IT4010 Research Project – 2025 July

Topic Assessment Form

Project ID : 25-26J-448

1. Topic (12 words max)

AI-Powered Educational Assistant for Sinhala Resources

2. Research group the project belongs to

CoEAI - Centre of Excellence for AI

3. Specialization of the project belongs to

Software Engineering (SE)

4. If a continuation of a previous project:

Project ID	
Year	

5. Brief description of the research problem including references (200 – 500 words max) – references not included in word count.

Sri Lanka's education system lacks effective AI tools for Sinhala-language learning. Current solutions face three key challenges:

1. **Unreliable AI Responses**: Existing chatbots often generate incorrect answers not supported by source materials, especially problematic for Sinhala content.
2. **Voice and Offline Limitations**: Most tools don't work without internet or support Sinhala voice commands effectively, particularly for Sri Lankan accents.
3. **Inefficient Grading**: Teachers waste time manually evaluating answers, as automated systems struggle with Sinhala's linguistic nuances.

Additionally, poor OCR performance on Sinhala handwritten materials hinders digital conversion. While some Sinhala NLP research exists, no complete solution combines accurate Q&A, voice interaction, and automated grading in an offline mobile application. This project will develop an AI assistant that:

- Provides strictly source-based answers
- Works offline with Sinhala voice support
- Automates answer evaluation

The solution will address critical gaps in Sri Lanka's digital education infrastructure.

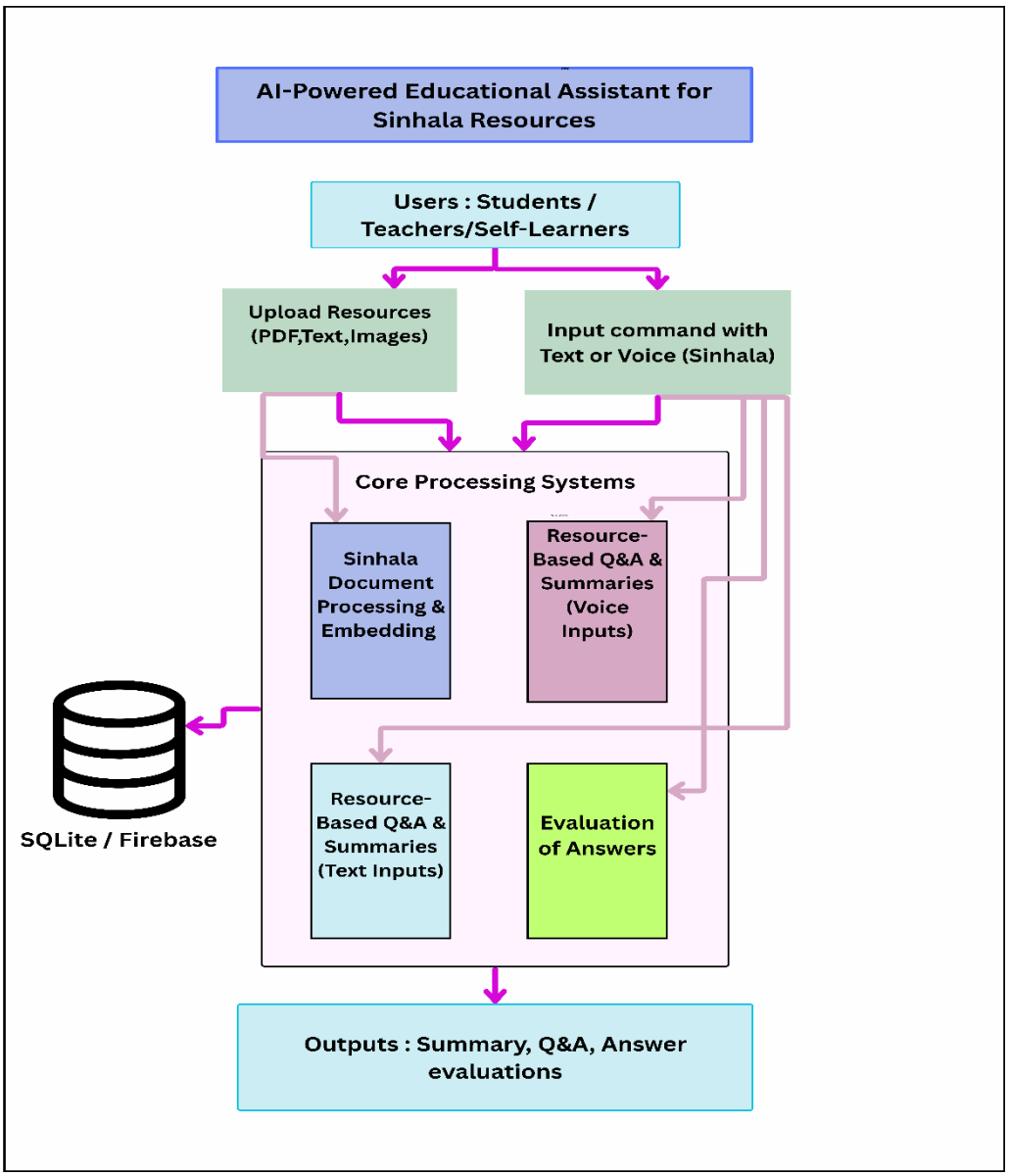
References

1. University of Moratuwa (2021). NLP for Sinhala
2. Google AI (2023). Whisper Speech Recognition
3. Sri Lankan Ministry of Education (2022). Digital Learning Survey

6. Brief description of the nature of the solution including a conceptual diagram (250 words max)

The solution combines:

1. **Sinhala Document Processing & Embedding:** Combines enhanced Tesseract OCR with custom preprocessing for handwritten Sinhala text, generating embeddings via a fine-tuned Llama 2 model stored in SQLite for offline access.
2. **Voice-Based Q&A:** Uses Whisper-based speech recognition fine-tuned for Sri Lankan accents, paired with offline/online capabilities
3. **Text-Based Q&A & Summaries:** Implements a Sinhala-specific RAG pipeline with strict source constraints to prevent hallucinations, alongside contextual summarization tools.
4. **Answer Evaluation:** Leverages embeddings and rule-based checks to grade Sinhala answers semantically, providing explainable feedback and adaptive thresholds.



7. Brief description of specialized domain expertise, knowledge, and data requirements (300 words max)

This research project demands specialized expertise across multiple AI domains to develop a comprehensive Sinhala educational assistant. The solution requires advanced NLP capabilities for fine-tuning language models (like Llama 2) to handle Sinhala's unique morphology, particularly for the answer evaluation system which combines semantic similarity analysis with rule-based checks to assess student responses. For document processing, expertise in optimizing Tesseract OCR with custom preprocessing pipelines is essential to accurately digitize both typed and handwritten Sinhala materials, which are then converted into searchable embeddings using fine tuned models. The voice interface component necessitates speech processing specialization to adapt Whisper models for Sri Lankan accents that function reliably offline. Simultaneously, the text based Q&A system requires sophisticated RAG implementation with strict source constraints to prevent hallucinations while generating educational summaries.

Critical data requirements include extensive collections of scanned textbooks and handwritten notes (10,000+ samples) for OCR training, along with thousands of annotated Q&A pairs for RAG development. The system needs 500+ hours of diverse Sinhala speech recordings with regional accent variations for voice model training, plus graded student answers with teacher rubrics to refine the evaluation module. Mobile implementation through Flutter requires optimization of on device models for offline functionality while maintaining performance. The project emphasizes privacy preserving data practices and continuous educator collaboration to ensure practical utility, with rigorous benchmarking against local educational standards. This multidisciplinary approach combines cutting-edge AI techniques with deep understanding of Sri Lanka's linguistic and pedagogical context, addressing critical challenges in offline access, voice interaction, and reliable content processing to create a truly localized educational tool.

8. Objectives and Novelty

Main Objective

1. Develop an **offline-capable** AI educational assistant for Sinhala-medium students/teachers.
2. Enable **multimodal input/output** (text, voice, PDF/image uploads).
3. Ensure **resource-constrained accuracy** (answers strictly from uploaded documents).
4. Advance **Sinhala NLP** in OCR, Q&A, question generation, and evaluation.

Member Name with Registration No	Sub Objective	Tasks	Novelty
Ranaweera P.H.K IT22234452	Sinhala Document Processing (Including Handwritten Text) & Embedding	Develop specialized preprocessing pipelines for Sinhala OCR that handle unique challenges like connected letters and vowel modifiers Create handwriting recognition enhancements through custom-trained models on Sri Lankan writing samples Optimize Llama 2 embeddings for Sinhala educational content by fine-tuning on textbook corpora Engineer a local SQLite database system that efficiently stores and retrieves document embeddings for offline use Build semantic search functionality that understands educational context and conceptual relationships	**Handwritten Sinhala OCR:** Custom preprocessing techniques improve accuracy for Sri Lankan handwriting, a challenge ignored by most OCR tools. **Offline Embedding Storage:** Lightweight embeddings (e.g., quantized Llama 2) are cached in SQLite, enabling resource-based Q&A without internet. **Unified Pipeline:** Combines typed, scanned, and handwritten inputs into a single workflow, reducing manual effort for teachers.
Sathsara T.T.D IT22362476	Resource-Based Q&A & Summaries (Voice Inputs)	Adapt Whisper speech recognition through accent-specific training using Sri Lankan voice samples Develop a hybrid TTS system that seamlessly switches between online and offline modes based on connectivity Optimize real-time processing for	**Accent-Aware STT:** Whisper is fine-tuned to better transcribe Sinhala spoken in Sri Lankan dialects and accents. **Voice Feedback with Citations**: The system speaks the answer and includes source citations (e.g., “As per Page 14...”), enhancing user trust.

		mobile devices to ensure instant response during lessons Create specialized pronunciation handling for complex academic and technical terms in Sinhala	**Offline/Hybrid Voice Processing:** Lightweight Sinhala STT/TTS models are included for use in low-connectivity or rural environments. **Audio Interface Integration:** A natural, intuitive voice interface is developed for Sinhala educational Q&A workflows.
Jayananda L.V.O.R IT22161406	Resource-Based Q&A & Summaries (Text Inputs)	Develop constrained RAG pipeline for Sinhala Implement source-bound answer generation Create educational context-aware summarization Build document retrieval system Develop hallucination prevention mechanisms Develop a constrained RAG architecture that strictly binds answers to retrieved source materials Implement advanced retrieval algorithms that understand Sinhala morphology and synonyms Create context-aware summarization that organizes content	**Zero-Hallucination Guarantee:** The generation pipeline is constrained to ensure that all answers are derived strictly from retrieved documents. **Sinhala-Aware RAG:** Retrieval components are tuned to handle Sinhala's unique linguistic characteristics (e.g., agglutinative morphology), improving result quality. **Contextual Summarization:** Generates topic-specific, concise summaries of Sinhala content, optimized for student and teacher use. **High-Precision Evaluation:** Uses educational rubrics and NLP metrics (BLEU, ROUGE) to assess and refine summarization and Q&A performance.
Lokuhewage M.M IT22003478	Evaluation of Answers	Develop semantic analysis models using XLM-Roberta to compare student answers against reference materials through embedding similarity calculations	**Paraphrase-Aware Grading**: Evaluates Sinhala answers for meaning (not just keywords) via embeddings. **Explainable Feedback:**

		Implement rule-based validation checks that verify the presence of key concepts and required terminology in responses Create a dynamic grading system with configurable strictness Design an interactive interface that displays automated evaluations with override capabilities and suggestion features Generate detailed feedback reports highlighting missing concepts, partial matches, and suggested improvements	Highlights missing concepts or deviations from source material. **Adaptive Thresholding:** Adjusts grading strictness based on document complexity (e.g., exams vs. notes).

9. Individual component description of how it is complied with the specialization.

Member Name with Registration No	Description
Ranaweera P.H.K IT22234452	**Sinhala Document Processing (Including Handwritten Text) & Embedding** **Specialization Compliance: AI/ML and Data Engineering** This component aligns perfectly with AI/ML and data engineering specializations through: • **OCR Optimization**: Requires expertise in computer vision and preprocessing techniques to enhance Tesseract's performance for Sinhala text, particularly challenging handwritten forms • **Embedding Engineering**: Involves creating efficient vector representations of Sinhala text, demanding knowledge of transformer architectures and dimensionality reduction • **Offline Storage Design**: Needs database optimization skills to implement SQLite caching of embeddings for offline access • **Pipeline Architecture**: Demands data engineering skills to build a robust document processing workflow that handles multiple input formats The specialist working on this component will apply their knowledge of machine learning models, data preprocessing, and storage systems to create a seamless document ingestion pipeline. In addition to core development, containerize, tests them locally, and pushes versioned Docker images to Amazon ECR for integration into the deployment pipeline.
Sathsara T.T.D IT22362476	**Resource-Based Q&A & Summaries (Voice Inputs)** **Specialization Compliance: Speech Processing and Mobile Development** This component aligns with speech processing and mobile development through: • **Accent Adaptation**: Meticulously adjusting Whisper speech recognition models to accurately interpret the rich variety of Sri Lankan accents and dialects • **Hybrid TTS Architecture:** Building a sophisticated text-to-speech system that intelligently switches between high-quality online synthesis and reliable offline operation • **Real-time Optimization:** Engineering highly efficient voice processing pipelines within Flutter that deliver responsive performance even on budget mobile devices • **Educational UI Design:** Creating natural voice interaction experiences specifically tailored to classroom environments and teacher needs The specialist will apply speech processing knowledge to handle Sinhala voice inputs and mobile development skills to create a responsive interface. In addition, setup the Kubernetes environment using Amazon EKS, writes deployment YAML files, configures services and ingress, and ensures smooth deployment using kubectl.

Jayananda L.V.O.R IT22161406	**Resource-Based Q&A & Summaries (Text Inputs)** **Specialization Compliance: NLP and Backend Development** This component matches NLP and backend development specializations by requiring: • **RAG Implementation**: Needs expertise in retrieval-augmented generation systems and LLM fine-tuning • **Query Processing**: Requires advanced NLP skills to handle Sinhala language peculiarities in question parsing • **API Development**: Demands backend skills to create efficient endpoints for text-based queries • **Database Integration**: Needs knowledge of PostgreSQL and Firebase for document storage and retrieval • **Hallucination Prevention**: Requires specialized NLP techniques to enforce strict source-based responses The specialist will leverage their NLP knowledge to build accurate Q&A systems and backend skills to ensure smooth integration with other components. Alongside this, implements Jenkins CI/CD pipelines to automate image building and deployment processes, pushing services to Amazon ECR and deploying them to the EKS cluster with GitHub integration.
Lokuhewage M.M IT22003478	**Evaluation of Answers** **Specialization Compliance: AI/ML and Educational Technology** This component matches AI/ML and educational technology specializations by involving: • **Semantic Analysis**: Requires expertise in NLP similarity metrics and embedding comparisons • **Adaptive Grading**: Needs ML skills to implement dynamic evaluation thresholds • **Feedback Systems**: Demands educational technology knowledge to design useful teacher feedback mechanisms • **Explainable AI**: Requires skills in creating interpretable evaluation reports • **Workflow Integration**: Needs understanding of teacher-student interactions in educational settings The specialist will use their AI/ML knowledge to build accurate evaluation systems and edtech expertise to ensure practical utility for educators. In addition to development, validates the deployed system on EKS, applies monitoring tools, and documents the complete deployment process, including scaling strategies and troubleshooting steps.

10. Supervisor details

SLIIT UNI
THE KNOWLEDGE UNIVERSITY

IT4010 Research Project – 2025 July

Topic Assessment Form

10. Supervisor details

	Title	First Name	Last Name	Signature
Supervisor	Prof.	Dilshan	De Silva	
Co-Supervisor	Ms.	Chamali	Pabasara	27.06.2025 (on behalf of Ms. Chamali)
External Supervisor				
Summary of external supervisor's (if any) experience and expertise				

This part is to be filled by the Topic Screening Staff members.

a) Does the chosen research topic possess a comprehensive scope suitable for a final-year project?

Yes		No	

b) Does the proposed topic exhibit novelty?

Yes		No	

c) Do you believe they have the capability to successfully execute the proposed project?

Yes		No	

d) Do the proposed sub-objectives reflect the students' areas of specialization?

Yes		No	

e) Supervisor's Evaluation and Recommendation for the Research topic:

Acceptable: Mark/Select as necessary

Topic Assessment Accepted	
Topic Assessment Accepted with minor changes*	
Topic Assessment to be Resubmitted with major changes*	
Topic Assessment Rejected. Topic must be changed	

* Detailed comments given below

Comments

Staff Member's Name	Signature

*__Important__:

1. According to the comments given by the evaluator, make the necessary modifications and get the approval by the **Evaluator**.
2. If the project topic is rejected, identify a new topic, and request the RP Team for a new topic assessment.