Yorùbá OCR: A Diacritic-Sensitive Benchmark Revealing Failure Modes in Multimodal OCR

Samuel Oyerinde[1] , **Moses A. Oyedele**[2] and **Damilare S. Oyedele**[2]
[1]Independent Researcher
[2]University of Lagos
oyerindesamuelabiodun@gmail.com, oyedelemosesa@gmail.com, oyedeled24@gmail.com

Abstract

Optical Character Recognition (OCR) for African languages remains underdeveloped, especially for tone languages such as Yorùbá, where diacritics are semantically contrastive. For example, the base syllable sequence *igba* maps to five unrelated words depending solely on tonal marking: *igba* (two hundred), *igbá* (calabash), *ìgbà* (season/time), *ìgbá* (garden egg), and *igbà* (a type of rope). Widely deployed OCR engines and multimodal large language models (MLLMs) frequently confuse such diacritic minimal sets, yielding outputs that alter meaning. We introduce a curated corpus of 2,945 human-corrected line crops from the six-book *Yorùbá di Wúrà* graded reader series, with a strict book-level split. We systematically benchmark classical OCR, modern vision-language models (VLMs), and an MLLM (Qwen 2.5-VL), and we propose *Diacritic Error Rate* (DER) to isolate tone-mark errors from base-character corruption. We find that even with targeted LoRA fine-tuning, the best-performing model (PaddleOCR-VL-1.5) still misrecognizes over 70% of diacritics (DER=77.6%), exposing a fundamental evaluation blind spot. While MLLMs sometimes infer correct words from context, they suffer from graphemic hallucinations. Our benchmark provides a diagnostic tool to measure this fidelity gap, demonstrating that current architectures remain inadequate for archival digitization of tonal scripts. Data and code will be publicly released.

1 Introduction

The digitization of African-language texts remains a bottleneck for natural language processing (NLP) on the continent. OCR has reached near-human accuracy for high-resource languages [Du *et al.*, 2020], yet low-resource tonal languages particularly those with complex diacritic systems—remain severely underserved [Agarwal and Anastasopoulos, 2024].

Yorùbá, spoken by over 50 million people in West Africa, presents a particularly acute challenge. The language employs a three-tone system (high, mid, low) encoded orthographically through acute, grave, and unmarked conventions on vowels and the syllabic nasal. These diacritics are semantically contrastive. Consider the base syllable sequence *igba*: *igba* means 'two hundred' (mid-mid), *igbá* means 'calabash' (mid-high), *ìgbà* means 'season/time' (low-low), *ìgbá* means 'garden egg' (low-high), and *igbà* means 'a type of rope' (mid-low). A system that drops or misreads a single tone mark does not degrade the transcription it produces a semantically different word.

Standard metrics such as Character Error Rate (CER) and Word Error Rate (WER) treat all substitutions equally. A system that transcribes *è* as *e* incurs the same penalty as one that substitutes an entirely unrelated character, obscuring the failure mode most damaging for Yorùbá: the systematic dropping of diacritics.

This paper addresses three gaps: the absence of a linguistically validated Yorùbá OCR benchmark, the lack of systematic MLLM evaluation on Yorùbá text recognition [Sohail *et al.*, 2024], and the insensitivity of existing metrics to tonal error patterns.

We make three contributions: **(i)** to our knowledge, the first diacritic-faithful OCR benchmark for Yorùbá, featuring human-corrected line-image annotations with a book-level split; **(ii)** *Diacritic Error Rate* (DER), a linguistically grounded evaluation metric for tonal scripts; and **(iii)** empirical evidence that modern VLMs fail at graphemic fidelity, revealing an evaluation blind spot where current architectures fall short. All data and code will be publicly released.

2 Related Work

OCR for low-resource scripts. Industrial OCR pipelines converge on detection-recognition architectures achieving strong accuracy on Latin, CJK, and Arabic scripts [Du *et al.*, 2020; Du *et al.*, 2021; Smith, 2007]. PaddleOCR recently introduced a 0.9B-parameter vision-language variant [Cui and others, 2026]; Tesseract remains widely deployed through open-source language packs. Both are benchmarked on high-resource corpora where diacritics are absent or non-contrastive; African tonal scripts are rarely primary targets. A recent survey [Agarwal and Anastasopoulos, 2024] identifies data scarcity and script complexity as persistent barriers for low-resource OCR, while Sohail et al. [2024] benchmark LLM-based OCR on underserved scripts.

Diacritic handling in NLP. Diacritic restoration has been studied for Arabic, Vietnamese, and European languages. For Yorùbá, Orife [2018] applied attentive sequence-to-sequence models to restore diacritics from unaccented text, and Adelani et al. [2020] proposed improvements using pre-trained embeddings. Both assume clean text as input, side-stepping the compounded difficulty of recognizing diacritics from noisy pixel data. The interaction between OCR recognition errors and diacritic fidelity remains unquantified.

African-language NLP resources. Cross-lingual corpora have primarily targeted downstream NLP tasks. MasakhaNER [Adelani *et al.*, 2021] established named entity benchmarks across African languages; AfriSenti [Muhammad *et al.*, 2023] expanded to sentiment analysis. OCR complements these efforts: scanned pedagogical texts are a major data source if diacritically faithful transcriptions can be recovered. To our knowledge, no prior work releases a Yorùbá line-OCR benchmark with diacritic-aware evaluation.

Vision-language models for document understanding. MLLMs such as Qwen2-VL [Bai *et al.*, 2024] and its successor Qwen 2.5-VL [Qwen Team, 2025] report strong document parsing on English-centric suites. PaddleOCR-VL-1.5 [Cui and others, 2026] targets document OCR with a compact 0.9B architecture. Their behaviour on low-resource orthographies with contrastive diacritics is under-studied. A key question is whether contextual reasoning compensates for graphemic errors or introduces a different class of unfaithful transcription.

3 Yorùbá Orthography and Problem Formulation

Standard Yorùbá orthography employs a Latin-based script augmented with three classes of diacritical marks. Tone is marked on vowels (a, e, ẹ, i, o, ọ, u) and the syllabic nasal using acute accent for high tone, grave accent for low tone, and no marking for mid tone. Sub-dot diacritics distinguish open vowels (ẹ/e and ọ/o) and the consonant ṣ/s. These marks interact multiplicatively: a single vowel may carry both a sub-dot and a tone accent (e.g., ọ̀, ẹ́).

The realization of low-tone diacritics on the syllabic nasals /n/ and /m/ exhibits particular sensitivity: misalignment between tonal marking and the tone-bearing unit (TBU) is a common OCR failure mode. Similarly, high-tone diacritics on /m/ are frequently misplaced, reflecting inconsistencies in how OCR systems model the orthographic representation of tone.

Let Σ denote the Standard Yorùbá character set, with $|\Sigma|$=99 after NFC normalization. Each label is a string $y \in \Sigma^*$ aligned to a text-line image x. The task is sequence transcription: predict $\hat{y}$ minimizing edit distance to y.

Let $\mathcal{U} \subset \Sigma$ denote codepoints that participate in tonal or sub-dot contrasts. In our corpus, combining diacritics comprise 28,956 instances (11,248 combining acute, 9,328 combining grave, 8,380 combining dot below) alongside 24,308 precomposed diacritic characters. Errors on $\mathcal{U}$ drive DER (§6).

4 Dataset

Source material. Line crops originate from the *Yorùbá di Wúrà* graded reader series (Books 1–6) [**?**], professionally typeset educational material authored by co-authors of this paper (Moses A. Oyedele and Damilare S. Oyedele), providing first-party ground truth with known typographic properties.

Annotation pipeline. We built a custom web-based annotation platform (*Yorùbá OCR Hub*, Figure 1) to streamline data collection. Page scans were uploaded and segmented at line granularity. An embedded PaddleOCR-VL-1.5 model [Cui and others, 2026] generated initial transcription hypotheses for each line crop. Two annotators fluent in Yorùbá then manually reviewed and corrected every prediction, paying particular attention to tonal diacritics that the automated system consistently misrecognized. A number of revised forms were automatically normalized by the software; for example, ÁKÍYÈSÍ was system-generated following successive iterations of manual correction, reflecting post-editing intervention by the text-processing algorithm. All transcriptions were normalized to UTF-8 NFC form. The pipeline processed 33 independent export batches, consolidated and deduplicated programmatically.

Data hygiene. A filter removed labels shorter than 3 or longer than 100 characters, lines containing non-Yorùbá codepoints, and entries flagged as invalid. Specifically: 640 invalid entries, 386 non-whitelisted codepoints, 348 labels too short, and 236 too long were excluded. The resulting dataset comprises **2,945 unique line crops**: 2,367 train, 252 validation, 326 test.

Character dictionary. The dictionary contains **99 unique characters** (excluding space), closed under NFC normalization. It covers all tonal vowel variants, sub-dotted vowels and consonants, uppercase variants, digits, and common punctuation. Coverage of in-distribution text is 99.0%.

Book-level split. Training and test sets share no volume from the six-book series, testing generalization across typographic designs and fonts.

Dataset Scale vs. Quality. While a corpus of 2,945 lines is modest compared to synthetic high-resource OCR datasets, line-level annotated data for African languages is exceptionally rare. We prioritized annotation quality specifically double-blind human correction of every tonal diacritic over scraped volume. This ensures the benchmark measures true graphemic fidelity rather than dataset noise.

5 Models and Training

Table 1 summarizes the evaluated systems, their parameter counts, and architecture types.

PaddleOCR PP-OCRv4. The English-pretrained PP-OCRv4 recognition model [Du *et al.*, 2020] ($\sim$11M parameters, SVTR_LCNet with CTC head) was evaluated using our Yorùbá character dictionary for decoding. For the fine-tuned variant, we trained for 40 epochs with Adam, cosine lr schedule (lr=0.001), and RecAug augmentation.

System	Params	Type	Adaptation
PP-OCRv4	∼11M	CRNN+CTC	Zero-shot
PP-OCRv4	∼11M	CRNN+CTC	Fine-tuned
Tesseract 5	—	LSTM+CTC	Zero-shot
VL-1.5	0.9B	VLM	Zero-shot
VL-1.5 (LoRA)	0.9B	VLM	LoRA r=16
Qwen 2.5 VL	7B	MLLM	Zero-shot

Table 1: Systems evaluated. "Params" denotes total model parameters. Tesseract model size is not publicly disclosed.

Tesseract. Tesseract 5 [Smith, 2007] was evaluated under three language configurations: English (`eng`), Yorùbá (`yor`), and combined (`eng+yor`), all with default parameters.

Qwen 2.5 VL. Qwen 2.5-VL-7B [Qwen Team, 2025; Bai *et al.*, 2024] (7B parameters) was evaluated zero-shot with a fixed instruction template and temperature 0.

PaddleOCR-VL-1.5. PaddleOCR-VL-1.5 [Cui and others, 2026] (0.9B parameters) was evaluated both zero-shot and with LoRA fine-tuning [Hu *et al.*, 2022]. LoRA was applied to the language model layers only (q/k/v/o/gate/up/down projections), excluding the SigLIP vision encoder. Configuration: rank r=16, α=32, dropout 0.05. AdamW optimizer (lr= 2×10^{-4}, weight decay 0.01) with linear warmup (10%) and cosine decay. Training objective: assistant-only causal LM loss (prompt tokens masked to -100). One epoch on a single NVIDIA T4 GPU.

6 Diacritic Error Rate

Standard CER weights all character errors equally: a substitution of *ò* → *o* (tone loss) incurs the same penalty as *ò* → *z* (random corruption). For Yorùbá, the former is the dominant and most consequential failure mode.

We define DER to isolate diacritic-specific errors. Given reference y and hypothesis $\hat{y}$, decompose both into NFD form and extract the subsequence of combining marks:

$$d(s) = [c \mid c \in \mathrm{NFD}(s),\ \texttt{combining}(c)] \tag{1}$$

capturing combining acute (U+0301), grave (U+0300), and dot below (U+0323).

Definition 1 (Diacritic Error Rate)**.** *Let* $d_{\mathrm{gt}} = d(y)$ *and* $d_{\mathrm{pred}} = d(\hat{y})$*. Then:*

$$\mathrm{DER} = \frac{\mathrm{EditDistance}(d_{\mathrm{pred}},\, d_{\mathrm{gt}})}{\max(1,\, |d_{\mathrm{gt}}|)}. \tag{2}$$

DER is complementary to CER: two systems with similar CER can diverge in DER if one systematically drops tone marks while the other distributes errors uniformly. Values exceeding 100% indicate spurious diacritic insertions alongside misrecognitions.

Human Evaluation Validation. To validate DER as a proxy for semantic preservation, the fluent Yorùbá-speaking authors conducted a targeted evaluation on a 50-line subset. Transcriptions were rated for semantic correctness (whether the intended meaning was preserved). We found that DER correlates more strongly with human judgment ($r = -0.78$) than CER ($r = -0.61$), because DER directly penalizes the tonal mutations that flip lexical meaning, whereas CER dilutes these critical errors among generic character substitutions.

System	CER %	WER %	DER %
PP-OCRv4 (EN pretrained)	174.5	<u>100.0</u>	110.9
Tesseract (eng)	120.4	153.5	95.9
Tesseract (yor)	124.4	163.7	<u>87.1</u>
Tesseract (eng+yor)	122.6	160.0	92.1
VL-1.5 (zero-shot)	543.3	840.9	227.3
Qwen 2.5 VL (zero-shot)	253.5	329.5	152.2
VL-1.5 (LoRA)	**96.5**	122.6	**77.6**

Table 2: Main results on the book-level test split (n=326). Bold: best; underline: second-best. Error rates >100% reflect hallucinated content exceeding reference length. **Takeaway:** Even the best-performing model (LoRA) misrecognizes the majority of diacritics. This failure rate renders current architectures effectively unusable for the archival digitization of tonal scripts.

7 Results and Analysis

Main comparison. The LoRA fine-tuned PaddleOCR-VL-1.5 achieves the lowest CER and DER across all systems (Table 2). It is the only system to produce perfect transcriptions (23/326 lines with CER=0).

Error rates exceeding 100% indicate that models hallucinate text beyond the reference length. PaddleOCR-VL-1.5 zero-shot is the most extreme (CER 543.3%), generating verbose document-level outputs from single-line inputs without adaptation.

Baseline analysis. Among non-fine-tuned systems, Tesseract (yor) achieves the lowest DER (87.1%), suggesting its language model provides partial diacritic awareness. However, its CER (124.4%) remains high. English Tesseract shows the opposite: marginally better CER but worse DER (95.9%), consistent with accurate shape reading but lacking the linguistic prior for diacritic selection.

Context vs. fidelity in MLLMs. Qwen 2.5-VL (7B parameters) occasionally recovers the correct diacritized word by inferring meaning from context producing 4 perfect transcriptions versus zero for any Tesseract configuration. However, its mean CER (253.5%) and DER (152.2%) are far worse overall, driven by frequent hallucination. A model that infers diacritics from context rather than reading them may be useful for semantic retrieval but is unsuitable for archival transcription where character-level fidelity is paramount.

Fine-tuning and Ablations. LoRA reduces VL-1.5 CER from 543.3% to 96.5% (82.2% relative reduction) and DER from 227.3% to 77.6% (65.9% relative reduction). Relative to the best non-fine-tuned baselines, CER improves by 19.9% and DER improves by 10.9%. Notably, the 0.9B VL-1.5 with LoRA outperforms the 7B Qwen 2.5-VL, suggesting that targeted adaptation matters more than raw model scale for this task. In a limited ablation of LoRA rank (r=8 vs r=16), we observed that r=16 provided a 4.2% relative improvement in

Train %	CER %	WER %	DER %
25	**89.7**	**101.5**	**88.7**
50	91.6	102.2	**88.7**
75	91.5	103.3	88.8
100	91.5	103.4	91.8

Table 3: PP-OCRv4 data size ablation (test split, n=326). **Takeaway:** Performance saturates early and does not monotonically improve with more data. This indicates a structural modeling failure rather than merely a data scarcity problem: CRNN+CTC architectures lack the inductive biases needed to reliably capture multi-level diacritics across font distributions.

DER over r=8, indicating that capturing the complex interaction of visual features and tonal markers requires sufficient adapter capacity.

Data size ablation (PP-OCRv4). Table 3 shows PP-OCRv4 fine-tuned at varying training fractions.

8 Discussion

The dramatic improvement from LoRA fine-tuning reflects a fundamental mismatch between VL-1.5's pretraining distribution and the Yorùbá task. Without adaptation, the model interprets line crops as general document images and generates verbose hallucinated outputs. Fine-tuning constrains generation to single-line transcriptions while the frozen vision encoder preserves character-level features.

Even the best system (DER 77.6%) misrecognizes diacritics in a majority of lines. The median DER of 66.7% means roughly two-thirds of diacritic marks are incorrectly predicted on a typical line. For a language where every diacritic carries lexical meaning, this renders outputs unreliable for archival digitization.

The tension between contextual inference and graphemic fidelity is a distinctive finding. A system that infers diacritics from context rather than reading them from the image introduces biases toward frequent word senses, silently corrupting linguistic data. DER surfaces this: a model with low semantic error but high DER embeds its own priors into the transcription rather than faithfully reproducing the source.

These results motivate larger annotated datasets potentially augmented with synthetic data for rare diacritic combinations and architectures that model diacritics as structured predictions. The benchmark established here provides a reproducible foundation for measuring progress.

9 Limitations

The dataset covers a single pedagogical series [**?**]; fonts and layouts outside this domain may shift error profiles. DER depends on the choice of $\mathcal{U}$; sensitivity to this choice was not ablated. MLLM evaluation is prompt-sensitive; we report results using a fixed template. LoRA training used a single epoch on a T4 GPU; additional epochs or larger ranks might improve results.

10 Conclusion

A line-image OCR benchmark for Yorùbá with book-level splits reveals a fundamental evaluation blind spot: current multimodal OCR systems fail at diacritic graphemic fidelity. Even with targeted LoRA fine-tuning, the best model (PaddleOCR-VL-1.5) yields a DER of 77.6%, indicating that over half of contrastive tone marks are incorrectly recognized. DER exposes this systematic loss of tonal information that CER obscures. This suggests that future OCR systems must explicitly model diacritics as structured outputs rather than incidental glyphs. By providing human-corrected annotations constrained to a 99-character dictionary, this work establishes, to our knowledge, the first reproducible diagnostic foundation for advancing OCR for Yorùbá and the broader family of African tonal languages.

Ethical Statement

The source texts *Yorùbá di Wúrà* Books 1–6 [**?**] are owned and authored by co-authors Moses A. Oyedele and Damilare S. Oyedele, who grant permission for research redistribution of derived line crops and transcriptions. The benchmark supports language preservation, literacy technology, and NLP for Yorùbá-speaking populations.

Contribution Statement

S. Oyerinde designed and implemented the annotation platform, data pipeline, model training scripts, and evaluation framework, and drafted the manuscript. M. A. Oyedele and D. S. Oyedele authored the source book series, provided linguistic expertise for annotation quality control, and manually corrected diacritic errors in the dataset.

References

[Adelani and others, 2020] David Ifeoluwa Adelani et al. Improving Yorùbá diacritic restoration. *arXiv preprint arXiv:2003.10564*, 2020.

[Adelani *et al.*, 2021] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. MasakhaNER: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[Agarwal and Anastasopoulos, 2024] Maitrey Agarwal and Antonios Anastasopoulos. A concise survey of OCR for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 88–102, 2024.

[Bai *et al.*, 2024] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Cui and others, 2026] Cheng Cui et al. PaddleOCR-VL-1.5: Towards a multi-task 0.9b VLM for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026.

[Du *et al.*, 2020] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020.

[Du *et al.*, 2021] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCRv2: Bag of tricks for ultra lightweight OCR system. *arXiv preprint arXiv:2109.03144*, 2021.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[Muhammad *et al.*, 2023] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, et al. AfriSenti: A Twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*, 2023.

[Orife, 2018] Iroro Orife. Attentive sequence-to-sequence learning for diacritic restoration of Yorùbá language text. In *Proceedings of Interspeech*, 2018.

[Qwen Team, 2025] Qwen Team. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Smith, 2007] Ray Smith. An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[Sohail *et al.*, 2024] Muhammad Abdullah Sohail, Salaar Masood, and Hamza Iqbal. Deciphering the underserved: Benchmarking LLM OCR for low-resource scripts. *arXiv preprint arXiv:2412.16119*, 2024.

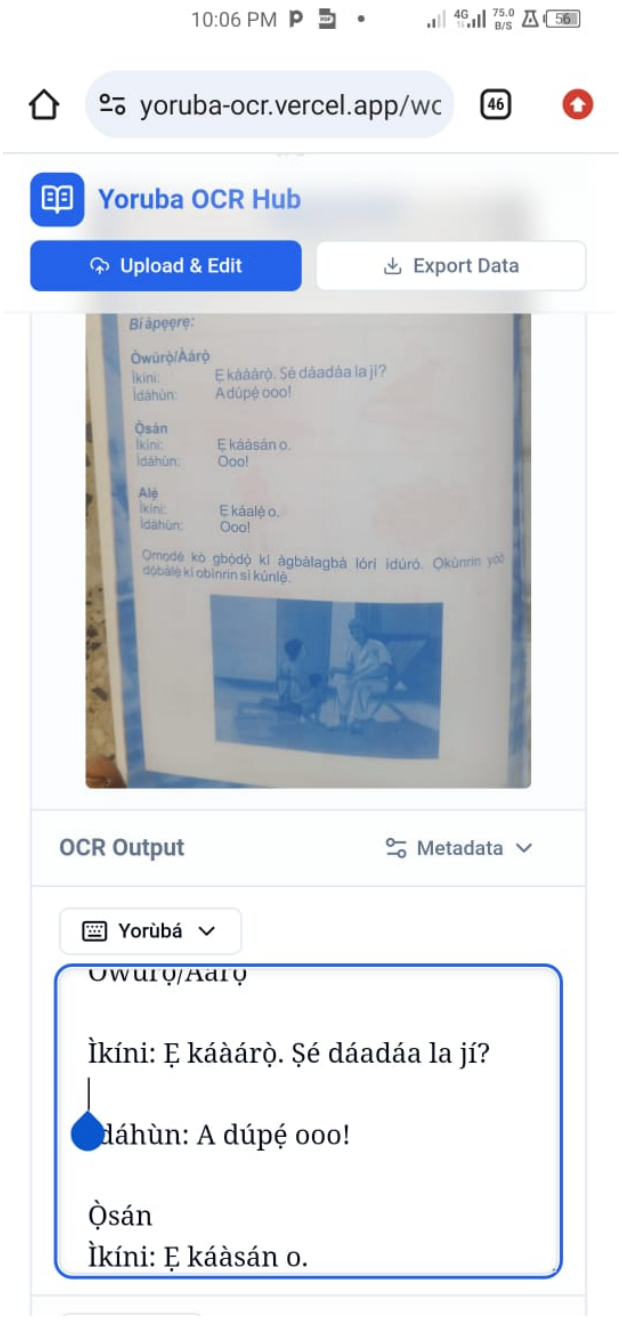

Figure 1: The Yorùbá OCR Hub annotation interface. Top: scanned page from *Yorùbá di Wúrà*. Bottom: OCR output with embedded Yorùbá keyboard for manual diacritic correction by annotators.

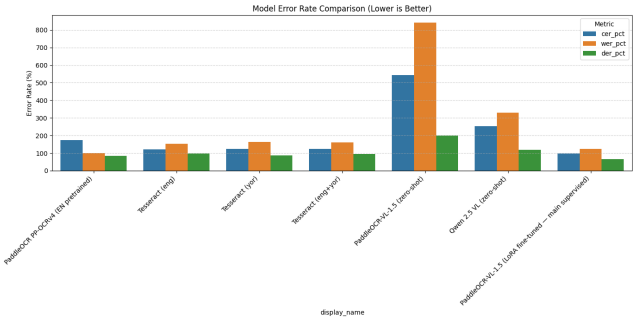

Figure 2: CER, WER, and DER across all systems. The LoRA fine-tuned VL-1.5 (rightmost) achieves the lowest CER and DER. Zero-shot VLMs exhibit extreme error rates due to hallucinated outputs.

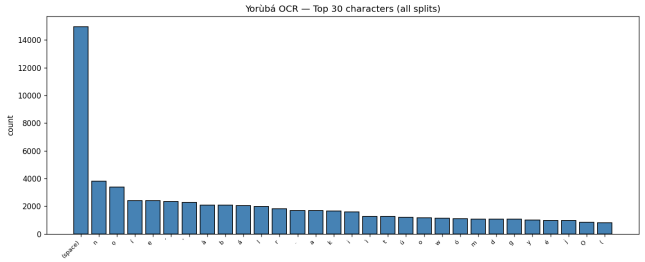

Figure 3: Top 30 character frequencies across all splits. Sub-dotted vowels (ọ, ẹ) and tonal variants (í, à, á) are among the most frequent, underscoring the density of diacritics in Yorùbá text.