ADScAI Applied Data Science & Artificial Intelligence Symposium

Developing a Question Answering System for the Sri Lankan School Education System*

Y.M.W.H.M.R.P.J.R.B. Kiridana
Dept. of Computer Science Engineering
University of Moratuwa
Katubedda, Sri Lanka
punsisi.24@cse.mrt.ac.lk

Gihan Dias
Dept. of Computer Science Engineering
University of Moratuwa
Katubedda, Sri Lanka
gihan@cse.mrt.ac.lk

***Abstract:*The integration of Artificial Intelligence (AI) into education unlocks opportunities for personalized learning. However, low-resource languages such as Sinhala currently lack robust Natural Language Processing (NLP) tools. This paper proposes a question-answering system tailored for the Sri Lankan school curriculum, designed to retrieve curriculum-based answers from structured educational materials. To address the digital divide in rural areas, an offline-accessible version is planned. The study outlines a framework for development and a proposed evaluation using standard NLP metrics and user feedback, aiming to create a scalable, effective tool for Sinhala-language learners.**

***Keywords*—Low-Resource Languages, Natural Language Processing (NLP), Educational Technology, Retrieval-Augmented Generation (RAG), Zero-Shot Learning, Domain-Specific Training**

I. Introduction

Artificial Intelligence (AI) has enhanced education through intelligent tutoring and automated question-answering (QA) systems. However, low-resource languages like Sinhala remain underrepresented due to limited datasets and adaptations [1]. A QA system tailored for Sri Lanka's school curriculum will offer students an opportunity to obtain precise answers to their questions on the curriculum. Existing tools, primarily English-focused and internet-dependent, exclude many rural students.

This study proposes a Sinhala-language QA system trained on school textbooks to deliver curriculum-aligned answers. The approach leverages fine-tuned large language models, retrieval-augmented generation (RAG), and parameter-efficient techniques to ensure accuracy and efficiency. An offline version is planned for inclusivity. This paper outlines the proposed framework and evaluation strategy.

II. Literature Review

Recent advances in Natural Language Processing (NLP) have transformed educational technology, particularly through Large Language Models (LLMs) like GPT-4, mT5, and XLM-R, which excel in tasks such as question answering and text generation [2]–[4]. The Transformer architecture, with its self-attention mechanism, underpins these models, enabling efficient text processing [5]. However, low-resource languages like Sinhala face challenges due to limited training data, impacting model performance [1].

To address such challenges, frameworks like adaptMLLM enhance LLM fine-tuning for low-resource languages [6], while techniques such as pruning and knowledge distillation improve computational efficiency [7]. Retrieval-Augmented Generation (RAG) combines retrieval and generation for contextually accurate responses [8], and tools like the PDF-to-Tree framework structure unstructured documents for QA tasks [9]. Additionally, visual IDEs like AI2Apps and automated prompt engineering (APE) streamline development and optimization [10] [11]. These technologies inform the proposed system's design and deployment.

III. Methodology

This study proposes a Sinhala-language QA system through a pipeline that processes user queries to generate curriculum-aligned answers. The workflow, illustrated in Figure 1, integrates data preparation, query processing, and response generation, tailored for the Sri Lankan school curriculum.

A. Data Preparation and Vectorization

Grade 11 Science textbooks and teacher guides in PDF format will be processed using the PDF-to-Tree framework [9] to convert text into a hierarchical structure (e.g., chapters, sections). The structured content will be vectorized using XLM-R Transformer embeddings and stored in a vectorized database using FAISS [12] for efficient similarity search.

B. Query Processing

The system will process user queries in Sinhala through the following steps:

- **Translation**: Queries will be translated from Sinhala to English using mBART [4] to leverage high-performance English-based models.

https://doi.org/10.31705/ADScAI.2025.21

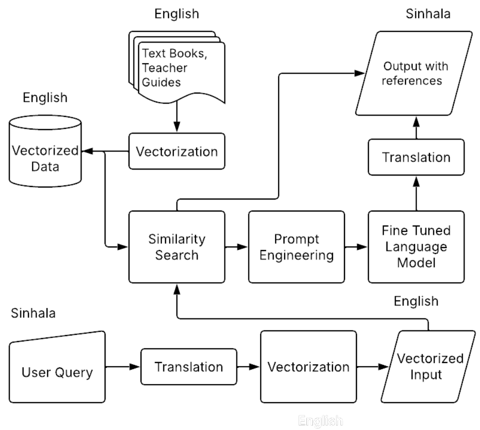

Fig. 1. Proposed QA System Workflow

- **Vectorization**: The translated query will be vectorized using Transformer embeddings, creating a vectorized input for further processing.

C. Answer Generation

The vectorized input will be processed to generate answers:

- **Fine-Tuned Language Model**: A fine-tuned XLM-R model [4], optimized with LoRA [13], will interpret the vectorized input.
- **Prompt Engineering**: Automated Prompt Engineering (APE) [11] will optimize prompts to improve the model's response quality.
- **Similarity Search**: The system will perform a similarity search against the vectorized database using FAISS [12] to retrieve relevant curriculum content.

D. Output Generation

The retrieved content will be processed to produce the final output:

- **Translation**: The generated answer in English will be translated back to Sinhala using mBART [4].
- **Output with References**: The system will provide the answer in Sinhala, including references to the textbook sections used, ensuring transparency and educational value.

E. Interface Deployment

A web-based interface will be built using AI2Apps [10], with an offline desktop version for rural access, enabling students to input queries and receive structured responses.

IV. Evaluation

The proposed QA system will be evaluated using a two-pronged approach: quantitative NLP metrics and qualitative user feedback. This section outlines the planned methodology, addressing performance and usability across the pipeline.

A. Quantitative Evaluation

The system's performance will be assessed at multiple stages of the pipeline:

- **Translation Accuracy**: The Sinhala-to-English and English-to-Sinhala translation steps will be evaluated using BLEU and METEOR scores [4], ensuring linguistic fidelity.
- **Similarity Search Effectiveness**: The FAISS-based similarity search will be measured using mean reciprocal rank (MRR) to assess the relevance of retrieved content.
- **Overall QA Performance**: Precision, recall, and F1-score will evaluate the system's answer accuracy against a test set of curriculum-based questions, manually annotated with expected answers from the Grade 11 Science textbook.

The system's performance will be compared to a baseline English-only QA model (e.g., a RAG implementation without Sinhala adaptation) to highlight improvements in language-specific retrieval.

B. Qualitative Evaluation

User feedback will be collected from a sample of Grade 11 students, teachers via surveys, focusing on answer relevance, system usability, and offline accessibility. Open-ended questions will identify areas for refinement, such as response time, translation quality, or interface design. This dual approach ensures the system meets both technical and educational goals.

References

[1] P. J. et al., "The state and fate of linguistic diversity and inclusion in the nlp world," *Proceedings of ACL*, 2020.
[2] T. B. et al., "Language models are few-shot learners," vol. 33, 2020, pp. 1877–1901.
[3] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, S. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *Journal of Machine Learning Research (JMLR)*, 2021.
[4] A. Conneau, A. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of NeurIPS*, 2019.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," 2017.
[6] Y. W. et al., "adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds," *arXiv preprint arXiv:2305.10988*, 2023.
[7] S. T. Sreenivas, S. Muralidharan, R. Joshi, M. Chochowski, M. Patwary, M. Shoeybi, B. Catanzaro, J. Kautz, and P. Molchanov, "Llm pruning and distillation in practice: The minitron approach," *arXiv preprint arXiv:2408.11796*, 2024.
[8] P. L. et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of NeurIPS*, 2020.
[9] Y. Z. et al., "Pdf-to-tree: Parsing pdf text blocks into a tree," *Findings of EMNLP*, 2024.
[10] X. Pang, Z. Li, J. Chen, Y. Cheng, Y. Xu, and Y. Qi, "Ai2apps: A visual ide for building llm-based ai agent applications," *arXiv preprint arXiv:2404.04902*, 2024.
[11] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," 2022.
[12] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," in *IEEE Transactions on Big Data*, 2017.
[13] E. H. et al., "Lora: Low-rank adaptation of large language models," in *Proceedings of ICLR*, 2021.