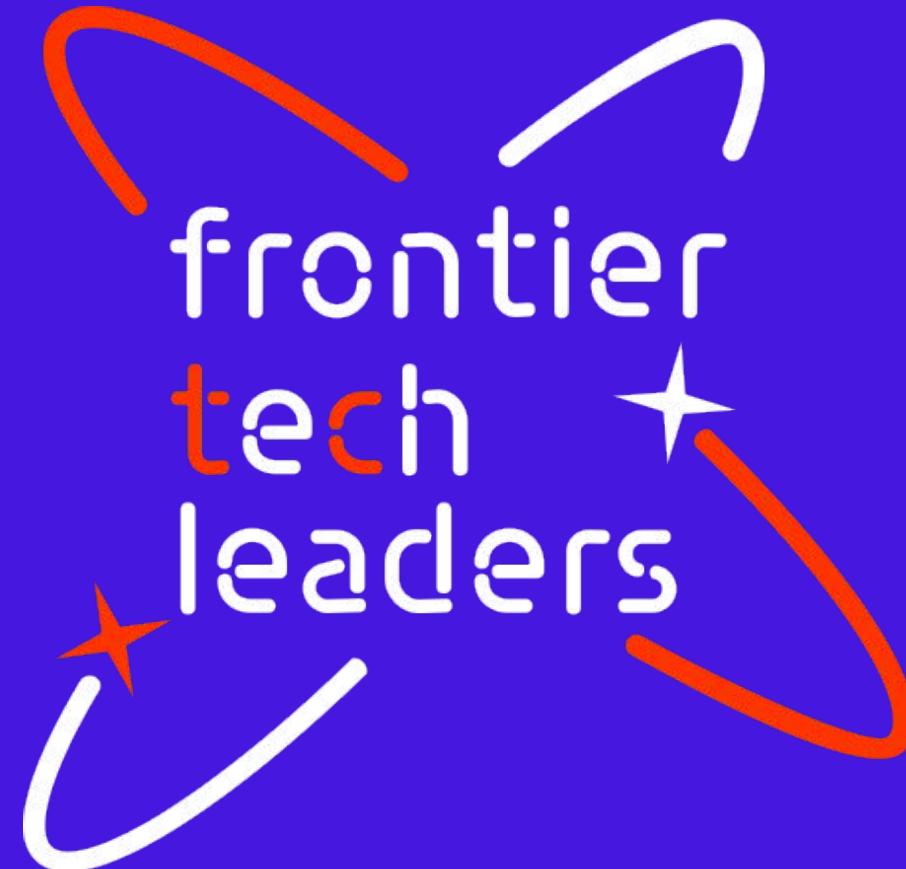


[S D G]
[A I C]
[L A B]



icpso
ISTANBUL INTERNATIONAL CENTER FOR PRIVATE SECTOR IN DEVELOPMENT

UN
DP



A new generation
of **tech**
specialists





[S D G]
[A I O]
[L A B]

UN
TECHNOLOGY
UNITED NATIONS TECHNOLOGY BANK
FOR LEAST DEVELOPED COUNTRIES

icpso
ISTANBUL INTERNATIONAL
CENTER FOR PRIVATE SECTOR
IN DEVELOPMENT

UN
DP

Lecture Companion: AI-Powered Translation and Summarization Tool for Burmese Learners

Team Members:

*Lwin Naing Kyaw
Sai Aike Sam
Nang Hom Phoung
Ingyin Khin
Hom Nan Thawe Htun*



Outline

- **Concept note and implementation plan:**
 - Background
 - Objectives
 - SDG Relation
- **Data**
 - Data Collection
 - Exploratory Data Analysis (EDA) and Feature Engineering
- **Model Selection and Training**
 - Model Evaluation and Hyperparameter Tuning
 - Model Refinement and Testing
- **Results**
- **Deployment**
- **Conclusion & Future Work**

Background

Educational Challenge in Myanmar

- Higher education in Myanmar depends heavily on English textbooks and lectures.
- Limited English proficiency, creating a significant comprehension barrier and widening learning inequality.

Language and Technical Barriers

- Morphological complexity, and segmentation challenges.
- Existing translation tools perform poorly on academic and technical Burmese content.

33 Consonants:

က ခ ဗ ဣ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

န ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

12 Vowels:

ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ ဗ

4 Medials:

ဗ ဗ ဗ ဗ

Myanmar Digits:

၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉

Pali:

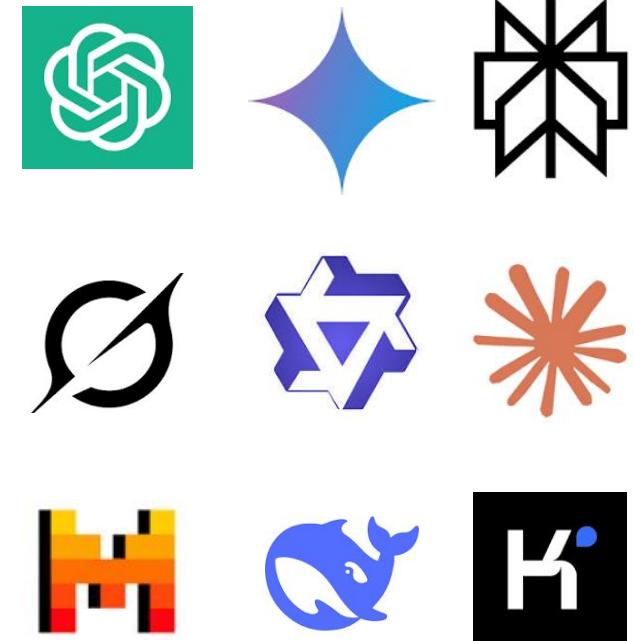
၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉

Background



Emerging Opportunity

- Advances in large language models now enable more accurate transcription, translation, and summarization for low-resource languages.
- These technologies pave the way for automated, accessible lecture support for Myanmar learners.



Objectives



- **Develop an End-to-End Pipeline:** Create a complete system that automatically transcribe, translate, and summarize English lectures into Burmese.
- **Enhance Accessibility:** Provide B1-level simplified summaries to make complex STEM topics understandable for intermediate English learners.
- **Enable Active Learning:** Implement a Retrieval-Augmented Generation (RAG) Q&A system allowing students to ask questions in Burmese and get grounded answers.
- **Vocab building:** Build a CEFR classifier to predict vocabulary's difficulty and provide learners with the meaning in their local language.

SDG Relation



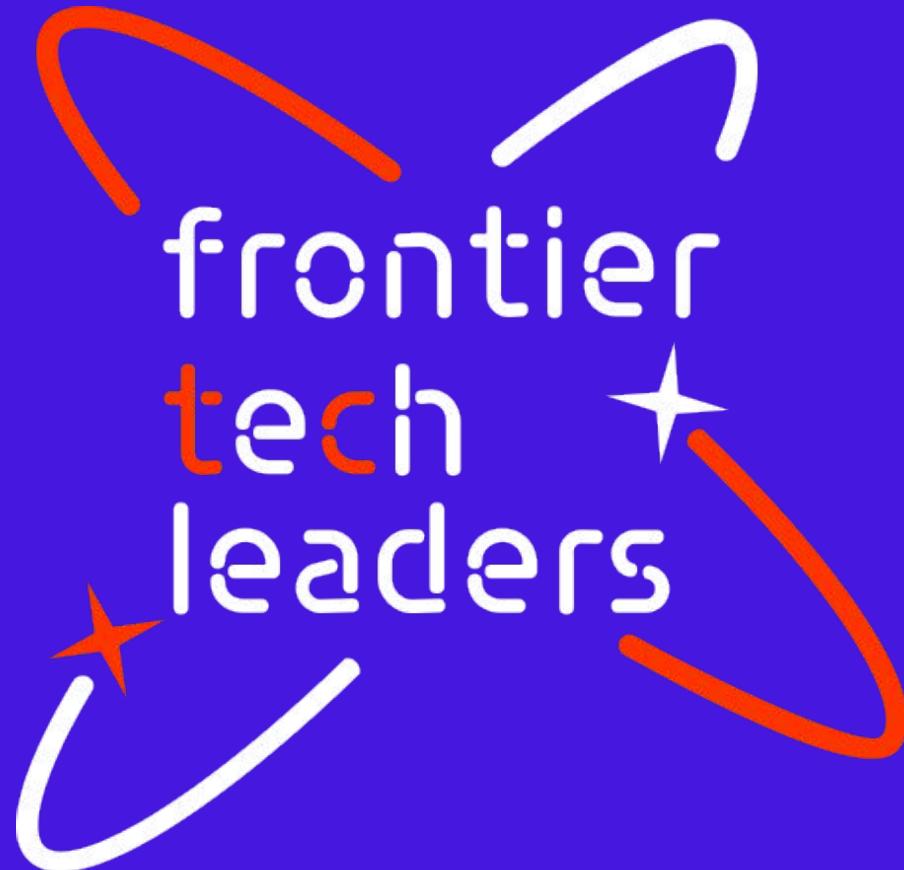
- Improve accessibility, inclusivity, and comprehension of higher-education materials for Burmese learners through translation, simplification, and AI-assisted learning tools.



- Reduces educational inequality by empowering students from rural or low-proficiency backgrounds to access the same content as English-fluent peers.



- Supports digital innovation by adapting localized AI infrastructure (ASR, NMT, RAG) for a low-resource language setting.



Data



Data Overview



Data Sources

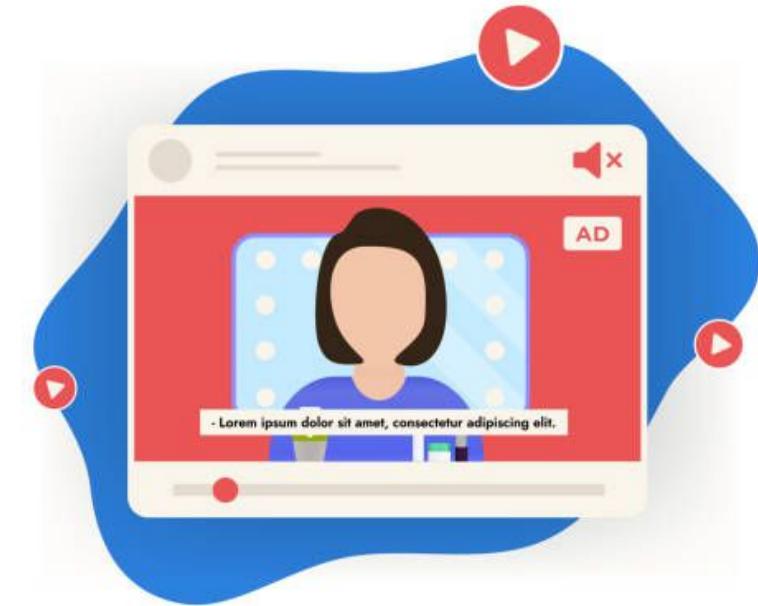
- **Primary:** YouTube Lectures (e.g., Andrew Ng's Machine Learning course).
- **Supplementary:** Transcripts from MOOC platforms (Coursera, edX) and podcasts.

Data Format

- **Video Upload:** primarily youtube url links
- **Transcript Upload:** pdf, txt, srt, vtt

Dataset Characteristics

- Diverse accents and speaking rates (mean ~148 WPM).
- High density of technical STEM vocabulary and academic sentence structures.



Data Collection & Processing



Preprocessing Pipeline

- **Step 1 (Ingestion):** Automated fetch of official captions via *YouTubeTranscriptApi* (prioritizing manual captions over auto-generated ones).
- **Step 2 (Normalization):** Text is normalized to remove timestamps and non-speech artifacts before processing.

Handling Missing Data (The "Fallback" Strategy)

- **Challenge:** Many educational videos lack official subtitles (Missing Data).
- **Solution:** Implemented a **Robust Fallback Mechanism**:
 - * If captions are missing → Download audio via *yt-dlp*.
 - * Process audio with **faster-whisper (ASR)** to generate a fresh transcript.
- **Result:** Ensures 100% data availability even for uncaptioned lectures.

Exploratory Data Analysis (EDA) & Feature Engineering



EDA Findings

- **Segment Suitability:** Text chunking strategy (500 chars) yields segments of 8-18 seconds, ideal for LLM processing.
- **Topical Cohesion:** High cosine similarity (0.7-0.8) between adjacent segments confirms retrieved context will be coherent.

Feature Engineering (for CEFR Classifier)

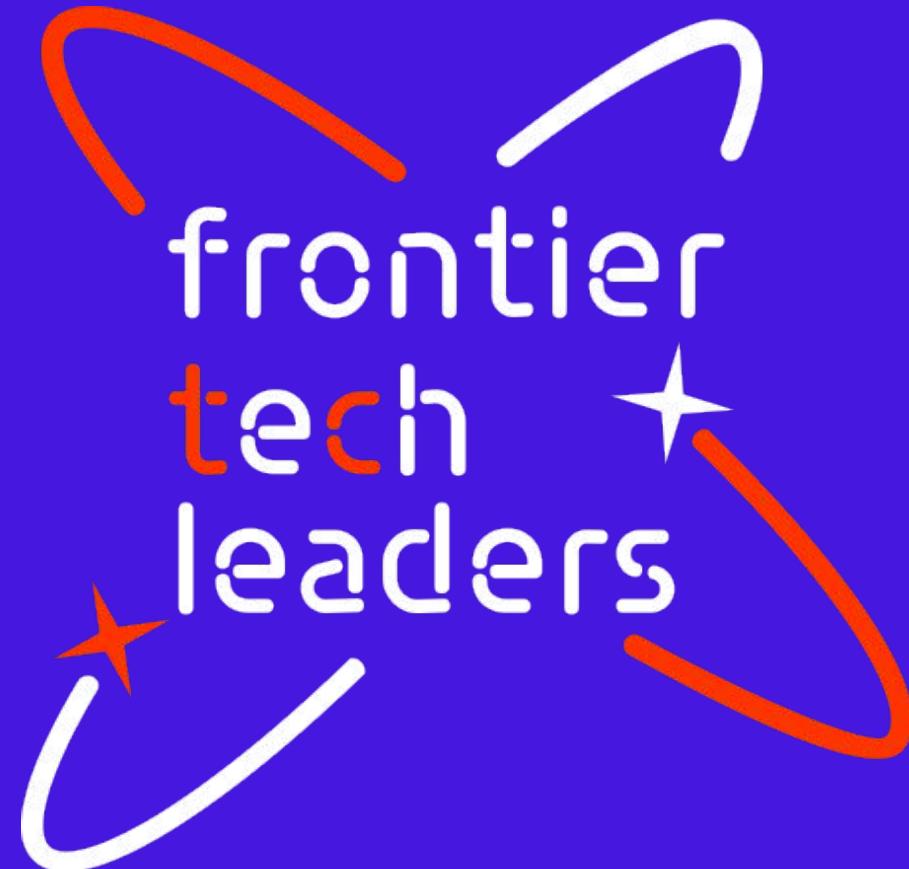
- **Feature:** Character N-Grams (2 to 5 chars).
- **Rationale:** N-grams capture morphological patterns (suffixes/prefixes) better than whole words for determining difficulty in unseen vocabulary.

[S D G]
[A I O]
[L A B]

UN
TECHNOLOGY
UNITED NATIONS TECHNOLOGY BANK
FOR LEAST DEVELOPED COUNTRIES

icpso
ISTANBUL INTERNATIONAL
CENTER FOR PRIVATE SECTOR
IN DEVELOPMENT

UN
DP



Machine Learning Model



Model Selection and Training/Testing



Pipeline 1: Core Processing

- **ASR:** *faster-whisper* (Robust to accents, 4x faster than original Whisper).
- **Translation/Summarization:** *Gemini 2.5 Flash* (High speed, long context window, strong instruction following for simplification).
 - Fallback to Mistral open source llm model: *mistralai/Mistral-7B-Instruct-v0.2*

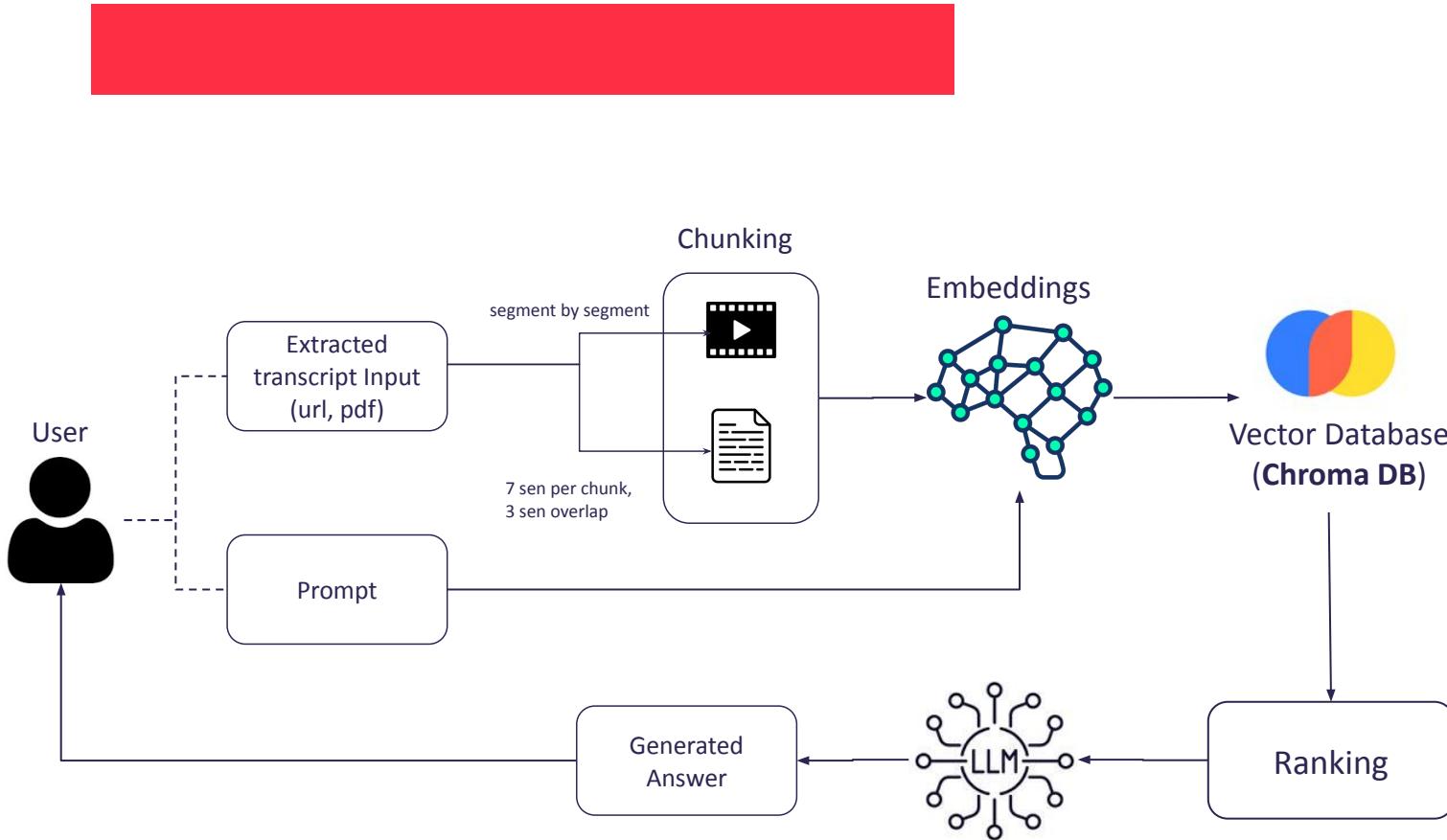
Pipeline 2: RAG System

- **Embeddings:** *paraphrase-multilingual-MiniLM-L12-v2* (Maps English and Burmese to shared semantic space).
- **Retrieval:** *Chroma Database* (Metadata storage: *start_time, end_time, source_type*)

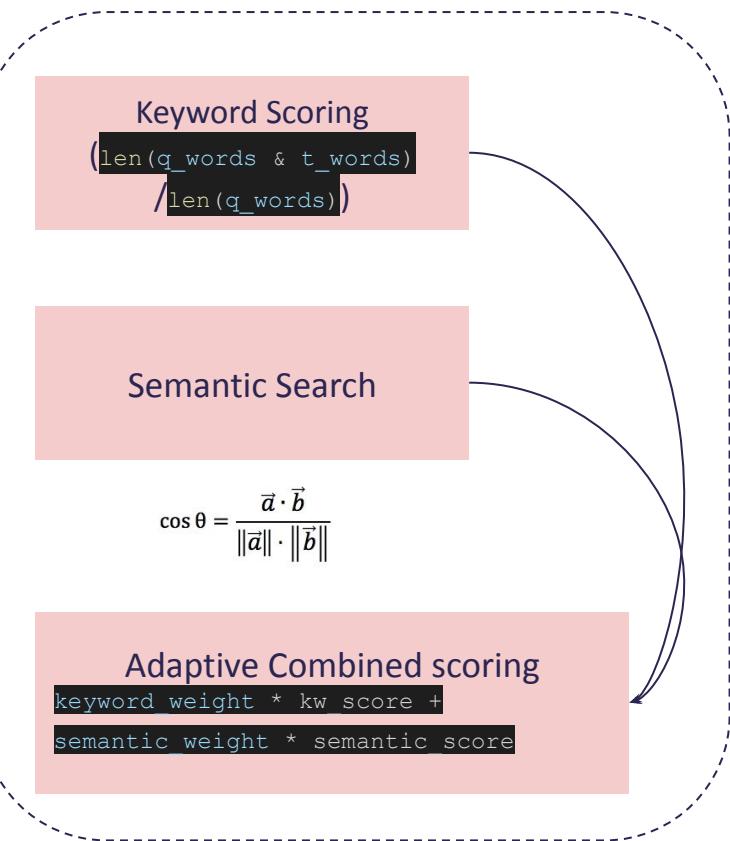
Pipeline 3: CEFR Classifier

- **Model:** *SGDClassifier* (Logistic Regression).
- **Selection:** Outperformed Naive Bayes and LinearSVC in GridSearch with an F1-weighted score of ~0.33.

RAG Architecture



Two-stage ranking system



Classifier: Model Comparison

Model	F1 Macro (CV)	F1 Weighted (CV)	Accuracy (CV)	Notes
SGD_Logistic	0.3188	0.3257	0.3269	Selected best model
SGD_Hinge	0.3179	0.3213	0.3199	Similar but less stable
ComplementNB	0.3048	0.3154	0.3217	Strong baseline; struggles on high-level CEFR
LogReg_SAGA	0.2783	0.3027	0.3189	Underfits
LinearSVC	0.2710	0.2927	0.3427	Uneven performance across classes, likely due to margin-based optimization being sensitive to class imbalance.

Evaluation & Hyperparameter Tuning



Evaluation Strategy

- **Translation:** Back-translation BLEU score (~15.69) and Cosine Similarity (~0.352) confirm semantic preservation despite language differences.
- **RAG System:** Grounding scores (0.42–0.52) indicate answers are factually supported by the lecture context.

Hyperparameter Tuning

- Focused on the CEFR Classifier using *GridSearchCV*.
- **Best Params:** *n-gram range (2,5), alpha 1e-05*.
- **Insight:** While simple, the n-gram approach has limits; future work requires contextual embeddings (BERT/RoBERTa).

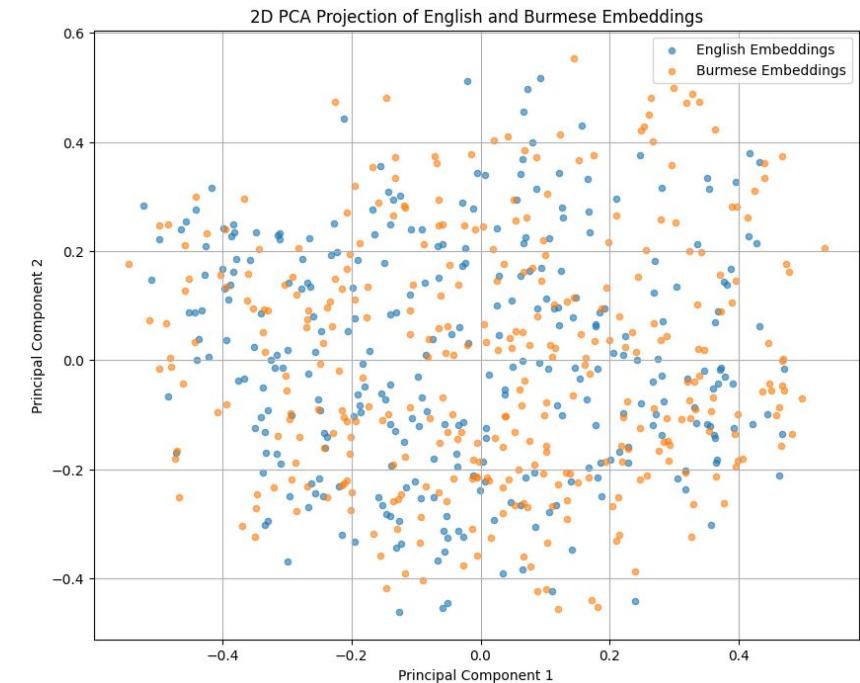
Model Refinement & Testing

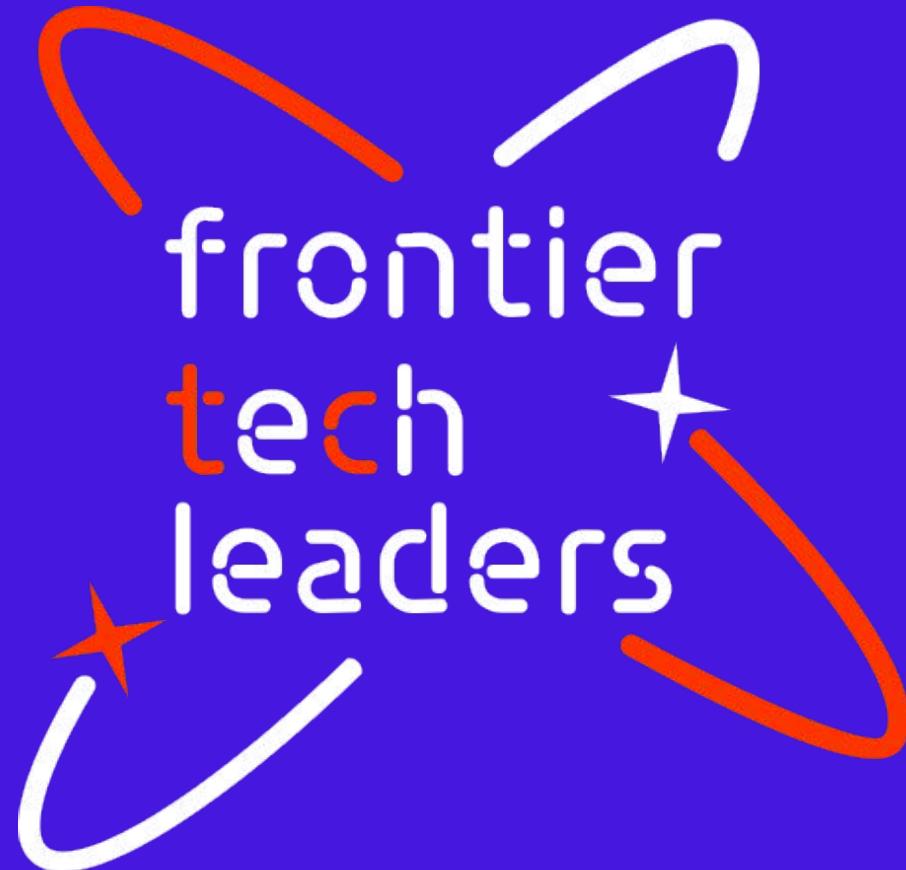
Refinement:

- **Prompt Engineering:** Tuned Gemini prompts to prioritize "B1-level" output and maintain technical terms (e.g., "Linear Regression").
- **RAG Optimization:** Fixed retrieval to top-6 chunks to balance context window vs. noise.

Testing Results:

- End-to-end pipeline successfully processed unseen lectures.
- Visualizations (PCA Plot) show English and Burmese embeddings occupy shared semantic space, validating the cross-lingual search capability.





Result



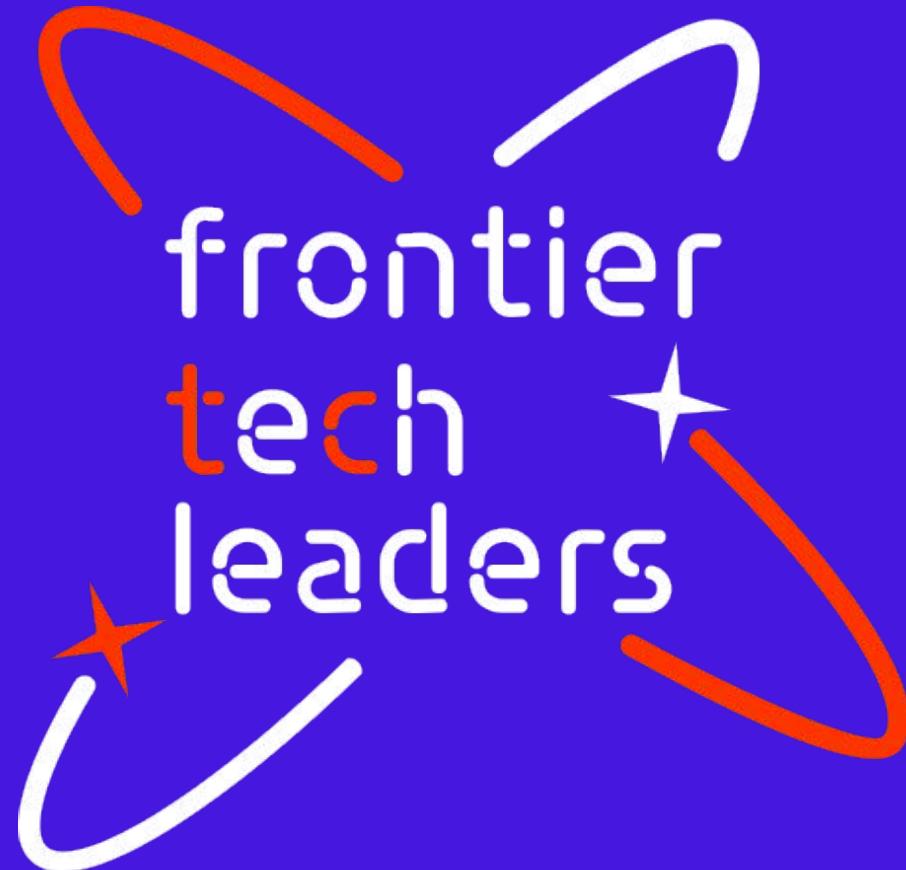
Evaluation Results

Functional Prototype Achieved:

- Successfully built an end-to-end pipeline processing YouTube URLs into:
 - **Bilingual Transcript:** English + Burmese side-by-side.
 - **Simplified Summary:** B1-level Burmese bullet points.
 - **Vocabulary List:** Interactive CEFR-graded glossary.
 - **Interactive Q&A:** RAG system answering queries in Burmese.

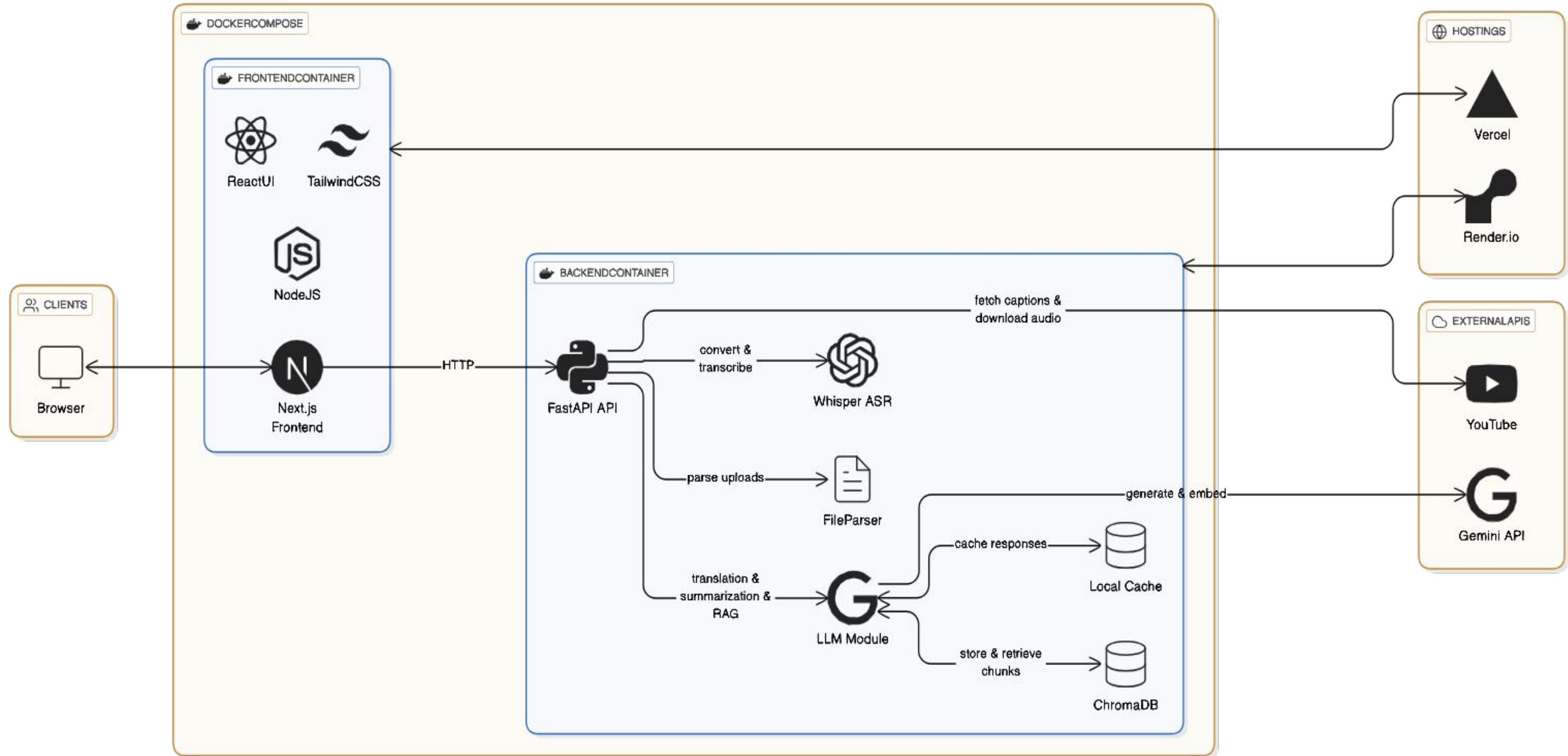
Performance Metrics:

- **Translation & Simplification:** Maintained **~0.352** Cosine Similarity, ensuring semantic meaning is preserved despite simplification.
- **RAG System Reliability:** Achieved **90%+ Self-Consistency** and **0.42–0.52 Grounding Scores**, proving answers are factually supported by the video.
- **Safety:** System successfully refuses off-topic questions (Refusal Score = 1.0 where appropriate).



Deployment





Conclusion & Future Work

- AI powered solution that can reliably translate, simplify, and retrieve English lecture content for Burmese learners.
- Establishes a strong foundation for improving equitable access to higher education for students with varying background.

Future Work

- Scalable, production-ready solution
- Making the solution available for other domains such as economic and healthcare.
- Adding in human evaluation loop to ensure translation integrity of Burmese translations and annotations.
- Working towards solution's offline availability.

Deployment

- *Demo link:*
<https://ftlgr3.vercel.app/>

Demo Video



Thank
you!

