AfriVox: Probing Multilingual and Accent Robustness of Speech LLMs

Anonymous ACL submission

Abstract

Recent advances in multimodal large language models (LLMs) have enabled impressive speech recognition and translation capabilities, yet these models remain poorly evaluated in low-resource settings, particularly for African languages and non-native English accents. In this work, we systematically compare state-of-the-art speech-based LLMs with traditional Automatic Speech Recognition (ASR) systems across transcription and translation tasks involving dialectally diverse African speech. To support reproducible evaluation, we introduce AfriVox, a novel open-source benchmark comprising medical and non-medical speech samples spanning 20 African languages and 100+ African English accents. Our findings reveal substantial performance disparities, underscoring the limitations of current LLMs in handling underrepresented linguistic varieties. To address this, we fine-tune the newly released Qwen-2.5-Omni for multilingual transcription and translation using NaijaVoices, a 1,800-hour Nigerian speech corpus. Fine-tuning via instruction-tuned, LoRA-based parameter-efficient methods yields a 54% reduction in Word Error Rate (WER) and a 21% average improvement in BLEU scores over baseline models. Our results demonstrate that multimodal LLMs can be effectively adapted for low-resource speech tasks using lightweight techniques. This work provides a foundation for scalable speech technology development in underrepresented languages and informs future research in inclusive multimodal learning.

1 Introduction

Recent rapid LLM advancements have enabled multimodal data processing (McKinzie et al., 2024; Cappellazzo et al., 2024). LLMs like GPT-4o (Hurst et al., 2024), Gemini (Team et al., 2024), and SALMONN (Yu et al.) now take native speech input, bypassing text altogether, showing promising performance across multiple languages and accents (Kwak and Pardos, 2024).

Despite these advancements, the performance of these multimodal models on low-resource languages remains underexplored (Liu and Niehues, 2024; Yin et al., 2024; McKinzie et al., 2024). In Nigeria alone, over 200 million people communicate in Igbo, Hausa, Yoruba, and Pidgin, yet off-the-shelf ASR and translation systems exhibit high error rates, code-switching failures, and dialectal bias (Ogunmodimu, 2015).

Several studies have explored unimodal speech models for African languages (e.g., Whisper, MMS, AfricanHubert, Seamless by Meta (Radford et al., 2023; Denisov and Vu, 2024; Alabi et al., 2024; Barrault et al., 2023)). However, the performance of multimodal speech LLMs for several African languages remains an open question (Yin et al., 2024). Multimodal LLMs with capabilities to handle multiple data types - text, images, audio, video - tasks simultaneously hold significant promise beyond communication, particularly in enhancing access to accurate and personalized information (Lyu et al., 2023). Therefore, understanding their ability to process spoken and indigenous languages from African-accented countries is essential to promote inclusive speech-driven AI in Africa (Sanni et al., 2025a).

In this work, we investigate the generalizability and robustness of speech- and multimodal LLMs to African languages and non-native English accents, comparing them with traditional unimodal ASR models. Our results reveal wide performance gaps with African languages and dialects. To address this gap, we fine-tuned the Qwen 2.5 Omni model on 3 African languages for transcription and translation, applying parameter-efficient fine-tuning (PEFT) (Ding et al., 2023; Han et al., 2024; Ding et al., 2023) achieving a 54% relative reduction in WER and an 21-point BLEU gain for transcription and translation respectively. As a final contribution, we release 2 diverse benchmark sets to measure progress on African languages: (i) a multilingual translation test set for 20 African languages, and (ii) a multilingual transcription test set for those same 20 languages, all curated from a wide array of sources. Our work aims to provide valuable insights for building more inclusive, multilingual voice-native systems by establishing a strong baseline for evaluating unimodal and multimodal speech LLMs in low-resource settings and demonstrating the potential of instruction tuning to improve their performance.

2 Related works

Prior work suggests that three main trends –scaling laws, reinforcement learning, and the emergence of self-supervised learning– are responsible for the current advances in speech-large language models (LLMs) (Sanni et al., 2025b; Liu and Niehues, 2024; Wang et al., 2024; Johnson et al., 2014). However, these performance gains are dominated by high-resource languages, particularly English (Olatunji et al., 2023; Radford et al., 2023), with these gains remaining unevenly distributed. Training sets are dominated by English and other high-resource languages or multilingual corpora with limited coverage for African languages and dialects (Shanbhogue et al., 2023; Lam-Yee-Mui et al., 2023; Hamed et al., 2022). As a result, while speech-based LLMs excel in challenging tasks such as open-domain question answering and conversational interactions (Wu et al., 2024; Nachmani et al., 2023), their applicability to the rich linguistic landscapes of Africa remains underexplored (Reitmaier et al., 2022). Furthermore, accent mismatch, code-switching, and sparse training data significantly impact model performance for African languages (Tachbelie et al., 2014; Sanni et al., 2025a).

Recent multimodal LLMs now integrate speech and text in unified architectures. Examples include Google's AudioPaLM (Rubenstein et al., 2023; Wang et al., 2024) which combine a PaLM-based LLM with a wav2vec-style speech encoder; Meta AI's SeamlessM4T (Barrault et al., 2023) which offers an all-in-one solution for speech-to-text, speech-to-speech, text-to-speech, and text-to-text, and Alibaba's Qwen-Audio (Chu et al., 2023), which scales audio-language pretraining across 30+ tasks, achieving breakthrough performance in speech based tasks (Wang et al., 2024).

Given these multimodal capabilities, fine-tuning such massive models for each new downstream task incurs prohibitive memory and compute costs (Han et al., 2024). Parameter-efficient fine-tuning (PEFT) has been proposed as a possible way to address this challenge by updating only a small subset of parameters, thus reducing resource overhead (Ding et al., 2023). Such strategies include adapters (Han et al., 2024), which insert lightweight bottleneck modules into each Transformer layer; LoRA (Karimi Mahabadi et al., 2021), which updates low-rank matrices (0.1–1 % of parameters) that can be merged into the backbone at inference; hybrid methods such as QLoRA—combining 4-bit quantization with LoRA on a single GPU—have further pushed this efficiency frontier (Dettmers et al., 2023). Together, these PEFT methods enable rapid, cost-effective adaptation of multimodal LLMs in resource-constrained and low-data regimes (Dettmers et al., 2023).

3 Methodology

3.1 Datasets

This work evaluates speech-based LLMs and unimodal ASR models on low-resource African languages and explores the benefits of fine-tuning multimodal LLMs. To support these tasks, we curated and open-sourced two datasets categories: African Accented English Speech (AES) and Multilingual African Speech (MLS) for benchmarking and model evaluations, while using the open-sourced NaijaVoices datasets for fine-tuning.

3.1.1 African Accented English Speech (AES)

We compiled speech from the NCHLT (Barnard et al., 2014), AfriSpeech (Olatunji et al., 2023), Common Voice 17 (filtered for African accents) (Ardila et al., 2020). The combined dataset consisted of 63.2 hours of speech from 2,000+ speakers across 12 countries and 108 distinct accents (Table 1).

3.1.2 Multilingual African Speech (MLS)

This group of datasets comprises 20 African languages across 7 public and private datasets, designed for ASR and AST benchmarking (Tables 2 and 3). For transcription, we included NCHLT, Common Voice 17, FLEURS, OpenSLR, BibleTTS, NaijaVoices[1], FISD[2], MedConv-Transcribe [4]. For translation, we included FLEURS, CoVoST(), NaijaVoices, IWSLT-LRST, MedConv-Translate [5].

3.1.3 NaijaVoices Dataset

For fine-tuning, we utilize the NaijaVoices dataset (Emezue et al., 2025): a 1,800-hour corpus with 600 hours each for Igbo, Hausa, and Yoruba. It includes 5,000+ speakers with balanced gender and age distributions (Table 3).

3.2 Data Quality and Ethics

All audio files are mono-channel WAV at 16kHz. Public datasets contain predefined transcripts. Parliamentary recordings were manually transcribed by native speakers and quality-checked; only those with over 80% reviewer approval were retained.

Dataset	Hours	Speakers	Accents
NCHLT	2.24	8	1
AfriSpeech	18.68	750	108
CV-17 En-Afr	0.11	46	9
Afrispeech-Parl (Sanni et al., 2025a)	42.17	∼1651	4
Total	**63.20**	**∼2455**	**108**

Table 1: Summary of African-accented English speech datasets.

Language	Region	Language Family	# Speakers
afr	South	IndoWest (Germanic)	7.2M
aka	West	Niger-Congo (Kwa)	24M
amh	East	Afro-Asiatic (Semitic)	35M
arz	North	Afro-Asiatic (Semitic)	78M
fra	West	Indo-European (Romance)	320M
ful	West	Niger-Congo (Atlantic)	36.8M
gaa	West	Niger-Congo (Kwa)	0.7M
hau	West	Afro-Asiatic (Chadic)	54M
ibo	West	Niger-Congo (Volta-Niger)	31M
kin	East	Niger-Congo (Bantu)	15M
lug	East	Niger-Congo (Bantu)	5.6M
nso	South	Niger-Congo (Bantu)	4.6M
sna	South	Niger-Congo (Bantu)	8.4M
sot	South	Niger-Congo (Bantu)	5.6M
swa	East	Niger-Congo (Bantu)	87M
tsn	South	Niger-Congo (Bantu)	8.2M
twi	West	Niger-Congo (Kwa)	4.4M
xho	South	Niger-Congo (Bantu)	8M
yor	West	Niger-Congo (Yoruboid)	45M
zul	South	Niger-Congo (Bantu)	13.6M

Table 2: Language, region, family, and number of speakers.

3.3 Models

For the evaluation task, we assessed five unimodal models for ASR and three for Automatic Speech

Dataset	Num Langs	Hours	Speakers
NCHLT	6	12.75	36
CV-17	10	16.89	670
FLEURS	13	14.44	1595
OpenSLR	3	0.31	372
Bible TTS	3	0.47	3
NaijaVoices[1]	3	1800	5000
FISD[2]	3	0.05	23
MedConv [3]	19	36.63	1179
	Total Hours	**1878.52**	

Table 3: Summary of multilingual speech datasets.

Translation (AST). MMS was excluded from translation evaluation as it was not trained for this task, and Parakeet-TDT is a monolingual ASR model. Four multimodal LLMs were evaluated for ASR: SeamlessM4T (Barrault et al., 2023), Gemini 2.0 Flash (Team et al., 2024), GPT-4o Audio Preview and Qwen2.5-Omni-7B (Chu et al., 2024). We utilized the pre-trained models or API endpoints without additional fine-tuning. Notably, only Qwen2.5-Omni-7B is open-source; the others are accessible via API. Therefore, we used Qwen 2.5 omni (Yang et al., 2025) for the PEFT fine-tuning.

4 Experiments

4.1 Experimental Setup

We evaluate both base and fine-tuned models across two tasks: Automatic Speech Recognition (ASR) and Automatic Speech Translation (AST). Inference is performed in two modes: using the base model's default settings and using the same setup with a fine-tuned model. For each task, we test three prompting strategies (detailed in Appendix A). All models use standard inference parameters unless otherwise noted. Inference was conducted on a single NVIDIA T4 for ASR and an NVIDIA A100 for the AST model with the largest memory footprint.

4.2 Fine-tuning Details

Due to our limited compute budget, we fine-tuned Qwen2.5-Omni-7B on approximately 280 hours per language from the NaijaVoices dataset using LoRA (rank 8, alpha 32), applied to all linear layers while freezing the vision encoder. We trained for three epochs using a learning rate of 1e-4 and a warmup ratio of 0.05. We used bfloat16 precision, a per-device batch size of 4, and gradient accumulation steps of 16. Training was conducted on four NVIDIA 3090 GPUs, with evaluations and check-

[1] `https://huggingface.co/datasets/naijavoices/naijavoices-dataset`

[2] `https://github.com/Ashesi-Org/Financial-Inclusion-Speech-Dataset`

[3] URL to be added after anonimity period

points every 500 steps. Prompt formatting details are included in Appendix A.

4.3 Post-processing.

To ensure fair comparisons, we normalize the output before scoring. For African-accented English ASR, we use a custom cleaning function to remove filler words, extraneous whitespace, and punctuation inconsistencies. For multilingual ASR, we apply Whisper's `BasicTextNormalizer`, removing diacritics to mitigate variability from inconsistent labeling. For AST, we use Moses tools (`MosesPunctNormalizer` and `MosesTokenizer`) for consistent punctuation and tokenization across languages.

4.4 Evaluation Metrics.

We apply a consistent evaluation protocol to both base and fine-tuned models across ASR and AST tasks. ASR performance is measured using Word Error Rate (WER) (Klakow and Peters, 2002), defined as the total number of substitutions, deletions, and insertions divided by the number of words in the reference. For AST, we report BLEU (Papineni et al., 2002), chrF (Popović, 2015), and two African-centric AfriCOMET-STL (Wang et al., 2023), which evaluate semantic adequacy using multilingual and single-task learning, respectively. We use AfriComet-STL as our main metric after conducting human-evaluation to identify which metric best evaluates the translation quality. The results from human-evaluation can be found in Appendix 16

5 Results and Analysis

Tables 4 and 5 present the transcription results on the African-Accented English Speech and Multilingual African Speech datasets. Results presented are for single runs. The results indicate that, in most cases, unimodal models outperformed the multimodal models. While Table 7 show multimodal models edges over unimodal models on the speech translation task. Additionally, Table 6 shows the comparison between the results of the base and fine-tuned Qwen 2.5 Omin model. A detailed breakdown of results by individual languages is provided in Appendix A. We provide the following analysis based on the findings from our experimental results.

Model	Lib	Af	NC	CV	Parl
Canary	1.48	38.03	**10.05**	**8.41**	27.38
Parakeet	**1.40**	34.96	11.33	9.48	21.89
Whisper M	3.02	30.81	10.17	12.39	28.53
Whisper L	2.01	**26.49**	10.10	12.54	**19.29**
MMS	12.63	61.19	32.11	23.09	107.41
M4T	2.89	49.75	32.96	10.40	54.68
Gemini	3.03	28.12	14.19	13.76	21.63
GPT-Aud.	5.26	36.54	86.52	26.76	41.88
Qwen2	1.60	49.61	25.14	11.16	57.43

Table 4: Word Error Rates (WER) across African-accented English speech data sources and Librispeech test-clean [Lib]. Af: Afrispeech, NC: NCHLT, CV: Common Voice, Parl: Parliamentary Proceedings (Panayotov et al., 2015), models in top are unimodal ASRs while those below are multimodal LLMs

5.1 Accent Robustness Gaps for African Speech

Across all models, WER on African-accented English and true African languages is dramatically higher than on native English or French as shown in Tables 4 and 5. For example Whisper Large-v3's WER increases from 2.01% on LibriSpeech to 26.49% on Afrispeech (Nigerian accents)- a more than ten-fold increase (Table 4). Likewise, MMS-1B-All—despite multilingual pretraining—yields 61.19% WER on Afrispeech, compared to 12.63% on LibriSpeech (Table 4). On individual languages such as Hausa and Yoruba, error rates often exceed 100% (e.g., 180.29% and 213.88% WER for Whisper Medium on Swahili and Yoruba respectively; Table 5), indicating severe misrecognitions. These findings highlight that simply including African data in pretraining does not guarantee accent robustness; improving performance in low-resource settings may require targeted accent adaptation and balanced data sampling.

5.2 Noise and Speaker-Overlap Vulnerability

When evaluated on the noisy parliamentary proceedings dataset, all models experienced substantial WER inflation. Whisper Large-v3's WER rose from 10.10% on NCHLT to 19.29% on Parlimentary audio, while GPT-4o Audio-Preview's WER soared to 41.88% (Table 4). Overlapping speech and background chatter proved especially challenging: systems often failed to segment speakers or filter noise, resulting in garbled transcripts or placeholder outputs ("cannot transcribe this audio"). Interestingly, Gemini-2.0 (Flash) remained comparatively robust, achieving a 21.63% WER—close to

Model	eng	fra	afr	aka	ara	fra	hau	ibo	kin	lug	sna	swa	xho	yor	zul
Canary	**3.03**	**4.06**	-	-	-	9.67	-	-	-	-	-	-	-	-	-
Whis. M	6.80	8.90	68.87	-	39.49	13.95	180.29	-	-	-	193.21	117.7	-	213.88	-
Whis. L	3.53	5.38	45.43	-	29.72	**9.31**	95.11	-	-	-	110.35	62.75	-	93.77	-
MMS	17.63	19.3	48.73	**62.92**	44.94	33.93	**40.47**	**50.33**	**36.73**	28.85	**30.7**	28.37	**42.24**	39.59	43.19
Qwen2.5	16.32	10.43	-	-	-	24.14	-	-	-	-	-	-	-	-	-
M4T	4.14	5.38	**18.41**	-	51.26	15.9	-	70.03	-	**16.39**	76.05	**16.25**	-	**37.43**	52.53
GPT-Aud.	9.63	22.71	84.36	104.02	31.88	22.29	118.6	112.23	135.75	131.19	90.51	73.96	130.79	101.14	135.84
Gemini	6.59	5.49	28.68	76.56	**16.11**	10.13	48.52	81.91	78.81	80.18	50.64	22.4	51.92	67.36	**35.71**

Table 5: Word Error Rates (WER) on Multilingual African Speech. Columns left of the vertical line show baseline performance on Multilingual LibriSpeech (Pratap et al., 2020), while those to the right display results for a selected subset of the 20 evaluated languages. A dash (–) means the model does not support that language, models in top are unimodal ASRs while those below are multimodal LLMs

Whisper's 19.29%—and outperforming other multimodals by 10+ points (Table 5). These results highlight that specialized acoustic models retain an advantage under adverse conditions, but some multimodal architectures can match that resilience if they incorporate sufficient noisy-audio training or robust front-ends.

5.3 Multimodal Models Struggle with Verbatim Transcription.

While multimodal models offer multiple avenues for language processing, they often struggle with verbatim transcription, which is key in ASR tasks. Instead of transcribing the exact spoken content, these models sometimes paraphrase the speech or generate descriptions of either the speech content or the audio's characteristics. In some cases, they fail to produce a transcription altogether, generating placeholders such as "cannot transcribe this audio." This behavior suggests that multimodal models prioritize high-level understanding over word-for-word transcription, making them less reliable for tasks requiring precise transcriptions. Figure 1 illustrates some of the common failure modes.

Example 1 [Af]: Paraphrasing and Audio Description
Reference: Adana spoke with doctor
Qwen2-Audio: A woman is saying Adana spoke with doctor

Example 2 [Parl.]: Content Description
Reference: We had legislation in front of this house to push down funds to the lowest levels of service delivery in the counties, namely the wards. What we have discussed this morning is that a lot of areas are against.
GPT Audio: The audio content discusses legislation aimed to allocate funds to the lowest levels of service delivery in counties, specifically the wards. It indicates that there is some disagreement or istance to this approach in various areas.

Figure 1: Examples of paraphrasing and audio description.

5.4 Multimodal Models Offer Better Language Coverage

Table 5 shows that multimodal models can support a much wider range of African languages compared to unimodal models. For instance, Gemini and SeamlessM4T achieve moderate-quality transcriptions for multiple African languages. Gemini is able to achieve this without needing explicit language prompts (i.e., there is no need to write the prompt in the language of the audio or supply a language ID). In contrast, some unimodal models demonstrate little to no support for these languages, underscoring a critical gap in language coverage.

5.5 Model Performance Across Languages

ASR: While MMS-1b (with language adapters) delivers the best overall performance for transcription, a closer examination reveals that different models excel in specific areas. SeamlessM4Tv2, for example, shows particularly strong results for Southern and Eastern African languages, providing clues about the language distribution in its training data. MMS performs best or remains competitive across most languages demonstrating stronger generalizability potential. These performance nuances suggest that model design, data, and training strategy can be optimized to tackle specific linguistic challenges in African languages–a promising direction for future research. Some examples of ASR outputs from the models are shown in Figure 2.

5.6 Performance Contrast with High-Resource Languages

The English and French WERs in Table **??** highlight a significant performance divide between high-resource and African languages. For example, our results show a significant gap in performance on native vs African-accented French. The gap worsens considerably as we examine other relatively large

Example 1: Background Noise
Reference: Uso wao ni kijvu zaidi kuliko mvesui.
Whisper Large-v3: kwa hivyo kwa hivo kw hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo.

Example 2: Word substitution
Reference: A adalai Hausawa ana ẏwa yara masu kaciya a cikin sa safar bakaahwi.
Gemini2.0: *A daddare* Hausawa ana yiwa yara masu kaciya in san ke shakar bakwai.

Example 3: Wrong language
Reference: awon obinrin naa na je isu.
GPT-Audio (French): malheureusement je ne peux pas repondre a des questions ou identifier des locuteurs à partir d'un echantillon vocal.
Translated to English: Unfortunately, I cannot answer questions or identify speakers from a voice sample.

Figure 2: Examples of ASR outputs from unimodal and multimodal models.

African languages like Swahili and Hausa, each spoken by over 50m people across 4+ countries. Our results reinforce the need for targeted improvements, as advances in ASR have yet to close the performance gap for African languages.

5.7 Noise & Environment Robustness

Across all datasets, models performed worst on the parliamentary proceedings dataset, despite containing accents present in other datasets. This suggests that the primary challenge was not linguistic variation but rather the presence of background noise and overlapping speech, which were mostly absent in the other datasets. Notably, unimodal ASR models maintained a lower WER in these conditions, while multimodal models like Gpt-4o-audio-preview exhibited significant performance degradation. The resilience of Gemini 2.0 Flash in this setting is noteworthy, as it remains competitive with ASR models despite being a multimodal model.

5.8 Unimodal vs. Multimodal Model AST Performance

Our evaluation highlights a significant performance gap between traditional unimodal models and modern multimodal models, particularly in handling African languages. Unimodal models like Whisper often struggle with these languages, frequently producing incoherent or untranslated outputs (See Table 7). For instance, Whisper Large-v3 consistently yields very low BLEU and CHRf scores across several languages, indicating minimal overlap with the reference translations and poor semantic capture.

In contrast, multimodal models demonstrate markedly better performance, especially on low-resource languages. Models such as Google's Gemini-2.0 (flash) achieve substantially higher scores, showing a clear advantage over Whisper in both Yoruba and Hausa, among others (See Table 7). Even multimodal models that are not the top performers—like Meta's SeamlessM4T (Large-v2)—outperform unimodal baselines across the board. Notably, SeamlessM4T performs competitively despite being trained on less data than Gemini or GPT-4. On higher-resource languages such as French and Arabic, its scores closely match those of larger models, and on low-resource languages like Shona, it often outperforms them. These results demonstrate that multimodal training significantly enhances translation quality, allowing models to generalize better and provide more accurate outputs even with limited language-specific data.

5.9 Impact of In-Domain Fine-Tuning on Qwen2.5-Omni

Fine-tuning Qwen-2.5 Omni on a subset of the NaijaVoices corpus yields dramatic improvements in both WER and translation quality (Table 7). Igbo WER plunges from 198 to 42 (-79%), Hausa from 127 to 51 (-60%), and Yoruba from 121 to 71 (-41%), while AfriComet-STL for those languages nearly triples (Igbo 0.18 $\rightarrow$ 0.54, Hausa 0.19 $\rightarrow$ 0.39, Yoruba 0.20 $\rightarrow$ 0.29), as seen in Table 6. These gains indicate that even modest, language-specific data can unlock large pretrained models' latent capacity for under-represented languages.

Table 6: Qwen-Omni2 ASR (WER score) and AST (AfriComet-STL) Performance Before and After Fine-Tuning

Language	ASR (WER)		AST (STL)	
	Base	Finetuned	Base	Finetuned
Hausa	127	**51**	0.19	**0.39**
Igbo	198	**42**	0.18	**0.54**
Yoruba	121	**71**	0.20	**0.29**

5.10 ASR Failures

Our evaluation revealed several common transcription failure modes across models. A primary issue was *phonetic confusions*, where accent variation led models to misinterpret spoken words, resulting in erroneous transcriptions. This was especially prevalent in non-standard pronunciations. We also observed *hallucinations*, notably in Whisper and Canary models, where silent segments were filled with repetitive or unrelated text, inflating WER scores. Additionally, Whisper models occasionally exhibited *skipped segments*, omitting significant portions at the beginning of audio clips—likely a result of internal heuristics ignoring initial speech.

The large multimodal models (Gpt-4o-audio & Gemini) sometimes introduce *contextual errors*, such as inserting additional phrases or paraphrasing content, which diverges from strict transcription standards. Furthermore, other models (e.g., Canary 1B) expanded acronyms (e.g., "HIV" as "human immunodeficiency virus"), which conflicted with domain conventions where abbreviations are standard, artificially increasing WER. Lastly, GPT-4o-Preview frequently failed to transcribe short samples—particularly from the NCHLT dataset—responding with messages indicating an inability to transcribe the content.

Possible Benchmark Contamination Issues: NCHLT and Common-Voice were released several years ago (old). Afrispeech and the private parliamentary proceedings are more recent (new). The 2-7x gap in performance of unimodal and multimodal models on the new vs old data suggests that model exposure to old datasets may convey a false sense of generalizability that new datasets expose. All models perform worst on the noisy challenging parliamentary dataset suggesting limitations with their use in real-world settings. This underscores the value of newer and more representative benchmarks in the speech domain.

5.11 AST Failures

Contextual Miss-Translation by Multimodal Models: In contrast to Whisper, the multimodal models produced meaning translations. Models like GPT-4 (audio), Gemini-2.0, and SeamlessM4T generally succeeded in translating entire sentences from the audio, even for more low-resource languages like (Ga) in contrast to Whisper. This highlights the multimodal models' strength in handling sentence-length context – they rarely got "stuck" partway through a translation. When errors did occur in the multimodal outputs, the problem was omitting or mistranslating important words. A common issue was the selection of an incorrect synonym or a phrase that slightly shifted the nuance of the source. This led to translations with significant information gaps. Such substitutions can affect fidelity – the translation is understandable and contextually plausible, but not exactly what a human translator would pick. Despite this, these errors are relatively minor compared to the complete failures seen in unimodal outputs. The higher AfriComet and CHRf scores for multimodal models (Table 7 & Appendix 18) support this: even if BLEU penalizes synonym mismatches, the character n-gram overlap remains high, indicating that translations captured most of the content. Overall, the multimodal systems demonstrated far better sentence-level translation quality, preserving context and structure, with errors generally confined to fine-grained lexical nuances.

Hallucination Patterns: Omission vs. Speculative Completions: While multimodal models like SeamlessM4T and Gemini reduce random errors compared to unimodal models, they are not entirely free from hallucinations. A notable issue we observed is over-generation—the model adds contextually relevant but unspoken content. For example, when a Yoruba speaker poses a question, SeamlessM4T might translate the question into English and then generate a plausible answer (see fig 4, even though none was provided. This suggests the model is attempting to be helpful or complete the conversation, behaving more like a dialogue agent than a strict translator.

This differs from hallucinations in unimodal models like Whisper, which tend to produce unrelated or nonsensical outputs (fig. 4) (Koenecke et al., 2024). In contrast, multimodal hallucinations often feel coherent and related, making them more subtle yet still problematic, as they introduce information not present in the original speech. These behaviors may originate from exposure to instruction-tuned or conversational training data. As such models are deployed in real-world translation tasks, it's critical to identify and correct these tendencies- users need accurate translations, not the model's assumptions or commentary.

Limited Robustness to Heavily Noisy Inputs: All models, regardless of architecture, showed different range of robustness when faced with very noisy or challenging audio. In our tests with overlapping speakers, background chatter, or poor audio quality, translation performance degraded substantially across the board. Often, the models would fail to disentangle multiple speakers or filter out noise, resulting in jumbled output. A common

Table 7: AfriComet-STL scores across the languages for each model. "–" means the models doesn't support the language. The the higlighted scores are the best score per language

Language	Canary 1b	Whisper medium	Whisper large-v3	Qwen2.5	SeamlessM4T Large-v2	Gpt-4o audio-preview	Gemini-2.0 flash
Afrikaans	-	0.57	0.65	-	0.73	0.71	**0.80**
Akan	-	-	-	-	-	0.34	**0.38**
Amharic	-	0.23	0.27	-	0.64	0.42	**0.79**
Arabic	-	0.65	0.70	-	0.80	0.81	**0.85**
French	0.65	0.70	0.73	**0.8**	0.79	0.78	**0.80**
Fulani	-	-	-	-	0.19	0.30	**0.35**
Ga	-	-	-	-	-	0.24	**0.29**
Hausa	-	0.16	0.19	-	0.17	0.37	**0.65**
Igbo	-	-	-	-	0.25	0.29	**0.37**
Kinyarwanda	-	-	-	-	-	0.29	**0.54**
Luganda	-	-	-	-	0.57	0.47	**0.59**
Pedi	-	-	-	-	-	0.31	**0.39**
Sesotho	-	-	-	-	0.23	0.35	**0.50**
Shona	-	0.18	0.21	-	**0.73**	0.47	0.61
Swahili	-	0.32	0.42	-	-	0.76	**0.81**
Tswana	-	-	-	-	**0.56**	0.32	0.46
Twi	-	-	-	-	**0.41**	0.33	0.32
Xhosa	-	-	-	-	-	0.35	**0.66**
Yoruba	-	0.18	0.20	-	-	0.36	**0.49**
Zulu	-	-	-	-	-	0.40	**0.71**

Example 1: Altered meaning

Reference: be careful not to allow fabric to become too hot which can cause shrinkage or in extreme cases scorch

SeamlessM4T-v2: be careful not to overheat the cloth which can cause itching or burn if it is to thick

Example 2: Altered meaning

Reference: on 15 august 1940 the allies invaded southern france the invasion was called operation dragoon

Whisper L.: name of the operation was given to the king in 1940 and was first introduced in southern france it was later called operation dragon

Example 3: Noisy samples

Gpt-4o-audio: I'm sorry, I cannot identify speakers

Figure 3: Examples of AST outputs from unimodal and multimodal models.

Example 1: Hallucination

Reference: Go and forgive your father

SeamlessM4T–v2: i m not going to be able to do it

Example 2: Token Repitition

Reference: the three kingdoms was one of the bloodiest eras in ancient china's history thousands of people died fighting to sit in the highest seat in the grand palace at xi'an

Whisper L: hello my name is meta i am from okanlala i am from kokor i am from the village of dáiwa and the next day the next day the next day the next dáy the next dáy the next dáy the next day the next day

Figure 4: Examples of AST outputs from unimodal and multimodal models.

failure mode under heavy noise was partial transcription without translation for Gemini and Gpt-4o audio. Other times, Whisper and GPT-4 audio, would latch onto a few words they could recognize and simply repeat them or present them in the original language, rather than translating. These issues show that while our models perform well on clean, single-speaker audio, real-world conditions with noise or speaker overlap remain very challenging. Improving noise robustness – perhaps via data augmentation (Puvvada et al., 2024)– is an important direction.

Limitations

One key limitation of this study lies in the use of pre-trained models without any fine-tuning or adaptation to the African linguistic context. While this approach allowed for consistent benchmarking across systems, it may have disadvantaged models that require domain-specific calibration to perform optimally in low-resource or accented speech settings. In our future work, we plan to fine-tune speech models to improve their performance on African languages and African-accented speech. Additionally, reliance on older benchmark datasets such as NCHLT and Common Voice raises concerns about possible benchmark contamination, as these datasets may have been included in the pretraining corpus of some models. This could lead to inflated performance estimates and reduce confidence in the models' generalizability to newer, more representative data. Furthermore, the evaluation employed a uniform prompting strategy across all languages and models, using simple instructions like "Transcribe this audio." While this ensured comparability, it may have constrained the performance of models that rely on task-specific or few-shot prompting strategies to fully leverage their multimodal or contextual capabilities.

References

Jesujoba O Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages. *arXiv preprint arXiv:2409.20201*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The nchlt speech corpus of the south african languages. In *4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 194–200.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2024. Large language models are strong audio-visual speech recognition learners. *arXiv preprint arXiv:2409.12319*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Pavel Denisov and Ngoc Thang Vu. 2024. Teaching a multilingual large language model to understand multilingual speech via multi-instructional training. *arXiv preprint arXiv:2404.10922*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Sewade Ogun, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Bunmi Akinremi, David Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2022. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. *Computer Speech & Language*, 72:101278.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14:1–14.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Yerin Kwak and Zachary A Pardos. 2024. Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*.

Léa-Marie Lam-Yee-Mui, Waad Ben Kheder, V. Le, C. Barras, and J. Gauvain. 2023. Multilingual models with language embeddings for low-resource speech recognition. *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*.

Danni Liu and Jan Niehues. 2024. Recent highlights in multilingual and multimodal speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 235–253.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, and 1 others. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*.

Morakinyo Ogunmodimu. 2015. Language policy in nigeria: Problems, prospects and perspectives. *International Journal of Humanities and Social Science*, 5(9):154–160.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. In *Interspeech 2024*, pages 3964–3968.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra D. Kayande, Emmanuel Ayodele, Naome A. Etori, Michael S. Mollel, Moshood Yekini, Chibuzor Okocha, Lukman E. Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. 2025a. Afrispeech-dialog: A benchmark dataset for spontaneous english conversations in healthcare and beyond. *Preprint*, arXiv:2502.03945.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra D. Kayande, Emmanuel Ayodele, Naome A. Etori, Michael S. Mollel, Moshood Yekini, Chibuzor Okocha, Lukman E. Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. 2025b. Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. *arXiv preprint*. ArXiv:2502.03945 [cs].

Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Soumya Saha, Dan Zhang, and Ashwinkumar Ganesan. 2023.

Improving low resource speech translation with data augmentation and ensemble strategies. pages 241–250.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2014. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language – Amharic. *Speech Communication*, 56:181–194.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, and 1 others. 2023. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *arXiv preprint arXiv:2311.09828*.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024. Just asr+ llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A speech understanding and generation llm in a codec-free full-duplex framework.

A Appendix

A.1 Automatic Speech Recognition

A.1.1 ASR Prompts

For automatic speech recognition (ASR), we evaluate three prompting strategies. The first employs a simple instruction: "Transcribe this audio." The second includes language specificity: "Transcribe the entire audio in {source_language}." The third is a few-shot variant of the second prompt, which provides two audio-transcription exemplars as demonstrations to guide the model's output.

A.2 Automatic Speech Translation

A.2.1 AST Prompting Strategies

We evaluate three AST prompting strategies:

1. **Zero-shot translation:**
   "Given audio in `{source_language}`*, translate to English."*

2. **Zero-shot transcriptiontranslation:**
   "Given audio in `{source_language}`*, first transcribe the speech, then translate the transcript into English."*

3. **Few-shot variants:**
   For each of the above prompts, we prepend two example audio–translation pairs to provide in-context demonstrations of the desired behavior.

We found the Zero-shot transcriptiontranslation gives the best result as it encourages the model to understand the audio by first transcribing, before attempting to translate.

A.2.2 Fleurs dataset

Table 8: WER scores for each models per language

Language	Canary-1b	Whisper medium	Whisper large-v3	MMS-1b all	Qwen2.5	Seamless-M4T Large-v2	Gpt-4o audio-preview	Gemini-2.0 flash
English (M. Lib)	3.03	6.80	3.53	17.63	16.32	4.68	9.63	6.63
French (M. Lib)	4.06	8.90	5.38	19.30	10.43	6.82	22.71	5.23
Spanish (M. Lib)	-	-	-	17.35	-	6.76	21.25	**3.22**
Afrikaans	-	68.87	45.43	48.73	-	18.41	84.36	18.02
Akan	-	-	-	62.92	-	-	104.02	67.04
Amharic	-	447.26	165.83	67.52	-	44.05	245.4	55.88
Arabic	-	39.49	29.72	44.94	-	51.26	31.88	14.44
French	9.67	13.95	9.31	33.93	24.14	15.90	22.29	9.12
Fulani	-	-	-	56.78	-	86.85	157.03	66.11
Ga	-	-	-	-	-	-	172.73	87.27
Hausa	-	180.29	95.11	40.47	-	-	118.60	38.48
Igbo	-	-	-	50.33	-	70.03	112.23	66.68
Kinyarwanda	-	-	-	36.73	-	-	135.75	58.44
Luganda	-	-	-	28.85	-	16.39	131.19	59.89
Pedi	-	-	-	41.43	-	-	119.29	70.69
Sesotho	-	-	-	-	-	-	158.21	59.30
Shona	-	193.21	110.35	30.7	-	76.05	90.51	38.84
Swahili	-	117.7	62.75	28.37	-	16.25	73.96	25.88
Tswana	-	-	-	-	-	-	133.46	54.85
Twi	-	-	-	51.09	-	-	98.86	67.13
Xhosa	-	-	-	42.24	-	-	130.79	39.32
Yoruba	-	213.88	93.77	39.59	-	37.43	101.14	43.42
Zulu	-	-	-	43.19	-	52.53	135.84	30.02

Table 9: FLEURS performance across models by language

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash
Afrikaans	0.44	0.31	0.26	0.19	0.32	0.14
Amharic	4.42	2.06	0.35	0.86	1.18	0.19
Arabic	–	0.11	0.36	0.09	0.07	0.04
Fulani	–	–	0.57	–	1.57	0.75
Hausa	1.58	0.86	0.31	–	1.01	0.35
Igbo	–	–	0.45	1.03	1.11	0.66
Luganda	–	–	0.46	0.38	0.89	0.53
Pedi	–	–	0.31	–	1.10	0.90
Shona	2.22	1.17	0.30	0.76	0.97	0.54
Swahili	0.99	0.42	0.22	0.12	0.30	0.12
Xhosa	–	–	0.45	–	1.25	0.57
Yoruba	2.04	0.87	0.34	0.31	0.83	0.42
Zulu	–	–	0.40	0.51	1.11	0.32

Table 10: Word Error Rate (WER) scores for each model for the Intron-medical dataset. "–" indicates unsupported languages.

Language	Canary 1b	Whisper medium	Whisper large-v3	MMS-1b all	Qwen2.5	SeamlessM4T-v2 Large	GPT-4o audio-preview	Gemini-2.0 flash
Afrikaans	–	0.53	0.33	0.37	–	0.15	0.47	0.18
Akan	–	–	–	0.63	–	–	1.04	0.77
Arabic	–	0.46	0.33	0.75	–	–	0.33	0.24
French	0.13	0.16	0.11	0.42	0.24	0.17	0.12	0.08
Hausa	–	1.30	0.94	0.43	–	–	1.26	0.40
Igbo	–	–	–	0.54	–	0.69	1.04	0.77
Kinyarwanda	–	–	–	0.47	–	–	1.34	0.65
Pedi	–	–	–	0.47	–	–	1.24	0.77
Sesotho	–	–	–	–	–	–	1.73	0.78
Shona	–	1.50	1.01	0.32	–	0.75	0.80	0.45
Swahili	–	1.12	0.48	0.34	–	0.19	0.43	0.16
Tswana	–	–	–	–	–	–	1.36	0.73
Twi	–	–	–	0.51	–	–	1.03	0.81
Xhosa	–	–	–	0.44	–	–	1.23	0.47
Yoruba	–	1.57	0.89	0.43	–	0.30	1.35	0.54
Zulu	–	–	–	0.48	–	0.52	1.29	0.35

Table 11: ALFFA performance across models by language

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash
Amharic	4.28	1.56	0.76	0.24	2.80	2.80
Swahili	1.33	0.73	0.41	0.26	0.94	0.94

Table 12: Ashesi Financial Inclusion performance across models by language

Language	MMS-1b all	Gpt-4o audio-preview	Gemini-2.0 flash
Akan	0.78	1.33	0.94
Ga	–	1.73	1.15
Twi	0.75	1.84	1.50

Table 13: Common Voice performance across models by language

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o-audio-preview	Gemini-2.0 flash
Afrikaans	0.52	0.38	0.27	0.14	0.57	0.18
Amharic	5.14	1.83	0.53	0.93	1.84	1.30
Arabic	0.36	0.18	0.28	0.68	0.32	0.12
Hausa	2.70	0.91	0.27	–	1.09	0.41
Igbo	–	–	0.61	0.43	2.46	0.82
Kinyarwanda	–	–	0.33	–	1.36	0.84
Luganda	–	–	0.29	0.16	1.32	0.81
Swahili	1.21	0.71	0.25	0.14	0.92	0.26
Twi	–	–	0.58	–	1.23	0.93
Yoruba	2.94	0.99	0.39	0.40	0.96	1.04

Table 14: NCHLT performance across models by language

Language	Whisper-medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash
Afrikaans	0.99	0.68	0.71	0.25	1.52	0.49
Pedi	–	–	0.42	–	1.19	0.91
Sesotho	–	–	–	–	1.33	1.04
Tswana	–	–	–	–	1.28	0.85
Xhosa	–	–	0.32	–	1.71	0.57
Zulu	–	–	0.28	0.56	2.08	0.45

Table 15: NaijaVoices performance across models by language

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash
Hausa	1.86	0.97	0.39	–	1.20	0.52
Igbo	–	–	0.49	0.66	1.18	0.87
Yoruba	2.13	0.98	0.44	0.45	1.07	0.78

Language	**Metric**	**Fluency** r	**Adequacy** r
Akan	BLEU	–0.09	0.58
	ChrF	–0.24	0.68
	AfriComet-STL	0.07	0.61
Igbo	BLEU	0.10	0.63
	ChrF	–0.11	0.69
	AfriComet-STL	–0.04	0.93
Pedi	BLEU	0.05	0.78
	ChrF	0.26	0.68
	AfriComet-STL	0.38	0.61
Shona	BLEU	0.38	0.44
	ChrF	0.48	0.73
	AfriComet-STL	0.67	0.86
Swahili	BLEU	0.43	0.47
	ChrF	0.56	0.70
	AfriComet-STL	0.67	0.76
Twi	BLEU	0.43	0.34
	ChrF	0.44	0.36
	AfriComet-STL	0.52	0.60
Yoruba	BLEU	0.30	0.61
	ChrF	0.40	0.76
	AfriComet-STL	0.47	0.70

Table 16: Pearson correlations (r) between automatic metrics and human evaluations of fluency and adequacy.

Table 17: BLEU performance across models by language

Language	Canary 1b	Whisper medium	Whisper large-v3	Qwen2.5	SeamlessM4T Large-v2	Gpt-4o audio-preview	Gemini-2.0 flash
Afrikaans	–	19.39	23.2	–	27.62	31.59	38.76
Akan	–	–	–	–	–	2.44	5.15
Amharic	–	0.8	0.71	–	15.61	4.2	24.88
Arabic	–	17.97	20.34	–	27.69	31.06	34.68
French	24.46	27.39	28.92	41.40	33.38	41.27	43.57
Fulani	–	–	–	–	0.58	1.05	2.41
Ga	–	–	–	–	–	0.49	1.06
Hausa	–	0.71	0.71	–	0.31	6.23	21.06
Igbo	–	–	–	–	1.92	2.97	5.82
Kinyarwanda	–	–	–	–	–	1.99	10.91
Luganda	–	–	–	–	15.97	7.77	13.79
Pedi	–	–	–	–	–	3.19	6.34
Sesotho	–	–	–	–	–	4.11	11.23
Shona	–	0.4	0.52	–	2.11	6.78	12.56
Swahili	–	2.84	5.47	–	23.27	26.78	32.62
Tswana	–	–	–	–	–	3.72	9.59
Twi	–	–	–	–	–	2.83	2.48
Xhosa	–	–	–	–	-	4.71	19.9
Yoruba	–	0.24	0.37	–	14.39	4.89	11.77
Zulu	–	–	–	–	8.17	6.57	22.9

Table 18: CHrF performance across models by language

Language	Gemini-2.0 flash	GPT-4o audio-preview	SeamlessM4T-v2 Large	Whisper Large	Whisper Medium	Canary-1b	Qwen2.5
Afrikaans	64.33	56.39	–	–	–	–	–
Akan	29.86	25.01	56.13	–	–	–	–
Amharic	56.62	29.62	–	50.33	45.58	–	–
Arabic	63.10	59.26	43.48	17.06	13.57	–	–
French	66.56	64.40	55.53	47.85	44.38	54.12	64.94
Fulani	27.56	23.82	63.72	58.61	57.19	–	–
Ga	20.08	19.09	16.25	–	–	–	–
Hausa	48.48	29.81	–	–	–	–	–
Igbo	32.10	25.40	13.47	13.29	7.78	–	–
Kinyarwanda	37.69	23.62	18.52	–	–	–	–
Luganda	44.23	35.56	44.21	–	–	–	–
Pedi	34.63	27.51	–	–	–	–	–
Sesotho	38.00	26.71	–	–	–	–	–
Shona	42.07	33.56	21.65	15.59	12.76	–	–
Swahili	61.74	55.90	53.39	30.00	22.13	–	–
Tswana	35.52	25.11	–	–	–	–	–
Twi	24.22	23.15	–	–	–	–	–
Xhosa	48.82	28.54	40.53	14.29	10.45	–	–
Yoruba	38.45	28.37	–	–	–	–	–
Zulu	52.76	31.54	32.79	–	–	–	–

Table 19: Multi-metric performance across models for FLEURS

Language	Gemini-2.0 flash		GPT-4o audio-preview		SeamlessM4T-v2 Large		Whisper Large		Whisper Medium	
	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF	BLEU	ChrF
Amharic	29.44	62.09	5.60	33.25	21.24	50.16	1.20	19.06	1.08	16.30
Arabic	33.25	66.44	30.66	63.85	33.86	62.88	18.83	50.45	18.07	48.54
Fulani	2.41	27.56	1.05	23.82	0.58	16.25	–	–	–	–
Hausa	17.68	50.09	6.07	34.25	0.48	16.79	0.16	15.18	0.22	10.13
Igbo	5.54	34.91	2.48	27.37	1.17	17.99	–	–	–	–
Luganda	13.79	44.23	7.77	35.56	15.97	44.21	–	–	–	–
Pedi	6.30	36.41	2.95	28.84	–	–	–	–	–	–
Shona	12.20	43.54	6.15	34.43	2.67	25.44	0.79	17.46	0.55	14.62
Swahili	30.70	62.10	23.89	55.24	28.41	57.03	4.48	29.04	2.54	20.40
Xhosa	20.09	51.51	4.19	29.77	–	–	–	–	–	–
Yoruba	10.21	40.15	4.23	30.70	13.25	41.04	0.62	16.73	0.41	12.20
Zulu	21.54	53.45	5.86	33.00	7.67	34.19	–	–	–	–

Table 20: Multi-metric performance across models for Intron (part 1: Gemini–Whisper Medium).

Language	Gemini-2.0 flash		GPT-4o audio-preview		SeamlessM4T-v2 Large		Whisper Large		Whisper Medium	
	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF
Afrikaans	38.76	64.33	31.59	56.39	27.62	56.13	23.20	50.33	19.39	45.58
Akan	5.15	29.86	2.44	25.01	–	–	–	–	–	–
Amharic	16.45	45.29	1.39	22.12	6.07	29.50	0.12	13.29	0.31	7.98
Arabic	24.75	55.28	21.98	52.07	15.99	44.95	13.55	41.54	10.78	36.94
French	32.49	60.96	28.99	57.45	20.07	50.06	23.95	53.37	21.31	51.01
Ga	1.06	20.08	0.49	19.09	–	–	–	–	–	–
Hausa	23.18	48.70	6.48	28.76	0.19	11.88	0.16	12.52	0.15	6.34
Igbo	5.69	29.50	2.99	23.62	2.05	17.18	–	–	–	–
Kinyarwanda	10.91	37.69	1.99	23.62	–	–	–	–	–	–
Pedi	6.40	31.04	3.61	24.81	–	–	–	–	–	–
Sesotho	11.23	38.00	4.11	26.71	–	–	–	–	–	–
Shona	12.98	40.15	7.55	32.42	1.15	16.26	0.23	13.34	0.25	10.40
Swahili	30.45	58.71	23.52	51.43	19.82	49.07	6.51	30.33	4.00	21.80
Tswana	9.59	35.52	3.72	25.11	–	–	–	–	–	–
Twi	2.48	24.22	2.83	23.15	–	–	–	–	–	–
Xhosa	19.76	46.48	5.11	27.47	–	–	–	–	–	–
Yoruba	14.37	39.68	5.61	27.77	14.01	40.44	0.11	12.72	0.08	8.35
Zulu	24.01	52.14	7.17	30.20	8.60	31.48	–	–	–	–

Table 20: (continued) Multi-metric performance across models for Intron (part2: Canary1b & Qwen).

Language	Canary1b		Qwen2.5	
	BLEU	CHrF	BLEU	CHrF
French	13.78	44.46	41.40	64.94

Table 21: Multi-metric performance across select models by NaijaVoices

Language	Gemini		GPT-4o-audio preview		SeamlessM4T v2 Large		Whisper Large		Whisper Medium	
	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF
Hausa	19.15	44.84	5.61	25.34	0.17	12.69	0.17	12.52	0.11	8.29
Igbo	6.97	28.67	4.35	22.91	4.22	22.80	–	–	–	–
Yoruba	9.92	32.57	4.88	24.32	16.34	39.61	0.11	11.52	0.11	10.33

Table 22: Multi-metric performance across models by IWSLT_LRST

Language	Gemini-2.0 flash		GPT-4o audio-preview		SeamlessM4T-v2 Large		Whisper Large		Whisper Medium	
	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF
Swahili	37.22	65.60	33.74	62.25	25.15	57.15	4.32	30.09	1.68	23.38

Table 23: Multi-metric performance across models by Covost (part 1: Gemini–Whisper Medium).

Language	Gemini-2.0 flash		GPT-4o audio-preview		SeamlessM4T-v2 Large		Whisper Large		Whisper Medium	
	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF	BLEU	CHrF
Arabic	51.72	70.78	45.97	64.50	37.07	62.11	30.92	54.18	28.03	50.48
French	44.40	66.91	42.19	64.83	34.35	64.56	29.32	58.98	27.84	57.57

Table 23: (continued) Multi-metric performance across models by Covost (part 2: Canary-1b & QWEN).

Language	Canary-1b		QWEN	
	BLEU	CHrF	BLEU	CHrF
French	25.03	54.72	41.40	64.94