

EduX-RAG: Retrieval Augmented Generation Framework for Cross-Lingual Educational Chatbots

Johanne Medina

*Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
jomedina@hbku.edu.qa*

Soon Gyo-Jung

*Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
sjung@hbku.edu.qa*

Joni Salminen

*School of Marketing and Communication
University of Vaasa
Vaasa, Finland
jonisalm@uwasa.fi*

Kholoud Khalil Aldous

*Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
kkaldous@hbku.edu.qa*

Bernard J. Jansen

*Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
jjansen@acm.org*

Abstract—Developing cross-lingual educational chatbots presents significant challenges when handling dynamic language switching and content retrieval across different languages. We introduce EduX-RAG, a framework that integrates Retrieval-Augmented Generation (RAG) with prompt engineering to enable cross-lingual conversations in educational chatbots. EduX-RAG addresses the complexities of managing multilingual content and cross-lingual queries within a single interaction. We validate the effectiveness of EduX-RAG using 13 languages, including high- and low-resource languages and commonly confused language pairs. Deployed in a fully functional educational chatbot, EduX-RAG demonstrates strong cross-lingual retrieval capabilities with a high answer rate of 92%, even when query and material languages differ. Overall, EduX-RAG delivers robust cross-lingual capabilities, achieving a Language Identification Accuracy of 90%, Material Citation Accuracy of 99%, and a Response Similarity Score of 70%.

Index Terms—LLM, chatbot, RAG, prompt-engineering, education

I. INTRODUCTION

Large Language Models (LLMs) have significantly enhanced educational chatbots, enabling interactive and personalized learning experiences [1]–[3]. However, ensuring their reliability remains a pressing challenge [4] as hallucinations, inconsistent tone, and formatting issues can introduce misinformation and undermine trust in educational contexts [5]–[7].

One major yet underexplored requirement in this domain is cross-lingual robustness. This is defined as the ability of chatbots to support multilingual learners who may switch languages mid-conversation or pose questions in one language (e.g., their native language) while learning in a different medium. This is especially relevant in multilingual classrooms and international academic institutions where linguistic diversity is the norm [8]–[10]. Examples include Arabic-speaking students in international, English-medium classrooms or Estonian-speaking students in Finnish schools.

To address this gap, we develop EduX-RAG, a framework that integrates Retrieval-Augmented Generation (RAG)

with dynamic prompt engineering to enable reliable, context-aware, and linguistically adaptive educational chatbots. RAG enhances LLMs by incorporating external knowledge sources, allowing the model to retrieve relevant and up-to-date information for more accurate, context-specific responses [11]–[13]. What distinguishes EduX-RAG is its end-to-end design for multilingual education: it handles material ingestion across multiple languages, identifies the query language even in code-switched input, retrieves context from multilingual sources, and adapts prompts in real time, without relying on pre-aligned translations or language-specific retraining.

We demonstrate the real-world applicability of EduX-RAG through its deployment in CIPHERBOT, an AI educational assistant used in schools and universities across Qatar, Finland, China, Vietnam, Korea, and the United States [7], [14]. While CIPHERBOT serves as a case study, EduX-RAG is broadly applicable to any educational assistant operating in multilingual contexts.

To validate EduX-RAG, we conduct a targeted evaluation using educational materials in 13 different languages. Our assessment spans both high-resource and low-resource languages, including closely related language pairs such as Finnish–Estonian, Tagalog–Spanish, Arabic–Persian, and Korean–Japanese, among others. We evaluate the system’s ability to identify language accurately, retrieve the correct content cross-lingually, respond appropriately in dynamic multilingual exchanges, and manage subtle distinctions between similar languages.

Our research addresses two primary research questions (RQs):

RQ1: *How can a RAG-based framework effectively manage the challenges of multilingual content retrieval in educational settings across high- and low-resource languages?*

RQ2: *How can prompt engineering techniques be optimized to improve language-switching accuracy, minimize language mismatches and ensure contextually appropriate multilingual*

responses in educational chatbots?

Our major contributions are summarized as follows:

- We introduce EduX-RAG, a novel framework that unifies multilingual material ingestion, cross-lingual context retrieval, and dynamic prompt engineering for educational chatbots.
- We deploy the system in real-world multilingual classrooms and demonstrate its practical utility in diverse linguistic settings.
- We design an evaluation methodology based on functional capabilities including language identification, cross-lingual retrieval, dynamic language switching, and disambiguation of similar languages rather than relying on existing baselines.
- We conduct a comprehensive evaluation using materials in 13 languages, including both high- and low-resource languages, and report performance across a wide range of multilingual scenarios.

This work contributes a reusable, modular framework and a practical evaluation methodology for cross-lingual educational AI systems. By focusing on multilingual content alignment and dynamic language use, EduX-RAG provides a foundation for building more inclusive, equitable, scalable, and reliable learning tools.

II. RELATED WORK

The integration of LLMs into educational assistants and chatbots has opened up new possibilities for interactive learning, intelligent tutoring, and scalable student support [15]–[18]. The release of ChatGPT in late 2022 accelerated their adoption across educational contexts, bringing renewed attention to the potential and limitations of conversational AI in classrooms [2], [19], [20]. Despite growing enthusiasm, existing work highlights persistent concerns regarding reliability, learner engagement, and pedagogical alignment.

Several studies have emphasized the need for learner-centric and pedagogically grounded chatbot designs. Bahja et al. [21] and Ilieva et al. [22] underline the importance of tailoring chatbot systems to users’ educational needs, technological skills, and learning contexts. However, as noted in recent reviews [23], [24], most chatbot deployments remain exploratory, with limited research on how to structure learning interactions or measure long-term impact.

Reliability and evaluation are particularly pressing in high-stakes educational environments. LLM-based systems are prone to hallucinations, factual inconsistencies, and formatting issues that reduce their trustworthiness [5], [25]. To address these challenges, new evaluation frameworks have emerged. Factored evaluation mechanisms and benchmarks provide metrics for assessing content fidelity and response relevance [5], [26], [27]. However, few of these efforts account for language variation or multilingual contexts.

Multilingual capability remains a critical and underexplored aspect of educational chatbots [28]. Most current systems are either monolingual, offer limited bilingual support, or are typically focused on language learning or translation tasks

[29]–[31]. Even LLM-powered teaching assistants have been shown to struggle in multilingual classrooms [32], [33]. These systems cannot often handle dynamic language switching or retrieve class-specific content across multiple languages [34], [35]. Some progress has been made in other domains (e.g. healthcare) where cross-lingual retrieval and response generation have received more attention [8], [36], but education has lagged behind [37].

Research has begun exploring direct use of ChatGPT [38] and retrieval-augmented techniques for multilingual generation [39], [40], but such systems often rely on pre-translated corpora and rarely address the need for dynamic language adaptation during user interaction.

Our work addresses these gaps by introducing EduX-RAG, a cross-lingual framework that integrates multilingual material ingestion, dynamic query language identification, cross-lingual retrieval, and prompt-level response control. Unlike prior approaches, EduX-RAG is designed specifically for educational settings and enables seamless language transitions within a single interaction, without relying on pre-aligned translations or language-specific retraining. This positions EduX-RAG as a practical and extensible solution for multilingual learning environments.

III. EDUX-RAG FRAMEWORK

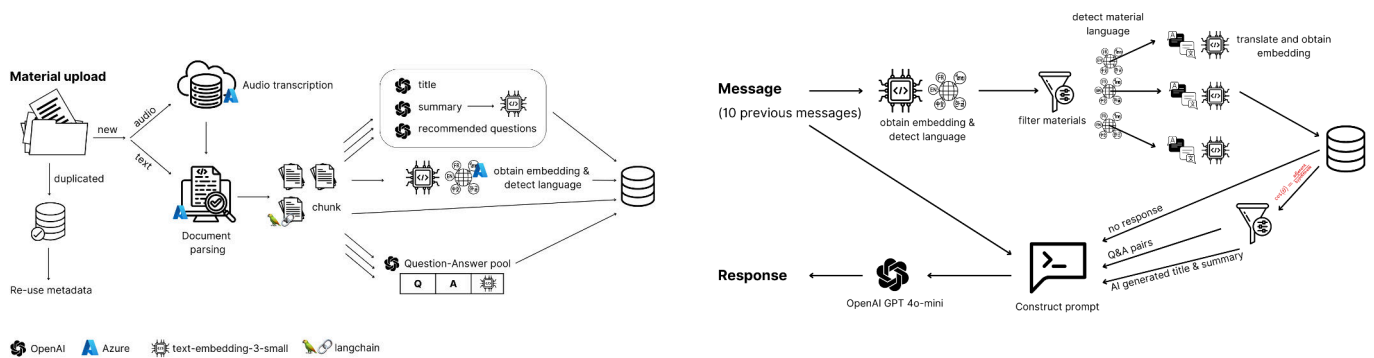
The EduX-RAG framework addresses the complexities of cross-lingual educational environments through three key components: (1) Materials Processing Pipeline, (2) Query Handling Pipeline, and (3) Dynamic Prompt Engineering. These components work together to process multilingual educational content efficiently, delivering accurate, cross-lingual responses in real-time.

A. Materials Processing Pipeline

As shown in Figure 1a, when new instructional material is added to the RAG, a content fingerprint (e.g., hash) is computed to detect duplicates and avoid redundant processing. This deduplication step is particularly valuable for large-scale deployments, as parsing and chunking educational content can be computationally expensive. If the material has already been processed, previously generated metadata is reused to accelerate subsequent query resolution.

The pipeline first detects the material’s format and routes it through an appropriate preprocessing module. For audio inputs, an automatic speech recognition (ASR) component transcribes the audio into text using a batched inference mechanism. For text-based formats such as PDFs, Word documents, presentations, or markdown files, a document parsing module extracts structured content, including text blocks, tables, and embedded metadata. Optical Character Recognition (OCR) is applied as needed to handle scanned or image-based documents¹

¹In our deployment, we use Microsoft Azure Document Intelligence and Azure Speech Services for parsing and transcription. However, our framework is modular and supports interchangeable components for multiple services.



(a) Materials processing pipeline. Uploaded materials are parsed, transcribed if necessary, segmented into overlapping chunks, and embedded for retrieval. Processed outputs, including structured metadata and vector representations, are stored in a vector-supporting database.

(b) Query handling pipeline. Incoming queries are embedded and language-tagged, filtered by class context, aligned to the language of indexed materials, and matched with relevant chunks using semantic similarity. Retrieved content is used to construct a context-rich prompt for grounded response generation.

Fig. 1: Overview of the EduX-RAG framework. (a) The materials pipeline automates the ingestion and processing of educational content in various modalities, generating structured metadata and retrieval-ready embeddings. (b) The query pipeline dynamically processes user messages, identifies relevant multilingual content based on semantic similarity, and constructs contextual prompts for the language model.

Once parsed, the material is segmented into overlapping chunks to facilitate context-aware retrieval. We use a sliding-window approach that retains a portion of the preceding content to preserve local coherence and continuity. This overlap is tuned to balance semantic coverage with computational efficiency, as overly small overlaps can fragment the material and weaken context while larger overlaps may introduce redundancy and processing overhead [41]. After chunking, each segment is processed in parallel to extract auxiliary metadata for downstream indexing and retrieval tasks.

Key tasks in this pipeline include:

- **AI-generated Metadata.** Using OpenAI’s GPT-4o mini, titles, summaries, and recommended questions are generated for each material. These additional details supplement the content and allow for faster answer retrieval.
- **Embedding and Language Detection.** Embeddings for each chunk are obtained, and the material’s language is detected, both stored as metadata. Embeddings for the AI-generated summaries are also created, which are crucial for the materials filtering step discussed in Section III-B
- **Question-Answer (QA) Pool Generation.** A pool of possible questions and corresponding answers is generated for each chunk using the OpenAI GPT-4o mini model. Embeddings of the questions are saved for use in the cosine similarity calculations outlined in Section III-B. This QA pool serves as the primary knowledge base from which the system retrieves information.

B. Query Handling Pipeline

Upon receiving a user query, the system computes its embedding to facilitate semantic retrieval. We use a multilingual embedding model selected for its balance of efficiency

and cross-lingual performance.² Concurrently, the query’s language is automatically identified using a language detection module.³ These components are modular and can be substituted based on deployment constraints or model availability.

To ensure contextual relevance, the retrieval process begins by filtering the corpus to include only materials associated with the user’s class or domain. This scoped retrieval ensures that the chatbot only considers materials uploaded by the relevant instructor, aligning with the principles of RAG. Since both the user query and the indexed materials may exist in different languages, the system aligns them through a lightweight translation step. Although the embedding model supports multilingual input, empirical evidence suggests that aligning both inputs in the same language before embedding often improves retrieval accuracy [42].

Semantic similarity is computed between the query embedding and precomputed embeddings of question-answer (QA) pairs drawn from the processed educational materials (see Section III-A). Cosine similarity is used as the distance metric, with a threshold of 0.8 applied to select only highly relevant matches. This threshold is calibrated to strike a balance between precision and recall, as setting it too low risks retrieving irrelevant material, while overly high values may exclude semantically useful content that is not an exact match.

Once the most relevant QA pairs are identified, their content is used to construct a context-aware prompt for generation. This prompt includes the user query, the immediate conversation history (typically the previous ten turns), and summaries or titles associated with the retrieved QA pairs. This aggrega-

²In our deployment, we use OpenAI’s *text-embedding-3-small* model.

³Implemented using Azure’s Language Detection API.

gated context is passed to the language model to produce a grounded and coherent response tailored to the user’s question.

C. Dynamic Prompt Engineering

Dynamic prompt engineering involves creating adaptable prompts that adjust based on user input and context, enabling the framework to generate accurate question-answer pairs and to handle multilingual queries effectively. This section details the structure and application of our dynamic prompts, which are designed to address specific needs in educational contexts.

1) *Creating Question-Answer Pairs*: This prompt is designed to understand the material, formulate relevant questions, and identify concise, informative answers that match the intent of the educational content. The full prompt is shown in Figure 2a. Components of the prompt include:

- **Content Introduction.** The prompt begins by presenting the educational material within a defined text block, identified as content.
- **Language Specification.** It specifies the language of the content, ensuring both questions and answers are generated in the same language.
- **Focused and Exhaustive Question Generation.** The prompt generates only content-answerable questions, continuing until all possible questions are exhausted.
- **Output Format.** Responses are structured in JSON, with each Q&A pair as an object, ensuring seamless chatbot integration.
- **Temperature and Sampling Settings.** The model is set with a temperature of 0 and a top_p value of 0.1.

2) *Processing and Answering Queries*: The query processing prompt is designed to guide the chatbot in answering questions with contextually relevant and language-specific responses. Instructing the chatbot to act as a teaching assistant helps ensure that responses are comprehensive, context-aware, and tailored to the student’s level and the language of the query. The system retrieves relevant materials and utilizes both the current query and historical context to formulate accurate and helpful answers. Components of the prompt include:

- **Teaching Assistant Role Definition.** The prompt starts by setting the chatbot’s role as CIPHERBOT, a teaching assistant for a specific class and educational level.
- **Class Instruction Contextualization.** The chatbot references class-specific details, dynamically generated by a modular function to ensure course-aligned responses.
- **Student Personalization.** The chatbot addresses students by name, enhancing engagement and a sense of personalization through an instruction function.
- **School Level Adaptation.** A function adjusts detail, vocabulary, and response length to match the student’s educational level.
- **Language Specification.** A function ensures responses match the query language, maintaining linguistic consistency in multilingual interactions.
- **Follow-Up and Formatting Guidance.** The prompt enforces JSON formatting and allows follow-up questions when relevant.

IV. FRAMEWORK VALIDATION

A. Data Gathering

To evaluate the cross-lingual capabilities of EduX-RAG, we selected 13 academic journals, each written in a different language and covering distinct, non-overlapping topics. With the selection of languages, we sought to include both widely spoken languages, such as English, Chinese, and Arabic, and languages commonly confused due to linguistic similarities, as identified in [43] (see Figure 3).

B. Evaluation Metrics

To assess the effectiveness of the EduX-RAG framework, we define a set of evaluation metrics that reflect its ability to manage complex multilingual interactions in educational settings. These metrics are designed to evaluate core capabilities such as accurate language detection, robust content retrieval across languages, and coherent handling of language switching and mixed-language inputs. Together, they provide a structured foundation for evaluating system performance in realistic, multilingual classroom environments.

1) *Language Identification*: Accurate language identification is a foundational requirement for any multilingual educational chatbot. It enables appropriate content retrieval and response generation by ensuring that the system correctly interprets both the language of instructional materials and student queries.

We evaluate three widely used language detection tools: Python’s `langdetect` library [44], Microsoft Azure Language Detection API [45], and OpenAI’s GPT-4o mini model [46]. Each tool is tested on a multilingual dataset (see Table 3) consisting of questions with known ground-truth language labels. Performance is evaluated using the following metrics:

- **Language Identification Accuracy (LIA)**: The proportion of queries where the predicted language matches the ground truth. Defined as:

$$LIA = \frac{\text{Number of language match}}{\text{Total number of queries}}$$

- **Precision and Recall (per language)**: Precision measures the proportion of correctly identified instances among all predictions for a given language, while recall measures the proportion of correctly identified instances among all actual instances of that language.

2) *Query Resolution Evaluation (Answered and Unanswered)*: We evaluate EduX-RAG’s ability to resolve user queries across both language-matched and cross-lingual scenarios by measuring how effectively it distinguishes between answerable and unanswerable questions. This assessment captures not only the system’s response generation capacity but also its self-awareness in recognizing gaps in knowledge.

- **Answered Question Rate (AQR)**: The proportion of user queries that receive a substantive, informative response. Formally:

$$AQR = \frac{\text{Number of answered queries}}{\text{Total number of queries}}$$

```

(start of TEXT )
(content)
(end of TEXT )
Content Introduction
TASK:
Generate a question answering pool in {language} for comprehensive understanding on the given
TEXT.
Language Specification
Generate only the questions which are answerable only based on the given TEXT.
Do not generate about the references or citations.
Generated questions should not refer the given TEXT as 'in the text', instead include
keywords.
Focused and Exhaustive
Question Generation
Use the same (or majority) language for question and answer as the TEXT.
Respond in the following JSON format:
- questions (array of question object (question, answer))
-- question (string) : the question
-- answer (string) : the answer {questions}
Your response in JSON:
""" questions_instruction = f""" You have generated already the below questions, so do not
generate the same questions:
== {questions} ==
Output Format
Do not generate questions if you cannot generate anymore, just return an empty list
"""

```

(a) Prompt used for generating synthetic question–answer pairs from parsed educational materials. This step enables the construction of a searchable QA pool to support retrieval during user interactions.

```

You are a teaching assistant, CIPHERBOT, your role is to assist (school_level_label) students of
the class, {class_name}, in understanding "
"by providing answers only derived from the retrieved materials and the given context, including
your previous responses and the student's messages.\n"
Teaching Assistant Role Definition
"{class_instruction}"
Class Instruction Contextualization
"{student_name_instruction}"
Student Personalization
"{school_level_instruction}"
School Level Adaptation
"when student requests for further information or more detailed explanations, you must look into
previous context for a corresponding answer.\n"
"language_response_instruction":
"language_response_instruction":
"If the question cannot be answered from the retrieved materials or the given context, "
"be sweet and apologetic and guide them towards asking more appropriate questions.\n"
)
- Your final output should be in this JSON format:\n'
{"response": "<your response with citations>", "follow_up_questions": [{"question1}>", "
question2>", "<question3>"}]\n'
- Make your response in markdown formatting in the right places, including line changes, bullet
points, numbered points, and bold text to improve clarity and readability (without title).\n"
- Generate maximum 3 follow-up questions without bullet/numbered points which are contextually
appropriate and answerable based on the retrieved materials while ensuring that the their
language matches response's language.\n"
Follow-up and Formatting Guidance
)

```

(b) Prompt structure used for responding to user queries. It integrates the query, prior conversation history, and retrieved question–answer pairs to ensure accurate, context-aware, and grounded responses.

Fig. 2: Prompt engineering in EduX-RAG. (a) illustrates the generation of question–answer pairs during material preprocessing to support downstream retrieval, (b) shows the construction of response prompts that combine the user query with relevant context retrieved from the indexed materials.

- **Unanswered Question Rate (UQR):** The proportion of queries for which the system explicitly indicates its inability to provide an answer (e.g., through fallback responses such as “I’m not sure how to answer that”). This reflects the framework’s capacity for calibrated uncertainty:

$$\text{UQR} = \frac{\text{Number of unanswered queries}}{\text{Total number of queries}}$$

- **Classification Accuracy (CA):** A weighted metric that combines the accuracy of both answered and unanswered predictions, taking into account the relative proportion of each:

$$\text{CA} = (\text{AQR} \times \text{Acc}_{\text{answered}}) + (\text{UQR} \times \text{Acc}_{\text{unanswered}})$$

3) *Multilingual Content Retrieval Accuracy:* Given that EduX-RAG is designed to operate across diverse languages, the accuracy of content retrieval is critical to its overall performance. We evaluate the framework’s ability to retrieve relevant educational materials and align them with user queries, regardless of language.

- **Material Citation Accuracy (MCA):** The proportion of responses in which the correct source material is cited, based on known ground-truth associations between questions and materials:

$$\text{MCA} = \frac{\text{Correctly cited materials}}{\text{Total material references}}$$

- **Response Similarity Score (RSS):** Measures the semantic similarity between the system-generated response and the reference answer, independent of language. This is computed using cosine similarity between the embedding vectors of the generated and expected answers:

$$\text{RSS} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

where \mathbf{A} and \mathbf{B} are embedding vectors of the generated and ground-truth responses, respectively.

4) *Language Switching Handling:* In multilingual educational settings, students often switch languages mid-conversation to improve understanding or clarify concepts. To support such fluid interactions, a robust chatbot must dynamically detect and adapt to language changes within ongoing dialogues.

We evaluate this capability using the **Language Consistency Score (LCS)**, which measures how consistently the chatbot responds in the same language as the user query. For example, if a student transitions from English to Spanish during a conversation, the system should detect the switch and respond in Spanish. The LCS is computed as:

$$\text{LCS} = \frac{\text{Number of language-matched responses}}{\text{Total number of responses}}$$

A higher score indicates better alignment between the response language and the user’s intended language, reflecting the chatbot’s ability to manage dynamic language switching accurately.

5) *Common Language Confusion Management:* A critical challenge in cross-lingual systems is distinguishing between closely related or commonly confused languages, such as Finnish–Estonian or Arabic–Persian. These language pairs often share syntactic structures, vocabulary, or phonetic similarities, making accurate identification and retrieval particularly difficult.

Rather than introducing a separate metric, we evaluate confusion management implicitly through existing metrics above, specifically, LIA, MCA, and RSS. By including these linguistically challenging language pairs in our test suite, we assess whether EduX-RAG can maintain retrieval and response quality without misclassification or degradation. This ensures the framework’s robustness in handling real-world multilingual interactions involving subtle linguistic distinctions.

Region	Language	Topic	Journal Title	Total Question-Answer Pool
East Asian	Chinese	5G Network	5G 网络技术研究现状和发展趋势	162
	Japanese	Society	性犯罪、性暴力に対して日本社会はどう変わったか	131
	Korean	IoT	IoT 디바이스를 활용한 교통안전 시스템의 구현 및 실험	104
South-East Asian	Tagalog	Politics	Ang Estetika, Politika, at Ideolohiyang Pilipino sa Arkitekturang Filipino: Ang Kaso ng Ilang Antigong Bahay sa Malolos, Bulacan	355
	Indonesian	Asian History	Isu Tabungan Orang Melayu dalam Sejarah Pengerjaan Haji Era Kapal Laut	322
Afro-Asiatic	Arabic	Education	التحول الرقمي وأثره على التعليم العالي اليمني من وجهة نظر طلبة كلية التربية جامعة تعز	214
	Persian	Humanities	کمال آدمی و ارتباط آن با تعریف انسان از منظر ملاصدرا	405
	Somali	History	Qaaccessa Qabiyee Faaruu Loonii Hawaasa Oromoo Booranaa	276
European	Spanish	Data Science	Compartir datos de investigación: reflexiones desde el ethos de la ciencia de Robert K. Merton	150
	Finnish	Personal Work Goals	Palkkioiden ja optimistisuuden yhteydet henkilökohtaisiin työtavoitteisiin	287
	Estonian	Literacy	Kuidas lugeda kaaneteksti? Algupäraste proosaraamatute kaanetekstide analüüs	333
	Norwegian	Literature	Det absurde ved å være fanget i blodbankende materie	261
Americas	English	Entrepreneurship	The interplay of context and entrepreneurship: the new frontier for contextualisation research	145
				3145

Fig. 3: Testing Dataset. List of the 13 multilingual journal used in the evaluation of the EduX-RAG framework, showing the region, languages grouped by region, the journal titles, and the number of questions with answers CIPHERBOT generated for each. The 3145 questions were used in the cross-lingual evaluation presented in this research.

V. RESULTS

A. Language Identification

We evaluated three language identification tools across 3,145 multilingual QA pairs (Fig. 3) using the material’s original language as ground truth. As shown in Table I, OpenAI’s model achieved the highest raw accuracy (91.7%), while Azure’s Language Detection API demonstrated superior precision (86.1%) and recall (81.8%). This suggests that while OpenAI may better match the majority language, Azure is more consistent across a diverse language set and better at avoiding misclassifications, making it more suitable for deployment in educational contexts.

Notably, all tools struggled with linguistically similar pairs such as Malay–Indonesian and Somali–Oromo. These results underscore the challenge of differentiating between closely related languages and highlight the need for future work

TABLE I: Language Identification Tools Comparison. Highest values in bold.

Tool	Accuracy	Precision	Recall
LangDetect	0.884	0.762	0.757
Azure API	0.900	0.861	0.818
OpenAI	0.917	0.714	0.667

focused on fine-grained disambiguation in low-resource language settings.

B. Query Resolution Evaluation: Answered and Unanswered

EduX-RAG successfully answered 92% of all queries, demonstrating its ability to retrieve relevant information even in cross-lingual scenarios. Only 7.79% of queries were classified as unanswered, often due to insufficient overlap between query content and indexed material. This high answer rate

illustrates the system’s effectiveness in managing diverse multilingual inputs.

To validate classification accuracy, we reviewed unanswered queries with high embedding similarity scores (above 0.7) and found that 31% may have been misclassified. A majority (85%) of these occurred in Chinese-related queries, suggesting language-specific retrieval or encoding issues that warrant further investigation. Additionally, 28.2% of answered queries had lower-than-expected similarity scores (below the 25th percentile), suggesting that while answers were generated, they may have lacked alignment with the most relevant material.

Despite these edge cases, the overall classification accuracy of 71.57% remains strong given the complexity of multilingual retrieval. Future work can explore adaptive thresholds or confidence-based mechanisms to better distinguish ambiguous cases.

C. Multilingual Content Retrieval Accuracy

1) *Material Citation Accuracy:* To evaluate EduX-RAG’s ability to accurately cite educational materials across multiple languages, we analyzed the responses to 577 multilingual questions. The goal was to assess how well EduX-RAG retrieved the correct source material in its responses, even when the query language differed from the language of the educational content. Out of the 577 questions processed, only one instance of material mis-reference was observed. Table II details the conversation where this mis-referencing occurred. Regardless, the overall material citation accuracy remains extremely high, with this being the only observed mis-reference out of 577 interactions, yielding 99% accuracy on this metric.

2) *Response Similarity Score:* Figure 4 illustrates the distribution of cosine similarity scores between generated and reference answers. Answered queries consistently show higher similarity values, validating the effectiveness of the retrieval and response generation pipeline. Unanswered queries show more variance, but some still exceed the 0.7 similarity threshold, further supporting our earlier observation of conservative misclassification.

The mean similarity score across all responses was 0.70 (SD = 0.20), with a peak of 0.98, indicating high semantic alignment even under challenging multilingual conditions. These results demonstrate that EduX-RAG generates content that remains semantically faithful to ground-truth responses across languages.

D. Language Switching Handling

EduX-RAG achieved a 73% language consistency score in our evaluation of language switching, indicating that the system responded in the same language as the query in the majority of cases. The remaining 27.21% (157 out of 577 cases) were language mismatches, most frequently involving Tagalog, Finnish, Estonian, Korean, and Spanish.

These results suggest that while the system handles most mid-conversation language transitions effectively, further refinement is needed for real-time language switch detection,

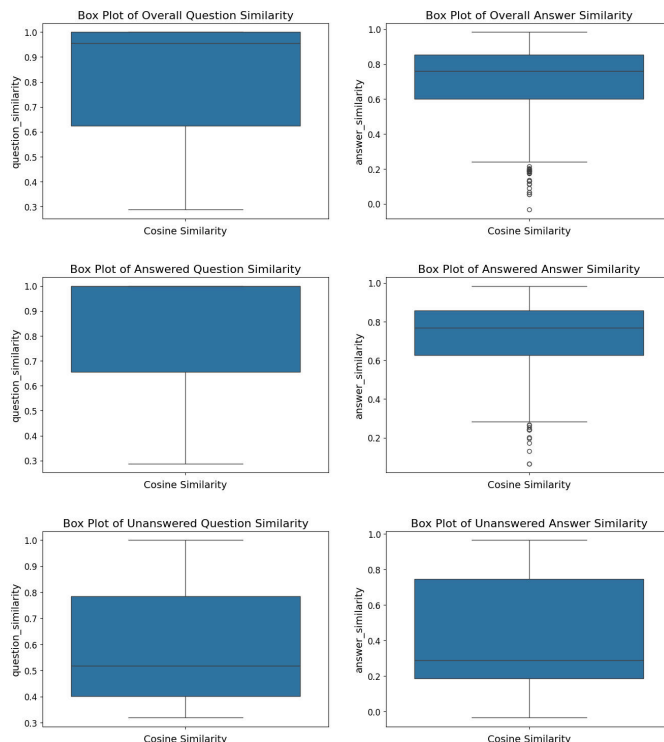


Fig. 4: Box plots for the different Question and Answer similarity scores.

especially for less frequently seen or morphologically complex languages. Enhancing language continuity in response generation remains an important direction for future work.

VI. DISCUSSION

EduX-RAG was designed to address two central research questions: whether RAG framework can effectively manage multilingual content in educational settings (**RQ1**), and whether prompt engineering techniques can support language-switching in dynamic, cross-lingual conversations (**RQ2**).

For **RQ1**, our results strongly support the viability of a RAG-based approach for multilingual educational dialogue. The framework achieved a 92% answered question rate and 99% material citation accuracy, even when the query and content were in different languages. This shows that embedding-based retrieval is not only language-agnostic in theory but functionally reliable in practice when paired with multilingual educational materials. Moreover, the high performance across both high-resource and closely related language pairs (e.g., Finnish–Estonian, Arabic–Persian) confirms the robustness of the retrieval pipeline in linguistically ambiguous contexts, which was a known challenge for prior chatbot systems. The 90% language identification accuracy further contributed to this performance by ensuring that queries were routed through the appropriate language-specific retrieval processes. While this step may seem straightforward, it is critical in multilingual deployments where incorrect detection could lead to retrieval from entirely irrelevant content domains. These results confirm

Question	Baseline Question	Question Similarity	Answer	Baseline Answer	Answer Similarity	Question Language	Baseline Language	Session Topic	Referenced Material
研究创业学强调什么方法？	What methodological approach is emphasized for studying entrepreneurship?	0.652	研究创业学强调多种**方法论**，包括：- **访谈**：通过与研究对象进行对话，获取定性数据；**观察**：直接观察创业活动和环境，以获得真实的情境数据；**焦点小组讨论**：集中一组人讨论特定主题，收集不同观点和经历。这些方法有助于深入理解创业过程及其背景，推动创业研究的发展 [1]	An ethnographic approach, particularly action research, is emphasized for studying entrepreneurship.	0.546	zh_chs	en	English - Entrepreneurship	Somali - History

TABLE II: Only mis-referenced response from the chatbot. In this case, the question was posed in Chinese, but the material referenced in the response came from a Somali history journal rather than the correct English journal on entrepreneurship.

that RAG can be extended beyond monolingual environments if the pipeline is carefully structured to support language awareness at each step.

In addressing **RQ2**, EduX-RAG’s dynamic prompt engineering was effective in preserving conversational coherence during mid-session language transitions. A 73% language consistency score indicates that, in most cases, the chatbot responded in the same language as the query, even in settings involving code-switching or rapid language shifts. This level of control, achieved through prompt design rather than retraining, shows promise for systems that need to adapt flexibly to multilingual input without tightly coupling to a single model per language. However, the 27% mismatch rate suggests that generation remains a more brittle component than retrieval in this architecture. This likely reflects the current limitations of generative models, particularly in cross-lingual settings where the system must infer not only intent but also linguistic preference. Nevertheless, the use of prompts for controlling response language rather than hard-coded templates or rule-based post-processing offers a scalable direction for managing multilingual dialogue.

VII. CONCLUSION

This work introduced EduX-RAG, a framework that combines Retrieval-Augmented Generation with prompt engineering to support multilingual educational chatbots. Evaluated across 13 languages and deployed in multilingual academic environments, EduX-RAG demonstrated strong performance in retrieval accuracy, language identification, and multilingual response generation. Its design enables context-aware, grounded responses, even in complex cases involving language switching or closely related language pairs. While challenges remain particularly in generation-level language control and system generalization to underrepresented languages, EduX-RAG provides a solid foundation for building more inclusive and adaptable AI-driven learning tools. By prioritizing retrieval and linguistic flexibility, it offers a practical step toward more accessible educational technologies.

REFERENCES

- [1] C. Kooli, “Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions,” *Sustainability*, vol. 15, no. 7, p. 5614, Mar. 2023. [Online]. Available: <https://www.mdpi.com/2071-1050/15/7/5614>
- [2] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, “What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education,” *Smart Learning Environments*, vol. 10, no. 1, p. 15, Feb. 2023. [Online]. Available: <https://slejournal.springeropen.com/articles/10.1186/s40561-023-00237-x>
- [3] B. Li, H. Fei, F. Su, F. Li, and D. Ji, “Integrating discourse features and response assessment for advancing empathetic dialogue,” *Information Processing & Management*, vol. 61, no. 5, p. 103803, Sep. 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0306457324001626>
- [4] J. Liu, C. Yan, W. Liu, X. Liu, Y. Ding, and Y. Zhou, “Research on learner “emotion-behavior-ability” characteristics based on mooc online education user profiles,” *Information Processing & Management*, vol. 62, no. 3, p. 104026, 2025.
- [5] B. Abeysinghe and R. Circi, “The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches,” Jun. 2024, arXiv:2406.03339 [cs]. [Online]. Available: <http://arxiv.org/abs/2406.03339>
- [6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, Dec. 2023, arXiv:2202.03629 [cs]. [Online]. Available: <http://arxiv.org/abs/2202.03629>
- [7] J. Salminen, S.-G. Jung, J. Medina, K. Aldous, J. Azem, W. Akhtar, E. Häyhänen, and B. J. Jansen, “Communication Design for an Educational AI Chatbot: Analyzing CIPHERbot’s Communication Style and Challenges,” in *Proceedings of the 27th International Academic Mindtrek Conference*. Tampere Finland: ACM, Oct. 2024, pp. 176–187. [Online]. Available: <https://dl.acm.org/doi/10.1145/3681716.3681727>
- [8] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. De Choudhury, and S. Kumar, “Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries,” in *Proceedings of the ACM Web Conference 2024*. Singapore Singapore: ACM, May 2024, pp. 2627–2638. [Online]. Available: <https://dl.acm.org/doi/10.1145/3589334.3645643>
- [9] L. P. Ojha, J. Burton, and P. I. De Costa, “Critical teacher education for equitable learning in multilingual classrooms: A possible way forward,” pp. 197–207, 2024.
- [10] S. Zhu, L. Pan, D. Jian, and D. Xiong, “Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models,” *Information Processing & Management*, vol. 62, no. 3, p. 104078, 2025.
- [11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.”
- [12] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, “Retrieval-Augmented Generation for AI-Generated Content: A Survey,” Jun. 2024, arXiv:2402.19473 [cs]. [Online]. Available: <http://arxiv.org/abs/2402.19473>
- [13] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, “Active Retrieval Augmented Generation,” Oct. 2023, arXiv:2305.06983 [cs]. [Online]. Available: <http://arxiv.org/abs/2305.06983>
- [14] J. Salminen, S.-g. Jung, J. Medina, K. Aldous, J. Azem, W. Akhtar, and B. J. Jansen, “Using CIPHERbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot,” in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*.

- Atlanta GA USA: ACM, Jul. 2024, pp. 279–283. [Online]. Available: <https://dl.acm.org/doi/10.1145/3657604.3664690>
- [15] H. Guldal and E. O. Diner, “Can rule-based educational chatbots be an acceptable alternative for students in higher education?” *Education and Information Technologies*, vol. 30, no. 3, pp. 3979–4012, 2025.
- [16] L. J. Labrague and S. Al Sabei, “Integration of ai-powered chatbots in nursing education: a scoping review of their utilization, outcomes, and challenges,” *Teaching and Learning in Nursing*, vol. 20, no. 1, pp. e285–e293, 2025.
- [17] S. Abedu, A. Abdellatif, and E. Shihab, “LLM-Based Chatbots for Mining Software Repositories: Challenges and Opportunities,” in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*. Salerno Italy: ACM, Jun. 2024, pp. 201–210. [Online]. Available: <https://dl.acm.org/doi/10.1145/3661167.3661218>
- [18] Z. Chu, S. Wang, J. Xie, T. Zhu, Y. Yan, J. Ye, A. Zhong, X. Hu, J. Liang, P. S. Yu *et al.*, “Llm agents for education: Advances and applications,” *arXiv preprint arXiv:2503.11733*, 2025.
- [19] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, “The role of large language models in personalized learning: a systematic review of educational impact,” *Discover Sustainability*, vol. 6, no. 1, pp. 1–24, 2025.
- [20] B. Dong, J. Bai, T. Xu, and Y. Zhou, “Large Language Models in Education: A Systematic Review,” in *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, vol. 37. Xi’an, China: IEEE, Apr. 2024, pp. 131–134. [Online]. Available: <https://ieeexplore.ieee.org/document/10589960/>
- [21] M. Bahja, R. Hammad, and G. Butt, “A User-Centric Framework for Educational Chatbots Design and Development,” in *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, C. Stephanidis, M. Kurosu, H. Degen, and L. Reinerman-Jones, Eds. Cham: Springer International Publishing, 2020, vol. 12424, pp. 32–43, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-030-60117-1_3
- [22] G. Ilieva, T. Yankova, S. Klisarova-Belcheva, A. Dimitrov, M. Bratkov, and D. Angelov, “Effects of Generative Chatbots in Higher Education,” *Information*, vol. 14, no. 9, p. 492, Sep. 2023. [Online]. Available: <https://www.mdpi.com/2078-2489/14/9/492>
- [23] G.-J. Hwang and C.-Y. Chang, “A review of opportunities and challenges of chatbots in education,” *Interactive Learning Environments*, vol. 31, no. 7, pp. 4099–4112, Oct. 2023. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/10494820.2021.1952615>
- [24] C.-C. Lin, A. Y. Q. Huang, and S. J. H. Yang, “A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022),” *Sustainability*, vol. 15, no. 5, p. 4012, Feb. 2023. [Online]. Available: <https://www.mdpi.com/2071-1050/15/5/4012>
- [25] A. Salemi and H. Zamani, “Evaluating Retrieval Quality in Retrieval-Augmented Generation,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Washington DC USA: ACM, Jul. 2024, pp. 2395–2400. [Online]. Available: <https://dl.acm.org/doi/10.1145/3626772.3657957>
- [26] J. Saad-Falcon, O. Khatlab, C. Potts, and M. Zaharia, “ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems,” Mar. 2024, arXiv:2311.09476 [cs]. [Online]. Available: <http://arxiv.org/abs/2311.09476>
- [27] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, “Evaluation of Retrieval-Augmented Generation: A Survey,” Jul. 2024, arXiv:2405.07437 [cs]. [Online]. Available: <http://arxiv.org/abs/2405.07437>
- [28] J. Guo, Y. Li, and K. Tan, “Triplet-modality group-guided incremental distillation with regularized group semantic consistency for multi-modal neural machine translation,” *Information Processing & Management*, vol. 62, no. 5, p. 104149, 2025.
- [29] G. Bilquise, S. Ibrahim, and K. Shaalan, “Bilingual AI-Driven Chatbot for Academic Advising,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, 2022. [Online]. Available: <http://thesai.org/Publications/ViewPaper?Volume=13&Issue=8&Code=IJACSA&SerialNo=8>
- [30] P. Polakova and B. Klimova, “Implementation of AI-driven technology into education – a pilot study on the use of chatbots in foreign language learning,” *Cogent Education*, vol. 11, no. 1, p. 2355385, Dec. 2024. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/2331186X.2024.2355385>
- [31] L. Rodrigues, C. Xavier, N. Costa, H. Batista, L. F. B. Silva, W. Chaleghi de Melo, D. Gasevic, and R. Ferreira Mello, “Llms performance in answering educational questions in brazilian portuguese: A preliminary analysis on llms potential to support diverse educational needs,” in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 2025, pp. 865–871.
- [32] B. Alsafari, E. Atwell, A. Walker, and M. Callaghan, “Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants,” *Natural Language Processing Journal*, vol. 8, p. 100101, Sep. 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2949719124000499>
- [33] D. Hennekeuser, D. D. Vaziri, D. Golchinfar, D. Schreiber, and G. Stevens, “Enlarged Education – Exploring the Use of Generative AI to Support Lecturing in Higher Education,” *International Journal of Artificial Intelligence in Education*, Aug. 2024, publisher: Springer Science and Business Media LLC. [Online]. Available: <https://link.springer.com/10.1007/s40593-024-00424-y>
- [34] S. Athanassopoulos, P. Manoli, M. Gouvi, K. Lavidas, and V. Komis, “The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom,” *Advances in Mobile Learning Educational Research*, vol. 3, no. 2, pp. 818–824, Aug. 2023. [Online]. Available: <https://www.syncsci.com/journal/AMLER/article/view/AMLER.2023.02.009>
- [35] Y. A. Mohamed, A. Khanan, M. Bashir, A. H. H. M. Mohamed, M. A. E. Adiel, and M. A. Elsadig, “The Impact of Artificial Intelligence on Language Translation: A Review,” *IEEE Access*, vol. 12, pp. 25 553–25 579, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/10438431/>
- [36] S. Dharshini, R. Venkatesan *et al.*, “Mindmate: Ai-powered multilingual mental health chatbot with personalized voice and text support with rasa and streamlit,” in *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2025, pp. 1104–1109.
- [37] D. Kennedy, E. Halim, A. R. Condrobimo, D. Syamsuar, and F. Ferdianto, “Overcoming Language Barriers in MOOCs with Artificial Intelligence: An AI-based Approach for Multilingual Education,” in *2023 Eighth International Conference on Informatics and Computing (ICIC)*. Manado, Indonesia: IEEE, Dec. 2023, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/10381900/>
- [38] H. Yu, Y. Guo, H. Yang, W. Zhang, and Y. Dong, “Can chatgpt revolutionize language learning? unveiling the power of ai in multilingual education through user insights and pedagogical impact,” *European Journal of Education*, vol. 60, no. 1, p. e12749, 2025.
- [39] N. Chirkova, D. Rau, H. Djean, T. Formal, S. Clinchant, and V. Nikoulina, “Retrieval-augmented generation in multilingual settings,” Jul. 2024, arXiv:2407.01463 [cs]. [Online]. Available: <http://arxiv.org/abs/2407.01463>
- [40] S. Kumar, V. Balloli, M. Ranjit, K. Ahuja, T. Ganu, S. Sitaram, K. Bali, and A. Nambi, “Bridging the Gap: Dynamic Learning Strategies for Improving Multilingual Performance in LLMs,” May 2024, arXiv:2405.18359 [cs]. [Online]. Available: <http://arxiv.org/abs/2405.18359>
- [41] Z. Zhong, H. Liu, X. Cui, X. Zhang, and Z. Qin, “Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation,” Jun. 2024, arXiv:2406.00456 [cs]. [Online]. Available: <http://arxiv.org/abs/2406.00456>
- [42] X. Zhang, N. Thakur, O. Ogundepo, E. Kamaloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin, “Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages,” Oct. 2022, arXiv:2210.09984 [cs]. [Online]. Available: <http://arxiv.org/abs/2210.09984>
- [43] H. Skirgrd, S. G. Roberts, and L. Yencken, “Why are some languages confused for others? Investigating data from the Great Language Game,” *PLOS ONE*, vol. 12, no. 4, p. e0165934, Apr. 2017. [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0165934>
- [44] PyPi, “Python LangDetect Library.” [Online]. Available: <https://pypi.org/project/langdetect/>
- [45] Microsoft, “Language Detection - Azure AI Language Service,” 2024. [Online]. Available: <https://learn.microsoft.com/en-us/azure/ai-services/language-service/language-detection/overview>
- [46] OpenAI, “OpenAI 4o-mini,” 2024. [Online]. Available: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>