Lueji: A Swahili-Language Medical Chatbot for Low-Resource Specialties in Sub-Saharan Africa

DIVIN KAYEYE KABEYA[1], WITESYAVWIRWA VIANNEY KAMBALE[3], KUTABUNA KUBAKISA CHRISTIAN[1], ISAAC LUKUSA KAYEMBE[1], MAHMOUD HAMED[2], KYANDOGHERE KYAMAKYA[1,2]

[1]Faculté Polytechnique, Université de Kinshasa (UNIKIN),
Kinshasa,
DEMOCRATIC REPUBLIC OF THE CONGO

[2]Institute for Smart Systems Technologies,
Universität Klagenfurt,
AUSTRIA

[3]Faculty of Information and Communication Technology,
Tshwane University of Technology,
Pretoria, SOUTH AFRICA

Abstract: Poor digital infrastructure and a lack of local language resources result in Sub-Saharan Africa having limited access to specialized medical information. Swahili, spoken by over 100 million people, is crucial for the democratization of digital healthcare. The lack of an extensive Swahili medical corpus complicates the development of credible and culturally tailored language models. Large Language Models (LLMs) promise to be a boon for medical chatbots, providing accessible, context-aware, and language-appropriate health information. However, fine-tuning techniques have constraints, including diminished factual robustness and a larger risk of medical hallucinations. The Swahili-speaking medical chatbot Lueji, initially based on a fine-tuned model, is extended with a Retrieval-Augmented Generation (RAG) architecture to address these limitations. A FAISS-based semantic retriever utilizing a Swahili-translated Huatuo-26M Chinese medical corpus and the UlizaLlama model, tailored for African languages, is employed in the proposed system. This hybrid methodology improves factual dependability and contextual accuracy while retaining generative fine-tuning-achieved verbal fluency. The BLEU, ROUGE, and GLEU metrics were used to quantify lexical consistency, structural similarity, and text quality. The fine-tuned model outperforms the RAG-based version on typical text similarity measures (BLEU-1 = 0.2217, ROUGE-1 = 0.3168, GLEU = 0.1077 vs. 0.1912, 0.2916, 0.0918). A qualitative investigation reveals that the RAG architecture significantly reduces hallucinations and enhances clinical factuality. Low-resource environments present a fundamental trade-off between linguistic fluency and factual accuracy. The study gives fresh empirical insights into the balance between fine-tuning and retrieval augmentation for low-resource medical LLMs. It lays the groundwork for reliable, hybrid, and culturally inclusive African medical chatbots.

Key-Words: Swahili Chatbot, Retrieval-Augmented Generation (RAG), Fine-tuning, Low-Resource Languages, Medical Artificial Intelligence, UlizaLlama, Generative Language Models.

Received: April 19, 2025. Revised: July 7, 2025. Accepted: August 14, 2025. Published: February 24, 2026.

1 Introduction

1.1 Background and Motivation

In many regions of Sub-Saharan Africa, chronic shortages of medical specialists, particularly in fields such as oncology, dermatology, and otorhinolaryngology, persistently hinder timely diagnosis and adequate care. These specialties, though critical, remain structurally underserved. Africa has fewer than one dermatologist per million people. In ENT, the average clinician density is just 0.18 per 100,000, equivalent to one specialist for every 556,000 individuals, compared to 5.7 per 100,000 in Europe, [1]. In oncology, the burden of non-communicable diseases, such as cancer, is steadily rising [2]. On average, the annual workload for an oncologist in Africa amounts to 325 consults, which is nearly double the median of 175 consults observed among their colleagues globally[3]. Meanwhile, the continent is facing a steep rise in non-communicable diseases, with projections forecasting over 1.27 million new cancer cases annually in Africa by 2030, [4].

Artificial Intelligence (AI), and more specifically Large Language Models (LLMs), offer a promising pathway to bridge this gap, [5]. Their ability to process, generate, and contextualize medical information at scale opens up new opportunities for

deploying conversational agents that can assist with symptom triage, deliver health education, and provide preliminary guidance to underserved populations, [6]. Yet, for these systems to be effective and inclusive, they must operate in languages that patients understand, [7], [8]. Swahili, spoken by over 100 million people across East and Central Africa, stands as a key linguistic gateway. Despite its widespread use, it remains vastly underrepresented in biomedical datasets, making it difficult to develop robust clinical models that align with local linguistic and cultural realities.

In this landscape, building a Swahili-language medical chatbot is both a technical challenge and a public health imperative. The Lueji system was originally developed through supervised fine-tuning, showing strong performance in generating fluent and clinically relevant responses, [9]. However, like many static models, it faced limitations in factual currency and traceability. To address these issues, this study proposes a new architecture based on Retrieval-Augmented Generation (RAG), which integrates a dedicated Swahili medical corpus as a dynamic knowledge source. This hybrid configuration aims to enhance factual accuracy, bolster clinical credibility, and maintain high linguistic fluency while laying the groundwork for the broader deployment of domain-specific AI tools in African healthcare systems.

1.2 Problem Statement and Research Questions

Despite the significant progress achieved with fine-tuned Swahili-language medical models, two major challenges persist in limiting their clinical applicability. First, the lack of high-quality Swahili medical corpora hinders the development of comprehensive and diverse knowledge bases, leading to incomplete or inconsistent responses. Second, the reliance on static, internalized model parameters leads to factual drift and an inability to access up-to-date medical knowledge, [9]. These issues often manifest as imprecise or erroneous outputs, undermining both user trust and patient safety in medical contexts.

To address these challenges, this research investigates the integration of a Retrieval-Augmented Generation (RAG) architecture within a low-resource linguistic setting, in order to evaluate its capacity to enhance factual grounding while preserving linguistic fluency. The following research questions guide this study:

- RQ1: To what extent does a RAG architecture alter the balance between linguistic quality and clinical reliability compared to a purely fine-tuned model?
- RQ2: What measurable trade-offs emerge between textual fluency and factual accuracy in a Swahili-language medical LLM?
- RQ3: Which retrieval and context-filtering strategies maximize the relevance of extracted information from a translated medical corpus while minimizing hallucinations?
- RQ4: To what extent can this hybrid framework be scalable, resource-efficient, and transferable to other African languages and specialized medical domains?

These questions position the study as a comparative analysis between the fine-tuning and retrieval-augmentation paradigms, aiming to identify their respective strengths, weaknesses, and complementarities in the design of African medical chatbots.

1.3 Contributions of This Paper

This work extends and deepens prior research on Lueji by introducing and critically evaluating a Retrieval-Augmented Generation (RAG) approach for Swahili-language medical dialogue systems. Its main contributions are as follows:

1. Development of a complete and reproducible Swahili RAG architecture. The paper presents an implementation that integrates the UlizaLlama generative model with a FAISS-based semantic retriever indexing a Swahili-translated medical corpus originally derived from the Chinese Huatuo-26M dataset. The whole pipeline, including data segmentation, normalization, and validation, is detailed to ensure reproducibility in low-resource environments.
2. Comparative evaluation of fine-tuning and RAG using quantitative and qualitative criteria. The study systematically compares both approaches using automatic metrics (BLEU, ROUGE, GLEU) and qualitative assessments focused on clinical factuality, hallucination reduction, and linguistic consistency.
3. Analysis of trade-offs between linguistic performance and factual reliability. The results reveal that while the fine-tuned model maintains stronger lexical and structural coherence, the RAG architecture substantially improves factual accuracy and minimizes clinical hallucinations. The paper offers a mechanistic explanation of these differences and proposes potential optimization strategies,

including reranking, similarity thresholding, and contextual segmentation.

4. Empirical contribution to multilingual medical AI research. By analyzing the behavior of a RAG-based system in a low-resource linguistic context, this work provides concrete insights for the development of hybrid, culturally inclusive, and clinically reliable medical chatbots, contributing to the advancement of digital health innovation across Africa.

The remainder of this paper is organized as follows: Section II reviews recent work on medical chatbots, retrieval-augmented generation architectures, and African NLP resources. Section III describes the construction of the Swahili medical corpus, including translation, thematic filtering, and validation procedures. Section IV presents the design of our RAG-based chatbot system, detailing each module of the architecture. Section V outlines the metrics used to assess performance and results of the system. Finally, Section VII concludes the study and highlights future research directions.

2 Related Works

Recent years have seen a surge in the development of large language models for healthcare applications, including clinical question answering, medical chatbots, and biomedical information retrieval. However, most of these advances have been limited to high-resource languages such as English and Chinese, with little attention given to African languages or low-resource settings. This review highlights the novelty of our contribution in bridging biomedical knowledge and conversational AI in Swahili.

2.1 Medical Chatbots and Clinical QA Systems

The proliferation of large language models (LLMs) has driven a new wave of medical conversational systems designed to enhance patient interaction, provide diagnostic support, and promote health education. These systems employ various strategies, including supervised fine-tuning, retrieval-augmented generation (RAG), reinforcement learning, and parameter-efficient tuning, to enhance factual accuracy and contextual relevance. Notable examples include Google's Med-PaLM 2, which integrates retrieval-based reasoning and multitask training to approach expert-level performance on clinical benchmarks, [10]. ChatDoctor similarly combines domain-specific fine-tuning on doctor–patient dialogues with a retrieval mechanism for factual grounding, improving the model's ability to generate context-aware and trustworthy responses, [11].

In parallel, open-source initiatives like PMC-LLaMA and DoctorGLM have demonstrated the potential of lightweight and scalable medical LLMs, [12], [13]. These models are fine-tuned on biomedical literature or structured Q&A corpora and adapted using efficient methods such as LoRA or quantization, enabling deployment in resource-constrained settings without sacrificing performance. Systems such as CancerBot and MQ-RAG extend this paradigm to domain-specific applications in oncology and dermatology, [14], [15]. By incorporating dense retrieval, contextual reranking, and targeted prompting, these systems dXiong2023DoctorGLMeliver grounded responses tailored to complex clinical scenarios.

Hybrid architectures further showcase the versatility of combining modular components. For instance, dual-pipeline systems pair retrieval engines with compact generative modules, such as NanoGPT, to support differential diagnosis or decision support, [16]. Crucially, a growing body of research focuses on language and cultural adaptation. Efforts targeting Amharic, Mandarin, or other underrepresented languages have shown that medically capable LLMs can be developed through a blend of real consultations, synthetic data, and feedback-driven refinement. Models like HuatuoGPT demonstrate how alignment with local clinical norms enhances trust and usability, [17].

2.2 Retrieval-Augmented Generation for Low-Resource Languages

Retrieval-Augmented Generation (RAG) approaches have recently demonstrated their effectiveness in low-resource language processing, circumventing the limitations posed by insufficient supervised training data and restricted semantic coverage. Systems developed for Persian, Arabic, Punjabi, Singlish, and Bangla have shown that integrating a semantic retrieval engine often optimized via multilingual embeddings or hybrid indices allows generated responses to be grounded in reliable documents while enhancing their factual accuracy, [18], [19], [20], [21].

These architectures often rely on translated corpora, language-specific models, and cross-lingual retrieval or contextual reranking strategies. In [22], the authors propose a language model tailored for Amharic question-answering tasks, combining Retrieval-Augmented Generation to integrate knowledge from diverse documents with fine-tuning using the LoRA method on a large Amharic dataset.

These studies highlight the structuring role of RAG as a viable alternative to fine-tuning

in low-resource linguistic and medical contexts. They provide a methodological foundation for applications in languages such as Swahili, where combining a structured corpus, a multilingual dense encoder, and a specialized generative engine can ensure high-quality conversational service, while guaranteeing adaptability, traceability, and clinical relevance of responses.

2.3 Swahili and African NLP Resources

In recent years, there has been a marked increase in NLP research targeting African languages, with the development of pretrained language models such as SwahBERT, AfriBERTa, and AfroLM. These models, whether monolingual or multilingual, have demonstrated that strong performance on tasks such as classification, NER, or QA can be achieved even in the absence of large corpora, [23], [24], [25]. Their effectiveness relies on strategies such as language-specific fine-tuning, pruning of irrelevant vocabulary, and active learning on limited corpora, enabling efficient adaptation to under-resourced local linguistic contexts.

Concurrently, significant efforts have been made to create annotated datasets in African languages. Resources such as KenSwQuAD, [26], MasakhaNER, [27], and AfriQA, [28], have established a foundation for more rigorous evaluation of systems in these languages. These datasets are often augmented through machine translation or cross-lingual transfer techniques, addressing the scarcity of native content in technical or specialized domains.

However, a critical gap remains regarding specialized corpora, particularly in the biomedical domain. The vast majority of prior work focuses on general language, effectively excluding applications such as diagnostic support or health education. Our contribution addresses this gap by creating the first structured biomedical corpus in Swahili, integrated into a RAG-based system for medical question answering. This resource represents a significant step toward more targeted and contextually relevant applications for African settings, facilitating access to clinical information in a local lingua franca

3 Dataset Construction

In low-resource linguistic environments, such as Swahili-speaking medical contexts, the scarcity of specialized datasets presents a significant challenge to the development of advanced natural language processing systems. To address this challenge, we constructed a Swahili medical corpus from the Chinese resource Huatuo-26M using a systematic pipeline involving translation, cleaning, validation, and structuring. The overall workflow is illustrated in Figure 1, which details the sequential stages: thematic filtering, automatic translation using GPT-4o-mini, followed by a three-step post-processing phase comprising cleaning, normalization, and partial human validation. This corpus constitutes the first large-scale biomedical dialogic database in the Swahili language.

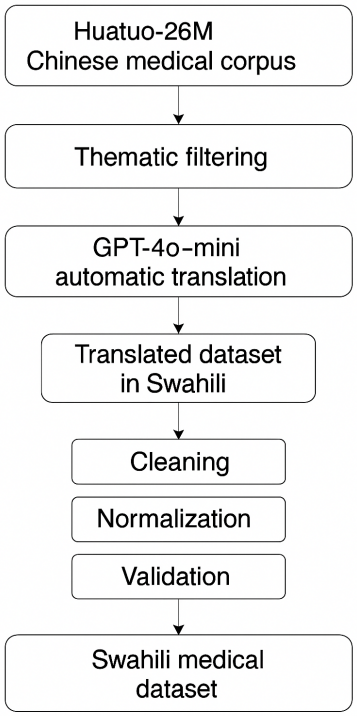


Fig. 1: Pipeline for the translation of the Huatuo dataset

3.1 Origin and Structure of the Source Corpus

The original corpus used in this study, Huatuo-26M, is a foundational resource in the development of our Swahili-language medical conversational system. Designed in 2023, this large-scale corpus was created to address the lack of structured data for training language models specialized in the medical field in Chinese. It stands out for its scale, thematic granularity, and pragmatic orientation towards simulated doctor-patient interactions, [29]. This subsection provides a detailed overview of its characteristics, organization, and the reasons behind its adoption as the foundation of the Swahili medical corpus.

In our project, we specifically used the Lite version of this corpus, titled Huatuo26M-Lite, available on the Hugging Face platform. This refined version is the result of a filtering process, qualitative enhancement, and rewriting by ChatGPT-type

language models, aiming to produce a higher-quality, more stable, and more usable subset than the raw version.

3.1.1 Initial Objectives and Positioning of Huatuo-26M

The Huatuo-26M project is part of a broader effort to develop large-scale medical textual resources specifically designed for training large language models in the clinical domain. The initial goal of its creators was twofold:

- to provide a realistic training base for health-oriented dialogue models
- to enable comparative evaluation of these models in tasks involving clinical understanding, reasoning, and question answering.

Unlike traditional corpora built from scientific articles or clinical guidelines, Huatuo-26M focuses on real conversational language, as it appears on online medical consultation platforms. As such, it offers linguistic and pragmatic diversity, which is conducive to generating contextualized, nuanced responses aligned with patients' phrasing.

3.1.2 Data Origin and Statistical Distribution

Table 1 summarizes the composition of the Huatuo-26M corpus by source type. The dataset is primarily composed of real-world patient-doctor interactions, supplemented by structured medical knowledge and simplified encyclopedic content, providing a diverse yet clinically grounded foundation for downstream applications.

Table 1. Data source distribution of the Huatuo-26M corpus

Data Source	Estimated Proportion	Description
Online medical consultations	∼95%	Real dialogues between patients and professionals via health forums or chats
Structured knowledge bases	∼3%	Definitions, care protocols, and decision trees extracted from API databases
Medical encyclopedic articles	∼1.4%	General and educational health content simplified for broader audiences

The Lite version used in our project is a filtered and enhanced subset, containing 178,000 Q-A pairs in Chinese. This dataset underwent a rigorous purification process, including:

- Deduplication
- Cleaning of syntactic and lexical duplicates
- Extraction of the most frequent questions
- Quality evaluation by ChatGPT
- Full rewriting of responses using ChatGPT, based on the original answers
- Selection of entries with the highest quality scores

These steps resulted in an optimized dataset, more reliable than the original, with improved semantic stability and demonstrably better readability and relevance, as validated by human annotation.

3.1.3 Internal Structuring of Entries

Each entry follows a stable [Q-A] format:

Question (Q): The initial request made by the patient, usually in spontaneous language, reflects their symptoms, medical history, emotions, or subjective concerns.

Answer (A): The healthcare professional's response, provided in the form of an explanation, diagnostic hypothesis, therapeutic advice, or a referral to specialized care.

Q-A pairs are accompanied by a set of metadata enabling their indexing based on medical specialty, associated pathology, quality score (assigned after validation), and a unique identifier, and a JSON structure exploitable via the datasets library

3.1.4 Semantic Richness and Clinical Diversity

Patient questions in the corpus span a broad range of symptoms, including dermatological disorders, chronic pain, organ dysfunction, infectious conditions, and psychiatric disturbances, as shown in Table 2. Many entries include rich context, such as duration, progression, previous treatment attempts, and patient demographics. On the response side, diversity is equally present, encompassing pathophysiological explanations, diagnostic hypotheses, lifestyle and dietary advice, therapeutic guidance, and specialized care recommendations. This combination makes the corpus highly exploitable for tasks such as symptom classification, contextual response generation, or the development of clinical support systems in local languages.

Table 2. Distribution des entrées par spécialité médicale

Medical Specialty	Number of Entries
Gynecology and Obstetrics	34 313
Internal Medicine	29 677
Dermatology & Sexually Transmitted Diseases	24 668
Pediatrics	21 202
ENT (Otorhinolaryngology)	13 791
Oncology	10 107
Neurosciences	10 009
Surgery	9 577
Male Health	8 109
Infectious Diseases & Immunology	5 028
Dentistry (Stomatology)	3 223
Psychology	2 989
Traditional Chinese Medicine	2 493
Reproductive Health	1 363
Others	1 133
Emergency Medicine	21

3.1.5 Pragmatic Style and Compatibility with Chatbots

One of the corpus's main strengths lies in its ability to reflect natural interactions, close to everyday spoken language. Patient questions exhibit high lexical and grammatical variability, often including approximate formulations, inconsistencies, or slips, all phenomena typical in real-world usage. Responses, though more standardized, often adopt a pedagogical or reassuring tone.

This linguistic realism makes the corpus particularly suitable for training or evaluating medical conversational models, which must be robust to patient language variability while maintaining high clinical coherence.

Even in its Lite version, the corpus has certain limitations. The authors note:

- Uneven coverage of complex pathologies
- Dependence on external tools (ChatGPT) for rewriting, which may introduce biases or simplifications
- Absence of native multilingual content

This requires careful adaptation when transferring the resource to languages with limited resources. For our Swahili project, these limitations were addressed through a controlled neural translation pipeline, rigorous data cleaning, and partial human validation.

3.2 Thematic Selection and Targeted Extraction

As part of building a Swahili-language medical chatbot suited to under-resourced healthcare systems, we undertook a strategic thematic selection of clinical domains to guide the construction of our dataset. This selection was driven by a combination of factors, including the significant disease burden across sub-Saharan Africa, the widespread shortage of trained specialists, and the structural suitability of specific medical dialogues found within the Huatuo26M-Lite corpus. These considerations led to the identification of three priority disciplines that form the core of our dataset :

- Otorhinolaryngology (ENT): 13,791 dialogues
- Dermatology and sexually transmitted infections (STIs): 24,668 dialogues
- Oncology: 10,107 dialogues

These domains address the region's urgent health challenges. In ENT, the WHO reports that over half of African countries have fewer than one ENT specialist per million, [30], while the prevalence of hearing loss continues to rise sharply, [31]. In dermatology, infectious skin conditions account for up to 80% of pediatric consultations in some rural areas, [32]. In oncology, recent projections indicate that cancer incidence in the region is expected to double by 2040, [33], a trend exacerbated by a severe shortage of oncologists and specialized infrastructure.

This triple gap of high morbidity, low specialist coverage, and unmet needs in therapeutic education fully justifies the integration of an automated support system, such as a medical chatbot operating in a local language. Furthermore, these specialties are well-represented in the Huatuo26M-Lite dataset, with rich symptom narratives and medically coherent responses, facilitating adaptation to a user-centric architecture.

The final sub-corpus comprises 48,566 medical dialogues, each formatted as a [Patient Symptoms / Medical Response] pair and fully translated into Swahili. This targeted extraction aligns the dataset with Lueji's specialization objectives, while minimizing translation, preprocessing, and fine-tuning costs, and maximizing clinical relevance in high-impact areas.

3.3 Machine Translation into Swahili

In a context characterized by the absence of structured biomedical corpora in the Swahili language, the entire set of 48,566 dialogues extracted from the Chinese sub-corpus Huatuo26M-Lite was automatically translated into Swahili. For this task, we employed

the multilingual GPT-4o-mini model. Its main advantage lies in the balance between linguistic quality and reduced operational cost, making it an ideal candidate for large-scale processing under budget constraints.

The translation was conducted via batch processing using OpenAI's official API, enabling fast and stable execution of the workflow while minimizing associated costs. Particular attention was paid to preserving the original dialogic structure to maintain the interactional logic of the corpus. Preserving this conversational dynamic is crucial to ensure compatibility with the training objectives of a medical dialogue agent. Despite the model's technical advancements, certain limitations were observed. These include incomplete translations, residual Chinese characters in specific segments, lexical ambiguities resulting from the still-unstable status of medical terminology in Swahili, and, in some complex cases, diminished clinical precision.

These observations motivated the implementation of a post-processing pipeline, detailed in the following section, designed to correct errors, standardize terminology, and validate the linguistic coherence of the final corpus. Overall, this machine translation step constitutes a critical milestone in the development of medical conversational resources for under-resourced African languages, confirming the viability of LLM-assisted linguistic transfer approaches.

3.4 Cleaning Pipeline and Linguistic Validation

To ensure the linguistic, typographic, and semantic quality of the translated corpus, a post-processing pipeline was implemented downstream of the automatic translation process. This pipeline consists of a sequence of modular stages, each designed to address a specific class of recurrent errors commonly observed in outputs from multilingual neural models. Figure 2 illustrates the overall workflow, which targets successive layers of noise to elevate the corpus to a quality level suitable for clinical language processing.

The first step involved detecting and removing partially translated pairs, typically identifiable by the presence of residual Chinese characters, non-Latin symbols, or syntactically incoherent segments. This was followed by typographic normalization, including the systematic correction of punctuation, sentence-initial capitalization, and the standardization of dialogue markers, such as quotation marks and dashes. A final filtering phase removed entries that were non-informative, off-topic, or inconsistent with basic clinical logic

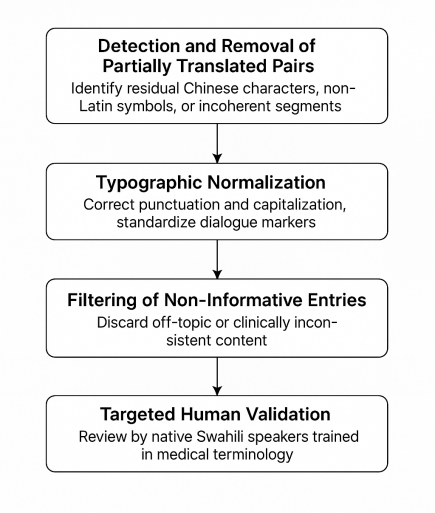


Fig. 2: Cleaning pipeline

In parallel with these automated procedures, a targeted human validation was performed on a random sample of 1,000 dialogues. Native Swahili speakers reviewed these with training in medical terminology. The revision focused on refining technical terms, correcting poorly rendered idiomatic expressions, and harmonizing therapeutic concepts. This process proved essential for improving linguistic fluency and ensuring the corpus's readability without compromising the integrity of the clinical content. Ultimately, this step marks a critical milestone toward finer alignment between the target language, expected medical register, and culturally grounded Swahili usage in both professional and lay communication.

3.5 Use Perspectives and Critical Assessment of the Corpus

The Swahili-language medical corpus developed in this study constitutes a pioneering resource for advancing language models tailored to clinical settings in low-resource environments. It stands out for its rare combination of semantic depth, thematic alignment with under-documented medical specialties, and linguistic accessibility in a major African vehicular language. Its structure, based on rigorously translated and cleaned clinical question–answer pairs, makes it suitable for a variety of applications, including fine-tuning, supervised generation, and semantic indexing. Furthermore, the foundational properties of the Huatuo26M source corpus support its relevance as a vehicle for transfer learning toward other languages or modeling paradigms, including zero-shot and controlled generation scenarios.

Nevertheless, several methodological and epistemic limitations must be considered. Although

dense and thematically filtered, the corpus remains static by design, making it increasingly vulnerable to obsolescence due to the rapid evolution of clinical guidelines. In addition, certain imperfections stemming from neural translation, particularly in oncology terminology, idiomatic phrasing, or lexicon domains poorly represented in Swahili, can introduce semantic bias or degrade medical accuracy. Human validation, while targeted and methodologically structured, was limited to a relatively small sample, which constrains quality assurance across the full dataset.

Within this context, the corpus primarily serves as a robust starting point for locally grounded biomedical AI initiatives. It provides a reliable foundation for model training, comparative benchmarking, and terminology documentation in Swahili, while laying the groundwork for a broader multilingual medical data ecosystem. Its future deployment would benefit from ongoing efforts in annotation, content updating, and disciplinary expansion to further enhance its scientific value and clinical relevance within African healthcare settings.

4 Experimental Framework and Evaluation Procedures

4.1 Objective and General Principle

The primary objective of this experimental phase is to design and evaluate a Swahili-language medical conversational system capable of generating clinically relevant and linguistically appropriate responses, without relying on supervised fine-tuning of the underlying language model. The system is built on a Retrieval-Augmented Generation (RAG) architecture, which combines the generative capabilities of a large language model (LLM) with a dense, semantically indexed document retrieval. This setup externalizes medical knowledge into an explicit knowledge base, enabling dynamic query-based access without altering the model's internal parameters.

This architectural choice is motivated by a combination of technical, operational, and contextual factors. In low-resource environments such as those found in East Africa, fine-tuning even a quantized LLM remains computationally expensive, difficult to replicate locally, and inflexible in the face of rapidly evolving medical knowledge. Moreover, the corpus used in this study was structured to consist of autonomous and retrievable knowledge units, making the RAG paradigm particularly well-suited: documents are dynamically retrieved from a vector database and injected into the model prompt at inference time, without any modification to the model's architecture.

Finally, the RAG approach offers specific advantages in the multilingual and low-resource context of medical Swahili. By shifting specialization to the document layer rather than embedding it in the model, it enables the deployment of an efficient, adaptable, and linguistically targeted conversational system. This separation of responsibilities, with explicit retrieval on one side and fluent generation on the other, offers a pragmatic response to the methodological and infrastructural constraints of medical AI in African language contexts.

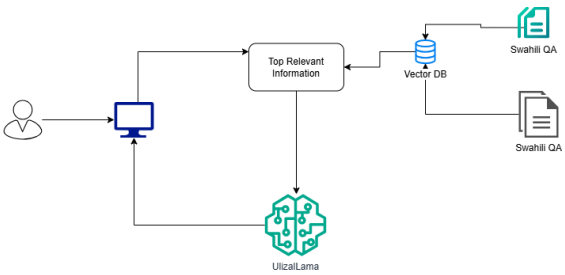


Fig. 3: RAG Pipeline

Figure 3 illustrates the functional architecture of the implemented RAG system. Starting from a query formulated in Swahili, the system encodes the question using a multilingual dense model, then performs a similarity search within a FAISS vector database containing indexed medical dialogues. The closest documents are initially retrieved and then refined through a reranking process, resulting in the selection of the most relevant contexts. These excerpts are subsequently injected into a structured prompt, along with system instructions, the user's question, and an explicit generation marker. The prompt is then submitted to a generative model, which produces a fluent response.

4.2 Generative Component

The generative engine of our RAG architecture is based on the UlizaLlama model, an optimized variant of the LLaMA-2 family, specifically adapted for the Swahili language. UlizaLlama is a 7-billion-parameter model developed by Jacaranda Health as part of its initiative to equip African languages with high-performance language models. This model represents the first LLM specifically trained for Swahili, a language that has so far been underrepresented in the large-scale training corpora of foundational models.

UlizaLlama demonstrates superior performance compared to the standard LLaMA-2 on core natural language processing tasks: improved accuracy in classification and translation, reduced latency, and greater robustness in production environments. Table 3 presents a comparative evaluation of the

two models across these dimensions, highlighting the performance gains that make UlizaLlama a particularly well-suited generative backbone for low-resource, Swahili-language medical applications.

Table 3. Comparison of Model Performance for Swahili

Model	Classification (accuracy %)	Translation (accuracy %)	Response Time (ms)
UlizaLlama	95	92	200
Meta/LLaMA2	90	88	500

UlizaLlama was pre-trained on a dedicated corpus of 321 million tokens exclusively in Swahili, using a custom tokenizer comprising 20,000 language-specific lexical units. This linguistic specialization step aims to address the limitations of generic models, such as LLaMA2, which exhibit significant performance degradation on African languages. The model was then fine-tuned through supervised instruction tuning on bilingual question–answer pairs (Swahili–English), using the LoRA (Low-Rank Adaptation) technique, which enables lightweight and modular integration of new knowledge.

4.2.1 Controlled Generation Parameters

To guide the model's output and ensure a balance between contextual fidelity, semantic accuracy, and controlled variability, a specific set of generation parameters was defined. These hyperparameters are presented in the Table 4 below.

Table 4. Model Generation Parameters

Parameter	Value	Role
`temperature`	0.2	Reduces generation randomness to favor deterministic responses.
`top_p`	0.75	Applies probabilistic regularization.
`num_beams`	1	Disables exhaustive search to speed up generation.
`max_new_tokens`	256	Limits the size of the generated response to avoid rambling.

In addition, to optimize performance in low-resource environments, the model is quantized to 4-bit precision using the bitsandbytes library. This significantly reduces memory requirements by approximately 3 to 4 times, enabling the model to run on consumer-grade GPUs with little VRAM. This optimization is essential for local deployment scenarios, especially in African contexts where access to high-performance cloud infrastructure remains limited.

The prompt provided to the model is explicitly structured and includes the following components:

- System instructions defining the role of a medical assistant;
- Document excerpts retrieved from the vector search, with annotations;
- The user's query;
- A response marker (`### JIBU ###`) used to guide the starting point of the generation.

This guided generation architecture ensures the coherence of the generated content while constraining the model's behavior.

4.2.2 Advantages for Localized Medical Deployments

The integration of UlizaLlama into our system offers several strategic advantages:

- Native linguistic specialization in Swahili, ensuring better alignment with local usage;
- Deployment on local servers, reducing dependence on proprietary or external infrastructure;
- Full control over patient data (no transfer to the cloud);
- No retraining required, significantly lowering maintenance costs.

These features make UlizaLlama particularly well-suited for conversational systems deployed in African contexts, where language constraints, limited computational resources, and data sovereignty concerns necessitate carefully targeted technical solutions.

4.3 Corpus Encoding and Indexing

The effectiveness of a RAG-based system relies heavily on its ability to accurately identify, for each user query, the most semantically relevant document fragments. This operation directly depends on the quality of the semantic encoding mechanism and the structure of the search index. In our system, the entire corpus was vectorized using a multilingual dense encoder, then indexed within a FAISS vector database.

4.3.1 Semantic Encoding with BAAI/bge-m3

We employed the BAAI/bge-m3 model, a state-of-the-art multilingual dense encoder optimized for semantic retrieval, reranking, and multilingual semantic matching tasks. This model is based on a BERT-style transformer architecture and was trained on multiple similarity and information retrieval tasks spanning over 100 languages. It demonstrates strong performance on low-resource language pairs, particularly in African contexts, and provides solid coverage for Swahili.

Each document in the corpus is structured and encoded into a 768-dimensional dense vector. This vector captures the overall semantic representation of the dialogue, integrating medical terminology, symptomatic context, and conversational structure. The encoder is used in sentence-transformers mode, with frozen parameters (no fine-tuning), ensuring interlingual consistency and reproducibility of the processing pipeline.

4.3.2 Vector Preprocessing and Normalization

Once generated, all document vectors were subjected to L2 normalization, ensuring that the Euclidean norm of each vector equals 1. This step is essential to guarantee that the similarity measure used during retrieval, namely, the cosine similarity between vectors, is mathematically equivalent to a simple dot product, which enhances both search accuracy and execution speed within dense retrieval libraries.

Normalization was performed using the faiss.normalize_L2 method, in accordance with FAISS implementation guidelines for angular similarity-based indexes.

4.3.3 Indexing with FAISS (Facebook AI Similarity Search)

To store and query the vector representations of the corpus, we used the FAISS library, which is designed for efficient similarity search in high-dimensional vector spaces. Specifically, we selected the IndexFlatIP structure, which performs a brute-force linear search using inner product similarity, particularly well-suited for cases where vectors are normalized.

The final index contains 48,566 vectors, corresponding to the full set of thematically filtered medical dialogues. Metadata associated with each vector (ID, medical specialty, source, raw text) was stored in a parallel .jsonl file, enabling efficient retrieval of document content at inference time.

4.4 Document Retrieval and Reranking

The document retrieval phase is the functional core of the RAG pipeline. Its objective is to identify, from a user query formulated in Swahili, the most relevant fragments of medical dialogues present in the indexed corpus. This step is essential for injecting precise, localized, and contextually coherent information into the generative model's prompt.

4.4.1 Initial Retrieval: Vector Similarity

When the user enters a question or a description of symptoms, it is first encoded using the same dense model as the one used for the corpus, namely BAAI/bge-m3. The text is converted into a 768-dimensional vector and then normalized (L2) to ensure geometric consistency with the pre-indexed vectors stored in the FAISS database.

The search is then performed on the FAISS IndexFlatIP index by measuring the cosine similarity between the vectorized query and the set of encoded documents. This operation returns an ordered set of documents, with the most similar ones ranked at the top of the list. In our case, we selected a top-20 initial set, which offers a good trade-off between informational coverage and processing time. The set of 20 returned documents includes :

- The unique ID of the fragment
- The raw similarity score (inner product)
- The original metadata

4.4.2 Contextual Re-ranking

To refine the selection, the top 20 results from the initial retrieval are subjected to a contextual reranking phase, which aims to prioritize documents based on their semantic relevance more precisely. This step uses the model BAAI/bge-reranker-v2-m3, a multilingual reranker trained on pairs of texts (query + passage) to predict a contextual similarity score.

The reranker processes each pair [user query, retrieved document] and assigns a more nuanced matching score than the initial inner product. Unlike pure vector search, the reranker considers syntactic relationships, the logical structure of the statements, and the exact match of mentioned medical entities.

The top 5 highest-scoring documents from this reranking are retained for the next phase (injection into the generative prompt). This reduced number helps limit the cognitive load on the model while maintaining a sufficient level of informational diversity.

4.4.3 Prompt Construction and Controlled Generation

The final step of the RAG pipeline involves constructing a structured prompt to guide the model in generating a controlled and context-aware response. The prompt plays a central role in shaping the model's behavior, ensuring that it maintains an

appropriate tone, adheres to its intended role, and produces responses grounded in relevant information. The prompt is organized into four logically ordered segments. It begins with system instructions that define the model's role as a Swahili-speaking medical assistant tasked with providing first-line health information. These instructions emphasize clarity, empathy, and informativeness, while clearly stating that the system is not a replacement for professional medical advice.

Next, the five documents retrieved and reranked earlier are inserted into the prompt as annotated reference blocks. Each includes a clinical question and a corresponding answer from the corpus, labeled with unique identifiers (e.g., ref:1, ref:2) to help the model trace its sources and rely explicitly on them when generating a response. The user's original query, typically written in natural Swahili, follows this context and is introduced as a direct information request. Placing the question after the context ensures that the model has access to the most relevant information before responding. Finally, a trigger marker (### JIBU ###) indicates exactly where the model should begin its output. This separator helps maintain a clear boundary between the input and the generated response, ensuring that the answer remains focused, coherent, and aligned with the user's intent.

For safety and ethical clarity, every generated response is systematically followed by a medical disclaimer in Swahili, dynamically appended at the end of the response block. The disclaimer takes the following form:

> *"Tafadhali kumbuka: Hili si jibu la daktari halisi. Tafuta ushauri wa mtaalamu wa afya kwa masuala yoyote ya kiafya."*

This disclaimer serves as a reminder that the system provides informative guidance rather than a formal diagnosis, in accordance with safety standards for AI systems applied in healthcare contexts. It also helps raise user awareness about the limitations of an automated tool, preserving trust without creating misleading expectations.

4.5 Evaluation Metrics

The performance of an automatic text generation system is typically assessed using standard metrics that measure how closely the generated output matches a reference response, often referred to as the gold standard. In this study, evaluation is conducted exclusively through automated methods, specifically ROUGE, BLEU, and GLEU. These metrics respectively capture lexical coverage, phrasal fidelity, and local generation robustness, [34]. All three are based on n-grams, meaning they compare sequences of n consecutive words between the generated response and the reference texts to quantify similarity.

4.5.1 BLEU (Bilingual Evaluation Understudy)

The BLEU metric is a widely adopted automatic evaluation method for assessing the quality of generated text, originally designed for machine translation and now extensively used in evaluating chatbots and generative models, [35]. It measures the n-gram precision, i.e., the proportion of n-grams in the generated output that also appear in a human reference. The BLEU score aggregates these weighted n-gram precisions and applies a brevity penalty to discourage excessively short outputs.

The general formula for BLEU can be expressed as:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where:

- p_n denotes the modified precision for n-grams of order n,
- w_n represents the weight assigned to each order (typically uniform),
- BP is the brevity penalty.

Variants such as BLEU-1 through BLEU-4 evaluate precision across different n-gram sizes, with higher orders capturing more complex syntactic and lexical structures. BLEU-1 focuses on unigram precision (lexical alignment), whereas BLEU-4 enforces stricter sequence matching, reflecting greater syntactic coherence.

4.5.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

The ROUGE metric is a family of measures developed to evaluate the quality of generated text, originally for automatic summarization and now widely used in natural language generation. Unlike BLEU, which emphasizes precision, ROUGE focuses on recall, assessing how much of the reference content is captured by the generated text. The general form of the ROUGE-N metric is expressed as:

$$\frac{\sum_{\text{ngram}\in\text{Ref}} \min(\text{Count}_{\text{ref}}(\text{ngram}), \text{Count}_{\text{gen}}(\text{ngram}))}{\sum_{\text{ngram}\in\text{Ref}} \text{Count}_{\text{ref}}(\text{ngram})}$$

ROUGE includes several variants such as ROUGE-1, ROUGE-2, and ROUGE-L, which

measure lexical and sequence-level overlap between generated and reference texts. While it effectively captures content recall, ROUGE does not account for semantics, paraphrasing, or multiple valid references. Nevertheless, it remains a standard recall-oriented baseline for evaluating generative language models.

4.5.3 GLEU (Google BLEU)

This metric, introduced by Google Research, is an alternative to BLEU designed to address its limitations in tasks with multiple valid outputs, such as machine translation and dialogue generation. Unlike BLEU, which emphasizes precision, GLEU symmetrically incorporates both precision and recall, making it more robust when diverse reference formulations exist. It is particularly effective for evaluating short or conversational text outputs, [34].

The general form of the GLEU score is expressed as:

$$\min\left(\frac{|\text{n-grams}(C)\cap\text{n-grams}(R)|}{|\text{n-grams}(C)|},\frac{|\text{n-grams}(C)\cap\text{n-grams}(R)|}{|\text{n-grams}(R)|}\right)$$

where C denotes the candidate (generated text) and R the reference.

GLEU penalizes both missing reference content and unnecessary words, while being more tolerant of paraphrasing than BLEU. However, it does not directly capture semantic similarity and may still be sensitive to tokenization or limited reference availability.

In our evaluation, GLEU is used to assess the robustness of the Swahili responses generated by the RAG system on short and variable-length queries. Its inclusion complements BLEU and ROUGE by offering a more balanced judgment of phrasing variability and recall precision trade-offs. The combination of BLEU, ROUGE, and GLEU provides a complementary and multi-faceted framework for evaluating the linguistic and lexical quality of the system's generated responses. BLEU offers a precision-oriented view, ROUGE emphasizes content recall, and GLEU balances both perspectives for short and variable utterances. While each metric has known limitations in capturing semantic equivalence or pragmatic adequacy, their joint application remains a standard and reproducible approach for benchmarking generative language models.

In the following section, we apply these metrics to quantitatively compare the performance of our retrieval-augmented generation (RAG) architecture with that of the previously developed supervised fine-tuning approach. The results highlight the respective strengths and limitations of each strategy in the context of Swahili-language medical question answering.

5 Experimental Results and Comparative Analysis

This section presents the quantitative results obtained from the evaluation of the Retrieval-Augmented Generation (RAG) system developed based on the Swahili-speaking medical corpus. The objective is to assess the quality of the system's generated responses using standard text generation evaluation criteria by applying previously introduced automatic metrics.

Following a description of the evaluation protocol and the characteristics of the test set employed, we report the scores achieved by the RAG system according to various metrics (ROUGE, BLEU, GLEU), and subsequently discuss their implications. A comparative analysis with the fine-tuned supervised model will be conducted at the end of the section to contextualize the two approaches from both qualitative and operational standpoints.

5.1 Evaluation Framework

The experimental evaluation of the RAG system was conducted on a test set comprising 500 real-world medical queries in Swahili, covering the three clinical specialties selected during the corpus construction: otorhinolaryngology, dermatology, sexually transmitted infections, and oncology. These questions, formulated in natural language, simulate typical use cases in frontline healthcare contexts.

For each question, a validated reference answer was pre-established and stored in a structured, aligned file, enabling automatic comparison between the system's outputs and the expected responses. This setup ensures a rigorous and reproducible assessment of generative behavior.

The quantitative analysis relies exclusively on automatic metrics, computed through a dedicated evaluation pipeline. The metrics used are summarized in the Table 5 below:

Table 5. Evaluation Metrics for Generated Text

Metric	Purpose
`ROUGE 1, 2, L`	Measure lexical overlap between the generated response and the reference.
`BLEU 1-4`	Quantify phrasal precision using n-grams of increasing size (from 1 to 4).
`GLEU`	Evaluate local robustness by accounting for both precision and recall on short sequences.

This framework enables a standardized and objective evaluation of the system's generative capabilities.

5.2 Quantitative Results

5.2.1 BLEU Metric Results

The Table 6 reports the BLEU scores obtained by the RAG system across its four standard variants. These scores measure n-gram precision, i.e., the proportion of lexical segments (ranging from 1 to 4 words) in the generated response that also appear in the reference answer.

Table 6. BLEU scores achieved by the RAG system

Metric	Score (%)
BLEU-1	19.12
BLEU-2	8.18
BLEU-3	3.95
BLEU-4	2.18

The results reveal a performance pattern that is characteristic of generative systems operating in low-resource language settings:

- The BLEU-1 score (19.12%) indicates that nearly one out of five words generated by the system also appears in the reference response. While modest, this level of lexical overlap is consistent with expectations for open-ended conversational tasks, particularly in specialized domains such as medicine.
- The steady decline across BLEU-2, BLEU-3, and BLEU-4 scores reflects the increasing difficulty for the model to produce longer phrasal segments that precisely match the reference formulations. This degradation is attributable to several factors, including syntactic variability inherent to the Swahili language structure, the lexical richness of the medical domain, where multiple paraphrastic or interchangeable formulations are valid, and the Absence of supervised fine-tuning, with the model relying solely on contextually retrieved documents.
- The low BLEU-4 score (2.18%) should not be interpreted as a strict indicator of failure, but rather as a symptom of natural lexical divergence between generated and reference answers, both of which may be semantically valid while differing stylistically.
- Lastly, the gap between BLEU-1 and BLEU-4 confirms that while the model partially captures the expected vocabulary, it often fails to reproduce exact structural patterns. These findings will be further contextualized in a subsequent subsection, which compares the results with those of the fine-tuned model.

5.2.2 ROUGE Metric Results

The following Table 7 summarizes the ROUGE scores obtained by the RAG system. ROUGE metrics evaluate lexical coverage between generated responses and reference answers. Unlike BLEU, which is precision-oriented, ROUGE focuses on recall, that is, the extent to which the system successfully reproduces the expected content from the reference answers.

Table 7. ROUGE scores achieved by the RAG system

Metric	Score (%)
ROUGE-1	29.16
ROUGE-2	6.64
ROUGE-L	18.11

- The ROUGE-1 score (29.16%) indicates that, on average, nearly 30% of the words present in the reference answers are also found in the system-generated outputs. This suggests a relatively strong ability to recover core lexical content, particularly domain-specific terms such as medical entities, symptoms, and common clinical instructions.
- By contrast, the ROUGE-2 score (6.64%) reveals a limited recovery of exact bigram sequences. This reflects the high degree of formulation variability in generated responses, which is common in natural language generation, particularly when the model paraphrases or syntactically reformulates retrieved content rather than replicating it verbatim.
- The ROUGE-L score (18.11%), which measures the longest common subsequence between generated and reference texts, indicates a partial structural alignment. This suggests that the model is capable of capturing the overall phrasal logic or discourse structure expected in the responses, even without strictly mimicking the reference wording.

Overall, the ROUGE results confirm that the system is effectively leveraging retrieved textual fragments to produce responses that contain a substantial portion of the expected content. These findings support the hypothesis that the RAG pipeline provides sufficient lexical grounding, even in low-resource multilingual settings.

5.2.3 GLEU Metric Result

The GLEU metric, which combines both precision and recall over short sequences, was used to

evaluate the phrasal robustness of the system in a conversational context. It is particularly suitable for tasks involving the generation of concise medical responses, where multiple lexical formulations can be equally valid and effective.

Table 8. GLEU score achieved by the RAG system

Metric	Score (%)
GLEU	9.18

- The GLEU score (9.18%), as illustrated in Table 8, reflects the system's ability to produce lexical sequences that partially overlap with reference responses, while tolerating legitimate paraphrasing and rewording. This metric serves as an indicator of local syntactic robustness.
- It confirms that despite variability in wording or word order, a non-negligible portion of the generated expressions aligns with those found in the references.

In the context of a low-resource language with limited standardization in formulation, this score suggests that the system remains reasonably aligned with expected phrasing while maintaining stylistic flexibility.

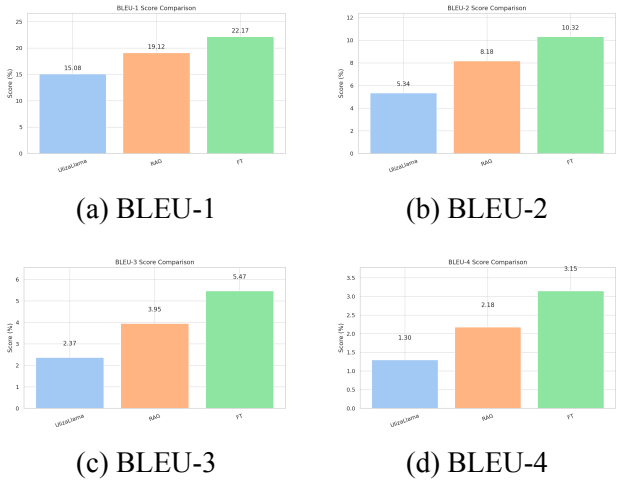
5.3 Comparison of Approaches

To complement the quantitative results, we conducted a comparative analysis of the three evaluated systems: UlizaLlama (baseline), UlizaLlama with RAG, and the fine-tuned model Lueji. The aim was to highlight their respective behaviors, strengths, and limitations in generating medical responses in Swahili. This comparison is based not only on aggregated metric scores but also on qualitative observations drawn from real test samples.

The goal is to assess how each system responds to medical queries in terms of fluency, completeness, domain alignment, and operational viability. While the fine-tuned model unsurprisingly yields the highest overall scores, the RAG architecture demonstrates promising trade-offs between relevance and deployment cost.

5.3.1 Analysis of BLEU Scores

The BLEU scores obtained across the four configurations (BLEU-1 to BLEU-4) reveal progressive differences in lexical precision and phrasing among the three models evaluated. As shown in Figure 4, Lueji (the fine-tuned model) consistently achieves the highest scores, closely followed by the RAG architecture, with UlizaLlama (baseline) trailing significantly behind.



(a) BLEU-1 (b) BLEU-2

(c) BLEU-3 (d) BLEU-4

Fig. 4: BLEU scores for selected models

BLEU-1 indicates that Lueji retrieves 22.17% of unigrams from the reference, reflecting its ability to reuse clinical terms and common response formulations accurately. RAG follows with 19.12%, suggesting effective keyword integration via document retrieval. UlizaLlama remains below at 15.08%, producing more generic and less medically aligned outputs.

As we progress to BLEU-2, BLEU-3, and BLEU-4, the scores decline across all models, reflecting the inherent difficulty of replicating longer, exact phrasing. However, the gap between FT and RAG narrows modestly, while Uliza continues to decline more sharply. This behavior is expected, given that neither RAG nor Uliza undergoes supervised training and relies more on local patterns or context injection.

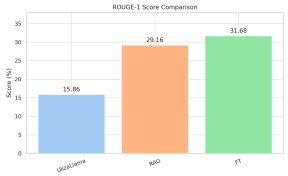
The BLEU-4 score of the fine-tuned model (3.15%) may appear modest in absolute terms, but it remains consistently higher than that of RAG (2.18%) and Uliza (1.30%), confirming its superior capacity for sentence-level structuring. Still, the proximity between RAG and FT in BLEU-2 and BLEU-3 is noteworthy, especially considering that RAG leverages no supervised training and instead relies on domain-informed content injection through semantic retrieval.

In summary, the BLEU scores support a clear ranking in terms of surface-level fluency and lexical alignment: FT > RAG > Uliza. Yet they also underscore the RAG system's ability to bridge a significant part of the gap between a fully fine-tuned model and a zero-shot baseline, particularly for short to medium spans of medical content.

5.3.2 Analysis of ROUGE Scores

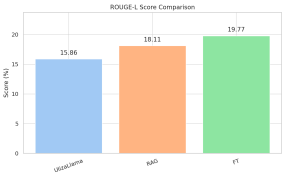
The ROUGE metrics offer a complementary perspective to BLEU by emphasizing content

recall rather than precision. They measure the extent to which the generated responses successfully retrieve the expected clinical information present in the reference answers. As such, ROUGE scores are particularly well-suited for evaluating responses that may be lexically diverse but conceptually aligned.



(a) ROUGE-1

(b) ROUGE-2



(c) ROUGE-L

Fig. 5: ROUGE scores for selected models

We can see in Figure 5, ROUGE-1, based on unigram recall, shows that the fine-tuned model (Lueji) achieves the highest score at 31.68%, indicating that nearly one-third of the lexical content present in the reference responses is captured. The RAG system achieves a performance of 29.16%, highlighting its ability to leverage relevant medical terms from the retrieved corpus. In contrast, UlizaLlama remains significantly behind at 15.86%, reflecting its generic generation behavior and lack of domain anchoring.

For ROUGE-2, which focuses on bigram recall and therefore measures short-span contextual fidelity, Lueji scores 8.38%, compared to 6.64% for RAG and 4.23% for Uliza. The moderate drop from ROUGE-1 is expected, given the increased difficulty of matching consecutive terms in natural language generation. Yet, the fact that RAG outperforms Uliza by over 2 points confirms its added value in capturing phrase-level continuity.

ROUGE-L, which relies on the longest common subsequence (LCS) between the generated and reference responses, captures structural similarity. Lueji again leads with 19.77%, followed by RAG at 18.11%, and Uliza at 15.86%. This suggests that both Lueji and RAG are capable of reproducing partial response skeletons or medical discourse patterns, even if not identically phrased.

Taken together, the ROUGE scores reinforce the interpretation drawn from BLEU: the fine-tuned model excels in both breadth and structure of content recovery. At the same time, the RAG configuration closes a significant part of the gap without requiring supervised training. The results also suggest that the Swahili corpus used in RAG indexing provides a sufficiently rich and structured knowledge base to support the generation of informative responses.

5.3.3 Analysis of GLEU Score

The GLEU metric, designed to balance both precision and recall of short sequences, provides a useful measure of robustness in contexts where multiple valid phrasings are possible, such as question answering in natural language. In contrast to BLEU or ROUGE, GLEU penalizes both over-generation and under-generation, making it particularly relevant for assessing conversational agents responding to compact or syntactically flexible queries.

As illustrated in Figure 6, the fine-tuned model achieves the highest GLEU score of 10.77%, confirming its ability to produce concise yet accurate responses that closely align with the structure and vocabulary of the reference. The RAG system achieves a performance level of 9.18%, maintaining a competitive level despite the absence of supervised learning. The baseline UlizaLlama model trails behind with 6.85%, a score that reflects its lack of specialization and tendency to generate more generic outputs.

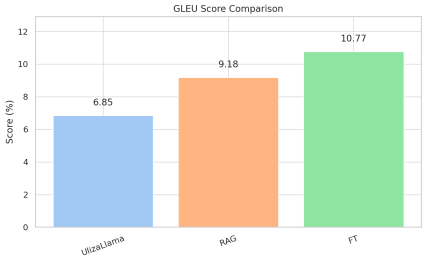


Fig. 6: GLEU Scores for selected models

Interestingly, the gap between RAG and Lueji in GLEU is smaller than in other metrics, suggesting that RAG performs relatively well on sentence segments and localized phrasing. This supports the hypothesis that the injection of targeted, semantically relevant content from the retrieval step helps stabilize and contextualize responses, even if the overall discourse structure remains simpler.In the Swahili medical domain, where the reference corpus is relatively constrained and linguistic variation is common, GLEU offers an additional layer of interpretability. It confirms that RAG is capable of producing short, meaningful, and contextually grounded responses, which are often sufficient in low-resource clinical assistance scenarios.

5.3.4 Case-Based Qualitative Comparison

To complement the quantitative metrics, we present a representative user query from the evaluation set and compare the responses generated by each system: the fine-tuned model, the RAG-enhanced UlizaLlama, and the baseline UlizaLlama. For reference, we also include the corresponding entry from the expert dataset.

User Query (Swahili):

User Input

Siku za hivi karibuni nimechoka, kila wakati ninakutana na jasho, sasa sijui ni vipi nilivyoambukizwa, miguu yangu imeanza kupata vidonda vidogo, na inakatika sana, nikikwaruza inabaki na alama nyekundu, na kuna nguo za nyeupe karibu nayo. Nimeenda hospitalini, wamesema ni psoriasis, sijui ugonjwa huu unatibiwaje, nasikia si rahisi kuponya kabisa. Tafadhali daktari, psoriasis inatibiwaje? Asante.

Lueji (Fine-Tuned):

Response Fine-Tuning (Lueji)

Psoriasis ni ugonjwa wa ngozi wa muda mrefu, ambao husababishwa na mambo mbalimbali kama vile urithi, kinga, na mazingira. Njia za matibabu ya psoriasis ni nyingi, ikiwa ni pamoja na matibabu ya dawa, phototherapy, na matibabu ya ndani. Matibabu ya dawa yanajumuisha dawa za nje na dawa za mdomo, kama vile corticosteroids, non-steroidal drugs, na immunosuppressants. Phototherapy inajumuisha matibabu ya mwanga wa ultraviolet, matibabu ya mwanga wa narrowband, na matibabu ya mwanga wa ultraviolet mwishoni. Matibabu ya ndani yanajumuisha matibabu ya nje, matibabu ya ndani, na dawa za jadi. Hata hivyo, matibabu ya psoriasis yanahitaji kuchaguliwa kulingana na hali ya mgonjwa na ushauri wa daktari. Hivyo basi, inashauriwa kwa wagonjwa kutembelea hospitali kwa ajili ya uchunguzi wa kina na matibabu. Wakati huo huo, wagonjwa wanapaswa kuzingatia kupumzika na lishe bora, na kuepuka kula vyakula vyenye pilipili na vichocheo.

RAG (UlizaLlama + Retrieval):

Response RAG System

Psoriasis ni ugonjwa wa muda mrefu wa ngozi unaosababishwa na uvimbe wa ngozi, ambao unaweza kusababisha dalili kama vile madoa mekundu, ganda la rangi ya fedha, na maumivu. Matibabu ya psoriasis yanatofautiana kulingana na ukali wa ugonjwa na hali ya mtu binafsi. Kwa kawaida, tiba za psoriasis zinajumuisha dawa za nje, dawa za mdomo, na phototherapy. Matibabu ya psoriasis pia yanapaswa kujumuisha utunzaji wa usafi wa ngozi, kuepuka vitu vinavyosababisha mzio, na kula chakula chenye lishe. Ikiwa psoriasis haijatibiwa, inaweza kusababisha maumivu zaidi na uharibifu wa ngozi, kwa hivyo ni muhimu kutafuta ushauri wa daktari wa ngozi.

UlizaLlama (Baseline):

Response UlizaLlama (Untrained)

Psoriasis ni ugonjwa wa kuambukiza wa ngozi ambao husababisha vipele, madoa, na vipele vya rangi ya manjano au ya njano kwenye ngozi na/au nywele. Ingawa hakuna tiba ya tiba ya psoriasis, matibabu yanayojumuisha creams za topical, dawa za kuzuia vipele, na mikakati ya mtindo wa maisha inaweza kusaidia kudhibiti dalili.

Reference (Expert Dataset):

Response Expert Dataset

Psoriasis ni ugonjwa wa ngozi wa muda mrefu, na matibabu yake ni ngumu kweli. Wagonjwa wengi wanasema, ingawa wanaweza kudhibiti hali hiyo kwa muda mfupi, kwa muda mrefu hali hiyo bado inarudi, na kila wakati inaporudi inaweza kuwa mbaya zaidi. Dawa za kawaida zinaweza tu kupunguza dalili, lakini kuna uwezekano mkubwa wa kurudi tena, jambo ambalo linawatatiza wagonjwa sana. Matibabu ya dawa za nje yanaweza tu kutibu dalili bila kutatua chanzo cha tatizo, na pia kuna madhara makubwa, kama vile kuunda rangi kwenye ngozi, na kusababisha ngozi kuwa kavu na rough. Kutumia dawa hizi kwa muda mrefu bila mwelekeo inaweza kusababisha psoriasis kuharibika hata kubadilika, na kuwa vigumu kutibu. Hivyo, katika matibabu ya psoriasis, inapendekezwa kuchagua kwa makini matibabu ya dawa, na pia wakati wa matibabu, kuzingatia kudumisha unyevu wa ngozi, na kuepuka kutumia vipodozi na bidhaa za usafi zenye kemikali kali. Aidha, unaweza pia kujaribu baadhi ya mbinu za matibabu za jadi za Kichina, kama vile soaking therapy na topical therapy. Hata hivyo, mpango maalum wa matibabu unahitaji kuamuliwa kulingana na hali yako binafsi, na inapendekezwa ushauri wa daktari mtaalamu kwa ajili ya uchunguzi na matibabu.

Interpretation The fine-tuned Lueji model produces a relatively structured and medically informed response, generally in line with reference expectations. The RAG-enhanced UlizaLlama system demonstrates solid clinical relevance and appropriate terminology, though it remains slightly shorter and less personalized. The baseline UlizaLlama model generates an inaccurate and overly generic answer, falsely describing psoriasis as an infectious disease. These results confirm that both Lueji and RAG provide clinically usable outputs, while only the former reaches full fluency and depth.

Table 9 highlights the distinct generative behaviors of each system. The fine-tuned model offers the highest level of coherence, medical precision, and structural completeness, aligning closely with reference responses in both content and tone. The RAG system, although slightly more concise, successfully integrates domain-relevant terminology and delivers technically sound outputs, making it a viable alternative in resource-constrained settings. By contrast, the baseline UlizaLlama model generates inconsistent and sometimes inaccurate answers, underscoring the risks of unadapted models in clinical applications. This comparative analysis confirms that retrieval augmentation, although not matching fine-tuning in narrative fluency, substantially improves response quality and medical reliability compared to non-specialized language models.

Table 9. Qualitative comparison of system responses on a representative Swahili medical query

Model	Summary of Response
Expert Dataset	Provides a detailed and cautious overview of psoriasis management. Acknowledges the risk of relapse, outlines the drawbacks of topical medication, warns of potential side effects, and emphasizes the importance of personalized care. Includes mention of traditional medicine.
Fine-Tuned	Highly structured and medically rich. Covers the pathophysiology of psoriasis, multiple treatment options (topicals, phototherapy, systemic drugs), personalized care, and lifestyle advice. Reads like a concise, evidence-informed medical summary.
RAG	Balanced and clinically relevant. Includes core symptoms and treatment modalities, emphasizing hygiene and avoidance of allergens. Slightly more generic but still technically sound and locally applicable.
UlizaLlama (Baseline)	Superficial and partially inaccurate. Incorrectly describes psoriasis as infectious, includes questionable symptoms, and lacks nuance in treatment. Generic formulation with factual errors.

6 Conclusion and Future Directions

A clear ambition drove this study: to propose a modular and scalable methodology for the specialization of language models in healthcare, particularly within multilingual and low-resource settings. By focusing on a Retrieval-Augmented Generation (RAG) architecture, we demonstrated that it is possible to achieve a balanced compromise

between linguistic performance, clinical robustness, and technical sustainability without relying on computationally intensive fine-tuning procedures.

The proposed RAG pipeline integrates several complementary components, including dense multilingual semantic encoding, vector-based indexing, contextual reranking, and generation through a quantized model. This framework enables context-aware responses to complex medical queries by leveraging a structured question-answer corpus that was carefully translated and cleaned. The resulting system produces concise, clinically informed, and linguistically coherent outputs aligned with the expected professional register.

From a quantitative standpoint, the BLEU, ROUGE, and GLEU scores obtained were comparable to those achieved by a fully fine-tuned supervised model. This proximity underscores the potential of the RAG paradigm for efficiently reusing domain-specific content and for capturing relevant formulations through semantic retrieval, even in the absence of direct task-specific training. In contrast to the standalone UlizaLlama model, the RAG configuration enhanced both clinical density and terminological precision, with measurable improvements in factual fidelity.

The practical advantages of such an approach are substantial. The model does not require retraining when the underlying corpus evolves, enabling rapid and cost-effective updates of medical knowledge. The entire pipeline can be deployed locally, ensuring data sovereignty, a crucial requirement in the healthcare domain. Furthermore, the modular and extensible architecture facilitates adaptation to additional medical specialties and regional contexts. Nevertheless, this first iteration presents limitations that warrant further investigation. The source corpus, despite rigorous preprocessing, remains partially derived from machine translation, which introduces potential biases and inaccuracies, particularly in technical terminology, cultural idioms, and pragmatic nuances. Similarly, while the UlizaLlama generative component performs adequately, its ability to structure complex medical reasoning remains below that of specialized models, even when provided with enriched contextual information.

Future work will focus on several complementary directions. A priority is to enhance corpus quality through the collection of locally sourced clinical dialogues in hospital settings, which will be annotated and validated by Swahili-speaking healthcare professionals. Such efforts would enrich the clinical signal, improve stylistic and cultural alignment, and mitigate the translation-related artifacts observed in this version. In parallel, experiments with other open-source generative models better suited for cross-lingual transfer could further improve linguistic coverage and factual robustness. A second research direction involves expanding the thematic scope. Beyond the three current medical specialties, integrating modules that address maternal health, pediatrics, tropical diseases, and mental health would enable the system to reflect regional epidemiological priorities better. This effort will require close collaboration with local healthcare institutions to identify pressing clinical needs and co-develop relevant terminological and knowledge resources. Finally, ethical and community-centered dimensions must be embedded into future development. Offline mobile deployment, adaptation to local dialects, and active involvement of physicians, nurses, community health workers, and patients in the system's design and evaluation are essential to ensure acceptability, usability, and contextual relevance across real-world African healthcare environments.

In summary, this study validates the feasibility of an effective Swahili-language medical chatbot built upon a lightweight, modular, and updatable RAG pipeline. It lays the groundwork for a localized medical AI ecosystem conceived not as a top-down technological imposition, but as a tool for dialogue, empowerment, and capacity-building in support of African healthcare systems.

References:

[1] B. Petrucci et al., "The global otolaryngology-head and neck surgery workforce," *JAMA Otolaryngology–Head & Neck Surgery*, vol. 149, no. 10, pp. 904–911, 2023. DOI: 10.1001/jamaoto.2023.2339. [Online]. Available: https://doi.org/10.1001/jamaoto.2023.2339.

[2] M. M. Willie, "Examining the scarcity of oncology healthcare providers in cancer management: A case study of the eastern cape province, south africa," *Open Health*, vol. 6, no. 1, p. 20 250 058, 2025, (Accessed: May 02, 2025). DOI: 10.1515/ohe-2025-0058. [Online]. Available: https://doi.org/10.1515/ohe-2025-0058.

[3] V. Vanderpuye et al., "Cancer care workforce in africa: Perspectives from a global survey," *Infectious Agents and Cancer*, vol. 14, no. 1, p. 11, 2019. DOI: 10.1186/s13027-019-0227-8. [Online]. Available: https://doi.org/10.1186/s13027-019-0227-8.

[4] B. S. Sylla and C. P. Wild, "A million africans a year dying from cancer by 2030: What can cancer research and control offer to the continent?" *International Journal of Cancer*, vol. 130, no. 2, pp. 245–250, 2012. DOI: 10.1002/ijc.26333. [Online]. Available: https://doi.org/10.1002/ijc.26333.

[5] Kumar and S. Joshi, "Applications of AI in healthcare sector for enhancement of medical decision making and quality of service," in Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiang Mai, Thailand: IEEE, 2022, pp. 37–41. DOI: 10.1109/DASA54658.2022.9765041. [Online]. Available: https://doi.org/10.1109/DASA54658.2022.9765041.

[6] J. C. L. Chow and K. Li, "Large language models in medical chatbots: Opportunities, challenges, and the need to address AI risks," *Information*, vol. 16, no. 7, p. 549, 2025. DOI: 10.3390/info16070549. [Online]. Available: https://doi.org/10.3390/info16070549.

[7] H. Al Shamsi, A. G. Almutairi, S. Al Mashrafi, and T. Al Kalbani, "Implications of language barriers for healthcare: A systematic review," *Oman Medical Journal*, vol. 35, no. 2, e122, 2020. DOI: 10.5001/omj.2020.40. [Online]. Available: https://doi.org/10.5001/omj.2020.40.

[8] A. E. Babalola, V. Johnson, A. Oromakinde, et al., "The role of local languages in effective health service delivery," *Discover Public Health*, vol. 22, no. 1, p. 59, 2025. DOI: 10.1186/s12982-025-00429-5. [Online]. Available: https://doi.org/10.1186/s12982-025-00429-5.

[9] D. K. Kabeya, W. V. Kambale, J.-G. M. Mboma, V. N. Bendo, S. K. Kasereka, and K. Kyamakya, "Designing a swahili-speaking medical chatbot for oncology, dermatology, and otorhinolaryngology care in low-resource settings," in *Proceedings of the 2025 Conference on Information Communications Technology and Society (ICTAS)*, Durban, South Africa, 2025, pp. 1–6. DOI: 10.1109/ICTAS64866.2025.11155337. [Online]. Available: https://doi.org/10.1109/ICTAS64866.2025.11155337.

[10] K. Singhal et al., "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, Mar. 2025. DOI: 10.1038/s41591-024-03423-7. [Online]. Available: https://doi.org/10.1038/s41591-024-03423-7.

[11] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge," Cureus, vol. 15, no. 6, e40704, 2023. DOI: 10.48550/arXiv.2303.14070 [Online]. Available: https://doi.org/10.48550/arXiv.2303.14070.

[12] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: Toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833–1843, Apr. 2024, ISSN: 1527-974X. DOI: 10.1093/jamia/ocae045. eprint: https://academic.oup.com/jamia/article-pdf/31/9/1833/58868261/ocae045.pdf. [Online]. Available: https://doi.org/10.1093/jamia/ocae045.

[13] H. Xiong et al., *DoctorGLM: Fine-tuning your chinese doctor is not a herculean task*, arXiv preprint, arXiv:2304.01097 (Accessed: December 04, 2025), 2023. [Online]. Available: https://arxiv.org/abs/2304.01097.

[14] Azam, Z. Naz, and M. U. G. Khan, "Cancerbot: A retrieval-augmented generation based cancer chatbot using large language models," in Proceedings of the 2024 18th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 2024, pp. 1–6. DOI: 10.1109/ICOSST64562.2024.10871155. [Online]. Available: https://doi.org/10.1109/ICOSST64562.2024.10871155.

[15] C. Li et al., "Development of a meta-question enhanced retrieval-augmented generation model and its application in dermatology," in *Proceedings of the 2024 17th International Conference on Advanced Computer Theory and Engineering (ICACTE)*, Xiamen, China: IEEE, 2024, pp. 281–285. DOI: 10.1109/ICACTE62428.2024.10871274. [Online]. Available: https://doi.org/10.1109/ICACTE62428.2024.10871274.

[16] S. P. Bande, V. N, and P. K. K, "Augmenting medical diagnostics with AI: A dual approach using RAG-based chatbots and NanoGPT models," in *Proceedings of the 2025 6th International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, India: IEEE, 2025, pp. 1–6. DOI: 10.1109/RAIT65068.2025.11088917. [Online].

Available: https://doi.org/10.1109/RAIT65068.2025.11088917.

[17] H. Zhang et al., "HuatuoGPT, towards taming language models to be a doctor," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, 2023, pp. 10 859–10 885. DOI: 10.18653/v1/2023.findings-emnlp.725. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.725.

[18] M. Alshammary, M. N. Uddin, and L. Khan, "RFPG: Question-answering from low-resource language (Arabic) texts using factually aware RAG," in *Proceedings of the 2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC)*, Cincinnati, OH, USA: IEEE, 2024, pp. 107–116. DOI: 10.1109/CIC62241.2024.00023. [Online]. Available: https://doi.org/10.1109/CIC62241.2024.00023.

[19] H. Hosseini, M. S. Zare, A. H. Mohammadi, A. Kazemi, Z. Zojaji, and M. A. Nematbakhsh, "PersianRAG: A retrieval-augmented generation system for the Persian language," in *Proceedings of the 2024 15th International Conference on Information and Knowledge Technology (IKT)*, Isfahan, Iran: IEEE, 2024, pp. 272–278. DOI: 10.1109/IKT65497.2024.10892726. [Online]. Available: https://doi.org/10.1109/IKT65497.2024.10892726.

[20] S. M. M. R. J. Senanayaka, A. W. A. D. N. D. Abeysekara, and M. G. N. N. Premadasa, "SingRAG: A translation-augmented framework for code-mixed Singlish processing," in *Proceedings of the 2024 9th International Conference on Information Technology Research (ICITR)*, Colombo, Sri Lanka: IEEE, 2024, pp. 1–6. DOI: 10.1109/ICITR64794.2024.10857714. [Online]. Available: https://doi.org/10.1109/ICITR64794.2024.10857714.

[21] J. Singh and R. Thakur, *Quantum-RAG and PunGPT2: Advancing low-resource language generation and retrieval for the Punjabi language*, arXiv preprint, arXiv:2508.01918 (Accessed: December 04, 2025), 2025. [Online]. Available: https://arxiv.org/abs/2508.01918.

[22] B. Bogale, T. Tegegne, S. Teferra, et al., *RAG based QA for low-resource languages*, Research Square preprint, (Accessed: December 04, 2025), 2024. DOI: 10.21203/rs.3.rs-5360450/v1. [Online]. Available: https://doi.org/10.21203/rs.3.rs-5360450/v1.

[23] G. Martin, M. E. Mswahili, Y.-S. Jeong, and J. Woo, "SwahBERT: Language model of Swahili," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, 2022, pp. 303–313. DOI: 10.18653/v1/2022.naacl-main.23. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.23.

[24] S. A. Sani, S. H. Muhammad, and D. Jarvis, *Investigating the impact of language-adaptive fine-tuning on sentiment analysis in the Hausa language using AfriBERTa*, arXiv preprint, arXiv:2501.11023 (Accessed: December 04, 2025), 2025. [Online]. Available: https://arxiv.org/abs/2501.11023.

[25] F. P. Dossou et al., "Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages," in Proceedings of the 29th International Conference on Computational Linguistics, (Accessed: December 04, 2025), Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 5235–5269. [Online]. Available: https://doi.org/10.18653/v1/2022.sustainlp-1.11.

[26] B. W. Wanjawa, L. D. A. Wanzare, F. Indede, O. Mconyango, L. Muchemi, and E. Ombui, "KenSwQuAD: A question answering dataset for swahili low-resource language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1–20, 2023. DOI: 10.1145/3578553. [Online]. Available: https://doi.org/10.1145/3578553.

[27] Adelani et al., "MasakhaNER: Named entity recognition for african languages," Transactions of the Association for Computational Linguistics, vol. 9, pp. 1116–1131, 2021. DOI: 10.1162/tacl_a_00416. [Online]. Available: https://doi.org/10.1162/tacl_a_00416.

[28] O. Ogundepo et al., "AfriQA: Cross-lingual open-retrieval question answering for african languages," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada: Association for Computational Linguistics, 2023, pp. 14 445–14 465. DOI: 10.18653/v1/2023.findings-emnlp.997. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.997.

[29] J. Li et al., *Huatuo-26m, a large-scale chinese medical qa dataset*, 2023. arXiv: 2305.01526 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2305.01526.

[30] S. O. Ayodele and S. K. Aremu, "The cost of setting up an ENT endoscopic practice in lower middle-income countries of sub-saharan africa," *Journal of the West African College of Surgeons*, vol. 12, no. 2, pp. 104–108, 2022. DOI: 10.4103/jwas.jwas_57_22. [Online]. Available: https://doi.org/10.4103/jwas.jwas_57_22.

[31] K. Khoza-Shangase, "Occupational noise regulation and hearing conservation in african LMICs: A narrative policy and implementation review," Environmental Disease, vol. 8, no. 3, pp. 69–77, 2023. DOI: 10.4103/ed.ed_12_25. [Online]. Available: https://doi.org/10.4103/ed.ed_12_25.

[32] S. K. Kiprono, J. W. Muchunu, and J. E. Masenga, "Skin diseases in pediatric patients attending a tertiary dermatology hospital in northern tanzania: A cross-sectional study," *BMC Dermatology*, vol. 15, no. 1, p. 16, 2015. DOI: 10.1186/s12895-015-0035-9. [Online]. Available: https://doi.org/10.1186/s12895-015-0035-9.

[33] W. Ngwa et al., "Cancer in sub-saharan africa: A lancet oncology commission," *The Lancet Oncology*, vol. 23, no. 6, e251–e312, 2022. DOI: 10.1016/S1470-2045(21)00720-8. [Online]. Available: https://doi.org/10.1016/S1470-2045(21)00720-8.

[34] W. Li, L. Yu, M. Wu, J. Liu, M. Hao, and Y. Li, "Doctorgpt: A large language model with chinese medical question-answering capabilities," in *Proceedings of the 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, Shenzhen, China, 2023, pp. 186–193. DOI: 10.1109/HDIS60872.2023.10499472. [Online]. Available: https://doi.org/10.1109/HDIS60872.2023.10499472.

[35] J.-H. Jung, D. Kim, K.-B. Lee, and Y. Lee, "Performance evaluation of large language model chatbots for radiation therapy education," Information, vol. 14, no. 7, p. 397, 2023. DOI: 10.3390/info16070521. [Online]. Available: https://doi.org/10.3390/info16070521.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.