AfriVox: Probing Multilingual and Accent Robustness of Speech LLMs

Busayo Awobade[1,9], Mardhiyah Sanni[1,7,9], Tassallah Abdullahi[2,9], Chibuzor Okocha[3,9], Kelechi Ezema[4,9], Devendra Deepak Kayande[5,9], Lukman E. Ismaila[6,9], Tobi Olatunji [1,8,9], Gloria Ashiya Katuka[9]

[1]Intron, [2]Brown University, [3]University of Florida, [4]University of Colorado Boulder,
[5]IIIT Allahabad, [6]John Hopkins University,
[7]University of Edinburgh, [8]Georgia Tech, [9]BioRAMP.

busayo@intron.io, tobi@intron.io

Abstract

Recent advances in multimodal and speech-native large language models (LLMs) have delivered impressive speech recognition, translation, understanding, and question-answering capabilities for high-resource languages. However, African languages and non-native French or English accents remain dramatically underrepresented in benchmarks limiting the understanding and applicability of leading LLMs for millions of francophone and anglophone users in low-resource settings. We present AfriVox, an open-source benchmark (including novel domain-specific and unscripted datasets) across 20 African languages, African-accented French, Arabic, and 100+ African English accents, contrasting leading multimodal speech LLMs with traditional unimodal automatic speech transcription (ASR) and translation (AST) models. Our analysis reveals significant language coverage variation, surprising LLM translation performance gains (e.g. Gemini), robustness concerns with unscripted speech, and substantial performance disparities for "supported" African languages. We profile the strengths, limitations, and language support of each model, and conduct the first targeted fine-tuning of a modern speech LLM (Qwen2.5-Omni) for three Nigerian languages, exceeding SOTA, and achieving up to 54% relative WER reduction and significant BLEU gains, offering practical guidance for implementers seeking to serve local language users.

1 Introduction

The transformative impact of LLMs in global technology—especially speech-enabled and multimodal LLMs—has opened new frontiers for human-computer interaction (AlSaad et al., 2024). Major recent breakthroughs, such as OpenAI's GPT-4o (Hurst et al., 2024), Google Gemini (Google DeepMind, 2024), and Meta's SeamlessM4T (Barrault et al., 2023), have enabled voice-based applications that promise to make information and services more accessible, especially in regions where text literacy and high-resource language proficiency may be limiting factors (Peng et al., 2025).

Across Africa, LLM-powered systems are already being deployed in sectors like health, agriculture, and financial inclusion, operating in large languages via text interfaces (Olatunji et al., 2023; Nazi and Peng, 2024; Al-Garadi et al., 2025). However, as voice-native and multilingual LLMs have rapidly improved (Bai et al., 2024; Google DeepMind, 2024), technology implementers across Africa are eager to shift towards more natural, relatable, and intuitive speech-driven interfaces that truly reflect users' language preferences and linguistic diversity (Sanni et al., 2025).

Despite this demand, no comprehensive benchmark exists that systematically evaluates modern speech LLMs on African languages and accents (Adelani et al., 2025; Ojo et al., 2025). Existing benchmarks such as MLS, mSTEB, NaijaVoices, and ML-SUPERB 2.0 include very limited African language coverage and lack recent domain-specific, real-world unscripted speech, especially for emerging LLM architectures (Pratap et al., 2020a; Beyene et al., 2025; Emezue et al., 2025; Shi et al., 2024). Most performance claims are based on high-resource languages, providing little actionable guidance to African technology teams deciding whether to trust LLMs for local deployment (Reid et al., 2021).

To bridge this gap, we introduce AfriVox, a unified benchmark suite aggregating and extending multiple African speech datasets and releasing two novel datasets under a CC-BY-NC-SA license including (1) **Afrispeech-Parliamentary**–parliamentary speech from 4 countries, and (2) **Afrivox-Medical**–health-focused conversations in 20 African languages. Additionally, we conduct the first systematic, reproducible evaluation of state-of-the-art speech LLMs and unimodal models across

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2672–2690
March 24-29, 2026 ©2026 Association for Computational Linguistics

20 languages and 100+ English, French, and Arabic accents.

We use AfriVox to answer two critical questions for implementers: (1) Which speech LLMs reliably support certain African languages and (2) How do leading multimodal LLMs compare with traditional leading ASR/AST models on understanding realistic African speech? Should implementers switch from unimodal ASR models to LLMs? Additionally, we provide detailed error analysis and practical guidance, including fine-tuning experiments with Qwen2.5-Omni on major Nigerian languages using only moderate data.

2 Related works

Recent years have seen remarkable progress in speech and multimodal large language models (LLMs) (Yu et al.), driven by advances in self-supervised learning, scaling laws, and reinforcement learning techniques (Ghosh et al., 2024). However, these improvements have disproportionately benefited high-resource languages, with African languages still underrepresented in both model training and evaluation (Adelani et al., 2025; Ojo et al., 2025).

Multilingual Speech Benchmarks: Benchmarks such as MLS (Multilingual LibriSpeech) (Pratap et al., 2020a), mSTEB (Beyene et al., 2025), and ML-SUPERB 2.0 (Shi et al., 2024) have provided valuable evaluation resources, but offer limited coverage of African languages, and their data is primarily read speech or synthetic in nature. ML-SUPERB 2.0 and mSTEB in particular have improved multilingual evaluation rigor, yet it covers only a handful of African languages and lacks representation of diverse accents and real-world conversational domains (Pratap et al., 2020a). Our benchmark, AfriVox, addresses these gaps by including (a) a broader and more granular set of African languages and accents, (b) domain-specific, real-world audio (e.g., parliamentary sessions, healthcare dialogues), and (c) explicit evaluation of both unimodal and state-of-the-art multimodal LLMs.

Speech and Multimodal LLMs: Large-scale unimodal models such as Whisper (Radford et al., 2023a), MMS (Denisov and Vu, 2024), and Parakeet (Galvez et al., 2024) have demonstrated robust speech recognition performance in high-resource settings, but their reliability in African language tasks remains largely anecdotal (Ojo et al., 2025). Recent multimodal models—including Google AudioPaLM (Rubenstein et al., 2023), Meta SeamlessM4T (Barrault et al., 2023), Qwen-Audio (Chu et al., 2024), and Gemini (Google DeepMind, 2024)—promise to unify speech, text, and translation tasks, but have yet to be systematically benchmarked on African data (Adelani et al., 2025).

Parameter-Efficient Fine-Tuning (PEFT): Scaling LLMs for downstream tasks in low-resource settings can be prohibitively expensive. PEFT approaches such as LoRA (Karimi Mahabadi et al., 2021), Adapters (Han et al., 2024), and QLoRA (Dettmers et al., 2023) enable practical model adaptation by training only a small subset of parameters. However, most prior studies have focused on high-resource or Asian languages (Bai et al., 2024); little is known about their impact on speech LLMs for African contexts (Emezue et al., 2025).

African Speech Datasets: Public African speech corpora including NCHLT (Barnard et al., 2014), CommonVoice (Ardila et al., 2020), and FLEURS (Conneau et al., 2023) have played a vital role, but coverage, accent diversity, and domain relevance remain limited. Recent datasets such as AfriSpeech (Olatunji et al., 2025) and NaijaVoices (Emezue et al., 2025) have begun to address these challenges. Our work builds on and expands these efforts, contributing new datasets and a unified benchmark for comprehensive, reproducible evaluation.

To our knowledge, Afrivox is the first to aggregate and compare both unimodal and multimodal speech LLMs across 20+ African languages and 100+ English accents providing practical, data-driven guidance for implementers on the suitability of LLMs vs. traditional ASR for local deployment

3 Methodology

Dataset	Hours	Speakers	Accents
NCHLT	2.24	8	1
AfriSpeech-200	18.68	750	108
CV-17 En-Afr	0.11	46	9
Afrispeech-Parl	42.17	∼1651	4
Total	**63.20**	**∼2455**	**108**

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

Dataset	# 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	2.98	200
FISD[2]	3	0.05	23
AfriVox-Medical [3]	19	36.63	1179
	Total Hours	**81.5**	

Table 2: Summary of multilingual speech datasets.

Language	Region	Language Family	# Speakers
Afrikaans	South	IndoWest (Germanic)	7.2M
Akan	West	Niger-Congo (Kwa)	24M
Amharic	East	Afro-Asiatic (Semitic)	35M
Egyptian Arabic	North	Afro-Asiatic (Semitic)	78M
French	West	Indo-European (Romance)	320M
Fula	West	Niger-Congo (Atlantic)	36.8M
Gaa	West	Niger-Congo (Kwa)	0.7M
Hausa	West	Afro-Asiatic (Chadic)	54M
Ibo	West	Niger-Congo (Volta-Niger)	31M
Kinyarwanda	East	Niger-Congo (Bantu)	15M
Luganda	East	Niger-Congo (Bantu)	5.6M
Northern Sotho	South	Niger-Congo (Bantu)	4.6M
Shona	South	Niger-Congo (Bantu)	8.4M
Southern Sotho	South	Niger-Congo (Bantu)	5.6M
Swahili	East	Niger-Congo (Bantu)	87M
Tswana	South	Niger-Congo (Bantu)	8.2M
Twi	West	Niger-Congo (Kwa)	4.4M
Xhosa	South	Niger-Congo (Bantu)	8M
Yoruba	West	Niger-Congo (Yoruboid)	45M
Zulu	South	Niger-Congo (Bantu)	13.6M

Table 3: Multilingual Speech Datasets: Language, region, family, and number of speakers.

3.1 Benchmark Design and Datasets

We design the AfriVox benchmark to evaluate speech LLMs and ASR/AST models on realistic African language and accent use-cases. This benchmark unifies and expands existing corpora, incorporating both new and public datasets to maximize coverage and relevance.

3.1.1 African-Accented English Speech (AES)

Afrivox-AES combines NCHLT (Barnard et al., 2014), AfriSpeech-200 (Olatunji et al., 2025), Common Voice 17 (Ardila et al., 2020)(filtered for African accents), and a newly-curated AfriSpeech-Parl dataset [4] comprised of human transcribed Parliamentary Proceedings (publicly available) from 4 African countries (Ghana, Kenya, Nigeria, and South Africa). Speech was transcribed by native speakers. AES is over 63 hours, with 2,000+ speakers from 12 countries, and 108 distinct African English accents (Table 1).

Curation and Quality Control: Common Voice was filtered for African accents using available speaker metadata and manual accent validation. Transcribed Parliamentary recordings were quality-controlled by graduate-level native speakers. Similarly to Common Voice (Ardila et al., 2020), the reviewers listened to the audio and verified that the transcripts were accurate and gave a positive (thumbs up) or negative (thumbs down) rating per pair of audio-transcripts. The reviewers manually rated 10-20% of the clips per contributor, strictly rejecting samples for language mismatch, content errors, or unintelligibility. Only contributors with a validated accuracy rate exceeding 80% from reviewers were included in this release to ensure benchmark reliability.

3.1.2 Multilingual African Speech (MLS)

Afrivox-MLS comprises existing open source transcription datasets–NCHLT, Common Voice 17 (filtered for African languages), FLEURS (Conneau et al., 2023), OpenSLR, BibleTTS (Meyer et al., 2022), FISD (Asamoah Owusu and Omane Boateng, 2022), NaijaVoices(Emezue et al., 2025)–and newly-created AfriVox-Medical[3], a health-related read-speech multilingual translation and transcription dataset of simulated text conversations across 20 languages).

To create Afrivox-Medical we extracted over 40,000 utterances from popular open source TEXTUAL doctor-patient conversation datasets (Korfiatis et al., 2022; Fareez et al., 2022; He et al., 2020), and had bilingual native speakers translate them to African languages and read out the text translation in their native language, creating a parallel dataset of English text, local language text, and local language voice, providing both transcription and translation data.

For translation, we include FLEURS (Conneau et al., 2023), CoVoST (Wang et al., 2020), NaijaVoices (Emezue et al., 2025), IWSLT-LRST (Cettolo et al., 2017), and AfriVox-Medical[3].

MLS represents 20 languages across 8 datasets, 3,000+ speakers, 81.5 hours of audio (Tables 3 and 2). Coverage includes both high-population and

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

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

[3] https://huggingface.co/datasets/intronhealth/afrivox

[4] https://huggingface.co/datasets/intronhealth/afrispeech-parliament

low-resource languages, and features diverse linguistic families (Niger-Congo, Afro-Asiatic, etc.). Details on the number of hours per language are included in Appendix Table 9.

Postprocessing, Ethics and Quality Control: All audio files are mono-channel WAV at 16kHz. All data is either open-source or collected with explicit consent. Using the same platform and process described for Afrispeech Parl above, QC for Afrivox-Medical was performed by native speakers, and ONLY contributors with 80% correct translations and recordings were included in the final dataset. All contributors and reviewers were fairly compensated via a crowdsourcing platform[5].

Annotator characteristics and instructions are included in Appendix section A.1.

3.2 Models Evaluated

We benchmarked a mix of unimodal and multimodal models: **Unimodal ASR:** Canary (Puvvada et al., 2024), Parakeet (NeMo and Suno.ai, 2023), Whisper (Medium/Large) (Radford et al., 2023b), MMS (with/without language adapters) (Pratap et al., 2024), Sahara-v2* (research-preview model)[6]. **Unimodal AST [X->En]:** Whisper, MMS. **Multimodal LLMs:** Google Gemini-2.0-Flash(Google DeepMind, 2024), Google Gemini-3.0-Flash[7], OpenAI GPT-4o (OpenAI et al., 2024), Alibaba Qwen2.5-Omni (Chu et al., 2024), Meta SeamlessM4T (Aharoni et al., 2019), Meta Omnilingual ASR (team et al., 2025), all meeting the LLM criteria (Wikipedia contributors, 2025)– characterized by (a) text generation, (b) vast training data, (c) large parameter size–and supporting at least one additional modality (e.g. speech, images) beyond text.

Languages supported for each model are presented in Table 6. “Covered languages" are defined either by the model’s documentation of supported languages (e.g. MMS, Whisper), prior work (Beyene et al., 2025; Adelani et al., 2025), or manual testing for text generation and question-answering in local languages (e.g. GPT-4, Gemini) where LLMs show clear understanding of the local language and can chat fluently in African languages (e.g. Yoruba, Kinyarwanda, Swahili, etc) as shown in (Beyene et al., 2025; Adelani et al., 2025).

Models were chosen for their reported state-of-the-art performance, public availability, language coverage (supporting one or more African languages), or relevance to real-world deployment in Africa. All were used in their pre-trained, off-the-shelf forms unless otherwise specified.

3.3 Fine-Tuning

To demonstrate the utility of localized datasets for adapting speech LLMs to African languages we fine-tuned Qwen2.5-Omni (Chu et al., 2024) on the NaijaVoices dataset–1,800 hours, 5,000+ speakers, balanced by gender and age, spanning 3 Nigerian languages– Hausa, Igbo, and Yoruba. Qwen2.5-Omni is a 10B multimodal and multilingual LLM selected for PEFT due to its open-source availability, multilingual support, and relatively small size (compute limitations).

Fine-tuning: We fine-tuned on four NVIDIA 3090 GPUs with approximately 280 hours of speech per language, 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 with bfloat16 precision, with a batch size of 256. Prompt formatting details are included in the Appendix A

3.4 Evaluation

Tasks include (1) **Automatic Speech Recognition (ASR)** where audio was transcribed into native script, (2) **Automatic Speech Translation (AST)**, and translating local language audio into English text.

All models were tested with consistent, standardized prompts (zero-shot and few-shot) for fairness and reproducibility (see Appendix A for details). Model outputs were normalized for punctuation, casing, and removing diacritics to ensure comparability.

All code, model configurations, and new data will be open-sourced; results are reported for single runs.

3.4.1 Metrics and Human Evaluation

Transcriptions were evaluated using Word Error Rate (WER), while translations were evaluated using BLEU (Papineni et al., 2002), chrF (Popović, 2015), and AfriCOMET-STL (Wang et al., 2023).

Human evaluation of translations is a very expensive and time-consuming process. Since prior

[5] https://speech.intron.health/

[6] https://www.intron.io/

[7] https://gemini.google.com/app

work has established the effectiveness of more recent automated metrics like COMET (Rei et al., 2020; Wang et al., 2023), our goal was to determine if such published SOTA automated metrics were reliable for our analysis. Graduate-level native-speakers manually reviewed and scored 50 randomly sampled model translations per language for Fluency and Adequacy. We hypothesize that metrics are reliable where, per language, correlation is moderately to strongly positive (Appendix Table 19).

3.4.2 Addressing Benchmark Contamination

We note and analyze the potential for older public datasets to appear in model pretraining, and explicitly distinguish between “old” (NCHLT, Common-Voice) and “new” (AfriSpeech-200, Afrispeech-Parl) data in analysis to measure true generalization.

4 Results and Analysis

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

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

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 6 shows multimodal models edges over unimodal models on the speech translation task. Additionally, Table 8 shows the comparison between the results of the base and fine-tuned Qwen 2.5 Omni 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:

4.1 Widespread Variation in African Language and Accent Performance and Support

Table 4 and 5 reveal that, despite recent advances and better coverage of African languages, both unimodal and multimodal speech models exhibit substantial performance gaps on African languages and non-native English accents when compared with large languages and native accents (Multilingual Librispeech). Wide variation within models exist, most evident with Seamless and Whisper for supported languages. Consistent with multilingual claims in its documentation, Gemini outperforms GPT-4o by a wide margin sometimes with 2-4x better WER.

Unusually High Error Rates for Supported Languages: On African-accented English, state-of-the-art unimodal ASR models (e.g., Whisper Large-v3) display a 10–15x increase in Word Error Rate (WER) compared to standard benchmarks—for example, WER rises from 2.0% (LibriSpeech) to 26–38% (AfriSpeech, NCHLT) (Table 4). For African languages, WERs routinely exceed 50% and, for some languages (e.g., Yoruba, Hausa, Swahili), surpass 100%, despite self-reported "support" for these languages (Table 5), indicating nearly unintelligible output. These results suggest that simply including African data in pretraining does not provide performance guaranties.

Potential Benchmark Contamination: As shown in Table 4, performance on "New" datasets significantly lags "Old" suggesting potential exposure of such models to existing benchmarks.

Multimodal Model Language Coverage: Multimodal LLMs (e.g., Gemini, SeamlessM4T, GPT-4o) support more African languages than most unimodal ASR/AST models and can be prompted without explicit language labels, but their accuracy often lags behind unimodal models for transcription. SeamlessM4Tv2, for example, shows particularly strong results for Southern and Eastern African languages, providing clues about the language distribution in its training data.

4.2 Transcription vs. Translation: Unimodal and Multimodal Model Trends

Transcription (ASR): As shown in Table 5, unimodal models, especially Sahara v2* and the Omni-

		Whisper	Whisper	MMS-1b		Seamless	Gpt-4o	Gemini-3.0	Omni-ASR	Sahara
Language	Canary-1b	medium	large-v3	all	Qwen2.5	M4T L-v2	aud-prev	flash	7B-CTC	v2*
English (M. Lib)	**3.03**	6.80	3.53	17.63	16.32	4.68	9.63	-	-	-
French (M. Lib)	**4.06**	8.90	5.38	19.30	10.43	6.82	22.71	-	-	-
Spanish (M. Lib)	-	-	-	17.35	-	**6.76**	21.25	-	-	-
Afrikaans	-	68.87	45.43	48.73	-	18.41	84.36	**16.47**	21.8	34.58
Akan	-	-	-	62.92	-	-	104.02	**46.74**	52.53	65.39
Amharic	-	447.26	165.83	67.52	-	44.05	245.4	57.54	**23.64**	27.5
Arabic	-	39.49	29.72	44.94	-	51.26	31.88	**12.13**	22.92	15.34
French	9.67	13.95	9.31	33.93	24.14	15.90	22.29	**6.63**	33.41	10.96
Fulani	-	-	-	56.78	-	86.85	157.03	61.62	**53.41**	-
Ga	-	-	-	-	-	-	172.73	74.55	101.82	**10.91**
Hausa	-	180.29	95.11	40.47	-	-	118.60	31.37	41.87	**19.5**
Igbo	-	-	-	50.33	-	70.03	112.23	50.02	57.42	**23.17**
Kinyarwanda	-	-	-	36.73	-	-	135.75	40.44	26.6	**10.48**
Luganda	-	-	-	28.85	-	**16.39**	131.19	35.14	22.41	19.42
Pedi	-	-	-	41.43	-	-	119.29	39.78	42.82	**23.7**
Sesotho	-	-	-	-	-	-	158.21	161.41	64.26	**19.8**
Shona	-	193.21	110.35	30.7	-	76.05	90.51	70.56	**20.29**	29.26
Swahili	-	117.7	62.75	28.37	-	16.25	73.96	15.83	17.35	**11.51**
Tswana	-	-	-	-	-	-	133.46	54.85	48.6	**22.6**
Twi	-	-	-	51.09	-	-	98.86	78.09	52.81	**10.63**
Xhosa	-	-	-	42.24	-	-	130.79	33.82	32.34	**28.36**
Yoruba	-	213.88	93.77	39.59	-	37.43	101.14	27.35	31.96	**17.66**
Zulu	-	-	-	43.19	-	52.53	135.84	26.85	31.05	**16.08**

Table 5: **WER (%) by model and language on the Multilingual African Speech transcription dataset**. Bold values mark the lowest (best) WER for each language. "-" indicates no available result. The first section of the table shows baseline performance on Multilingual LibriSpeech (Pratap et al., 2020b)

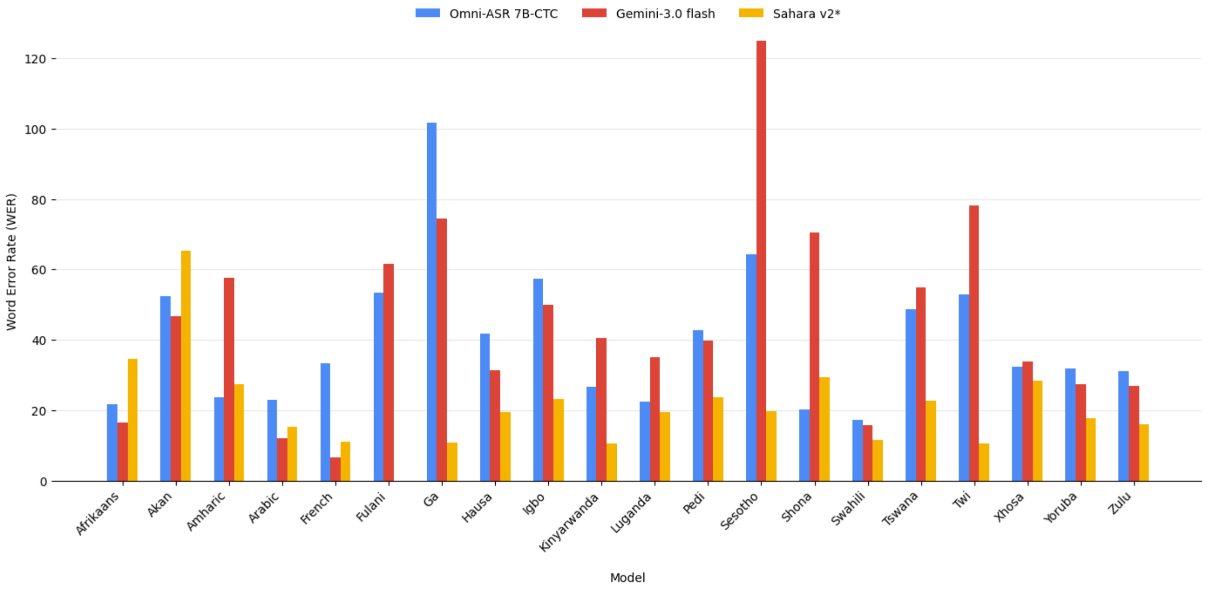


Figure 1: WER for Best Performing Models on Multilingual African Speech Dataset

ASR model, outperform multimodal LLMs for exact transcription in most African languages. Gemini stands out, outperforming both models across 4 languages, indicating progress towards more inclusive multimodal LLMs. However, with WERs still over 20% for several languages and accented speech, top ASR models and LLMs still struggle with accent/language diversity and noisy or spontaneous speech.

Translation (AST): Multimodal models (Table 6), especially Gemini and SeamlessM4T, significantly outperform unimodal baselines on low-resource African language audio-to-English translation. They achieve higher BLEU and AfriCOMET-STL scores, and provide more semantically faithful translations, particularly on longer, context-rich utterances. Appendix Table 19 shows AfriCOMET-STL's correlation with human evaluation.

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	**0.80**	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**

Table 6: **AfriComet-STL scores** across the languages for each model. "–" means the language is not supported by the model. The highlighted scores are the best score per language

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash	Qwen2.5
Hausa	186.23	96.99	**39.37**	–	119.74	52.16	126.81
Igbo	–	–	**48.81**	66.27	117.84	87.32	198.68
Yoruba	213.41	97.51	**44.05**	44.62	107.25	78.46	120.84

Table 7: **Transcription WER % for each model–language pair on the NaijaVoices subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

Language	ASR (WER)		AST (STL)	
	Base	Finetuned	Base	Finetuned
Hausa	126.81	**50.54**	0.19	**0.39**
Igbo	198.68	**42.41**	0.18	**0.54**
Yoruba	120.84	**71.29**	0.20	**0.29**

Table 8: Qwen-Omni2 ASR (WER score) and AST (AfriComet-STL) Performance Before and After Fine-Tuning

4.3 Robustness to Real-World Speech

Realistic Noisy Conditions: As shown in Table 4 When compared with "Old" datasets, all models perform 7-70% worse on the parliamentary proceedings dataset, which contains high ambient noise, overlapping speakers, and real-world accented spontaneous speech. Here, WERs for even the best models double relative to clean, read speech. MMS is most notable in this regard, with a 5x collapse in WER, likely demonstrating an overreliance on clean/read speech during training.

Accent and Dialect Variability: As shown in clean (Afrispeech) and noisy (Parl) settings (4), accented English speech transcription is still an unsolved problem. Prior work (Radford et al., 2023a; Galvez et al., 2024; Puvvada et al., 2024) and public leaderboards (Srivastav et al., 2023) show that the selected models perform up to 50% better on noisy native English speech, e.g. Vox-populi (Parliamentary) (Wang et al., 2021), AMI (meetings) (von Platen et al., 2019), and Earnings-21 (Del Rio et al., 2021).

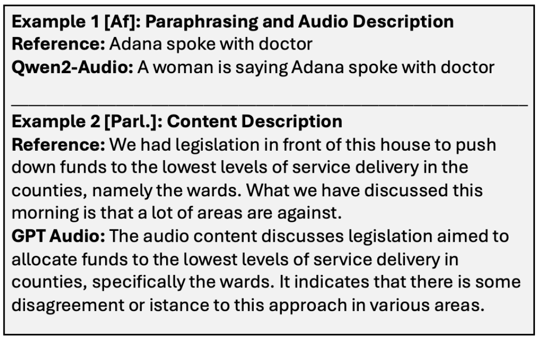
Accented French (Table 5) further reinforces the observation that performance degrades for underrepresented accents and dialects even beyond English. Inclusion of accent-diverse data exposes weaknesses in nearly all models, with WER dropping by roughly 2x.

4.4 Fine-Tuning Unlocks Substantial Gains

Parameter-Efficient Fine-Tuning (PEFT): Table ?? zooms in on model performance for the 3 languages selected for fine-tuning. Although all 3 languages were unsupported by Qwen2.5-Omni, Table 8 shows that fine-tuning on just 280 hours per language from NaijaVoices yields a 54% reduction in WER and up to 21-point gains in BLEU for Igbo, Hausa, and Yoruba, exceeding SOTA (MMS) on Igbo. Although results appear expected, this is one of the first open-source multimodal LLMs to support these 3 African languages. AfriCOMET-STL (translation performance) more than doubles for all three languages, exceeding SOTA (Gemini) on Igbo. These results demonstrate that, even with moderate in-domain data, open-source speech LLMs can be rapidly adapted for African languages using PEFT, offering a viable path for local teams.

4.5 Error Analysis

Verbatim vs. Paraphrase: Multimodal models frequently paraphrase or summarize rather than provide exact transcriptions (Figure 2), which is unsuitable for many ASR use cases. In contrast, unimodal ASR models are more likely to attempt verbatim output, albeit with higher rates of insertion and substitution errors on low-resource languages.

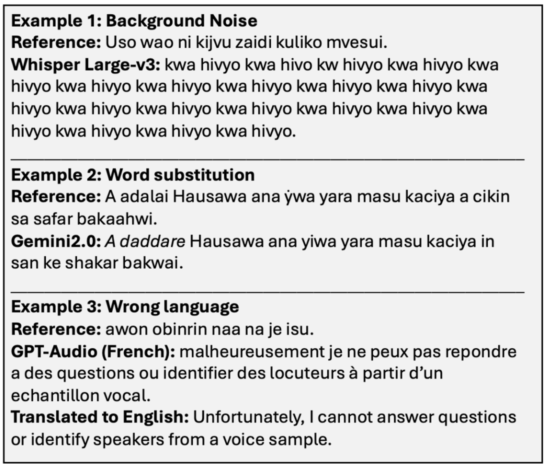


Example 1 [Af]: Paraphrasing and Audio Description
Reference: Adana spoke with doctor
Qwen2-Audio: A woman is saying Adana spoke with doctor

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 2: Examples of paraphrasing and audio description.

Hallucinations: Similar to findings in (Barański et al., 2025; Serai et al., 2022; Koenecke et al., 2024) on hallucinations in neural speech recognition models, both Whisper and Canary sometimes hallucinate content—repeating text (Oscillations (Frieske and Shi, 2024)) or filling silent segments with unrelated words as shown in Figure 3. Multimodal models are prone to "helpful" completions (Figure 3), such as generating plausible answers to questions not present in the audio.

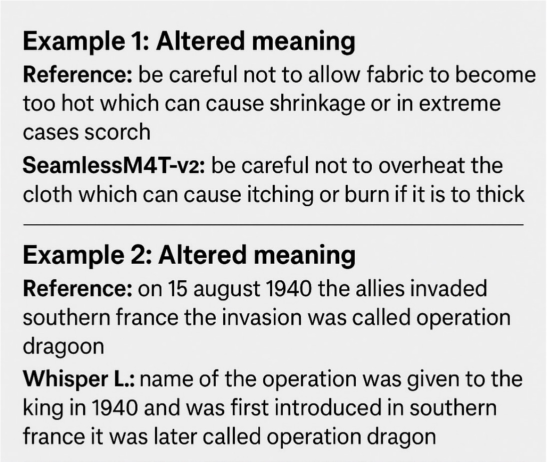


Example 1: Background Noise
Reference: Uso wao ni kijvu zaidi kuliko mvesui.
Whisper Large-v3: kwa hivyo kwa hivo kw hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo kwa hivyo.

Example 2: Word substitution
Reference: A adalai Hausawa ana ƴwa 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 3: Examples of oscillations, hallucination, word substitutions, and language mismatch in ASR outputs from unimodal and multimodal models.



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

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

Contextual Mistranslations: In AST tasks, multimodal models occasionally substitute synonyms or miss important words (Figure 4), producing contextually plausible but non-literal translations—highlighted by AfriCOMET-STL (Table 6, which better captures adequacy than BLEU alone.

Noise Sensitivity: All models suffer from degraded output under overlapping speech and real-world noise, with frequent failures to segment speakers or filter background sounds, indicating model's failure to adequately generalize to real-world spontaneous speech.

4.6 Implications for Inclusive Voice Technology

Our findings have clear implications for implementers, researchers, and product teams:

Model Selection: For applications requiring exact transcription—such as legal or medical records—unimodal ASR models remain preferable where they support the target language. However, for conversational interfaces or translation tasks, recent multimodal LLMs (e.g. Gemini) offer broader language coverage and better semantic translation, even in low-resource settings.

Fine-Tuning Value: The dramatic improvements achieved with PEFT fine-tuning on Qwen2.5-Omni (Table 8) highlight a promising pathway for African NLP practitioners. Moderate, domain-specific datasets can unlock substantial gains, making open-source LLMs much more practical for local deployment.

Benchmark Relevance: Our analysis underscores the need for modern, representative benchmarks like AfriVox. Results on older datasets (e.g., CommonVoice, NCHLT) often overestimate model performance due to likely benchmark contamination; newer, more challenging datasets like AfriSpeech-200 and Afrispeech-Parliamentary expose the true generalization gap.

Language and Accent Prioritization: Error patterns suggest that models benefit from balanced, accent-diverse training and evaluation data. Developers should prioritize expanding coverage of dialects and spontaneous speech, not just major languages.

5 Conclusion

This work introduces AfriVox, the first comprehensive benchmark designed to evaluate both unimodal and multimodal speech LLMs on African languages and accented English, directly addressing the urgent need for evidence-based guidance as voice-driven AI proliferates across Africa. Our systematic comparison of state-of-the-art models reveals that, despite recent advances, major gaps persist in model accuracy, language support, and robustness—particularly for spontaneous speech, diverse dialects, and real-world conditions.

Limitations

While AfriVox makes an important step toward rigorous, inclusive benchmarking for African speech technologies, several methodological constraints should be noted:

Dataset Representation: Although AfriVox covers more African languages and accents than prior work, many of Africa's 2,000+ languages remain unrepresented or covered by small sample sizes. Dialectal and spontaneous speech diversity is still far from exhaustive.

Benchmark Contamination: Some older public datasets (e.g., CommonVoice, NCHLT) may overlap with pretraining data for popular models, possibly inflating apparent model performance relative to unseen, truly out-of-domain audio. Our results on newly-curated datasets are more reliable but still limited by size and scope.

Evaluation Scope: Most evaluations focus on transcription and direct audio-to-English translation. We do not benchmark the full range of speech LLM multimodal abilities (e.g., dialog, spoken question answering), nor do we exhaustively test different prompting strategies or task configurations due to compute constraints.

Fine-Tuning Experiments: Our parameter-efficient fine-tuning is limited to three Nigerian languages, using moderate (not minimal) amounts of labeled data. Results may not generalize to ultra-low-resource languages or domains with dramatically less data available.

Noise and Real-World Testing: While AfriVox includes challenging real-world audio, our robustness analysis is not exhaustive. Further work should explore adversarial noise, code-switching, and multi-speaker dialog in more depth.

Despite these constraints, AfriVox establishes a practical and extensible blueprint for ongoing evaluation and improvement of speech and text LLMs in Africa. We hope this work will catalyze further open data sharing, community-driven evaluation, and development of voice AI systems that genuinely serve Africa's linguistic diversity.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. *arXiv preprint*. ArXiv:2406.03368 [cs].

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* *Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammed Al-Garadi, Tushar Mungle, Abdulaziz Ahmed, Abeed Sarker, Zhuqi Miao, and Michael E. Matheny. 2025. Large Language Models in Healthcare. *arXiv preprint*. ArXiv:2503.04748 [cs].

Rawan AlSaad, Alaa Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *Journal of Medical Internet Research*, 26(1):e59505. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

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.

Korsah A. Quartey B. Nwolley Jnr. S. Sampah D. Adjepon-Yamoah D. Asamoah Owusu, D. and L. Omane Boateng. 2022. GitHub - Ashesi-Org/Financial-Inclusion-Speech-Dataset: A speech dataset to support financial inclusion created by Ashesi University and Nokwary Technologies with funding from Lacuna Fund. [link].

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, and 1 others. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. *arXiv preprint arXiv:2506.08400*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

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.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

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.

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.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, and 1 others. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Rita Frieske and Bertram E Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *arXiv preprint arXiv:2401.01572*.

Daniel Galvez, Vladimir Bataev, Hainan Xu, and Tim Kaldewey. 2024. Speed of light exact greedy decoding for rnn-t speech recognition models on gpu. *arXiv preprint arXiv:2406.03791*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S. Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *arXiv preprint*. ArXiv:2406.11768 [cs].

Google DeepMind. 2024. Gemini 2.0 flash: Built for the agentic era.

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*.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, and 1 others. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

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*.

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.

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.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Primock57: A dataset of primary care mock consultations. *arXiv preprint arXiv:2204.00333*.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, and 1 others. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *arXiv preprint arXiv:2207.03546*.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI. Issue: 3.

NVIDIA NeMo and Suno.ai. 2023. Parakeet tdt 1.1b: An asr model with fastconformer and tdt decoder.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? *arXiv preprint*. ArXiv:2311.07978 [cs].

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.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood Yekini, Jonas Kemp, Katherine Heller, Jude Chidubem Omeke, Chidi Asuzu MD, Naome A. Etori, Aimérou Ndiaye, Ifeoma Okoh, and 7 others. 2025. AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset. *arXiv preprint*. ArXiv:2411.15640 [cs].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

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.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2025. A Survey on Speech Large Language Models. *arXiv preprint*. ArXiv:2410.18908 [eess] version: 3.

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, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020a. MLS: A Large-Scale Multilingual Dataset for Speech Research. ArXiv:2012.03411 [eess].

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020b. 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. 2023a. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. *arXiv preprint*. ArXiv:2109.04715 [cs].

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. 2025. Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. *arXiv preprint*. ArXiv:2502.03945 [cs].

Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:890–900.

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung-yi Lee, and Shinji Watanabe. 2024. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. *arXiv preprint*. ArXiv:2406.08641 [cs].

Vaibhav Srivastav, Somshubra Majumdar, Nithin Koluguri, Adel Moumen, Sanchit Gandhi, and 1 others. 2023. Open automatic speech recognition leaderboard. `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Patrick von Platen, Chao Zhang, and Philip Woodland. 2019. Multi-span acoustic modelling using raw waveform signals. *arXiv preprint arXiv:1906.11047*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

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*.

Wikipedia contributors. 2025. Large language model — Wikipedia, the free encyclopedia. [Online; accessed 6-October-2025].

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 Annotator Characteristics and Instruction

Annotators were graduate-level native speakers between 22-40 years of age located in respective countries e.g. Nigeria, Ghana, Kenya, South Africa.

Transcription annotators were instructed to transcribe audio files verbatim paying attention to

Table 9: Dataset Summary by Language and Domain

Language	Medical (hrs)	Non-medical (hrs)	Total (hrs)	Num Speakers	Num Samples
Afrikaans	1.79	2.27	4.05	42	1406
Akan	0.67	0.58	1.24	15	411
Amharic	0.37	0.26	0.62	8	214
Arabic	1.47	1.13	2.60	26	799
French	0.30	0.21	0.51	9	135
Ga	0.00	0.01	0.01	1	5
Hausa	3.81	1.71	5.53	125	1869
Igbo	1.66	1.42	3.08	37	970
Kinyarwanda	2.80	2.20	5.00	48	1172
Pedi	2.18	1.94	4.13	33	1121
Sesotho	2.56	2.11	4.68	28	1356
Shona	2.81	2.18	4.99	41	1114
Swahili	3.35	2.57	5.92	121	1377
Tswana	2.09	3.56	5.65	51	1573
Twi	0.58	0.13	0.71	4	339
Xhosa	3.31	2.70	6.01	58	1799
Yoruba	1.56	1.25	2.81	78	890
Zulu	3.95	2.75	6.70	73	2331

named entities and other domain specific terms, adding speaker tags where multiple speakers exists, and ignoring segments with indistinguishable or overlapping speech, ignoring disfluencies and filler words.

Translators were instructed to translate to the best of their ability including incomplete or unintelligble segments, refraining from adding to the meaning of the source text. Code switched or non-existent words in the target language were to be used in the source form. Numbers, dates, and other non-text expressions were to be translated to the nearest form in the target language.

A.2 Automatic Speech Recognition

A.2.1 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.3 Automatic Speech Translation

A.3.1 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.3.2 Performance Across Sources

Language	Whisper medium	Whisper large-v3	MMS-1b all	SeamlessM4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash	Gemini-3.0 flash	Sahara V2*
Afrikaans	44.49	30.93	26.48	18.64	32.20	13.77	**13.56**	36.41
Amharic	441.81	205.81	34.71	86.45	118.45	**19.10**	48.77	23.84
Arabic	–	11.06	36.28	9.29	6.64	**4.42**	7.08	21.43
Fulani	–	–	**56.78**	–	157.03	74.62	61.62	-
Hausa	158.21	86.13	31.39	–	100.85	34.92	26.88	**24.52**
Igbo	–	–	44.60	102.95	110.63	66.07	53.62	**38.36**
Luganda	–	–	45.77	**37.62**	89.34	52.98	40.44	47.96
Pedi	–	–	31.29	–	110.12	90.11	**30.46**	34.13
Shona	222.30	116.51	**29.60**	76.46	97.43	54.45	99.75	33.21
Swahili	99.04	41.51	22.22	11.98	29.92	12.37	**8.94**	12.16
Xhosa	–	–	44.58	–	124.79	56.94	36.59	**22.36**
Yoruba	204.21	87.18	34.29	31.03	82.98	42.04	**24.93**	26.38
Zulu	–	–	40.30	50.56	110.88	32.03	**24.44**	22.48

Table 10: **WER % for each model–language pair on the FLEURS subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

Language	Canary 1b	Whisper medium	Whisper large-v3	MMS-1b all	Qwen2.5	SeamlessM4T-v2 Large	GPT-4o audio-preview	Gemini-2.0 flash	Gemini-3.0 flash	Sahara V2*
Afrikaans	–	52.54	32.7	36.99	–	14.69	47.48	17.64	**11.14**	32.01
Akan	–	–	–	62.90	–	–	103.97	76.53	**46.71**	55.18
Arabic	–	45.74	33.10	75.29	–	–	32.76	23.67	18.71	**16.70**
French	13.14	16.32	10.65	41.74	24.00	16.80	12.11	8.02	**7.45**	10.02
Hausa	–	129.55	93.68	43.22	–	–	125.96	39.55	29.67	**21.96**
Igbo	–	–	–	53.61	–	68.97	104.18	77.30	49.62	**24.35**
Kinyarwanda	–	–	–	46.65	–	–	134.26	65.19	36.11	**11.37**
Pedi	–	–	–	46.67	–	–	124.27	76.72	40.82	**23.59**
Sesotho	–	–	–	–	–	–	172.76	77.59	214.12	**23.28**
Shona	–	150.31	101.27	32.33	–	75.46	80.30	45.04	26.76	**16.22**
Swahili	–	112.09	48.11	34.17	–	18.87	42.74	16.30	15.76	**9.86**
Tswana	–	–	–	–	–	–	135.98	72.81	31.66	**30.14**
Twi	–	–	–	50.55	–	–	102.58	80.66	39.83	**12.46**
Xhosa	–	–	–	43.62	–	–	122.86	46.54	**30.51**	34.39
Yoruba	–	157.12	88.98	43.05	–	30.44	134.79	54.02	24.59	**16.21**
Zulu	–	–	–	48.41	–	52.49	129.38	35.19	26.77	**10.41**

Table 11: **WER % for each model–language pair on the Intron-AfriVox subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

Language	Whisper medium	Whisper large-v3	MMS-1b all	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash	Gemini-3.0 flash	Sahara V2*
Amharic	427.57	155.51	76.16	**23.94**	280.17	280.17	67.61	36.77
Swahili	132.67	73.47	40.56	26.39	93.58	93.58	21.16	**20.37**

Table 12: **WER % for each model–language pair on the ALFFA subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

Language	Canary 1b	Whisper medium	Whisper large-v3	MMS-1b all	Qwen	Seamless-M4T-v2 Large	Gpt-4o audio-preview	Gemini-2.0 flash	Gemini-3.0 flash	Sahara V2*
French	**5.49**	7.69	11.10	24.53	24.00	14.82	34.55	12.67	5.64	12.1

Table 13: **WER % for each model–language pair on the OpenSLR subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

	MMS-1b	Gpt-4o	Gemini-2.0	Gemini-3.0	Sahara
Language	all	audio-preview	flash	flash	V2*
Akan	**77.78**	133.33	94.44	66.67	85
Ga	–	172.73	114.55	74.55	**10.91**
Twi	75.00	184.38	150.00	84.38	**6.25**

Table 14: **WER % for each model–language pair on the Ashesi Financial Inclusion subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

	Whisper	Whisper	MMS-1b	Seamless-M4T-v2	Gpt-4o-	Gemini-2.0	Gemini-3.0	Sahara
Language	medium	large-v3	all	Large	audio-preview	flash	flash	V2*
Afrikaans	52.30	37.65	27.09	**13.80**	57.07	17.55	6.64	31.82
Amharic	513.92	183.28	**52.69**	92.51	183.54	130.17	28.16	33.91
Arabic	36.24	18.33	27.66	68.27	31.73	11.94	8.43	**2.30**
Hausa	270.36	91.49	27.20	–	109.09	40.53	21.84	**15.91**
Igbo	–	–	60.71	42.86	246.43	82.14	46.43	**15.91**
Kinyarwanda	–	–	32.75	–	136.35	84.26	42.15	**12.81**
Luganda	–	–	28.51	**15.97**	132.04	80.73	35.04	18.84
Swahili	120.74	71.30	24.50	14.11	92.47	26.33	15.92	**11.81**
Twi	–	–	57.53	–	123.29	93.15	**49.32**	65.75
Yoruba	294.01	99.43	38.63	39.91	96.48	103.57	25.62	**24.57**

Table 15: **WER % for each model–language pair on the Common Voice subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

	Whisper-	Whisper	MMS-1b	Seamless-M4T-v2	Gpt-4o	Gemini-2.0	Gemini-3.0	Sahara
Language	medium	large-v3	all	Large	audio-preview	flash	flash	V2*
Afrikaans	99.00	68.31	71.32	**25.01**	151.50	48.94	26.23	27.68
Pedi	–	–	42.03	–	119.29	90.75	48.87	**19.82**
Sesotho	–	–	–	–	133.43	104.33	71.96	**14.01**
Tswana	–	–	–	–	127.82	85.19	181.46	**9.24**
Xhosa	–	–	31.93	–	171.43	56.70	37.52	**15.27**
Zulu	–	–	28.10	56.43	208.26	44.64	31.73	**18.45**

Table 16: **WER % for each model–language pair on the NCHLT subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

	Whisper	Whisper	MMS-1b	Seamless-M4T-v2	Gpt-4o	Gemini-2.0	Gemini-3.0	Sahara
Language	medium	large-v3	all	Large	audio-preview	flash	flash	V2*
Hausa	186.23	96.99	39.37	–	119.74	52.16	29.13	**16.73**
Igbo	–	–	48.81	66.27	117.84	87.32	49.60	**15.26**
Yoruba	213.41	97.51	44.05	44.62	107.25	78.46	35.95	**14.67**

Table 17: **WER % for each model–language pair on the NaijaVoices subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

	Whisper	Whisper	MMS-1b	Seamless-M4T-v2	Gpt-4o	Gemini-2.0	Gemini-3.0	Sahara
Language	medium	large-v3	all	Large	audio-preview	flash	flash	V2*
Hausa	112.01	102.16	**39.37**	–	110.46	104.58	104.93	103.57
Twi	–	–	–	51.53	89.81	78.04	51.58	**4.82**
Yoruba	118.50	106.66	24.63	27.23	84.70	44.94	16.77	**12.41**

Table 18: **WER % for each model–language pair on the BibleTTS subset of the Multilingual African Speech transcription dataset**; the lowest (best) WER per language is shown in bold. "-" indicates the language is not supported by the model.

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
Average	BLEU	0.23	0.52
	ChrF	0.40	0.66
	AfriComet-STL	**0.48**	**0.70**

Table 19: Pearson correlations (r) between automatic metrics and human evaluations of fluency and adequacy for automatic speech translation.

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 20: **BLEU scores for each model–language pair on the Multilingual African Speech translation dataset**; the highest (best) BLEU score per language is shown in bold. "-" indicates the language is not supported by the model.

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	56.13	50.33	45.58	–	–
Akan	**29.86**	25.01	–	–	–	–	–
Amharic	**56.62**	29.62	43.48	17.06	13.57	–	–
Arabic	**63.10**	59.26	55.53	47.85	44.38	–	–
French	**66.56**	64.40	63.72	58.61	57.19	54.12	64.94*
Fulani	**27.56**	23.82	16.25	–	–	–	–
Ga	**20.08**	19.09	–	–	–	–	–
Hausa	**48.48**	29.81	13.47	13.29	7.78	–	–
Igbo	**32.10**	25.40	18.52	–	–	–	–
Kinyarwanda	**37.69**	23.62	–	–	–	–	–
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	–	–	–	–	–
Yoruba	38.45	28.37	**40.53**	14.29	10.45	–	–
Zulu	**52.76**	31.54	32.79	–	–	–	–

Table 21: **CHrF scores for each model–language pair on the Multilingual African Speech translation dataset**; the highest (best) CHrF score per language is shown in bold. "-" indicates the language is not supported by the model.

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 22: **BLEU & CHrF scores for each model–language pair on the FLEURS subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.

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 23: **BLEU & CHrF scores for each model–language pair on the Intron-AfriVox subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.

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

Table 23: **BLEU & CHrF scores for each model–language pair on the Intron-AfriVox subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.

Language	Gemini		GPT-4o-audio preview		SeamlessM4T v2 Large		Whisper Large		Whisper Medium		Qwen Omni	
	BLEU	ChrF	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	0.25	13.19
Igbo	**6.97**	**28.67**	4.35	22.91	4.22	22.80	–	–	–	–	0.26	12.59
Yoruba	9.92	32.57	4.88	24.32	**16.34**	**39.61**	0.11	11.52	0.11	10.33	0.24	13.12

Table 24: **BLEU and ChrF scores for each model–language pair on the NaijaVoices subset of the Multilingual African Speech Translation dataset**. The highest (best) BLEU and ChrF score per language is shown in bold, with the ChrF score further underlined. "–" indicates the language is not supported by the model.

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 25: **BLEU & CHrF scores for each model–language pair on the IWSLT_LRST subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.

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 26: **BLEU & CHrF scores for each model–language pair on the Covost subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.

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

Table 26: **BLEU & CHrF scores for each model–language pair on the Covost subset of the Multilingual African Speech translation dataset**; the highest (best) BLEU & CHrF score per language is shown in bold with the CHrF score further underlined. "-" indicates the language is not supported by the model.