

Development and Evaluation of HEAL A Bilingual English-Luganda RAG Chatbot for Disease Surveillance in Uganda

Mugume Twinamatsiko Atwine

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Timothy Mwanje Kintu

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Ibra Lujumba

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Ibrahim Mbabali

Makerere University

Khalifan Muwonge

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Lawrence Muwonge

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Lydia Nakiire

Infectious Diseases Institute

Vivian Ntono

Infectious Diseases Institute

Mike Nsubuga

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Ronald Galiwango

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Stella Lunkuse

Infectious Diseases Institute

Grace Kebirungi

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Daudi Jjingo

djjingo@idi.co.ug

African Centers of Excellence in Bioinformatics and Data Intensive Sciences

Keywords: Artificial intelligence, Disease Surveillance, Large Language Models, Uganda, Pandemic Preparedness, Linguistic Equity, IDSR, Retrieval-Augmented Generation

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10357473/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Frontline health workers in Uganda are often unable to use disease surveillance guidelines effectively: the Pdf documents are long, technical, and written in English, leaving Luganda-speaking workers without practical access. Off-the-shelf large language models generate plausible but ungrounded advice that can contradict national protocols. We developed HEAL, a bilingual chatbot using retrieval-augmented generation to deliver Uganda's Integrated Disease Surveillance and Response guidelines in English and Luganda, built on GPT-4-turbo with a fine-tuned MBart translation model trained on parallel English-Luganda sentence pairs. Six disease surveillance experts rated HEAL against three commercial models on field-generated questions across relevance, coverage, and coherence, while twenty frontline health workers gave qualitative feedback. HEAL scored higher than all baseline models in pooled analysis, with reviewers rating it highest most consistently. Grounding a large language model in national guidelines alongside a local-language translation layer produced higher expert ratings than commercial alternatives, an approach adaptable to other languages and guidelines.

Introduction

Background

Pandemics hit low- and middle-income countries (LMICs) hardest, in part because frontline health workers in these settings lack practical access to the surveillance guidelines they are expected to follow [1]. Uganda illustrates the problem clearly. The country sits within a tropical belt that produces recurring outbreaks of Ebola, Marburg, and other zoonotic diseases [2–5]. The 2022 Sudan ebolavirus outbreak resulted in 142 confirmed cases and 55 deaths [6], despite years of training investment by Uganda's Ministry of Health [7] — an outcome that points to persistent gaps in early detection and rapid response.

Like most African nations, Uganda follows the WHO's Integrated Disease Surveillance and Response (IDSR) framework [8]. The national IDSR guidelines compile standardized procedures for detecting, reporting, and responding to epidemic-prone diseases [9], but multiple barriers limit their frontline use [10]. The document spans hundreds of pages of dense, technical English. Health workers lack time to search through it during emergencies. Many community health workers and some district surveillance officers have limited English proficiency, particularly where training is infrequent. Critical information is scattered across sections, forcing users to synthesize material from multiple places under time pressure. High staff turnover and insufficient training resources mean that workers often rely on memory rather than the guidelines themselves.

Large language models (LLMs) can process complex documents and generate contextually relevant responses [11, 12], which makes them a plausible tool for closing the gap between comprehensive guidelines and actual frontline use. But most LLMs are optimized for high-resource languages like English, leaving out roughly 2 billion speakers of under-resourced languages [13, 48, 49]. In health

settings where miscommunication can be fatal, deploying an English-only tool to a Luganda-speaking workforce does not solve the access problem — it relocates it.

Prior Work

Retrieval-augmented generation (RAG) pairs a language model with an external knowledge source to improve factual grounding and reduce hallucination [14]. Recent advances include multi-task relevance sampling [15], chunking strategies that preserve semantic coherence [16], iterative refinement frameworks that generate sub-queries until sufficient evidence is gathered [17], and heterogeneous designs that decouple retrieval and generation representations [18, 23]. Evaluation suites such as RAGAS now provide reference-free metrics for retrieval quality [19].

Most of this work assumes English or other high-resource languages. Multilingual RAG approaches either translate queries or retrieve across languages, but both face coverage and consistency problems [20, 21]. Cross-lingual RAG improves performance by translating retrieved documents into a common language before generation [21], and prompt-guided retrieval augmentation handles non-knowledge-intensive tasks with minimal training overhead [22]. Transfer-learning methods such as parent-child training and back-translation have shown that modest seed data can yield substantial gains for low-resource neural machine translation [24–26], but few of these techniques have been applied to indigenous African languages in a health context.

In health applications specifically, LLMs have been used to automate literature reviews [27] and answer medical questions, where small fine-tuned models can approach larger baselines when augmented with domain-specific content [28, 29]. AI-driven decision-support systems for pandemics combine reinforcement learning with real-time monitoring, though adoption remains limited by infrastructure constraints [30]. Ethical analyses of these systems stress the need for transparency, bias mitigation, and readability [31, 32].

What remains missing is a system that integrates an indigenous African language directly into the generation pipeline and evaluates it with field-generated questions from working surveillance officers rather than static benchmarks.

Study Objectives

We developed HEAL (AI for Health Equity), a bilingual chatbot that uses GPT-4-turbo with RAG to make Uganda's IDSR guidelines conversationally accessible in English and Luganda. This study addressed three questions: (1) How does a RAG-grounded chatbot perform compared to commercial LLMs in supporting disease surveillance? (2) Is such a tool acceptable and feasible among frontline health workers in Uganda? (3) Is it feasible to integrate an indigenous Ugandan language into an LLM-based health tool for disease surveillance? To our knowledge, this is the first application of generative AI to

Ugandan IDSR guidelines and the first integration of a Ugandan indigenous language into an LLM-based health system.

METHODS

Data Preparation

The primary data source was Uganda's official National Technical Guidelines for Integrated Disease Surveillance and Response (IDSR), provided by the Ministry of Health [9]. The guidelines detail standardized procedures for surveillance, case definitions, laboratory confirmation, and response protocols for all epidemic-prone diseases monitored in Uganda. We converted the original PDF to JSON to preserve the document's hierarchical structure and enable downstream NLP processing.

Translation of IDSR Guidelines

We translated the full IDSR guidelines into Luganda, the most widely spoken indigenous language in central Uganda. Translation followed a hybrid pipeline: initial automated translation using the Google Cloud Platform translation API, followed by review from two bilingual language professionals who checked medical accuracy and cultural appropriateness. The translations were iteratively refined for domain-specific terminology and idiomatic expressions, with a final review by native Luganda speakers with healthcare backgrounds.

Parallel Corpus and Dataset Splits

The resulting parallel corpus contained 59,699 English–Luganda sentence pairs drawn from IDSR guidelines, Ministry of Health training materials, and selected public-domain health education texts. We partitioned the corpus at the document level into training (80%), validation (15%), and test (5%) splits before cleaning to prevent leakage of near-identical sentences across sets.

Cleaning removed 364 sentence pairs that exhibited quality issues: 358 from training (0.75%), 6 from validation (0.07%), and none from test. The final corpus contained 59,333 clean parallel sentence pairs (Table 1).

Data Split	Initial pairs	Pairs removed	Final pairs	Purpose
Training	47,759	358	47,401	Used for model parameter learning.
Validation	8,952	6	8,946	Used for hyperparameter tuning and early stopping.
Test	2,986	0	2,986	Held back for final, unbiased evaluation of the trained model.
Total	59,699	364	59,333	Overall size of the curated and partitioned dataset.

Table 1: Parallel corpus composition before and after quality filtering, with dataset split sizes and purposes.

Across splits, English sentences had a mean length of approximately 10 tokens (median 9; 90th percentile 14; maximum 187), while Luganda sentences averaged 8.2 tokens (median 7; 90th percentile 13; maximum 196). Most sentences were short operational phrases: roughly 66% of English training sentences contained 10 or fewer tokens and about 30% fell in the 11–20 token range, with a small tail exceeding 30 tokens. Validation and test sets followed a similar distribution (**Table 2**).

Split	≤ 10	11-20	21-30	31-50	> 50
train	31549	14422	1306	422	60
val	5915	2702	244	80	11
test	1972	901	83	27	3

Table 2: Distribution of sentence pairs by token count of the longer sentence in each pair.

Stratified Partitioning

To avoid biased evaluation and ensure that rare but high-stakes surveillance content appeared in all three splits, we re-partitioned the corpus using stratified random sampling rather than naive random line splitting. Each English sentence was tagged with three types of metadata: sentence length bin (≤ 10 , 11–20, 21–30, 31–50, > 50 tokens), disease mention (case-insensitive keyword matching against Uganda's IDSR priority disease list, including cholera, measles, malaria, acute flaccid paralysis, viral haemorrhagic fevers, MDR/XDR tuberculosis, Guinea worm, maternal and perinatal deaths, and acute viral hepatitis), and functional category (case definition, reporting, laboratory, or training, assigned by keyword groups and prioritized in that order). Strata were defined as the Cartesian product of these three dimensions, and sentence pairs were randomly assigned within each stratum while maintaining the 80/15/5 global proportions.

In the final test set, disease-related coverage was broad but sparse by design: tuberculosis appeared in approximately 2.1% of English test sentences, while cholera, measles, malaria, acute viral hepatitis, and maternal/perinatal death each appeared in 0.1–0.3%. Functionally, about 3.4% of test sentences were reporting-related, 1.2% case definitions, 1.3% laboratory-related, and 0.6% training-related. Full stratification details and keyword lists are provided in Multimedia Appendix 1.

Question-Answer Dataset

Separately from the translation corpus, the study team curated a dataset of question-answer pairs from the IDSR guidelines to support the RAG pipeline's knowledge base. The dataset covered all epidemic-prone diseases and represented realistic queries frontline health workers would encounter, ranging from basic case definitions and reporting requirements to multi-step response protocols and outbreak management scenarios. Each question-answer pair was validated by two team members for accuracy and relevance to practical surveillance activities.

Translation Model

We used the Sunbird/mbart-mul-en sequence-to-sequence model as our starting point[41]. This model extends mBART and includes Luganda among its pretraining languages, though its Luganda data come from general-domain texts with no health-sector specificity. Our work therefore represents domain adaptation of a Luganda-aware model to IDSR and public health content, not construction of a Luganda MT system from scratch.

We fine-tuned the model in both directions using the hyperparameters in (Table 3). Training was performed on a single NVIDIA A100 32GB GPU with early stopping (patience of 3 epochs) based on validation chrF. Optimization used AdamW with a learning rate of 1×10^{-6} and a batch size of 32, with a fixed random seed of 42 for reproducibility.

Table 3
Training hyperparameters for the bidirectional English–Luganda translation model.

Model	Learning rate	Epochs	Hardware
EN→LUG	1×10^6	5	1 × NVIDIA A100 32GB GPU
LUG→EN	1×10^6	20	1 × NVIDIA A100 32GB GPU

The EN→LUG model converged in 5 epochs; the LUG→EN model required 20. Post-training evaluation for catastrophic forgetting showed perplexity increases below 50% for both models, indicating adequate retention of previously learned translation capacity.

We evaluated both models on the held-out test set using BLEU, chrF++, COMET, and BERTScore,[43–46] following current recommendations for neural MT evaluation in low-resource settings..

Retrieval-Augmented Generation Architecture

For all experiments, HEAL’s retrieval-augmented generation (RAG) pipeline was implemented using the open-source Danswer backend configured as follows. We used the $\text{then} \frac{r}{e} - \text{small}$ sentence-transformer as a bi-encoder for both indexing and query encoding, producing 384-dimensional document embeddings with a maximum context window of 512 tokens per chunk. Source documents from the IDSR guidelines were split into overlapping segments of 512 encoder tokens with 5% token overlap (≈ 26 tokens), with optional 150-token “mini-chunks” enabled for finer-grained retrieval when indexing long sections. We used Danswer’s default which exposes a single logical index ($\text{danswer} \in \text{dex}$) backed by Vespa for both keyword and dense vector search. (Fig. 1)

Retrieval Augmented Generation Architecture

At query time, each question is embedded with the same bi-encoder and used to retrieve up to 50 candidate chunks via semantic search, with a cosine-similarity distance cutoff of 0.1 to filter clearly irrelevant results. The top 15 candidates are then re-ranked using an ensemble of cross-encoder reranking models, each applied with a maximum input length of 512 tokens. For answer generation, we pass retrieved chunks into the LLM until reaching a budget of approximately 2,560 encoder tokens (five 512-token chunks) for direct QA and 1,536 tokens for chat, ensuring a controlled and reproducible context length per query.

To evaluate the retrieval component, we constructed a small gold-standard set of IDSR questions with manually annotated relevant guideline sections and used Danswer’s built-in metrics handlers ($\text{RetrievalMetricsCont} \in \text{er}$, $\text{RerankMetricsCont} \in \text{er}$, $\text{LLMMetricsCont} \in \text{er}$) to log, for each query, the IDs, URLs, snippet text, similarity scores, and token usage of the top retrieved and re-ranked chunks. On this set we compute standard ranking-based retrieval metrics (e.g., recall at k and mean reciprocal rank) to quantify how often at least one gold-standard section appears within the top-k retrieved chunks, and we qualitatively examine cases where the model cites passages that are not among the highest-scoring retrieved segments. Citation and provenance correctness are assessed by checking, for each generated answer, whether the URLs and quoted guideline snippets presented to the user correspond to retrieved chunks that actually support the answer, and by flagging instances where the answer text or citations hallucinate unsupported content.

We employed OpenAI’s GPT-4-turbo via API for augmented generation due to its superior long-context handling capabilities and performance on domain-specific tasks.

Evaluation Framework

Study Design

We employed a blinded, comparative evaluation design to assess the performance of HEAL against three commercial large language models: GPT-4, GPT-3.5, and Gemini. The evaluation consisted of two components: (1) quantitative assessment of response quality by disease surveillance experts, and (2) qualitative evaluation of system usability and acceptability by frontline health workers.

Question Collection and LLM Response Generation

During a two-week pilot deployment, Ministry of Health field workers (surveillance officers and epidemiological officers) used the HEAL platform to generate questions relevant to their daily disease surveillance activities. Questions and their corresponding HEAL responses were stored and reviewed to remove duplicates and queries irrelevant to disease surveillance in Uganda. From this collection, 50 representative questions spanning a broad range of surveillance topics were selected for comparative evaluation.

These 50 questions were then posed to all four LLMs (HEAL, GPT-4, GPT-3.5, and Gemini) to generate responses. All responses were stored and edited to remove markdown formatting or other indicators that could be attributed to a specific LLM, ensuring evaluator blinding.

Reviewer Selection and Training

Six reviewers from Uganda's Ministry of Health and the Infectious Diseases Institute, all with more than three years of disease surveillance experience, independently evaluated the responses. Prior to evaluation, reviewers received training on how to assess LLM-generated responses using a simplified scoring rubric (Supplementary Materials). Reviewers received anonymized, randomly ordered responses for each question, maintaining complete blinding to model identity.

Reviewers were selected through a combination of purposive selection by role: the Ministry of Health and the Infectious Diseases Institute identified and provided reviewers based on their availability and relevant surveillance experience. All six reviewers who began the evaluation completed it, with one exception: one reviewer was unable to submit scores for one LLM's set of responses due to an unrelated field deployment for surveillance activity during the evaluation period; this missing data point was accommodated by the linear mixed model, which does not require complete cases across all conditions.

Evaluation Criteria

Based on a framework for evaluating LLM-generated health information proposed by Awasthi et al. (2025), we assessed responses across three criteria using 5-point Likert scales:

Relevance

Accuracy, logical reasoning, and helpfulness of the response in addressing the specific query. This criterion captures elements of bias, hallucinations, and toxicity that reduce the usefulness of the response.

Coverage

Holistic completeness of the response such that it addresses all key points and relays essential knowledge elements from the IDSR manual without significant missingness.

Coherence

Grammatical correctness and organizational quality of the generated responses.

Each reviewer scored all responses for each LLM against the information contained in Uganda's 2021 Third Edition of National Technical Guidelines for Integrated Disease Surveillance and Response (IDSR), which serves as the reference standard for training disease surveillance personnel in Uganda.

Evaluation Protocol

We employed a fully crossed design (Reviewer × Question × LLM) to minimize potential order effects and fatigue bias. Each reviewer evaluated all 50 questions for all 4 LLMs, yielding 1,200 total observations (6 reviewers × 4 LLMs × 50 questions). The presentation order of both LLM responses and questions was randomized separately for each reviewer to prevent systematic bias.

Each reviewer completed a structured feedback form for each LLM's set of responses, yielding 24 completed forms in total (6 reviewers × 4 LLMs).

Qualitative Evaluation

Qualitative evaluation was conducted in two phases to assess system usability, acceptability, and real-world utility:

Phase 1

Usability Testing: Recent graduate nurses and doctors (n = 10) were given access to HEAL for two weeks to test its utility in clinical scenarios arising in their daily work.

Phase 2

Field Testing: Surveillance officers (n = 10) used HEAL in real-world surveillance activities over two weeks.

Both groups underwent a detailed hour-long onboarding session introducing the HEAL project objectives and chatbot mechanics. Following onboarding, participants engaged with HEAL for one week, querying it on various priority diseases and conditions outlined in the IDSR guidelines. After this hands-on interaction period, two Focus Group Discussions (FGDs) were conducted with each group, each lasting at least 45 minutes. During these discussions, the research team gathered feedback on HEAL's benefits, challenges, and recommendations for improvement, with focus on ease of use, willingness to use, accuracy, and trustworthiness.

Statistical Analysis

For quantitative evaluation, the Friedman test assessed differences in scores across LLMs for each reviewer. Post-hoc pairwise comparisons used the Nemenyi test to identify specific significant differences between models. To explore inter-rater agreement, Kendall's W coefficient of concordance assessed agreement among reviewers, with effect sizes interpreted as small ($W < 0.1$), moderate ($0.1 \leq W < 0.3$), or large ($W \geq 0.3$). Krippendorff's alpha evaluated inter-rater reliability, with $\alpha \geq 0.67$ indicating acceptable reliability. All analyses were conducted using R version 4.3.0 with packages 'coin' (Friedman and Nemenyi tests), 'irr' (Kendall's W), and 'icr' (Krippendorff's alpha). Statistical significance was set at $\alpha = 0.05$.

No evaluation data were excluded from the primary analysis. As a robustness check, we repeated all analyses after removing statistical outliers; results were consistent regardless of their inclusion or exclusion. All statistical tests were two-sided.

The linear mixed model was fit in R using the lme4 package [57], with significance testing via lmerTest [58], estimated marginal means via emmeans [59], and Friedman-statistic plot annotations via ggpubr [60] and rstatix [61].

Generative AI use disclosure: This study evaluates large language models (GPT-4-turbo, GPT-3.5, GPT-4, Gemini) as its research subject, and GPT-4-turbo serves as HEAL's generation backbone; a fine-tuned MBart model performs translation. During preparation of this manuscript, the authors used Claude (Anthropic) to assist with editing and restructuring portions of the manuscript text, and used Consensus to identify relevant academic literature; all AI-suggested content, citations, and edits were reviewed and evaluated by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

RESULTS

Participants

There were 23 frontline health workers (doctors, nurses and surveillance officers) who participated across the two evaluation phases. Quantitative evaluation included six expert reviewers. Qualitative evaluation involved 10 recent graduates (Phase 1) and 10 experienced professionals (Phase 2). Participant characteristics are detailed in Supplementary Table 1.

Translation Model Performance

The fine-tuned MBart model was evaluated on the held-out test set of 2,986 sentence pairs across five standard metrics: BLEU, ChrF, chrF++, METEOR, and BERTScore. Results are reported separately for both translation directions in (Table 4). Across all metrics, the Luganda-to-English (LUG→EN) direction consistently outperformed English-to-Luganda (EN→LUG), with BLEU scores of 42.44 versus 29.80

respectively – a gap of 12.64 points. This asymmetry is consistent with established patterns in low-resource neural machine translation, where the higher-resource target language (English) benefits from stronger pre-training representations. Despite this gap, the EN→LUG direction achieved a BLEU score of 29.80 and BERTScore F1 of 0.5468, establishing a meaningful baseline for translation into a morphologically complex, low-resource language.

To contextualise these scores: a BLEU of 42.44 (LUG→EN) falls within the range considered high quality for MT systems and is consistent with practical usability, while a BLEU of 29.80 (EN→LUG) indicates that the clinical content of guidelines is conveyed accurately, though naturalness of Luganda phrasing may vary – a finding consistent with the translation quality concerns raised in our qualitative evaluation.

Table 4
Translation quality metrics for the bidirectional English–Luganda neural machine translation model evaluated on the held-out test set (n = 2,986 sentence pairs). Segment-level metrics are averaged across individual sentences with 95% confidence intervals derived from bootstrap resampling (1,000 iterations). Corpus-level metrics are computed as single aggregate scores; no standard method for CI derivation at the corpus level is established. – = CI not applicable.

Metric	Level	Eng-Lug	95% CI	Lug-En	95% CI
BLEU	Segment	29.80	[28.87,30.72]	42.44	[42.18, 44.37]
BLEU	Corpus	27.25	-	42.44	
ChrF	Corpus	59.37	-	61.42	
chrF++	Corpus	55.41	-	61.47	
METEOR	Segment	0.4459	[0.4356, 0.4563]	0.6370	[0.6273, 0.6467]
BERT Score F1	Segment	0.5468	[0.5375, 0.5560]	0.7604	[0.7529, 0.7680]

The LUG→EN direction achieved strong performance across all metrics (Table 4), with a segment-level BLEU of 42.44 and BERTScore F1 of 0.7604 – scores comparable to those reported for well-resourced language pairs. This outcome reflects the advantage of a high-resource target language during fine-tuning.

The EN→LUG direction was more challenging, as expected for a morphologically complex, low-resource target language. Segment-level BLEU of 29.80 and BERTScore F1 of 0.5468 establish a meaningful baseline, with the corpus-level BLEU of 27.25 providing the aggregate performance estimate.

Across all metrics, LUG→EN outperformed EN→LUG by 12.64 BLEU points at the segment level – a gap consistent with established patterns in low-resource NMT where the higher-resource target language benefits from stronger pre-training representations (Table 4).

Inter-rater Reliability

We examined the inter-reviewer agreement to establish the reliability of the scores within the criteria used to assess the responses. Krippendorff's alpha analysis revealed poor inter-rater reliability across all evaluation criteria, with all values either negative or near-zero (α range: -0.107 to 0.017), falling well below the acceptable threshold of $\alpha \geq 0.67$ (Krippendorff, 2004). We used Krippendorff's alpha (ordinal method) to estimate inter-reviewer reliability across all LLM responses per criterion, and separately for each LLM within each criterion. Negative alpha values indicate systematic disagreement among raters, suggesting inconsistent application of the evaluation rubric.

The low reliability observed suggests that the source of disagreement is not specific to any particular LLM's responses, but rather in how reviewers interpreted and applied the evaluation criteria. We tested whether the disagreement could stem from use of the Likert scale - that is, whether reviewers agreed on which LLM performed better or worse, but consistently gave different values on the Likert scale. For example, Reviewer A consistently scores above or below another reviewer across all LLMs, so they assign different absolute scores after reaching a conclusion about the rank of the LLMs. Pairwise Spearman correlation values among reviewers were low but HEAL was ranked the highest across all evaluation criteria. This suggests that reviewers could recognise the quality of the responses for a given LLM but had difficulty scoring ambiguous responses with generalised information. This variability likely reflects differences in reviewer expertise, subjective interpretation of evaluation criteria, and varying priorities regarding what constitutes optimal surveillance guidance. Although human evaluation is considered a gold-standard form of assessment, the approach suffers from high variance in rater reliability, which stems from ambiguity of the criteria used from the same scoring rubric [33–39]. We suggest that more rigorous training for human evaluators with concrete examples on how to score LLM-generated responses using any chosen metric is needed.

Overall Performance

Aggregate Performance Across Reviewers

HEAL achieved the highest aggregate median scores for three of six reviewers (Reviewers 1, 4, 5); GPT-4 ranked highest for Reviewer 2 and GPT-3.5 for Reviewer 6, with Reviewer 3 showing no distinction among models. Three reviewers (1, 5, 6) showed statistically significant differences across LLMs (Friedman test, $p < 0.05$), with HEAL significantly outperforming all three commercial models for Reviewers 1 and 5. Full per-reviewer statistics, effect sizes, and post-hoc comparisons are reported in Multimedia Appendix 1

Statistical testing confirmed significant differences in aggregate scores for three reviewers: Reviewer 1 showed highly significant differences ($\chi^2=49.1$, $p = 1.22 \times 10^{-10}$) with a moderate effect size (Kendall's $W = 0.328$); Reviewer 5 showed significant differences ($\chi^2=14.3$, $p = 0.003$) with a small effect size ($W = 0.095$); and Reviewer 6 showed significant differences ($\chi^2=10.8$, $p = 0.013$) with a small effect size ($W = 0.074$). Full Friedman test results are reported in Supplementary Table 1. Post-hoc pairwise analysis revealed that for Reviewers 1 and 5, HEAL significantly outperformed all three commercial LLMs ($p <$

0.001 for all comparisons). For Reviewer 6, significant differences emerged between HEAL and GPT-3.5 ($p = 0.04$) and between GPT-3.5 and Gemini ($p = 0.04$).

Pooled Analysis Using Linear Mixed Models

To account for reviewer discordance and provide a unified assessment of LLM performance, we evaluated the performance of different LLMs using linear models with fixed and random effects. We used a linear model to test whether LLM identity can predict the total score assigned to each question across criteria. We used GPT3 as the baseline because it was released earlier than the other models in our study. We observed that all LLMs were scored significantly higher than GPT3 in the order HEAL > Gemini > GPT4 > GPT3, and LLM identity explained 5.59% of the variance in the total score after correcting for model complexity.

After post hoc pairwise comparisons using Tukey's multiple comparisons of means, we observed that all LLMs were significantly better than GPT3 ($p < 0.05$), but GPT4 and Gemini were statistically indistinguishable ($p = 0.90$). We further tested whether these LLM differences remained significant after accounting for reviewer discordance by adding reviewers as a random intercept. We observed that the LLMs' rankings did not change and that reviewer identity accounted for ~ 17% of the variance in total scores.

We added two more controls to the linear mixed model: the question ID as a random effect to account for differences in question difficulty, and the length of the response as a fixed effect to test whether reviewers inadvertently rewarded the verbosity rather than the quality of the LLM response. We found that question difficulty explained additional variance (4.03%), and response length was not a significant predictor of total scores ($p = 0.434$), suggesting that verbosity bias was not a major driver of scores assigned by the reviewers. This null result is consistent with the broader literature on verbosity bias, where automated and human evaluators alike can conflate response length with quality [50, 51]; controlling for it here suggests HEAL's higher scores were not simply an artifact of producing longer responses.

HEAL received higher scores than the other three models across most reviewers (Fig. 2). Random effects were evenly spread among the reviewers, with Reviewer 2 scoring 2.7 points below the mean, while Reviewer 6 scored 2.18 points above the mean. The variance explained by reviewers (13–17%) was notably larger than that explained by question difficulty (4.03%), confirming that the identity of the reviewer mattered more than the question that was being evaluated. (Fig. 2)

Criterion-Level Performance

Analysis Across Individual Reviewers

We assessed performance across three evaluation criteria: Coverage, Relevance, and Coherence. Individual reviewer analyses revealed varying patterns of statistical significance:

Coverage

Only reviewers 1 and 6 identified significant differences in coverage scores. Reviewer 1 found HEAL significantly superior to all other models ($p < 0.001$ for all comparisons, $W = 0.24$). Reviewer 6 reported significant differences between HEAL and GPT3.5 ($p = 0.02$), GPT3.5 and GPT4 ($p < 0.001$), and GPT3.5 and Gemini ($p = 0.02$).

Relevance

Three reviewers (1, 5, and 6) detected significant differences, with reviewers 1 and 5 showing moderate effect sizes ($W = 0.34$ and 0.30 , respectively). HEAL significantly outperformed all commercial models for reviewers 1 and 5 ($p < 0.01$), while reviewer 6 found significant differences only between GPT4 and Gemini ($p = 0.088$).

Coherence

Four reviewers (1, 3, 5, and 6) identified significant differences in coherence. Notably, reviewer 1 demonstrated a large effect size ($W = 0.51$), with HEAL significantly surpassing all other models ($p < 0.01$). Reviewer 6 showed moderate effect size ($W = 0.18$) with significant differences between HEAL and GPT3.5 ($p = 0.012$) and GPT3.5 and Gemini ($p = 0.001$).

All these results are available in the supplementary information.

Pooled Criterion-Level Analysis

Further, we assessed the performance of LLMs using criterion-level scores in the linear mixed model framework. We observed that the ranking of the LLMs did not change, and HEAL outperformed baseline GPT3 by approximately 3 points. The variance explained by reviewers decreased to 13%, suggesting that reviewers agreed more on a criterion level.

To this model, we added Criterion identity as a fixed effect to investigate whether evaluation criteria differ in their average score levels, and whether LLM differences hold across all criteria simultaneously. We observed that across all LLMs, Coherence attracted the highest scores from reviewers, followed by Relevance and Coverage (Fig. 3). These results suggest that although all LLMs could produce well-constructed sentences and paragraphs, they generally struggled to provide sufficient contextual depth in response to the user's query. LLM effects and ranking were the same as the model with criterion identity as the fixed effects, and all criterion main effects were significant.

LLM × Criterion Interactions

Further, we included an interaction term between LLM and Criterion identity to evaluate whether any given LLM improves one criterion at the expense of another, while accounting for reviewer severity and question-level difficulty. We observed that the fixed effects for HEAL improved to 3.8 points above GPT3

while those for Gemini and GPT4 decreased slightly to 1.8 and 1.60 points, respectively. However, this did not change the ranking of the LLM performance.

Overall, the LLM x Criterion interaction ($F(6, 3585) = 2.949, p < .001$) was significant, suggesting that LLM performance/ranking was not consistent across evaluation criteria, with HEAL driving this effect. All LLM x Criterion interactions were non-significant except HEAL x Coverage interaction. These observations suggest that all models outperformed GPT3 consistently across all criteria except HEAL which underperformed when reviewers scored it for coverage. This difference could result from discordance among reviewers, with some scoring the breadth of the response and others the depth. The coverage mean score was the lowest across all criteria (Fig. 3).

Qualitative Findings

Overall, participants found the tool's rapid and accurate responses valuable, with experienced professionals successfully integrating it into daily surveillance, training, and reporting. This consistent performance helped build trust, even among initially skeptical surveillance officers. However, significant challenges emerged, particularly poor Luganda translation quality, mobile interface limitations, and critical barriers related to internet dependency and device availability for rural health teams. Key suggestions for enhancement focused on developing offline functionality, conversational memory, and voice input to improve access and utility.

Discussion

HEAL scored higher than GPT-3.5, GPT-4, and Gemini across relevance, coverage, and coherence when evaluated by six Ministry of Health surveillance experts on 50 field-generated questions. LLM identity explained 5.59% of score variance in the pooled linear mixed model after adjusting for reviewer effects, and three of six reviewers gave HEAL the highest median scores, with two showing highly significant differences ($p < 0.001$). Counter to expectation, HEAL's advantage was least consistent for coverage — the criterion RAG grounding should theoretically help most, since retrieval is meant to improve completeness against the source guideline. Only two of six reviewers found HEAL significantly ahead of all three commercial models on coverage, compared to broader and stronger effects on relevance and coherence. This may reflect reviewer disagreement about what "complete" coverage means for an open-ended surveillance question, rather than a genuine limitation of the RAG architecture — a possibility consistent with the poor inter-rater reliability reported below. The LUG→EN and EN→LUG translation models achieved BLEU scores of 42.44 and 29.80 respectively, sufficient for practical use but with clear room for improvement. Qualitative feedback from 20 frontline health workers confirmed that users found the system accurate and time-saving, but also flagged problems with Luganda spelling recognition and medical terminology coverage that would need to be resolved before deployment to workers with limited English.

The infrastructure barriers that emerged from qualitative evaluation — internet connectivity, device availability, and ongoing API costs — are not unique to HEAL. Any cloud-dependent health AI tool

deployed in rural Uganda will face the same constraints, and they should be treated as first-order design problems rather than afterthoughts.

These findings sit within a broader context of RAG-based health AI research. RAG has been shown to improve factual grounding and reduce hallucination in knowledge-intensive tasks [14], and recent work has extended these benefits through iterative refinement [17], heterogeneous retrieval-generation architectures [18], and cross-lingual retrieval [21]. However, nearly all of this work has been evaluated on English-language benchmarks or high-resource language pairs. HEAL differs in two respects. First, it grounds generation in a specific national guideline document rather than a general medical corpus, which means the system can provide Uganda-specific case definitions, reporting timelines, and response protocols that commercial LLMs cannot reliably produce. Second, it integrates Luganda — a language with limited digital resources and virtually no health-domain NLP tools — directly into the response pipeline through a fine-tuned MBart model. Our own earlier work used an intent-classification chatbot (built on the DIET architecture via the Rasa framework) to deliver bilingual English-Luganda pandemic communication during COVID-19 [47]; HEAL differs by grounding a generative large language model in a structured national guideline document through RAG, rather than classifying user intents against a fixed set of pre-written responses

The closest precedents are multilingual RAG systems that translate queries or retrieved passages before generation [21, 22], including region-specific bilingual health chatbots for infectious disease [53], Swahili medical triage [54], and low-resource retrieval models for Kinyarwanda [55], though none has been evaluated by domain experts using field-generated questions in the way this study does; a broader scoping review of health chatbots in Africa similarly identifies language accessibility as an unresolved gap [56]. Most health-focused LLM evaluations rely on static benchmarks such as medical licensing exams [12] or curated QA datasets [28]. Our use of questions drawn from actual surveillance officer queries and scored by working surveillance experts provides a more ecologically valid test of whether these systems can support real operational decisions.

The low inter-rater reliability we observed (Krippendorff's alpha below zero) is consistent with broader findings that human evaluation of LLM outputs is poorly standardized [33, 34]. This is not a solved problem in the field, and the absence of validated rubrics for scoring LLM-generated health guidance remains a barrier to rigorous comparative evaluation[52].

Several limitations qualify these findings. The inter-rater reliability problem is the most serious methodological limitation. With Krippendorff's alpha values below zero, the expert scores reflect substantial disagreement about what constitutes a good response. This could mean the rubric was insufficiently operationalized, that reviewers needed more calibration training, or that response quality is genuinely difficult to rate on a single scale. Future work should develop rubrics with concrete anchor examples and conduct calibration sessions before scoring begins.

The sample size constrains generalizability. Six reviewers scoring 50 questions across four LLMs yields 1,200 observations — enough to detect the effects we report, but not enough to stratify by disease

category, question complexity, or reviewer background. A larger validation study covering more of the IDSR guideline content is needed.

The Luganda translation quality, while a first for this language in a health AI context, is not yet reliable enough for deployment to users who cannot cross-check against the English output. The EN→LUG BLEU of 29.80 reflects both the difficulty of the translation direction and the limited parallel training data available. Improving this will require larger domain-specific parallel corpora, community-driven data collection, and collaboration with Luganda linguists to capture medical terminology and dialectal variation.

HEAL's dependence on cloud-hosted GPT-4-turbo raises data sovereignty and privacy concerns for sensitive health information. It also creates a recurring cost that may not be sustainable at scale. Locally hosted open-source alternatives should be explored for production deployment[42].

Finally, this study measured short-term performance and acceptability. We do not know whether HEAL changes surveillance quality, outbreak detection speed, or health worker competency over time. Implementation science research with longitudinal outcome measures would be needed to answer those questions.

The study also cannot claim generalizability beyond Uganda. Other LMICs have different linguistic landscapes, health system structures, and infrastructure conditions. The methodological approach – RAG grounding in national guidelines, bilingual translation, expert evaluation with field-generated questions – should transfer, but the specific results may not.

In summary, grounding an LLM in national disease surveillance guidelines and pairing it with a local-language translation layer produced higher expert ratings than three commercial alternatives across relevance, coverage, and coherence. Frontline health workers found the tool useful and accurate, though infrastructure gaps and translation quality remain barriers to deployment. The approach is replicable for other guideline sets and language pairs, but moving from proof-of-concept to operational use will require offline capability, locally hosted models, improved translation corpora, and validated evaluation rubrics. Longer-term studies should measure whether tools like HEAL actually improve surveillance outcomes, not just whether experts rate their responses well.

Abbreviations

AI

Artificial Intelligence

API

Application Programming Interface

BLEU

Bilingual Evaluation Understudy

EN

English
GPU
Graphics Processing Unit
HEAL
AI for Health Equity
IDSR
Integrated Disease Surveillance and Response
LLM
Large Language Model
LMICs
Low- and Middle-Income Countries
LUG
Luganda
MBart
Multilingual Bidirectional and Auto-Regressive Transformer
MT
Machine Translation
NLP
Natural Language Processing
RAG
Retrieval-Augmented Generation

Declarations

Acknowledgments

We gratefully acknowledge the Ministry of Health Uganda surveillance officers who contributed questions and feedback during HEAL's development. We thank the Infectious Diseases Institute reviewers for their careful evaluation. This work was supported by the Bill & Melinda Gates Foundation (Investment ID INV-062695); the funder had no involvement in the study design, data collection, analysis, interpretation of results, or the writing of the manuscript.

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

Conflicts Of Interest

None Declared.

Data Availability

Anonymized evaluation data supporting the findings of this study are available upon reasonable request and completion of a data use agreement [40].

Code Availability

Code for the HEAL chatbot is available upon reasonable request and completion of a data use agreement [40].

Ethical Considerations

This study received ethical approval from the Infectious Diseases Institute Institutional Review Board (Protocol IDI-REC-2023-63). All participants provided informed consent before participation. Data were anonymized and stored securely in compliance with Uganda's data protection regulations.

References

1. Jayatilleke K. Challenges in Implementing Surveillance Tools of High-Income Countries (HICs) in Low Middle Income Countries (LMICs). *Curr Treat Options Infect Dis* [Internet]. 2020 [cited 2025 Nov 12];12(3):191–201. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7453076/>
2. Nyakarahuka L, Shoemaker TR, Balinandi S, Chemos G, Kwesiga B, Mulei S, et al. Marburg virus disease outbreak in Kween District Uganda, 2017: Epidemiological and laboratory findings. Van Griensven J, editor. *PLoS Negl Trop Dis* [Internet]. 2019 Mar 18 [cited 2025 Nov 12];13(3):e0007257. Available from: <https://dx.plos.org/10.1371/journal.pntd.0007257>
3. Okware SI. Three ebola outbreaks in Uganda 2000-2011 [Internet] [Doctoral thesis]. The University of Bergen; 2015 [cited 2025 Nov 12]. Available from: <https://bora.uib.no/bora-xmlui/handle/1956/9239>
4. Sospeter SB, Udohchukwu OP, Ruaichi J, Nchasi G, Paul IK, Kanyike AM, et al. Ebola outbreak in DRC and Uganda; an East African public health concern. *Health Science Reports* [Internet]. 2023 Aug [cited 2025 Nov 12];6(8):e1448. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/hsr2.1448>
5. Bwire G, Sartorius B, Guerin P, Tegegne MA, Okware SI, Talisuna AO. Sudan Ebola virus (SUDV) outbreak in Uganda, 2022: lessons learnt and future priorities for sub-Saharan Africa. *BMC Med* [Internet]. 2023 Apr 13 [cited 2025 Nov 12];21(1):144. Available from: <https://bmcmmedicine.biomedcentral.com/articles/10.1186/s12916-023-02847-1>
6. Kabami Z, Ario AR, Harris JR, Ninsiima M, Ahirirwe SR, Ocerro JRA, et al. Ebola disease outbreak caused by the Sudan virus in Uganda, 2022: a descriptive epidemiological study. *The Lancet Global Health* [Internet]. 2024 Oct [cited 2025 Nov 12];12(10):e1684–92. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2214109X24002602>
7. Zalwango JF, Naiga HN, Nsubuga EJ, Akunzirwe R, Buhuguru R, Zalwango MG, et al. Understanding the delay in identifying Sudan Virus Disease: gaps in integrated disease surveillance and response and community-based surveillance to detect viral hemorrhagic fever outbreaks in Uganda,

- September 2022. BMC Infect Dis [Internet]. 2024 July 30 [cited 2025 Nov 12];24(1):754. Available from: <https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-024-09659-5>
8. World Health Organization. Integrated disease surveillance and response technical guidelines: Booklet Two: sections 1, 2, and 3, 3rd ed [Internet]. 2019 [cited 2025 Nov 12]. Available from: <https://www.who.int/publications/i/item/WHO-AF-WHE-CPI-01-2019>
 9. Ministry of Health Uganda. National Technical Guidelines for Integrated Disease Surveillance and Response [Internet]. Ministry of Health Uganda; 2021 [cited 2025 Nov 12]. Available from: https://www.afro.who.int/sites/default/files/2021-09/2_Uganda%203rd%20IDSR%20Tech%20Guideline_PrintVersion_10Sep2021.pdf
 10. Wolfe CM, Hamblion EL, Dzotsi EK, Mboussou F, Eckerle I, Flahault A, et al. Systematic review of Integrated Disease Surveillance and Response (IDSR) implementation in the African region. Hodges MH, editor. PLoS ONE [Internet]. 2021 Feb 25 [cited 2025 Nov 12];16(2):e0245457. Available from: <https://dx.plos.org/10.1371/journal.pone.0245457>
 11. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med [Internet]. 2024 Feb [cited 2025 Nov 12];177(2):210–20. Available from: <https://www.acpjournals.org/doi/10.7326/M23-2772>
 12. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature [Internet]. 2023 Aug 3 [cited 2025 Nov 12];620(7972):172–80. Available from: <https://www.nature.com/articles/s41586-023-06291-2>
 13. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns [Internet]. 2025 Jan [cited 2025 Nov 12];6(1):101118. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2666389924002903>
 14. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.
 15. Hofstätter S, Chen J, Raman K, Zamani H. Multi-task retrieval-augmented text generation with relevance sampling. arXiv preprint arXiv:2207.03030. 2022.
 16. Merola C, Singh J. Reconstructing context: evaluating advanced chunking strategies for retrieval-augmented generation. In: Wang Z, et al., editors. Knowledge-Enhanced Information Retrieval. KEIR 2025. Lecture Notes in Computer Science, vol 16086. Cham: Springer; 2026. p. 1-13.
 17. Aghajani Asl M, Asgari-Bidhendi M, Minaei-Bidgoli B. FAIR-RAG: faithful adaptive iterative refinement for retrieval-augmented generation. arXiv preprint arXiv:2510.22344. 2025.
 18. Yang P, Li X, Hu Z, Wang J, Yin J, Wang H, He L, Yang S, Wang S, Huang Y, Qi T. HeterRAG: a heterogeneous retrieval-augmented generation framework with decoupled knowledge representations. arXiv preprint arXiv:2504.10529. 2025.
 19. Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: Aletras N, De Clercq O, editors. Proceedings of the 18th Conference of the European

- Chapter of the Association for Computational Linguistics: System Demonstrations; 2024 Mar; St. Julians, Malta. Association for Computational Linguistics; 2024. p. 150-8.
20. Joty S, Nakov P, Màrquez L, Jaradat I. Cross-language learning with adversarial neural networks. In: Levy R, Specia L, editors. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017); 2017 Aug; Vancouver, Canada. Association for Computational Linguistics; 2017. p. 226-37.
 21. Ranaldi L, Haddow B, Birch A. Multilingual retrieval-augmented generation for knowledge-intensive question answering task. In: Findings of the Association for Computational Linguistics: EACL 2026; 2026 Mar 24-29; Rabat, Morocco. Association for Computational Linguistics; 2026. p. 697-716.
 22. Guo Z, Cheng S, Wang Y, Li P, Liu Y. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023; 2023 Jul; Toronto, Canada. Association for Computational Linguistics; 2023. p. 10896-912.
 23. Chen J, Zhang R, Guo J, de Rijke M, Liu Y, Fan Y, Cheng X. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23); 2023 Jul 23-27; Taipei, Taiwan. New York: ACM; 2023. p. 1448-57.
 24. Kocmi T, Bojar O. Trivial transfer learning for low-resource neural machine translation. In: Bojar O, Chatterjee R, Federmann C, Fishel M, Graham Y, Haddow B, et al., editors. Proceedings of the Third Conference on Machine Translation: Research Papers; 2018 Oct; Brussels, Belgium. Association for Computational Linguistics; 2018. p. 244-52.
 25. Niu X, Denkowski M, Carpuat M. Bi-directional neural machine translation with synthetic parallel data. In: Birch A, Finch A, Luong T, Neubig G, Oda Y, editors. Proceedings of the 2nd Workshop on Neural Machine Translation and Generation; 2018 Jul; Melbourne, Australia. Association for Computational Linguistics; 2018. p. 84-91.
 26. Tars S, Fishel M. Multi-domain neural machine translation. In: Pérez-Ortiz JA, Sánchez-Martínez F, Esplà-Gomis M, Popović M, Rico C, Martins A, et al., editors. Proceedings of the 21st Annual Conference of the European Association for Machine Translation; 2018 May; Alicante, Spain. 2018. p. 279-88.
 27. Ali NF, Mohtasim MM, Mosharrof S, Krishna TG. Automated literature review using NLP techniques and LLM-based retrieval-augmented generation. In: 2024 International Conference on Innovations in Science, Engineering and Technology (ICISSET); 2024 Oct 26-27; Chittagong, Bangladesh. IEEE; 2024.
 28. Khlaut J, Dancette C, Ferreres E, Alaedine B, Herent H, Manceron P. Efficient medical question answering with knowledge-augmented question generation. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Bitterman D, editors. Proceedings of the 6th Clinical Natural Language Processing Workshop; 2024 Jun; Mexico City, Mexico. Association for Computational Linguistics; 2024. p. 10-20.

29. Li Z, Fan S, Gu Y, Li X, Duan Z, Dong B, Liu N, Wang J. FlexKBQA: a flexible LLM-powered framework for few-shot knowledge base question answering. *Proc AAAI Conf Artif Intell.* 2024;38(17):18608-16.
30. Periañez A, Fernández Del Río A, Nazarov I, Jané E, Hassan M, Rastogi A, Tang D. The digital transformation in health: how AI can improve the performance of health systems. *Health Syst Reform.* 2024;10(2):2387138.
31. Lin Z. Beyond principlism: practical strategies for ethical AI use in research practices. *AI Ethics.* 2025;5(3):2719-31.
32. Qiu W, Huang T, Rullo R, Kuang Y, Maatouk A, Ramos SR, Ying R. RephQA: evaluating readability of large language models in public health question answering. *arXiv preprint arXiv:2509.16360.* 2025.
33. Awasthi R, Bhattad A, Ramachandran SP, Mishra S, Khanna AK, Cywinski JB, et al. Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization. *Npj Health Systems.* 2025;2(1):40. Available from: <https://doi.org/10.1038/s44401-025-00043-2>
34. Bojić L, Zagovora O, Zelenkauskaitė A, Vuković V, Čabarkapa M, Veseljević Jerković S, et al. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports.* 2025;15(1):11477. Available from: <https://doi.org/10.1038/s41598-025-96508-3>
35. Deutsch D, Dror R, Roth D. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics.* 2021;9:1132–1146. Available from: https://doi.org/10.1162/tacl_a_00417
36. Fabbri AR, Kryściński W, McCann B, Xiong C, Socher R, Radev D. SummEval: re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics.* 2021;9:391–409. Available from: https://doi.org/10.1162/tacl_a_00373
37. Howcroft DM, Belz A, Clinciu MA, Gkatzia D, Hasan SA, Mahamood S, et al. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In: Davis B, Graham Y, Kelleher J, Sripada Y, editors. *Proceedings of the 13th International Conference on Natural Language Generation.* Association for Computational Linguistics; 2020. p. 169–182. Available from: <https://doi.org/10.18653/v1/2020.inlg-1.23>
38. Krippendorff K. *Content Analysis: An Introduction to Its Methodology.* SAGE; 2004.
39. Wang Y, Huang J, Du L, Guo Y, Liu Y, Wang R. Evaluating large language models as raters in large-scale writing assessments: a psychometric framework for reliability and validity. *Computers and Education: Artificial Intelligence.* 2025;9:100481. Available from: <https://doi.org/10.1016/j.caeai.2025.100481>
40. ACE Uganda. HEAL project [Internet]. GitHub. Available from: https://github.com/aceuganda/HEAL_project
41. Akera B, Mukiibi J, Naggayi LS, Babirye C, Owomugisha I, Nsumba S, Nakatumba-Nabende J, Bainomugisha E, Mwebaze E, Quinn J. Machine translation for African languages: community

- creation of datasets and models in Uganda. In: Proceedings of the 3rd Workshop on African Natural Language Processing; 2022
42. Akera B, Ouma EN, Yiga G, Walukagga P, Natukunda P, Saaka T, Nsumba S, Nabukeera LT, Muhanguzi J, Sekalala I, Namara NJ, Bainomugisha E, Mwebaze E, Quinn J. Sunflower: a new approach to expanding coverage of African languages in large language models. arXiv preprint arXiv:2510.07203. 2025
 43. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul; Philadelphia, PA. Association for Computational Linguistics; 2002. p. 311–8.
 44. Popović M. chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation; 2015 Sep; Lisbon, Portugal. Association for Computational Linguistics; 2015. p. 392–5.
 45. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; 2005 Jun; Ann Arbor, MI. Association for Computational Linguistics; 2005. p. 65–72.
 46. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: 8th International Conference on Learning Representations; 2020 Apr 26–30; Addis Ababa, Ethiopia.
 47. Atwine MT, Jjingo D, Nsubuga M, Serunjogi R, Mbabaali I, Lujumba I, David B, Kintu T, Galiwango R, Grace K. A bi-lingual chatbot implementation for pandemic response using the transformer-based approach. PLOS Digit Health. 2026 Apr 1;5(4):e0001256.
 48. Hussen KY, Sewunetie WT, Ayele AA, Imam SH, Alemu EN, Muhammad SH, Yimam SM. The state of large language models for African languages: progress and challenges. arXiv preprint arXiv:2506.02280. 2025.
 49. Alabi JO, Hedderich MA, Adelani DI, Klakow D. Charting the landscape of African NLP: mapping progress and shaping the road ahead. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025 Nov; Suzhou, China. Association for Computational Linguistics. p. 27807–41.
 50. Saito K, Wachi A, Wataoka K, Akimoto Y. Verbosity bias in preference labeling by large language models. arXiv preprint arXiv:2310.10076. 2023.
 51. Dubois Y, Galambosi B, Liang P, Hashimoto TB. Length-controlled AlpacaEval: a simple way to debias automatic evaluators. arXiv preprint arXiv:2404.04475. 202
 52. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, Osterhoudt H, Wu X, Visweswaran S, Fu S, Mathur P, Cacciamani GE, Sun C, Peng Y, Wang Y. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024 Sep 28;7(1):258.

53. Dairo O, et al. Development of a bilingual diagnostic chatbot (English and Yoruba) for infectious disease symptom checks using large language models with retrieval augmented generation. In: 2024 Proceedings of the International Conference on Artificial Intelligence and Robotics (MIRG-ICAIR); 2024.
54. Kabeya DK, et al. Designing a Swahili-speaking medical chatbot for oncology, dermatology, and otorhinolaryngology care in low-resource settings. In: 2025 Conference on Information Communications Technology and Society (ICTAS); 2025.
55. Nzeyimana A. KinyaColBERT: a lexically grounded retrieval model for low-resource retrieval-augmented generation. arXiv preprint. 2025.
56. Phiri M, Munoriyarwa A. Health chatbots in Africa: scoping review. J Med Internet Res. 2023;25:e35573.
57. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015;67(1):1-48.
58. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. Journal of Statistical Software. 2017;82(13):1-26.
59. Lenth R, Piaskowski J. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 2.0.2. 2026.
60. Kassambara A. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.3. 2026.
61. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.3. 2025.

Figures

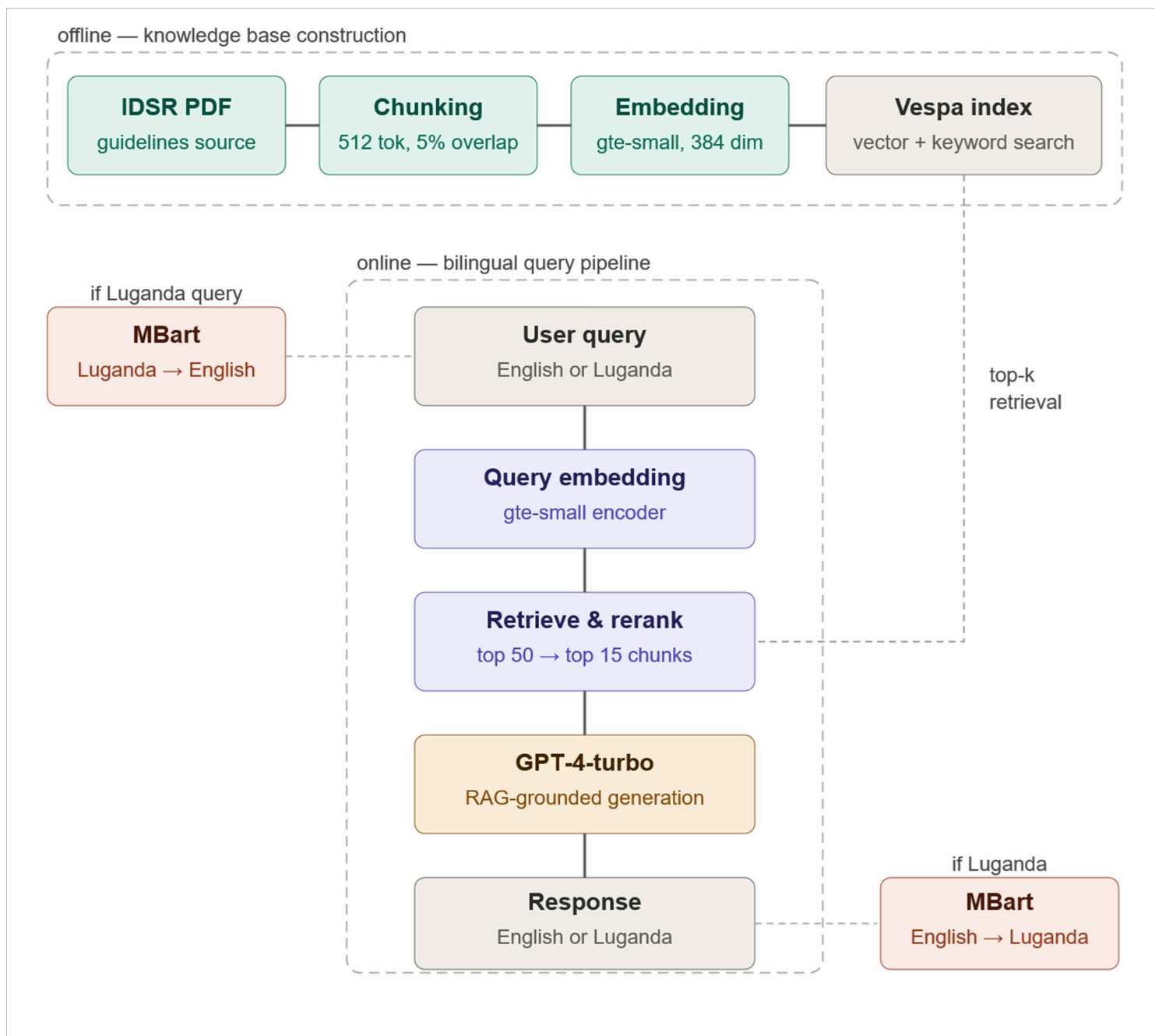


Figure 1

HEAL system architecture. The offline pipeline (top) processes Uganda's IDSR guidelines PDF into a Vespa vector index via chunking and gte-small embedding. At query time (online pipeline, centre), user queries are embedded using the same encoder and matched against stored chunks via semantic search; the top 50 candidates are re-ranked to 15 and passed as context to GPT-4-turbo for grounded response generation. Luganda queries and responses are handled via MBart bidirectional translation (dashed paths), enabling full bilingual access. Solid lines indicate the primary English pipeline; dashed lines indicate the conditional Luganda pathway.

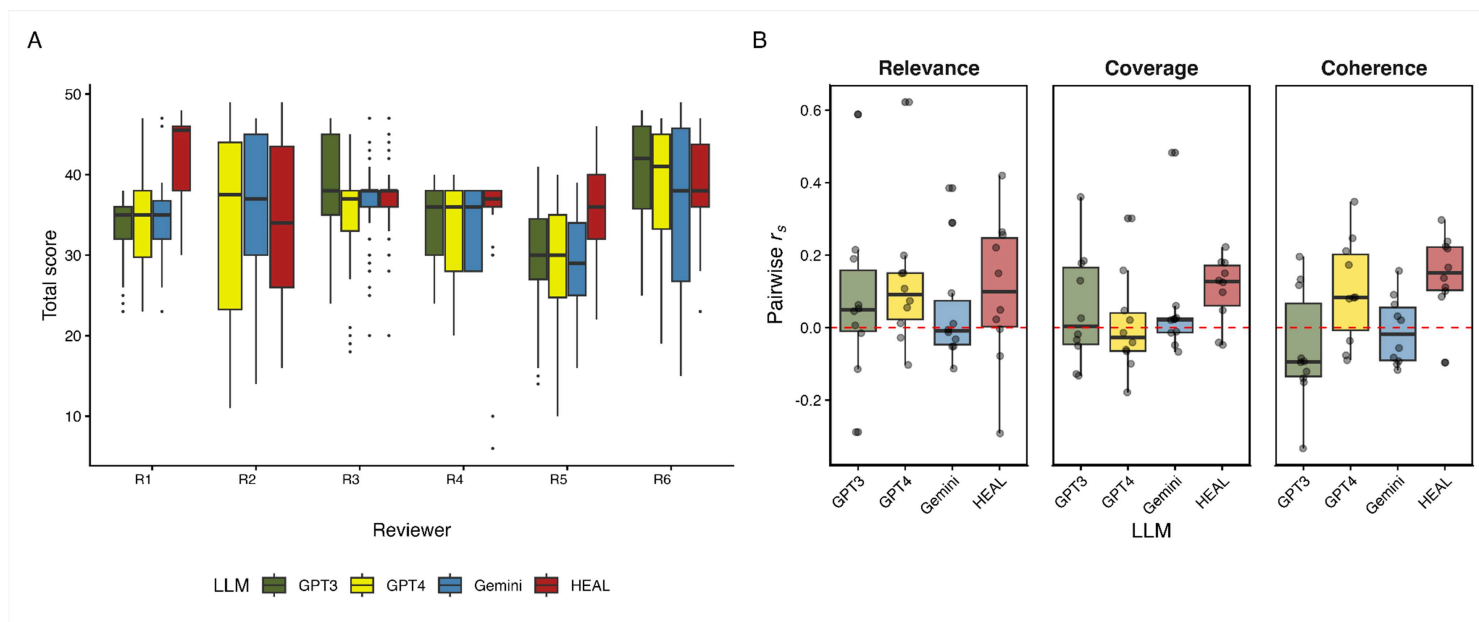


Figure 2

LLM performance across reviewers and evaluation criteria. (A) Distribution of aggregate total scores for each LLM across six expert reviewers (R1–R6). Dashed boxes indicate reviewers with no statistically significant differences between models. (B) Pairwise Spearman correlation coefficients (r_s) between reviewer rankings for GPT-3, GPT-4, Gemini, and HEAL across Relevance, Coverage, and Coherence criteria. The red dashed line indicates $r_s = 0$.

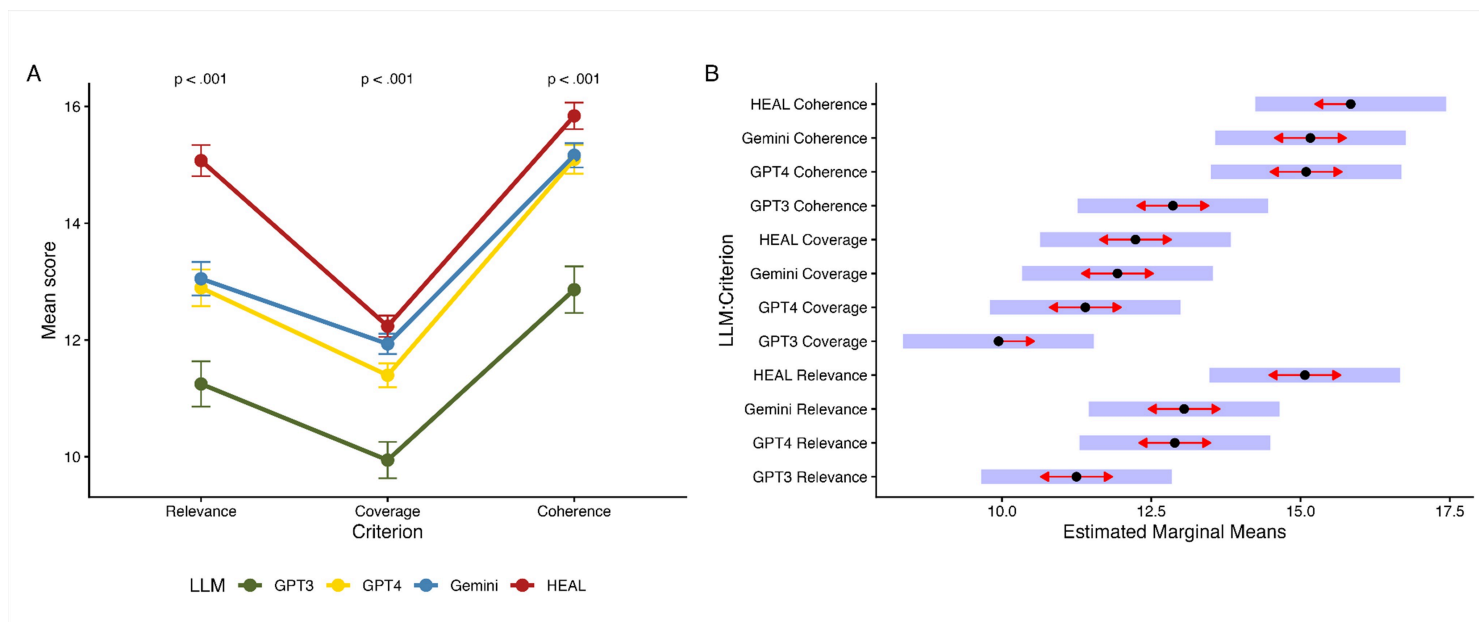


Figure 3

Criterion-level LLM performance from the linear mixed model. (A) Mean scores ($\pm$ SE) for GPT-3, GPT-4, Gemini, and HEAL across Relevance, Coverage, and Coherence criteria. Significant LLM $\times$ Criterion

interactions are indicated ($p < 0.001$). (B) Estimated marginal means with 95% confidence intervals for each LLM–Criterion combination; red arrows indicate significant pairwise differences.