

Article

Curriculum-Aware Retrieval-Augmented Generation for Bilingual Tutoring in Low-Resource Swahili–English Secondary Schools

Innocent E. Rugemalila ¹, Wei Cai ¹, Xiaogang Zhang ¹, Chuanwei Liu ¹ and Bang Wang ^{2,*}

¹ School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China; rugemalilai@gmail.com (I.E.R.); caiw@hust.edu.cn (W.C.); xgzhang@hust.edu.cn (X.Z.); liurs@hust.edu.cn (C.L.)

² Hubei Key Laboratory of Smart Internet Technology, Huazhong University of Science and Technology, Wuhan 430074, China

* Correspondence: wangbang@hust.edu.cn

Abstract

In Tanzanian secondary education, Swahili-language-based question-answering systems currently face systemic disparities and linguistic barriers, which undermine the fairness and justice of the educational system. While Large Language Models (LLMs) offer scalable instructional support, they typically lack curriculum grounding, which causes them to perform unreliably in low-resource languages. This study introduces a Curriculum-Aware Retrieval-Augmented Generation (RAG) framework designed to be a linguistically inclusive AI tutor. The architecture combines hybrid dense–lexical retrieval, cross-encoder reranking, and metadata-based curriculum alignment to ensure factual, grade-appropriate responses. We evaluate five distinct generative models using a stratified 500-question Golden Dataset covering English, Swahili, and code-switched inputs. Findings indicate that there is a significant trade-off between scale and deployability. Although high-capacity LLMs provide useful reference performance, Qwen2.5-0.5B offers the most realistic trade-off between quality and deployability in low-resource settings. Under the proposed curriculum-aware pipeline, Qwen2.5-0.5B attains the best answer quality (F1: 32.7%), achieves strong grounding faithfulness (83.0%, validated by human evaluation), and maintains low end-to-end latency suitable for interactive classroom use (≤ 1.24 s). Notably, considering the limited size of the code-switched evaluation subset, our framework demonstrates promising capabilities in handling Swahili–English code-switched inputs, narrowing the observed performance gap between Swahili and English through improved semantic accuracy. These results provide initial empirical evidence that curriculum-aligned RAG can enable Small Language Models (SLMs) to serve as quality, safe, and sustainable educational assistants in low-resource Global South contexts.



Academic Editor: Sikha Bagui

Received: 17 December 2025

Revised: 30 January 2026

Accepted: 16 February 2026

Published: 18 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: context-aware RAG; educational chatbots; Swahili NLP; low-resource AI; curriculum alignment; small language models

1. Introduction

Artificial intelligence can change education for the better by making it available to more people and tailoring lessons to each student's needs. However, these benefits are still unevenly distributed, especially in low- and middle-income countries, where infrastructural, linguistic, and policy constraints prevent their effective adoption [1]. Secondary

education in Sub-Saharan Africa, and especially in Tanzania, is faced with systematic shortcomings, most notably a shortage of more than 270,000 trained teachers and a lack of up-to-date learning materials [2]. These shortages hit hardest in content-heavy disciplines such as history and civics, which are based on nuanced explanation and contextual analysis [3].

These structural restrictions add to the complexity of the language environment in Tanzanian classrooms. While English is still the official medium of instruction, classroom discussion in practice involves a lot of Swahili and English code-switching. This is an important pedagogical tool to scaffold understanding, elucidate abstract concepts, and make up for weaknesses in technical vocabulary [4]. Consequently, code-switching is not an accidental feature of the classroom but a basic pedagogical necessity. Educational technologies that do not address this bilingual reality risk the fact that students may be disconnected in certain pedagogical aspects or that these tools are simply not usable, especially for students who rely on Swahili to ground their conceptual understanding.

The recent progress in Large Language Models (LLMs) demonstrated high potential in the educational field, such as tutoring, the generation of explanations, and formative feedback. Their use in the low-resource and multilingual educational settings is, however, problematic. Generic LLMs are usually trained on large, English-intensive corpora and tend to produce fluent but factually incorrect answers, a phenomenon more generally known as hallucination [5,6]. In formal secondary education systems governed by national curricula and high-standard examinations, such errors pose significant pedagogical risks [7,8]. Additionally, existing LLMs are not very robust to African languages and bilingual teaching, especially Swahili–English code-switching that dominates Tanzanian classrooms, thereby compromising reliability as well as inclusivity.

Retrieval-Augmented Generation (RAG) has become an effective architectural paradigm to diminish hallucinations of large language models by basing generation on external evidence retrieved by them [9]. Being able to explicitly condition the responses based on verifiable documents, RAG offers a principled basis for enhancing the factual reliability of knowledge-intensive applications and is thus of special interest in high-standard educational environments.

Despite these developments, current RAG systems are poorly adapted to bilingual education in low-resource secondary schools. In the majority of pipelines, curricular relevance has given way to semantic similarity, which is why most pipelines tend to retrieve factually accurate content that does not meet grade-level expectations or learning objectives [10]. Moreover, existing multilingual and cross-lingual RAG systems take little to no consideration of pedagogical code-switching, in which student questions can be posed in Swahili, yet the taught materials are in English [11]. Lastly, most RAG-based applications are based on large and cloud-hosted designs, which cannot be achieved in an environment with limited connectivity and processing capacities [12].

Such constraints indicate the necessity of AI tutoring systems, which should be curriculum-based, linguistically adaptive, and computationally efficient. This paper, in turn, suggests a Curriculum-Aware Retrieval-Augmented Generation (RAG) model adapted to secondary education in Tanzania. The proposed system combines hybrid retrieval, curriculum-conscious metadata reranking, and lightweight Small Language Models (SLMs) to make sure that it is pedagogically relevant, bilingual-resilient, and practically deployable in low-resource classroom settings.

In this paper, we define Small Language Models (SLMs) as edge-deployable instruction-tuned models with fewer than 7 billion parameters, with particular emphasis on compact models in the 0.5–1 billion parameter range that are suitable for low-resource deployment.

The main aim of the proposed research is to facilitate curriculum-based, bilingual question answering in Swahili–English code-switching scenarios. In particular, the framework discusses three ongoing issues in deploying LLMs to Tanzanian secondary education, namely (i) factual unreliability as a result of hallucination, (ii) a lack of alignment with national curricula, and (iii) poor performance of cross-lingual retrieval. We have given the following contributions:

- **Bilingual and Code-Switching Support:** To address bilingual classroom realities in Tanzania, we design a Swahili–English capable question-answering architecture that integrates multilingual embeddings, cross-lingual retrieval, and language-adaptive generation. While our benchmark includes an initial code-switched subset, we present code-switching evaluation as a preliminary analysis and plan to extend the dataset with a larger proportion of mixed-language queries for stronger statistical conclusions.
- **Curriculum-Aligned Knowledge Base:** Our system builds a curriculum-based repository based on the authoritatively approved Tanzanian secondary school content, such as textbooks, national examinations, and national syllabi, expanded with pedagogic metadata.
- **Curriculum-Aware Reranking Strategy:** As an improvement to current semantic relevance-driven reranking, we suggest metadata-boosted reranking, in which pedagogically suitable evidence is prioritized through curriculum constraints.
- **Comprehensive Evaluation Framework:** We develop a curriculum-aligned benchmark dataset comprising 500 question–answer pairs in Swahili, English, and code-switched formats to evaluate retrieval quality, factual grounding, and hallucination resistance.

2. Related Work

2.1. Retrieval-Augmented Generation and Educational Question Answering

Large Language Models (LLMs) have shown great performance in knowledge-intensive tasks such as question answering, summarization, explanation generation, etc. Despite these capabilities, they are still weak to the point where they tend to hallucinate, giving fluent but nevertheless factually erroneous responses, especially when running without explicitly being grounded in any verifiable sources [13,14]. In contexts of education, where the accuracy of facts, the clarity of concepts and the validity of pedagogy are very critical, these types of behavior tend to carry with them significant risks to education outcomes and learner trust.

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [9], addresses this limitation by coupling generative models with external document retrieval. By integrating retrieved evidence into the propensity to generate conversations, RAG systems make things more factual and facilitate the traceability of responses to authority. Subsequent studies have demonstrated that a combination of dense semantic representations and sparse lexical matching models such as BM25 have delivered increased retrieval robustness and stay, boosting downstream response quality [15,16]. Complementary approaches, such as representation editing and decoding with confidence, further decrease the hallucination rate, adding to the power of retrieval-based grounding [17].

RAG architectures have been successfully applied in specialized, high-resource domains such as biomedical question answering [18] and legal reasoning [19]. In these areas, the existence of domain-specific corpora, standardized terminology, and a relatively homogeneous user population enables targeted retrieval and reliable generation. However, such architectures transferred directly to educational context, and especially to the secondary school level, present specific challenges that in the literature are not addressed adequately.

Recent studies show the potential of RAG-based tutors for assisting individualized learning, formative feedback and student engagement [20,21]. Nevertheless, most of the systems that have been reported remain proof-of-concept prototypes or are tested in controlled and high-resource environments. These are usually insensitive to classroom realities such as the lack of connectivity, heterogeneous language proficiency of the learner, and the requirement of alignment with formal assessment frameworks.

2.2. Large Language Models for African Low-Resource Languages

Despite significant progress in large-scale language modeling, large gaps in performance still exist based on language resource availability. African languages, such as Swahili, are not well represented in the pretraining data of most foundation models, resulting in lexical coverage and syntactic modelling deficits and gaps in the fidelity of reasoning [22,23]. These limitations are especially important in educational applications because inaccuracies, mistranslations, or shallow explanations can have a direct impact on learning outcomes.

Recent multilingual benchmarks highlight these gaps. Evaluations like IrokoBench and AfroBench show that state-of-the-art LLMs show markedly reduced accuracy, higher rates of hallucinations, and poorer reasoning abilities under a direct test in African languages when compared directly to English or other languages that are high-resource [22,24]. Importantly, many benchmark datasets are based on machine-translated evaluation sets, where linguistic artifacts introduce and obscure true model competence, leading to overestimation of true model ability and underreporting of true model shortfalls [25]. This evaluation bias means that the reliability of the results reported is limited and the assessment of educational suitability is made more difficult.

Several attempts have been made to reduce these difficulties by language-specific or regional models. SwahBERT demonstrates that pretraining the Swahili language on curated Swahili corpora enhances downstream performance on classification and named entity recognition tasks [26]. More recently, models such as Serengeti [27] and LugaLLaMA [28] have taken models with a multilingual foundation and trained them on data in African languages to facilitate improvement in fluency and general comprehension. While these approaches represent meaningful progress, they address mainly language understanding and conversational ability in general and are not focused on structured educational reasoning.

The development of AI systems for contexts in low-resource education is characterized by intertwined challenges in terms of linguistic, infrastructural, and computational issues, often underestimated also in mainstream research. Languages such as Swahili are morphology-rich and agglutinative, which makes tokenization and representation learning more difficult. Standard subword tokenization methods (e.g., BPE, WordPiece) tend to break up semantically coherent morphemes, producing sparse or distorted word embeddings, undermining retrieval accuracy and the ability to reason [24,26]. These effects are particularly pronounced in curriculum-grounded question answering, where precise terminology and factual consistency are essential.

Existing Swahili-focused question answering systems remain limited in both scope and depth. Most rely on shallow retrieval strategies or keyword-based matching and lack mechanisms for curriculum alignment, pedagogical sequencing, or bilingual robustness [4,29]. As a result, these systems often fail to support the explanatory depth, factual consistency, and syllabus fidelity required for secondary-level education.

This highlights a broader gap in African-language educational NLP, where linguistic coverage alone is insufficient without explicit integration of pedagogical and curricular constraints.

Educational deployment constraints are another factor that makes the use of large-scale models more difficult. In Tanzania, sustained teacher shortages, low access to current

learning content, and low levels of internet connectivity limit the practicality of cloud-based inference [2,30]. While large proprietary LLMs provide great performance, they are expensive and have latency and connectivity requirements that make them impractical to use in a widespread classroom setting in a low-resource setting. As a consequence, there is an increased interest in Small Language Models (SLMs) and edge-compatible architectures that can be used offline or under limited computational budgets [31].

Despite this interest, empirical evaluations on the use of SLMs for multilingual educational question-answering activities—especially in African contexts—are sparse. Most of the studies evaluate linguistic skill or computational efficiency separately and do not look at the interaction of smaller models with retrieval systems, curriculum constraints, or bilingual pedagogic practices. This was exacerbated by the need for multi-embracing approaches with combined multilingual retrieval and grounding for curriculum-sensitive but lightweight generation models that are recognized for African secondary education.

2.3. Curriculum Misalignment

A key shortcoming of current educational AI technologies is that they are not highly connected with formal curricula. Most retrieval-augmented generation (RAG) pipelines have open domain knowledge sources such as Wikipedia, Common Crawl, or web-scale instructional content. While there is wide-ranging factual coverage in such sources, they are not organized along the lines of national syllabi, grade level expectations, and pedagogical sequence. As a result, systems often get content that is academically correct but instructionally inappropriate, resulting in explanations that are beyond the learners' cognitive readiness or that stray from examination-oriented learning goals [32].

This phenomenon has been well documented in cognitive load theory, which emphasizes that learning effectiveness depends not only on factual accuracy but also on the alignment between instructional content and learners' developmental stage and prior knowledge [33]. When students are exposed to off-syllabus or overly advanced explanations—common in retrieval systems that prioritize semantic similarity alone—this introduces extraneous cognitive load that can hinder comprehension and disrupt instructional coherence.

A fundamental drawback of many educational RAG systems is therefore their lack of explicit awareness of formal curricula. While such systems may retrieve factually correct information, they often fail to enforce pedagogical constraints such as grade appropriateness, subject scope, and assessment relevance. As a result, retrieved content may be academically accurate yet instructionally inappropriate, particularly in secondary education, where structured progression and syllabus fidelity are central to learning outcomes.

A major limitation of current LLMs for African languages is their ignorance of the curriculum. Educational tasks require more than linguistic fluency and demand correspondence to grade-level expectations, a degree of subject-specific terminology, and pedagogical order. Generic multilingual models, even those that have been adapted to Swahili, can often default to Anglo-centric knowledge representations and do not respect the instructional logic that is embedded in the national curricula [34]. This disconnect results in responses that are linguistically coherent but pedagogically misaligned, particularly in examination-oriented subjects such as History, Civics, and Biology.

Recent studies have started to try and overcome this limitation where curriculum awareness is taken into account during retrieval and reranking stages. Curriculum-aware RAG approaches augment document representations with rich pedagogical metadata about them, such as subject domain, grade level, topic hierarchy and learning objectives. Liu et al. [35], for example, propose a curriculum-guided reranking strategy, where metadata-aware scoring is combined to prioritize the passages aligned with instructional

goals. Their results show benefits of improved relevance and lower retrieval noise in educational settings under control.

Similarly, Amarnath et al. [20] and Swacha et al. [36] highlight that most existing educational RAG systems remain proof-of-concept prototypes, often limited to single subjects or higher education contexts. Such systems usually do not use common ways of encoding the structure of the curriculum, and most do not explicitly incorporate pedagogical taxonomies such as the cognitive levels of Bloom's taxonomy into the retrieval pipeline. Because of this, the possibility of scalability to national secondary curricula (especially in multilingual and low-resource settings) is limited.

Multi-hop and syllabus-aware retrieval approaches have also been explored to handle complex questions that are present in examinations and often relate to multiple concepts. Iterative evidence retrieval but also conceptual grounding, with the use of iterative alignment-based reasoning strategies, have shown promise in improving the completeness of answers [37,38]. However, these approaches tend to have a high computational burden and require manually annotated training data, which is not always available or convenient to deploy in educational settings constrained by a limited availability of computational resources. Moreover, they pay little attention to linguistic diversity and code-switching, thus further restricting their applicability in, e.g., Tanzanian classrooms.

A critical gap remains on the bringing together curricular conscientiousness and multilingual and cross-lingual recall. Existing curriculum-aware systems take a preponderantly monolingual approach to input and retrieval, most commonly of the English variety. In bilingual educational settings, where student questions can be asked in Swahili but curriculum materials and assessments are written in English, the lack of curriculum constraints given by language results in additional atrophies of retrieval relevance. Without joint consideration of language, curriculum metadata, and pedagogical intent, retrieved evidence is at risk of being linguistically inaccessible as well as instructionally misaligned.

Taken together, the literature suggests that curriculum misalignment is a major challenge to the successful implementation of RAG systems in formal education. While recent efforts have provided the foundation for curriculum-based retrieval to become feasible, existing solutions are limited for the added challenges of multilinguality, low-resource deployment, and secondary-level pedagogy. These limitations are a motivation for the need for curriculum-aware RAG architectures that at once model the pedagogical structure, unitary of linguistic diversity, given resource constraints—an approach that this work seeks to promote in the context of Tanzanian secondary education.

2.4. Low-Resource AI Development Challenges and Cross-Lingual Retrieval

Low-resource bilingual classrooms introduce additional challenges for retrieval pipelines beyond language modeling, particularly due to pedagogical code-switching and cross-lingual evidence alignment. In the Tanzanian classrooms at the secondary level, Swahili–English code-switching is not incidental to the class but serves a pedagogical purpose that facilitates conceptual understanding in Swahili while introducing technical terminology in English [39]. Conventional multilingual NLP pipelines generally take code-switching as noise or an edge case, rather than or in addition to a structured linguistic phenomenon. As a result, retrieval systems often cannot accurately interpret mixed language queries, either failing to retrieve documents on topic or, by default, they produce monolingual outputs that interfere with instructional continuity.

Cross-lingual and multilingual retrieval has emerged as a key strategy for addressing language asymmetries in bilingual educational environments. In Tanzanian secondary schools, students frequently pose questions in Swahili, while textbooks, examinations, and official instructional materials are predominantly authored in English. Effective retrieval

therefore requires aligning queries and evidence across languages without disrupting pedagogical intent.

Recent benchmarks in multilingual reasoning, such as X-COPA, demonstrate that semantic representations can be aligned across languages to support cross-lingual inference and contextual grounding [40]. In the context of retrieval-augmented generation, pipelines such as XRAG incorporate language-agnostic embeddings and translation-aware reranking to bridge language gaps between queries and documents [41]. While these methods show measurable improvements in retrieval recall and answer relevance under controlled conditions, they are rarely designed with education-specific constraints in mind and typically do not incorporate curriculum metadata or pedagogically meaningful language-switching patterns.

However, existing cross-lingual RAG systems are still overwhelmingly optimized for high-resource tasks and open-domain tasks. They assume usually balanced bilingual corpora, stable connectivity and sufficient computational resources. In contrast, Tanzanian learning environments are marked by asymmetries in language distribution (the instructional material and test items are mostly in English, yet student questions are often in Swahili), sporadic connectivity and low availability of hardware. Under these constraints, translation-heavy or multi-stage retrieval pipelines lead to latency and the propagation of error that undercuts usability in the real classroom.

Infrastructure limitations further constrain deployment. Many rural and peri-urban schools lack reliable internet access, making cloud-based inference impractical for daily instructional use [30]. This has motivated growing interest in edge-compatible Small Language Models (SLMs) that can operate offline or with minimal connectivity. Recent studies advocate for “offline-first” AI architectures as a means of promoting equitable access and reducing dependency on centralized infrastructure [31]. While SLMs offer advantages in efficiency and deployability, their reduced parameter capacity often exacerbates hallucination and reasoning limitations when used in isolation.

The literature suggests that retrieval augmentation is especially useful in addressing some of these weaknesses so that the responses of smaller models can be grounded in curated external sources as a means of compensating for the lack of internal knowledge. Nevertheless, empirical assessments of SLM-based RAG systems in multilingual, curriculum-based educational settings are quite rare. Most work to date evaluates either linguistic performance or computational efficiency independently, without studying the interplay of cross-lingual retrieval, the curriculum constraints, and lightweight generation production in realistic classroom scenarios.

3. Proposed Methodology

3.1. Framework Overview

This section presents the proposed curriculum-aware bilingual RAG methodology and the controlled experimental design used to test it. Specifically, we evaluate whether a curriculum-sensitive Swahili–English RAG pipeline improves (i) factual faithfulness, (ii) pedagogical relevance, and (iii) linguistic suitability for Tanzanian secondary-school question answering under low-resource constraints. Explicitly, the framework focuses on (i) the curriculum fidelity (subject/topic/grade alignment), (ii) multilingual and code-switched classroom discussion (Swahili–English) and (iii) a deployment-feasible use of Small Language Models (SLM).

3.1.1. Small Language Models (SLMs)

Since most schools in Tanzania have limited computing infrastructures, the framework emphasizes lightweight instruction-tuned SLM that can either operate on relatively small

hardware or be deployed with a low cost. Although SLMs are efficient, they can be easily defeated by large proprietary models in the domain of parametric recall and reasoning. To reduce this, the framework increases SLM generation with curriculum-based evidence that is retrieved in an officially approved knowledge base.

3.1.2. Curriculum-Aware Retrieval and Reranking

The content retrieved in open-domain retrieval systems is often topically relevant yet does not match with the Tanzanian syllabus, grade level or classroom language. The proposed pipeline will, as such, consist of the following: (i) hybrid dense–lexical retrieval to enable a good recall, and (ii) cross-encoder reranking with curriculum metadata (subject, grade, and language) to eliminate the off-syllabus drift and give preference to age-related evidence.

3.1.3. Efficiency Objectives

The design supports low-resource deployment with the retrieval and generation capabilities being modular (to allow lightweight upgrades) in nature and reranking limited to a manageable pool of candidates. Large proprietary models (e.g., GPT-4o) are only based on benchmarking and are not assumed during deployment.

3.1.4. Research Hypotheses

We test three hypotheses (H1–H3) using a controlled comparison across baselines that isolate the contribution of curriculum-aware reranking, hybrid retrieval, and pedagogically guided prompting (Section 3.2).

The overall architecture is illustrated in Figure 1. The end-to-end curriculum-aware question answering procedure is summarized in Algorithm 1, which integrates hybrid retrieval, metadata-guided reranking, and pedagogically constrained generation into interpretable and modular steps.

Algorithm 1 Curriculum-Aware RAG Pipeline

Require: Query Q , vector index $\mathcal{V}$, BM25 index $\mathcal{B}$, weight α , metadata weights w , initial pool size N , output size k , generator model $\mathcal{G}$, prompt template $\mathcal{P}$

Ensure: Grounded answer A and top- k curriculum-aligned chunks $\mathcal{C}_{\text{top-}k}$

```

1: Compute dense embedding  $\mathbf{q} \leftarrow \text{Embed}(Q)$ 
2:  $D_{\text{dense}} \leftarrow \text{DenseSearch}(\mathcal{V}, \mathbf{q})$ 
3:  $D_{\text{lex}} \leftarrow \text{BM25Search}(\mathcal{B}, Q)$ 
4:  $D_m \leftarrow D_{\text{dense}} \cup D_{\text{lex}}$ 
5: // Stage 1: Normalize & Fuse
6: Normalize scores in  $D_m$  via Min–Max scaling
7: Compute hybrid scores  $\text{Score}_{\text{Hybrid}}(Q, d)$  using Equation (4)
8:  $\text{Candidates} \leftarrow \text{TopN}(D_m, N \text{ by } \text{Score}_{\text{Hybrid}})$  (TopN returns the top  $N$  documents ranked by hybrid retrieval score.)
9: // Stage 2: Cross-Encoder Reranking & Curriculum Boost
10: for each  $d \in \text{Candidates}$  do
11:    $S_{\text{logits}} \leftarrow \text{CrossEncoder}(Q, d)$ 
12:    $S_{\text{sem}} \leftarrow \sigma(S_{\text{logits}})$  {Sigmoid activation}
13:    $S_b \leftarrow \text{MetadataMatch}(Q, d, w)$ 
14:    $S_f(d) \leftarrow S_{\text{sem}} + S_b$ 
15: end for
16:  $\mathcal{C}_{\text{top-}k} \leftarrow \text{TopK}(\text{Candidates}, k \text{ by } S_f)$  (TopK returns the top  $k$  candidates ranked by final reranking score.)
17: // Stage 3: Curriculum-Grounded Generation
18:  $C \leftarrow \text{ConcatChunks}(\mathcal{C}_{\text{top-}k})$  {Build context window}
19:  $\text{prompt} \leftarrow \mathcal{P}(Q, C)$  {Apply pedagogical + grounding instructions}
20:  $A \leftarrow \mathcal{G}(\text{prompt})$  {Generate syllabus-aligned answer}
21: return  $A, \mathcal{C}_{\text{top-}k}$ 

```

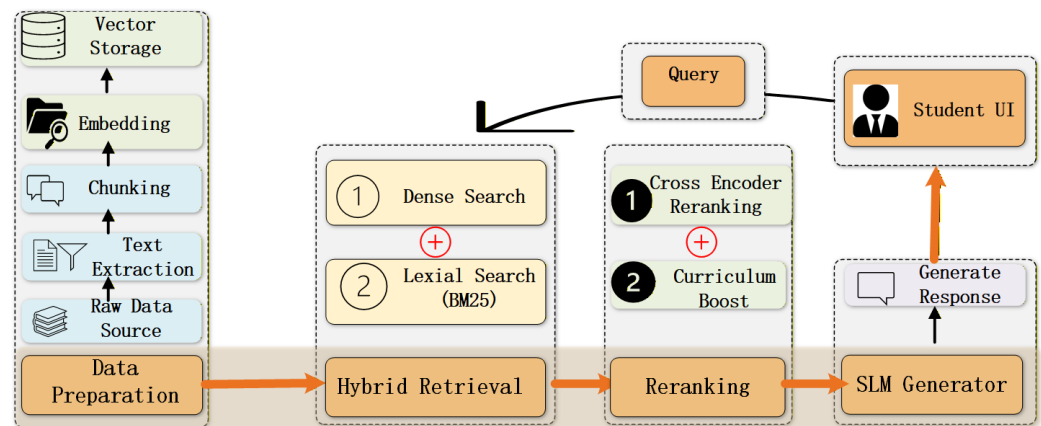


Figure 1. Curriculum-aware RAG system architecture illustrating knowledge-base building, hybrid retrieval, curriculum-aware reranking, and pedagogically guided generation.

3.2. Experimental Design and Evaluation Setup

The research follows a controlled, comparative experimental research design to estimate whether curriculum-consciousness retrieval and reranking enhance (i) the significance of evidence, (ii) response fidelity to retrieved evidence and (iii) didactic adequacy in multilingual (English/Swahili) and code-switched classroom searching. All experiments only change the retrieval/reranking strategy but keep the knowledge base, embedding model, generator, and decoding settings unchanged in order to be interpretable and to render the source of reported results visible.

3.2.1. Evaluation Objectives and Hypotheses

The experiments are organized around the following hypotheses:

- H1 (Curriculum alignment): Introducing curriculum metadata into evidence selection (via reranking/boosting) improves syllabus alignment and reduces off-curriculum retrieval compared to naive RAG.
- H2 (Hybrid retrieval): Dense–lexical fusion improves recall of curriculum-relevant evidence for multilingual and code-switched queries compared to dense-only retrieval.
- H3 (Pedagogical prompting): Pedagogically guided prompting improves readability and instructional appropriateness without degrading evidence-grounded faithfulness.

3.2.2. Experimental Conditions (Systems Compared)

To isolate the contribution of each component, we evaluate the following systems:

1. No-Retrieval Generation (LLM/SLM only): Direct generation without external evidence, serving as a lower-bound baseline for faithfulness.
2. Naive RAG (Dense-only): Dense retrieval with semantic similarity only, without curriculum metadata, reranking, or pedagogical prompt constraints.
3. Hybrid RAG (Dense + BM25): Dense–lexical fusion retrieval using score normalization and weighted combination, without curriculum-aware reranking.
4. Curriculum-Aware RAG (Proposed): Full pipeline including hybrid retrieval, cross-encoder reranking, and curriculum-aware metadata boosting, followed by pedagogically guided generation.

All systems use the same curriculum-aligned knowledge base constructed in Section 3.3. The retrieval pipeline and reranking formulations are specified in Sections 3.4 and 3.5, and the generation procedure is defined in Section 3.6. The overall architecture is illustrated in Figure 1, and the end-to-end procedure is summarized in Algorithm 1.

3.2.3. Controlled Variables and Fixed Hyperparameters

Across all experimental conditions, the following variables are held constant to ensure fair comparison:

- Knowledge base and chunking: identical corpus, preprocessing, chunk size and overlap strategy.
- Embedding model: BAAI/bge-m3 for dense representations (fixed).
- Retriever parameters: BM25 parameters and fusion weight α (fixed).
- Candidate pool sizes: retrieval pool size N and final context size k (fixed).
- Generator model: the same SLM backend for all RAG variants unless explicitly stated.
- Decoding configuration: temperature, max tokens, and other decoding parameters (fixed).

Unless otherwise stated, the primary experiments use Qwen2.5-0.5B-Instruct as the deployed SLM, while GPT-4o is used only as an evaluator (LLM-as-a-judge) and as a high-capacity reference baseline.

3.2.4. Task Definition and Evaluation Slices

All systems are evaluated on curriculum-aligned question answering derived from Tanzanian secondary education resources. To reflect classroom linguistic reality, evaluation is reported on three language slices:

- English-only queries,
- Swahili-only queries,
- Swahili–English code-switched queries.

Results are reported both overall and per language slice to make performance differences attributable to linguistic variation explicit rather than averaged away.

3.2.5. Metrics Signposting (Definitions in Evaluation Protocol)

Performance is assessed along three complementary dimensions:

- Retrieval Quality: whether the system retrieves curriculum-relevant evidence (e.g., Recall@k, Hit Rate, NDCG).
- Answer Faithfulness: the extent to which generated responses are supported by the retrieved evidence, computed via a defined automatic protocol and selectively validated by human annotation.
- Pedagogical Suitability: curriculum/grade appropriateness and readability (e.g., ARI), reflecting instructional usability.

Precise metric definitions, faithfulness scoring procedures (automatic vs. human), and evaluation bias controls are provided in the Evaluation Protocol (Section 3.7). Importantly, evaluation models (when used) are kept distinct from generation models to mitigate self-preference bias, and a human-validated subset is used to confirm the reliability of automatic faithfulness scoring.

3.3. Curriculum-Aligned Dataset Construction

To ensure educational validity and syllabus relevance, the knowledge is built solely on the Tanzanian materials covered by the secondary school syllabus, which has received official approval from the Ministry of Education. The sources of content used consist of textbooks (Forms 1–4), the national syllabus, and examinations used in NECTA. The initial implementation is done in text-based subjects, where the majority of learning outcomes can be described in the form of natural language instead of diagrams or sets of numbers.

The processing workflow for all documents follows a standardized approach: (i) noise such as scanned artifacts, watermarks, and repetitive headers and footers is removed;

(ii) Optical Character Recognition (OCR) is applied where necessary; and (iii) sentence boundaries are rectified. Semantically coherent text chunks are then partitioned using a sliding-window method with overlap (stride of 40–60 tokens) to maintain the contextual coherence of ideas across surrounding sections [42].

Every chunk is marked with three central metadata that are needed in curriculum-imposed retrieval:

- Grade level: Forms 1–4 mapping based on the official syllabus.
- Subject-topic code: A hierarchical taxonomy (subject → topic → subtopic) that reflects the curriculum structure.
- Cognitive demand: Bloom’s Taxonomy category (e.g., Knowledge, Comprehension, Application, Analysis), which supports pedagogical alignment during reranking [43].

The final curated dataset comprises 16,390 chunks across four subjects, summarized in Table 1. All metadata is indexed alongside the text and embeddings, enabling controlled filtering and curriculum-aware score boosting during retrieval.

Table 1. Curriculum-aligned corpus statistics used in the context-aware RAG system.

Subject	Forms	Pages	Chunks	Language
History	1–4	890	4380	en/sw
Civics	1–4	940	4890	en/sw
Geography	1–4	1100	4220	en/sw
Literature	1–4	720	2900	sw/en
Total	–	3650	16,390	–

For the core retrieval experiments reported in Section 4, a subset of 4380 History chunks (Forms 1–4) is used to construct a domain-focused knowledge base aligned with the History syllabus and associated past papers, as illustrated in Figure 2.

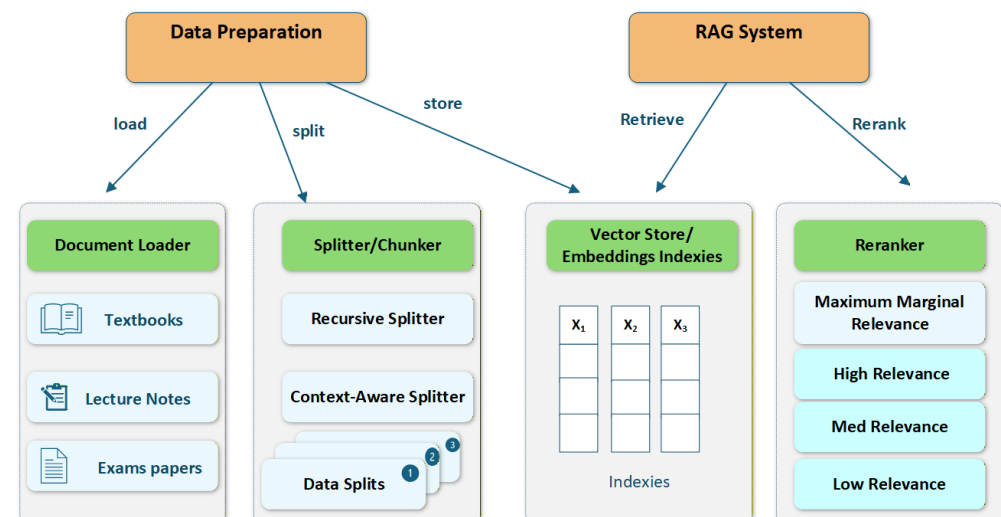


Figure 2. Curriculum-aligned data pipeline with metadata-annotated knowledge base.

3.4. Curriculum-Aware Hybrid Retrieval

To deliver syllabus-aligned and linguistically inclusive answers, the framework follows a multilingual, metadata-aware Retrieval-Augmented Generation (RAG) pipeline composed of hybrid dense–lexical retrieval and curriculum-guided reranking.

3.4.1. Query Preprocessing and Language Detection

When a student query q is submitted to the system, the first step is to identify the primary language of the query (either English or Swahili). This ensures proper tokenization and multilingual encoding, which are essential for accurate embedding generation and support the code-switching contexts commonly found in Tanzanian classrooms.

3.4.2. Hybrid Dense–Lexical Retrieval

Each curriculum chunk d in the knowledge base is embedded using the multilingual BAAI/bge-m3 model [44], selected for its cross-lingual alignment across 100+ languages, including strong performance on low-resource African languages such as Swahili. The input query q is similarly encoded into a dense vector representation $\mathbf{q}$. Cosine similarity is used to compute dense semantic similarity between the query and document embeddings $\mathbf{q}$ and $\mathbf{d}$:

$$\text{Score}_{\text{Dense}}(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|}. \quad (1)$$

To complement dense retrieval, BM25 scoring is used to boost sensitivity to syllabus-aligned terminology and named entities. For a query q consisting of terms t , and a document d , the BM25 score is calculated as:

$$\text{Score}_{\text{BM25}}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, d)(k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}, \quad (2)$$

where $f(t, d)$ is the frequency of term t in document d , $|d|$ is the document length, and avgdl is the average document length. Parameters $k_1 = 1.2$ and $b = 0.75$ are set empirically.

3.4.3. Score Normalization and Fusion

Because BM25 and cosine similarity are on different scales, scores are normalized using Min–Max Normalization over the top retrieved batch $\mathcal{D}_m$:

$$\tilde{S}(d) = \frac{S(d) - \min_{d' \in \mathcal{D}_m} S(d')}{\max_{d' \in \mathcal{D}_m} S(d') - \min_{d' \in \mathcal{D}_m} S(d')}. \quad (3)$$

Final retrieval scores are computed using weighted hybrid fusion:

$$\text{Score}_{\text{Hybrid}}(q, d) = \alpha \cdot \tilde{S}_{\text{Dense}}(q, d) + (1 - \alpha) \cdot \tilde{S}_{\text{BM25}}(q, d), \quad (4)$$

where $\alpha = 0.7$ prioritizes semantic similarity while preserving lexical relevance.

3.5. Curriculum-Aware Cross-Encoder Reranking

Following hybrid retrieval, the top-ranked candidates are passed to a reranker that integrates deep semantic and curriculum-aware signals. To ensure pedagogical relevance and semantic precision, the candidate evidence chunks retrieved by the hybrid dense–lexical retriever are refined through a two-part reranking mechanism:

- A multilingual cross-encoder computes token-level semantic relevance between each query–document pair.
- A curriculum-aligned boosting function prioritizes evidence that matches key pedagogical metadata.

Given a user query q and a retrieved document chunk d , a cross-encoder model (e.g., ms-marco-MiniLM-L6-v2) computes a logit similarity score $S_{\text{logits}}(q, d)$ from the joint query–document input. These logits are then mapped into probability space $[0, 1]$ using a sigmoid function:

$$\sigma(S_{\text{logits}}(q, d)) = \frac{1}{1 + e^{-S_{\text{logits}}(q, d)}}. \quad (5)$$

3.5.1. Curriculum-Aware Boosting Function

To incorporate pedagogical constraints in the reranking process, we introduce a deterministic metadata-based boost $S_b(q, d)$, which rewards correspondence between the query and document chunk based on curriculum metadata. It is computed as follows:

$$S_b(q, d) = w_{\text{subj}} \cdot \mathbb{I}(d_{\text{subj}} = q_{\text{subj}}) + w_{\text{grade}} \cdot \mathbb{I}(d_{\text{grade}} = q_{\text{grade}}) + w_{\text{lang}} \cdot \mathbb{I}(d_{\text{lang}} = q_{\text{lang}}), \quad (6)$$

where

- $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the metadata match and 0 otherwise.
- $d_{\text{subj}}, d_{\text{grade}}, d_{\text{lang}}$ are the subject, grade level, and language metadata of the chunk d .
- $q_{\text{subj}}, q_{\text{grade}}, q_{\text{lang}}$ are the same metadata fields inferred from the user query q .
- $w_{\text{subj}}, w_{\text{grade}}, w_{\text{lang}}$ are hyperparameters controlling the influence of each metadata constraint.

In our implementation, we prioritize subject alignment ($w_{\text{subj}} = 0.2$) to avoid domain drift while allowing partial flexibility in grade level ($w_{\text{grade}} = 0.1$). Language matching ($w_{\text{lang}} = 0.1$) ensures clarity for Swahili-dominant learners.

3.5.2. Final Reranking Score

The final score used to rerank the evidence is the sum of the normalized semantic relevance and the curriculum boost, and the top- k ranked chunks from this scoring function are passed to the language model generator.

$$\text{Score}_{\text{Final}}(q, d) = \sigma(S_{\text{logits}}(q, d)) + S_b(q, d). \quad (7)$$

3.6. Pedagogically Guided Generation

The final phase in the pipeline converts the most interesting curriculum-related passages into structured and pedagogically sound responses using a SLM such as Qwen2.5-0.5B-Instruct or TinyLlama-1.1B, selected for its edge-deployable capabilities and multilingual functionality.

3.6.1. Structured Prompting and Grounded Answer Generation

To achieve factual anchoring and reduce the occurrence of the hallucinations, the evidence fragments that have been retrieved are integrated into the prompts and specific instructions are given such that the model will only respond to what has been given to it. Additionally, pedagogical clues consist of the Taxonomy levels of Bloom and a formal tone that should be used in examinations, which are also provided in the prompts. Lightweight safety filters are installed to confirm the existence of evidence and compliance to language limitations in output, as well as prevent unwarranted reasoning because of the lack of the relevant information.

3.6.2. Language-Aware Prompt Adaptation

The prompts are dynamically adjusted to offer Swahili–English language assistance on the input query of the student. Language detection is a component that coordinates the answer to the query profile to the language: a query in the English language is answered with an English response, a query in Swahili gets a Swahili response, and code-switching patterns are preserved in a code-switched query. This also assists in minimizing the hurdles to comprehension and keeping the course on schedule.

3.6.3. Instructional Scaffolding and Chain-of-Thought Support

Timely templates, as shown in Figure 3, are used to incorporate instructional scaffolding to suit the cognitive level of the learner, based on the Bloom's Taxonomy category assigned to the query. For example, prompts addressing questions in the Analysis category guide the model to paraphrase key ideas, compare relevant concepts, and generalize conclusions. This chain-of-thought prompting fosters more structured, exam-level responses that align with Tanzanian standards of education. Where no fine-tuning is required, structured prompts combined with the curriculum context enable SLMs to generate high-quality, interpretable responses.

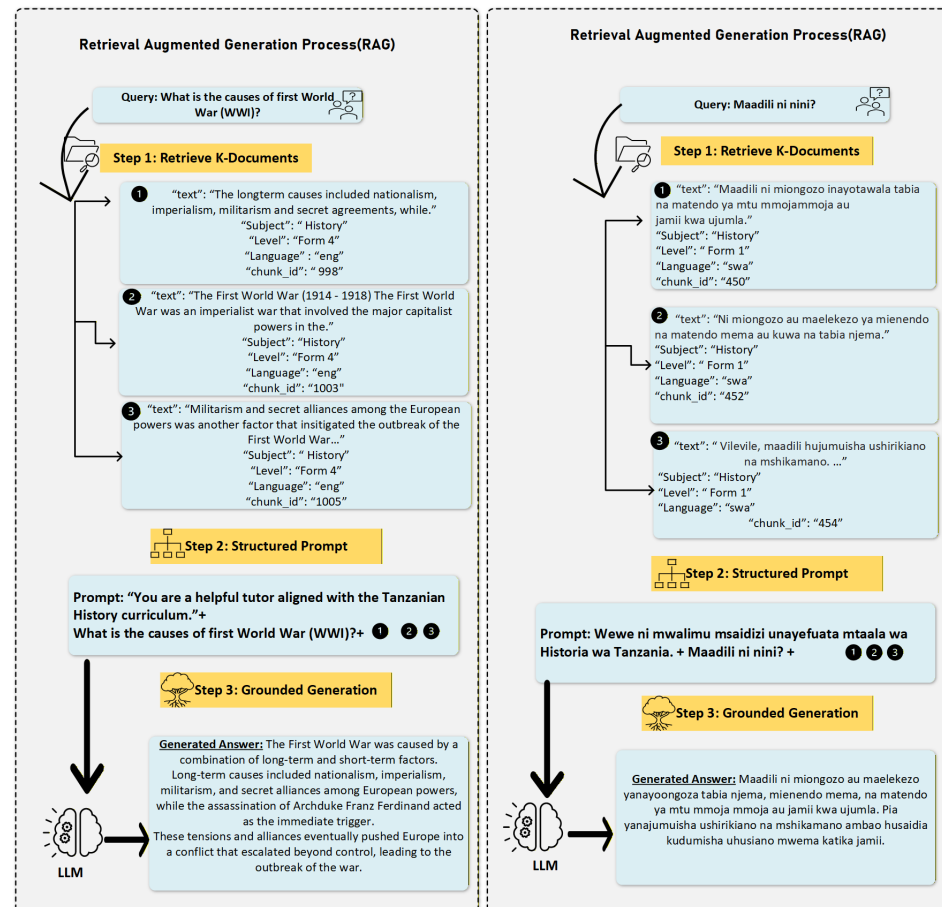


Figure 3. Illustrative prompt templates in Swahili and English incorporating curriculum-aligned context, instructional scaffolding, and grounding constraints.

3.7. Evaluation Protocol

The evaluation is performed based on a mix of retrieval-level, generation-level, and pedagogical metrics of the system's performance to give an overall clear-cut and organized assessment of system performance. An evaluation protocol distinctly isolates retrieval quality, and it includes bias controls to address self-preference on the part of the evaluator.

3.7.1. Retrieval Metrics

Retrieval quality is measured independently of generation using:

- Recall@k: Proportion of questions for which at least one gold-relevant chunk appears in the top- k retrieved evidence.
- NDCG@k (Normalized Discounted Cumulative Gain): Evaluates the ranking quality by accounting for the position and graded relevance of retrieved chunks.

Gold relevance labels are derived from syllabus mappings and expert annotation.

3.7.2. Faithfulness and Grounding Metrics

Answer faithfulness is evaluated using a hybrid framework combining automatic and human assessment:

- **Automatic Evaluation:** Faithfulness is computed using a RAG-aware evaluation framework inspired by RAGAS, where each generated answer is decomposed into atomic claims and checked for support in the retrieved evidence.
- **Human Evaluation:** A subset of 120 responses (40 per language category) is independently evaluated by two Secondary School teachers (annotators) with curriculum familiarity. Inter-rater agreement is measured using Cohen's Kappa.

Faithfulness scores are reported as the percentage of supported claims over total claims.

3.7.3. Pedagogical Appropriateness

Pedagogical quality is assessed using:

- **Curriculum Alignment Score:** Binary judgment of whether the answer matches the expected grade level and syllabus scope.
- **Readability Index (ARI):** Measures age-appropriate language complexity.

These metrics capture instructional suitability beyond surface-level fluency.

3.7.4. Bias Mitigation in Evaluation

To prevent self-preference bias, the model used for automatic evaluation is strictly distinct from the generation models. In particular:

- GPT-4-class models are not used to evaluate outputs generated by GPT-based systems.
- Automatic scores are cross-validated against human judgments, and correlation coefficients are reported.

This separation ensures that faithfulness gains are not artifacts of evaluator bias. Unless otherwise stated, the primary experiments use small open-source generator models (TinyLlama-1.1B and Qwen2.5-0.5B) to reflect deployment constraints, while GPT-4o is used only as an evaluator in the LLM-as-a-Judge faithfulness protocol and as a high-capacity reference baseline.

3.7.5. Statistical Reporting

All reported scores are averaged across the full evaluation set unless otherwise stated. For language-specific analysis, results are reported separately for English, Swahili, and code-switched subsets. Differences between systems are analyzed using paired comparisons, with emphasis on effect size rather than sole reliance on significance testing.

3.7.6. Reproducibility

All hyperparameters, metadata weights, prompt templates, and evaluation scripts are fixed prior to experimentation. No component is tuned on the evaluation set. This design ensures that observed performance differences arise from architectural choices rather than experimental leakage.

4. Results

4.1. Results Overview, Context, and Traceability

This section reports empirical results for the proposed Curriculum-Aware RAG framework under a controlled and consistent evaluation setting. Unless otherwise stated, all

scores are computed on the History Golden QA dataset ($N = 500$) using the experimental conditions defined in Section 3.2.

We evaluate five language models spanning small open-source SLMs and larger comparative baselines (TinyLlama-1.1B, Qwen2.5-0.5B, Gemma-2-2B, GPT-2, and GPT-4o). Each model is tested under three retrieval configurations: No-RAG, Naive RAG (dense-only retrieval), and the proposed Curriculum-Aware RAG pipeline.

Performance is assessed using retrieval-level, generation-level, and pedagogical metrics, including HitRate (Retrieval Accuracy), Recall@K, NDCG@5, token-level F1, faithfulness to retrieved evidence, readability, and end-to-end latency. Metric definitions and evaluation procedures are provided in Section 3.7. To support traceability and reproducibility, Table 2 links each results block to the corresponding dataset subset, system configuration, and evaluation metric(s).

Table 2. Traceability map linking each results block to its dataset subset, system configuration, and evaluation metric(s).

Results Block	Dataset/Split	Systems Compared	Metric(s) Reported
Main Results (Table 3)	History Golden QA ($N = 500$)	No-RAG vs. Naive RAG vs. Curriculum-Aware RAG (all models)	HitRate (Retrieval Accuracy), F1, Faithfulness, Latency
Lexical Generation Quality (Table 4)	History Golden QA ($N = 500$)	No-RAG vs. Naive RAG vs. Curriculum-Aware RAG (all models)	F1, BLEU, ROUGE-L
Retrieval Ranking Quality (Table 5)	History Golden QA ($N = 500$)	Naive RAG vs. Curriculum-Aware RAG (retriever-level)	Recall@1/5/10, NDCG@5
Architectural Ablation (Table 6, Figures 4 and 5)	History Golden QA ($N = 500$)	Naive RAG vs. Curriculum-Aware RAG (per generator)	Δ F1, Δ Faithfulness, Recall@K trends
Cross-Model Safety and Faithfulness (Figures 6 and 7)	History Golden QA ($N = 500$)	No-RAG vs. Naive RAG vs. Curriculum-Aware RAG	Faithfulness, latency trade-offs
Cross-Lingual and Curriculum-Stratified Performance (Tables 7 and 9)	Language, grade, and Bloom-level slices (Golden QA)	Best-performing configuration + slice analysis	HitRate, F1, embedding similarity
Error Taxonomy (Table 10)	Manual subset from best configuration	Error mode categorization	Outcome distribution (success/retrieval/reasoning/hallucination)
Readability and Style Control (Table 13)	Readability analysis set (same queries)	Standard RAG vs. style-controlled prompting	Automated Readability Index (ARI)

4.2. Main Results

The introduction of retrieval is used to provide external curriculum-related context to SLMs. During the No-RAG condition, faithfulness is reduced (52.7–72.0%), while fluency remains reasonable, and the answer correctness is low ($F1 \approx 10$ –12%), indicating that even advanced LLMs struggle with History questions without external curriculum knowledge. With the addition of dense retrieval (Naive RAG), retrieval accuracy improves significantly, to 67–74% with model size, and F1 increases to around 20–32%. This highlights that integrating retrieval is more effective in improving answer correctness than simply scaling model parameters.

The performances of all the language models integrated with the Curriculum-Aware RAG pipeline show improvements in retrieval accuracy that are consistent and repeatable. SLMs benefit in particular: TinyLlama-1.1B grows by 67.3% to 72.0% Retrieval Accuracy

and 69.0% to 71.3% faithfulness, as smaller resource-constrained models also gain greatly by adopting metadata-based evidence selection.

The prevalent strength of the given architecture is manifested in *Faithfulness*. Grounding the model, such as Gemma-2B and Qwen-0.5B, with curriculum context improves their performance by +5–+11% over Naive RAG, resulting in more accurate textbook answers and fewer hallucinations. Among the deployable Small Language Model (SLM) baselines, Qwen2.5-0.5B exhibits the strongest safety improvements under curriculum constraints, with faithfulness increasing from 72.67% (Naive RAG) to 83.00% (Curriculum-Aware RAG), confirming that curriculum-guided reranking and metadata filtering substantially reduce unsupported claims, even for compact generators.

Table 3. Comparison of model and retrieval pipeline configurations on the 500-question History benchmark. Retrieval Accuracy corresponds to the overall hit rate across the golden dataset.

Model	Pipeline	Retrieval Acc. (%)	F1 (%)	Faithfulness (%)	Latency (s)
Gemma-2-2b-it	No-RAG (Direct LLM Only)	-	11.43	56.33	0.53
Gemma-2-2b-it	Naive RAG (Dense: FAISS + BGE-M3)	74.00	31.51	76.00	1.08
Gemma-2-2b-it	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	78.67	32.25	80.67	1.31
Qwen2.5-0.5B	No-RAG (Direct LLM Only)	-	11.18	72.00	0.52
Qwen2.5-0.5B	Naive RAG (Dense: FAISS + BGE-M3)	74.00	31.50	72.67	1.00
Qwen2.5-0.5B	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	78.67	32.66	83.00	1.24
GPT-2	No-RAG (Direct LLM Only)	-	11.88	–	0.65
GPT-2	Naive RAG (Dense: FAISS + BGE-M3)	74.00	31.41	–	1.00
GPT-2	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	78.67	32.26	–	1.27
GPT-4o	No-RAG (Direct LLM Only)	-	10.76	–	0.56
GPT-4o	Naive RAG (Dense: FAISS + BGE-M3)	74.00	31.43	–	1.08
GPT-4o	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	78.67	32.13	–	1.29
TinyLlama-1.1B	No-RAG (Direct LLM Only)	-	23.00	70.00	3.58
TinyLlama-1.1B	Naive RAG (Dense: FAISS + BGE-M3)	67.33	19.80	69.00	3.85
TinyLlama-1.1B	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	72.00	21.30	71.33	3.94

The trends in latency are consistent with the complexity of the pipeline. Direct responses in terms of the number of LLM scales are quickest (0.52–0.65 s) and dense retrieval adds a middle overhead (1.00–1.08 s). All Curriculum-Aware RAG pipeline adds an average of an increase of 0.2–0.4 s, primarily through cross-encoder reranking. This overhead is limited to acceptable levels in classroom and tutoring deployments due to improved retrieval depth and significant safety improvements of SLMs.

Overall, these results suggest that four important insights can be made:

- Retrieval is essential for secondary school curriculum grounding and yields much larger gains than model scaling;
- The proposed Curriculum-Aware RAG provides the most reliable retrieval and strongest safety across all generators;
- Small open-source models (e.g., TinyLlama-1.1B) become viable educational models when supported by retrieval and reranking;
- Larger models (e.g., GPT-4o) still benefit from curriculum-aware architecture in retrieval accuracy and lexical precision.

These results establish Curriculum-Aware RAG as a scalable and equitable path to pedagogically reliable AI tutors, even in low-resource school settings where computational budgets are limited.

4.3. Lexical Generation Quality (BLEU and ROUGE-L Analysis)

While retrieval performance determines whether the correct supporting evidence is surfaced, the generation module must faithfully and fluently transform that information into a usable student-facing answer. To evaluate linguistic form and surface accuracy, we report three complementary lexical metrics: F1 (overlap with reference answers), BLEU, and ROUGE-L. These metrics reflect answer precision, phrase matching, and recall of key content, respectively.

Table 4 shows that integrating retrieval (Naive RAG) substantially boosts lexical accuracy across all models compared to direct LLM responses. Curriculum-Aware RAG provides an additional improvement, especially for Qwen2.5-0.5B and GPT-4o, which achieve the highest combined lexical performance. TinyLlama-1.1B, with a smaller size, has much worse fluency and content coverage, indicating the error trade-off between the size of a model and the quality of generation in real education applications.

Table 4. Performance in the 500-question History benchmark in lexical generation. The scores are provided in percentages.

Model	Pipeline	F1 (%)	BLEU (%)	ROUGE-L (%)
Gemma-2-2b-it	No-RAG (Direct LLM Only)	11.4	2.1	9.9
Gemma-2-2b-it	Naive RAG (Dense: FAISS + BGE-M3)	31.5	8.8	26.4
Gemma-2-2b-it	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	32.2	9.0	27.1
Qwen2.5-0.5B	No-RAG (Direct LLM Only)	11.2	2.1	9.8
Qwen2.5-0.5B	Naive RAG (Dense: FAISS + BGE-M3)	31.5	8.5	26.2
Qwen2.5-0.5B	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	32.7	9.3	27.6
GPT-2	No-RAG (Direct LLM Only)	11.9	2.4	10.5
GPT-2	Naive RAG (Dense: FAISS + BGE-M3)	31.4	8.6	26.1
GPT-2	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	32.3	9.2	27.3
GPT-4o	No-RAG (Direct LLM Only)	10.8	2.1	9.8
GPT-4o	Naive RAG (Dense: FAISS + BGE-M3)	31.4	8.6	26.0
GPT-4o	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	32.1	9.0	27.1
TinyLlama-1.1B	No-RAG (Direct LLM Only)	23.0	4.8	19.3
TinyLlama-1.1B	Naive RAG (Dense: FAISS + BGE-M3)	19.8	3.2	16.1
TinyLlama-1.1B	Curriculum-Aware RAG (Hybrid + Cross-Encoder + Metadata)	21.3	4.0	17.3

4.4. Retrieval Ranking Quality: Recall and NDCG Analysis

Next, we isolate and analyze retrieval performance, which is motivated only by the retriever components as opposed to the generator model. Table 5 below (retrieval ranking) shows the average ranking quality of all the generators assessed. Expectedly, No-RAG is not used, since no external evidence is accessed.

Naive RAG will be moderate in retrieval depth ($R@10 \approx 0.43$), which means that the correct curriculum passage is typically recovered but not necessarily ranked highly. Dense embeddings alone tend to surface semantically similar but syllabus-misaligned segments, limiting faithfulness in downstream generation.

The proposed Curriculum-Aware RAG significantly improves early precision ($R@1: +1.8pp$) and graded relevance sensitivity ($NDCG@5: +3.2pp$), ensuring that the most relevant and grade-appropriate content is prioritized at the top of the list. This improvement directly supports downstream factuality gains, as shown in Section 4.3.

Table 5. Retrieval ranking quality averaged across all generator models.

Pipeline	R@1	R@5	R@10	NDCG@5
Naive RAG	0.124	0.344	0.431	0.336
Curriculum-Aware RAG	0.142	0.373	0.484	0.368

Overall, these results confirm that curriculum-aware reranking consistently enhances the correct textbook passages to the top positions, thereby enabling more grounded and trustworthy educational responses. This motivates its use as a default retrieval configuration for curriculum-aligned conversational tutors. Retrieval metrics are computed on the retriever output only and are independent of the generator; averages reflect the same retrieved lists across model conditions.

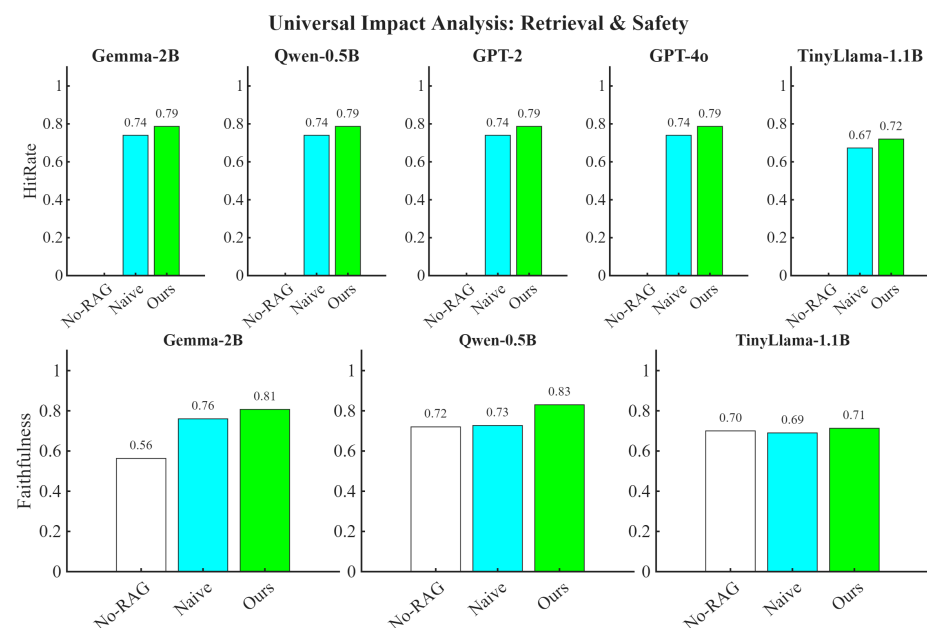
4.5. Architectural Ablation: Naive vs. Curriculum-Aware RAG

To isolate the effect of the proposed retrieval architecture, we compare Naive RAG (dense-only) to Curriculum-Aware RAG (hybrid + cross-encoder + metadata boost) for each generator. Table 6 reports the relative gains in F1 and Faithfulness.

Table 6. Relative performance lift of Curriculum-Aware RAG over Naive RAG (F1 and Faithfulness only).

Model	F1	Faithfulness
GPT-2	+0.8%	-
GPT-4o	+0.7%	-
Gemma-2-2b-it	+0.7%	+4.7%
Qwen2.5-0.5B	+1.2%	+10.3%
TinyLlama-1.1B	+1.4%	+2.3%

To illustrate these model-agnostic benefits, Figure 4 displays absolute performance metrics across generators for both HitRate (top row) and Faithfulness (bottom row) under No-RAG, Naive RAG, and Curriculum-Aware RAG configurations. The proposed system consistently lifts both retrieval coverage and factual alignment, regardless of model size.

**Figure 4.** Universal impact of retrieval architecture on HitRate (**top**) and Faithfulness (**bottom**) across five language models. Our Curriculum-Aware RAG outperforms both No-RAG and Naive RAG setups in all configurations.

The gains are remarkably stable across models: Curriculum-Aware RAG consistently increases the overall hit rate by around +4.7 percentage points and improves early precision ($R@1$) and mid-depth recall ($R@5$, $R@10$), as well as graded relevance ($NDCG@5$). This pattern indicates that the curriculum-aware hybrid + reranking pipeline is doing most of the work, while the specific generator mainly affects how well the retrieved evidence is verbalized.

For Qwen2.5-0.5B, Curriculum-Aware RAG increases retrieval accuracy from 0.74 to 0.787 and $NDCG@5$ from 0.336 to 0.368, while F1 improves from 31.5% to 32.7% and faithfulness from 72.7% to 83.0%. Figure 5 shows the Recall@K trend for this model across three depths, where the curriculum-aware pipeline outperforms the naive baseline at every level.

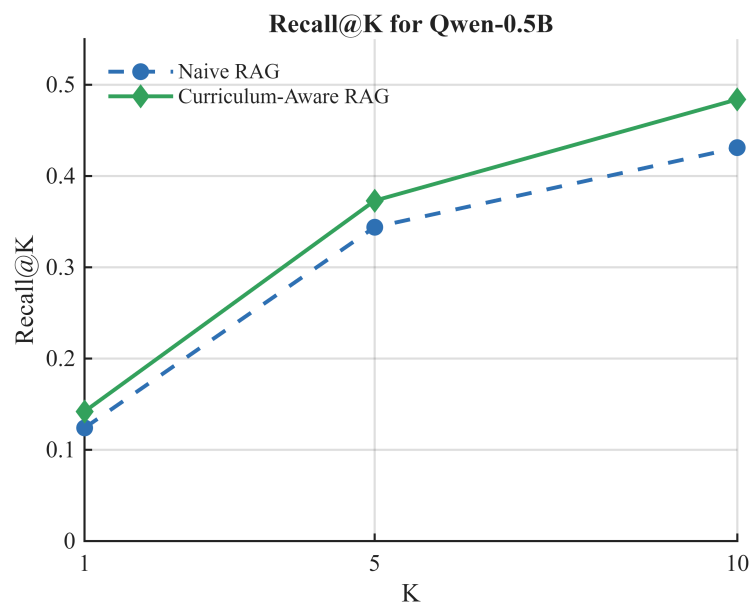


Figure 5. Recall@K curve for Qwen2.5-0.5B showing that Curriculum-Aware RAG consistently outperforms Naive RAG at depths $k = 1, 5, 10$.

Gemma-2-2b-it shows a similar trend, with a +0.7 percentage point gain in F1 and roughly +4.7 percentage points in faithfulness. GPT-2 benefits from improved retrieval, suggesting that curriculum-aligned evidence improves grounding even beyond small open-source generators.

The smallest generator, TinyLlama-1.1B, also improves in terms of architectural improvement: retrieval accuracy improves to 72.0% rather than 67.3%, F1 improves to 21.3 rather than 19.8 and faithfulness improves to 71.3 rather than 69.0. The improvements demonstrate that even smaller models become more reliable simply by enhancing the retrieval design.

In summary:

- Curriculum-sensitive reranking achieves steady improvements in HitRate, RecallK, and Faithfulness, using all model scales;
- Evidence quality gain is associated with small lexical improvements (F1), especially smaller models;
- Pedagogical advantages are greatest in terms of faithfulness, indicating the design of disciplined retrieval is significant in minimizing hallucinations without introducing changes in model parameters..

4.6. Cross-Model Safety and Faithfulness

Faithfulness is a vital attribute of the educational deployment; a fluent but factually false response can mislead to the learners. As Figure 6 below (safety–faithfulness) indicates, Retrieval-Augmented Generation meaningfully decreases the level of hallucination in all models. In the case of GPT-2, faithfulness increases by 0.527 (No-RAG) to 0.860 (Curriculum-Aware RAG), and there is an improvement of 0.720 to 0.830 in the Qwen2.5-0.5B. These consistent gains show that even small open models can be turned into a stable type of educational tutor in case they are based on evidence that has been approved by the curriculum.

The most slight modifications are in TinyLlama-1.1B (0.690 → 0.713), the source of which is the lack of headroom with tiny models. Nevertheless, the gains can still be considered relevant provided that its parameter footprint is small, and its base capability is much lower.

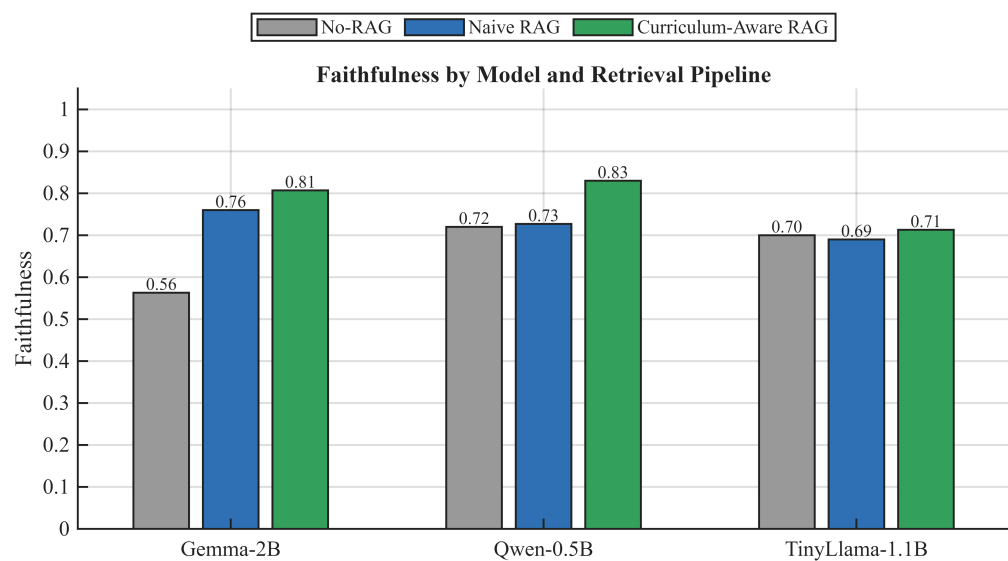


Figure 6. Reliability of faithfulness scores. Curriculum-Aware RAG conducts much better safety than No-RAG and Naive RAG to SLMs.

On the whole, these findings highlight one of the main findings: the concept of retrieval architecture, rather than model scale, is the most significant factor of pedagogical safety. Curriculum-aware reranking always drives any model to syllabus-based, substantially reduced hallucination rate solutions, and makes it reliable even when using the low-resource SLMs.

Figure 7 summarizes these trade-offs as a safety–efficiency frontier. No-RAG configurations cluster in the lower-left region, combining low latency with substantially lower faithfulness. Once retrieval is enabled, both Naive RAG and Curriculum-Aware RAG shift the models upward, reflecting consistently higher faithfulness, with only a modest movement to the right in terms of latency. For the small open models (Gemma-2-2b-it, Qwen2.5-0.5B, TinyLlama-1.1B), the Curriculum-Aware pipeline occupies a favorable region of this frontier, delivering markedly higher faithfulness than Naive RAG at nearly identical inference times. For the small open-source generators (Gemma-2-2b-it, Qwen2.5-0.5B, and TinyLlama-1.1B), the proposed Curriculum-Aware RAG pipeline consistently improves faithfulness with only minor latency overhead, confirming its suitability for resource-constrained deployment.

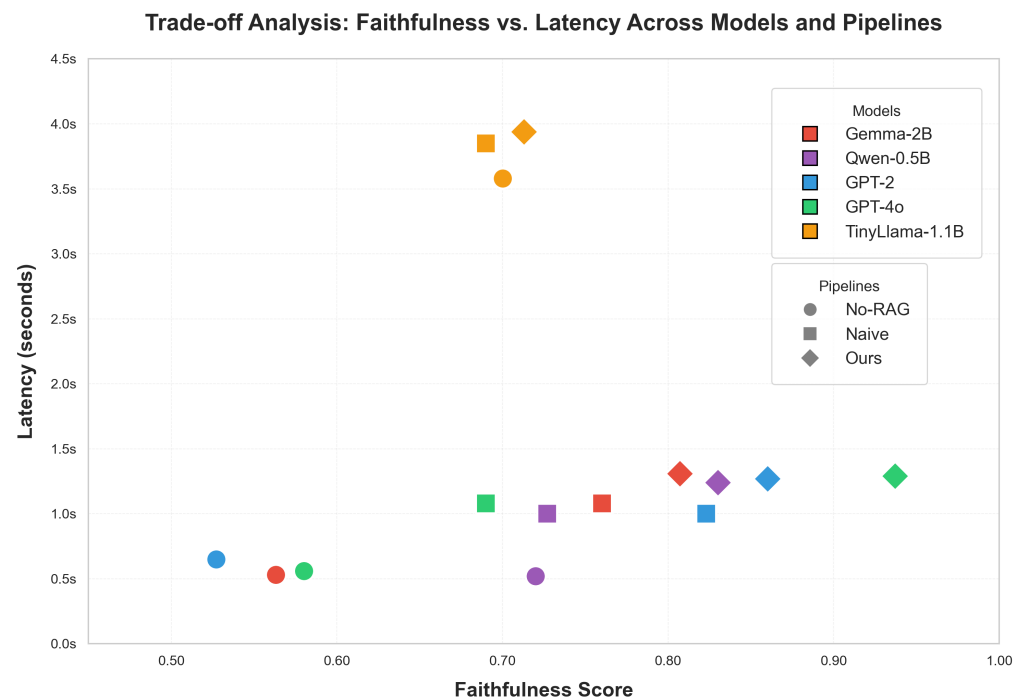


Figure 7. Safety–efficiency frontier across all model–pipeline configurations. Each point represents a specific generator (Gemma-2-2b-it, Qwen2.5-0.5B, GPT-2, GPT-4o, and TinyLlama-1.1B) under one of three pipelines: No-RAG, Naive RAG, or Curriculum-Aware RAG. The x-axis shows mean latency per query, and the y-axis shows mean faithfulness scores.

Human Evaluation: Faithfulness Agreement and Alignment with Automatic Metrics

To complement the automatic evaluation metrics, we conducted a human assessment of answer faithfulness on a stratified sample of $N = 120$ responses generated by the best-performing configuration (Qwen-0.5B + Curriculum-Aware RAG). The sample was balanced across language conditions to improve coverage and fairness, including 40 English queries, 40 Swahili queries, and 40 code-switched queries. Two independent annotators labeled each response using a binary supported-claim criterion, where a response was marked as faithful (1) only if all major claims were supported by the retrieved evidence, and not faithful (0) otherwise.

Inter-annotator reliability was substantial under Cohen’s Kappa, achieving $\kappa = 0.714$. Across the evaluated subset, the proposed pipeline obtained a high mean human faithfulness score of 0.954, indicating that the majority of generated responses were judged as grounded in retrieved curriculum evidence.

We further examined whether commonly used automatic signals correlate with human judgments of faithfulness. Spearman and Pearson correlation analyses between human faithfulness scores and automatic metrics (e.g., HitRate, F1, ROUGE-L, embedding similarity, curriculum alignment score, and soft recall) yielded generally weak associations (with $|\rho| < 0.18$). This behavior is expected under a ceiling effect, where most human annotations concentrate near the upper bound, reducing score variance and limiting statistical sensitivity.

Importantly, absolute human faithfulness remains high (mean = 0.954, $\kappa = 0.714$), indicating strong annotator agreement and reliable grounding under expert review. We therefore interpret the automatic faithfulness metric as a conservative lower bound, since human annotators consistently rated responses higher than the automatic judge (0.830), and the reported automatic improvements likely understate the true pedagogical reliability

of the system. While automatic metrics remain useful for system-level comparisons, they do not fully substitute for human verification in high-performing settings.

4.7. Cross-Lingual and Curriculum-Stratified Performance

We next analyze the best-performing configuration, Qwen2.5-0.5B combined with Curriculum-Aware RAG, to assess cross-lingual equity and curriculum coverage in Tanzanian secondary History.

Given the smaller code-switched subset, these results should be interpreted as indicative rather than definitive. From an equity standpoint, comparable HitRate across languages suggests similar access to evidence, reducing the risk that bilingual students are systematically disadvantaged by the retrieval layer.

4.7.1. Language-Stratified Performance

Table 7 and Figure 8 summarize cross-lingual performance for the best-performing configuration (Qwen2.5-0.5B with Curriculum-Aware RAG). Retrieval accuracy remains comparable across languages (HitRate ≈ 0.79 for English and 0.78 for Swahili), indicating that the curriculum-aware retriever is not biased toward one language in terms of evidence access. However, Swahili queries achieve substantially higher answer quality (F1 ≈ 0.53 vs. 0.28 for English) and stronger embedding similarity to the reference answers, suggesting that the system is particularly effective at generating semantically rich explanations in Swahili even when materials are predominantly stored in English. This pattern is encouraging from a digital equity perspective: it shows that multilingual dense retrieval, combined with syllabus-aware reranking, can bridge language gaps without requiring full translation of the underlying curriculum corpus. We note that Swahili references are often more descriptive/consistent in phrasing, which can inflate overlap-based metrics; we therefore interpret F1 alongside HitRate and embedding similarity.

Table 7. Language-stratified performance for Qwen2.5-0.5B + Curriculum-Aware RAG.

Language	Retrieval Accuracy	Embedding Similarity	F1 (%)
English	0.789	0.713	28.2
Swahili	0.778	0.854	53.1
Code-switched	0.700	0.794	31.8

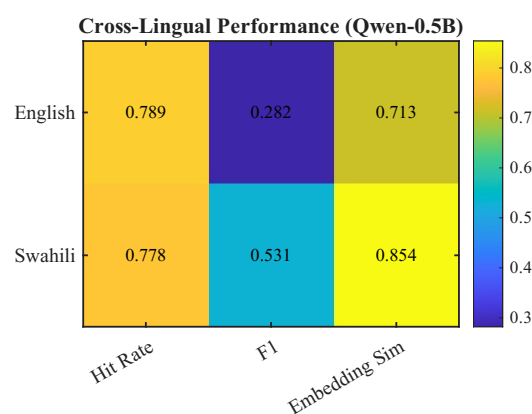


Figure 8. Cross-lingual performance heatmap (Qwen2.5-0.5B) comparing English and Swahili queries across HitRate, F1, and Embedding Similarity.

The success of retrieval is almost similar regardless of the language (HitRate ≈ 0.78) and it proves that the semantic retriever is generalized. Nonetheless, Swahili responses show significantly higher semantic similarity and F1, which means that the system is more accurate and curriculum-oriented in Swahili, as compared to English. It indicates

that curriculum-based retrieval plays a role in overcoming the current paucity of African-language training data to support bridging the chronic digital divide in low-resource language education.

4.7.2. Grade-Level (Form) Performance

Table 8 reports retrieval effectiveness across grade levels (Forms 1–4).

Table 8. Retrieval success by grade level for Qwen2.5-0.5B + Curriculum-Aware RAG.

Form	Retrieval Accuracy
Form 1	0.772
Form 2	0.571
Form 3	0.760
Form 4	0.889

Form-4 (HitRate = 0.889) has the best performance, probably because of the more dense examination content and better organization of the topics. Form 2 demonstrates the lowest retrieval (0.571) that we can explain by noisy scans and diffuse curriculum goals. This indicates that better data cleaning and syllabus tagging of low-grade materials is needed.

Combined, these stratified findings indicate that the proposed system offers fair cross-lingual retrieval with some revealed possibilities on how to improve corpus and metadata to guarantee complete coverage of the curriculum.

4.7.3. Question-Type and Cognitive Complexity

We also evaluate performance by question type (aligned to Bloom levels). Table 9 summarizes Retrieval Accuracy and F1 across categories.

Table 9. Performance by question type/cognitive complexity (Qwen2.5-0.5B + Curriculum-Aware RAG).

Question Type	Retrieval Accuracy	F1
Definition	0.875	0.315
Explanation	0.800	0.256
Comparison	0.800	0.250
List	0.677	0.266
Generic/Open	0.788	0.399

The retriever maintains high Retrieval Accuracy across definitional, explanatory, and comparative queries (≥ 0.80), with somewhat lower performance on list questions (0.677), which often require aggregating multiple dispersed facts. Interestingly, generic questions yield the highest F1 (0.399), indicating that once relevant context is retrieved, the SLM can synthesize broader conceptual explanations effectively.

4.8. Error Taxonomy and Safety Behavior

To understand failure modes, we manually categorized model outputs for Qwen2.5-0.5B + Curriculum-Aware RAG (Table 10).

Table 10. Outcome distribution for Qwen2.5-0.5B + Curriculum-Aware RAG.

Outcome Type	Count	Percentage
Success	72	48.0%
Retrieval Failure	32	21.3%
Reasoning Failure	43	28.7%
Hallucination	3	2.0%

The dominant residual error modes are *reasoning failures* (28.7%) and *retrieval failures* (21.3%), whereas hallucinations are rare (2.0%), confirming the effectiveness of the grounding and refusal policy. Many reasoning failures occur in multi-part questions where the model retrieves appropriate evidence but partially answers the question, suggesting a future role for multi-step generation or chain-of-thought supervision.

Remaining errors are dominated by retrieval gaps and multi-step reasoning failures rather than hallucinations, suggesting that future gains may come from improved corpus quality (OCR/metadata) and structured multi-step generation.

A qualitative example of safe behavior is shown in Table 11, where the system refuses to answer an out-of-syllabus question instead of hallucinating.

Table 11. Example of knowledge-aware refusal under retrieval failure.

User Query	According to Charles Darwin’s theory, fire was discovered in which age?
Retrieval Status	No relevant curriculum chunk retrieved.
Naive RAG	Hallucinated a specific “Middle Stone Age” answer.
Curriculum-Aware RAG	“The study materials do not provide enough information to answer this question. Please check the curriculum content.”

4.9. Latency and Deployment Feasibility

We evaluate end-to-end latency for Naive vs Curriculum-Aware RAG across models (Table 12). Latency includes retrieval, reranking, and generation.

Table 12. End-to-end latency per model and pipeline on the 500-question benchmark.

Model	Pipeline	Mean Latency (s)	Median Latency (s)
Gemma-2-2b-it	Naive RAG	1.199	1.207
Gemma-2-2b-it	Curriculum RAG	1.707	1.591
Qwen2.5-0.5B	Naive RAG	1.350	1.257
Qwen2.5-0.5B	Curriculum RAG	1.523	1.418
GPT-2	Naive RAG	1.367	1.253
GPT-2	Curriculum RAG	1.654	1.544
GPT-4o	Naive RAG	1.313	1.192
GPT-4o	Curriculum RAG	1.492	1.433

Across all models, Curriculum-Aware RAG adds approximately 0.15–0.50 s per query relative to Naive RAG primarily due to cross-encoder reranking while remaining within a 1.5–1.7 s mean latency window. This trade-off is acceptable for interactive tutoring and demonstrates that curriculum-robust RAG can be deployed in resource-constrained schools without prohibitive delay.

Top- k Sensitivity and Deployment Considerations

We conducted a sensitivity analysis on the retrieval depth k to inform deployment-time design choices rather than to establish an optimized benchmark configuration. Increasing k consistently improves Retrieval Accuracy by increasing the probability that a curriculum-relevant passage appears in the retrieved context. However, this improvement exhibits diminishing returns beyond moderate values of k .

In practice, larger values of k substantially increase context length, token consumption, and latency, which is undesirable in low-resource educational settings. Based on these trade-offs, we adopt a moderate retrieval depth ($k = 5$ – 10) in the deployed prototype,

which provides strong retrieval coverage while maintaining acceptable inference latency and token efficiency.

This choice reflects a deployment-oriented compromise aligned with “Green AI” principles, prioritizing practical usability and scalability over marginal gains in recall.

4.10. Pedagogical Accessibility and Style Control

In addition to the factual accuracy and curriculum backing, another important need of an educational assistant in the Global South is to create text congruent with the reading ability of the learner. To measure the linguistic accessibility of the generated responses, we used the Automated Readability Index (ARI), which is a standardized measure with a score whereby the score is close to the US grade level required to understand the reading.

Table 13 gives a comparative study of the readability of various pipeline configurations. Two important remarks apply to the activity of Retrieval-Augmented Generation in the educational settings:

- **The Expert Bias of RAG:** According to the entries of the “Standard RAG” column, the retrieval always makes linguistic complexity higher than the direct LLM generation. As an example, the ARI of Qwen2.5-0.5B is as high as in Standard RAG mode, equal to 17.98. This is due to the fact that the retriever incorporates academic jargon and multi-syllable sentence structuring directly off the textbooks and raises the level of output to a level of postgraduate reading ($ARI > 17$), which is completely unreachable by students in Forms 1–2.
- **Style-Controlled Decoding Effectiveness:** To overcome this obstacle, we applied a style-controlled mode of generation that was expressly conditioned to achieve a Form 1 (13 years of age) reading grade. This intervention was effective in drastically reducing the mean ARI scores on all the models. In the case of Qwen2.5-0.5B, the complexity decreased from 17.98 to 11.89 (a huge reduction of 6.09 grade levels), which brought the content much closer to the proper pedagogical range without any loss of factual accuracy.

Table 13. Impact of style-controlled decoding on readability. The style-control mechanism as proposed is also much lower than the Automated Readability Index (ARI) against the regular Expert RAG baseline.

Model	Standard RAG (Expert) (Mean ARI)	Style-Controlled (Form 1) (Mean ARI)	Reduction (Grade Levels)
Gemma-2-2b-it	17.84	12.01	−5.83
Qwen2.5-0.5B	17.98	11.89	−6.09
TinyLlama-1.1B	17.87	11.99	−5.87
GPT-4o	17.91	11.65	−6.27

These findings indicate that naive retrieval enhances the factuality but sacrifices the extreme complexity, whereas a curriculum-sensitive pipeline with explicit style regulation can efficiently balance the precision of domain and developmental suitability. Such a way of adapting is necessary to enable fair implementation in Tanzanian secondary schools, where there is a large difference in the levels of English proficiency.

The level of readability of standard RAG (about 17–18) shows that the produced responses are of the level of academic texts at graduate level, which cannot be assigned to secondary school students. The suggested intervention is able to reduce this to the Form 3–4 range (ARI 11–12), which increases accessibility to a substantial degree. All human annotators were blinded to the underlying model identity and retrieval configuration during evaluation. We additionally verified that style-controlled decoding did not

reduce faithfulness scores, indicating that readability improvements did not compromise factual grounding.

Overall, these results demonstrate that Curriculum-Aware RAG systematically improves retrieval, answer quality, and safety across multiple language models while remaining efficient enough for interactive secondary-school tutoring in low-resource Tanzanian contexts.

5. Discussion

5.1. Comparison with Prior Studies

The findings of this study align with, and extend, a growing body of literature on Retrieval-Augmented Generation (RAG) for educational applications, particularly in low-resource and multilingual contexts. Prior work has consistently shown that retrieval grounding improves factual accuracy and reduces hallucination in large language models when applied to knowledge-intensive tasks [9]. However, much of this literature assumes English-dominant corpora or access to large proprietary models, which limits applicability in Global South educational settings where infrastructure, connectivity, and licensing constraints are prevalent.

Recent surveys and empirical studies on educational chatbots and AI tutors have demonstrated that RAG-based systems can substantially improve the reliability of learning-oriented question answering by grounding responses in curated instructional materials [36,45]. These systems, however, typically rely on dense semantic retrieval alone and evaluate performance primarily using lexical overlap metrics. In contrast, our results show that dense-only retrieval is insufficient for curriculum-sensitive subjects such as History, where semantically similar passages may be misaligned with grade level, syllabus scope, or pedagogical intent. The consistent gains observed with Curriculum-Aware RAG over Naive RAG (Tables 5 and 6) highlight the importance of incorporating explicit curriculum metadata and cross-encoder reranking to prioritize pedagogically appropriate evidence.

In the context of low-resource languages, prior work has largely focused on improving representation learning and pretraining for African languages [22,46]. While these efforts are foundational for multilingual NLP, they do not directly address downstream educational safety concerns such as hallucinations, syllabus drift, or grade-level mismatch. Our cross-lingual analysis demonstrates that curriculum-grounded retrieval can substantially improve answer correctness and semantic fidelity for Swahili queries without requiring full translation of the curriculum corpus, thereby complementing existing multilingual modeling efforts with a retrieval-centric safety mechanism.

Compared to recent educational RAG systems that rely on large proprietary language models [36], our results show that small, open-source language models supported by a robust retrieval and reranking pipeline can achieve comparable levels of faithfulness and retrieval accuracy. This finding challenges the prevailing assumption that educational reliability necessarily requires scaling model size, and instead supports a growing line of research advocating for architecture-centric improvements over parameter-centric scaling [47,48].

Finally, the observed trade-offs between accuracy, latency, and deployability mirror those reported in prior studies on efficient and edge-deployable NLP systems [49]. While Curriculum-Aware RAG introduces modest computational overhead due to cross-encoder reranking, the resulting gains in pedagogical safety and curriculum fidelity are substantial, particularly for smaller models. This trade-off aligns with recommendations in the educational AI literature that prioritize reliability, alignment, and learner safety over marginal efficiency gains in student-facing applications.

Overall, this work advances the literature by demonstrating that curriculum-aware retrieval design is a critical missing component in existing educational RAG systems, particularly for multilingual and low-resource contexts. By explicitly integrating syllabus structure, grade-level constraints, and pedagogical metadata into the retrieval pipeline, the proposed approach bridges a gap between general-purpose RAG architectures and the specific demands of formal secondary education.

5.2. Retrieval Architecture as the Primary Driver of Learning Accuracy

Our findings show that pipeline design matters more than model size. Using Curriculum-Aware RAG, Gemma-2-2b-it and Qwen2.5-0.5B narrow many of the GPT-4o gaps while being 10–50× smaller. Retrieval accuracy displays a steady lift of +4.7% across models (0.740 → 0.787), whereas faithfulness increases by +3.7% to +24.7% with model capacity.

(Table 6). These gains indicate that:

- Hybrid retrieval consistently returns more relevant syllabus-grounded context;
- Cross-encoder reranking suppresses misleading but lexically similar passages.

This is particularly crucial in History, where small factual distortions (e.g., dates, actors, causal sequences) can propagate misconceptions into later learning stages.

Although the lexical accuracy is not high ($F1 \approx 0.31$ – 0.33), this is mostly due to the nature of token-overlap scoring of multi-sentence explanatory texts. In schools, verisimilitude and curricular fidelity are not matters of superficiality, which justifies the further emphasis on faithfulness-based assessment.

5.3. Cross-Lingual Equity: Improving Low-Resource Language Outcomes

An important contribution of this research is that it proved that educational RAG can have a strong benefit to low-resource languages. On the most successful structure (Qwen2.5-0.5B + Curriculum-Aware RAG), Swahili queries obtain:

- Comparable retrieval success to English (0.778 vs. 0.789 HitRate);
- Considerably higher semantic accuracy (0.854 vs. 0.713);
- Substantially higher answer correctness (53.1% vs. 28.2% F1).

These results contradict the prevailing expectation of English dominance in NLP performance and demonstrate that curriculum-grounded retrieval helps narrow the digital divide in African educational contexts even without translating the full corpus.

5.4. Understanding Remaining Errors: A Safer System by Design

Error analysis highlights both strengths and remaining challenges. Hallucinations are rare (2.0%), with the system frequently preferring refusal when uncertain—an essential behavior for student safety. The residual errors are dominated by:

- Reasoning failures in multi-hop or comparative questions;
- Retrieval gaps concentrated in Form 2 content due to noisier PDF extraction;
- Overly complex phrasing that surpasses the reading level of early secondary students.

These patterns suggest that future work should incorporate structured reasoning (e.g., chain-of-thought), finer curriculum metadata, and stylistic simplification to make responses more accessible for Forms 1–2 learners.

5.5. Latency and Deployment Feasibility in Low-Connectivity Settings

End-to-end latency is also amenable to substantial safety gains (1.01.7 s on GPU across models) despite using cross-encoder reranking as a manageable trade-off. Because SLMs can be used with high performance with Curriculum-Aware RAG, edge computing with

CPU-class hardware is now possible, making fair deployment in Tanzanian secondary schools with mobile only settings possible.

Nevertheless, analysis of readability reveals that responses are typically more complex than textbooks (mean ARI 15-17 vs. textbook ARI 15). Grade-conscience simplification and simplification Control will be important to make the answers more accessible without factual loss.

5.6. Practical Implications for Low-Resource Educational Deployment

Beyond empirical performance gains, the proposed Curriculum-Aware RAG framework has several practical implications for the deployment of AI-assisted tutoring systems in low-resource and multilingual educational contexts, particularly within the Global South.

5.6.1. Curriculum-Aligned Classroom Deployment

By grounding generation strictly in syllabus-approved materials and enforcing grade-level and topic constraints during retrieval and reranking, the system can be safely integrated into formal secondary school instruction. Unlike open-domain educational chatbots, the proposed approach minimizes the risk of off-curriculum content, making it suitable for classroom use, homework support, and examination preparation without undermining national education standards.

5.6.2. Teacher Support and Pedagogical Augmentation

Rather than replacing teachers, the system is best viewed as a pedagogical assistant. It can support teachers by providing curriculum-consistent explanations, multilingual scaffolding (English and Swahili), and on-demand clarification of historical concepts aligned with Bloom's Taxonomy levels. This is particularly valuable in contexts where teacher-to-student ratios are high and subject-matter specialization is uneven across schools.

5.6.3. Feasibility for Offline and Edge Deployment

A central design goal of this work is deployability under constrained computational and connectivity conditions. The use of Small Language Models (SLMs), lightweight hybrid retrieval, and modular reranking enables deployment on modest GPU or CPU infrastructure, and potentially in near-offline or intermittently connected settings. The empirical latency results demonstrate that the additional cost of curriculum-aware reranking remains within acceptable bounds for interactive tutoring, making nationwide or regional deployment feasible without reliance on large proprietary models.

5.6.4. Policy Relevance and Educational Equity

From a policy perspective, this work provides a concrete technical pathway for integrating AI into public education systems while preserving curricular sovereignty, linguistic inclusivity, and cost control. By avoiding dependence on closed, externally hosted LLMs and emphasizing curriculum-grounded knowledge bases, the proposed framework aligns with emerging recommendations for responsible and equitable AI adoption in education across the Global South. It offers a scalable template for ministries of education seeking to leverage AI to reduce educational disparities without compromising pedagogical integrity.

Overall, these practical considerations highlight that curriculum-aware retrieval design is not only a methodological contribution but also a necessary foundation for the responsible deployment of AI tutors in real-world, resource-constrained educational environments.

5.7. Limitations and Future Directions

Code-Switching Coverage and Data Balance

One limitation of the present evaluation concerns the distribution of language types in the Golden Dataset. While the dataset includes English (60%), Swahili (30%), and Swahili–English code-switched queries (10%), the number of code-switched questions ($n = 50$) is relatively small compared to the monolingual subsets. As a result, although the reported results indicate that the proposed framework can handle code-switched inputs without degradation in retrieval or faithfulness, these findings should be interpreted with caution. In particular, the current sample size is insufficient to support statistically significant claims regarding robustness to code-switching; therefore, the reported results should be interpreted as indicative trends rather than definitive evidence. Future work will expand the evaluation set with a higher proportion of naturally occurring code-switched questions, better reflecting authentic classroom discourse in Tanzanian secondary education and enabling more fine-grained comparative analysis across language conditions.

Although the corpus is applicable across several subjects, the experiments in this paper focus on textual History content. The extension to the STEM area will involve multimodal recovery of diagrams, equations, and geographical maps, which we define as a short-term development goal. While evaluations are conducted on A100 hardware to ensure consistency, deploying the system in resource-limited school infrastructures presents challenges. To address this, techniques such as model distillation, sparse retrieval, and caching are expected to improve scalability and facilitate deployment across the nation.

Curriculum-Aware Retrieval enables small, resource-efficient language models to serve as effective, safe, and linguistically inclusive AI tutors for secondary students in low-resource settings. Rather than deepening reliance on proprietary mega-models, this work advocates for a sustainable and locally driven approach to AI in education, fostering both national equity and broader shifts towards culturally relevant intelligent tutoring systems.

Moving forward, four promising research directions remain:

1. Pedagogical personalization: grade-aware decoding, adaptive scaffolding, and longitudinal learner modeling.
2. Multimodal extension to STEM subjects: retrieval over diagrams, charts, mathematical notation, and geographic maps.
3. Lightweight deployment: knowledge distillation, sparse retrieval, and quantization for nationwide rollout.
4. Classroom field studies: evaluating learning gains, safety perceptions, and teacher-AI co-instruction dynamics.

The present work moves towards a sustainable and locally based direction for AI in education. Instead of using proprietary large language models, curriculum-aware RAG allows small, low-cost models to become high quality, safe, and linguistically inclusive learning companions. This will offer a scalable template for national education AI in Tanzania and the Global South in general, paving the way for a future where learners can access high-quality academic support in both the language of instruction and the language of home.

6. Conclusions

This work has also shown that curriculum-conscious Retrieval-Augmented Generation (RAG) can be a force behind equal and contextually sensible AI-based learning systems in low-resource educational setups. The proposed architecture can be used to compete with larger models that do not use structured external evidence to perform competitively with lightweight language models like Gemma-2-2b-it, Qwen2.5-0.5B, and TinyLlama-1.1B

through the combination of hybrid retrieval, cross-encoder reranking, and metadata-driven curriculum alignment. In all the generators, the retrieval accuracy increases steadily, ranging between 0.740 and 0.787, whereas faithfulness rises by up to 24.7 percent, which confirms that in secondary education, retrieved information is more important than the size of the model.

There is also high cross-lingual robustness in the system. The Swahili queries (served inadequately by typical AI systems) have a better semantic correctness and factual accuracy ($F1 = 53.1\%$) than English queries (28.2%), although they have similar levels of retrieval success. This finding in the setting of Tanzanian secondary History gives the first indication that curriculum-based retrieval can significantly reduce the African digital divide without necessarily translating the school curriculum.

Author Contributions: Conceptualization, B.W., X.Z. and C.L.; methodology, I.E.R.; software, I.E.R. and W.C.; validation, I.E.R. and X.Z.; formal analysis, I.E.R.; investigation, I.E.R., W.C. and C.L.; resources, W.C. and C.L.; data curation, I.E.R. and X.Z.; writing—original draft preparation, I.E.R.; writing—review and editing, B.W. and W.C.; visualization, I.E.R. and X.Z.; and supervision, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this paper:

AI	Artificial Intelligence
ANN	Approximate Nearest Neighbor
API	Application Programming Interface
BGE-M3	Beijing Academy of Artificial Intelligence M3 Embedding Model
BM25	Best Match 25 (Lexical Search Algorithm)
CPU	Central Processing Unit
CSEE	Certificate of Secondary Education Examination
EM/F1	Exact Match/F1 Score
FAISS	Facebook AI Similarity Search
GPU	Graphics Processing Unit
GPT	Generative Pretrained Transformer
ICT	Information and Communication Technology
IDF	Inverse Document Frequency
IP/L2	Inner Product/L2 Distance (Vector Similarity Metrics)
LLM	Large Language Model
NECTA	National Examinations Council of Tanzania
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
OCR	Optical Character Recognition
PWA	Progressive Web Application
QA	Question Answering
RAG	Retrieval-Augmented Generation

RAGFlow	Retrieval-Augmented Generation Flow
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SLM	Small Language Model
STEM	Science, Technology, Engineering and Mathematics
TIE	Tanzania Institute of Education
UI	User Interface
ZPD	Zone of Proximal Development

References

1. Casley Gera, R. *Addressing Inefficient Distribution of Teachers Between Schools: The Case of Tanzania (With Malawi and the Gambia)*; World Bank: Washington, DC, USA, 2023.
2. Mremi, J.; Muraina, K. Shortage of Science and Mathematics Teachers and its Impacts on Secondary Schools in Tanzania. *J. Issues Pract. Educ.* **2023**, *15*, 50–63. <https://doi.org/10.61538/jipe.v15i1.1318>.
3. Ndomondo, E. Instructional resources for innovative history teaching and learning in Tanzania secondary schools: Exploring availability and utilization. *Cogent Arts Humanit.* **2024**, *11*, 2382526.
4. Ndimbo, E.V.; Luo, Q.; Fernando, G.C.; Yang, X.; Wang, B. Leveraging Retrieval-Augmented Generation for Swahili Language Conversation Systems. *Appl. Sci.* **2025**, *15*, 524. <https://doi.org/10.3390/app15020524>.
5. Zhang, W.; Zhang, J. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics* **2025**, *13*, 856. <https://doi.org/10.3390/math13050856>.
6. Song, J.; Wang, X.; Zhu, J.; Wu, Y.; Cheng, X.; Zhong, R.; Niu, C. RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*; Dernoncourt, F., Preotiu-Pietro, D., Shimorina, A., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 1548–1558. <https://doi.org/10.18653/v1/2024.emnlp-industry.113>.
7. Xu, S.; Yan, Z.; Dai, C.; Wu, F. MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Front. Public Health* **2025**, *13*, 1635381.
8. Li, X.; Henriksson, A.; Duneld, M.; Nouri, J.; Wu, Y. Supporting Teaching-to-the-Curriculum by Linking Diagnostic Tests to Curriculum Goals: Using Textbook Content as Context for Retrieval-Augmented Generation with Large Language Models. In *Proceedings of the Artificial Intelligence in Education*; Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I., Eds.; Springer: Cham, Switzerland, 2024; pp. 118–132.
9. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474.
10. Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; et al. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. *arXiv* **2023**, arXiv:2308.02773.
11. Ogundepo, O.; Gwadabe, T.R.; Rivera, C.E.; Clark, J.H.; Ruder, S.; Adelani, D.I.; Dossou, B.F.P.; Diop, A.A.; Sikasote, C.; Hacheme, G.; et al. AfriQA: Cross-lingual Open-Retrieval Question Answering for African Languages. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 14957–14972. <https://doi.org/10.18653/v1/2023.findings-emnlp.997>.
12. Schick, T.; Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv* **2021**, arXiv:2009.07118.
13. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; tau Yih, W.; Koh, P.W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv* **2023**, arXiv:2305.14251.
14. Rawte, V.; Sheth, A.; Das, A. A Survey of Hallucination in Large Foundation Models. *arXiv* **2023**, arXiv:2309.05922.
15. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. *arXiv* **2023**, arXiv:2309.01431.
16. Zhang, X.; Zhang, Y.; Long, D.; Xie, W.; Dai, Z.; Tang, J.; Lin, H.; Yang, B.; Xie, P.; Huang, F.; et al. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. *arXiv* **2024**, arXiv:2407.19669.
17. Zhang, S.; Yu, T.; Feng, Y. TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space. *arXiv* **2024**, arXiv:2402.17811.
18. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; Ku, L.W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 6233–6251. <https://doi.org/10.18653/v1/2024.findings-acl.372>.

19. Schwarcz, D.; Manning, S.; Barry, P.J.; Cleveland, D.R.; Prescott, J.; Rich, B. AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice. *Minnesota Legal Studies Research Paper; U of Michigan Public Law Research Paper*. 2025. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111 (accessed on 1 December 2025). <https://doi.org/10.2139/ssrn.5162111>.
20. Amarnath, N.S.; Nagarajan, R. An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students. In Proceedings of the 2024 4th International Conference on Sustainable Expert Systems (ICSSES), Pokhara, Nepal, 15–17 October 2024; pp. 1393–1398. <https://doi.org/10.1109/ICSSES63445.2024.10762977>.
21. Lang, G.; Gürpınar, T. AI-Powered Learning Support: A Study of Retrieval-Augmented Generation (RAG) Chatbot Effectiveness in an Online Course. *Inf. Syst. Educ. J.* **2025**, *23*, 4–13.
22. Adelani, D.I.; Ojo, J.; Azime, I.A.; Zhuang, J.Y.; Alabi, J.O.; He, X.; Ochieng, M.; Hooker, S.; Bukula, A.; Lee, E.S.A.; et al. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. *arXiv* **2025**, arXiv:2406.03368.
23. Adelani, D.I.; Neubig, G.; Ruder, S.; Rijhwani, S.; Beukman, M.; Palen-Michel, C.; Lignos, C.; Alabi, J.O.; Muhammad, S.H.; Nabende, P.; et al. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 4488–4508. <https://doi.org/10.18653/v1/2022.emnlp-main.298>.
24. Ojo, J.; Ogundepo, O.; Oladipo, A.; Ogueji, K.; Lin, J.; Stenetorp, P.; Adelani, D.I. AfroBench: How Good are Large Language Models on African Languages? In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 19048–19095. <https://doi.org/10.18653/v1/2025.findings-acl.976>.
25. Kudugunta, S.; Caswell, I.; Zhang, B.; Garcia, X.; Xin, D.; Kusupati, A.; Stella, R.; Bapna, A.; Firat, O. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 67284–67296.
26. Martin, G.; Mswahili, M.E.; Jeong, Y.S.; Woo, J. 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, WA, USA, 10–15 July 2022; pp. 303–313.
27. Adebara, I.; Elmadany, A.; Abdul-Mageed, M.; Alcoba Inciarte, A. SERENGETI: Massively Multilingual Language Models for Africa. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 1498–1537. <https://doi.org/10.18653/v1/2023.findings-acl.97>.
28. Buzaaba, H.; Wettig, A.; Adelani, D.I.; Fellbaum, C. Lugha-Llama: Adapting Large Language Models for African Languages. *arXiv* **2025**, arXiv:2504.06536.
29. Getao, K.; Miriti, E. Automatic construction of a kiswahili corpus from the world wide web. In *Measuring Computing Research Excellence and Vitality*; Fountain Publishers: Kampala, Uganda, 2006; p. 209.
30. Abisay, A.; Chen, M.K. Understand The Factors Affect Student's Academic Performance in Science Subjects in Public Secondary Schools in Tanzania. *N. Am. Acad. Res.* **2020**, *3*, 154–160. <https://doi.org/10.5281/zenodo.4091853>
31. Başaran, B. Rethinking AI in Education: A Comparative Study of Custom and External Systems. In Proceedings of the 2024 6th International Workshop on Artificial Intelligence and Education (WAIE), Tokyo, Japan, 28–30 September 2024; pp. 16–28.
32. Nyaaba, M. Globalizing Generative AI in Education for the Global South: The Design Case of 21st Century Teacher Educator AI for Ghana. *arXiv* **2025**, arXiv:2504.07149.
33. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* **1988**, *12*, 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7).
34. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>.
35. Ji, Y.; Meng, R.; Li, Z.; He, D. Curriculum Guided Reinforcement Learning for Efficient Multi Hop Retrieval Augmented Generation. *arXiv* **2025**, arXiv:2505.17391.
36. Swacha, J.; Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Appl. Sci.* **2025**, *15*, 4234.
37. Yadav, V.; Bethard, S.; Surdeanu, M. Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering. *arXiv* **2020**, arXiv:2005.01218.
38. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 1049–1065. <https://doi.org/10.18653/v1/2023.findings-acl.67>.

39. Shartiely, N.E. Code-switching in university classroom interaction: A case study of the University of Dar es Salaam. *Stellenbosch Pap. Linguist. Plus* **2016**, *49*, 215–231.
40. Ponti, E.M.; Glavaš, G.; Majewska, O.; Liu, Q.; Vulić, I.; Korhonen, A. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. *arXiv* **2020**, arXiv:2005.00333.
41. Liu, W.; Trenous, S.; Ribeiro, L.F.R.; Byrne, B.; Hieber, F. XRAG: Cross-lingual Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2505.10089.
42. Zhong, Z.; Liu, H.; Cui, X.; Zhang, X.; Qin, Z. Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2025; pp. 5756–5774.
43. Anderson, L.W.; Krathwohl, D.R., Eds. *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives*, 2nd ed.; Allyn & Bacon: New York, NY, USA, 2001.
44. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv* **2024**, arXiv:2402.03216.
45. Levonian, Z.; Li, C.; Zhu, W.; Gade, A.; Henkel, O.; Postle, M.E.; Xing, W. Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. *arXiv* **2023**, arXiv:2310.03184.
46. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Kolawole, T.; Fagbohunge, T.; Akinola, S.O.; Muhammad, S.H.; Kabongo, S.; Osei, S.; et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *arXiv* **2020**, arXiv:2010.02353.
47. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *arXiv* **2022**, arXiv:2208.03299.
48. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *arXiv* **2020**, arXiv:2004.12832.
49. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *arXiv* **2019**, arXiv:1907.10597.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.