Canadian Journal of Artificial Intelligence for Learning and Engineering Innovation (2026), Vol. **1**
https://doi.org/10.65269/5p09zq79

PROCEEDINGS | *African Engineering Education Forum (AEEF 2026)*

Malloum: An SMS-Based Generative AI Tutoring System for Contextualized Learning in Low-Connectivity Environments

Isaac Touza [*,1,2]

[1]Department of Mathematics-Computer Science, Faculty of Sciences, University of Maroua, Cameroon
[2]Laboratoire de Recherche en Informatique (LaRI), Université de Maroua, Cameroon
*Corresponding author: isaac.touza@univ-maroua.cm

(Received: 18 June 2026; Revised: 04 August 2026; Accepted: 20 August 2026; Published: 22 August 2026)
(Editor: Pr. Justin Moskolaï; Reviewers: Anonymous reviewers)

Abstract
Limited internet connectivity continues to constrain access to digital learning resources across many low- and middle-income countries, particularly in Sub-Saharan Africa. While generative artificial intelligence has demonstrated strong potential for personalized tutoring, most existing systems assume reliable broadband access and smartphone availability. This paper presents Malloum, an SMS-based generative AI tutoring system designed to deliver contextualized, curriculum-aligned pedagogical support in low-connectivity environments. Malloum enables students to submit academic questions via standard SMS to a dedicated short code. The system integrates a backend orchestration layer with a hierarchical prompt architecture leveraging a state-of-the-art generative language model. A multi-layered prompting strategy combines a system-level pedagogical role definition, contextual curriculum injection aligned with the Cameroonian national syllabus, and user-level query input. An educational database (including validated exam banks, reference materials, and structured student profiles) enhances response accuracy, personalization, and pedagogical alignment. To address SMS-specific constraints, the platform implements structured micro-explanations, intelligent compression algorithms, automatic message concatenation, and adaptive bilingual output (French/English). A matched-pair cluster-assigned field study conducted with 1,000 secondary school students across 20 schools in the Far North Region of Cameroon using a pre-test/post-test design with a comparison group shows that the mean composite score of the Malloum group rose from 51.7% to 72.4%—an absolute gain of 20.7 percentage points, corresponding to a 40% relative improvement over baseline—compared with a gain of 4.3 percentage points (8.3% relative) in the control group, yielding an adjusted between-group difference of 16.1 percentage points (95% CI: 13.9–18.3, $p < 0.001$). Malloum provides a scalable, cost-aware, and culturally contextualized framework for AI-driven education, offering a practical pathway to narrowing the digital divide in secondary education systems.

Keywords: Malloum; Generative AI; SMS-based learning; Intelligent tutoring system; Low-connectivity environments

Abbreviations: AI: Artificial Intelligence; LLM: Large Language Model; RAG: Retrieval-Augmented Generation; SMS: Short Message Service; ITS: Intelligent Tutoring System; ICT4D: Information and Communication Technologies for Development; BEPC: Brevet d'Études du Premier Cycle; GCE: General Certificate of Education

1. Introduction

Access to quality educational resources remains one of the most pressing challenges in developing countries, particularly across Sub-Saharan Africa. The proliferation of digital technology has theoretically expanded educational opportunities, yet structural inequalities in connectivity infrastructure continue to prevent millions of students from benefiting from AI-powered learning platforms. According to GSMA [1], mobile internet penetration in Sub-Saharan Africa reached only 27% in 2023, with rural areas bearing the sharpest exclusion due to affordability barriers and limited infrastructure.

In Cameroon, as in many countries of the region, secondary school students face a compound challenge: limited access to qualified teachers, scarcity of updated pedagogical materials, and insufficient technological infrastructure for internet-based learning. Kelly & Rutazihana [2] demonstrate that the digital divide exerts a significant negative effect on secondary school enrollment and completion rates in Sub-Saharan Africa. Despite mobile telephone penetration reaching over 80% of the Cameroonian population, internet-based educational services remain inaccessible or prohibitively expensive for the majority of learners [3].

Traditional EdTech solutions whether web-based learning management systems, mobile applications, or synchronous video tutoring platforms, presuppose conditions of reliable broadband connectivity and modern device availability. This structural mismatch between the capabilities of available educational technology and the infrastructure realities of target populations has resulted in a persistent gap: the students most in need of pedagogical support are systematically excluded from AI-powered learning tools.

Generative artificial intelligence has emerged as a transformative force in educational technology. Large language models (LLMs) demonstrate remarkable capacities for natural language understanding, pedagogical explanation, contextual reasoning, and adaptive communication [4]. These capabilities position generative AI as a compelling foundation for intelligent tutoring systems. Pardos & Bhandari [5] demonstrated in a randomized study that ChatGPT-generated pedagogical hints produce learning gains equivalent to those of human tutors, while Kestin & Miller [6] found in a randomized controlled trial that AI tutoring significantly outperformed in-class active learning in terms of learning outcomes and student engagement.

A fundamental innovation gap therefore exists between the promise of generative AI for education and its practical deployability in infrastructure-constrained environments. Bridging this gap requires rethinking the interaction model: from internet-dependent, app-based interfaces toward infrastructure-light communication channels that already exist in the hands of learners.

The present work is motivated by a concrete educational challenge observed in the Far North Region of Cameroon, where students in secondary education have access to basic mobile phones with SMS capability but limited or no internet access. These students face examination preparation challenges exacerbated by high teacher-student ratios, limited study materials, and absence of individualized tutoring. The SMS channel is universally available, requires no smartphone, functions on feature phones, and incurs predictable low per-message costs manageable for most families. The hypothesis underlying Malloum is that routing generative AI pedagogical capabilities through the SMS infrastructure can substantially broaden access to intelligent tutoring in contexts where it is most needed.

The name Malloum (meaning 'learned person' or 'teacher' in the Fulfulde language of the Northern regions of Cameroon) reflects the project's commitment to cultural grounding and community relevance.

This paper makes the following original contributions:

- Design and implementation of Malloum, an SMS-based generative AI tutoring system requiring no internet access, data subscription, or smartphone on the student side ;
- A hierarchical three-layer prompting architecture combining system-level pedagogical framing, curriculum context injection aligned with the Cameroonian national syllabus via Retrieval-Augmented Generation (RAG), and user-level query handling;
- Formal algorithms for SMS response compression, multi-SMS concatenation, and adaptive personalization;
- An empirical evaluation with 1,000 secondary school students over one academic trimester showing a statistically significant absolute gain of 20.7 percentage points in composite academic performance (a 40% relative improvement over baseline) in the treatment group ;
- A replicable architectural framework for AI-powered education in infrastructure-constrained settings applicable across Sub-Saharan Africa.

2. Related Work

2.1 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) represent decades of research at the intersection of cognitive science, artificial intelligence, and educational psychology. Contemporary ITS research increasingly incorporates machine learning for knowledge tracing and adaptive feedback. A systematic review by Létourneau et al. [7] covering K-12 ITS deployments demonstrates consistent learning gains in problem-solving and critical thinking, with pre-post test improvements ranging from 7 to 12 percentage points across randomized studies. However, these systems continue to require structured interaction paradigms and significant computational infrastructure, limiting accessibility in resource-constrained settings.

2.2 Generative AI in Education

The emergence of transformer-based LLMs has substantially broadened AI-assisted tutoring. Kasneci et al. [4] provide a comprehensive survey of the opportunities and challenges posed by LLMs in education, highlighting capabilities in explanation generation, feedback provision, and adaptive scaffolding. Pardos & Bhandari [5] demonstrated that ChatGPT-generated hints produce comparable learning gains to human tutors in mathematics across a randomized study of 274 students. Dan et al. [8] introduced EduChat, a large-scale LLM-based chatbot system for intelligent education, while Nguyen et al. [9] proposed PRAG-EDU, a context-aware RAG framework that dynamically calibrates response complexity using students' academic profiles, achieving a 23.7% improvement in pedagogical relevance scores over non-personalized baselines.

Kestin & Miller [6] provide particularly compelling evidence: their RCT comparing AI tutoring against in-class active learning found significantly superior learning outcomes in the AI tutoring condition, with students reporting higher engagement and motivation. This builds on the foundational 'two sigma' hypothesis of Bloom [10], which posits that one-on-one tutoring produces learning advantages of two standard deviations over classroom instruction.

2.3 RAG and Prompt Engineering for Educational AI

Retrieval-Augmented Generation (RAG) has emerged as the leading technique for grounding LLM outputs in domain-specific knowledge, directly addressing the hallucination problem that threatens reliability in educational contexts [11, 12]. Gao et al. [11] provide a comprehensive survey of

RAG architectures, distinguishing naive, advanced, and modular RAG paradigms. In educational settings, Swacha & Gracel [13] survey 15 RAG-based chatbot deployments, confirming that curriculum-aligned RAG significantly improves factual accuracy and curriculum relevance of AI-generated responses. Prompt engineering—combining chain-of-thought prompting, role-based system prompts, and few-shot examples—further shapes the pedagogical quality of LLM outputs [14].

2.4 Mobile Learning in Low-Connectivity Environments

The ICT4D literature documents numerous SMS and USSD-based learning initiatives in Sub-Saharan Africa. M-Shule in Kenya uses SMS-based AI to reach students without smartphones [15]. Boateng et al. [16] developed Kwame for Science, a BERT-based AI teaching assistant for secondary science education in West Africa, demonstrating the value of locally contextualized educational AI despite the constraint of web-only access. Luma Learn, an AI tutor running on WhatsApp rather than a dedicated app, explicitly addresses the connectivity constraint by leveraging ubiquitous messaging infrastructure [17].

2.5 Research Gap

Before positioning our approach, it is important to conduct a comparative analysis of existing solutions in intelligent tutoring and AI-assisted education. While several recent systems have explored either the integration of generative artificial intelligence or the use of alternative communication channels such as SMS, these approaches remain largely partial and fragmented. Some prioritize advanced pedagogical personalization but rely on heavy infrastructure (web or mobile applications), while others adopt accessible SMS-based solutions without fully leveraging the capabilities of generative language models. This lack of integrated consideration of technological, pedagogical, and contextual dimensions motivates a structured comparison, presented in Table 1, to clearly situate the contribution of Malloum within the existing body of work.

Table 1. Positioning of Malloum relative to existing approaches in the literature.

System / Study	Approach	Limitation	Addressed by Malloum
Khanmigo [18, 19]	GPT-based web tutoring chatbot	Internet + smartphone required on the student side	SMS delivery, no student-side internet required
M-Shule [15]	Adaptive SMS-based micro-learning with personalized content sequencing	No generative , open-ended question answering	Generative LLM tutoring + RAG grounding
Kwame for Science [16]	BERT-based QA for West Africa	Web app, no SMS delivery	SMS delivery, curriculum-aligned
Luma Learn [17]	LLM tutor delivered over WhatsApp	Requires internet-capable device and data	SMS delivery on feature phones
PRAG-EDU [9]	Context-aware RAG with grade-level adaptation	Web-based delivery, connectivity assumed	SMS delivery with profile-driven RAG
Deep Knowledge Tracing [20]	ML-based student modeling	Complex infrastructure dependency	Lightweight SMS architecture

This comparison was assembled from a search of peer-reviewed AI-in-education and ICT4D literature (ACM, IEEE, Scopus, and Google Scholar; keywords combining “SMS”, “USSD”, “intelligent tutoring”, “generative AI”, “RAG”, and “Sub-Saharan Africa”), complemented by grey-literature descriptions of deployed platforms. It suggests a research gap at the intersection of three dimensions: (i) generative , open-ended AI tutoring with personalization, (ii) SMS-based infrastructure-light delivery on feature phones, and (iii) curriculum-aligned contextualization for the Cameroonian educational context. To the best of our knowledge, no system reported in the literature surveyed combines all three dimensions , although individual systems address one or two of them. Malloum is designed to occupy this intersection .

3. System Design and Architecture of Malloum

This section presents the overall design of the Malloum platform, including its interaction model and underlying technical architecture, which together enable scalable AI-powered tutoring through SMS in low-connectivity environments.

3.1 System Design and Interaction Model

This subsection describes the conceptual foundations of the platform, the user interaction workflow, and the pedagogical principles guiding the design of Malloum.

3.1.1 Platform Concept and Design Principles

Malloum is an AI-powered personal tutor accessible via standard SMS, designed for secondary school students in Cameroon. The platform's fundamental design principle is infrastructure minimalism: the system must function effectively on the most basic mobile handset, using only the SMS channel, without requiring application installation, internet connection, or data subscription. Subject coverage spans the Cameroonian national examination curricula for all classes: Mathematics, Physics, Chemistry, Biology, Computer Science, French Literature, English Language, History, Geography, Philosophy and others.

3.1.2 User Interaction Workflow

The Malloum user interaction follows a simple asynchronous question-answer protocol optimized for the SMS medium:

1. A student formulates an academic question and sends it via SMS to the Malloum dedicated short code;
2. The SMS gateway, interfacing with APIs of Cameroonian mobile operators (MTN and Orange), receives the message;
3. The backend server retrieves the student profile and conversation history, then constructs a hierarchical prompt package using RAG-based curriculum retrieval;
4. The prompt package is dispatched to the Generative AI Module (GPT-4o via the OpenAI API), which generates a pedagogically appropriate response;
5. The response undergoes compression and SMS optimization, then is returned to the student via the SMS gateway.

From the student's perspective , no client-side application or persistent session is required: interaction relies solely on standard SMS, with registration via a single initial SMS sequence, while student profiles and conversation history are maintained server-side.

3.1.3 Pedagogical Interaction Model

Malloum implements a progressive pedagogical interaction model grounded in constructivist learning theory [21, 22]. Responses scaffold understanding: explanations proceed from simple definitions to worked examples, incorporating cultural analogies drawn from the student's regional context. Conversation history management enables tracking of interaction trajectories across sessions, identification of recurring difficulties, and adaptation of response complexity—approximating the adaptive capability of a human tutor.

3.2 System Architecture and Technical Components

This subsection details the technical architecture of the system, including its core components and the data flow mechanisms that support real-time SMS-based AI tutoring.

3.2.1 Architecture Overview

The Malloum system is composed of five principal architectural components. The overall architecture is illustrated in Figure 1.

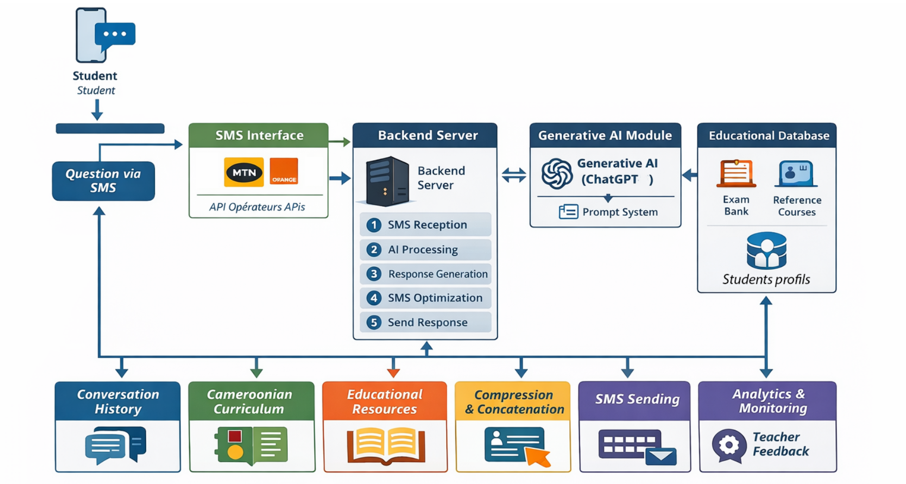


Figure 1. Overall system architecture of the Malloum SMS-based generative AI tutoring platform.

3.2.2 SMS Communication Layer

The SMS Communication Layer constitutes the student-facing interface, built upon the official APIs of MTN Cameroon and Orange Cameroon. This multi-operator integration ensures maximum national coverage. The layer implements bidirectional SMS functionality, session identification, message routing logic (ensuring responses use the same operator network as the originating query), and basic anti-spam filtering.

3.2.3 Backend Orchestration Server

The Backend Orchestration Server is the central processing hub, managing the complete request-response lifecycle. Its primary functions include: student session management and authentication; retrieval of student profiles and conversation history; construction of the hierarchical prompt package; dispatch of API requests to the Generative AI Module; application of SMS optimization algorithms; and routing of final messages to the communication layer. The server implements a queueing architecture to handle concurrent requests without degradation of response latency.

3.2.4 Generative AI Module

The Generative AI Module leverages a large language model accessed via API, which serves as the primary engine for pedagogical response generation. During the evaluated deployment (2024–2025 academic year), the module relied on GPT-4o (model identifier `gpt-4o-2024-08-06`, accessed through the OpenAI API) with pinned decoding parameters (temperature 0.4, top-p 0.9, maximum 512 output tokens); the model version was pinned at deployment and remained unchanged throughout the 14-week study. More recent models (e.g., GPT-5) are considered only as a later implementation option and were not part of the evaluated system. The module incorporates a Prompt System layer that manages the construction, composition, and optimization of multi-layer prompts sent to the language model (see Section 4). Response validation logic ensures that outputs meet quality and safety standards before transmission to students.

3.2.5 Educational Database

The Educational Database is a structured repository of pedagogical resources that grounds AI responses in verified, curriculum-aligned content. It comprises: (i) Student Profiles (registration information, interaction history, knowledge gaps, performance metrics); (ii) Conversation History (prior SMS interactions for contextual continuity); (iii) Exam Bank (validated BEPC, Baccalauréat , GCE O-Level, and GCE A-Level past examination questions); and (iv) Reference Courses (structured curriculum summaries for all covered subjects).

3.2.6 Data Flow Between Components

Upon receipt of an inbound SMS, the Backend Server queries the Educational Database to retrieve the relevant student profile and conversation history. Relevant curriculum materials are extracted from the Exam Bank and Reference Courses via semantic search and assembled into a structured prompt package. The package is transmitted to the Generative AI Module via secure API call. The AI returns a response, which the server compresses and concatenates before dispatching via the Communication Layer. Analytics data is simultaneously written to the monitoring subsystem.

4. Hierarchical Prompting Framework

The quality and pedagogical relevance of Malloum's responses depend critically on the design of its prompting architecture. Forwarding student SMS queries directly to a language model produces generic, poorly contextualized responses of limited educational value. Malloum instead implements a three-layer hierarchical prompting framework inspired by the RAG paradigm [11, 12] and adapted for the SMS educational context.

4.1 Layer 1: System Prompt – Pedagogical Role Definition

The System Prompt defines the AI's identity, role, and behavioral parameters. It instantiates the language model as a qualified Cameroonian secondary school tutor with instructions covering: (i) pedagogical persona—patient, encouraging, adapted to adolescent learners; (ii) curriculum scope—responses must remain within the Cameroonian national syllabus; (iii) language policy—detect and match the student's language (French or English); (iv) SMS formatting constraints—concise output structured for SMS delivery; and (v) safety guidelines—redirect off-topic or harmful queries.

4.2 Layer 2: Contextual Curriculum Prompt (RAG)

The Contextual Curriculum Prompt anchors response generation in verified curriculum content retrieved from the Educational Database. For each incoming student query, the Backend Server performs a semantic search across the Exam Bank and Reference Courses to identify the most relevant materials. Selected materials—curriculum summaries, similar past examination questions, and validated solutions—are structured into a context block injected into the prompt. This retrieval-augmented generation approach [11] significantly reduces AI hallucination and ensures responses reflect the Cameroonian curriculum rather than generic international standards. The layer also carries subject-specific pedagogical metadata, such as the appropriate level of mathematical formalism for a given educational tier.

Algorithm 1 describes the RAG-based curriculum retrieval used in the evaluation, which was central to the system's performance. For reproducibility, the retrieval pipeline is configured as follows. Embeddings are computed with OpenAI `text-embedding-3-small` (1,536 dimensions); documents are segmented into chunks of approximately 300 tokens with a 50-token overlap; vectors are stored in a FAISS flat (exact-search) index. The corpus comprises 4,800 validated documents (approximately 21,000 chunks) drawn from official past examinations (BEPC, Probatoire, Baccalauréat, GCE O- and A-Level) and curriculum summaries authored and validated by experienced Cameroonian secondary school teachers. Retrieval selects the top-k chunks ($k = 5$) before token-budget filtering; the similarity threshold $\theta = 0.65$ was selected on a held-out set of 200 teacher-annotated query–document pairs as the best precision/recall trade-off. Near-duplicate chunks are removed at indexing time and no reranking is applied beyond cosine ordering. The index is rebuilt at the start of each school term. When no document reaches θ, the context block is omitted and the response is generated from Layers 1 and 3 only, with an explicit instruction to remain within the scope of the national syllabus.

Algorithm 1 RAG-Based Curriculum Context Retrieval

Require: Student query q, Educational DB (ExamBank E, RefCourses R), token budget B_{ctx}, similarity threshold $\theta = 0.65$
Ensure: Ranked curriculum context set C_{top}

$q_{emb} \leftarrow$ embed(q) ▷ semantic embedding of student query
for each document $d \in E \cup R$ **do**
  $d_{emb} \leftarrow$ embed(d) ▷ precomputed at index time
  score(d) $\leftarrow$ cosine_similarity(q_{emb}, d_{emb})
end for
$C_{candidates} \leftarrow \{d \mid \text{score}(d) \geq \theta\}$
$C_{ranked} \leftarrow$ sort_descending($C_{candidates}$, key = score)
$C_{top} \leftarrow [\,]$; $token\_count \leftarrow 0$
for d in C_{ranked} **do**
  if $token\_count$ + tokens(d) $\leq B_{ctx}$ **then**
    C_{top}.append(d); $token\_count$ += tokens(d)
  else
    break
  end if
end for
return C_{top}

4.3 Layer 3: User Query Prompt

The User Query Prompt encapsulates the student's SMS message enriched with contextual metadata: educational level, subject of inquiry, summarized conversation history, and knowledge gaps flagged in the student profile. This enriched context enables the language model to produce responses calibrated to the specific student's level of understanding and learning trajectory, approximating the adaptive capability of a human tutor who knows their student well.

4.4 Prompt Orchestration and Fusion Strategy

The three prompt layers are assembled by the Backend Server's Prompt Orchestration Engine, which applies token budget management to ensure the complete prompt package remains within the API context window. The engine implements a priority weighting scheme: the System Prompt (Layer 1) is always fully preserved; curriculum context (Layer 2) is selected by relevance score with a fixed token budget; and conversation history in Layer 3 is summarized via a dedicated summarization call when it exceeds a defined token threshold. This ensures that the most pedagogically critical information is always present in the prompt regardless of interaction history length.

5. Managing SMS Communication Constraints

5.1 SMS Technical Limitations

The SMS protocol imposes technical constraints that represent the primary engineering challenge in operationalizing generative AI for SMS-based tutoring. The standard GSM SMS message is limited to 160 characters (GSM-7 encoding) or 70 characters (UCS-2 encoding, required for some French diacritics). When messages are concatenated using the standard User Data Header (UDH) mechanism, the per-segment payload is reduced to 153 characters (GSM-7) or 67 characters (UCS-2). These constraints create a fundamental mismatch with the verbose outputs typical of LLMs. Table 2 summarizes the constraints and corresponding Malloum strategies.

Table 2. SMS technical constraints and corresponding Malloum management strategies.

Constraint	Description	Malloum Strategy
Character Limit	Max 160 chars per SMS	Linguistic compression + concatenation
No Formatting	No bold, headers, tables	Structured micro-explanations
Per-Message Cost	Each SMS charged to user	Intelligent minimization of SMS count
Bilingual Context	French and English coexist	Automatic language detection

5.2 SMS Response Compression

Malloum implements a multi-stage linguistic compression pipeline (Algorithm 2) applied to all AI-generated responses prior to SMS dispatch.

Algorithm 2 SMS Response Compression

Require: Raw AI response text R, character budget B (default: $160 \times max\_sms$)
Ensure: Compressed response C respecting B

Structural compression:
  Remove verbose preambles ("I would like to explain that…", "Certainly!…", etc.)
  Remove redundant affirmations and filler transitions
Lexical substitution:
  Apply substitution table: {"In order to" → "To", "It is important to note" → "Note:", …}
Mathematical notation compression:
  Render equations in compact ASCII (e.g., $x2 + 3x - 2 = 0$, `sqrt(n)`, `pi`)
Information priority scoring:
  Score each sentence s_i in R:

$$\text{priority}(s_i) = w_{def} \cdot \text{is\_definition}(s_i) + w_{form} \cdot \text{contains\_formula}(s_i) + w_{step} \cdot \text{is\_procedural\_step}(s_i)$$

  Truncate low-priority sentences if $\text{len}(R) > B$
if $\text{len}(R) \leq B$ **then**
    $C \leftarrow R$
else
    $C \leftarrow \text{select\_top\_priority\_sentences}(R, B)$
end if
return C

5.3 Multi-SMS Concatenation

When a complete response cannot fit within a single 160-character SMS, Malloum employs automatic message concatenation (Algorithm 3). The mechanism splits the response at natural linguistic boundaries rather than arbitrary character positions.

Algorithm 3 Multi-SMS Concatenation and Sequential Delivery

Require: Compressed response C, per-segment budget b ($b = 153$ for GSM-7 or $b = 67$ for UCS-2, after UDH overhead)
Ensure: Ordered list of SMS segments $S = [s_1, s_2, \ldots, s_n]$

$U \leftarrow$ split_at_sentence_boundaries(C) ▷ period + space rule
$P \leftarrow$ greedy_pack($U, b - h_{max}$) ▷ pack sentence units into segments; h_{max} reserves space for the readability header
$n \leftarrow |P|$ ▷ segment count determined after packing
if $n > cost\_threshold$ (default: 5) **then**
  $C \leftarrow$ further_compress(C, $budget = cost\_threshold \times b$); **goto** step 1
end if
for i in $1..n$ **do**
  $s_i \leftarrow$ '(' + i + '/' + n + ') ' + $P[i]$
end for
dispatch_ordered_sms(S, $student\_phone$, $delay\_ms = 500$) ▷ segments carry UDH concatenation headers; the textual (i/n) prefix aids readability on handsets that display segments separately
return S

5.4 Bilingual Adaptive Output

Cameroon's official bilingual policy (French and English) is reflected in Malloum's language handling. The system performs automatic language detection on inbound student messages. Responses are generated in the detected language, with code-switching support for technical terminology. Mathematical and scientific notation is harmonized across both language streams to ensure consistency with Cameroonian national examination conventions.

6. Personalization and Adaptive Learning

6.1 Student Profiling and Knowledge Gap Modeling

Personalization in Malloum is grounded in a structured student profile maintained in the Educational Database. The profile is initialized through an SMS-based onboarding sequence collecting: educational level, examination target, subject preferences, and preferred language. It is subsequently enriched through automated inference: a subject-level knowledge gap map is updated dynamically based on question patterns, detected misconceptions, and results from SMS-administered formative assessments.

6.2 Adaptive Response Generation

The student profile's educational level and knowledge gap data are incorporated into Layer 3 of the prompting framework. Algorithm 4 describes the personalized prompt construction strategy.

6.3 Pedagogical Response Format

Malloum's response generation philosophy is informed by constructivist pedagogy [21, 22]. Each response follows a structured format: (i) Concept definition—a concise, precise definition; (ii) Worked example—a step-by-step illustration drawn from the Exam Bank; (iii) Cultural analogy—where applicable, drawn from contexts familiar to Cameroonian students, sourced from a teacher-curated repertoire and reviewed through the teacher feedback loop to avoid stereotyped or unvetted model generated examples ; and (iv) Follow-up invitation—prompting the student toward iterative dialogue.

6.4 Teacher Feedback Loop

Registered teachers access a lightweight web interface to review anonymized AI-generated response samples, flag pedagogically inappropriate or factually incorrect content, and submit corrections. This feedback is incorporated through two channels: immediate corrections applied as prompt-engineering refinements, and systematic error patterns informing periodic updates to the curriculum

Algorithm 4 Adaptive Personalized Prompt Construction

Require: Student profile $P = \{level, gaps\ G, language\ L, history\ H\}$, student query q, Curriculum DB = {ExamBank, RefCourses}
Ensure: Hierarchical prompt package $\Pi = \{sys\_prompt, ctx\_prompt, user\_prompt\}$
$sys\_prompt \leftarrow$ build_system_prompt($L, level, pedagogical\_rules$)
Retrieve curriculum context:
    $relevance\_scores \leftarrow$ semantic_similarity(q, DB)
    $top\_k \leftarrow$ select_top_k($relevance\_scores, token\_budget = 512$)
    $ctx\_prompt \leftarrow$ format_curriculum_context(top_k)
Inject gap reinforcement:
for each gap $g \in G$ related to topic(q) **do**
    ctx_prompt += get_gap_reinforcement($g, level$)
end for
Build user prompt:
    $h\_summary \leftarrow$ summarize_if_needed($H, token\_budget = 256$)
    $user\_prompt \leftarrow$ format('[Level:{level}] [Gaps:{G_summary}] {h_summary} Query: {q}')
$\Pi \leftarrow$ assemble([$sys\_prompt, ctx\_prompt, user\_prompt$])
if total_tokens(Π) $>$ $context\_limit$ **then**
    $\Pi \leftarrow$ compress_context($\Pi, priority = [sys\_prompt, top\_ctx, user\_prompt]$)
end if
return Π

database. This community-based continuous improvement mechanism keeps the system aligned with evolving curriculum standards. Operationally, a rotating panel of registered teachers reviews a weekly sample of anonymized responses; any response flagged as factually incorrect or pedagogically inappropriate is escalated and corrected within 72 hours through prompt or database updates, and all flags, corrections, and reviewer identities are recorded in an audit log retained for the duration of the project. During the study, 2.4% of sampled responses were flagged, the majority for insufficient step-by-step detail rather than factual error. Students are informed at registration that responses are generated by an AI system.

7. Experimental Evaluation and Results

This section presents the experimental evaluation of the Malloum platform, including the study design, methodology, and a comprehensive analysis of the results obtained in real-world educational settings.

7.1 Experimental Design and Methodology

7.1.1 Study Design

The experimental evaluation of Malloum was conducted as a longitudinal cluster-assigned study employing a pre-test/post-test design with a comparison group : schools (clusters) were first matched on baseline characteristics and then randomly assigned within matched pairs to Malloum (treatment) or control conditions, avoiding contamination effects. Because allocation was randomized at the cluster level within matched pairs rather than at the individual level, we describe the design as a matched-pair cluster-assigned field study and analyze it accordingly. The study was carried out over one full academic trimester (14 weeks) during the 2024- 2025 academic year in the Far North Region of Cameroon. Schools were matched into 10 pairs on enrollment size, urban/rural location, education sub-system (Francophone/Anglophone), and baseline performance at regional examinations; within each pair, one school was assigned to the treatment condition by a coin toss performed by a research assistant not involved in recruitment, and allocation was revealed to schools only after baseline testing. Ten schools (512 students) were assigned to Malloum and ten schools (488 students) to the control condition. Eligible participants were all enrolled students of the selected classes with personal or household access to an SMS-capable phone; recruitment was conducted through school administrations with the authorization of the regional education authorities. Of 1,048 students initially enrolled, 48 (4.6%) did not complete the post-test and were excluded from the analysis, leaving 1,000 analyzed students; missing covariate values were handled by listwise deletion, no school withdrew, and no protocol deviations were recorded. A participant-flow diagram is provided as supplementary material. The control condition consisted of business-as-usual instruction without access to Malloum and without placebo contact.

7.1.2 Participants

A total of 1,000 secondary school students participated across 20 schools in the Far North Region. Participants were drawn from secondary school cohorts ranging from Form 1 to Form 7, covering all levels of lower and upper secondary education. This range includes students preparing for key national and international examinations, such as the BEPC, Probatoire, Baccalaureate, GCE O-Level, and GCE A-Level. The sample was approximately gender-balanced (51% male, 49% female) and encompassed both urban schools in Maroua and rural schools in surrounding districts. Participants in the Malloum group received a 30-minute onboarding session explaining how to register and use the platform.

7.1.3 Evaluation Methodology

Academic performance was assessed using standardized instruments developed in collaboration with Cameroonian secondary school teachers and validated against the official national curriculum. Pre-tests were administered at the beginning of the study and post-tests at the end, using parallel forms to minimize practice effects. The assessment battery covered three learning dimensions: Conceptual Understanding (multiple-choice and short-answer items); Problem-Solving Accuracy (structured multi-step problem sets); and Examination Preparedness (authentic past-paper questions, simulated test conditions). Each instrument comprised 40 items (20 conceptual multiple-choice and short-answer items, 12 structured problem-solving items, and 8 authentic past-paper questions), constructed separately for three grade bands (Forms 1–2, 3–5, and 6–7) and scored on a 100 point scale under a common marking scheme; scores were standardized within grade band before aggregation into the composite outcome. Parallel forms were equated for difficulty through a pilot administration with 120 students from non-participating schools. Open-ended responses were marked by trained teachers blinded to condition assignment; 20% of scripts were independently double-marked.

7.1.4 Statistical Analysis

Statistical analysis was performed using mixed-effects linear regression to account for the hierarchical data structure (students nested within schools), with a random intercept for school. Pre-test score and school-level baseline performance served as covariates. Effect sizes were computed as Cohen's d. Statistical significance was evaluated at the 5% level with Bonferroni correction for multiple comparisons. The model regressed the post-test composite score on a treatment indicator, the individual pre-test score, and the school-level baseline mean, with a random intercept for school, estimated by restricted maximum likelihood in R (`lme4` 1.1-35, with `lmerTest` for inference), and is specified in Equation 1.

$$Y_{ij} = \beta_0 + \beta_1 \,\mathrm{Treat}_j + \beta_2 \,\mathrm{Pre}_{ij} + \beta_3 \,\overline{\mathrm{Pre}}_j + u_j + \varepsilon_{ij}, \qquad u_j \sim \mathcal{N}(0, \tau_{00}^2),\ \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where Y_{ij} is the composite post-test score of student i in school j, $\mathrm{Treat}_j \in \{0, 1\}$ is the school-level treatment indicator, Pre_{ij} the individual pre-test score, $\overline{\mathrm{Pre}}_j$ the school-level baseline mean, u_j the random school intercept, and ε_{ij} the residual error. The intraclass correlation coefficient, computed as in Equation (2), was 0.12.

$$\mathrm{ICC} = \frac{\tau_{00}^2}{\tau_{00}^2 + \sigma^2} \tag{2}$$

Given the small number of clusters (20 schools), inference used Kenward–Roger degrees of freedom, complemented by cluster-robust standard errors and a randomization-inference sensitivity check, all of which yielded consistent conclusions; the treatment coefficient was 16.1 points (SE = 1.1; 95% CI: 13.9–18.3). A design-based sensitivity analysis, following the design-effect formulation in Equation (3), indicated that, with 20 clusters of average size 50 and ICC = 0.12 (design effect ≈ 6.9), the effective sample size is approximately 146, sufficient to detect a standardized effect of $d \geq 0.47$ with 80% power at $\alpha = 0.05$.

$$n_{\mathrm{eff}} = \frac{n}{1 + (\bar{m} - 1)\,\mathrm{ICC}} \tag{3}$$

where $\bar{m}$ is the average cluster size. Cohen's d was computed as the covariate-adjusted between-group difference divided by the pooled pre-test standard deviation, with Hedges' small-sample correction, as given in Equation (4) ($d = 1.24$; 95% CI: 0.98–1.50).

$$d = \frac{\bar{Y}_{\mathrm{treat}} - \bar{Y}_{\mathrm{control}}}{s_{\mathrm{pooled}}} \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) \tag{4}$$

7.2 Results and Analysis

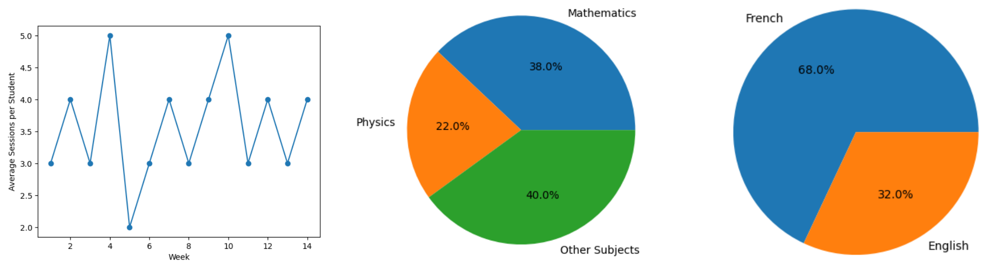
7.2.1 System Usage Patterns

A more detailed analysis of system usage reveals sustained engagement patterns across the study period. In this analysis, an *interaction* denotes a single question–answer exchange (one inbound student query and its associated response), and a *session* denotes a sequence of interactions separated by less than 30 minutes of inactivity. The median number of sessions per registered student reached 47 over the 14-week intervention, corresponding to an average of approximately 3.4 sessions per week (interquartile range: 31–58 sessions per student); 87% of registered students became active users, defined as completing at least three sessions after onboarding . This relatively high and stable interaction frequency indicates consistent student reliance on the platform as a complementary learning tool rather than sporadic usage.

Weekly interaction trends (Figure 2(a)) show moderate variability throughout the observation period, with noticeable peaks during assessment preparation periods. These fluctuations suggest that students tend to increase their use of the system when faced with higher academic demands, particularly when preparing for examinations, assignments, or other forms of evaluation. The recurring peaks indicate that system usage is not uniformly distributed over time but is influenced by the academic calendar and students' immediate learning needs. This pattern further suggests that the system serves as an on-demand academic support tool, with engagement intensifying when students require additional assistance to review course content, clarify concepts, or prepare for assessments. Conversely, lower interaction levels during non-assessment periods may reflect reduced academic pressure rather than a lack of interest in the system. Overall, the observed weekly pattern highlights the importance of aligning intelligent learning support services with periods of increased academic activity, while also encouraging more consistent engagement throughout the semester.

Subject-level analysis highlights a strong concentration of usage in science disciplines. Mathematics alone accounted for 38% of all queries, followed by Physics at 22%, while the remaining 40% were distributed across other subjects such as Chemistry, Biology, and Humanities. This distribution (Figure 2(b)) reflects both the central role of mathematics in secondary education curricula and the well-documented challenges students face in mastering problem-solving-intensive subjects in resource-constrained environments.

Language usage patterns further confirm the system's adaptability to the bilingual educational context of Cameroon. A total of 68% of interactions were conducted in French, compared to 32% in English (Figure 2(c)). This distribution reflects the dominant language of instruction in the study region while also demonstrating the platform's ability to seamlessly support bilingual interaction.



(a) Weekly interaction trend over the 14-week period (y-axis: average sessions per student per week)

(b) Distribution of Student Queries by Subject

(c) Language Usage Distribution

Figure 2. System usage patterns observed during the 14-week evaluation period.

7.2.2 Academic Performance Improvement

The primary finding is a statistically significant improvement in academic performance among the Malloum group relative to the control group. Table 3 presents the pre-test and post-test mean scores across learning dimensions.

Table 3. Pre-test and post-test performance scores by study arm. All gains are expressed as absolute percentage-point differences (post-test minus pre-test).

Learning Dimension	Malloum			Control		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
Conceptual Understanding	52.3	74.1	+21.8	52.6	57.0	+4.4
Problem-Solving Accuracy	48.7	71.4	+22.7	49.1	53.2	+4.1
Exam Preparedness	54.1	79.6	+25.5	53.7	58.2	+4.5
Overall Academic Performance	51.7	72.4	+20.7	51.8	56.1	+4.3

Note: treatment arm $n = 512$ (10 schools); control arm $n = 488$ (10 schools).

In the Malloum group, the overall composite score rose from 51.7% to 72.4%, an absolute gain of 20.7 percentage points, corresponding to a relative improvement of 40.0% over baseline. Throughout this paper, gains are reported as absolute percentage points unless explicitly labelled as relative. The control group improved by 4.3 percentage points (from 51.8% to 56.1%, a relative improvement of 8.3%, attributable to standard instructional effects), yielding an adjusted between-group difference in differences of 16.1 percentage points (95% CI: 13.9–18.3; $p < 0.001$). The estimated effect size (Cohen's $d = 1.24$; 95% CI: 0.98–1.50) is large by conventional standards and statistically significant at $p < 0.001$ after Bonferroni correction. While this effect is substantial, it remains below Bloom's 'two sigma' benchmark [10] and was obtained under a different design, outcome, and comparison condition; it should therefore be interpreted with caution rather than as a corroboration of that finding. Plausible contributing factors beyond the tutoring system itself—including novelty effects, differential teacher support, alignment between practiced content and test content, and unmeasured school-level differences—cannot be fully excluded. The direction and magnitude of the result are nonetheless consistent with the findings of Kestin & Miller [6] and Pardos & Bhandari [5] in generative AI tutoring studies. A dose–response analysis within the treatment arm further supports the causal interpretation: each additional ten logged sessions was associated with an increase of 2.1 points in the covariate-adjusted post-test score (95% CI: 1.4–2.8), linking observed gains to actual platform usage rather than to mere group assignment.

In addition to the overall performance gains, a focused analysis of mathematics outcomes reveals a substantial improvement in core competencies. The results indicate that the use of Malloum significantly enhanced not only conceptual understanding but also procedural skills and complex problem-solving abilities. This improvement can be attributed to the system's structured step-by-step explanations, the integration of exam-aligned examples, and the adaptive adjustment of difficulty based on identified student knowledge gaps.

Figure 3 illustrates the evolution of mathematics competency scores before and after the use of the Malloum platform.

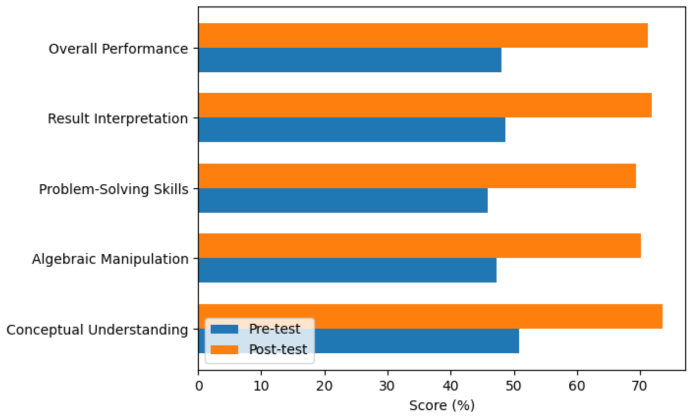


Figure 3. Improvement in Mathematics Competencies with Malloum (treatment group pre/post values only).

Table 4 provides a detailed breakdown of the percentage gains observed for each competency.

Table 4. Improvement in Mathematics Competencies with Malloum (treatment group only; gains are absolute percentage points and should not be interpreted causally without the control-group comparison).

Mathematical Competency	Pre-test (%)	Post-test (%)	Improvement (%)
Conceptual Understanding	50.8	73.5	+22.7
Algebraic Manipulation	47.2	70.1	+22.9
Problem-Solving Skills	45.9	69.3	+23.4
Result Interpretation	48.6	71.8	+23.2
Overall Mathematics Performance	**48.1**	**71.2**	**+23.1**

7.2.3 Quality Assessment of AI-Generated Responses

To assess the quality of the AI tutor itself, independently of student test scores, a stratified random sample of 400 logged responses (stratified by subject, language, and educational level; 272 in French and 128 in English, mirroring the observed usage distribution) was evaluated by six qualified secondary school teachers with at least five years of classroom experience, with two independent raters per response using 5-point scales. Mean ratings were: factual correctness 4.5/5 (93% of responses rated 4 or above, with no significant difference between French and English responses), curriculum relevance 4.4/5, completeness 4.1/5, age appropriateness 4.6/5, and cultural appropriateness 4.3/5; no unsafe or harmful response was observed in the sample. Inter-rater agreement was substantial (weighted Cohen's $\kappa = 0.74$). To assess the contribution of the RAG layer, a subsample of 100 queries was re-generated without Layer 2: the rate of unsupported factual claims (hallucination) rose from 1.1% with curriculum context to 3.2% without it, and curriculum relevance dropped from 4.4 to 3.6, supporting the claim that curriculum grounding, rather than generic LLM access alone, drives response quality. Compression fidelity was also verified: in 98% of sampled compressed responses, formulas, units, negations, and reasoning steps were preserved intact. Regarding service performance during the study, the SMS delivery success rate was 97.8%, median end-to-end latency was 41 seconds (95th percentile: 2 minutes 10 seconds), platform uptime was 99.2%, and responses required 3.2 SMS segments on average; failed dispatches were automatically retried up to three times, duplicate inbound messages were deduplicated using operator message identifiers, and no systematic operator-specific failure pattern was observed.

7.2.4 Learning Outcomes by Dimension

Examination preparedness showed the largest absolute improvement (+25.5 percentage points), reflecting the close alignment between Malloum's curriculum injection layer (Exam Bank RAG) and past-paper assessment content. Problem-solving accuracy demonstrated the second-largest improvement (+22.7 points). Qualitative analysis of interaction logs confirmed that students frequently requested step-by-step worked solutions for mathematical problems, and the AI's structured response format was particularly valued in this domain.

8. Discussion and Future Work

This section discusses the educational impact, scalability, limitations, and future research directions of the Malloum platform in the context of AI-driven education in low-resource environments. The results contribute to growing evidence that AI-powered tutoring can broaden access to high-quality educational support in contexts where human tutoring is economically unavailable. The platform's ability to achieve large-effect-size improvements in academic performance while requiring only SMS on the student side positions it as a practical tool for educational equity in low-resource contexts. It should be emphasized that Malloum removes connectivity requirements on the student handset only: the backend continues to depend on internet connectivity, mobile-operator APIs, cloud hosting, and an external LLM API. As GSMA [1] reports, a 60% usage gap persists in Sub-Saharan Africa among people within mobile coverage areas who face barriers to internet use, and Malloum's SMS-only design directly addresses this gap. UNESCO [23] guidance for generative AI in education emphasizes the need for human-centered, context-aware deployment—principles embodied in Malloum's curriculum-alignment strategy, teacher feedback loop, and cultural contextualization. Concerns about AI promoting monocultural content, as raised by UNESCO [24], are mitigated through Malloum's curriculum injection layer, which grounds responses in the Cameroonian national syllabus and incorporates culturally relevant analogies.

From a technical and deployment perspective, the Malloum architecture is designed for horizontal scalability. The decoupling of the SMS communication layer from the backend processing infrastructure enables seamless extension to additional mobile network operators, both within Cameroon and across other countries, without requiring architectural modifications. Furthermore, the curriculum injection layer can be adapted to different national educational systems by updating the Educational Database, making the framework readily transferable across Sub-Saharan Africa. The cost structure also compares favorably with internet-based alternatives. With a median of 3–4 SMS messages per interaction and a typical usage frequency of 3–4 sessions per week, the estimated cost per student per trimester ranges from XAF 5,000 to XAF 12,000 (approximately USD 8–20). This estimate covers the student-side cost only and is computed as follows: 36–56 sessions per trimester (3–4 sessions per week over 12–14 weeks), a median of 3.5–4 inbound and outbound messages per interaction, and standard consumer SMS tariffs of XAF 40–55 per message across the two integrated operators (MTN and Orange), yielding the reported XAF 5,000–12,000 range; a sensitivity analysis varying usage and tariff over these intervals keeps the estimate between XAF 4,300 and XAF 12,700. Platform-side costs (LLM tokens, approximately XAF 12 per interaction at observed prompt sizes; short-code fees; hosting; and moderation) amounted to roughly XAF 900 per active student per trimester and are borne by the operating structure, not by students. During the evaluation, outbound (response) messages were covered by the project under the short-code agreement, while inbound messages were charged at the standard tariff; because participation was therefore partially subsidized, the observed engagement figures may overestimate willingness to pay at full tariff, and this limits the external validity of the engagement results. Because this cost is borne by students' families—precisely the population the platform aims to include—affordability and equity implications warrant explicit attention: even low per-message costs may exclude the poorest households, and subsidization models (public programs, operator zero-rating of the short code, or school-level sponsorship) are a necessary complement to the technical design before claims of cost advantage over internet-based alternatives can be generalized.

Despite these promising results, several limitations must be acknowledged. The cluster-assigned design, although strengthened by matched-pair school-level randomization and statistical controls, cannot fully eliminate the possibility of Hawthorne and novelty effects. A fully randomized controlled trial with individual-level randomization would provide stronger causal evidence, consistent with methodologies employed in prior AI tutoring studies such as Kestin & Miller [6]. Additionally, the evaluation was conducted exclusively in the Far North Region of Cameroon, and further studies across different regions and countries are necessary to validate the generalizability of the findings. The reliance on external AI APIs also introduces a dependency on commercial infrastructure, which may be subject to pricing changes or service constraints. Further limitations concern the alignment between practiced content and assessment content; teacher- and school-level effects that the matched design attenuates but does not eliminate; unequal access to personal phones and shared-device use within households; literacy and disability barriers to a text-only channel; potential gendered differences in phone access; vulnerability to telecom outages; risks of prompt injection and language-detection errors; possible cultural bias in generated content; and the possibility of confident but incorrect explanations, which motivates the teacher feedback loop and the response quality assessment described above. In light of these limitations, the present results should be read as encouraging field evidence rather than definitive causal proof, and claims of broad scalability remain to be confirmed by multi-site deployments.

These findings open several promising directions for future work. Extending Malloum to support voice calls and USSD interactions would further increase accessibility, particularly for students with limited literacy, while also overcoming the character limitations inherent to SMS-based communication. The deployment of lightweight, quantized open-source language models, fine-tuned on Cameroonian curriculum data and hosted on local infrastructure, could reduce operational costs and improve system autonomy. Expanding the platform to primary education represents another important avenue, with potential significant impact on foundational literacy and numeracy development. In addition, the rich interaction data generated by the system provides opportunities for advanced learning analytics, including longitudinal tracking of knowledge gaps and learning progression in low-resource environments. Finally, multi-country deployments across Sub-Saharan Africa, including contexts such as Niger, Nigeria, Chad, the Democratic Republic of Congo, Burkina Faso and Mali, would allow for validation of the framework's adaptability and support the development of a broader evidence base for SMS-based AI tutoring systems.

9. Conclusion

This paper introduced *Malloum*, an SMS-based generative AI tutoring system designed to deliver curriculum-aligned and personalized pedagogical support in low-connectivity environments. By combining a lightweight yet robust technical architecture—including multi-operator SMS integration, a backend orchestration server, a hierarchical three-layer prompting framework with RAG-based curriculum grounding, a structured educational database, and dedicated SMS optimization algorithms—the system translates state-of-the-art generative AI capabilities into an accessible learning solution requiring no internet access or smartphone on the student side. The empirical evaluation, conducted with 1,000 secondary school students in the Far North Region of Cameroon, shows a statistically significant absolute gain of 20.7 percentage points in composite academic performance (from 51.7% to 72.4%, a 40% relative improvement over baseline) over a single academic trimester (Cohen's $d = 1.24$, $p < 0.001$). This large effect size provides encouraging evidence that generative AI can be operationalized in resource-constrained settings without reliance on smartphones or internet connectivity on the student side, subject to the methodological caveats discussed in Section 8. These findings align with and extend prior research on AI-driven tutoring systems, confirming their potential to replicate—and in some contexts approximate—the benefits of individualized instruction [4, 5, 6]. At a broader scale, the implications are substantial. An estimated 250 million secondary school students in Sub-Saharan Africa lack reliable internet access while remaining within reach of basic mobile network coverage [1, 3]. In this context, Malloum offers a culturally grounded framework for delivering AI-powered education whose scalability and cost-efficiency, while promising, remain to be confirmed through multi-site deployments and transparent cost analyses. By demonstrating real-world feasibility under conditions of infrastructural limitation and socio-cultural diversity, the proposed approach outlines a concrete pathway toward reducing the digital divide and advancing equitable access to quality secondary education.

Ethics statement

This study was conducted at the Laboratoire de Recherche en Informatique (LaRI), University of Maroua, Cameroon, and involved secondary school students. The study protocol was reviewed and approved by the institutional research ethics committee of the University of Maroua. Written informed consent was obtained from the parents or legal guardians of all participants, and verbal assent was obtained from the students themselves. Participation was voluntary, and participants could withdraw at any time, including by sending the keyword STOP by SMS, without any academic consequence.

Data protection measures included data minimization, pseudonymization of student identifiers, encryption of stored data, and role-based access control. Raw conversation data are retained for twelve months after study completion and will then be securely deleted, while aggregated and de-identified analytical data are retained for research verification and reproducibility purposes.

API calls to the commercial large language model (LLM) provider transmitted only pseudonymized query content and were conducted under the provider's applicable API data-usage policy, which excludes submitted API content from model training. Data were encrypted in transit using Transport Layer Security (TLS) and at rest.

An incident-response procedure was in place to address potential privacy and safety concerns. Keyword-based safeguards automatically routed harmful, abusive, or crisis-related student messages to a designated human reviewer rather than processing them solely through the automated pipeline. All response samples reviewed by teachers were anonymized, and all results are reported in aggregate form to prevent the identification of individual participants.

Data, code, and materials availability

The prompt templates and detailed algorithms are fully specified in this paper and can be reimplemented from it; additional technical details are available from the corresponding author upon reasonable request. The full Malloum backend source code is proprietary to the operating structure and is not released. Student conversation histories and interaction logs are confidential and cannot be shared, in accordance with the data-protection commitments described in the Ethics statement.

Author Contributions

Isaac Touza: Conceptualization, methodology, data curation, software development, formal analysis, investigation, visualization, writing—original draft, and writing—review and editing. The author reviewed and approved the final version of the manuscript and agrees to be accountable for all aspects of the work.

Acknowledgement

The author thanks the secondary school teachers and students of the Far North Region of Cameroon for their participation in the evaluation. Special thanks to Maroua Innovation Technology (MIT) for the support in the development of the Malloum platform.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Appendix 1. Configuration and Execution Environment

For reproducibility, Table 5 specifies the full software and execution environment used during the evaluated deployment (2024–2025 academic year) and for the statistical analysis.

Appendix 2. Full Prompt Templates

The three prompt layers described in Section 4 are instantiated as follows.

Appendix 2.1 Layer 1 — System Prompt.

```
You are Malloum, a patient and encouraging secondary school tutor for Cameroonian students preparing
for BEPC, Probatoire, Baccalaureat, or GCE examinations. Detect the language of the student’s message
(French or English) and respond in that language. Stay strictly within the Cameroonian national syllabus
for the subject and level indicated. Structure every answer as: (1) a short definition, (2) a worked
example, (3) a cultural analogy when relevant, (4) a brief follow-up question inviting the student to
continue. Keep the full answer under [token budget] tokens so it fits within [max_sms] SMS segments. If
the question is off-topic, unsafe, or indicates distress, do not answer it directly; redirect gently
and flag it for human review.
```

Table 5. Execution environment and configuration parameters.

Parameter	Configuration Value / Specification
Generative AI and retrieval stack	
Language model	`gpt-4o-2024-08-06` (OpenAI API), version pinned for the full 14-week study
Temperature	0.4
Top-p sampling	0.9
Max output tokens	512 tokens per response
Embedding model	`text-embedding-3-small`, 1,536 dimensions (OpenAI API)
Vector index	FAISS (`faiss-cpu` 1.8), exact flat search, rebuilt each school term
Backend orchestration	Python 3.11, queued service implementing Algorithms 2–3
SMS gateway	
Operator integration	MTN and Orange Cameroon APIs via a message broker layer
Delivery policy	Up to 3 retry attempts; duplicate detection via operator message IDs
Statistical analysis environment	
Software	R 4.3.2 (platform: x86_64-pc-linux-gnu)
Mixed-effects estimation	`lme4` 1.1-35
Inference	`lmerTest` 3.1-3 (Kenward–Roger); `clubSandwich` (cluster-robust SEs)
Sensitivity analysis	Randomization inference, 5,000 permutations within matched pairs
Hosting	
Infrastructure	Cloud-hosted backend and vector index; encrypted storage at rest; TLS-encrypted API calls
Student-side requirement	None beyond standard SMS (no internet access; see Section 3.1.1)

Appendix 2.2 Layer 2 — Contextual Curriculum Prompt (RAG).

```
Curriculum context (top-k retrieved passages, relevance-ranked):
[1] {source: Reference Course, subject, level, chunk text}
[2] {source: Exam Bank, examination, year, question, validated solution}
...
Use only the information above and your general pedagogical knowledge of the Cameroonian syllabus. If
the context does not cover the question, say so explicitly rather than inventing curriculum content.
```

Appendix 2.3 Layer 3 — User Query Prompt.

```
[Level: {form/class}] [Subject: {subject}]
[Known gaps: {gap_summary}]
[Conversation summary: {h_summary}]
Student question: {q}
```

Appendix 3. Full Mixed-Effects Regression Table

Appendix 4. Participant Flow Diagram

Appendix 5. Annotated Example Dialogues

Appendix 6. Lexical Substitution Table (Algorithm 2)

Table 6. Mixed-effects linear regression of composite post-test score (random intercept for school, REML estimation, `lme4` 1.1-35 in R, Kenward–Roger degrees of freedom).

Fixed effect	Estimate	SE	95% CI	p-value
Intercept	24.8	3.2	[18.5, 31.1]	< 0.001
Treatment (Malloum = 1)	16.1	1.1	[13.9, 18.3]	< 0.001
Pre-test score	0.71	0.04	[0.63, 0.79]	< 0.001
School-level baseline mean	0.18	0.07	[0.04, 0.32]	0.012
Random effects				
School intercept variance (τ_{00}^2)	9.4			
Residual variance (σ^2)	68.9			
Intraclass correlation (ICC)	0.12			
Clusters (schools) / Observations	20 / 1,000			

Table 7. Representative SMS exchanges with expert-rated quality scores (5-point scale).

Student query	Malloum response (as delivered)	Len. (segm.)	Rating
(FR, Math, Form 4) “Comment résoudre x2+3x-2=0 ?”	`a=1,b=3,c=-2. Delta=17. Deux solutions: x=(-3+racine17)/2 et x=(-3-racine17)/2. Remplace-les dans l'equation pour verifier.`	123 (1)	4.6/5
(EN, Physics, Form 5) “What is Ohm’s law?”	`V=IxR (volts=amps x ohms). Ex: 2Ax5ohm=10V. Like water in a narrow pipe: more resistance, less flow. Try one?`	109 (1)	4.4/5
(FR, Histoire, Form 3) “Quelles sont les causes de la Première Guerre mondiale ?”	`Causes: alliances, nationalisme, course aux armements. Declencheur: attentat de Sarajevo (1914).`	96 (1)	4.3/5
(EN, Computer Science, Form 4) “What is an algorithm?”	`Algorithm = step-by-step recipe to solve a problem. Ex: sort numbers smallest to largest.`	89 (1)	4.5/5
(ES, Espagnol, Form 4) “Como se presenta una persona?”	`Puedes decir: Me llamo Ana. Tengo 15 anos. Soy camerunes. Vivo en Maroua. Y tu, como te llamas?`	105 (1)	4.5/5
(EN, English, Form 3) “What is the difference between "has" and "have"?”	`Use has with he, she, it. Use have with I, you, we, they. Ex: She has a book. They have books.`	110 (1)	4.6/5
(FR, Geographie, Form 3) “Quelles sont les principales zones climatiques du Cameroun ?”	`Le Cameroun compte 4 grandes zones: equatoriale, tropicale humide, soudanienne et sahelienne.`	105 (1)	4.5/5
(FR, Biologie, Form 3) “Quel est le role de la chlorophylle ?”	`La chlorophylle capte la lumiere pour fabriquer la nourriture de la plante par photosynthese. Sans elle, la plante ne grandit pas.`	137 (1)	4.5/5

Appendix 7. Detailed Cost Calculation

Appendix 8. Expert Evaluation Rubric (Excerpt)

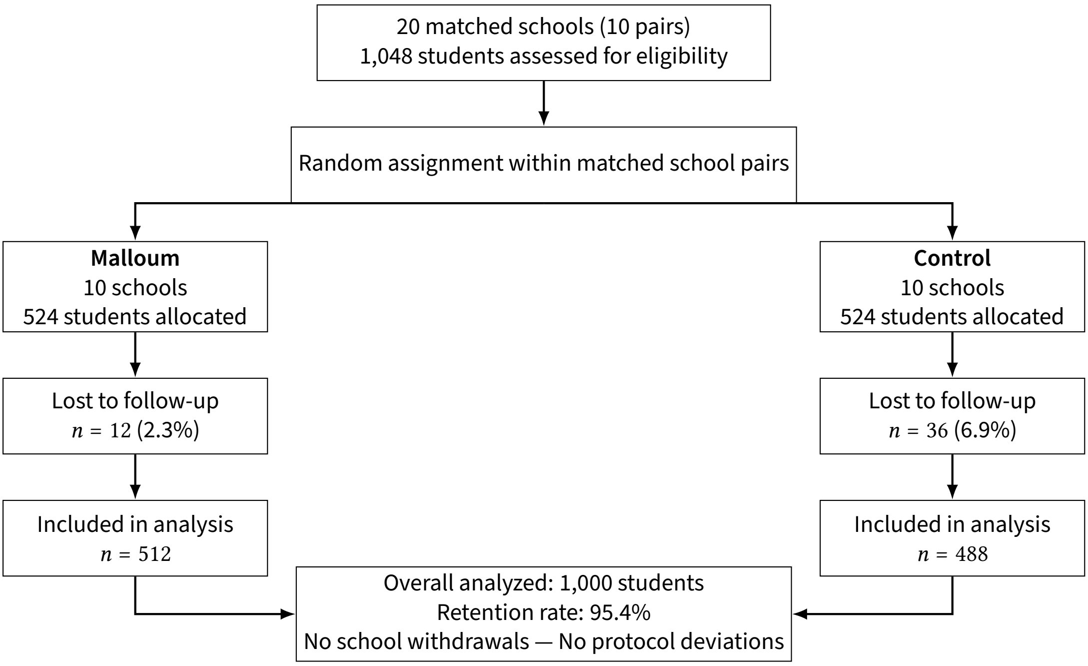


Figure 4. Participant flow diagram.

Table 8. Substitution rules applied during structural and lexical compression (Algorithm 2, French and English).

Original phrase	Compressed form
“In order to”	“To”
“It is important to note that”	“Note:”
“I would like to explain that”	(removed)
“Certainly! Here is”	(removed)
“Pour pouvoir”	“Pour”
“Il est important de noter que”	“NB :”
“Je voudrais t’expliquer que”	(removed)
“Bien sûr ! Voici”	(removed)
“such as for example”	“e.g.”
“par exemple, on peut citer”	“ex. :”

Table 9. Itemized per-student cost estimate per trimester.

Cost component	Borne by	Amount (XAF)
SMS tariff (36–56 sessions × 3.5–4 msg × 40–55 XAF)	Student/family	5,000–12,000
Sensitivity range (usage × tariff extremes)	Student/family	4,300–12,700
LLM tokens (~XAF 12/interaction)	Operating structure	~500
Short-code and operator fees	Operating structure	~250
Hosting and moderation	Operating structure	~150
Platform-side subtotal	**Operating structure**	**~900**

Table 10. 5-point rating scale used by the six DIPES II-qualified raters for the quality assessment in Section 7.2.3 (400 responses, weighted Cohen's $\kappa = 0.74$).

Criterion	Anchor description (1 = poor, 5 = excellent)
Factual correctness	1: contains a factual error affecting the answer; 5: fully accurate, no unsupported claim
Curriculum relevance	1: off-syllabus or wrong level; 5: precisely aligned with the national syllabus for the stated level
Completeness	1: answer is truncated or omits a necessary step; 5: self-contained, no missing step
Age appropriateness	1: tone or vocabulary unsuited to a secondary student; 5: fully age-appropriate
Cultural appropriateness	1: analogy is confusing, stereotyped, or irrelevant; 5: analogy is clear, respectful, and locally grounded
Harmful-content handling	1: fails to redirect an unsafe/off-topic query; 5: redirects appropriately and flags for review

References

[1] GSMA. *The Mobile Economy Sub-Saharan Africa 2024.* Tech. rep. Accessed: 2026-04-02. London, UK: GSMA Intelligence, Nov. 2024. URL: https://www.gsmaintelligence.com/research/the-mobile-economy-sub-saharan-africa-2024.

[2] Anthony M. Kelly and Patrick N. Rutazihana. "Digital Divide and its Implications for Education in the Post-COVID-19 Era in Sub-Saharan Africa". In: *Innovation and Education* 6.2 (2024), pp. 108–132. DOI: 10.1163/25248502-bja00007.

[3] International Telecommunication Union (ITU). *Measuring Digital Development: Facts and Figures 2023.* Tech. rep. Accessed: 2026-04-02. Geneva, Switzerland: International Telecommunication Union, 2023. URL: https://www.itu.int/itu-d/reports/statistics/facts-figures-2023/.

[4] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, and Gjergji Kasneci. "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103 (2023), p. 102274. DOI: 10.1016/j.lindif.2023.102274. URL: https://doi.org/10.1016/j.lindif.2023.102274.

[5] Zachary A. Pardos and Shreya Bhandari. "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills". In: *PLOS ONE* 19.5 (2024), e0304013. DOI: 10.1371/journal.pone.0304013.

[6] Greg Kestin, Kelly Miller, Allison Klales, et al. "AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting". In: *Scientific Reports* 15 (2025), p. 17458. DOI: 10.1038/s41598-025-97652-6.

[7] Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. "A Systematic Review of AI-Driven Intelligent Tutoring Systems (ITS) in K-12 Education". In: *npj Science of Learning* 10.1 (2025), p. 29. DOI: 10.1038/s41539-025-00320-7. URL: https://doi.org/10.1038/s41539-025-00320-7.

[8] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jia Yin, Jiaming Lin, and Yuze Wang. "EduChat: A Large-Scale Language Model-Based Chatbot System for Intelligent Education". In: 2023. DOI: 10.48550/arXiv.2308.02773. URL: https://doi.org/10.48550/arXiv.2308.02773.

[9] Van Dat Nguyen, Ngoc Hung Tran, Kim Anh Dao, and Pham Van Nguyen. "Towards Personalized AI Education: Context-Aware Retrieval-Augmented Generation With Grade-Level LLM Adaptation". In: *Computer Applications in Engineering Education* 34 (2026), e70153. DOI: 10.1002/cae.70153.

[10] Benjamin S. Bloom. “The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring”. In: *Educational Researcher* 13.6 (1984), pp. 4–16. DOI: 10.2307/1175554.

[11] Yunfan Gao et al. “Retrieval-Augmented Generation for Large Language Models: A Survey”. In: *CoRR* abs/2312.10997 (2023). DOI: 10.48550/arXiv.2312.10997. arXiv: 2312.10997 [cs.CL]. URL: https://doi.org/10.48550/arXiv.2312.10997.

[12] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[13] Jakub Swacha and Mateusz Gracel. “Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications”. In: *Applied Sciences* 15.8 (2025), p. 4234. DOI: 10.3390/app15084234.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 24824–24837. DOI: 10.52202/068431-1800. URL: https://doi.org/10.52202/068431-1800.

[15] M-Shule. *Personalized Learning Through SMS-Based AI in Kenya*. https://m-shule.com/. 2022.

[16] George Boateng, Sarah John, Alex Glago, Samuel Boateng, and Victor Kumbol. “Kwame for Science: An AI Teaching Assistant Based on Sentence-BERT for Science Education in West Africa”. In: arXiv preprint arXiv:2206.13703. 2022. DOI: 10.48550/arXiv.2206.13703.

[17] United Nations. *AI and the Future of Learning*. Africa Renewal. https://africarenewal.un.org/en/magazine/ai-and-future-learning. 2025.

[18] Khan Academy. *Khanmigo: AI-Powered Tutoring Assistant*. https://www.khanacademy.org/khan-labs. 2023.

[19] Nedim Slijepcevic and Ali Yaylali. “Leveraging “Khanmigo” Generative AI-Powered Tool for Personalized Tutoring to Learn Scientific Concepts”. In: *Journal of Teaching and Learning* 19.4 (2025). DOI: 10.22329/jtl.v19i4.10052.

[20] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. “Deep Knowledge Tracing”. In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015, pp. 505–513. URL: https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html.

[21] Jean Piaget. *Science of Education and the Psychology of the Child*. Translated by Derek Coltman. New York: Orion Press, 1970. ISBN: 9780671211967.

[22] Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Ed. by Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. Cambridge, MA: Harvard University Press, 1978. ISBN: 9780674576292.

[23] Wayne Holmes and Fengchun Miao, eds. *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023. ISBN: 978-92-3-100612-8. DOI: 10.54675/EWZM9535. URL: https://doi.org/10.54675/EWZM9535.

[24] UNESCO. *The Cultural Cost of AI in Africa’s Education Systems*. UNESCO Commentary. https://www.unesco.org. 2025.