# SmartEdQA: Transforming Learning with AI-Powered Question Answering for Students and Teachers in Underserved Areas

**delphine nyaboke**
Carnegie Mellon University Africa
Kigali, Rwanda
dnyaboke@andrew.cmu.edu

## Abstract

This project presents a question-answering (QA) system designed to empower students, and teachers in low-resource settings with access to critical educational information. The main objective is to facilitate informed decision-making by providing tailored support for career advising, revision, question-answering on specific topics, lesson planning, and quizzing. The system leverages open-book QA and machine reading comprehension (MRC) frameworks, enabling users to extract precise answers from provided documents. It supports two QA task domains - short answer and multiple choice. The system's performance is evaluated using Precision, Recall, F1 Score metrics, ensuring reliable and accurate responses. By bridging information gaps, this solution aims to enhance educational outcomes in resource-constrained environments. Here's the GitHub Link.

## 1 Introduction

### 1.1 Motivation

Nairobi has a population of 5 766 990, out of which 250 000+ of those live in Kibera Slums (World Population Review, 2024) and the teacher student ratio in public schools is 100:1 (The Borgen Project, 2024). No one should be punished because their parents / guardians are poor, being given an opportunity to get out, a choice of leveraging AI in this fourth industrial revolution would mean everything to that person.

### 1.2 Research Question

This study focused on addressing the following core questions:

a) Can a state-of-the-art large language model (LLM) be effectively tuned to handle the types of questions commonly posed by teachers and students?

b) Can state-of-the-art LLMs be integrated into a chatbot architecture that is both accessible and usable by teachers and students?

### 1.3 Potential Impact

The use of this AI-powered question answering would enable teachers and students to leverage AI in this fourth industrial revolution to enhance quality of education:

- the teachers are able to conduct quizzes and grade them
- the students are able to study on their own and improve their grades

These speaks of Sustainable Development Goal 4 Quality Education and Goal 10 Reduced Inequalities that, everyone should be able to be an activate participant in this revolution, and that these students are the AI Natives.

## 2 Related Work / Literature Review

Large Language Models(LLMs) can enhance personalized learning, improving teaching quality, and introducing new models of instruction. The benefits include personalised learning support where there's real time problem solving and adaptive feedback, teacher assistance and efficiency e.g through grading and content creation; assessment and feedback where the teacher's workload are reduced like use of ChatGPT, Khanmigo, MathGPT. (Xu et al., 2024).

Natural Language Processing can be employed in text embeddings and clusters, and can be seen in

large scale education surveys and those open ended student responses. (Katz et al., 2023).

LLMs can also be integrated into project based learning to enhance creativity, problem-solving, and collaboration among middle school students. It investigates both the opportunities and challenges of using LLMs in child-centered education and provides design recommendations for effectively implementing AI in classrooms. (Zha et al., 2024)

## 3 Methodology

### 3.1 Task: Input & Output

The project is designed to serve low-resourced communities in both rural areas and urban poor neighborhoods such as the Kibera Slums in Nairobi, Kenya. The primary beneficiaries are:

I. *Students*: Offered an interactive, engaging space to practice and reinforce learning.

II. *Teachers* Provided with tools to assess student understanding and enhance lesson planning.

### 3.2 Hypothesis

The hypothesis is that:

- ***Student Engagement & Learning*** - a state-of-the-art large language model (LLM) can be effectively tuned to create an interactive environment where students can apply what they have learned, with potential gamification to make learning both fun and effective. This will encourage summarizing, quizzing, and explaining content to reinforce reading comprehension and retention.

- ***Teacher Support & Assessment*** - a state-of-the-art LLM can be integrated into a chatbot architecture that is both accessible and usable by teachers, enabling them to design targeted quizzes, receive sample responses, and tailor lesson plans accordingly. The pilot will cover one key subject.

## 4 Proposed method

This project tackles:

I. *Open-book QA* - the task of question answering given one or more open book documents, and

II. *Machine Reading Comprehension(MRC)* - a model is given a question and a document and is required to extract the answer from the given document.

The following are the QA Task Domains covered:

- ***Short Answer*** - quick response questions to test immediate understanding.

- ***Multiple Choice*** - structured questions for recognition and recall.

### 4.1 Project Details

There are two types / Nature of Questions:

- **LLM-Generated Questions:**
  - Questions generated automatically by a Large Language Model (LLM) based on a given subtopic.

- **Human-Generated Questions:**
  - Questions written by users in response to:
    - A specific subtopic or concept.
    - A paragraph of content.
  - Types of user-generated questions include:
    i. Direct and straightforward factual questions.
    ii. Questions that involve an explanation or elaboration within the query.
    iii. Questions seeking general knowledge or conceptual understanding.

### 4.2 Project Architecture

The project had an Injection Pipeline, a Retrieval Augmented Generation (RAG) system with an LLM, a Rasa framework, and a combined RAG + LLM and Rasa.

#### 4.2.1 Injection Pipeline

Documents are processed by loading various formats (e.g., PDFs, JSON), splitting them into manageable text chunks, embedding them using a vector model, and storing the results in a vector database like Chroma.


Figure 1: Injection Pipeline: Loading, chunking, embedding, and storing document data.

#### 4.2.2 RAG + LLM

A LangChain RAG pipeline was used where a user query from a Jupyter Notebook triggers retrieval of relevant document chunks from a Chroma vector store. These are used to construct a prompt sent to an OpenAI LLM, which returns a contextualized answer.

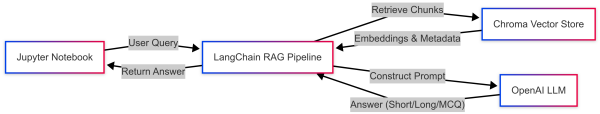


Figure 2: LangChain RAG Pipeline with OpenAI LLM and Chroma Vector Store.

#### 4.2.3 Rasa Only

Rasa is a conversational AI framework that handles dialogue management and executes custom actions based on user input. In this setup, Rasa receives a user message, triggers a custom action that invokes a LangChain-based RAG pipeline, and returns the generated response to the user via the chatbot interface.


Figure 3: Rasa integration with LangChain RAG pipeline and OpenAI LLM.

#### 4.2.4 Combined: RAG + LLM, and Rasa

This combined architecture showcases a unified system where both developers and end users interact with a LangChain RAG pipeline. Developers issue queries via a Jupyter Notebook interface, while end users communicate through a Rasa chatbot, both leveraging the same backend powered by Chroma for retrieval and OpenAI LLM for generating responses.

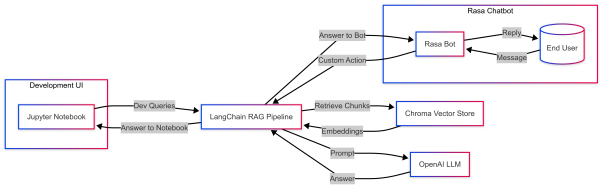


Figure 4: Combined architecture of RAG, LLM and Rasa

## 5 Experiment

### 5.1 Dataset

The concentration was on Senior 3 Biology (Rwanda's Education System), which is the same as Form 3 Biology (Kenya's Education System). The data was from:

- Rwanda Education Board
- Senior 3 Biology textbook

#### 5.1.1 Metrics and Significance Testing

Evaluation of the quality of LLM responses against golden references using the following metrics:

- **Precision:** Indicates how many tokens in the candidate (LLM response) have a high similarity to tokens in the reference (golden response). A high precision means that most of the words or phrases generated by the model are relevant to the reference.
- **Recall:** Measures how many tokens in the reference are captured by the candidate. High recall suggests that the candidate covers most of the important content or details in the reference.
- **F1 Score:** The harmonic mean of precision and recall. It provides a single overall measure of similarity that balances both metrics. A high F1 score indicates that the candidate is both comprehensive and accurate compared to the reference.
- **BERT Score:** Computes a similarity score using contextualized embeddings from BERT, capturing semantic similarity more effectively than simple token overlap. A high BERT score implies that the candidate responses closely match the semantic content of the golden responses.

## 6 Result & Analysis

### 6.0.1 Multiple Choice Questions

These are the results for Multiple Choice Questions (MCQs) that were generated by one LLM and answered by a different LLM i.e., DeepSeek LLM and ChatGPT, respectively incorporating BAAI/bge-base-en-v1.5 as an embedding model:

| Model | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| GPT-3.5-Turbo | 0.9998 | 0.9998 | 0.9998 | 89.74% |
| GPT-4 | 0.9999 | 0.9999 | 0.9999 | 94.87% |
| GPT-4-Turbo | 0.9979 | 0.9979 | 0.9979 | 97.95% |

Table 1: BERTScore and accuracy for MCQs.

### 6.0.2 Short Answer Questions

These are the results for Short Answer Questions that were generated by one LLM and answered by a different LLM i.e., DeepSeek LLM and ChatGPT, respectively incorporating BAAI/bge-base-en-v1.5 as an embedding model:

| LLM Model | Precision | Recall | F1 Score |
|---|---|---|---|
| GPT-3.5-Turbo | 0.9332 | 0.9508 | 0.9416 |
| GPT-4 | 0.8751 | 0.8715 | 0.8730 |
| GPT-4-Turbo | 0.8769 | 0.8863 | 0.8812 |

Table 2: BERTScore for short answer questions.

### 6.0.3 Human Generated Questions

These questions were generated by actual humans — a few students and teachers — and answered by an LLM (GPT-3.5-Turbo) using BAAI/bge-base-en-v1.5 as the embedding model.

| User Type | Precision | Recall | F1 Score |
|---|---|---|---|
| Straightforward Approach | 0.9422 | 0.9551 | 0.9483 |
| Explains Approach | 0.9391 | 0.9529 | 0.9457 |
| General Knowledge | 0.9259 | 0.9527 | 0.9388 |

Table 3: BERTScore for human-generated short answer questions.

### 6.0.4 Chatbot Responses

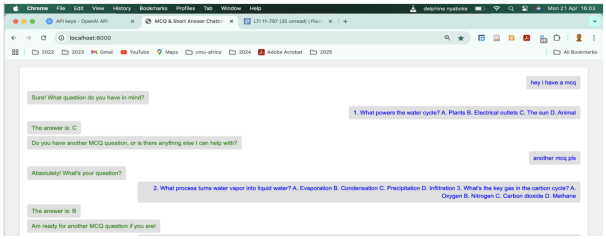


Figure 5: Testing the chatbot with different types of questions.

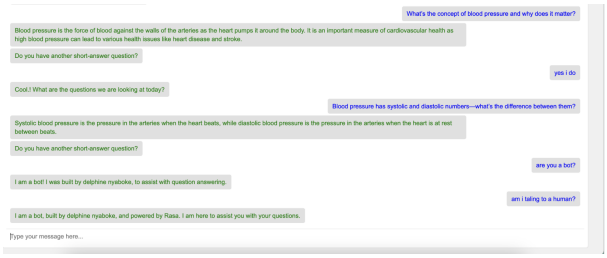


Figure 6: Example responses generated by the chatbot.

## 7 Conclusion

This project required 4 servers to be running concurrently, each with it's own virtual environment for the dependencies and libraries conflict mitigation. Postman was used to test the frontend and backend independently as apps before the complete integration.

### 7.1 Future Directions

To further enhance the educational impact of the system, the following future directions are proposed:

- **Subject Expansion:** Extend support to cover the full Biology curriculum from Junior 1 to Senior 6 (i.e., Form 1 through Form 4) for both Rwanda and Kenya respectively.
- **STEM Integration:** Incorporate additional STEM subjects such as Physics and Mathematics at the secondary school level.
- **Primary School Support:** Add modules for Mathematics and Science tailored for primary school learners.
- **Bilingual Accessibility:** Provide bilingual support in Kinyarwanda for Rwanda and Swahili

for Kenya to improve inclusivity and understanding.

- **Mobile-Responsive Interface:** Develop a responsive front-end interface that renders seamlessly on both Android and iOS devices.
- **Scalable Deployment:** Deploy the platform on cloud infrastructure (Azure, AWS, or GCP) to support large-scale testing with students and teachers.
- **Parental Engagement:** Integrate features that allow parents to monitor their children's academic progress and engagement.
- **Community Support Partnerships:** Collaborate with NGOs and CBOs to equip at least one school with internet access to interact with the chatbot.
- **Gamification and Feedback:** Gamify the application to drive engagement among students, teachers, and parents, while collecting feedback to improve the system continuously.
- **RAG Error Evaluation:** Monitor and record incorrect failure cases, noting their frequency, associated risks, and error types.
- **User Study:** Conduct human validation or certification of RAG-generated responses to ensure accuracy and reliability.

# 8 Limitations

While the proposed system demonstrates promising results,(and works on localhost) a few limitations remain:

- **Limited Curriculum Coverage:** The current implementation uses content from the Rwanda Education Board; textbooks approved by the Ministry of Education in Kenya are yet to be incorporated.
- **Model Training Constraints:** Training custom LLMs tailored to local educational content would require significant time, expertise, and computational resources.
- **Deployment Cost:** The financial cost of deploying and maintaining the system at scale—especially across different schools and regions—needs to be carefully considered.

## 8.1 Deployment Challenges

The following steps outline how to deploy the full system with RAG, LLM, Rasa, and a chatbot UI as separate concurrent services:

- **Containerize** each component (RAG pipeline, LLM API, Rasa server, and chat frontend) using Docker, exposing necessary ports for inter-service communication.
- **Launch all services concurrently** using Docker Compose or Kubernetes, ensuring containers are networked and can communicate via HTTP or gRPC.
- **Configure an API gateway or reverse proxy** (e.g., NGINX or Traefik) to route incoming traffic to the appropriate backend service (e.g., `/chat` to Rasa, `/query` to RAG).
- **Persist embedding storage** using the existing Chroma vector database to retain indexed document chunks across user sessions and service restarts.

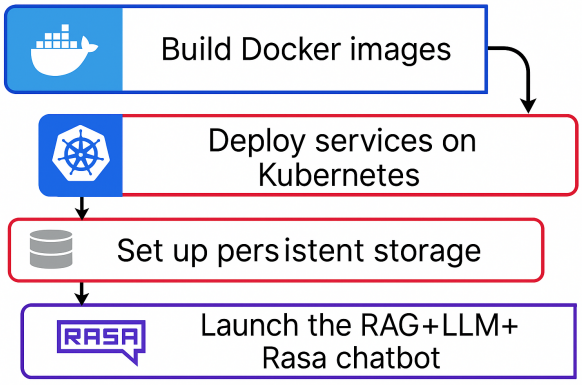


Figure 7: Deployment Steps

### 8.1.1 Sample Deployment Cost

The estimated monthly cost of deploying the system on a cloud platform like AWS is outlined below. These estimates assume moderate usage by several

schools, using t3-series EC2 instances, standard storage, and OpenAI GPT-3.5 API calls. Prices are based on publicly available AWS and OpenAI documentation as of April 2025.

| Service Component | Estimated Monthly Cost (USD) |
|---|---|
| EC2 Instance for RAG (t3.medium, 720 hrs) | $33.26 |
| EC2 Instance for Rasa (t3.medium, 720 hrs) | $33.26 |
| EC2 Instance for LLM proxy/API bridge | $33.26 |
| EC2 Instance for UI (t3.small, 720 hrs) | $17.81 |
| Amazon S3 for backups (10 GB) | $0.23 |
| EBS Storage (30 GB SSD) | $3.00 |
| OpenAI GPT-3.5 API usage (10k calls) | $15.00 |
| Data Transfer (30 GB out) | $2.70 |
| Load Balancer and Route 53 DNS | $15.00 |
| **Total Estimated Cost** | **$153.52/month** |

Table 4: Estimated monthly deployment cost on AWS with OpenAI integration. Based on pricing from AWS and OpenAI.

## Acknowledgments

I appreciate Prof Eric Nyberg for his guidance and mentorship, and Kimihiro Hasegawa for his incredible support throughout the course.

I also thank the efforts of few teachers and students that took their time, to provide insights during this project. This was very valuable, thank you.!

## References

A. Katz, U. Shakir, and B. Chambers. 2023. The utility of large language models and generative ai for education research. Accessed: 21 April 2025.

The Borgen Project. 2024. Education in kibera. Accessed: 21 April 2025.

World Population Review. 2024. Nairobi population 2024. Accessed: 21 April 2025.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024. Large language models for education: A survey. Accessed: 21 April 2025.

S. Zha, Y. Qiao, Q. Hu, Z. Li, J. Gong, and Y. Xu. 2024. Designing child-centric ai learning environments: Insights from llm-enhanced creative project-based learning. Accessed: 21 April 2025.