun 3.

[40] Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. Social Science & Medicine. 2022 Mar 1;296:114782.

[41] Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior. 2024 Nov 1;160:108352.

[42] Ahmed A, Leroy G, Sachdeva A, Harber P, Rains SA, Youn S, Barai P. Trust in Generative AI for Health Information Consumption and the Effect of Learned Dependency: An Experimental Study. arXiv preprint arXiv:2606.20605. 2026 May 21.

[43] Abby Sen A, Joy JM, Jennex ME. Towards Navigating Ethical Challenges in AI-Driven Healthcare Ad Moderation. Computers. 2025 Sep 11;14(9):380.

[44] Townsend BA, Sihlahla I, Naidoo M, Naidoo S, Donnelly DL, Thaldar DW. Mapping the regulatory landscape of AI in healthcare in Africa. Frontiers in Pharmacology. 2023 Aug 24;14:1214422.

[45] Munung NS, Staunton C, Mazibuko O, Wall PJ, Wonkam A. Data protection legislation in Africa and pathways for enhancing compliance in big data health research. Health research policy and systems. 2024 Oct 15;22(1):145.

[46] Townsend BA. Governance-by-design as an enabler of AI in digital health in sub-Saharan Africa. Law, Technology and Humans. 2026 Apr;8(1):8-22.