# Beyond Exact Match: Unmasking WER Inflation in Akan ASR through Variant-Aware Evaluation

Anonymous Author(s)

Affiliation Withheld for Double-Blind Review

**Abstract.** Standard Automatic Speech Recognition (ASR) evaluation relies on Word Error Rate (WER), which treats every token-level mismatch as an error regardless of linguistic validity. For dialectally diverse languages like Akan, phonological variations, spelling discrepancies, and code-switching markers are penalized identically to genuine misrecognitions, inflating reported error rates. We introduce a variant-aware WER (VA-WER) that normalizes predictions and references through a canonical lexicon compiled exclusively from training and validation data before scoring. Our framework classifies near-miss word pairs into a five-tier linguistic taxonomy using an LLM-as-a-Judge approach validated against native-speaker human annotators, then compiles frequency-based canonical lexicons. Across three Akan ASR models (Whisper-large-v3, Gemma-3n-2B, MMS-1b-all), we demonstrate that: (1) standard WER is inflated by 15.0%–17.0% (relative) due strictly to legitimate phonological and code-switching variations, rising to 22.4%–25.4% when human spelling inconsistencies are included; (2) the LLM judge achieves 81.4% agreement with native speakers when orthographic categories are collapsed, and sensitivity simulations ($N = 50$ Monte Carlo trials) confirm that metric reduction remains statistically robust ($p < 10^{-4}$) even under substantial boundary noise; and (3) lexicons from spontaneous speech corpora transfer to independent Akan datasets, confirming that captured variants represent systematic linguistic phenomena rather than corpus-specific noise. We advocate reporting VA-WER alongside standard WER in orthographically fluid languages to distinguish genuine model failures from benign dialectal variations.

**Keywords:** Automatic Speech Recognition · Word Error Rate · Low-Resource Languages · Akan · Orthographic Variation · Variant-Aware Evaluation

## 1 Introduction

The evaluation of Automatic Speech Recognition (ASR) systems relies almost universally on Word Error Rate (WER) [12, 13], which calculates the minimum word-level insertions, deletions, and substitutions required to align a machine transcription with a reference. While this exact-match standard functions adequately for high-resource languages with enforced orthographic rules, its application to low-resource and dialectally diverse languages routinely misrepresents

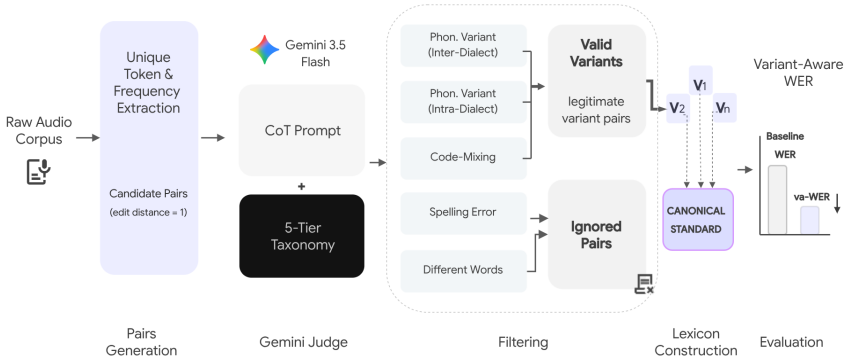



Fig. 1: End-to-end pipeline of the variant-aware evaluation framework: vocabulary extraction, near-miss candidate generation, and LLM-as-a-Judge taxonomy classification.

model performance [5, 9]. In African language ecologies where oral traditions dominate and formal orthographies remain fluid, standard WER penalizes valid regional and phonological variation identically to acoustic recognition failures. For example, in Akan, predicting "afuo" instead of reference "ɛfuo" (both meaning "farm") triggers a full substitution penalty despite identical semantic content.

In this paper, we show that standard WER is systematically inflated for Akan, conflating valid linguistic variation with model failure. We deconstruct the primary sources driving this inflation and propose targeted remedies. Sourcing a diverse corpus, we use an LLM-as-a-Judge framework validated by human annotators to automatically categorize error types and build canonical normalization lexicons. Our main contributions are:

- **Variant-Aware Framework & Leakage-Free Lexicon Generation:** We present an end-to-end framework for building frequency-based canonical lexicons for Akan, systematically isolating dialectal and orthographic variations across three datasets while restricting compilation strictly to training and validation splits.
- **Quantifying Metric Inflation:** We demonstrate how standard WER is systematically inflated, explicitly deconstructing inflation to isolate human spelling noise from valid linguistic variations (such as dialectal shifts and code-switching).
- **Statistical Robustness & Sensitivity Analysis:** We provide analytical bounds and Monte Carlo sensitivity simulations ($N = 50$) demonstrating that variant-aware error reduction remains statistically robust ($p < 10^{-4}$) even under substantial LLM classification noise.
- **Cross-Corpus Transferability:** We demonstrate that canonical lexicons transfer asymmetrically across Akan corpora, with spontaneous-speech lexicons providing substantially greater normalization than read-speech lexicons.

## 2 Previous Work

**WER Limitations and Invariant Scoring.** Standard WER penalizes all surface mismatches identically. Alternative metrics attempt to capture linguistic reality: embedding-based distances like SemDist [7] evaluate semantic preservation, while Character Error Rate (CER) reduces penalties for morphologically complex languages [6]. In Dravidian speech, SCRIBE demonstrates that valid compound merges inflate WER by up to 30% relative [8]. Closest to our work, OIWER [1] generates Indic spelling variants via LLMs, and SN-WER [11] applies multi-script transliteration. However, neither addresses African dialect continua where phonological shifts, not just script variation, drive metric inflation.

**African ASR and Akan Orthographic Fluidity.** Benchmarking across African languages shows steep accuracy degradations in multilingual foundation models like Whisper and MMS [12, 13]. Mensah et al. [9] benchmark seven Akan models and emphasize the need to decouple dialectal variants from acoustic failures, while Adeeba et al. [5] highlight standard WER's inadequacy for tonal, diacritically rich African languages. In Akan, orthographic fluidity stems from three historical literary traditions (Basel Mission 1842, Christaller 1875, and the Akan Orthography Committee 1978) that continue to coexist in modern usage [3]. Consequently, transcribers inevitably mix conventions, penalizing models that output legitimate regional alternations.

## 3 Taxonomy of Orthographic Variation in Akan

To deconstruct and analyze the inflation of standard WER, it is necessary to distinguish between genuine recognition failures and superficial script mismatches. Because Akan is a morphologically rich language characterized by distinct regional dialects (Asante Twi, Akuapem Twi, Fante) and fluid spoken patterns, human transcriptions often vary despite sharing high phonemic overlap. To decouple acoustic recognition accuracy from orthographic preferences, we establish a structured five-tier taxonomy:

1. **Intra-Dialect Phonological Variant:** Pronunciation shifts deviating from standard orthography. Typical cases include intervocalic rhotacism where 'd' becomes 'r' (e.g., standard "akwadaa" vs. spoken "akwaraa").
2. **Inter-Dialect Phonological Variant:** Systematic phonological differences between standard dialect groups where both are valid (e.g., Asante "adwuma" vs. Fante "edwuma").
3. **Spelling Error:** Violation of standardized orthographic conventions by transcribers without phonological justification (e.g., *bebree* vs. *beberee*).
4. **Code-Switching:** Phonetic adaptation of English loanwords into Akan orthography versus English spelling (e.g., "kar" vs. "car").
5. **Different Words:** Genuine semantic or grammatical errors altering word meaning (e.g., verb "ɛte" [sit] vs. "ɛye" [good]; plural "ntokua" vs. singular "tokua").

In our evaluation pipeline, categories (1), (2), and (4) form the basis of the strict VA-WER canonical lexicon; category (3) is evaluated in a separate ablation to isolate transcriber noise; and category (5) is strictly discarded.

## 4 Methodology

Our framework operates in four stages: candidate pair generation, LLM-Judge classification, canonical lexicon construction, and variant-aware WER computation (Figure 1). The evaluation pipeline operates on the offline transcription outputs of models fine-tuned on NVIDIA H100 GPUs. The metric computation itself is lightweight: candidate generation, graph-based canonicalization, and text normalization execute on commodity multi-core CPU, while classification is handled via the Gemini 3.5 Flash API, processing ∼44,000 candidate pairs (∼15M input tokens) across all corpora.

### 4.1 Vocabulary Extraction and Distance Computation

To prevent test-set leakage, unique tokens and corpus frequencies are extracted exclusively from training and validation splits; the test split remains completely unseen during candidate pair generation and lexicon compilation. A pairwise Levenshtein distance algorithm is executed across the training vocabulary to isolate all pairs within an exact edit distance of $d = 1$, capturing the systematic near-miss distribution (Appendix A).

### 4.2 LLM Judge Auto-rating and Reliability

An LLM judge (Gemini 3.5 Flash) classifies candidate pairs into the five-tier taxonomy using Chain-of-Thought (CoT) prompting with in-context sentence examples. For every candidate pair $(W_1, W_2)$, we provide two naturally occurring corpus sentences $(S_1, S_2)$ containing $W_1$ and $W_2$ respectively. The prompt enforces a four-step reasoning sequence: (1) identify character-level differences, (2) verify phonological or dialectal validity, (3) assess semantic and grammatical preservation across contexts, and (4) assign the final taxonomy category (Appendix H).

**Classification Noise and Sensitivity Analysis.** While the judge achieves 81.4% agreement with native speakers on collapsed orthographic categories (Section 6.2), we stress-tested its reliability through analytical bounding and Monte Carlo simulations ($N = 50$) (Appendix G). Under an adversarial bound treating every human-judge boundary disagreement as an erroneous match, relative WER reduction remains 8.23%–9.38% across architectures ($p < 10^{-4}$). In Monte Carlo trials purging 45.83% of the phonological lexicon to simulate boundary noise, VA-WER reductions remain substantial across architectures ($8.92\% \pm 1.20\%$ for Whisper, $7.94\% \pm 1.02\%$ for Gemma, $8.69\% \pm 1.10\%$ for MMS; all 95% CIs strictly positive). Thus, the conclusions of VA-WER are fundamentally structural rather than artifacts of classifier noise.

| Metric / Stage | WAXAL | FID | GhanaNLP |
|---|---|---|---|
| Unique vocabulary ($d = 1$, Train+Val) | 15,805 | 418 | 4,681 |
| Candidate near-miss pairs | 37,214 | 566 | 6,371 |
| Final Canonical Lexicon Size | 7,434 | 25 | 519 |

Table 1: Vocabulary extraction and resolved canonical lexicon sizes across corpora.

### 4.3 Canonical Lexicon Construction

All pairs classified as 'Different Words' are permanently discarded. For remaining pairs, we utilize a dual-routing strategy:

- **Strict va-WER Lexicon:** Retains strictly legitimate phonological variants and code-switching, excluding 'Spelling Errors'.
- **Ablation Lexicon:** Retains 'Spelling Errors' to quantify the volume of metric inflation caused by transcriber inconsistency.

A flat lexicon $L : \text{variant} \mapsto \text{canonical}$ is compiled using corpus frequency as the standardization criterion, resolving circular references and transitive chains via graph algorithms (Appendix C).

### 4.4 Normalization and Scoring

Both reference and prediction strings are normalized prior to scoring: each token $t$ found in $L$ is replaced with its canonical form $c(t)$. Standard WER is then computed on the normalized strings:

$$\text{WER}_{\text{norm}} = \text{WER}\left(\text{normalize(ref)}, \text{normalize(pred)}\right) \tag{1}$$

## 5 Experimental Setup

We evaluate on the held-out test split of WAXAL [2] across three foundation speech models fine-tuned on Akan: Whisper-large-v3, Gemma-3n-2B, and MMS-1b-all. Canonical lexicons are compiled from WAXAL, FID [10], and GhanaNLP [4] strictly using training and validation splits. We evaluate performance via absolute error recovery ($\Delta\text{Abs} = \text{WER}_{\text{std}} - \text{WER}_{\text{va}}$) and relative reduction ($\Delta\text{Rel} = \frac{\Delta\text{Abs}}{\text{WER}_{\text{std}}} \times 100\%$).

## 6 Results and Discussion

### 6.1 Lexicon Statistics

Table 1 reports candidate pair generation and resolved lexicon sizes. In spontaneous WAXAL speech, over 30% of candidate pairs represent valid linguistic variants, yielding 7,434 resolved canonical mapping entries. In contrast, scripted read-speech in FID yields only 25 valid mappings, demonstrating that collection modalities dictate transcription noise (detailed distribution in Appendix B).

| | Std. | Strict va-WER | | Fully-Norm. WER | | Points Recovered ($\Delta$) | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **WER** | **Score** | **$\Delta$ Rel** | **Score** | **$\Delta$ Rel** | **Phono.** | **Code-Sw.** | **Spelling** |
| Gemma-3n-2B | 40.00% | 34.02% | 14.95% | 31.02% | 22.45% | 5.87 | 0.11 | 2.54 |
| Whisper-large-v3 | 38.97% | 32.35% | 16.99% | 29.07% | 25.40% | 6.47 | 0.15 | 2.73 |
| MMS-1b-all | 39.54% | 33.14% | 16.19% | 30.01% | 24.10% | 6.26 | 0.14 | 2.60 |

Table 2: Standard WER vs. strict VA-WER and fully-normalized WER (including spelling) on the held-out WAXAL test set using the train+val canonical lexicon. Category recovery columns report isolated single-class evaluations; joint multi-class normalization yields an additional 0.46–0.55 point non-linear alignment synergy.

### 6.2 LLM Reliability Analysis

Evaluating the LLM judge against a 20-pair human-consensus gold standard revealed a 75.0% alignment rate (15/20 correct; 95% CI: [53.1%, 88.8%]). Independent evaluations on the calibration set prior to joint adjudication yielded a human Fleiss' Kappa of $\kappa = 0.2662$ (fair agreement), rising to $\kappa = 0.5087$ (moderate agreement) when grouping orthographic and phonological variants. This confirms that boundary subjectivity is intrinsic to native-speaker transcription rather than classifier error. Across the 236 complete unique human-annotated samples ($N = 236$), the judge achieved a raw 5-way accuracy of 58.90% (139/236; 95% CI: [52.53%, 64.98%]).Collapsing the orthographic categories raised classification accuracy to 81.36% (192/236; 95% CI: [75.90%, 85.81%], margin of error $\pm$4.9%; see Appendix F).

### 6.3 Quantifying WER Inflation

Table 2 presents our primary empirical findings across all three models on the held-out WAXAL test split. Standard WER over-reports error by 14.95% to 16.99% relative once phonological and code-switching variations are credited. Whisper-large-v3 shows the largest inflation (16.99% relative, 6.62 absolute points). Phonological variation is the dominant driver across all models, accounting for 5.87–6.47 absolute points, while code-switching accounts for a consistent residual of 0.11–0.15 points.

### 6.4 Cross-Corpus Lexicon Transferability

Table 3 evaluates whether the canonical lexicon transfers across datasets. The GhanaNLP lexicon delivers a 3.09%–3.48% relative reduction on WAXAL predictions despite being independently compiled from community text. Because GhanaNLP captures spontaneous speech, it intersects WAXAL's distribution far more effectively than FID (0.67%–0.78% reduction), suggesting that captured variants represent recurring linguistic phenomena rather than corpus artifacts, though the substantial vocabulary size disparity between corpora (7,434 vs. 25 entries) partially confounds this comparison (visualized in Appendix E).

| Model | Lexicon | Std. WER (%) | va-WER (%) | Overall Inflation | | Proportion of Inflation | |
|---|---|---|---|---|---|---|---|
| | | | | Abs (Δ) | Rel (%) | Phono. | Code-Sw. |
| Gemma-3n-2B | WAXAL | 40.00 | 34.02 | 5.98 | 14.95% | 5.87 | 0.11 |
| | GhanaNLP | 40.00 | 38.61 | 1.39 | 3.48% | 1.39 | 0.00 |
| | FID | 40.00 | 39.72 | 0.28 | 0.70% | 0.28 | 0.00 |
| Whisper-large-v3 | WAXAL | 38.97 | 32.35 | 6.62 | 16.99% | 6.47 | 0.15 |
| | GhanaNLP | 38.97 | 37.63 | 1.34 | 3.44% | 1.34 | 0.00 |
| | FID | 38.97 | 38.71 | 0.26 | 0.67% | 0.26 | 0.00 |
| MMS-1b-all | WAXAL | 39.54 | 33.14 | 6.40 | 16.19% | 6.26 | 0.14 |
| | GhanaNLP | 39.54 | 38.32 | 1.22 | 3.09% | 1.22 | 0.00 |
| | FID | 39.54 | 39.23 | 0.31 | 0.78% | 0.31 | 0.00 |

Table 3: Cross-corpus lexicon transferability on the WAXAL test set. Reduction scales with source corpus spontaneity and register diversity.

### 6.5 Human Spelling Inconsistency Ablation

Because spelling errors represent transcription defects rather than valid linguistic realisations, they are excluded from va-WER. However, isolating spelling inconsistency reveals that human transcription noise accounts for substantial uncredited performance (Table 2). Including spelling normalization reduces Whisper's reported WER to 29.07% (a 25.40% relative reduction). Decomposing Whisper's total 9.90 absolute point recovery, isolated category evaluations account for 6.47 points (65.4%) from phonological variants, 2.73 points (27.6%) from spelling errors, and 0.15 points (1.5%) from code-switching. The remaining 0.55 points (5.6%) arise from non-linear interaction synergies during joint Levenshtein dynamic programming alignment, where concurrent normalizations unlock shorter global edit paths.

## 7 Impact in African Contexts

Spoken language interfaces represent a vital gateway to civic and financial services across Africa, particularly where oral communication outpaces formal written literacy. However, standard exact-match WER penalizes ASR models for valid phonological shifts and nativized loanwords, inflating error rates by 15.0%–17.0% relative. In resource-constrained ecosystems, reporting an unadjusted 40% WER rather than an operational 32% risks false negatives during model evaluation, causing viable speech applications to be prematurely abandoned. By valuing dialect-inclusive transcriptions, variant-aware evaluation aligns benchmarking with living oral usage rather than rigid orthographic standards. Furthermore, the compiled canonical lexicons provide shared digital infrastructure to normalize noisy speech corpora and support broader cross-dialectal NLP across morphologically rich, multidialectal African languages (extended discussion in Appendix I).

## 8 Ethics Statement

Automated scoring of dialectal variants carries inherent risks of flattening linguistic diversity into canonical approximations. The compiled lexicons should be used exclusively for acoustic benchmarking and corpus normalization, not as instruments of cultural or orthographic standardization. All human annotation was performed by native Akan-speaking members of the research team.

## 9 Limitations

Our framework has four primary limitations. First, candidate classification relies on a proprietary LLM (Gemini 3.5 Flash), which may inherit biases or imperfect cultural-linguistic grounding in Akan. Second, collapsing dialectal variants into canonical forms erases stylistic and regional distinctions, making it unsuitable for accent or dialect identification tasks. Third, restricting candidate generation to an edit distance of $d = 1$ captures high-frequency phonetic and orthographic alternations efficiently, but omits multi-character morphological shifts. Finally, compiling canonical lexicons strictly from training and validation splits enforces a closed-vocabulary regime: novel dialectal variants unobserved during development bypass normalization and continue to incur exact-match penalties at test time.

## 10 Conclusion

We demonstrated that standard WER systematically misrepresents ASR performance in dialectally diverse languages like Akan by penalizing valid orthographic and phonological variations. Our variant-aware framework reveals that standard WER is inflated by up to 16.99% relative (6.62 absolute points) from legitimate phonological and code-switching variations alone, rising to 25.40% when transcriber spelling inconsistencies are included. Standardized read speech exhibits minimal inflation ($< 0.8\%$). We recommend reporting VA-WER alongside standard WER in orthographically variable languages: standard WER preserves historical comparability, while VA-WER isolates true acoustic model quality. The canonical lexicons generated by our pipeline can further serve to normalize noisy training transcriptions at their source. Future work should explore phone-level representations to evaluate recognition quality independently of writing system noise.

# Bibliography


[1] Bhogale, K.S., Javed, T., John, G.S., Rathi, D., Padmanaban, A., Parasa, N., Khapra, M.M.: Towards orthographically-informed evaluation of speech recognition systems for indian languages. arXiv preprint arXiv:2603.00941 (2026). https://doi.org/10.48550/arXiv.2603.00941, https://arxiv.org/abs/2603.00941

[2] Diack, A., Nelson, P., Agbesi, K., Nakalembe, A., MohamedKhair, M., Dube, V., Siyavora, T., Venugopalan, S., Hickey, J., Okonkwo, U., et al.: Waxal: A large-scale multilingual african language speech corpus. arXiv preprint arXiv:2602.02734 (2026)

[3] Dolphyne, F.: The Akan (twi-fante) Language: Its Sound Systems and Tonal Structure. Ghana Universities Press (1988), https://books.google.com.gh/books?id=K44OAAAAYAAJ

[4] GhanaNLP: Twi words speech-text parallel dataset. https://huggingface.co/datasets/michsethowusu/twi-words-speech-text-parallel (2025)

[5] Imam, S.H., Belay, T.D., Husse, K.Y., Ahmad, I.S., Abdulmumin, I., Umar, H.A., Bello, M.Y., Nakatumba-Nabende, J., Yimam, S.M., Muhammad, S.H.: Automatic speech recognition (asr) for african low-resource languages: A systematic literature review (2025), https://arxiv.org/abs/2510.01145

[6] James, J., Gopinath, D.P., et al.: Advocating character error rate for multilingual asr evaluation. arXiv preprint arXiv:2410.07400 (2024)

[7] Kim, S., Arora, A., Le, D., feng Yeh, C., Fuegen, C., Kalinli, O., Seltzer, M.L.: Semantic distance: A new metric for asr performance analysis towards spoken language understanding. In: Interspeech (2021), https://api.semanticscholar.org/CorpusID:233033738

[8] Manohar, K., Bhattacharya, A., Juvekar, K., Nethil, K.: Scribe: Diagnostic evaluation and rich transcription models for indic asr (2026), https://arxiv.org/abs/2605.20712

[9] Mensah, M.A., Wiafe, I., Ekpezu, A., Appati, J.K., Abdulai, J.D., Wiafe-Akenten, A.N., Yeboah, F.E., Odame, G.: Benchmarking akan asr models across domain-specific datasets: A comparative evaluation of performance, scalability, and adaptability. In: Proceedings of the Future Technologies Conference. pp. 210–225. Springer (2025)

[10] Owusu, D.A., Korsah, A., Sampah, D., Quartey, B., Nwolley Jnr, S., Adjepon-Yamoah, D.E., Boateng, L.O.: Financial inclusion speech dataset. Ashesi University and Nokwary Technologies, funded by Lacuna Fund (2022), https://github.com/Ashesi-Org/Financial-Inclusion-Speech-Dataset

[11] Pattnayak, P.: Sn-wer: Script-normalized wer for multi-script indic asr evaluation. In: Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026). pp. 249–256 (2026)

[12] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.N., Conneau, A., Auli, M.: Scaling speech technology to 1,000+ languages (2023), https://arxiv.org/abs/2305.13516

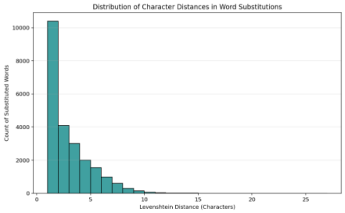
[13] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022), https://arxiv.org/abs/2212.04356

## A The Low-CER/High-WER Paradox and Near-Miss Concentration

Standard ASR evaluation suffers from a mathematical limitation when applied to dialectally diverse or orthographically fluid languages like Akan. As shown in Figure 2a, there is a severe distributional mismatch between Word Error Rate (WER) and Character Error Rate (CER). While CER remains skewed toward low error rates (10%–30%), standard WER is shifted toward 40%–50%. This demonstrates that the model produces grapheme sequences that are structurally and phonetically near-identical to references, but is heavily penalized by exact word-level matching.




(a) WER and CER Score Distributions




(b) Levenshtein Edit Distances

Fig. 2: Empirical evidence of metric divergence. (a) The Low-CER/High-WER paradox. (b) The near-miss syndrome, illustrating that word substitutions are concentrated at an exact edit distance of $d = 1$.

To isolate the root cause, Figure 2b details the Levenshtein character distance across all substituted word pairs. Over 10,000 mismatches occur at an exact edit distance of $d = 1$, while frequency drops exponentially for $d \geq 2$, confirming that errors are dominated by single-character phonological or typographical shifts.

## B Candidate Pair Taxonomy Distribution across Corpora

Figure 3 presents the complete distribution of candidate word pairs classified across our five-tier linguistic taxonomy for each corpus.

## C Canonical Lexicon Construction in Detail

We model the lexicon as a directed graph $G = (V, E)$ where nodes represent surface variant words and directed edges point toward canonical targets. For each candidate pair $(W_1, W_2)$, the word with higher corpus frequency is selected as canonical. The graph representation resolves three structural conflicts:

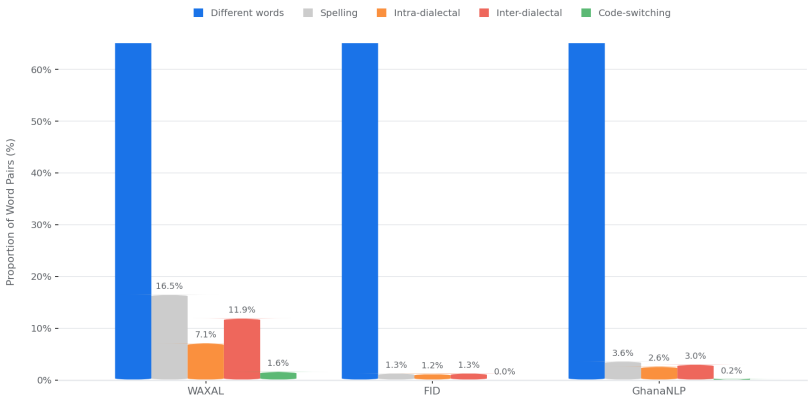



Fig. 3: Distribution of candidate word pairs across WAXAL, FID, and GhanaNLP datasets. Genuine semantic divergence ("Different words") constitutes the majority of raw near-miss pairs, while valid orthographic and phonological variants are heavily concentrated in spontaneous conversational speech.

- **Diverging Mappings:** When a variant maps to multiple distinct canonical targets (e.g., 'ebinom' mapping to 'binom' or 'ɛbinom'), conflicts are resolved by category prioritization (Phonological Variant > Code-Switching) and frequency.
- **Circular References:** When variants form closed loops ($A \to B \to C \to A$), cycles are broken by identifying the highest-frequency node across the cycle as the single canonical target.
- **Transitive Paths:** Multi-hop chains ($A \to B \to C$) are flattened via path compression so that all variants map directly to the root in $O(1)$ lookup time.

## D Surgical Traceback Discounting vs. Global Normalization

As an alternative to global string replacement, we explored an alignment-level surgical discounting algorithm (Figure 4). During Levenshtein alignment traceback between reference token $r$ and prediction token $p$ ($r \neq p$), we check whether $(r \to p) \in L$ or $(p \to r) \in L$. If registered in the canonical lexicon, the substitution cost is discounted to 0:

$$\text{cost}(r, p) = \begin{cases} 0 & \text{if } (r \to p) \in L \text{ or } (p \to r) \in L \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

Empirically, global string replacement and surgical discounting yield identical WER scores on our evaluation set, validating the algorithmic consistency of our framework.

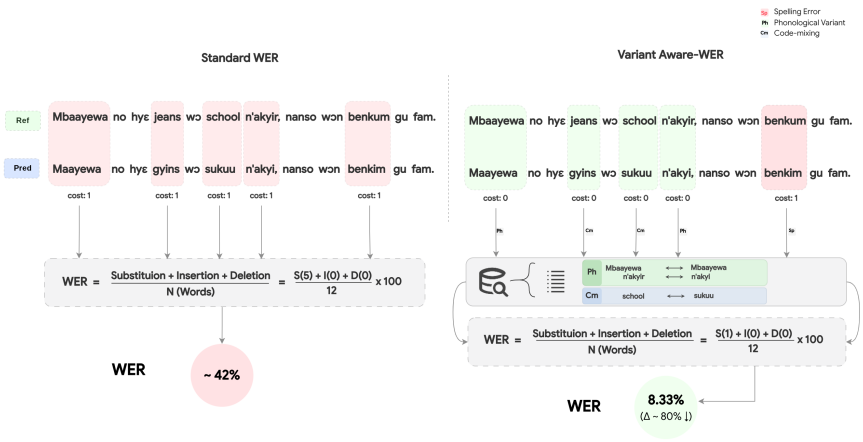



Fig. 4: Traceback comparison of standard exact-match WER (which penalizes valid phonetic approximations or dialectal shifts as substitution errors) versus our variant-aware surgical discounting evaluation framework (VA-WER) which discounts aligned variants to cost 0.

## E Cross-Corpus Transferability Chart

Figure 5 visualizes the cross-corpus relative WER reduction achieved when evaluating on WAXAL test predictions using canonical lexicons independently compiled from WAXAL, GhanaNLP, and FID corpora.

## F LLM Judge Performance vs. Human Raters

*Human Inter-Rater Reliability.* To measure baseline human agreement, three native Akan speakers independently evaluated the 20 overlapping calibration pairs prior to joint consensus panel adjudication. Across all five taxonomy categories, human raters achieved an overall Fleiss' Kappa of $\kappa = 0.2662$, indicating only fair agreement. While raters readily reached unanimous consensus on whether a token pair represented an acceptable linguistic variant versus a genuine semantic mismatch, distinguishing informal orthographic errors from dialectal phonological shifts proved highly subjective. Collapsing the three orthographic classes (Spelling Error, Inter-Dialect, and Intra-Dialect) increased Fleiss' Kappa to $\kappa = 0.5087$ (moderate agreement). Table 4 details per-category precision, recall, and F1 scores of the Gemini 3.5 Flash judge against each individual rater's unique sample partition ($N = 236$ total). Figure 6 illustrates the resulting confusion matrices pre- and post-category collapse.

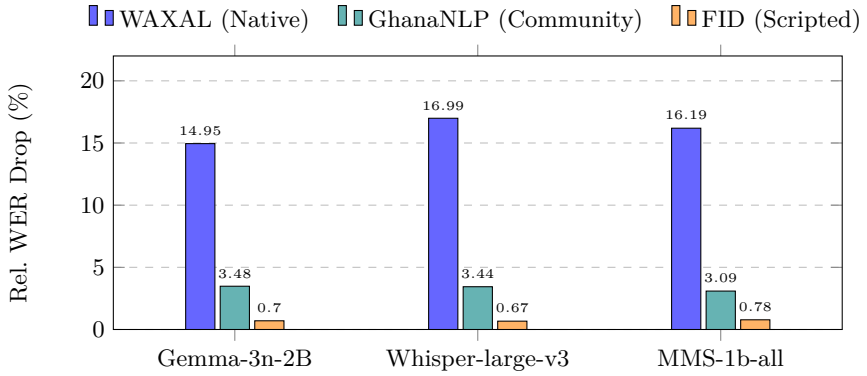



Fig. 5: Relative WER inflation reduction on the WAXAL test set across native, community, and scripted canonical lexicons.

| | Rater 1 (Unique Samples) | | | | Rater 2 (Unique Samples) | | | | Rater 3 (Unique Samples) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Error Category** | **P** | **R** | **F1** | *N* | **P** | **R** | **F1** | *N* | **P** | **R** | **F1** | *N* |
| Code-Switching | 0.69 | 1.00 | 0.82 | 9 | 0.55 | 0.86 | 0.67 | 7 | 0.65 | 1.00 | 0.79 | 15 |
| Different Words | 0.64 | 0.69 | 0.67 | 13 | 0.80 | 0.70 | 0.74 | 23 | 1.00 | 0.52 | 0.69 | 23 |
| Phonological Variant – Inter-Dialect | 0.47 | 0.60 | 0.53 | 15 | 0.71 | 0.67 | 0.69 | 15 | 0.27 | 0.57 | 0.36 | 7 |
| Phonological Variant – Intra-Dialect | 0.33 | 0.42 | 0.37 | 12 | 0.47 | 0.64 | 0.54 | 11 | 0.11 | 0.50 | 0.18 | 4 |
| Spelling Error | 0.62 | 0.36 | 0.45 | 28 | 0.85 | 0.71 | 0.77 | 24 | 0.73 | 0.27 | 0.39 | 30 |
| **Overall Accuracy** | | | **0.55** | | | | **0.70** | | | | **0.52** | |

Table 4: LLM Judge individual performance evaluation across isolated unique human rater sets. Metrics include Precision (**P**), Recall (**R**), F1-Score (**F1**), and sample count ($N$) for all raw uncollapsed categories.

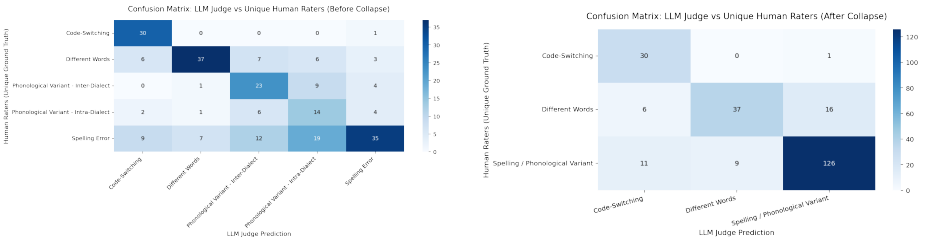
## G Metric Sensitivity Analysis under Classification Noise

A natural question regarding the LLM-as-a-Judge pipeline is whether classification errors between spelling and phonological variants distort VA-WER. While collapsing orthographic categories yields 81.43% accuracy, the uncollapsed 5-way accuracy is 58.90% (Figure 6, left). Because VA-WER retains phonological variants while strictly excluding spelling errors, misclassifications across this boundary could theoretically leak transcriber spelling noise into the lexicon or over-normalize distinct words.

To address this, we evaluate metric stability through both theoretical worst-case bounding and empirical Monte Carlo perturbation on the held-out test set.

*Empirical Boundary Error Rates.* From the human evaluation confusion matrix ($N = 236$, Figure 6), out of 96 word pairs classified by the judge as phonological variants:

– **Spelling Leakage ($\alpha_{\text{spelling}} = 32.29\%$):** 31 pairs were annotated by native raters as spelling errors.




(a) Uncollapsed 5-Way Matrix (b) Collapsed Orthographic Matrix

Fig. 6: Confusion matrices evaluating the LLM judge against unique human-annotated samples ($N = 236$).

- **Lexical Over-normalization ($\alpha_{\mathbf{dw}} = 13.54\%$):** 13 pairs were annotated as distinct lexical items.
- **Phonological Omission ($\beta_{\mathbf{phono}} = 12.50\%$):** Conversely, 8 out of 64 ground-truth phonological pairs were misclassified as spelling errors and omitted from the lexicon.

*Analytical Worst-Case Bounds.* Because misclassified spelling errors ($\alpha_{\text{spelling}} = 32.29\%$) and distinct lexical items ($\alpha_{\text{dw}} = 13.54\%$) reside directly within the candidate phonological lexicon, their maximum possible unearned reduction is bounded directly as a fraction of the phonological gap:

$$\Delta\text{va-WER}_{\text{leakage}} \leq (\alpha_{\text{spelling}} + \alpha_{\text{dw}}) \cdot \Delta_{\text{phono}} \tag{3}$$

For Whisper-large-v3 ($\Delta_{\text{phono}} = 6.47\%$), the upper bound on unearned reduction is $0.4583 \cdot 6.47\% = 2.97\%$ absolute WER. Even under this maximally pessimistic analytical bound—assuming every misclassified entry was active in test evaluation and ignoring omitted true phonological variants ($\beta_{\text{phono}} = 12.50\%$)—the true phonological and code-switching recovery remains at least $6.62\% - 2.97\% = 3.65\%$ absolute (9.38% relative drop; $p < 10^{-4}$). Across architectures, the guaranteed relative reduction remains between 8.23% and 9.38%, directly corroborating the empirical distributions from our Monte Carlo perturbations (Table 5).

*Empirical Monte Carlo Perturbation.* To complement the analytical bounds, we conducted an empirical Monte Carlo sensitivity simulation ($N = 50$ trials) on the held-out test predictions. In each trial, we randomly purged 45.83% of entries from the phonological lexicon (matching the combined empirical rate of human-annotated spelling and distinct-word errors) and recomputed VA-WER across all three models.

As shown in Table 5, the empirical variance is narrow ($\sigma \leq 0.47\%$). Even with nearly half the phonological lexicon randomly eliminated, the mean relative inflation reduction remains substantial (7.94%–8.92%), and the 95% confidence intervals strictly exclude zero ($p < 10^{-5}$). Furthermore, model rankings (Whisper > MMS > Gemma-3n) remained strictly invariant across all 50 iterations, confirming that the findings are robust to classifier boundary noise.

| Model | Std. WER | Reported va-WER | Reported Rel. Drop | Pessimistic Bound | Purged va-WER (Mean ± Std) | Purged Rel. Drop (95% CI) |
|---|---|---|---|---|---|---|
| Whisper-large-v3 | 38.97 | 32.35 | 16.99% | 9.38% | 35.50% ± 0.47% | [6.87%, 11.25%] |
| MMS-1b-all | 39.54 | 33.14 | 16.19% | 8.93% | 36.11% ± 0.43% | [6.59%, 10.56%] |
| Gemma-3n-2B | 40.00 | 34.02 | 14.95% | 8.23% | 36.83% ± 0.41% | [5.64%, 9.53%] |

Table 5: Sensitivity analysis under LLM classification noise. Columns show standard WER, reported va-WER, analytical worst-case bound, and the empirical Monte Carlo distribution under an aggressive 45.83% lexicon purge ($N = 50$).

## H LLM Judge Prompt Specification

**LLM Taxonomy Annotation Prompt Template**

**System Instructions:**
You are an expert ASR evaluator and native linguist specializing in the Akan macrolanguage (Asante Twi, Akuapem Twi, Fante).
Your task is to analyze a single-character difference between two words within the context of full Akan sentences. Classify the error into exactly one category:

1. **Phonological Variant:** Valid phonological alternation in spoken Akan (Intra-Dialect spoken rule or Inter-Dialect alternation).
2. **Spelling Error:** Violation of standardized Akan orthographic conventions without phonological justification (e.g., *bebree* vs. *beberee*).
3. **Code-Switching:** English loanwords phonetically adapted into Akan orthography (e.g., *kar* vs. *car*).
4. **Different Words:** Genuine semantic or grammatical errors altering word meaning (e.g., singular/plural shifts like *nipa* vs. *nnipa*).

**Chain-of-Thought Reasoning Protocol:**
For every word pair, you must think step-by-step in your `"explanation"` field:

- **Step 1:** Identify the exact character-level difference.
- **Step 2:** Evaluate whether the variation matches known phonological rules, dialectal shifts, or borrowing.
- **Step 3:** Determine if the grammatical syntax or semantic meaning of the sentence has changed.
- **Step 4:** Conclude with the categorical classification.

**Output Contract:** Return strictly a JSON dictionary with keys `"category"` and `"explanation"`.

## I Extended Discussion: Broader Impact in African NLP Ecosystems

*Aligning Models with Spoken Realities in Civic Services.* In civic deployments—such as rural healthcare triage or agricultural advisory hotlines—callers rarely

conform to institutional orthographic standards. A Central Region farmer speaking Fante will produce *edwuma* (work), while an Ashanti Region clinician expects *adwuma*. Standard exact-match metrics penalize models that recognize authentic spoken dialects whenever transcribers have recorded alternative regional standards. By incorporating variant-aware evaluation, voice interfaces can be assessed based on how well they respect living linguistic diversity rather than enforcing rigid, school-taught orthographic conformity.

*Canonical Lexicons as Digital Public Infrastructure.* Linguistic resource creation in Africa is frequently siloed, with teams duplicating effort to reconcile orthographic discrepancies. The canonical variant lexicons compiled by our pipeline serve as reusable digital public goods. Beyond speech evaluation, these mappings provide shared linguistic infrastructure directly applicable to: (1) cleaning and normalizing noisy speech corpora before pre-training; (2) improving cross-dialectal machine translation alignment; and (3) informing national language planning bodies, such as the Bureau of Ghana Languages, regarding the evolving boundaries between formal orthography and living oral usage.

*A Generalizable Blueprint for Multidialectal African Languages.* Although demonstrated on Akan, our evaluation framework addresses structural characteristics common across the Niger-Congo language phylum. Languages such as Yoruba, Igbo, Wolof, and Luganda exhibit wide dialect continuums, active vowel harmony systems, and rapid loanword assimilation, yet lack extensive standardized lexicons. Our prompt-guided candidate pair extraction and graph-based resolution pipeline provides a low-cost, reproducible methodology for researchers across the continent to build variant-aware evaluation layers without requiring years of manual lexicographical fieldwork.