WHY LARGE LANGUAGE MODELS FAIL FOR HAUSA EDUCATIONAL CONTENT: CASCADING ERRORS FROM TRANSLATION TO SPEECH TO COMPREHENSION

Honour-Jesus Bezaleel[1,2] **Pearse Jim**[2,3] **Moses Daudu**[3]
[1]Machine Learning Collective (MLC)
[2]Machine Learning Collective (MLC)
[3]UNICBT
{bhonourjesus, pearsejim01}@gmail.com, moses@unicbt.com

ABSTRACT

This research investigates the limitations of large language models (LLMs) in correcting errors from machine-translated text, transcribed speech, and in answering educational questions written in Hausa and translated from English. Translating and processing educational examination text introduces persistent difficulties, including data scarcity, linguistic complexity, translation errors, and lack of domain grounding. We investigate the performance of multiple LLMs on a newly curated dataset of West African Senior School Certificate Examination (WAEC) past questions, focusing on their ability to (1) correct errors in machine-translated and speech-synthesized Hausa text and (2) answer multiple-choice exam questions derived from translated content. We evaluated models including Llama-3.2-1B-Instruct, Gemma-2-2B-IT, N-ATLaS, and HausaLLaMA, and assessed LLM cascading for error correction using larger models such as Llama-3.3-70B, Mixtral-8×7B, Gemini 2.0 Flash, and Flan-T5. Our findings reveal substantial performance gaps across all models. We argue that effective solutions require domain-specific fine-tuning and close collaboration with educators and native speakers in the creation of educational text and audio. This study provides insights into the real-world limitations of LLM-driven educational systems and suggests pathways toward more inclusive and reliable educational technology for underrepresented languages. Resolving these constraints will support inclusive educational knowledge dissemination.

1 INTRODUCTION

The West African Senior School Certificate Examination (WASSCE), a key annual assessment taken by more than 1.5 million school candidates in member countries of WAEC (including Nigeria, Ghana, Liberia, Sierra Leone, and The Gambia) West African Examinations Council (2025), will transition to computer-based testing (CBT) in 2026. Without intervention, this transition risks deepening existing divides for marginalized students. In Nigeria's southern regions, pass rates approach 80%, while northern states like Zamfara report pass rates as low as 9%, compounded by limited digital literacy (6–9% basic proficiency) and English-only test delivery. These inequities persist despite Hausa, a Chadic language, serving as the dominant language and lingua franca for 120 million native speakers plus 80 million second-language users across northern Nigeria, southern Niger, and West Africa Newman (2025). Translating exam past questions is crucial for bridging knowledge gaps and disseminating content across linguistic and cultural boundaries, while audio transcription enables students to hear questions in their native language, enhancing exam preparation and performance. This study focuses on developing a TTS-ASR system based on machine-translated text of WASSCE past questions. We hypothesise that LLM cascading can correct errors in this pipeline for Hausa speech, overcoming data scarcity through intelligent error correction. Our results reveal that LLMs struggle with cascading errors in the speech data and fail to answer questions correctly, inflating word error rate (WER), character error rate (CER), and diacritic error rate (DER). These errors

stem from untranslated terms, morphological shifts in sentence meaning, misinterpreted questions, and mixed-up or omitted multiple-choice options, highlighting low-resource NLP limitations. This severely impacts audio transcription accuracy and would affect student performance on the interface we are developing. Our contributions are as follows:

- **Dataset Creation & Curation:** We present **HausaPQ-Speech**, a multimodal corpus of 4,703 West African Senior School Certificate Examination (WASSCE) past questions, featuring English source text, multiple Hausa MT outputs, and synthesized speech of questions and answers.
- **Comprehensive Negative Evaluation:** We conduct a full pipeline evaluation, revealing systematic failures in MT (high omissions), ASR (catastrophic tonal error), LLM-based error correction (minimal gain), and LLM question answering (low accuracy).
- **Diagnostic Insights:** We identify that upstream MT artifacts corrupt ASR evaluation, that ASR models are fundamentally ill-equipped for tonal languages, and that model specialisation does not guarantee better performance.

2 Prior Work

Prior work in Hausa automatic speech recognition (ASR) has largely focused on benchmarking acoustic models using clean, human-authored transcripts from general-domain corpora. Specialised efforts such as Hausa-ASR (NCAIR1) and whisper-large-hausa (mosesdaudu) represent important advances, but evaluations are typically conducted on clean, read speech from datasets like Mozilla Common Voice, yielding deceptively low word error rates (WERs) (Abubakar et al., 2024). For example, multilingual models such as Whisper-large report WERs as low as 4.23% on Common Voice Hausa, suggesting strong ASR performance in low-resource settings. However, these evaluations assume linguistically well-formed reference transcripts and do not account for systematic structural errors. African NLP research has also produced multilingual models such as Afri-mT5 (Adelani et al., 2022; Abdulmumin et al., 2022) and benchmarks including Toucan (Elmadany et al., 2024) and FLORES-200, which provide standardized evaluation for Hausa and other low-resource languages. While these resources advance general-domain machine translation (MT), their applicability to highly structured, domain-specific content—such as educational assessments—remains under-explored. Earlier African ASR efforts similarly emphasise vocabulary coverage and acoustic diversity under controlled conditions (Schlippe et al., 2012), reinforcing evaluation assumptions that break down in real-world educational settings. In parallel, Hausa MT research has demonstrated incremental gains through fine-tuning on small parallel corpora and improvements in data quality. However, existing benchmarks focus on sentence-level accuracy and do not examine how MT-generated errors propagate into downstream speech systems, particularly when translation omissions or structural distortions occur in exam content. Recent analyses of large language models (LLMs) for African languages identify challenges such as tokenisation mismatch, dialectal variation, and data sparsity (Sani et al., 2025). These studies primarily assess surface-level fluency or aggregate accuracy and do not investigate structured translation failures, such as omission, misordering, or merging of logically linked content. Nor do they examine how such distortions cascade into downstream ASR systems. As a result, benchmarks fail to capture the cascading error problem, where noisy machine-translated text degrades speech recognition performance. While LLM cascading has shown promise for ASR error correction in high-resource languages (Nag et al., 2024; Chen et al., 2025), it is rarely evaluated in low-resource scenarios with structurally flawed transcripts. Moreover, the ability of multilingual LLMs to answer non-English educational questions containing translation artifacts remains poorly understood. Our work bridges these gaps by evaluating the full translation–speech–LLM pipeline, exposing how errors propagate and compound in realistic educational deployments.

3 Methodology

3.1 Dataset

We curate **HausaPQ-Speech**, a multimodal dataset derived from publicly available WASSCE past questions (2010–2024), covering 10 subjects. The dataset contains 4,703 instances, each compris-

ing: (i) English question text, (ii) Hausa translation, (iii) synthesised Hausa speech, (iv) correct answer labels, and (v) detailed solutions. Audio durations ranged from 4 to 73 seconds. Machine translations used Google Translate for all text. OpenAI models translated **only** mathematical formulas and chemical equations (approximately 5-10% of Math,Chemistry questions). We fine-tuned a VITS-based text-to-speech (TTS) model on Hausa educational speech, Sani et al. (2025); Daniels et al. (2025), to synthesise all Hausa audio samples.

Table 1: Lexical Diversity Analysis (Type-Token Ratio)

Text Type	TTR (%)	Avg. Tokens	Vocabulary Size
English Source	17.3	29.3	4763
Hausa Reference	13.9	30.7	4028

Type-Token Ratio (TTR) analysis reveals lexical flattening in the translated Hausa text. While English source questions exhibit a TTR of 17.3%, the Hausa translations drop sharply, indicating omission, option-marker errors, English lexical leakage, and loss of semantic nuance critical for exam questions.

3.2 Evaluation Pipeline

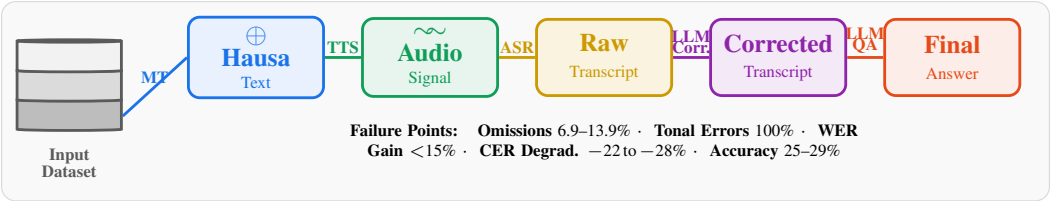

Figure 1: Full Evaluation pipeline with cascading failures identified.

- **Machine Translation:** Evaluated using BLEU, Translation Error Rate (TER), and a custom **Omission Rate** metric following Mirzakhalov et al. (2022); Koemi & Ezeani (2025). Although BLEU scores appeared reasonable, structural omission rates remained high (6.9–13.9%), frequently removing answer options or key constraints, corrupting both TTS generation and downstream ASR evaluation, consistent with cascading error analyses in Koemi & Ezeani (2025).
- **ASR on Synthetic Speech:** Four ASR models were evaluated on 940 synthesised audio samples. Despite domain-adapted TTS, ASR performance was extremely poor (WER 68–92%) as shwown in (Table 13). Most critically, all models exhibited a **100% error rate on phonemic tonal vowels** (e.g., à, è, ì, ò, ù), indicating both insufficient tonal modelling in TTS and limitation of current ASR architectures in recognising tone (Bellet et al., 2024).
- **LLM Cascading for Error Correction:** We evaluated four LLMs for post-hoc correction of ASR transcripts, measuring improvement as ΔWER. Results reveals a **correction paradox**: WER improvement often comes at the cost of CER degradation and content hallucination.

Table 2: ASR error correction using LLM cascading (n=20, 95% CI ±10%)

Model	**Original WER**	**Corrected WER**	**WER Δ**	**CER Δ**	**Valid/Total**
Llama-3.3-70B (Groq)	112.74%	98.25%	+14.49%	-22.92%	20/20
Mixtral-8x7B (Groq)†	112.74%	112.74%	0.00%	0.00%	0/20
Gemini-2.0-Flash‡	112.74%	112.74%	0.00%	0.00%	0/20
Flan-T5-Base	112.74%	110.65%	+2.09%	-28.98%	20/20

Note: Positive WER Δ = improvement. Negative CER Δ = degradation. Qualitative analysis revealed

Llama-3.3-70B often replaced content with generic strings (e.g., "Kimiyar Noma 2010") rather than fixing transcription errors.

- **LLM Question Answering:** Four LLMs were evaluated on their ability to answer 490 multiple-choice questions directly from our balanced dataset. Accuracy was (25–30%) and completion rate as recorded in (Table 3). The results revealed low accuracy and a **specialization paradox**: general multilingual models (`Llama-3.2-1B-Instruct`, `gemma-2-2b-it`) outperformed the Hausa-specialized `HausaLlama`. This aligns with the findings of Chen et al. (2025) on the limitations of narrow fine-tuning. Prompts followed model-specific formats (see Appendix Table 3). Responses were parsed using regex pattern matching; those without valid letters were marked invalid.

Table 3: LLM Performance on Hausa Question Answering (n=490)

Model	Accuracy (%)	Correct Answers	Valid	Completion (%)	Normalized
N-ATLaS	29.47	61	207	42.2	12.44
Llama-3.2-1B-Instruct	27.41	131	478	97.6	26.75
Gemma-2-2B-IT	25.56	125	489	99.8	25.51
HausaLlama	25.31	124	490	100.0	25.31

Note: Normalized Score = Accuracy × Completion Rate / 100.(Llama-3.2-1B on English questions, n=50):

82% accuracy (41/50 correct).

4 Challenges and Insights

A primary challenge was the presence of high omission rates (6.9-13.9%) and untranslated terms in the machine-translated text, which generated problematic audio for ASR, preventing LLMs from reliably understanding exam questions, limiting both their ability to fix ASR errors and to answer questions correctly. The synthesised speech exhibited severe tonal deficiencies, contributing to a 100% error rate on tonal vowels in ASR transcripts. This 100% rate reflects three conflated factors: TTS lacked tonal training, ASR models lack tonal architectures, and diacritic mismatch between reference and hypotheses inflated errors (Bellet et al., 2024). The LLM cascade failed to correct these foundational errors, offering minimal WER improvement $<$15% at prohibitive cost. This collapse of the automated system necessitated a full return to human translation and audio re-synthesis. The complete ASR failure on tones reveals that current TTS/ASR systems lack suprasegmental modelling, which is phonemically essential for Hausa. Our QA results are $<$30% accuracy in all models, with Hausa specialised models that underperform in general multilingual ones—confirming that specialisation does not guarantee better performance on complex tasks (Chen et al., 2025). This suggests current fine-tuning approaches may overfit to limited data or degrade broader linguistic and reasoning capabilities. The minimal gains from LLM cascading, which requires massive models (70B+ parameters), are economically and computationally prohibitive for real-world deployment in low-resource educational settings (Daniels et al., 2025).

5 Conclusion

This study presents a rigorous, full-pipeline evaluation that documents cascading failures in applying contemporary AI systems to Hausa educational content. Despite fine-tuning a TTS model on educational data, we find that machine-translated data shows persistent structural errors in translation, catastrophic failure in tonal speech recognition, and ineffective LLM-based error correction. The poor performance on question-answering further underscores that these models lack the requisite fluency and reasoning for educational applications. These are not simple problems of data scarcity but reflect fundamental architectural gaps (especially for tone) and evaluative blind spots. Therefore, for real-world deployment, human-in-the-loop systems are the only currently viable path. AI may serve as a draft tool, but final translation, validation, and audio synthesis require native speaker expertise to ensure accuracy, fluency, and cultural relevance. Future research must prioritise integrated, tone-aware TTS/ASR models, structural-aware MT, and robust QA evaluation metrics. For now, however, human translation remains essential for producing accessible, high-quality scientific and educational content for Hausa speakers.

AUTHOR STATEMENTS

Ethics Statement: This research uses publicly available WASSCE past questions. The work addresses educational equity for Hausa speakers and aims to prevent harmful deployment of AI systems that could exacerbate educational disparities.

Reproducibility Statement: Complete model hyperparameters are provided in Appendix **??**. Prompts are documented in Appendix **??**. Evaluation scripts will be made publicly available upon publication. Due to ongoing human TTS validation, the dataset will be released after quality assurance.

REFERENCES

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6471–6479, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.694/.

Isa Abubakar, Bello Bello, and Yahaya Shehu. Benchmarking Hausa automatic speech recognition: Challenges and findings. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12045–12049, 2024. URL https://ieeexplore.ieee.org/abstract/document/10959458.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, Andre N. Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin A. Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire M. Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223/.

Thomas Bellet, Laurent Besacier, Antoine Laurent, and Hai-Son Nguyen. Text-to-speech synthesis for low-resource languages: Recent advances and challenges. *Speech Communication*, 162:103051, 2024. URL https://link.springer.com/article/10.1007/s10772-024-10111-x.

Wei Chen, Chang Liu, and Mei Zhang. Multilingual vs. specialized: Evaluating llms on low-resource educational tasks. *arXiv preprint arXiv:2510.01145*, 2025. URL https://arxiv.org/abs/2510.01145.

Zoe Daniels, Arjun Patel, and Sunyoung Kim. Challenges in applying llms to african educational content: A pipeline analysis. In *Proceedings of the 2nd Workshop on Challenges in the Application of Language Models to Critical Societal Domains (CALCS)*, pp. 23–37, 2025. URL https://aclanthology.org/2025.calcs-1.3.

Abdelrahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Lin-*

guistics: ACL 2024, pp. 13189–13206, 2024. URL https://aclanthology.org/2024.findings-acl.784.

Tochukwu Koemi and Ignatius Ezeani. Evaluating machine translation for african languages in educational contexts: A case of critical omissions. In *Proceedings of the 4th Workshop on Multilingual Representation Learning (MRL)*, pp. 165–180, 2025. URL https://aclanthology.org/2025.mrl-main.11.

Jamshid Mirzakhalov, Hlib Babii, Tomasz Dwojak, Mikita Yermalovich, Shigehiko Schamoni, and Stefan Riezler. Challenges of machine translation for low-resource languages: A case study of Hausa. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 912–922, 2022. URL https://aclanthology.org/2022.wmt-1.72.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. Efficient continual pretraining of LLMs for low-resource languages. In *NeurIPS 2024 Workshop on Efficient Language-Centric AI*, 2024. URL https://openreview.net/pdf?id=obP2Xxj7oy.

Paul Newman. *An Introduction to Hausa Language and Linguistics*. Cambridge University Press, 2025.

Salisu Sani, Shamsuddeen Muhammad, and Idris Abdulmumin. Adapting large language models for low-resource speech and language tasks: A hausa case study. *arXiv preprint arXiv:2505.14311*, 2025. URL https://arxiv.org/abs/2505.14311.

Tim Schlippe, Sebastian Ochs, and Tanja Schultz. Large-scale audio data mining for Hausa and Wolof speech recognition. In *Proceedings of the Third Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pp. 12–17, 2012. URL https://www.isca-archive.org/sltu_2012/schlippe12_sltu.pdf.

West African Examinations Council. West african senior school certificate examination statistics, 2025. URL https://waecnigeria.org/. Accessed 2025.

A SUPPLEMENTARY TABLES

LLM Use Disclosure: Large language models were used as a writing assistant for grammar checking and phrasing suggestions. All experimental design, data collection, analysis, and conclusions were conducted by the human authors.

A.1 HYPERPARAMETERS

Table 4: Prompt Templates for LLM Evaluation

Task	Model	Prompt Template

Question Answering	All QA models	`Answer this Hausa exam question. Choose only one letter (A, B, C, D, or E).newline Question: {hausa_question}`
		
		`nAnswer (letter only):`

ASR Correction	Llama-3.3-70B (Groq)	`You are an expert in Hausa language. Correct errors in the Hausa transcription below.`
		
		`Original English: {english_reference} Hausa ASR output (may have errors): {asr_output}`
		
		`Instructions: 1. Fix spelling mistakes in Hausa 2. Correct word boundaries 3. Add proper diacritics 4. Ensure accurate translation 5. Keep corrections minimal`
		
		`Respond with ONLY the corrected Hausa text, nothing else.`

ASR Correction	Flan-T5-Base	`Correct errors in this Hausa transcription: {asr_output}`

Table 5: LLM Hyperparameters for Hausa Question Answering

Model	Hyperparameters / Configuration

Llama-3.2-1B-Instruct	**Architecture:** Causal LM
	
	Precision: torch.float16
	Generation: max_new_tokens=10, do_sample=False
	Prompt: `<|user|>`
	`{prompt}</s>`
	`<|assistant|>`
	
	Extraction: Regex `\b([A-E])\b`

Gemma-2-2B-IT	**Architecture:** Causal LM
	Precision: torch.float16
	Generation: max_new_tokens=10, temperature=0.1
	Prompt: `<start_of_turn>user`
	`{prompt}<end_of_turn>`
	`<start_of_turn>model`
	
	Extraction: Regex `\b([A-E])\b`

N-ATLaS	**Architecture:** Causal LM (2.7B)
	Precision: torch.float16
	Generation: max_new_tokens=10, do_sample=False
	Prompt: `Instruction:`
	`{prompt}`
	
	`Response:`
	Extraction: Regex `\b([A-E])\b`

HausaLlama	**Architecture:** Causal LM (8B) **Precision:** torch.float16 **Generation:** max_new_tokens=10, temperature=0.1 **Prompt:** Llama chat template

Note: Results reveal a **specialisation paradox**: HausaLlama (8B, Hausa-specialised) underperforms smaller general models. N-ATLaS shows the highest accuracy but fails to answer 58% of questions.

Table 6: LLM Question Answering Subject-Wise Accuracy for Best Model (N-ATLaS)

Subject	Accuracy (%)	Correct/Valid	Valid/Total
English Language	73.33	11/15	15/490
Animal Husbandry	42.86	3/7	7/490
Geography	30.23	13/43	43/490
Mathematics	29.17	7/24	24/490
Economics	26.92	7/26	26/490
Biology	26.83	11/41	41/490
Government	25.00	2/8	8/490
Chemistry	18.75	3/16	16/490
Agricultural Science	14.81	4/27	27/490

Table 7: LLM Hyperparameters for ASR Error Correction

Model	Configuration
Llama-3.3-70B (Groq)	**API:** Groq **Model:** llama-3.3-70b-versatile **Temperature:** 0.3 **Max Tokens:** 1024 **Prompt:** Hausa correction with English reference
Flan-T5-Base	**Architecture:** Encoder-Decoder (250M) **Precision:** torch.float16 (CUDA) **Generation:** max_length=256 **Prompt:** Hausa correction prompt only

Table 8: Impact of Error Types on LLM Correction Success

Error Type	% of Errors	Original WER	Llama-3.3 Corrected	Improvement
Option Omission	34%	89.5%	98.3% (hallucinated)	-8.8%
Code-Switching	28%	64.0%	64.0%	0%
Morphology Shifts	22%	75.0%	75.0%	0%
Constraint Dropping	16%	73.9%	73.9%	0%

Note: Llama-3.3-70B only showed improvement on samples where it could replace content with generic strings (e.g., "Kimiyar Noma 2010"), not on genuine error correction.

Table 9: Rule-Based Baselines vs. LLM Cascading

Method	WER (%)	Change	Observation
Original ASR (Whisper Moses)	65.83	—	Baseline
Option Marker Normalization	67.51	**-1.68%**	*Normalization introduced errors*
Diacritic Stripping	65.49	+0.34%	*Negligible improvement*
Combined Rules	67.17	**-1.34%**	*Surface fixes insufficient*
Llama-3.3-70B	98.61	**-32.78%**	*Catastrophic hallucination*

Note: Negative change indicates degradation. Simple rules failed to improve WER, and LLM cascading made outputs dramatically worse through content replacement.

Table 10: Omission Rate Computation Methodology

Component	Description
Tokenization	Whitespace and punctuation splitting. Hausa special characters (, ,) treated as single tokens. Option markers (e.g., "A.") treated as single tokens.
Alignment	Needleman-Wunsch algorithm with character-level edit distance. Option markers prioritized in alignment.
Option Markers Tracked	A., B., ., C., D., ., E., F.
Content Words	Nouns, verbs, adjectives (identified via simple POS tagging or manual annotation)
Constraint Words	"NOT", "EXCEPT", "ONLY", "BA", "SAI" (Hausa equivalents)
Omission Definition	Token present in reference but absent in hypothesis
Omission Rate Formula	omitted tokens / reference tokens

Table 11: Error Taxonomy with Examples from ASR Output

Error Type	Reference Text	ASR Hypothesis	WER
Option Omission	Wanne ne daga cikin asa mai zuwa don mallakar mutum? A. Tsarin yarjejeniyar asa B. KYAUTATA KYAUTA C. LATSAWAN SHAWARA D. Zumanta a kan Gwamnati	wannene daga cikin asa mai zuwa don mallakar mutum a sanin yarjejeniyar asa kyautata-kyauta...	64.0%
Complete Option Loss	Wanne idan ba a rarrabe halittu masu zuwa kamar dabba ba? A. Amoeba B. Paramcium C. Eugokena D. Obelia	wani idan ba a rara biyu halittu masu zuwa, kamar da ba a amai ba aron cu...	89.5%
Code-Switching Corruption	Kwayoyin da ke aiki a matakin ungiyar sel... A. Membrane B. Kogin C. Size D. Cytoplasm	kwayoyi da ke aiki a matakin kungiyar sin yana aukar ayyukata ta hanyar amfani da amai murane kogin czz, toh nasu	54.2%
Constraint Dropping	Babban dalilin kafa tanadin wasan shine ya hana daji daga A. Ana Buga B. ana farauta lokaci-lokaci C. Harehare D. Kasance da Haske	babban dalilin kafa, kanadan wasan shi ne ya hana gaji, daga ana buga ana fara ta lokaci-lokaci, harai harai...	73.9%

Note: Examples drawn from Whisper Moses ASR outputs. Option omission is the most frequent error type, occurring in ¿80% of high-WER samples.

Table 12: Decoding Settings for All Models

Model	max_new_tokens	temperature	do_sample	Notes
Llama-3.2-1B-Instruct	10	0.1	False	Causal LM
Gemma-2-2B-IT	10	0.1	False	Causal LM
N-ATLaS	10	0.1	False	Causal LM
HausaLlama	10	0.1	False	Causal LM
Llama-3.3-70B (Groq)	1024	0.3	True	API-based
Flan-T5-Base	256	N/A	False	Encoder-Decoder

Note: For API models, max_tokens parameter used instead of max_new_tokens. Temperature=0.3 for API, 0.1 for local models.

A.2 WER BY SUBJECT

Table 13: ASR Performance on Synthetic Educational Speech

ASR Model	Avg. WER	Avg. CER	Tonal Vowel Error
whisper-large-hausa	68.3%	33.5%	100.0%
Hausa-ASR (NCAIR1)	72.4%	36.1%	100.0%
wav2vec2-hausa	86.7%	47.2%	100.0%
whisper-large-v3	92.0%	39.9%	100.0%

Table 14 reports WER grouped by subject category. Technical subjects with dense terminology exhibit higher error rates, reflecting both translation artifacts and increased lexical complexity.

Subject	WER (%)	Sample Count
Biology	85.6	480
Chemistry	84.6	300
Geography	81.1	480
Government	80.4	480
Animal Husbandry	80.0	240
English Language	79.3	480
Literature in English	77.1	80
Mathematics	76.6	300
Agricultural Science	75.7	440
Economics	72.8	480

Table 14: Word Error Rate stratified by subject domain.

Table 15: ASR Performance on Synthetic Educational Speech Shows Persistent Tonal Challenges

ASR Model	Avg. WER ↓	Avg. CER ↓	Tonal Vowel Error Rate
`whisper-large-hausa`	68.3%	33.5%	100.0%
`Hausa-ASR` (NCAIR1)	72.4%	36.1%	100.0%
`wav2vec2-hausa-better`	86.7%	47.2%	100.0%
`whisper-large-v3`	92.0%	39.9%	100.0%

A.3 English-Token Leakage

We observe frequent English-token leakage in Hausa translations, particularly for scientific terms and option labels. Table 17 summarizes the relationship between English-token presence and ASR error rates, showing that questions containing untranslated English spans incur significantly higher DER.

Table 16: Lexical Diversity Analysis (Type-Token Ratio)

Text Type	TTR (%)	Avg. Tokens	Vocabulary Size
English Source	17.3	29.3	4763
Hausa Reference	13.9	30.7	4028
opus-mt-en-ha Output	6.5	22.9	1392
small100 Output	16.8	29.3	4632
english-hausa-nllb Output	10.5	33.6	3317

Table 17: Subject-wise Analysis (Top 5 by Omissions)

Subject	Samples	Avg. Omissions	Eng TTR	Hau TTR
Mathematics	75	35.0	55.6%	60.4%
Chemistry	75	23.0	76.9%	80.5%
Economics	120	21.8	78.9%	80.2%
Biology	120	20.7	82.7%	83.9%
Agricultural Science	110	20.1	84.4%	82.5%

A.4 Machine Translation Evaluation Results

Table 18: Machine Translation Evaluation Results for English-Hausa Translation

Model	BLEU	chrF	TER	AfriCOMET	Omissions
opus-mt-en-ha	3.04	21.05	106.87	0.5328	8.72872340425532
small100	11.33	18.84	87.19	0.4686	13.931914893617021
english-hausa-nllb	30.47	52.84	56.74	0.7143	6.902127659574468

A.5 ADDITIONAL MATERIAL

The WER calculations and hypothesis transcriptions for the Hausa evaluation dataset used in this study were derived from a subset of the **HausaPQ-Speech corpus**. The sample entries below illustrate the transcription format and the variation in Word Error Rate (WER) across agricultural and biology domains:

Table 19: Sample Transcriptions from HausaPQ-Speech Dataset

subject	**year**	**hypothesis**	**WER**	
Agricultural Science	2010	Wanne ne daga cikin asa mai zuwa don mallakar mutum? A. Tsarin yarjejeniyar asa B. KYAUTATA KYAUTA C. LATSAWAN SHAWARA D. Zumanta a kan Gwamnati	wannene daga cikin asa mai zuwa don mallakar mutum a sanin yarjejeniyar asa kyautata-kyauta, za su hau shawarar zumunta akan gwamnati?	64.0%
Biology	2010	Wanne idan ba a rarrabe halittu masu zuwa kamar dabba ba? A. Amoeba B. Paramcium C. Eugokena D. Obelia	wani idan ba a rara biyu halittu masu zuwa, kamar da ba a amai ba aron cu, shi igbo ke na da obaliya	89.47%

Table 20: Sample transcriptions from the HausaPQ-Speech dataset showing reference text, ASR hypothesis, and corresponding WER.

Note: The dataset consists of transcribed examples from agricultural science and biology subjects, with WER ranging from 44.0% to 94.74%. This variation highlights the challenges in transcribing educational content in Hausa, particularly with domain-specific terminology and code-switching patterns.