Beyond Aggregate WER: Mapping Linguistic Failure Modes to Data Prescriptions in African ASR

Anonymous ACL submission

Abstract

Aggregate Word Error Rate (WER) masks critical, real-world failure modes in Automatic Speech Recognition (ASR), a limitation that is particularly consequential for linguistically diverse African languages. In this work, we present the first systematic, fine-grained error taxonomy for African ASR, analyzing approximately 600,000 predictions across 24 languages and five distinct models. Moving beyond descriptive benchmarking, we propose a three-axis diagnostic framework that operationalizes specific error profiles—substitution/deletion/insertion (SDI) distributions, named entity transcription failures, and numerical formatting errors—into targeted data diagnostics. Our analysis reveals that deletion errors are disproportionately concentrated in morphologically rich languages and persist even at the 7B parameter scale, indicating a systematic failure mode that scaling alone cannot resolve. Furthermore, we demonstrate near-total failure in structured number transcription across all architectures and expose a broadly stable, model-agnostic hierarchy of difficulty for named entities (ORG > PER > LOC > MISC > DATE), with strict entity error rates ranging from 45.0% (DATE) to 67.1% (ORG), and tonal languages suffering the most severe degradation. Finally, we find preliminary evidence consistent with our framework using a regionally-tuned model, demonstrating a strong positive correlation ($r = 0.659, p = 0.001$) between a language's baseline deletion rate and its relative Word Error Rate reduction. These results suggest that fine-grained error profiles can diagnose and prioritize hypothesized linguistic data requirements.

1 Introduction

Automatic Speech Recognition (ASR) has transitioned from a specialized tool to a foundational interface across global enterprise and consumer domains. In customer support, it facilitates real-time intent detection and agent assistance (Nambiar et al., 2011); in healthcare, voice-enabled digital scribes alleviate clinician documentation burdens (van Buchem et al., 2021; Sanni et al., 2025); and in legal settings (Saadany et al., 2023), it powers complex proceedings transcription (Olatunji et al., 2023b; Ashungafac et al., 2025). Today, the challenge for practitioners is no longer training bespoke models but selecting the optimal pre-trained foundation system—ranging from self-supervised models like wav2vec 2.0 (Baevski et al., 2020) to multitask giants like Whisper (Radford et al., 2023) or Specialized speech recognition LLMs and general purpose multimodal LLMs (Adelani et al., 2025; Awobade et al., 2026).

However, the "scaling laws" that have perfected ASR for Western languages have not yet closed the persistent "generalization gap" for African linguistic contexts (Ashungafac et al., 2025). While global leaderboards often report Word Error Rates (WER) below 5% on standard English benchmarks , these models experience a 5x to 10x performance degradation when exposed to the rich phonetic and prosodic diversity of African accents (Olatunji et al., 2023b). This disparity is even more pronounced in indigenous African languages, where "supported" models often produce nearly unintelligible outputs with WERs exceeding 100% (Ashungafac et al., 2025).However, the dominant evaluation paradigm—aggregate Word Error Rate (WER)—provides an incomplete picture of model behavior in real-world deployment. A model achieving a seemingly acceptable WER may still catastrophically fail on the named entities, numerical expressions, and morphologically complex words that are most critical for downstream task correctness.

This limitation is particularly consequential for African language ASR, where models are evaluated on linguistically diverse datasets spanning multiple language families, morphological typologies, and tonal systems (Babatunde et al., 2023). Re-

cent benchmarks, such as AfriVox-v2 (Awobade et al., 2026), have documented the performance gap between African and high-resource language ASR. Yet, the literature has largely stopped short of characterizing *how* models fail—whether errors are substitutions, deletions, or insertions—and whether these failure patterns follow predictable linguistic structures.

In this work, we present the first systematic, fine-grained error taxonomy for African ASR across 24 languages, five models, and three error dimensions. Using approximately 600,000 reference and model hypothesis pairs from our results on the AfriVox-v2 benchmark dataset[1] (Awobade et al., 2026), we move beyond aggregate WER to propose a three-axis diagnostic framework that maps language-specific error profiles to hypothesized data requirements.

In summary, our core contributions are:

- **A Novel Diagnostic Framework:** We propose a three-axis evaluation framework (SDI distributions, named entities, and numerical formatting) that moves beyond aggregate WER to map specific linguistic failure modes to actionable data prescriptions in African ASR.

- **The Limits of Scaling on Morphology:** We demonstrate that deletion errors are disproportionately concentrated in morphologically rich languages and persist even at the 7B parameter scale, proving that model scaling alone is insufficient to resolve these systematic linguistic failures.

- **Exposure of Structured Formatting Collapse:** We reveal a near-total transcription failure on complex numerical formats across all architectures, alongside establishing a universal, model-agnostic hierarchy of difficulty for named entities (ORG $>$ PER $>$ LOC $>$ MISC $>$ DATE).

- **Empirical Framework Validation:** We validate the predictive power of our framework against Sahara-v2, demonstrating a strong positive correlation ($r = 0.659, p = 0.001$) between a language's baseline deletion rate and its relative WER reduction when explicitly tuned on regionally representative data.

[1] https://huggingface.co/datasets/intronhealth/afrivox-v2

2 Related Work

Beyond Word Error Rate The reliance on aggregate Word Error Rate (WER) as the primary evaluation metric in Automatic Speech Recognition (ASR) is well documented, yet it frequently masks critical linguistic failures. A 2025 systematic literature review confirmed that the vast majority of African ASR research relies exclusively on WER, with minimal application of linguistically informed evaluation metrics, particularly for tonal and morphologically rich languages (Imam et al., 2025). Recent work by Chen et al. (2026) addressed this limitation by proposing Feature Error Rate (FER), demonstrating that WER obscures specific phonological and tonal errors in languages such as Yoruba and Uneme. While their approach provides deep phonological insights for two languages, our work introduces a broader, computationally accessible word-level framework (SDI) scaled across 24 languages to directly guide data curation for practitioners.

African ASR Benchmarking The rapid expansion of African speech datasets has led to comprehensive evaluation benchmarks. Recent efforts by Elmadany et al. (2025); Awobade et al. (2026); Adelani et al. (2025); Ashungafac et al. (2025) benchmarked ASR models across many African languages, linking architectural design and pre-training coverage to linguistic complexity. While these efforts focus primarily on the impacts of fine-tuning and data scaling on aggregate performance, our work pivots from descriptive benchmarking to a diagnostic taxonomy, operationalizing specific error profiles into language-specific data diagnostics.

Entity and Numerical Transcription Accurate transcription of named entities and structured numerical formats remains a persistent challenge in low-resource settings. Olatunji et al. (2023a) highlighted this in *AfriNames*, demonstrating significant ASR degradation on African person names (e.g., "Ukachukwu", "Lakicia", "Ingabire") and proposing targeted data augmentation to mitigate the performance drop. However, *AfriNames* focuses exclusively on African-accented English and is restricted to person names. Our analysis significantly extends this scope by evaluating African languages across a complete entity typology (PER, LOC, ORG, MISC, DATE) and integrating it into a holistic diagnostic framework. Furthermore, our taxonomy introduces structured number transcription (e.g., ordinals, per-

centages)—an area with sparse literature in African ASR—as a critical failure dimension requiring dedicated linguistic formatting data.

3 Data and Models

3.1 Dataset

We utilize the AfriVox-v2 benchmark dataset[2], a comprehensive corpus of African speech (Awobade et al., 2026). The dataset spans 24 languages across major linguistic groupings—Niger-Congo, Afro-Asiatic, and Bantu—and includes African-accented French as a regionally localized variety distinct from its European counterpart and representative of Francophone African speech communities. The data represents diverse real-world domains, including agriculture, health, finance, and sports, allowing for robust cross-domain evaluation.

3.2 Evaluated Models

To ensure our error taxonomy captures architectural and scaling differences, we evaluate predictions from five distinct ASR systems:

- **Omni-CTC Suite**: Three general-purpose Conformer-CTC models at varying scales (300M, 1B, and 7B parameters) (team et al., 2025).
- **Gemini-3-Flash**[3]: A multimodal foundation model incorporating strong language modeling priors.
- **Sahara-v2**[4]: Sahara-v2 models explicitly fine-tuned on geographically representative African speech data.

4 Methodology

To move beyond aggregate Word Error Rate (WER), we propose a fine-grained, three-axis diagnostic framework analyzing error type distributions, named entity failures, and numerical transcription.

4.1 Error Type Distribution (SDI)

We compute word-level substitutions, deletions, and insertions using the `jiwer` library (Vaessen, 2018). To normalize for sequence length, error counts for each sample are divided by the total number of words in the reference text. We specifically isolate the **Deletion Rate** (D/N, where N is total reference words) as a primary metric for quantifying morphological dropping.

4.2 Named Entity Recognition (NER) Error Analysis

To extract named entities from reference transcriptions, we utilize a coverage-based ensemble of two NER models: the Masakhane afroxlmr-large model (Adelani et al., 2021) as the primary model for the 14 languages within MasakhaNER's coverage, and roberta-large-NER[5] as a fallback for the remaining 10 languages outside its coverage. Because ASR models frequently capture the phonetic shell of an entity while failing exact orthographic reconstruction, we evaluate entity preservation on two levels. **Strict Error** occurs when the hypothesized entity span fails to exactly match the reference string. **Partial Error** is recorded only when the model fails to capture even a substring or single token of the multi-word entity, indicating a total hallucination or complete omission.

4.3 Number Format Classification

We extract numerical tokens from the reference text using regex-based pattern matching. Numbers are classified into six structural formats: cardinal, ordinal, currency, date, percentages, and phone numbers. We then calculate the exact-match error rate to evaluate whether the corresponding ASR hypothesis correctly transcribes these structured formats. Languages for which the regex pipeline finds no matched tokens for a given format are excluded from that format's analysis—this is the basis for Bemba's N/A designation in Table 1.

4.4 Diagnostic Severity Thresholds

To construct the diagnostic profiles presented in Table 1, we categorized per-language performance into discrete severity bands. These thresholds were determined empirically based on the distributional quartiles of errors across all 24 languages and five models:

- **Deletion Severity:** Low ($< 4\%$), Med (4–7%), High ($> 7\%$).
- **Entity Error Severity:** Evaluated on Strict Error. Low ($< 40\%$), Med (40–55%), High ($> 55\%$).

[2] https://huggingface.co/datasets/intronhealth/afrivox-v2

[3] https://blog.google/products-and-platforms/products/gemini/gemini-3/

[4] https://www.intron.io/sahara-v2/

[5] https://huggingface.co/51la5/roberta-large-NER

- **Number Error Severity:** Evaluated on aggregate structured number error. Low ($< 40\%$), Med (40–80%), High ($> 80\%$).

By bounding these severities, we provide a standardized, model-agnostic mechanism to identify which linguistic failure modes require immediate data interventions for a given language.

4.5 Statistical Significance

To ensure the reliability of our findings, we compute 95% bootstrap confidence intervals ($n = 200$ resamples) for all reported WER and error rate metrics. Non-overlapping confidence intervals are used to establish the statistical significance of performance differences between models and language families.

5 Results

5.1 Error Distribution and the Deletion Bias

While aggregate WER implies that models struggle uniformly with specific languages, our fine-grained analysis reveals that the underlying error distributions are heavily partitioned by linguistic structure. Substitution is the dominant error class across the baseline Conformer-CTC architectures, accounting for 63–86% of total errors. This indicates that acoustic confusion and vocabulary mismatch remain the primary hurdles for African ASR.

However, deletion errors expose a systematic failure mode tied to morphological complexity. Deletion rates are disproportionately concentrated in agglutinative Bantu and Niger-Congo languages. For example, Pedi (9.6%), Sesotho (9.4%), and Twi (8.2%) exhibit deletion rates significantly higher than the cross-lingual average. To illustrate the qualitative character of these errors, Table 2 presents representative examples drawn from each error category. Substitution errors are often phonetically plausible but semantically disruptive: in Swahili, the reference clause **uuzaji wa** (sales of) is mistranscribed as **soko ya** (market of). Deletion errors are particularly concentrated in morphologically complex tokens: a long multi-word clause in Luganda (*balinga abava ku ssomero*) is entirely omitted from the hypothesis, consistent with the systematic morphological dropping we observe at scale.

One result warrants particular attention. Bemba exhibits a mean WER of 79.2%, more than 14 percentage points above the next worst performing language (Fulani, 64.7%), despite belonging to the same Bantu family as several mid-performing languages. This gap is not straightforwardly explained by our three-axis framework—Bemba shows high deletion severity but low entity error severity, and could not be evaluated on numerical formats due to insufficient test set tokens. Whether this outlier reflects a benchmark artifact in AfriVox-v2's Bemba subset, extreme pretraining underrepresentation across all five evaluated models, or a linguistic property not captured by our current diagnostic axes remains an open question. We report Bemba's results for completeness but caution against strong conclusions pending further investigation, and flag it as a priority case for future diagnostic work.

Crucially, this deletion bias persists regardless of model architecture or parameter count. While scaling the Omni-CTC model from 300M to 7B parameters reduces the deletion rate in Pedi from 13.5% to 9.1% (a 32% relative reduction), substitutions reduce by only 20% over the same scale. Despite this proportional improvement, absolute deletion rates remain heavily elevated even in the 7B parameter regime, suggesting that scaling alone cannot resolve the morphological dropping inherent to these languages.

5.2 Named Entity Transcription Failures

Proper noun recognition is a critical requirement for downstream ASR utility. Our alignment-based evaluation reveals that entity transcription failures are pervasive across African languages, while also exposing a systematic gap between strict preservation and partial preservation that has practical significance for downstream systems. Restricting our evaluation to languages with high-confidence automated NER capabilities, we observe that even the best-performing models fail on more than half of all transcribed entities.

The entity error hierarchy under strict evaluation follows a broadly stable, model-agnostic structure: ORG (67.1%) > PER (59.8%) > LOC (48.8%) > MISC (45.8%) > DATE (45.0%). Alongside strict error, we report partial error—the rate at which at least one entity token is lost—which is 20–34 percentage points lower across all types. This indicates that models frequently produce recognizable fragments of entities even when full preservation fails, as demonstrated by the phonetic mapping of *Mossack Fonseca* → *mossak fonseka* in Table 2. While this ranking holds robustly across the three Omni-CTC scales and Sahara-v2, Gemini-3-Flash shows partial reordering in the mid-tier entities

Language	Family	Avg WER	Deletion Severity	Entity Error Severity	Number Error Severity
Kinyarwanda	Bantu	24.3%	Low	Med	High
Swahili	Bantu	26.2%	Low	High	High
Pidgin	Niger-Congo	32.2%	Low	Med	Med
Tswana	Bantu	32.3%	Med	High	Med
Hausa	Afro-Asiatic	39.1%	Med	Med	Med
Xhosa	Bantu	41.2%	Med	Med	Med
Zulu	Bantu	42.4%	Low	Med	Med
Shona	Bantu	42.4%	High	Med	Med
Arabic	Afro-Asiatic	43.4%	Low	Low	High
Pedi	Bantu	46.2%	High	High	Med
Wolof	Niger-Congo	47.2%	Med	High	Med
Amharic	Afro-Asiatic	47.3%	Low	Med	High
Oromo	Afro-Asiatic	49.3%	Med	Med	High
Sesotho	Bantu	50.2%	High	High	Med
Luganda	Bantu	50.7%	Med	Med	High
Igbo	Niger-Congo	53.8%	Med	High	Med
Akan	Niger-Congo	56.3%	Med	Med	Med
Yoruba	Niger-Congo	57.8%	Med	High	Med
Twi	Niger-Congo	58.4%	High	Med	Med
Fulani	Niger-Congo	64.7%	Med	High	High
Bemba	Bantu	79.2%	High	Med	N/A*

Table 1: Diagnostic Framework Mapping: Per-Language Failure Profiles. Severity thresholds are defined based on distributional quartiles. Deletion: Low ($< 4\%$), Med (4–7%), High ($> 7\%$). Entity Error: Low ($< 40\%$), Med (40–55%), High ($> 55\%$). Number Error: Low ($< 40\%$), Med (40–80%), High ($> 80\%$). *Bemba lacked sufficient numerical tokens in the test set for reliable evaluation.

(PER, LOC), likely reflecting its stronger language modeling priors.

Furthermore, entity error rates are notably worse for tonal languages. Niger-Congo languages with complex tonal systems, including Yoruba (66.3%), Igbo (55.7%), and Akan (50.0%), exhibit elevated strict entity error rates. Conversely, Arabic (35.1%), Twi (40.9%), and Kinyarwanda (42.9%) achieve the lowest strict error rates among languages with large evaluation sets, aligning with the tighter 95% bootstrap confidence intervals observed for baseline architectures (e.g., Kinyarwanda achieving a tight 14.7–15.2% WER interval under the top-performing model).

5.3 Numerical Format Transcription

The transcription of structured numerical formats represents the most severe failure mode across all evaluated ASR systems. Our analysis reveals a catastrophic collapse in structured number transcription across all architectures.

While baseline cardinal numbers exhibit a high mean error rate of 53.1%, more complex numerical formats suffer from near-total transcription failure. Most notably, percentage transcriptions experienced a 100% error rate across evaluated models, failing to accurately map symbols or text equivalents. Similarly, phone number formatting and currency amounts yielded devastating mean error rates of 85.7% and 69.5%, respectively. This indicates that while current architectures can occasionally capture raw numeric values, they entirely lack the formatting priors necessary for domain-specific numerical transcription in African languages.

6 A Diagnostic Framework for Targeted ASR Improvement

The central limitation of aggregate WER is its inability to inform targeted model improvement. To bridge the gap between error analysis and model development, we propose a three-axis diagnostic framework that maps language-specific error profiles to hypothesized data requirements. Rather than uniformly prescribing "more data" for underperforming languages, this framework isolates specific linguistic failure modes to determine the *kind* of data likely required.

6.1 The Three Diagnostic Axes

Our framework categorizes ASR failures along three dimensions:

- **Error Type Distribution (SDI)**: A high substitution rate typically indicates generalized

Error Category	Language	Reference Text Span (REF)	Model Hypothesis Span (HYP)
Substitution	Swahili	**uuzaji wa** viatu ni biashara yenye...	**soko ya** viatu ni soko ambayo...
Deletion	Luganda	...tulaba abaana, **balinga abava ku ssomero** nga...	...tulaba abaana, (*[DEL]*) nga...
Entity (Strict)	Igbo	Dọkịta **Johnson** chọpụtara na...	dọkịta **jonsịn** chọpụtara na...
Entity (Partial)	Hausa	Malam **Isa** ya yi kira ga...	Malam **Isah** yayi kira ga...
Number Format	Igbo	...ka ha chekwaa opekata mpe **$10,000**	...ka ha chekwa opekata mpe **1000**
Insertion	Swahili	Hivi karibuni (Siku mbili zilizopita)...	hivi karibuni **fungwa mabano** siku mbili zilizopita **funga mabano**...

Table 2: Representative error examples carefully selected across diverse linguistic families to illustrate persistent, continent-wide ASR failure modes. Notably, models suffer catastrophic collapse on numbers ($10,000 → 1000), output spoken punctuation ("fungwa mabano" translating to open bracket), entirely drop complex multi-word clauses (Deletion in Luganda), and phonetically approximate named entities ("Johnson" → "jonsịn").

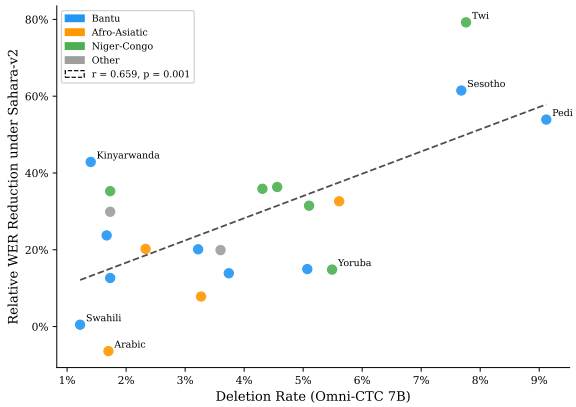

Figure 1: Correlation between baseline deletion rate (Omni-CTC 7B) and relative WER reduction achieved by the Sahara-v2 model ($r = 0.659, p = 0.001$).

acoustic confusion or vocabulary mismatch, requiring broader acoustic coverage. However, a disproportionately high deletion rate suggests models are systematically dropping morphological affixes, signaling a need for morphologically diverse, affix-rich training text rather than simply more audio hours.

- **Entity Failure Profile**: Entity errors expose specific vocabulary gaps. High error rates on person names (PER) suggest out-of-vocabulary (OOV) named entities, addressable via named entity lexicon augmentation. Conversely, high MISC errors point to a lack of domain-specific terminology, requiring targeted domain vocabulary coverage.

- **Number Format Profile**: Catastrophic failure on ordinals and percentages indicates a format mismatch, requiring training sets that explicitly map spoken number conventions to structured numerical text.

We apply this framework to our evaluated languages in Table 1, categorizing the severity of each failure mode to provide a per-language data diagnostic.

To illustrate the framework concretely, consider Pedi. Pedi exhibits the highest deletion rate in our dataset (9.6%), high entity error severity, and medium number error severity (Table 1). Under our framework, this profile generates a specific, prioritized data prescription: the dominant need is morphologically diverse training text that explicitly covers Pedi's agglutinative affix system—particularly verb extensions and noun class prefixes that CTC-based models systematically drop. A secondary need is named entity lexicon coverage, particularly for person names and organizational terms. Number formatting data, by contrast, is a low priority for Pedi relative to other languages. This stands in contrast to a language like Swahili (WER 26.2%), whose low deletion rate (1.7%) but high number error severity (89.9%) points to a fundamentally different data requirement: structured numerical formatting data mapping spoken Swahili number conventions to cardinal, ordinal, and currency text. The framework thus does not prescribe "more data" uniformly—it prescribes different data for each language based on its specific failure profile.

6.2 Preliminary Evidence from Regional Training

To explore the predictive utility of our diagnostic framework, we examine whether a language's baseline deletion rate correlates with its degree of improvement under Sahara-v2, a regionally-tuned

model trained on geographically representative African speech data. We treat this as a preliminary stress-test of our framework's central hypothesis: that deletion errors reflect morphological coverage gaps addressable through targeted training data.

Figure 1 plots the relationship between per-language deletion rate (Omni-CTC 7B) and relative WER reduction achieved by Sahara-v2. We observe a strong positive correlation ($r = 0.659, p = 0.001, n = 20$), indicating that languages with higher deletion rates tend to benefit substantially more from regional training. Languages such as Pedi (deletion = 0.091, WER reduction = 52.5%) and Twi (deletion = 0.077, WER reduction = 74.1%) show the largest gains, while low-deletion languages such as Arabic (deletion = 0.017) show minimal or even negative relative reduction (-6.8%).

An important confound, however, limits the causal interpretation of this result. Sahara-v2's training corpus is regionally representative but not publicly characterized at the per-language level. Languages exhibiting high deletion rates in our taxonomy—predominantly agglutinative Bantu and Niger-Congo languages—may simply be the same languages that are most underrepresented in general-purpose pretraining corpora and best covered by Sahara-v2's regional data. Under this alternative explanation, the correlation would reflect a shared underlying variable (language resource level) rather than confirming that deletion rate specifically diagnoses the relevant data gap. Disentangling these two explanations would require either access to Sahara-v2's per-language training distribution or a controlled experiment in which affix-rich data is added in isolation and deletion rates are measured directly.

We therefore interpret this correlation as consistent with our framework rather than as direct validation of it. The result is encouraging—the diagnostic signal aligns with observed improvement patterns—but the causal claim that targeting deletion errors produces WER gains requires future experimental confirmation.

7 Conclusion

In this work, we demonstrated that aggregate Word Error Rate (WER) systematically masks critical, linguistically patterned failure modes in African ASR. By analyzing approximately 600,000 reference and model hypothesis pairs across 24 languages, we established a fine-grained error taxonomy revealing that morphological complexity drives persistent deletion errors, structured numbers fail catastrophically across architectures, and named entity transcription follows a broadly stable, model-agnostic hierarchy of difficulty.

Moving beyond descriptive evaluation, we introduced a three-axis diagnostic framework that operationalizes these specific error profiles into targeted data diagnostics. We find that a language's baseline deletion rate correlates with its relative WER reduction when evaluated on a regionally-tuned model (Sahara-v2). Ultimately, our findings provide ASR practitioners with a quantifiable, language-specific roadmap for data curation, shifting the focus from simply scaling data volume to strategically addressing structural linguistic gaps.

Ethics Statement

Potential Risks and Artifact Intended Use: While our diagnostic framework aims to mitigate architectural biases against underrepresented African languages, inaccurate error diagnoses could lead to suboptimal dataset curation strategies. All existing software artifacts and public datasets utilized in this study (`jiwer`, MasakhaNER, AfriVox-v2) are employed strictly in accordance with their original open-source research licensing conditions. The error taxonomy and diagnostic thresholds introduced here are intended solely for academic research and model evaluation purposes.

Data Privacy, Anonymization, and Content Filtering: We rely exclusively on secondary data from the publicly released AfriVox-v2 benchmark. Personally identifiable information (PII) scrubbing and offensive content filtering were managed upstream by the dataset creators; our evaluation pipeline processes these pre-scrubbed transcriptions and positional alignments, posing zero additional downstream privacy risks.

Human Participants and Crowdsourcing: This work did not recruit, employ, or engage any human participants, annotators, or crowdsourced workers. All evaluations were executed through automated inference and algorithmic extraction pipelines, rendering participant instructions, consent protocols, recruitment mechanisms, and demographic compensation metrics not applicable to this study.

Statistical Transparency: To ensure reproducibility and scientific transparency, all quanti-

tative results in this paper are reported as mean values across the evaluated corpora. Performance variance is rigorously captured using 95% bootstrap confidence intervals calculated over 200 resamples, explicitly detailed in both our primary methodology section and the complete results tables.

Limitations

While our diagnostic framework provides empirically grounded data diagnostics, it fundamentally serves as a hypothesis-generating tool. The framework's core premise—that specific linguistic failure modes can be resolved by corresponding targeted data augmentations—is theoretically motivated and supported by our correlational analysis with Sahara-v2, but it requires direct experimental validation in future work. Future studies must demonstrate that explicitly adding affix-rich text or domain-specific lexicons causally reduces deletion and MISC entity errors, respectively.

Our entity error analysis uses positional matching between reference and hypothesis tokens, which assumes approximate word-level alignment—an assumption that may not hold for languages with high deletion rates, potentially inflating entity error rates for morphologically complex languages.

Additionally, our methodology relies on automated extraction pipelines rather than human-annotated ground truth. Our named entity error analysis utilizes a coverage-based ensemble of automated NER models. While this ensures broad language coverage, automated entity annotations may still introduce boundary errors in conversational speech, which could inflate reported ASR error rates. Future work should validate these findings using human-annotated spans. Similarly, our numerical format analysis utilizes regex-based pattern matching on reference texts, which captures conventionally formatted numbers but may miss inherently spoken number forms.

Finally, we note that ASR failure profiles may differ in code-switched speech, limiting generalizability beyond the evaluated domains.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, and 1 others. 2025. Irokobench: A benchmark for african languages in the age of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2732–2757, Mexico City, Mexico. Association for Computational Linguistics.

Gabrial Zencha Ashungafac, Mardhiyah Sanni, Busayo Awobade, Alex Gichamba, and Tobi Olatunji. 2025. Afrispeech-multibench: A verticalized multidomain multicountry benchmark suite for african accented english asr. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3642–3653.

Busayo Awobade, Gabrial Zencha Ashungafac, and Tobi Olatunji. 2026. Afrivox-v2: A domain-verticalized benchmark for in-the-wild african speech recognition. *Preprint*, arXiv:2605.03590.

Oreoluwa Boluwatife Babatunde, Emmanuel Akeweje, Sharon Ibejih, Victor Tolulope Olufemi, and Sakinat Oluwabukonla Folorunso. 2023. Automatic speech recognition for nigerian-accented english. In *Submitted to Deep Learning Indaba 2023*. Under review.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

Fei-Yueh Chen, Lateef Adeleke, and C. M. Downey. 2026. Linguistically informed evaluation of multilingual ASR for African languages. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 149–162, Rabat, Morocco. Association for Computational Linguistics.

AbdelRahim A. Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Mapping Africa's speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11028–11050, Suzhou, China. Association for Computational Linguistics.

Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie

Yimam, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition (asr) for african low-resource languages: A systematic literature review. *Preprint*, arXiv:2510.01145.

Ullas Nambiar, Tanveer Faruquie, L. Venkata Subramaniam, Sumit Negi, and Ganesh Ramakrishnan. 2011. Discovering customer intent in real-time for streamlining service desk conversations. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 1383–1388, New York, NY, USA. Association for Computing Machinery.

Tobi Olatunji, Tejumade Afonja, Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Chris Chinenye Emezue, Amina Mardiyyah Rufai, and Sahib Singh. 2023a. AfriNames: Most ASR Models "Butcher" African Names. In *Interspeech 2023*, pages 5077–5081.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, and 1 others. 2023b. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1599–1617.

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

Hadeel Saadany, Catherine Breslin, Constantin Orasan, and Sophie Walker. 2023. Better transcription of uk supreme court hearings. In *AI4AJ@ICAIL*.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Kayande, Emmanuel Ayodele, Naome Etori, Michael Mollel, and 1 others. 2025. Afrispeech-dialog: A benchmark dataset for spontaneous english conversations in healthcare and beyond. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.

Nikolay V. Vaessen. 2018. Jiwer: Similarity measures for automatic speech recognition evaluation. `https://github.com/jitsi/jiwer`.

Marieke M. van Buchem, Hileen Boosman, Martijn P. Bauer, Ilse M. J. Kant, Simone A. Cammel, and Ewout W. Steyerberg. 2021. The digital scribe in clinical practice: A scoping review and research agenda. *npj Digital Medicine*, 4:57.

8 Appendix

Table 3: Per-model, per-language WER and deletion rate (Del%) with 95% bootstrap confidence intervals. Languages sorted by average WER. Subscript/superscript denote CI lower/upper bounds. – indicates the model was not evaluated on that language.

Language	Omni-CTC 300M		Omni-CTC 1B		Omni-CTC 7B		Gemini-3-Flash		Sahara-v2	
	WER	Del%	WER	Del%	WER	Del%	WER	Del%	WER	Del%
Swahili	$35.8^{36.2}_{35.3}$	$2.6^{2.8}_{2.5}$	$29.3^{29.7}_{28.9}$	$1.8^{1.9}_{1.6}$	$26.9^{27.4}_{26.5}$	$1.2^{1.3}_{1.1}$	$19.8^{20.2}_{19.5}$	$0.9^{1.0}_{0.8}$	$19.3^{19.7}_{18.9}$	$1.8^{1.9}_{1.7}$
Kinyarwanda	$30.0^{30.4}_{29.6}$	$3.8^{4.0}_{3.7}$	$21.6^{22.0}_{21.2}$	$2.3^{2.3}_{2.2}$	$17.6^{18.0}_{17.3}$	$1.4^{1.5}_{1.3}$	$37.2^{37.5}_{36.9}$	$2.5^{2.5}_{2.4}$	$15.0^{15.2}_{14.7}$	$1.0^{1.1}_{1.0}$
Arabic	$53.9^{55.1}_{52.9}$	$3.5^{3.8}_{3.2}$	$42.9^{44.1}_{41.9}$	$1.7^{1.9}_{1.5}$	$41.1^{42.0}_{40.0}$	$1.7^{1.9}_{1.5}$	$35.2^{36.2}_{34.0}$	$0.8^{0.9}_{0.7}$	$43.9^{45.1}_{42.8}$	$1.8^{2.0}_{1.6}$
Tswana	$40.3^{40.8}_{39.7}$	$6.7^{7.0}_{6.5}$	$34.2^{34.7}_{33.6}$	$4.9^{5.1}_{4.7}$	$28.5^{29.0}_{27.9}$	$3.2^{3.4}_{3.0}$	$38.7^{39.3}_{38.0}$	$5.0^{5.2}_{4.8}$	$20.1^{20.5}_{19.6}$	$4.5^{4.7}_{4.3}$
Zulu	$49.1^{49.8}_{48.4}$	$3.4^{3.5}_{3.2}$	$44.7^{45.4}_{44.0}$	$2.5^{2.6}_{2.4}$	$40.6^{41.2}_{39.9}$	$1.7^{1.8}_{1.6}$	$42.9^{43.6}_{42.3}$	$2.4^{2.5}_{2.3}$	$34.7^{35.3}_{34.1}$	$3.5^{3.7}_{3.4}$
Pidgin	$46.1^{46.4}_{45.8}$	$3.5^{3.7}_{3.4}$	$31.3^{31.5}_{31.0}$	$1.8^{1.9}_{1.7}$	$29.5^{29.8}_{29.2}$	$1.7^{1.8}_{1.7}$	$28.4^{28.6}_{28.1}$	$0.9^{1.0}_{0.9}$	$25.8^{26.0}_{25.6}$	$1.5^{1.6}_{1.5}$
Hausa	$51.6^{52.4}_{51.0}$	$6.9^{7.2}_{6.6}$	$43.8^{44.4}_{43.0}$	$5.1^{5.4}_{4.8}$	$42.0^{42.7}_{41.2}$	$5.6^{6.0}_{5.2}$	$31.3^{32.0}_{30.7}$	$3.4^{3.6}_{3.2}$	$26.9^{27.4}_{26.4}$	$2.6^{2.7}_{2.4}$
Xhosa	$49.2^{50.0}_{48.6}$	$5.0^{5.2}_{4.8}$	$43.9^{44.6}_{43.3}$	$3.5^{3.7}_{3.3}$	$35.5^{36.1}_{34.7}$	$3.7^{4.0}_{3.5}$	$47.0^{47.7}_{46.3}$	$3.9^{4.0}_{3.7}$	$30.2^{30.8}_{29.7}$	$4.5^{4.8}_{4.3}$
Yoruba	$63.3^{63.7}_{62.9}$	$8.4^{8.5}_{8.2}$	$56.6^{57.0}_{56.2}$	$7.0^{7.1}_{6.8}$	$51.4^{51.9}_{51.0}$	$5.5^{5.6}_{5.3}$	$67.8^{68.7}_{67.2}$	$4.3^{4.4}_{4.2}$	$49.6^{50.0}_{49.1}$	$6.5^{6.7}_{6.3}$
Shona	$48.7^{49.9}_{47.5}$	$4.3^{4.6}_{3.9}$	$41.5^{42.8}_{40.2}$	$2.4^{2.7}_{2.2}$	$37.7^{38.9}_{36.5}$	$1.7^{1.9}_{1.4}$	$59.4^{61.1}_{57.5}$	$31.4^{33.6}_{28.8}$	$24.8^{25.9}_{23.6}$	$3.9^{4.2}_{3.5}$
Oromo	$49.6^{49.9}_{49.2}$	$5.4^{5.6}_{5.3}$	$41.5^{41.9}_{41.1}$	$3.8^{3.9}_{3.6}$	$39.0^{39.3}_{38.6}$	$3.3^{3.4}_{3.1}$	–†	$4.9^{5.0}_{4.8}$	$32.0^{32.4}_{31.7}$	$4.7^{4.9}_{4.5}$
Igbo	$60.9^{61.3}_{60.5}$	$7.4^{7.6}_{7.2}$	$55.5^{55.9}_{55.0}$	$6.7^{6.8}_{6.5}$	$50.3^{50.8}_{49.9}$	$5.1^{5.3}_{4.9}$	$62.0^{62.7}_{61.3}$	$5.6^{5.8}_{5.5}$	$40.4^{40.8}_{40.0}$	$5.1^{5.3}_{5.0}$
Amharic	$56.1^{56.4}_{55.7}$	$6.0^{6.1}_{5.8}$	$45.9^{46.3}_{45.5}$	$3.6^{3.7}_{3.5}$	$41.8^{42.1}_{41.4}$	$2.3^{2.4}_{2.2}$	–†	$2.6^{2.7}_{2.5}$	$31.1^{31.5}_{30.8}$	$4.6^{4.8}_{4.4}$
Luganda	$67.4^{68.2}_{66.7}$	$10.9^{11.4}_{10.4}$	$57.9^{58.6}_{57.1}$	$7.9^{8.3}_{7.5}$	$51.6^{52.4}_{50.8}$	$5.1^{5.4}_{4.7}$	$42.0^{43.1}_{40.9}$	$4.6^{4.8}_{4.3}$	$34.8^{35.5}_{34.0}$	$5.1^{5.4}_{4.8}$
French	$63.6^{66.2}_{61.3}$	$9.3^{11.2}_{7.7}$	$51.7^{54.1}_{49.3}$	$5.5^{6.7}_{4.4}$	$48.1^{50.7}_{46.2}$	$3.6^{4.7}_{2.5}$	$24.7^{27.3}_{22.4}$	$0.8^{1.1}_{0.5}$	$24.9^{27.3}_{22.4}$	$2.2^{3.0}_{1.5}$
Wolof	–	–	$46.6^{47.6}_{45.7}$	$5.7^{6.0}_{5.4}$	–	–	$58.9^{59.6}_{58.3}$	$4.9^{5.1}_{4.7}$	$36.1^{36.7}_{35.5}$	$6.1^{6.3}_{5.9}$
Pedi	$56.0^{56.9}_{55.0}$	$13.5^{14.1}_{12.9}$	$49.4^{50.6}_{48.1}$	$10.9^{11.5}_{10.4}$	$44.8^{45.9}_{43.6}$	$9.1^{9.8}_{8.5}$	$59.6^{60.7}_{58.5}$	$9.0^{9.6}_{8.4}$	$21.2^{22.1}_{20.3}$	$5.5^{5.9}_{5.1}$
Sesotho	$63.6^{64.1}_{63.0}$	$12.2^{12.5}_{11.8}$	$57.9^{58.4}_{57.3}$	$9.3^{9.7}_{9.0}$	$54.1^{54.7}_{53.6}$	$7.7^{8.1}_{7.3}$	$51.4^{52.1}_{50.8}$	$11.4^{11.9}_{11.0}$	$24.0^{24.5}_{23.6}$	$6.4^{6.7}_{6.1}$
Twi	$77.2^{79.2}_{75.2}$	$11.1^{12.3}_{10.0}$	$73.4^{75.5}_{71.0}$	$12.6^{14.0}_{11.3}$	$66.2^{68.5}_{63.7}$	$7.7^{8.8}_{6.8}$	$58.2^{60.6}_{55.8}$	$6.8^{7.8}_{5.9}$	$17.1^{18.9}_{15.3}$	$2.6^{3.6}_{1.9}$
Akan	$68.8^{69.5}_{68.2}$	$7.0^{7.3}_{6.7}$	$64.3^{64.9}_{63.7}$	$6.8^{7.2}_{6.5}$	$60.0^{60.7}_{59.4}$	$4.3^{4.6}_{4.0}$	$53.2^{53.9}_{52.4}$	$5.2^{5.4}_{4.9}$	$35.1^{35.8}_{34.4}$	$5.7^{6.0}_{5.4}$
Fulani	$68.6^{69.1}_{68.0}$	$8.6^{8.9}_{8.4}$	$66.4^{66.9}_{65.9}$	$8.7^{9.1}_{8.4}$	$65.5^{66.3}_{64.7}$	$4.6^{4.9}_{4.3}$	$79.2^{79.7}_{78.7}$	$7.2^{7.4}_{6.9}$	$44.0^{44.6}_{43.4}$	$4.3^{4.5}_{4.1}$
Bemba	–	–	–	–	–	–	$79.2^{80.7}_{77.7}$	$8.2^{8.6}_{7.9}$	–	–
Afrikaans	$55.8^{57.8}_{54.2}$	$2.3^{2.6}_{2.1}$	$37.0^{38.2}_{35.6}$	$2.3^{2.5}_{2.0}$	$33.8^{35.3}_{32.5}$	$1.7^{1.9}_{1.5}$	$60.4^{61.8}_{59.1}$	$6.2^{6.7}_{5.7}$	$23.3^{24.6}_{22.4}$	$3.7^{4.1}_{3.4}$

Language	**CTC 300M**	**CTC 1B**	**CTC 7B**	**STT v2**
Luganda	10.89%	7.91%	5.06%	5.09%
Yoruba	8.38%	6.96%	5.49%	6.51%
Igbo	7.45%	6.66%	5.10%	5.15%
Amharic	5.97%	3.62%	2.32%	4.57%
Zulu	3.37%	2.47%	1.73%	3.53%
Swahili	2.63%	1.76%	1.23%	1.80%

Table 4: Per-Model Deletion Rates across selected linguistically diverse African languages. While scaling from 300M to 7B parameters yields absolute reductions in deletion rates, morphologically rich languages (e.g., Luganda, Yoruba, Igbo) maintain disproportionately elevated deletion baselines ($\geq$ 5.0%) even at the 7B parameter scale, demonstrating that parameter scaling alone is insufficient to resolve this failure mode.