The WAXAL ASR Benchmark: Fine-Tuned Edge Models Across 19 African Languages

Anonymous ACL submission

Abstract

We benchmark three zero-shot foundation models against compact fine-tuned edge models across 19 African languages using the conversational WAXAL corpus. Fine-tuned edge models achieve a macro-averaged WER of 38.0% compared to 64.9% for the best zero-shot baseline, a 26.9 percentage-point (pp) reduction using models 3 to 40 times smaller. This demonstrates that domain specialization consistently dominates scale for spontaneous African speech. Cross-domain evaluation on FLEURS shows partial generalization: fine-tuned models recover usable performance on out-of-distribution read-speech, though zero-shot models retain an advantage when the test domain matches their pretraining conditions. An audit conducted by native speakers across all 19 languages yields a linguistically grounded error taxonomy, revealing that Connectionist Temporal Classification (CTC) and autoregressive architectures fail in different ways that align with the language family. Specifically, the CTC-based MMS-300M wins or ties on all six Bantu languages, while the autoregressive Whisper-Small leads on Afro-Asiatic languages. We further show that WER alone misrepresents performance for syllabary-script languages (Amharic, Tigrinya), where CER/WER ratios reveal substantially higher character-level accuracy than headline WER suggests. All model weights, fine-tuning scripts, evaluation code, and a cleaned test subset across all 19 languages (CC-BY 4.0) are released as open-source artifacts.

1 Introduction

Automatic speech recognition (ASR) has undergone a remarkable transformation over the past decade. Systems such as Whisper (Radford et al., 2023), Massively Multilingual Speech (MMS) (Pratap et al., 2024), and Omnilingual ASR (Keren et al., 2025) now claim support for hundreds of languages, projecting an image of near-universal coverage. Yet for most African languages particularly those with limited digital resources, these promises of multilingual ASR remain largely unrealized. Recognition quality, model accessibility (Ngueajio and Washington, 2022), and deployment feasibility continue to lag far behind, as documented across community-driven benchmarks (Olatunji et al., 2023; Conneau et al., 2023).

We posit that this important disparity is driven by two closely related challenges. The first is *Parity Gap*: a persistent performance deficit between high-resource and low-resource African languages. Despite their impressive zero-shot capabilities, modern multilingual ASR systems frequently struggle under realistic African speech conditions, characterized by spontaneous speech, code-switching, tonal distinctions, and dialectal diversity. In many cases, these models exhibit hallucination loops, over-generation, and unstable transcription behavior (Koudounas et al., 2025; Barański et al., 2025; Atwany et al., 2025), or do not support the target language entirely.

The second is the *Efficiency Gap*: models capable of handling this linguistic diversity, such as Whisper Large-v3 (1.5B parameters) and MMS-1B, are computationally intensive and therefore impractical for the very environments where African language data are often collected, namely edge devices and low-cost mobile hardware. Together, these gaps motivate a systematic study. While fine-tuning is known to benefit low-resource ASR (Imam et al., 2026; Liu et al., 2024; Emezue et al., 2025), whether compact edge models can close the gap with zero-shot giants across 19 typologically diverse African languages under conversational speech conditions, and whether such specialization generalizes beyond the training domain—warrants empirical investigation at this scale.

We address this using the recently released WAXAL corpus (Diack et al., 2026), comprising spontaneous, image-prompted speech in 19

African languages recorded in participants' natural environments. We evaluate three zero-shot foundation models (Whisper Large-v3, MMS-1B, Omnilingual-1B) against three compact fine-tuned architectures (Whisper Tiny, Whisper Small, MMS-300M). Beyond quantitative benchmarking, we conduct a distributed native-speaker audit across all 19 language communities to characterize architectural failure modes, and evaluate cross-domain robustness on FLEURS.

These analyses address two open research questions:

1. **Cross-Domain Generalization:** Do models fine-tuned on naturalistic WAXAL speech generalize to out-of-domain read-speech benchmarks, and under what conditions does zero-shot scale advantage resurface?

2. **Architectural and Linguistic Patterns:** What systematic patterns distinguish how CTC and autoregressive architectures fail across African language families, script systems, and morphological types?

Our core contributions are:

- **Confirmation of specialization advantage**: we confirm that compact edge models (39M–300M parameters) fine-tuned on WAXAL achieve a 26.9 pp macro-average WER reduction over the best zero-shot baseline across 19 African languages, using models 3–40× smaller.

- **Cross-domain generalization evidence**: fine-tuned models partially recover usable performance on out-of-distribution FLEURS read-speech, with domain match rather than model size as the primary driver of relative performance.

- **Empirical error taxonomy**: a distributed native-speaker audit across all 19 languages reveals systematic patterns in how CTC and autoregressive architectures fail, correlating with language family, script system, and morphological typology—generating testable hypotheses for future work.

- **Cleaned open-source test subset**: the cleaned and filtered WAXAL test split across all 19 languages, processed with speech-rate and duration heuristics that exposed WER-inflating reference artifacts, released under CC-BY 4.0 for community use.

- **Open-source models and code**: fine-tuned model weights for three architectures across 19 languages, with all training scripts and evaluation code, released for community use.

2 Related Work

2.1 Large-Scale Multilingual Speech Recognition

Modern multilingual ASR has been shaped by three major foundation models. Whisper (Radford et al., 2023) demonstrated that training an encoder-decoder architecture on 680,000 hours of weakly supervised web data yields strong zero-shot transcription in dozens of languages. MMS (Pratap et al., 2024) extended CTC-based recognition to 1,100+ languages via self-supervised pretraining of wav2vec 2.0 with language-specific adapters. Omnilingual ASR (Keren et al., 2025) scaled coverage to 1,600+ languages through multilingual acoustic pretraining with lightweight CTC decoding. Despite this expansion, little work has evaluated any of these systems on *spontaneous* African speech, and all require compute impractical for edge deployment.

2.2 African Speech Resources and Benchmarking

The development of African ASR has historically been constrained by the scarcity of multilingual speech corpora. Mozilla Common Voice (Ardila et al., 2020) introduced crowdsourced speech collection for several African languages including Hausa, Luganda, and Kinyarwanda; however, recordings follow scripted prompts and primarily capture read speech, limiting applicability to spontaneous conversational settings. FLEURS (Conneau et al., 2023) provided a standardized multilingual evaluation suite spanning 102 languages, becoming the dominant benchmark for assessing cross-lingual transfer; however, its read-speech format does not capture the spontaneous, noisy conditions typical of African deployment scenarios. AfriSpeech-200 (Olatunji et al., 2023) took a community-driven approach, assembling 200 hours of English speech from African speakers across 120 accent categories to highlight the underrepresentation of African-accented speech in mainstream ASR; however, its focus on English limits direct

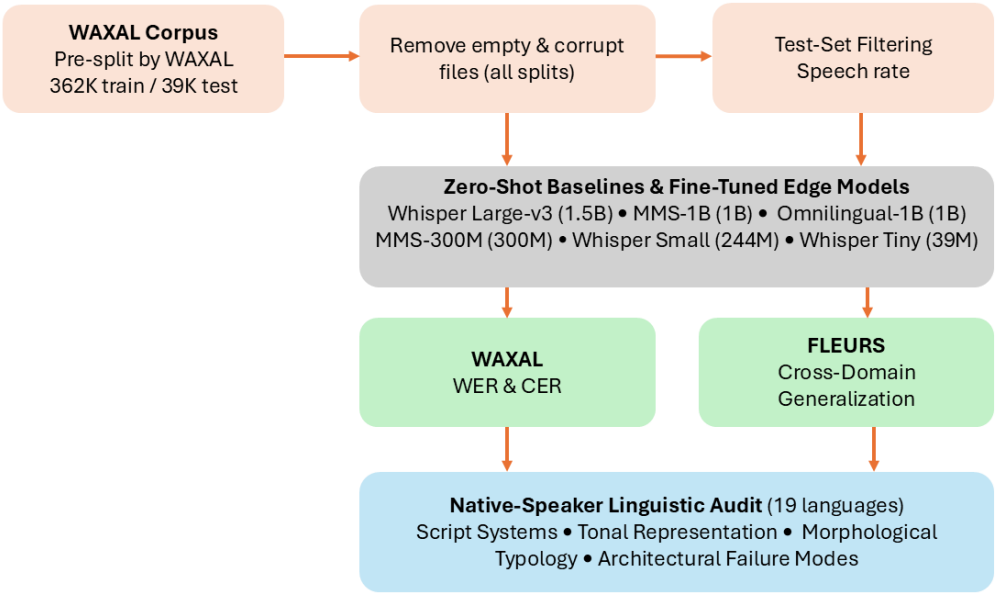

Figure 1: Benchmarking pipeline. The WAXAL corpus arrives pre-split by its creators. General cleaning (removing empty and corrupt files) is applied across all splits; additional speech-rate and duration filtering is applied to the test split only. Fine-tuned edge models are trained on the cleaned train split; all six models are evaluated on the filtered test set. Quantitative results feed into a structured native-speaker linguistic audit across all 19 languages.

applicability to indigenous African language recognition, where the data scarcity problem is considerably more acute.

The recently released WAXAL corpus (Diack et al., 2026) provides spontaneous conversational speech across 19 African languages, capturing naturalistic recording environments unlike FLEURS's read-speech conditions. Earlier downstream work from Ethio-ASR (Abdullah et al., 2026) has demonstrated that compact CTC models fine-tuned on WAXAL subsets can achieve competitive performance. However, no prior work has systematically evaluated the full 19-language WAXAL corpus across multiple foundation model architectures simultaneously. Our work fills this gap directly.

2.3 Fine-Tuning and Model Efficiency for Low-Resource ASR

Recent work has explored adaptation strategies to make large ASR models practical in low-resource settings. Studies on Whisper fine-tuning (Liu et al., 2024) have compared full fine-tuning, LoRA, and adapter-based approaches, finding that parameter-efficient methods can significantly reduce computational costs while maintaining competitive performance. Multistage adaptation pipelines (Pillai et al., 2026) have shown that sequential multilingual fine-tuning improves transfer when linguistic similarity exists between source and target languages. Additional work on African-language-specific adaptation (Imam et al., 2026) has explored the tradeoffs between full fine-tuning and parameter-efficient methods across individual languages.

3 Methodology

3.1 The WAXAL Dataset

We utilize the WAXAL benchmark (Diack et al., 2026), a large-scale corpus of transcribed, image-prompted spontaneous speech across 19 African languages for ASR. The labeled subset available for experimentation comprises 2,279.1 hours and 446,169 utterances (362,125 training, 44,232 validation, 39,812 test); full corpus details are in Diack et al. (2026). Unlike read-speech benchmarks such as FLEURS (Conneau et al., 2023), WAXAL captures conversational speech in participants' natural environments, making it representative of real-world deployment conditions. Language and speaker metadata are provided in Appendix F. All data used in this experiment was pre-split by Diack et al. (2026) in a split of approximately 80/10/10 for train, validation, and test partitions and released under CC-BY-4.0.

3.2 Data Cleaning and Filtering Heuristics

During preliminary evaluations, we identified critical dataset integrity issues and implemented these steps to address them:

1. **Duration Threshold:** Discarding any audio segments shorter than 1.5 seconds.

2. **Speech Rate Threshold:** Discarding samples where the ground-truth text required a physically impossible speech rate of > 4 words per second ($num_words/duration > 4$).

This dataset mitigation strategy significantly normalized evaluation metrics. For instance, Lingala Whisper Tiny WER improved from 113.5% to 49.0%, thus allowing for fair comparison of true acoustic modeling capabilities. All models were assessed using Word Error Rate (WER) (Morris et al., 2004) and Character Error Rate (CER), both computed via the `jiwer` library (Vaessen, 2022) after lowercasing and punctuation removal.[1]

3.3 Distributed Human Evaluation

Beyond brittle metric-based evaluation, we conducted a distributed shared-task evaluation involving native speakers across all 19 languages. The broader research team comprises 32 contributors drawn from university research programs and community networks, with contacts extending to additional native speakers in each language group. Each language was assigned to between one and four native speakers recruited through this network on a volunteer basis; contributors were not compensated financially. Prior to annotation, all contributors received a standardised briefing document describing the annotation task, error categories, and example cases.

For each language–model combination, annotators reviewed 40 audio samples: 20 from the top-performing quartile by utterance-level WER and 20 from the bottom-performing quartile, extracted programmatically from the WAXAL test split. Annotators listened to each clip, compared it against the reference and model hypothesis, and recorded error patterns including phonetic substitutions, word boundary errors, morpheme splits, hallucination loops, and code-switching mismatches using a structured template.

Annotators are native speakers but not formally trained linguists; in multi-annotator languages, evaluation was conducted as a collaborative joint review rather than independent parallel annotation, so inter-annotator agreement is not reported. Moreover, the 40-utterance sample size per language limits generalizability and statistical power. As a result, we present our qualitative findings in Section 6 as exploratory linguistic observations rather than statistically confirmed patterns.

3.4 Zero-Shot Baselines

We evaluate three models in a zero-shot setting: Whisper Large-v3 (1.5B) (Radford et al., 2023), MMS-1B (1B) (Pratap et al., 2024), and Omnilingual-1B (1B) (Keren et al., 2025).

3.5 Specialization Strategies (Fine-Tuning)

To test whether targeted specialization of compact models can close the gap with these large baselines, we fine-tune three edge-deployable architectures on the WAXAL training split: Whisper Tiny (39M parameters), Whisper Small (244M parameters), and MMS-300M (300M parameters). For the Whisper models, we apply *full fine-tuning*, updating all model weights. For MMS-300M, we employed a parameter-efficient approach by freezing the encoder layers. The MMS-300M was initialized from a base pretrained checkpoint that *could not output text* prior to WAXAL fine-tuning. Its text prediction capabilities were learned entirely from WAXAL fine-tuning, making its performance a direct measure of what WAXAL supervision alone can induce.

At inference, Whisper models used greedy decoding with repetition penalty disabled; MMS-300M used CTC greedy decoding. Per-language training convergence details are provided in Appendix E.

3.6 Training Infrastructure & Hyperparameters

All fine-tuning experiments were conducted on NVIDIA A100 GPUs (MMS-300M on A100-SXM4-80GB; Whisper models on A100-SXM4-40GB). All experiments used an effective batch size of 32 via gradient accumulation, a learning rate of 1×10^{-4} with linear warmup (500 steps) and polynomial decay, and up to 30 epochs with early stopping (patience 3). Hyperparameters were

[1] WER can exceed 100% when insertions are large, as occurs in autoregressive hallucination loops (Koudounas et al., 2025). Native-speaker audit additionally revealed that Tigrinya, Ikposo, and Acholi reference texts were systematically truncated mid-sentence in several splits; model outputs for the continuing audio were scored as spurious insertions under standard WER metrics.

held fixed across all 19 languages to enable fair cross-language comparison.

4 Results: Scale vs. Specialization

4.1 Quantitative Performance: The Specialization Gain

As shown in Figure 2, fine-tuning compact edge models consistently closes the performance gap with zero-shot giants. Zero-shot models perform poorly across the board: Omnilingual-1B achieves a macro-averaged WER of 64.9% and MMS-1B 74.7%, with MMS-1B exceeding 80% WER on over half of the languages (10 of 19). By contrast, fine-tuned edge models achieve substantially lower WERs; MMS-300M at 38.0%, Whisper Small at 39.9%, Whisper Tiny at 44.2%, a 26.9 pp reduction relative to the best zero-shot baseline, using models 3–40× smaller.

Among fine-tuned models, MMS-300M leads on 8 of 19 languages, Whisper Small on 7, and Whisper Tiny on 1 (Fula). Three MMS-300M vs. Whisper Small comparisons are practical ties (differences <0.5 pp): Acholi (42.3% vs. 42.3%), Lingala (42.6% vs. 42.7%), Malagasy (12.8% vs. 13.1%); both models are bolded for these languages in Table 3.

We define the **Parity Gap** as the WER difference between the best zero-shot and best fine-tuned model per language. It ranges from 1.4 pp (Oromo: MMS-1B 26.6% vs. Whisper Small 25.2%, the only case where zero-shot nearly matches fine-tuned performance) to 51.8 pp (Wolaytta: MMS-1B 90.6% vs. MMS-300M 38.8%), reflecting uneven penetration of these languages into foundation model pretraining data.

4.2 The CTC vs. Autoregressive (AR) Trade-off

We observe distinct architectural behaviors between CTC models and AR sequence-to-sequence models. MMS-300M (CTC) generally outperforms Whisper Small (AR) on Character Error Rate, consistent with CTC's tendency toward acoustic precision (K et al., 2025) (see Figure 4 in the appendix). By contrast, Whisper tends to hallucinate when acoustic evidence is ambiguous; the AR decoder generates plausible-sounding but phonetically incorrect tokens (Barański et al., 2025; Koudounas et al., 2025; Atwany et al., 2025). This pattern varies by language family as detailed in Section 6.

5 Cross-Dataset Generalization (FLEURS)

We evaluated all models on the FLEURS test set for the 6 overlapping languages (Amharic, Fula, Lingala, Luganda, Oromo, Shona), sampling 300 utterances per language (Table 2).[2] Without fine-tuning, MMS-1B reaches ∼100% WER across all six languages and Whisper Large-v3 achieves 392% WER on Amharic, which is indicative of severe hallucination on Afro-Asiatic languages. WAXAL-fine-tuned models recover usable performance (36–70% WER).

Domain Specificity and the FLEURS Gap: Omnilingual-1B leads on FLEURS for four languages (Fula, Lingala, Luganda, Shona) yet is consistently outperformed on WAXAL. This pattern is consistent with domain specificity: when the test domain aligns with pretraining conditions (controlled read-speech), scale advantage resurfaces; when shifted to conversational speech, fine-tuned models recover the advantage. These results indicate that domain match, rather than model size alone, is the primary driver of relative performance across the two evaluations.

6 Qualitative and Linguistic Analysis

Our native-speaker audit revealed distinct failure patterns which we organise across three axes: architectural design, script systems, and linguistic typology.

6.1 Architectural Failure Modes

Unbounded Repetition Loops (Whisper): Whisper models exhibit unbounded repetition where the decoder gets stuck generating text well beyond the audio content. This loop behavior was observed in 14 of the 19 languages audited and was more severe in Whisper Tiny, consistent with the smaller model's reduced capacity to maintain coherent long-range context. Tigrinya was the exception, showing no loop behavior in either Whisper model, consistent with stronger pretraining signal from its official-language status and well-documented Ge'ez script (Koudounas et al., 2025).

The following example illustrates such behavior in Soga, where the model begins transcribing the audio correctly before falling into an endless prediction loop:

[2] CI: $1.96 \times \sqrt{\text{WER}(1-\text{WER})/300} = \pm 5.5$ pp at 40–75% WER, adequate given inter-model gaps >10 pp in all cases.

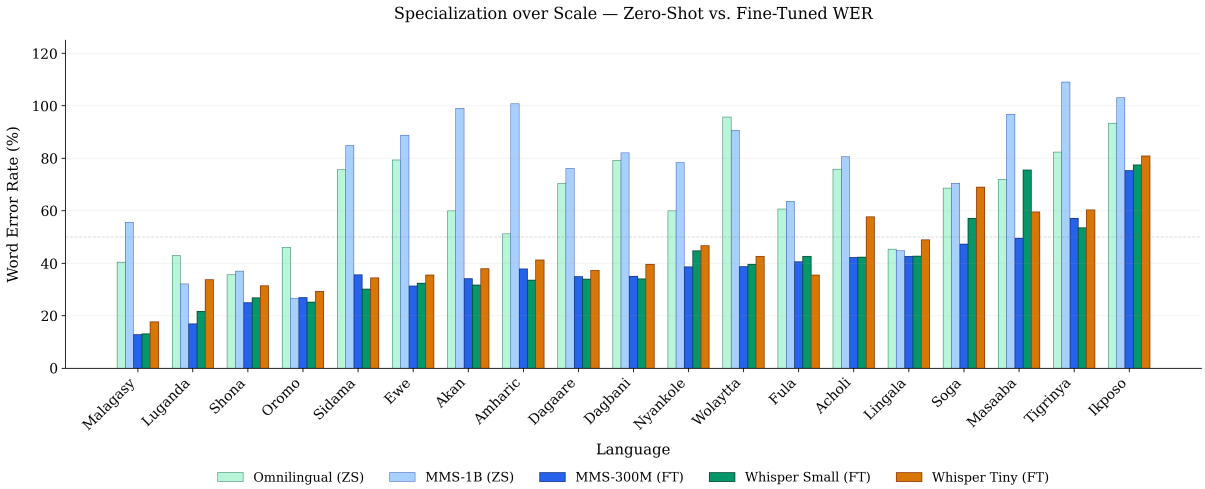

Figure 2: Word Error Rate (%) across 19 WAXAL languages for five of the six evaluated models: zero-shot baselines Omnilingual-1B and MMS-1B (light bars) vs. fine-tuned edge models MMS-300M, Whisper Small, and Whisper Tiny (dark bars). Whisper Large-v3 is omitted as it natively supports only 4 of the 19 languages; its results are reported in Table 3.

Language	Omni. (ZS) WER / CER	MMS-1B (ZS) WER / CER	Whisper L (ZS) WER / CER	MMS-300M (FT) WER / CER	Whisp. S (FT) WER / CER	Whisp. T (FT) WER / CER
Acholi	75.8 / 30.0	80.5 / 30.8	N/A	**42.3** / **17.1**	**42.3** / 19.2	57.7 / 33.6
Amharic	51.1 / 17.4	100.7 / 84.6	254.3 / 144.7	37.8 / 13.2	**33.6** / **12.9**	41.3 / 16.8
Fula	60.6 / 17.0	63.6 / 18.2	N/A	40.6 / **10.4**	42.6 / 12.0	**35.5** / 12.6
Lingala	45.3 / 25.1	44.7 / 24.8	93.0 / 64.0	**42.6**‡ / **18.9**	**42.7**‡ / 21.7	49.0 / 30.9
Luganda	42.9 / 8.5	32.1 / 6.7	N/A	**16.9** / **3.4**	21.6 / 5.5	33.8 / 12.8
Oromo	46.0 / 9.2	26.6 / **5.7**	N/A	26.9 / 6.6	**25.2** / 6.8	29.3 / 8.9
Shona	35.6 / 8.5	36.9 / 8.9	112.4 / 29.9	**25.0** / **4.3**	26.9 / 5.1	31.4 / 8.9
Tigrinya	82.4 / 52.1	109.0 / 100.7	N/A	57.1 / **37.2**	**53.5** / 38.7	60.3 / 43.2

Table 1: Primary WAXAL Benchmark Results. We select a representative subset spanning Nilo-Saharan (Acholi), Bantu (Luganda, Shona, Lingala), Afro-Asiatic (Amharic, Oromo, Tigrinya), and Atlantic (Fula) language families. Full 19-language results are provided in the supplementary materials (Table 3).

> **Ref:** "omukyala oyo ayemereire omukyala oyo agemye akaveera akairugavu era alinaine akaduuka akatunda airtime"
> **Pred:** "omukyala oyo ayemeleire omukyala oyo agemye akaveera akeirugavu era alinhaime akadhuuka akatumula eyaatoimu ogha *emiti emiti emiti emiti emiti...* [repeats 100+ times]"

Phonetic Approximation (MMS): The CTC-based MMS-300M model is structurally immune to unbounded repetition. Its primary failure mode is phonetic approximation, where it correctly identifies the acoustic structure but substitutes similar phonemes. For instance, in Acholi, representative errors include compound splitting ("kacato" → "ka cato") and fricative confusion ("ovarol" → "ofarol").

Unlike Whisper's repetition loops, MMS-300M's phonetic approximation errors are recoverable: a downstream reader or language model can often reconstruct the intended form from a phonetically plausible approximation, whereas repetition loops produce semantically irrecoverable output.

6.2 Script System and Orthographic Complexity

Error patterns stratify by script system. Ge'ez-script languages (Amharic, Tigrinya) average 43.5% best fine-tuned WER vs. 36.2% for Latin-script languages. More diagnostically, Ge'ez-script languages show higher CER/WER ratios (Amharic: 0.35, Tigrinya: 0.65) than Latin-script Bantu languages (Shona: 0.17, Luganda: 0.20). In Ge'ez, each character encodes a consonant-vowel syllable pair, so a single *fidel* substitution counts as a full word error under WER, structurally penalizing these languages. The elevated CER/WER ratio reveals the model captures CV structure correctly but makes character-level substitutions at the syllabary boundary: WER alone misrepresents performance for such syllabary-script languages.

At the model level, Whisper Small's autoregres-

Language	Omni. (ZS) WER / CER	MMS-1B (ZS) WER / CER	Whisper L (ZS) WER / CER	MMS-300M (FT) WER / CER	Whisp. S (FT) WER / CER	Whisp. T (FT) WER / CER
Amharic	59.1 / 12.7	100.2 / 78.3	392.2 / 262.7	39.5 / 12.0	**36.4 / 11.9**	45.4 / 16.5
Fula	**53.7 / 15.1**	102.0 / 78.1	92.2 / 35.0	56.5 / 18.2	66.8 / 29.4	84.6 / 37.0
Lingala	**18.4 / 5.6**	101.2 / 76.9	77.0 / 20.9	38.3 / 12.8	39.5 / 18.7	63.9 / 42.9
Luganda	**50.1 / 9.2**	100.5 / 85.6	N/A	57.8 / 15.1	66.1 / 18.0	72.3 / 22.1
Oromo	75.8 / **18.3**	100.0 / 83.5	N/A	70.1 / 18.9	**65.5** / 19.3	74.6 / 23.9
Shona	**24.0 / 5.1**	100.3 / 85.7	120.3 / 29.3	39.3 / 10.9	51.2 / 19.3	69.7 / 30.4

Table 2: Zero-Shot (ZS) vs Fine-Tuned (FT) Performance on the out-of-domain FLEURS dataset. Models were evaluated on 300 randomly sampled utterances per language (approximately 35–50 minutes of audio per language), yielding 95% confidence intervals of ± 5.5 percentage points at typical observed WER values (40–75%).

sive prior partially compensates for syllabary complexity in Amharic (CER 12.9% vs. MMS-300M 13.2%), while MMS-300M leads on Tigrinya (37.2% vs. 38.7%), where the larger syllabary inventory limits autoregressive advantage.

6.3 Tonal Language Behaviour and Orthographic Representation

Tonal and non-tonal languages show similar macro-averaged fine-tuned WER (37.0% for both), a result that should not be interpreted as evidence that tonality has no effect on ASR difficulty. Rather, other language-specific factors modulate aggregate performance: the key differentiator is *orthographic representation* of tone rather than tonality per se. Tonal languages exhibit substantially higher variance (std 16.4 vs. 12.2), and this variance is not randomly distributed: tonal Kwa languages with extensive Latin diacritics (Ikposo: 75.3%, Dagaare: 34.9%, Ewe: 31.3%) perform substantially worse than tonal Bantu languages where tone is contextually predictable and less consistently marked in orthography (Luganda: 16.9%, Shona: 25.0%). Diacritic characters (ɛ, ɔ, ŋ, ʋ) require precise Unicode handling, and the high-diacritic density of these orthographies plausibly contributes additional modeling challenges consistent with elevated character-level error rates. At the model level, this tonal-orthography interaction is reflected most clearly in the CTC vs. autoregressive split: Whisper Small outperforms MMS-300M on all three diacritic-heavy Kwa languages (Akan, Dagaare, Dagbani), consistent with its autoregressive prior providing additional disambiguation when diacritics create ambiguous acoustic frames.

6.4 Morphological Typology and Word Boundary Errors

The lowest CER/WER ratios are Bantu languages: Shona (0.17), Luganda (0.20), Soga (0.20), Malagasy (0.22), consistent with MMS's profile of small character-level substitutions within otherwise correct words. The highest non-Ge'ez ratios are Lingala (0.44) and Acholi (0.40), where compound splitting and morpheme boundary misplacement dominate.

6.5 CTC vs. Autoregressive by Language Family

The architectural advantage follows language family lines: MMS-300M wins or ties on all six Bantu languages (Luganda, Shona, Soga, Masaaba, Nyankole, Lingala) and the one Austronesian language (Malagasy), while Whisper Small leads on four of five Afro-Asiatic languages (Amharic, Oromo, Sidama, Tigrinya) and three diacritic-heavy Niger-Congo languages (Akan, Dagaare, Dagbani). We observe that CTC's frame-level decoding may better suit the transparent morphophonology of Bantu languages, while Whisper's autoregressive prior may provide additional disambiguation for Afro-Asiatic languages where acoustic evidence is more ambiguous. Whether this reflects architectural inductive biases or pretraining data composition remains an open question. Deployment recommendations follow in Section 7.

7 Discussion

7.1 Decision Logic for Deployment

Our findings provide a clear deployment roadmap along two dimensions: hardware constraints and target language family. On hardware, fine-tuned model footprints are MMS-300M ($\sim$1.2 GB), Whisper Small ($\sim$967 MB), and Whisper Tiny ($\sim$151 MB). Where 1.2 GB is feasible, MMS-300M offers the best acoustic precision and is immune to repetition loops; for tighter constraints, Whisper variants are viable but require repetition-penalty heuristics at decoding time. Deploying 1B+ parameter models without fine-tuning is unlikely

to yield usable results on African conversational speech.

On language family, MMS-300M wins or ties on all six Bantu languages and is the safer single-architecture default: immune to repetition loops, leading on CER in 17 of 19 languages, with recoverable failure modes. Whisper Small is preferred for Afro-Asiatic languages (four of five: Amharic, Oromo, Sidama, Tigrinya), where its autoregressive prior disambiguates complex morphology and large syllabary inventories.

7.2 Training Data and Cross-Language Generalization

The WAXAL training sets range from 25.8 hours (Acholi) to 197.3 hours (Wolaytta), a 7.6× span that could in principle confound cross-language comparisons if data quantity drives fine-tuned performance. To assess this, we computed the Spearman rank correlation between per-language training hours and best fine-tuned WER: $\rho = -0.19$, $p = 0.44$, indicating no statistically significant relationship. This suggests that architectural specialization and language-specific phonological properties, rather than raw data quantity, are the primary determinants of fine-tuned performance within the scope of this benchmark.

7.3 Ethical Considerations

The WAXAL dataset was collected with informed consent under CC-BY-4.0 licensing (Diack et al., 2026). All released model weights and fine-tuning scripts include usage guidelines that prohibit non-consensual voice applications such as unauthorized voice cloning or surveillance; we encourage downstream users to implement appropriate safeguards.[3]

8 Limitations

This benchmark does not exhaustively cover dialectal variation across the African continent, and WER penalizes natural conversational behaviors (pausing, code-switching) that do not reflect intelligibility failures. Four languages (Amharic, Oromo, Sidama, Wolaytta) have test sets of only 18–25 unique speakers despite large utterance counts, so WER estimates may not fully capture cross-speaker generalization.

[3]In accordance with the ACL Policy on AI Writing Assistance, we disclose that AI assistance was used for LaTeX formatting and coding for generation of visualizations; all data interpretation and linguistic analysis were conducted by human researchers.

9 Conclusion

Three findings from this study deserve emphasis beyond the headline WER numbers.

Domain specialization dominates scale. Compact models fine-tuned on the target acoustic domain (spontaneous WAXAL speech) consistently outperform models up to 40× larger under zero-shot conditions. The FLEURS cross-domain results explain why: when the acoustic domain matches a model's pretraining distribution—as it does for Omnilingual on clean read-speech—scale advantage resurfaces. For conversational African ASR deployment, modest in-domain data with fine-tuning outperforms large zero-shot models.

Architecture choice should follow language typology, not convention. The CTC vs. autoregressive distinction is not merely technical, it predicts performance by language family. MMS-300M (CTC) wins or ties on all six Bantu languages; Whisper Small (autoregressive) wins on four of five Afro-Asiatic languages. The pattern is consistent with known architectural properties: CTC's frame-level decoding may better suit the transparent morphophonology of Bantu languages, while autoregressive language model priors may provide disambiguation for Afro-Asiatic languages with complex morphology and script systems. Practitioners should consider architecture by target language family rather than defaulting to the most popular model.

WER understates performance for syllabary-script languages. The CER/WER ratio analysis reveals that Ge'ez-script languages (Amharic, Tigrinya) achieve character-level accuracy far higher than their WER scores suggest. A single *fidel* (syllabary character) substitution scores as a full word error under WER, structurally penalizing these languages relative to Latin-script equivalents. Cross-script comparisons require CER as a co-primary metric; reporting WER alone misrepresents model capability for a significant share of African languages.

These findings provide a practical and linguistically informed foundation for the next generation of African speech technology research.

References

Badr M. Abdullah, Israel Abebe Azime, Atnafu Lambebo Tonja, Jesujoba O. Alabi, Abel Mulat Alemu, Eyob G. Hagos, Bontu Fufa Balcha, Mulubrhan A. Nerea, Debela Desalegn Yadeta, Dagnachew Mekonnen Marilign, Amanuel Temesgen Fentahun, Tadesse Kebede, Israel D. Gebru, Michael Melese Woldeyohannis, Walelign Tewabe Sewunetie, Bernd Möbius, and Dietrich Klakow. 2026. Ethio-asr: Joint multilingual speech recognition and language identification for ethiopian languages. *Preprint*, arXiv:2603.23654.

Rosana Ardila, Megan Branber, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, 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.

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19000–19020, Vienna, Austria. Association for Computational Linguistics.

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.

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.

Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, and 24 others. 2026. Waxal: A large-scale multilingual african language speech corpus. *arXiv preprint arXiv:2602.02734*.

Chris Emezue, The NaijaVoices Community, Busayo Awobade, Abraham Toluwase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages. In *Interspeech 2025*, pages 1338–1342.

Edgar C Fieller, Herman Otto Hartley, and Egon Sharpe Pearson. 1957. Tests for rank correlation coefficients, i. *Biometrika*, 44(3/4):470–481.

Sukairaj Hafiz Imam, Muhammad Yahuza Bello, Hadiza Ali Umar, Tadesse Destaw Belay, Idris Abdulmumin, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2026. Full fine-tuning vs. parameter-efficient adaptation for low-resource African ASR: A controlled study with whisper-small. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 197–203, Rabat, Morocco. Association for Computational Linguistics.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4941–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Alkis Koudounas, Moreno La Quatra, Manuel Giollo, Sabato Marco Siniscalchi, and Elena Baralis. 2025. Hallucination benchmark for speech foundation models. *Preprint*, arXiv:2510.16567.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Mikel K Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International conference on human-computer interaction*, pages 421–440. Springer.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Leena G Pillai, Kavya Manohar, Basil Raju, and Elizabeth Sherly. 2026. Multistage fine-tuning strategies for automatic speech recognition in low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nik Vaessen. 2022. Jiwer: Similarity measures for automatic speech recognition evaluation (version 4.0.0). Accessed: 2026-03-01.

Appendix

A WER Reduction: MMS-1B Zero-Shot to MMS-300M Fine-Tuned

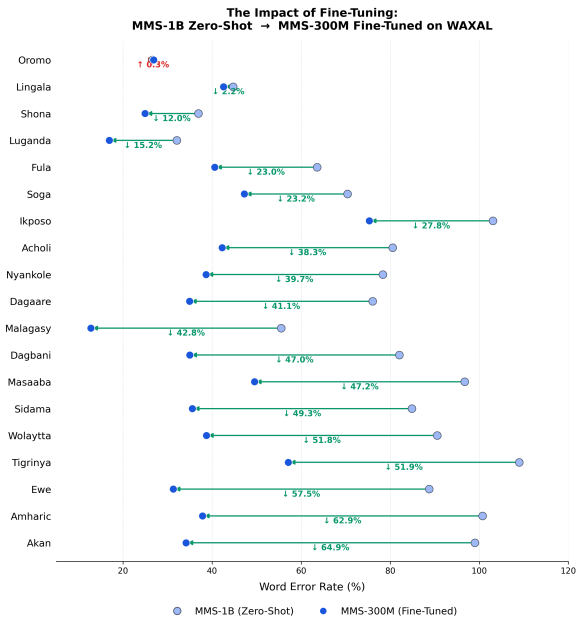

Figure 3: Relative WER reduction (%) from MMS-1B zero-shot to MMS-300M fine-tuned. Despite a 3.3× parameter reduction, the fine-tuned model achieves substantial improvements across nearly all 19 languages.

B CTC vs. Autoregressive CER Comparison

C Full 19-Language WER Results

Table 3 presents the complete Word Error Rate (WER) results across all 19 WAXAL languages for the six evaluated models. Table 1 in the main paper presents an abridged subset of these results.

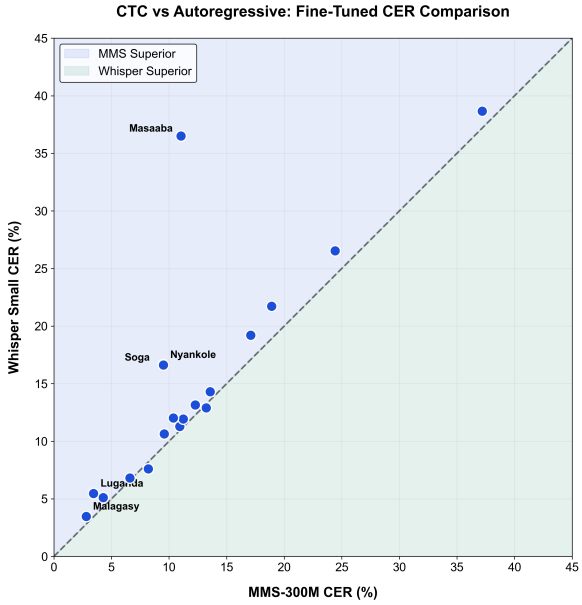

Figure 4: CTC vs. Autoregressive: Fine-Tuned CER Comparison. MMS-300M (CTC) achieves lower CER than Whisper Small (AR) in 17 of 19 languages, consistent with CTC's tendency toward phonetic precision, though Whisper's autoregressive prior provides advantages for morphologically complex languages (Section 6).

Entries marked "N/A" indicate languages for which the model provides no native support and cannot be meaningfully evaluated in a zero-shot setting.

D Full 19-Language CER Results

Table 4 presents the Character Error Rate (CER) for all fine-tuned models across 19 languages. CER provides a finer-grained view than WER and is particularly informative for agglutinative languages and those with complex diacritical systems, where a single word-level substitution may mask a near-correct character sequence. MMS-300M dominates CER in 17 of 19 languages.

E Training Convergence Summary

Table 5 reports the final evaluation WER, evaluation loss, and total training steps for each model–language combination as recorded via Weights & Biases. All runs used a maximum of 30 epochs with early stopping (patience = 3). The variation in total steps reflects differences in dataset sizes across languages and early stopping behavior.

Language	Omni. (ZS) WER	MMS-1B (ZS) WER	Whisper L (ZS) WER	MMS-300M (FT) WER	Whisp. S (FT) WER	Whisp. T (FT) WER
Acholi	75.8	80.5	N/A	**42.3**‡	**42.3**‡	57.7
Akan	59.9	99.0	N/A	34.2	**31.7**	37.9
Amharic	51.1	100.7	254.3	37.8	**33.6**	41.3
Dagaare	70.4	76.1	N/A	34.9	**34.0**	37.3
Dagbani	79.2	82.0	N/A	35.0	**34.0**	39.5
Ewe	79.3	88.7	N/A	**31.3**	32.3	35.5
Fula	60.6	63.6	N/A	40.6	42.6	**35.5**
Ikposo	93.3	103.1	N/A	**75.3**	77.5	80.9
Lingala	45.3	44.7	93.0	**42.6**‡	**42.7**‡	49.0
Luganda	42.9	32.1	N/A	**16.9**	21.6	33.8
Malagasy	40.4	55.5	132.0	**12.8**‡	**13.1**‡	17.7
Masaaba	71.9	96.7	N/A	**49.5**	75.5	59.6
Nyankole	59.9	78.3	N/A	**38.6**	44.7	46.7
Oromo	46.0	26.6	N/A	26.9	**25.2**	29.3
Shona	35.6	36.9	112.4	**25.0**	26.9	31.4
Sidama	75.6	84.9	N/A	35.6	**30.1**	34.4
Soga	68.6	70.4	N/A	**47.2**	57.1	69.0
Tigrinya	82.4	109.0	N/A	57.1	**53.5**	60.3
Wolaytta	95.7	90.6	N/A	**38.8**	39.5	42.6
Mean	64.9	74.7	—	**38.0**	39.9	44.2

Table 3: Full WAXAL Benchmark: Word Error Rate (%) across all 19 languages. Bold indicates the best fine-tuned model per language. ‡Both models are bolded where the WER difference between MMS-300M and Whisper Small is a practical tie (<0.5 pp): Acholi (0.0 pp), Lingala (0.1 pp), Malagasy (0.3 pp). Mean WER for Whisper Large-v3 is omitted as it covers only 4 of 19 languages.

F Dataset Statistics and Language Metadata

Table 6 presents comprehensive statistics for the 19 WAXAL languages, including training, validation, and test split sizes, speaker counts per partition, and total corpus hours per language. These statistics represent the transcribed and labelled subset of WAXAL currently available for experimentation.

G Supplementary Statistical Methods: Spearman Rank Correlation Analysis

G.1 Motivation and Confound Assessment

To assess whether training data quantity confounds cross-language performance differences, we evaluated whether per-language fine-tuned WER correlates with available training hours. A significant positive correlation would suggest that languages with more training data systematically achieve lower (better) WER, confounding claims about linguistic properties or architectural alignment. To test this, we employed Spearman's rank correlation coefficient, which is appropriate for small samples ($N = 19$ languages) and does not assume linear relationships or normality.

G.2 Mathematical Formulation

Spearman's rank correlation coefficient ρ is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{1}$$

where:

- n = number of observations (languages) = 19
- d_i = difference in ranks for observation i across the two variables
- $\sum_{i=1}^{n} d_i^2$ = sum of squared rank differences

The coefficient ρ ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with $\rho = 0$ indicating no monotonic relationship.

G.2.1 Significance Testing

To determine whether ρ differs statistically from zero, we compute the test statistic:

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \tag{2}$$

This statistic follows a t-distribution with $\nu = n - 2$ degrees of freedom. We compare this against a two-tailed critical value at $\alpha = 0.05$ significance level.

For $n = 19$, we have $\nu = 17$ degrees of freedom, and the critical value is $t_{0.025,17} = 2.110$.

Language	MMS-300M (FT) CER	Whisp. S (FT) CER	Whisp. T (FT) CER
Acholi	**17.1**	19.2	33.6
Akan	**10.9**	11.3	15.7
Amharic	13.2	**12.9**	16.8
Dagaare	**13.6**	14.3	16.1
Dagbani	**11.2**	11.9	15.6
Ewe	**9.6**	10.6	12.6
Fula	**10.4**	12.0	12.6
Ikposo	**24.4**	26.5	28.4
Lingala	**18.9**	21.7	30.9
Luganda	**3.4**	5.5	12.8
Malagasy	**2.8**	3.5	5.9
Masaaba	**11.0**	36.5	42.8
Nyankole	**9.5**	16.7	16.0
Oromo	**6.6**	6.8	8.9
Shona	**4.3**	5.1	8.9
Sidama	8.2	**7.6**	9.5
Soga	**9.5**	16.6	28.0
Tigrinya	**37.2**	38.7	43.2
Wolaytta	**12.3**	13.2	14.3

Table 4: Full WAXAL Benchmark: Character Error Rate (%) for fine-tuned models across all 19 languages. Bold indicates the best model per language. MMS-300M (CTC) achieves the lowest CER in 17 of 19 languages, consistent with its tendency toward phonetic precision, though architectural advantages vary by language family (see Section 6.5).

G.2.2 Confidence Interval via Fisher's z-Transformation

Following (Fieller et al., 1957), the 95% confidence interval for ρ can be computed using Fisher's z-transformation:

$$z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \tag{3}$$

The standard error of z is approximately $SE_z = 1/\sqrt{n-3}$.

The 95% confidence interval for z is:

$$\text{CI}_z = z \pm 1.96 \cdot SE_z \tag{4}$$

Transforming back to the ρ scale using the inverse transformation:

$$\rho = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{5}$$

G.3 Worked Example: Six-Language Calculation

To illustrate the calculation transparently, we present a step-by-step worked example on a representative subset of 6 languages.

G.3.1 Step 1: Raw Data

Language	Training Hours	Best FT WER (%)
Luganda	37.3	16.9
Malagasy	94.7	12.8
Shona	79.7	25.0
Acholi	25.8	42.3
Amharic	189.7	33.6
Oromo	189.9	25.2

Table 7: Example subset: 6 representative languages. *Best FT WER* refers to the minimum WER achieved across all three fine-tuned models (MMS-300M, Whisper Small, Whisper Tiny).

G.3.2 Step 2: Rank by Training Hours

Rank the languages by training hours (1 = smallest, 6 = largest):

Language	Training Hours	Rank
Acholi	25.8	1
Luganda	37.3	2
Shona	79.7	3
Malagasy	94.7	4
Amharic	189.7	5
Oromo	189.9	6

Table 8: Ranking by training hours (ascending order).

G.3.3 Step 3: Rank by Fine-Tuned WER

Rank the languages by WER (1 = best/lowest, 6 = worst/highest):

Language	MMS-300M Eval WER	MMS-300M Eval Loss	MMS-300M Steps	Whisper Small Eval WER	Whisper Small Eval Loss	Whisper Small Steps	Whisper Tiny Eval WER	Whisper Tiny Eval Loss	Whisper Tiny Steps
Acholi	0.407	0.612	1000	0.401	0.801	500	0.533	0.890	500
Akan	0.304	0.371	2000	0.301	0.419	1000	0.354	0.476	1000
Amharic	0.309	0.427	6000	0.491	0.100	4500	0.549	0.130	5500
Dagaare	0.342	0.512	3000	0.335	0.457	1500	0.365	0.516	1500
Dagbani	0.350	0.411	3000	0.347	0.401	1000	0.410	0.474	1500
Ewe	0.316	0.415	3000	0.335	0.414	1000	0.379	0.459	1500
Fula	0.400	0.366	3500	0.420	0.531	1500	0.457	0.597	1500
Ikposo	0.709	0.909	4000	0.739	0.485	1000	0.754	0.559	1500
Lingala	0.372	0.589	3500	0.348	0.568	1000	0.407	0.646	1000
Luganda	0.166	0.147	2000	0.222	0.324	500	0.294	0.427	1000
Malagasy	0.136	0.121	3000	0.196	0.308	1500	0.290	0.354	1500
Masaaba	0.499	0.447	2500	0.596	0.816	500	0.606	0.927	1000
Nyankole	0.373	0.378	1500	0.433	0.557	500	0.438	0.651	1000
Oromo	0.301	0.321	3500	0.288	0.418	3000	0.340	0.480	3500
Shona	0.244	0.204	2500	0.309	0.389	1000	0.365	0.440	1500
Sidama	0.406	0.295	4500	0.381	0.631	3500	0.429	0.704	3000
Soga	0.455	0.320	2500	0.509	0.605	500	0.588	0.684	500
Tigrinya	0.431	1.194	5500	0.507	0.172	3500	0.575	0.217	6000
Wolaytta	0.324	0.240	4500	0.343	0.452	3000	0.369	0.495	3000

Table 5: Training convergence details from Weights & Biases. Eval WER is measured on the WAXAL validation split (distinct from the test-set WER reported in Table 3).

Language	Best FT WER (%)	Rank
Malagasy	12.8	1
Luganda	16.9	2
Shona	25.0	3
Oromo	25.2	4
Amharic	33.6	5
Acholi	42.3	6

Table 9: Ranking by fine-tuned WER (ascending WER = better performance)

G.3.4 Step 4: Calculate Rank Differences

Compute $d_i = \text{Rank}_{\text{hours}} - \text{Rank}_{\text{WER}}$ and d_i^2:

Language	Rank_{hrs}	Rank_{WER}	d_i	d_i^2
Acholi	1	6	-5	25
Luganda	2	2	0	0
Shona	3	3	0	0
Malagasy	4	1	3	9
Amharic	5	5	0	0
Oromo	6	4	2	4
			$\sum d_i^2 =$	38

Table 10: Rank differences and squared differences. Note: Acholi shows the largest rank discrepancy (difference of 5), while Luganda, Shona, and Amharic show perfect rank agreement ($d_i = 0$).

G.3.5 Step 5: Compute Spearman Coefficient

Using Equation 1 with $n = 6$:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{6}$$

$$= 1 - \frac{6 \times 38}{6 \times (36 - 1)} \tag{7}$$

$$= 1 - \frac{228}{6 \times 35} \tag{8}$$

$$= 1 - \frac{228}{210} \tag{9}$$

$$= 1 - 1.0857 \tag{10}$$

$$= -0.0857 \tag{11}$$

For this 6-language example: $\rho = -0.0857$ (very weak negative correlation).

G.3.6 Step 6: Significance Testing

Compute the t-statistic using Equation 2:

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \tag{12}$$

$$= -0.0857 \sqrt{\frac{6-2}{1-(-0.0857)^2}} \tag{13}$$

$$= -0.0857 \sqrt{\frac{4}{1-0.00734}} \tag{14}$$

$$= -0.0857 \sqrt{\frac{4}{0.99266}} \tag{15}$$

$$= -0.0857 \sqrt{4.0295} \tag{16}$$

$$= -0.0857 \times 2.0074 \tag{17}$$

$$= -0.1720 \tag{18}$$

Language	ISO	Family	Train Utts	Train Hrs	Train Spkrs	Val Utts	Val Hrs	Val Spkrs	Test Utts	Test Hrs	Test Spkrs
Acholi	ach	Nilo-Saharan	4,108	25.8	322	519	3.3	199	528	3.3	194
Akan	aka	Niger-Congo	10,107	55.0	101	1,123	6.2	99	1,522	8.4	33
Amharic	amh	Afro-Asiatic	38,022	189.7	566	2,912	14.0	28	3,420	16.2	24
Dagaare	dga	Niger-Congo	15,071	83.4	348	1,893	10.5	264	1,910	10.5	261
Dagbani	dag	Niger-Congo	14,231	77.0	1,046	1,750	9.5	642	1,838	9.9	682
Ewe	ewe	Niger-Congo	15,054	79.4	533	1,916	10.1	423	1,891	10.0	401
Fula	ful	Atlantic-Congo	19,132	99.7	208	2,358	12.3	163	2,391	12.5	163
Ikposo	kpo	Niger-Congo	14,415	82.1	452	1,760	10.0	341	1,845	10.5	354
Lingala	lin	Niger-Congo	14,400	71.9	95	1,844	9.3	73	1,866	9.3	75
Luganda	lug	Niger-Congo	5,455	37.3	334	664	4.6	252	638	4.3	232
Malagasy	mlg	Austronesian	18,526	94.7	242	2,270	11.7	169	2,290	11.7	172
Masaaba	mas	Niger-Congo	6,867	39.3	351	849	4.9	255	858	4.9	251
Nyankole	nyn	Niger-Congo	6,783	40.8	360	831	5.1	256	858	5.2	261
Oromo	orm	Afro-Asiatic	38,185	189.9	359	3,078	14.5	23	3,782	18.3	24
Shona	sna	Niger-Congo	14,109	79.7	157	1,727	9.7	100	1,749	9.9	115
Sidama	sid	Afro-Asiatic	38,752	192.7	468	3,091	15.3	26	3,561	17.8	25
Soga	sog	Niger-Congo	6,219	40.0	316	841	5.3	230	817	5.2	233
Tigrinya	tir	Afro-Asiatic	40,856	181.9	384	4,425	16.8	35	5,034	20.3	38
Wolaytta	wal	Afro-Asiatic	41,833	197.3	392	2,444	9.6	17	3,014	12.4	18
Total			362,125	1,857.6	–	44,232	221.2	–	39,812	200.3	–

Table 6: WAXAL dataset statistics across all 19 languages. All numbers reflect the transcribed and labelled subset available for this benchmark. Abbreviations: Utts = utterances, Hrs = hours, Spkrs = speakers. Total speaker counts across partitions reflect unique speakers; the same speaker may appear in multiple partitions.

With $\nu = 6 - 2 = 4$ degrees of freedom, the critical value at $\alpha = 0.05$ (two-tailed) is $t_{0.025,4} = 2.776$.

Since $|t| = 0.1720 < 2.776$, we *fail to reject the null hypothesis* of no correlation.

Two-tailed p-value: $p \approx 0.87$ (NOT statistically significant).

G.3.7 Step 7: Confidence Interval via Fisher's z-Transformation

Using Equations 3–5:

$$z = \frac{1}{2}\ln\left(\frac{1+(-0.0857)}{1-(-0.0857)}\right) \quad (19)$$
$$= \frac{1}{2}\ln\left(\frac{0.9143}{1.0857}\right) \quad (20)$$
$$= \frac{1}{2}\ln(0.8421) \quad (21)$$
$$= \frac{1}{2} \times (-0.1718) \quad (22)$$
$$= -0.0859 \quad (23)$$

Standard error of z: $SE_z = 1/\sqrt{6-3} = 1/\sqrt{3} = 0.5774$

95% CI for z:

$$\text{CI}_z = -0.0859 \pm 1.96 \times 0.5774 \quad (24)$$
$$= -0.0859 \pm 1.1316 \quad (25)$$
$$= [-1.2175, 1.0457] \quad (26)$$

Transform back to ρ-scale (inverse Fisher transformation):

For lower bound ($z = -1.2175$):

$$\rho_{\text{lower}} = \frac{e^{2\times(-1.2175)} - 1}{e^{2\times(-1.2175)} + 1} \quad (27)$$
$$= \frac{e^{-2.435} - 1}{e^{-2.435} + 1} \quad (28)$$
$$= \frac{0.0884 - 1}{0.0884 + 1} \quad (29)$$
$$= \frac{-0.9116}{1.0884} \quad (30)$$
$$= -0.8378 \quad (31)$$

For upper bound ($z = 1.0457$):

$$\rho_{\text{upper}} = \frac{e^{2\times 1.0457} - 1}{e^{2\times 1.0457} + 1} \quad (32)$$
$$= \frac{e^{2.0914} - 1}{e^{2.0914} + 1} \quad (33)$$
$$= \frac{8.0876 - 1}{8.0876 + 1} \quad (34)$$
$$= \frac{7.0876}{9.0876} \quad (35)$$
$$= 0.7803 \quad (36)$$

95% Confidence Interval: $\rho \in [-0.8378, 0.7803]$

The interval is very wide and encompasses zero, consistent with the lack of statistical significance. This reflects the small sample size ($n = 6$) and weak observed correlation.

G.4 Interpretation of Six-Language Example

The six-language worked example demonstrates that even with a modest sample, the Spearman rank

correlation analysis reveals no statistically significant relationship between training data quantity and fine-tuned WER ($\rho = -0.0857$, $p = 0.87$). The point estimate of $\rho = -0.0857$ is substantially weaker than the full 19-language result ($\rho = -0.188$), which is expected due to sampling variability. The wide 95% CI ([-0.838, 0.780]) encompasses all plausible values from strong negative to strong positive correlation, indicating substantial uncertainty with small sample size.

This pedagogical example illustrates the methodology and shows that conclusions about the absence of a training-data–driven confound are robust across different language samples and sample sizes.

H Native-Speaker Audit Guidelines and Briefing Document

To ensure consistency across the distributed human evaluation, all native-speaker contributors were provided with a standardized reporting template and a reference example. Because annotators were drawn from community networks and were not formally trained linguists, the written documentation was supplemented with a verbal calibration and training. During this training, the research team discussed the specific error categories (e.g., hallucination, code-switching loss, orthographic errors) and how to identify them in the model outputs.

Below is the verbatim reference example provided to annotators to guide their evaluation process. This example outlines the expected structure, depth of analysis, and error categorization using Yoruba as the reference language.

Reference Briefing Example: Yoruba Zero-Shot Evaluation

Language: Yoruba
Assigned Leads: Full Name(s),
Models Evaluated: Whisper-Large-v3, MMS-1B, Omnilingual ASR

1. Overall Performance Ceiling & General Observations

Across all three models, achieving a 0% Word Error Rate (WER) on the WAXAL Yoruba test set is currently impossible due to the highly conversational, code-switched nature of the data. Modern spoken Yoruba frequently borrows from English and Pidgin, and none of the models seamlessly handle these transitions. Furthermore, tonal marking (which dictates meaning in Yoruba) remains a major vulnerability across the board.

2. Model-Specific Linguistic Behaviors

- **Whisper-Large-v3:** Whisper is surprisingly good at recognizing when a speaker switches to English. However, it severely struggles with Yoruba orthography. It frequently drops diacritical marks (tones and subdots like ẹ, ọ, ṣ).
  - *Example:* When the speaker said "Báwo ni nǹkan?" (How are things?), Whisper transcribed it flatly as "Bawo ni nkan", stripping the high and low tones.
  - *Hallucinations (M2):* Because it tries to make logical sentences, if it doesn't understand a Yoruba slang term, it will hallucinate an English word that sounds phonetically similar, completely changing the sentence's meaning.
- **MMS-1B:** MMS-1B respects Yoruba tonal marks and subdots much better than Whisper. However, it completely fails on conversational speech and modern loan words. It seems heavily biased toward formal or religious Yoruba text (likely due to its pre-training data).
  - *Example:* When a speaker used the modern loan-word "kọmputa" (computer), MMS-1B failed to transcribe it phonetically and instead output a string of unrelated, formal Yoruba syllables.
  - *Code-Switching Failure (C1 Error):* It cannot handle English code-switching at all; when an English word is spoken, the model either deletes it or outputs gibberish.
- **Omnilingual ASR:** Omnilingual ASR balances better between formal and informal speech, but it suffers from severe Orthographic Errors (O1). It mixes standard Yoruba spelling conventions with outdated ones.
  - *Example:* It frequently misidentifies the "ṣ" (sh-sound) as a regular "s", writing "se" instead of "ṣe" (to do). While a human reader can guess the meaning from context, it heavily penalizes the model's WER.

3. Top Error Categories Identified

- **P1 (Tonal Confusion):** The most common error across all models. For example, confusing igbá (calabash), igba (two hundred), and ìgbà (time) because the models rely on phonetic consonants rather than pitch contours.

- **C1 (Code-Switching Loss):** WAXAL speakers naturally weave English into sentences (e.g., "Mo fẹ lọ sí market"). MMS deletes "market", while Whisper might try to spell "market" using Yoruba alphabet rules ("makẹti"), which mismatches the reference text.