International Journal of Artificial Intelligence and Digital Innovation
https://journals.hu.edu.so/index.php/ijaidi

Cross-Lingual Lecture Note Generation for a Low-Resource Language: Design and User Evaluation of BaroPlatform AI for Somali-Speaking Students

Abdirahim Mohamud Iftin[1], Mohamed Abdi Mire[1], Faaiza Daahir Dhaqane[1], Hafso Hersi Omar[1], Yusuf Abas Mohamed[2]*

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Hormuud University, Mogadishu, Somalia
[2]Department of Data Science, Faculty of Computer Science and Information Technology, Hormuud University, Mogadishu, Somalia

*Corresponding author's email: yuzerzef2@gmail.com

Abstract
Somali is spoken by an estimated 25 million people yet remains among the least represented languages in speech and language technology. Somali-speaking university students consequently depend on English-medium lecture material while no AI-assisted learning tool produces study material in their language. This paper presents BaroPlatform AI, an end-to-end platform that converts English lecture video into structured Somali study material, and reports its evaluation with university students. The system ingests a YouTube link or an uploaded file, extracts audio with FFmpeg, transcribes it with OpenAI Whisper (whisper-1), and prompts Google Gemini (gemini-2.5-flash-lite, with gemini-2.5-flash as fallback) to generate Somali notes, a summary, key points, quiz items and transcript-grounded chatbot responses; Azure Cognitive Services synthesises Somali speech from the notes. The backend is FastAPI with PostgreSQL; the frontend is Next.js. Gemini was empirically rejected for transcription after it returned RECITATION safety blocks on widely distributed educational audio, an engineering finding reported here in full. Seven students at two Somali universities evaluated the deployed system using a 33-item instrument (Cronbach's $\alpha = 0.816$). Usability scored 4.57/5, note quality 4.02/5 and educational value 4.45/5; every participant agreed without exception that the platform reduced the time needed to prepare study notes ($M = 5.00$, $SD = 0.00$), every participant reported measurable time savings, and every participant indicated they would use the system regularly if deployed. Operational data show that while 66.7% of submitted lectures completed the full pipeline, 100% of lectures that reached the generation stage produced usable Somali notes, locating the reliability constraint in media handling rather than in cross-lingual generation. Among the note-quality items, the naturalness of the generated Somali scored below the construct mean ($M = 3.86$), consistent with the known limits of general-purpose multilingual models on low-resource languages. We argue that prompt-level cross-lingual generation is a viable interim strategy for languages that lack the annotated corpora required for fine-tuning.

Keywords: Somali language, low-resource languages, automatic speech recognition, large language models, cross-lingual generation, educational technology.

1. Introduction

Video lecture material is now a primary channel of academic content, and platforms such as YouTube, Coursera and Khan Academy have made an enormous volume of instruction freely available. The value a learner extracts from that material, however, depends on their fluency in the language it is delivered in. For a student who processes English as a second language, watching a dense lecture while simultaneously constructing notes imposes a compounding cognitive burden.

This burden is documented rather than assumed. Bui and Myerson [1] showed that lecture note-taking draws directly on working memory capacity, so that students with less available capacity produce poorer notes and recall less of the lecture. When the lecture is delivered in a second language, part of that same capacity is consumed by comprehension and translation before any note can be written. Canham et al. [2] examined this directly in bilingual students and found that the language chosen for note-taking measurably affects both note quality and memory for newly learnt content. Macaro et al. [3], reviewing English-medium instruction across higher education internationally, report persistent evidence that students following instruction in a second language face comprehension difficulties that institutional policy has generally failed to address.

This is the ordinary condition of university study in Somalia. English predominates as the medium of instruction in many Somali universities, while Somali remains the language in which a large proportion of students think and reason most fluently. The gap is not merely a matter of preference; on the evidence above it is a measurable constraint on how efficiently knowledge can be extracted from the material students are given.

Language technology has not closed this gap, and the reasons are structural. Joshi et al. [4] classified the world's languages by their representation in natural language processing resources and found that the overwhelming majority of languages, Somali among them, sit in categories with negligible labelled data and correspondingly little research attention, a distribution they argue is self-reinforcing. Blasi et al. [5] quantified the downstream consequence, demonstrating systematic inequality in the performance of language technologies across languages that tracks resource availability rather than linguistic difficulty. Automatic speech recognition has advanced substantially through transformer architectures and large-scale weakly supervised training [6], but those gains have accrued unevenly: surveys of ASR for African languages report scarce annotated speech, morphological complexity, dialectal variation and limited computing infrastructure as recurring constraints [7].

Somali has received some dedicated attention. Biswas et al. [8] demonstrated that semi-supervised acoustic and language model training measurably improves Somali recognition. That work targets the recognition layer alone, however. No system builds upward from recognition to the pedagogical artefacts a student actually studies from.

Nor do commercial tools fill the space. Otter.ai, Notta and Fireflies.ai combine recording, transcription and summarisation competently, and all target English-speaking business users. None generates Somali. Somali-speaking students sit outside the design intent of every widely available AI study tool — an instance of the general pattern Kasneci et al. [9] identify in their review of large language models in education, where capability concentrates in English and a small set of high-resource languages while other communities benefit correspondingly less.

This paper presents BaroPlatform AI, a deployed platform that closes the loop from lecture video to Somali study material, and reports what happened when Somali university students used it. Our contributions are four:

- An end-to-end architecture that takes a lecture video and returns structured Somali notes, a summary, key points, synthesised Somali audio, a transcript-grounded chatbot

and generated quiz items, deployed and publicly accessible rather than demonstrated in a laboratory.

- Empirical evidence that prompt-level cross-lingual generation with a general-purpose multilingual model produces educationally usable Somali without any language-specific fine-tuning — a practical route for languages whose annotated corpora do not yet support the alternative.
- A documented negative engineering result: Google Gemini proved unusable for transcription in this setting because it returns RECITATION safety blocks on widely distributed educational audio, a failure mode that is undocumented, silent, and directly consequential for anyone building lecture-processing systems.
- A user evaluation with Somali-speaking students at two institutions, reported with per-item statistics, an internal-consistency coefficient, and an explicit account of what a sample of this size can and cannot establish.

Section 2 situates the work in the literature. Section 3 describes the methodology, covering both the system and its evaluation. Section 4 reports results and analysis. Section 5 discusses them and states the limitations, and Section 6 concludes.

2. Related Work

2.1. Speech Recognition for Low-Resource and African Languages

Whisper [6] established that large-scale weak supervision across 680,000 hours of multilingual audio yields robust zero-shot transcription without task-specific fine-tuning. Its multilingual coverage is the property that matters here: it removes the requirement for a Somali-specific acoustic model as a precondition for building anything downstream.

That coverage is nonetheless uneven, and the unevenness is structural rather than incidental. Joshi et al. [4] show that resource availability across languages follows a steep and self-reinforcing distribution, and Blasi et al. [5] demonstrate that measured system performance tracks that distribution closely. Imam et al. [7] document the specific consequences for African languages: English corpora such as LibriSpeech run beyond 1,000 hours of labelled audio while comparable Somali resources do not exist, and morphological and dialectal complexity compound the scarcity. Pillai et al. [10] show that multistage fine-tuning materially improves low-resource recognition, which establishes the ceiling that an unfine-tuned pipeline such as ours accepts in exchange for requiring no training data at all.

Somali specifically has been addressed at the model level. Biswas et al. [8] applied semi-supervised acoustic and language model training to Somali and reported measurable improvement over the supervised baseline. We note this deliberately: claims that no Somali speech technology exists are inaccurate. What has not been built is the layer above: a system converting recognised speech into structured pedagogical material in Somali.

2.2. Summarisation and Note Generation with Large Language Models

Basyal and Sanghvi [11] compared several large language models on summarisation and found abstractive approaches, in which the model rewrites rather than extracts, consistently more readable and more information-dense than extractive baselines. This is the capability BaroPlatform AI depends on: generating Somali notes from an English transcript is an abstractive rewrite across a language boundary, not an extraction.

Saini et al. [12] proposed a multilingual framework for note-taking and content-based analysis of lectures, establishing feasibility for the general task across languages. Xie et al. [13] went further and measured outcomes, showing that an LLM-supported lecture summarisation system improved both knowledge recall and student satisfaction. Xie et al. supply the outcome-

level evidence that our own design is not powered to produce, and we return to this in Section 5.

For the assessment components, Bulathwela et al. [14] demonstrated scalable generation of educational questions with pre-trained language models, and Adigwe and Yuan [15] showed that few-shot prompting produces contextually appropriate teacher responses in student-teacher dialogue. Our quiz and chatbot modules are applications of both lines of work, extended across a language boundary they were not developed for.

2.3. Multilingual Models in Education

Kasneci et al. [9] reviewed the opportunities and risks of large language models in education and identified the asymmetry described above: these systems are built for English and a small set of high-resource languages. Alam et al. [16] examine LLM behaviour in multilingual and dialectal settings and document degraded performance as resource availability falls. The Gemini model family [17], [18] was trained jointly across text, image, audio and video and reports strong multilingual benchmark performance; whether that translates into pedagogically adequate Somali was, before this study, an open empirical question.

2.4. The Research Gap

The literature therefore establishes three things. Multilingual ASR is good enough to transcribe lecture audio without a language-specific model [6], [7]. Large language models summarise lectures well enough to improve measurable learning outcomes [13]. Somali speech recognition has been demonstrated at the model level [8].

What no prior work provides is the composition of these capabilities into a system that a Somali-speaking student can use, and evidence about what happens when they do. Specifically, we found no prior system that (i) accepts arbitrary lecture video and returns structured study material in Somali, (ii) does so without any Somali-specific model training, or (iii) has been evaluated with Somali-speaking students. This paper addresses all three.

Table 1. Feature comparison with widely available lecture and meeting transcription tools

Capability	**BaroPlatform AI**	**Otter.ai**	**Notta**	**Fireflies.ai**
Somali output	Yes — full notes	No	No	No
YouTube URL input	Yes	No	No	No
File upload	Yes	Limited	Yes	Yes
Structured study notes	Yes	Summary only	Summary only	Summary only
Key point extraction	Yes (Somali)	Yes (English)	Yes (English)	Yes (English)
Lecture chatbot	Yes (Somali)	No	No	Yes (English)
Quiz generation	Yes	No	No	No
Note audio playback	Yes (Somali TTS)	No	No	No
Primary audience	University students	Professionals	Enterprise	Enterprise

3. Methodology: System Design and Implementation

This section sets out how the study was carried out, in two parts. The first describes the system: its three-layer architecture, the processing pipeline that carries a lecture video from submission through to Somali study material, the model-selection decisions that shaped that pipeline — including one rejection we report as a finding in its own right — and the mechanism by which cross-lingual generation is performed. The second describes how the deployed system was evaluated: the research design, the participants and how they were selected, the questionnaire instrument, the procedure followed and the statistical treatment applied. Taken together, the two parts are intended to be sufficient for the system to be reconstructed and the evaluation repeated.

3.1. Architecture

The platform is organised in three layers. A Next.js and React frontend handles lecture submission, processing status and the learning dashboard. A FastAPI backend orchestrates the pipeline, executes long-running work through background tasks so media processing never blocks the request-response cycle, and enforces JWT-based role separation between student and administrator. PostgreSQL provides persistence through SQLAlchemy, with schema versioning through Alembic. The system is deployed and publicly accessible at baro-platform.elivateict.com.

3.2. Processing Pipeline

A submission enters as either a YouTube URL or an uploaded media file. For URLs, yt-dlp retrieves metadata and audio; uploads are written directly to the server. In both paths FFmpeg normalises the audio to mono WAV and segments long recordings to remain inside API request limits. The normalised audio goes to Whisper for transcription. The resulting English transcript is passed to Gemini under structured prompts requesting a Somali summary, key concepts, detailed explanatory notes and a key-point list, together with quiz items. Azure Cognitive Services then synthesises Somali speech from the generated notes using native Somali neural voices. Outputs are written to PostgreSQL and surfaced through the dashboard.

Table 2. Pipeline components and version identifiers

Stage	Component	Function
Media retrieval	yt-dlp	Fetches YouTube metadata and audio streams
Audio processing	FFmpeg	Normalises to mono WAV; segments long audio
Transcription	OpenAI Whisper (whisper-1)	English transcript with timestamped segments
Generation	Google Gemini 2.5-flash-lite (fallback: 2.5-flash)	Somali notes, summary, key points, quiz, chatbot
Speech synthesis	Azure Cognitive Services (so-SO-UbaxNeural, so-SO-MuuseNeural)	Somali audio rendering of generated notes
Persistence	PostgreSQL / SQLAlchemy / Alembic	Stores transcripts, notes, jobs, quiz attempts, logs
Orchestration	FastAPI background tasks	Asynchronous execution with per-stage progress

3.3. Transcription Model Selection: A Negative Result

Gemini was our initial candidate for transcription, since using a single provider for both transcription and generation would have simplified the architecture and reduced cost. We rejected it after testing.

On widely distributed educational audio — the exact category this system is built to process — Gemini returned RECITATION safety blocks rather than transcripts. The mechanism guards against reproducing memorised training content, and popular lecture recordings are precisely the material likely to appear in a model's training distribution. The failure is therefore systematic rather than incidental: its probability rises as source material becomes more widely used, which inverts the reliability property a lecture-processing system requires.

Whisper imposes no comparable restriction and transcribed the same material without incident. We report this because the failure is invisible until encountered, is not documented in the model card, and would cost any team building on this pattern the same investigation.

3.4. Cross-Lingual Note Generation

The system performs no discrete translation step. There is no intermediate Somali transcript and no dedicated translation model. Gemini receives the English transcript together with a structured prompt specifying the required Somali output schema, and generates directly into Somali. A quality-assurance layer inspects the returned structure and triggers a repair prompt where output is malformed or incomplete.

This distinction matters for interpreting our results. What is being evaluated is a model's capacity to reason across a language boundary and produce pedagogically organised text in the target language, not the fidelity of a translation pipeline. The corresponding weakness is that generation quality is bounded by the model's Somali competence, which no component of our system can improve.

3.5. Learning Features and Administration

The student workspace presents the transcript with timestamped segments alongside the generated Somali notes, a summary, extracted key points, audio playback, downloadable files, a transcript-grounded chatbot and a five-item multiple-choice quiz with Somali feedback. A Notes Library organises material by AI-assigned topic across lectures.

An administrative portal exposes job queues, per-stage processing state, user accounts, pipeline success rate, average processing time and system logs. We included this because AI pipelines fail in ways invisible to users and to operators without explicit instrumentation: a silent safety block, a rate limit and a malformed generation are indistinguishable from the outside, and all three occurred during development. The operational data reported in Section 4.4 come from this layer.

3.6. Evaluation Design

We conducted a descriptive mixed-methods evaluation of the deployed system, combining Likert-scale ratings with open-ended response. The design is observational: there is no control condition, no random assignment and no pre-test/post-test measurement. It establishes whether the system functions for real users and how they judge it. It does not establish learning-outcome effects, and no such claim is made.

The evaluation is framed by the Technology Acceptance Model [19], which holds that perceived usefulness and perceived ease of use are the principal determinants of whether a technology is adopted. The instrument maps onto both constructs: usability items address perceived ease of use, while note quality and educational value items address perceived usefulness, and the overall-assessment items capture behavioural intention to use. This framing

is what makes the divergence reported in Section 4.3 interpretable rather than merely anomalous.

3.7. Participants and Sampling

Participants were selected purposively against the criteria that define the system's target population: currently enrolled university students who study from English-medium lecture material and who use Somali academically. Selection also sought variation in academic level and language preference, so that responses would not reflect a single stage of study or a single relationship to Somali. The resulting group spans Year 2 to postgraduate level and four distinct language-preference profiles, as Table 3 shows.

Eight responses were collected. One response was submitted by a member of the development team and was excluded from all analysis, since a developer's assessment of their own system is not an independent observation. All seven remaining responses were retained, including the most critical response received. Analysis is reported on n = 7, drawn from two institutions: six participants at Hormuud University and one at Plasma University.

Table 3. Participant profile (n = 7)

Characteristic	Distribution	Observation
Academic level	Year 2: 1 (14%); Year 3: 1 (14%); Year 4: 4 (57%); Masters/PhD: 1 (14%)	Spans undergraduate to postgraduate; majority final-year, the primary target user
Institution	Hormuud University: 6 (86%); Plasma University: 1 (14%)	Two institutions represented
Technology comfort (1–5)	5: 4 (57%); 4: 1 (14%); 3: 2 (29%) M = 4.29	Range present, not uniformly expert
Prior AI tool use	Regularly: 6 (86%); Occasionally: 1 (14%)	Able to compare against alternatives
Study language preference	Somali: 2 (29%); Both equally: 2 (29%); Subject-dependent: 2 (29%); English: 1 (14%)	86% engage academically in Somali; four distinct profiles represented
Prior transcription tool use	No: 3 (43%); Aware but unused: 2 (29%); Yes: 2 (29%)	71% had no practical exposure

3.8. Instrument and Procedure

The instrument comprised 33 items administered through Google Forms in four groups: participant background (6 items), platform usability (7 Likert items, B1–B7), note quality (7 Likert items C1–C7 plus one open-ended item), and educational value (7 Likert items D1, D3–D8, plus one binary item D2), followed by three overall-assessment items and two open-ended items. Likert items used a five-point agreement scale (1 = strongly disagree, 5 = strongly agree). The full instrument appears in Appendix A.

Participants worked with the deployed system directly, submitting lectures of their own choosing and working through the generated transcript, notes, summary, key points, audio, chatbot and quiz. Each participant completed the questionnaire after using the platform. Responses were collected on 1 and 2 July 2026.

Internal consistency across the 21 Likert items was acceptable, with Cronbach's α = 0.816 [20]. Subscale coefficients were α = 0.679 (usability), α = 0.753 (note quality) and α = 0.604 (educational value); the lower subscale values reflect seven-item scales at n = 7 and are reported as indicative rather than as established reliability.

3.9. Analysis and Ethical Considerations

Likert responses were mapped to integers 1–5 and summarised by mean, standard deviation and the proportion agreeing or strongly agreeing. Categorical items were summarised by frequency. Internal consistency was computed as Cronbach's α. Open-ended responses were read and grouped by recurring theme. Given n = 7, no inferential tests were performed and none are reported; all statistics are descriptive.

Participation was voluntary and participants were informed of the purpose of the evaluation. Responses are reported only in aggregate. The system processes publicly available lecture content and user-uploaded files, applies role-based access control, and transmits no student data to third parties beyond the AI service providers the pipeline requires.

4. Results and Analysis

This section reports the questionnaire results across all three measured constructs, the overall assessment and qualitative responses, and the operational data recorded by the platform during the evaluation period. Analysis is presented alongside the figures rather than deferred, so that each result is interpreted in the context that produced it.

4.1. Questionnaire Results

Table 4 reports all 21 Likert items with means, standard deviations and agreement proportions, grouped by construct.

Table 4. Questionnaire results by construct (n = 7)

Item	Statement	M	SD	% ≥ 4
	Usability (perceived ease of use)			
B1	Easy to navigate without technical knowledge	4.43	0.53	100%
B2	Submitting a link or uploading a file was straightforward	4.43	0.53	100%
B3	Dashboard layout was clear and well organised	4.86	0.38	100%
B4	Transcript, notes and key points easy to locate	4.86	0.38	100%
B5	System responded quickly after submission	4.29	0.49	100%
B6	Progress tracking was informative and helpful	4.71	0.49	100%
B7	Could use the platform independently without guidance	4.43	0.79	86%
	Section mean	**4.57**		
	Note quality			
C1	Notes written in natural, correct Somali	3.86	0.69	71%
C2	Notes accurately captured the lecture's main ideas	3.86	0.69	71%

Item	Statement	M	SD	% ≥ 4
C3	Somali summary was concise and easy to understand	4.00	1.00	86%
C4	Key points were relevant and useful for revision	4.29	0.76	86%
C5	Structured format made notes easy to study from	3.71	0.49	71%
C6	Notes better than manual notes in the same time	4.00	0.82	71%
C7	Quiz questions structured like exam-ready questions	4.43	0.53	100%
	Section mean	**4.02**		
	Educational value (perceived usefulness)			
D1	Reduced the time needed to prepare study notes	5.00	0.00	100%
D2	Transcript alongside notes was useful (binary item)	100% yes	—	100%
D3	Chatbot helped understanding of lecture content	4.57	0.53	100%
D4	Quiz effectively tested understanding	4.57	0.53	100%
D5	Audio playback was a useful learning feature	4.43	0.98	71%
D6	Notes Library helped organise and revisit material	4.29	0.49	100%
D7	Useful for day-to-day university studies	3.71	1.11	57%
D8	Would recommend to other Somali-speaking students	4.57	0.53	100%
	Section mean (Likert items)	**4.45**		

Usability was the strongest construct at 4.57. Dashboard layout and content locatability both scored 4.86 — the highest ratings in the study after D1 — and every usability item drew agreement from every participant with a single exception. Students without technical training could operate the platform unaided, which is a precondition for the system being useful at all. Note quality was the weakest construct at 4.02 and the most informative. The naturalness and grammatical correctness of the generated Somali (M = 3.86) and the accuracy with which notes captured the lecture's main ideas (M = 3.86) both fell below the construct mean. Both drew agreement from 71% of participants, meaning two of seven were neutral or negative on each. The structural organisation of notes scored lowest in the construct (M = 3.71) though no participant rated it below 3, indicating a consistently moderate judgement rather than a divided one. The Somali summary showed the widest dispersion in this construct (SD = 1.00, range 2–5), suggesting that summary quality varies materially across lectures rather than being uniformly adequate. Quiz items were the construct's strongest result (M = 4.43, unanimous agreement).

Educational value averaged 4.45 and contains the study's single unambiguous result. Every participant strongly agreed that the platform reduced the time needed to prepare study notes from a lecture (M = 5.00, SD = 0.00). Unanimity at the scale maximum with zero variance is the strongest signal a sample of this size can produce, and it corresponds directly to the

problem the system was built to solve. All seven participants likewise answered the binary item D2 affirmatively: having the transcript available alongside the Somali notes was useful. The chatbot (M = 4.57) and quiz (M = 4.57) were rated as highly as recommendation intent (M = 4.57), indicating that supplementary features were valued alongside note generation rather than treated as peripheral.

4.2. Overall Assessment

All seven participants reported measurable time savings against studying from raw lecture video, and all seven indicated they would use the platform regularly if it were deployed at their institution. The mean overall rating was 4.29 (SD = 0.49); no participant rated the platform below 4.

Table 5. Overall assessment (n = 7)

Measure	Distribution	Summary
Time saved vs. raw video	< 30 min: 5 (71%); > 1 hour: 2 (29%)	100% reported time savings
Likelihood of regular use	Likely: 5 (71%); Very likely: 2 (29%)	100% positive intent
Overall rating (1–5)	4 — Good: 5 (71%); 5 — Excellent: 2 (29%)	M = 4.29, SD = 0.49

4.3. The Time-Saving and Daily-Use Divergence

One result requires separate treatment because it contradicts another. Item D1 achieved unanimous maximum agreement that the platform reduces note-preparation time. Item D7 — usefulness for day-to-day university studies — scored jointly lowest in the entire instrument at 3.71 (with C5), drew agreement from only 57% of participants, and carries the widest dispersion of any item (SD = 1.11, responses spanning 2 to 5).

Read through the acceptance framing set out in Section 3.6, these two items measure different things. D1 measures perceived usefulness for a bounded task; D7 measures perceived usefulness within an existing workflow. A tool can demonstrably save time on a task while failing to fit the routine in which that task sits, and our data indicate exactly that pattern. Three explanations are consistent with what we observed: processing latency of roughly three minutes per lecture may be acceptable for a trial but not for habitual use; the courses participants actually take may not be well served by the YouTube lecture content the system handles best; or the Somali quality gap registered by C1 may matter more under repeated study than during a single session. Our design cannot distinguish these, and the open-ended responses are too sparse to adjudicate. The divergence is nonetheless the most analytically important finding in the study, and Section V returns to how it could be resolved.

4.4. System Performance

Operational data recorded by the administrative portal during the evaluation period are reported in Table 6.

Table 6. Platform operational metrics, evaluation period (7 days, July 2026)

Metric	Value	Observation
Registered users	20	19 learners, 1 administrator
Lectures submitted	15	14 YouTube, 1 direct upload
Lectures completed	10	Ready for user access

Metric	Value	Observation
Failed lectures	5	Attributed to API concurrency limits
Pipeline success rate	66.7%	End-to-end completion
Notes generated	10	100% note conversion from completed jobs
Mean job time	2 min 43 s	Across completed lectures

The headline reliability figure — 66.7% end-to-end completion — is a prototype-grade result and we do not present it otherwise. Its composition, however, is more informative than the figure itself. Every lecture that reached the generation stage produced usable Somali notes: note conversion was 100% across completed jobs, with no partial or malformed outputs requiring discard. The five failures occurred upstream, in media retrieval and transcription, and were attributed to API concurrency limits rather than to pipeline logic or to generation quality. This separation matters for how the paper's central claim should be read. The novel component of this system is cross-lingual generation, and that component did not fail during the evaluation period. The reliability constraint sits in commodity media handling — a well-understood engineering problem addressable through request queuing and retry policy — rather than in the capability the study set out to test. Mean processing time of 2 minutes 43 seconds across completed lectures is consistent with cloud-based inference over long audio and is the quantity most directly implicated in the daily-use result discussed above.

Usage distribution is also worth noting: 14 of 15 submissions were YouTube links rather than uploads. Students given both options overwhelmingly chose the one requiring no local media, which has implications for where engineering effort should be directed.

4.5. Qualitative Findings

Open-ended responses were sparse but consistent with the quantitative pattern, and three themes recur. Cross-lingual accuracy was named as the distinguishing value. One participant reported that other AI tools "do not give correct and relevant information when I pass a YouTube link", whereas this system did — locating the contribution precisely at the cross-lingual step rather than at transcription or summarization individually. Another described the core value as being able to take a video and "have them translated into my language, Somali". The chatbot was named as the most valuable feature by one participant, on the grounds that help is available on demand.

Somali generation quality was raised directly and unprompted. One participant observed that "some Somali words are not considered standard or refined" — a qualitative statement of exactly what item C1 measures quantitatively, and the only substantive language criticism volunteered.

Two implementation defects were reported: low text contrast against dark backgrounds in dark mode together with a navigation state from which a page could not be exited, and "some strain on the system during use and when signing up". Both are interface and stability issues rather than architectural ones, and both were reported by participants who nonetheless rated the platform 4 overall.

5. Discussion

5.1. Cross-Lingual Generation Without Fine-Tuning

The central question this study addressed empirically is whether a general-purpose multilingual model produces Somali of sufficient quality to study from, given that fine-tuning is precluded by the absence of adequate Somali corpora [7], [8], [4]. The answer is a qualified yes. Note quality averaged 4.02, and the strongest note-related item — quiz question construction at 4.43

with unanimous agreement — indicates that the model's grasp of the content is more reliable than its command of the language it renders that content in.

That asymmetry is the finding. Content fidelity and language quality are dissociable, and they dissociate in a specific direction: participants rated the pedagogical structure of the output higher than its linguistic naturalness. This is consistent with Alam et al. [16], who report degraded surface performance for lower-resource languages while semantic capability persists, and with the broader pattern Blasi et al. [5] quantify across languages. It suggests that the practical constraint on systems of this kind is the target language's surface realisation rather than the model's understanding of the source material.

The operational data strengthen this reading. Generation succeeded on every job that reached it, which means the dissociation is not an artefact of cherry-picked successes: the notes participants rated were the notes the system produced, without exception. The implication for practice is that prompt-level cross-lingual generation is a viable interim strategy for low-resource languages. It is not equivalent to a fine-tuned system and our data do not suggest otherwise, but it is available now, at a total API cost of approximately US$40–50 across the entire development and evaluation period, and it produces material students rate as usable.

5.2. Resolving the Divergence

The contradiction between unanimous agreement on time saving and the weakest result in the instrument on daily usefulness is, in acceptance terms, a gap between perceived usefulness for a task and perceived usefulness in practice [19]. Both readings are supported by our data and our design cannot separate them.

This is precisely the divergence Xie et al. [13] were able to resolve, because their design measured knowledge recall rather than perception. A study measuring whether students who use the system perform better on recall tasks than students studying from raw video, and observing use across weeks rather than in a single session, would distinguish a genuine workflow-fit problem from an artefact of first exposure. We identify this as the single most valuable next study, and note that the instrument reported here is adequate as its perception-side component.

5.3. The Transcription Safety-Block Finding

Our rejection of Gemini for transcription has implications beyond this system. The RECITATION block is a content-provenance guard, and its probability of firing rises with how widely distributed the source audio is. For a lecture-processing system this is an adversarial property: the most pedagogically valuable material — popular, widely viewed lectures — is the most likely to be refused.

Kasneci et al. [9] discuss LLM reliability in education largely in terms of output correctness. This is a different failure class: not incorrect output but categorical refusal, triggered by a property of the input the operator cannot inspect in advance and producing no diagnostic. We report it because architectural separation of transcription from generation is a defensible design response, and because the failure is silent enough that a team could reasonably attribute it to their own code for some time.

5.4. Interpreting a Sample of Seven

Seven participants cannot support claims of generalisability, and none are made. What this sample does support is narrower and still meaningful. A unanimous result with zero variance on D1 is not a marginal effect that a larger sample might reverse in direction, though its magnitude would certainly move. The consistently moderate note-quality ratings, and an internal consistency coefficient of 0.816, indicate that participants discriminated between items rather than responding uniformly positively, which is the failure mode that would otherwise make a small purposive sample uninformative.

The instrument distinguishes a 5.00 from a 3.71 within the same construct and from the same respondents. That internal variation is why the note-quality and daily-use findings are treated here as signal rather than noise.

5.5. Limitations

Sample size and composition. Seven purposively selected participants cannot represent Somali university students generally, and six of the seven attend one institution. Participants were predominantly comfortable with technology and 86% already used AI tools regularly, so they are more capable than a representative student and their usability ratings should be read as an upper bound rather than a typical case. Absence of a control condition. Time savings and educational value are self-reported perceptions with no baseline comparison. No learning outcome was measured. Claims about efficacy cannot be made from this design and are not made.

No automatic quality metrics for Somali. BLEU, ROUGE and comparable measures are not validated for Somali, so note quality could be assessed only by human judgement. Item C1 therefore records a perception of linguistic quality rather than a measurement of it, and the specific error categories underlying that perception remain uncharacterised.

Pipeline reliability. A 66.7% end-to-end success rate is acceptable for a prototype and unacceptable for institutional deployment. The attribution of those failures to API concurrency limits rather than pipeline logic follows from the 100% note-conversion figure but has not been independently verified under controlled load. Single-session exposure. Every judgement in this study was formed in one session. Novelty effects inflate first-exposure ratings, and the divergence between D1 and D7 may partly reflect this. Cost at scale. Total API expenditure of approximately US$40–50 covered development and evaluation. Per-user cost under institutional load is unmeasured, and the economics of deployment are therefore unestablished.

6. Conclusion and Future Work

This paper presented BaroPlatform AI, an end-to-end platform converting English lecture video into structured Somali study material, and reported its evaluation with seven purposively selected students at two Somali universities. Three findings stand. First, prompt-level cross-lingual generation with a general-purpose multilingual model produces Somali study material students rate as usable (note quality 4.02/5) without any Somali-specific model training — a route available now for languages whose corpora cannot yet support fine-tuning, and one that did not fail on a single job during the evaluation period. Second, every participant without exception agreed the platform reduces the time required to prepare study notes, while agreement on its fit for day-to-day study was the weakest result in the instrument; time saving and workflow fit are separate properties and this system has demonstrated only the first. Third, Google Gemini is unsuitable for transcribing widely distributed educational audio because of RECITATION safety blocking, which makes architectural separation of transcription from generation a defensible design choice rather than a redundancy. The limits are equally clear. Seven participants at two institutions in a single session, with no control condition and no outcome measurement, establish feasibility and perceived value — not efficacy.

Four lines of work follow directly. A controlled study measuring knowledge recall between students using the platform and students studying from raw video, following Xie et al. [13], would convert perceived value into measured effect and resolve the divergence reported in Section 4.3. A structured assessment of generated Somali by native-speaker raters independent of the development team, scoring grammatical correctness, terminology appropriateness, fidelity, coverage and usability for revision, would characterise the specific weaknesses that item C1 registers only as a number. Raising end-to-end pipeline reliability substantially above 66.7% — a media-handling problem rather than a generation problem, on

the evidence in Section 4.4 — is a precondition for any institutional deployment. Finally, the architecture is Somali-specific only at the prompt layer; substituting prompts for Afar, Oromo or Tigrinya requires minimal engineering, though whether generation quality holds for those languages is an empirical question this study cannot answer.

Conflict of Interest

The authors confirm that there is no conflict of interest related to this manuscript.

References

[1] D. C. Bui and J. Myerson, "The role of working memory abilities in lecture note-taking," Learning and Individual Differences, vol. 33, pp. 12–22, 2014.

[2] A. Canham, M. Coumel, J. Manolova, and A. M. T. de Bruin, "Note taking in bilingual students: Does using a first or second language influence note-taking quality and memory of newly learnt content?," International Journal of Bilingual Education and Bilingualism, vol. 28, no. 1, 2025, doi: 10.1080/13670050.2024.2390862.

[3] E. Macaro, S. Curle, J. Pun, J. An, and J. Dearden, "A systematic review of English medium instruction in higher education," Language Teaching, vol. 51, no. 1, pp. 36–76, 2018.

[4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in Proc. 58th Annu. Meeting Assoc. Computational Linguistics (ACL), Jul. 2020, pp. 6282–6293.

[5] D. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic inequalities in language technology performance across the world's languages," in Proc. 60th Annu. Meeting Assoc. Computational Linguistics (ACL), Dublin, Ireland, May 2022, pp. 5486–5505.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Machine Learning (ICML), Honolulu, HI, USA, Jul. 2023, pp. 28492–28518.

[7] S. H. Imam, B. Sani, D. K. Gete, B. Y. Ahmed, I. S. Ahmad, I. Abdulmumin, S. M. Yimam, M. Y. Bello, and S. H. Muhammad, "Automatic speech recognition for African low-resource languages: Challenges and future directions," in Proc. 6th Workshop African Natural Language Processing (AfricaNLP), Assoc. Computational Linguistics, 2025.

[8] A. Biswas, R. Menon, E. van der Westhuizen, and T. Niesler, "Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training," in Proc. Interspeech 2019, Graz, Austria, Sep. 2019, pp. 3008–3012.

[9] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, and E. Hüllermeier, "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, art. 102274, 2023.

[10] L. G. Pillai, K. Manohar, B. K. Raju, and E. Sherly, "Multistage fine-tuning strategies for automatic speech recognition in low-resource languages," arXiv preprint arXiv:2411.04573, 2024.

[11] L. Basyal and M. Sanghvi, "Text summarization using large language models: A comparative study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT models," arXiv preprint arXiv:2310.10449, 2023.

[12] M. Saini, V. Arora, M. Singh, J. Singh, and S. O. Adebayo, "Artificial intelligence inspired multilanguage framework for note-taking and qualitative content-based analysis of lectures," Education and Information Technologies, vol. 28, no. 1, pp. 1141–1163, 2023.

[13] T. Xie, Y. Kuang, Y. Tang, J. Liao, and Y. Yang, "Using LLM-supported lecture summarization system to improve knowledge recall and student satisfaction," Expert Systems with Applications, vol. 269, art. 126371, 2025.

[14] S. Bulathwela, H. Muse, and E. Yilmaz, "Scalable educational question generation with pre-trained language models," arXiv preprint arXiv:2305.07871, 2023.
[15] A. O. I. Adigwe and Z. Yuan, "The ADAIO system at the BEA-2023 shared task on generating AI teacher responses in educational dialogues," in Proc. 18th Workshop Innovative Use of NLP for Building Educational Applications (BEA), Assoc. Computational Linguistics, 2023.
[16] F. Alam, S. A. Chowdhury, S. Boughorbel, and M. Hasanain, "LLMs for low resource languages in multilingual, multimodal and dialectal settings," in Proc. 18th Conf. European Chapter Assoc. Computational Linguistics (EACL), St. Julian's, Malta, Mar. 2024, pp. 27–33.
[17] Gemini Team, Google, "Gemini: A family of highly capable multimodal models," arXiv preprint arXiv:2312.11805, 2023.
[18] Gemini Team, Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," arXiv preprint arXiv:2403.05530, 2024.
[19] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, vol. 13, no. 3, pp. 319–340, 1989.
[20] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, vol. 16, no. 3, pp. 297–334, 1951.

Appendix A. Evaluation Instrument

Thirty-three items administered via Google Forms. Likert items use a five-point agreement scale (1 = strongly disagree, 5 = strongly agree).

Group	Items
Background	Full name; institution; academic level; technology comfort (1–5); prior AI tool use; study language preference; prior transcription tool use
B — Usability	B1 Easy to navigate without technical knowledge · B2 Submitting a link or file was straightforward · B3 Dashboard layout clear and well organised · B4 Transcript, notes and key points easy to locate · B5 System responded quickly after submission · B6 Progress tracking informative and helpful · B7 Could use independently without guidance
C — Note quality	C1 Notes written in natural, correct Somali · C2 Notes accurately captured main ideas · C3 Somali summary concise and easy to understand · C4 Key points relevant and useful for revision · C5 Structured format easy to study from · C6 Better than manual notes in the same time · C7 Quiz questions structured like exam-ready questions · Open: any errors, missing content or language problems?
D — Educational value	D1 Reduced time needed to prepare study notes · D2 Transcript alongside notes was useful (binary) · D3 Chatbot helped understanding · D4 Quiz tested understanding effectively · D5 Audio playback useful · D6 Notes Library helped organise and revisit · D7 Useful for day-to-day university studies · D8 Would recommend to other Somali-speaking students
Overall	Most valuable feature and why (open) · Time saved vs. raw lecture video · Likelihood of regular use if deployed · Overall rating (1–5) · Final suggestions or comments (open)