Reclaiming African Voices: Surveying Indigenous Writing Systems for Inclusive NLP

Mamady Traore, Ngoc Tan Le, Fatiha Sadat
Université du Québec à Montréal (UQAM)
Montréal, QC, Canada
traore.mamady@courrier.uqam.ca, le.ngoc_tan@uqam.ca, sadat.fatiha@uqam.ca

Abstract

Multilingual NLP has expanded rapidly through large-scale pretraining and cross-lingual transfer, yet this progress remains structurally uneven across writing systems. This survey reframes multilingual NLP around scripts rather than languages, arguing that writing systems constitute a critical and under-theorized axis of computational inequality. Focusing on African scripts—Indigenous (Vai, Ge'ez, Tifinagh), modern (ADLaM, N'Ko), and adapted Arabic-based (Ajami)—we analyze how script properties interact with digital infrastructure, tokenization, and downstream task performance. We organize the literature across four analytical layers: infrastructural (Unicode and input systems), representational (segmentation efficiency and vocabulary allocation), functional (task-level disparities), and epistemic (evaluation bias and the "low-resource" framing). Synthesizing evidence from 47 studies, we show that performance gaps across scripts arise primarily from engineering design choices rather than intrinsic linguistic complexity. We conclude by outlining a research agenda for native multiscript foundation models, including script-aware scaling laws, tokenizer equity metrics, and evaluation reform. We argue that multiscript equity is not a peripheral concern but a structural precondition for genuine multilingual inclusion.

Keywords: African Scripts, Writing Systems, NLP Decolonization, Tokenization Bias, Multilingual Modeling

1. Introduction

Natural Language Processing (NLP) has made substantial multilingual advances through large-scale pretraining and cross-lingual transfer. Yet these gains are uneven across writing systems. Model performance varies systematically with script representation, formatting conventions, and tokenization behavior, indicating that multilingual coverage does not automatically translate into equity at the script level (Reddy et al., 2026; Kanjirangat et al., 2025; Asprovska and Hunter, 2024).

Current NLP infrastructures remain implicitly Latin-centric. Prior work situates this imbalance within broader historical and structural dynamics. Yan and Xu (2024) argue that African NLP development reflects colonial-era language hierarchies and technological dependency, framing the challenge as infrastructural rather than purely technical. Adebara (2024) further contends that dominant evaluation paradigms reproduce Western standards, while Ògúnrmí et al. (2023) contend that the "low-resource" label itself reflects structural marginalization rather than a purely technical limitation.

Empirical work confirms that script properties directly influence model behavior. Reddy et al. (2026) show that Large Language Model (LLM) arithmetic accuracy declines substantially when numerals are presented in underrepresented scripts such as ADLaM and N'Ko, demonstrating sensitivity to script distribution in pretraining. More broadly, Shani et al. (2026) attribute cross-linguistic performance gaps to architectural and data design choices (particularly tokenization fragmentation, encoding imbalance, and skewed sampling) rather than linguistic complexity. Complementing this, Liu et al. (2025) demonstrate that explicitly incorporating script data during multilingual pretraining improves downstream performance, establishing script structure as a meaningful modeling variable.

Script effects also appear in cross-lingual alignment and digitization. Transliteration-based posttraining can partially reduce performance disparities between scripts (Xhelili et al., 2024), indicating that representation space alignment is script-sensitive. At the infrastructural level, limited standardization, encoding constraints, and lack of script-specific tooling continue to restrict NLP feasibility for traditions such as Wolofal (Le et al., 2025; Zaugg, 2020; Zaugg et al., 2022).

Collectively, these findings indicate that the issue is not solely linguistic scarcity but structural imbalance across writing systems. Multilingual models may include many languages yet allocate disproportionate representational capacity to dominant scripts (Liu et al., 2025; Teklehaymanot and Nejdl, 2025), with some requiring up to thirteen times more tokens for equivalent content (Petrov et al., 2023; Asprovska and Hunter, 2024).

This survey makes four primary contributions: (1) we introduce a script-centric analytical framework for multilingual NLP, positioning writing systems alongside language as a fundamental unit of analysis; (2) we propose a four-layer model to sys-

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 96–106
12 May 2026.

tematically diagnose script-induced inequities; (3) we synthesize empirical evidence demonstrating that performance disparities across African scripts arise primarily from engineering design choices, not linguistic complexity; (4) we articulate a research agenda toward native multiscript foundation models, including script-sensitive scaling, tokenizer equity metrics, and evaluation reform.

2. Methodology

This survey adopts a multidimensional methodology that integrates technical analysis with sociotechnical critique. We conducted a structured literature review across major computational linguistics venues in the ACL Anthology, including ACL, EMNLP, NAACL, and AfricaNLP workshops, as well as recent preprints (2024–2026) on arXiv. Additional sources were identified through IEEE Access, the ACM Digital Library, SpringerLink, Google Scholar, institutional thesis repositories, and specialized outlets such as the *International Journal of Writing Systems* were also consulted.

2.1. Keywords and Selection Criteria

Search terms were grouped into three Boolean clusters: (1)Script/Identity: “African scripts,” “Indigenous writing systems,” “ADLaM,” “Tifinagh,” “N’Ko,” “Ge’ez,” “Ajami/Wolofal,” “Vai”; (2)Computational Mechanics:“Tokenization bias,” “subword segmentation,” “BPE fragmentation,” “Unicode integration”; (3)Critical Frameworks: “Digital sovereignty,” “data colonialism,” “epistemic bias,” “decolonial NLP.”

Studies were included if they (a) presented empirical evidence of performance variation associated with African scripts, (b) proposed technical interventions for the digital representation of Indigenous African writing systems, or (c) examined the historical development or digital infrastructure supporting these scripts. The review prioritizes contemporary research (2020–2026), while retaining foundational work on script taxonomy and encoding history for contextual grounding. After duplicate removal and screening based on these criteria, a final corpus of $N = 47$ studies was selected for full-text analysis.

2.2. Key Findings and Categorization

Each selected study was systematically coded using a structured extraction matrix capturing research questions, targeted scripts, methodological approach, and contributions to the decolonization of NLP pipelines.

Findings were organized across four analytical layers: (1) Infrastructural: Unicode allocation, keyboard layouts, font availability, and rendering constraints; (2) Representational: tokenization efficiency and vocabulary distribution; (3) Functional: downstream task performance, including arithmetic reasoning, named entity recognition, machine translation, and sentiment analysis; and (4) Epistemic: data sovereignty, colonial hierarchies, and evaluation fairness.

Figure 1 illustrates the distribution of studies across these layers, and Table 1 summarizes representative works, scripts examined, and thematic focus within each category.

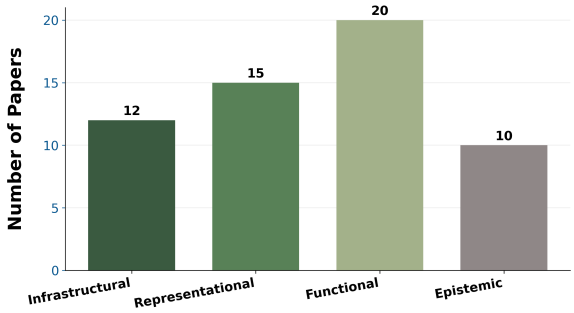

Figure 1: Distribution of surveyed studies across the four-layer analytical framework. Studies may span multiple layers, so counts are non-exclusive.

Layer	Primary Scripts Covered	Key Studies	Focus
Infrastructural	Ge’ez, Vai, ADLaM, N’Ko, Tifinagh, Ajami, Bamum	Kasonde (2025); Zaugg (2020); Simpson (2025); Graaf (2025)	Unicode encoding, font rendering, keyboard input, digital vitality, script history / taxonomy
Representational	ADLaM, N’Ko, Ge’ez	Ahia et al. (2024); Teklehaymanot and Nejdl (2025); Kanjirangat et al. (2025); Liu et al. (2025)	Tokenization bias, subword segmentation, vocabulary allocation, script-aware pretraining
Functional	ADLaM, Osmanya, N’Ko, Ge’ez, Ajami (Wolofal)	Reddy et al. (2026); Ojo et al. (2025); Le et al. (2025); Edman et al. (2025)	Downstream task performance: MT, NER, sentiment, arithmetic, benchmarks, HTR
Epistemic	Broad African / General Multilingual	Yan and Xu (2024); Adebara (2024); Ògúnrmí et al. (2023); Zaugg et al. (2022)	Data sovereignty, colonial hierarchies, evaluation fairness, decolonial frameworks

Table 1: Summary of surveyed literature by analytical layer, with representative studies and thematic focus

2.3. Limitations

The number of studies explicitly centered on Indigenous African scripts remains limited, and in many cases scripts appear only as secondary variables,

constraining direct cross-script comparison. The review also relies predominantly on Anglophone publication venues, which may overlook relevant scholarship published in Indigenous or regional languages not indexed in major databases.

3. Taxonomy of African Writing Systems and Computational Viability

African writing systems span a diverse landscape of typological families and historical trajectories, including long-standing Indigenous scripts, locally invented modern scripts (neo-traditional), and adapted traditions—most notably Latin and Arabic (*Ajami*)—integrated into African linguistic communities over centuries (Kelly, 2018; Voogt, 2014). This diversity carries direct computational consequences: script typology influences font rendering, keyboard design, Unicode inclusion, and the representational layers of NLP pipelines (Asprovska and Hunter, 2024; Kanjirangat et al., 2025; Simpson, 2025).

3.1. Typological Classification and NLP Implications

Historical and comparative studies situate Indigenous scripts like Vai and Bamum, and modern inventions like ADLaM, within a pattern of continuous African script innovation (Kelly, 2018; Voogt, 2014; Simpson, 2025). Voogt (2014) identifies a chronological shift from syllabic to alphabetic script design after the 1930s, driven by regional transmission patterns and missionary standardization rather than linguistic necessity.

Empirical research demonstrates that these scripts do not merely store information: they shape how models reason. Shifting from standard Hindu-Arabic digits to underrepresented scripts like Osmanya or N'Ko results in a measurable "script tax", where LLMs experience significant drops in arithmetic and logical accuracy despite identical underlying semantics (Reddy et al., 2026; Shani et al., 2026). This degradation is amplified by tokenization parity gaps: subword segmentation trained on Latin-dominant corpora fragments African-scripted text into disproportionately more tokens, increasing computational costs and reducing representational density (Ahia et al., 2024; Teklehaymanot and Nejdl, 2025; Asprovska and Hunter, 2024).

3.2. Infrastructural Constraints and Digital Survival

Indigenous and neo-traditional scripts face infrastructural marginalization. Digital survival depends on integration into global encoding standards, yet the path from script invention to Unicode inclusion is lengthy and institutionally mediated (Waddell, 2016; Simpson, 2025). Even after formal encoding, a lack of standardized keyboard layouts and font rendering engines often renders these scripts invisible to major datasets (Zaugg, 2020; Zaugg et al., 2022).

The case of Wolofal (*Wolof Ajami*) exemplifies the challenges of digraphia, where multiple scripts compete for the same language. Ajami functioned as a widespread literacy system in Senegal long before colonialism (Sall, 2020), yet its absence from standardized digital orthographies and the lack of Ajami-specific OCR mean that many historical documents remain computationally inaccessible (Le et al., 2025; Yousuf et al., 2026).

3.3. Orthographic Properties as Computational Bottlenecks

The structural properties of a script (grapheme composition, diacritics, and segmentation density) dictate how efficiently a tokenizer can process text (Kasonde, 2025). High character-to-token ratios in scripts like Ethiopic or Tifinagh lead to tokenization imbalance, which propagates bias into downstream tasks such as translation and summarization (Teklehaymanot et al., 2025). Addressing these gaps requires script-aware tokenization calibrated to the segmentation properties of each writing system (Teklehaymanot and Nejdl, 2025; Teklehaymanot et al., 2025).

4. Data Ecology and Corpus Provenance

The performance disparities observed across African writing systems are fundamentally rooted in data ecology. The availability, provenance, and digitization maturity of corpora determine whether a script is central or marginal to global NLP pipelines (Yan and Xu, 2024; Alabi et al., 2025; Hussen et al., 2025).

4.1. Historical Foundations and Digitization Gaps

Digital corpora for African languages are often constrained by colonial legacies. Long-standing Indigenous literacy traditions such as Ajami remain severely under-digitized due to historical suppression in favor of Latin-based systems, resulting in significant gaps in available textual resources (Sall, 2020). These gaps are compounded by limited Unicode support and font rendering infrastructure, which restrict script visibility in online repositories (Zaugg, 2020). Where digitization has been attempted, existing Arabic-trained OCR systems fail

to handle West African orthographic conventions, producing significant errors in manuscript processing (Yousuf et al., 2026).

4.2. Domain Concentration and Corpus Skew

African NLP datasets exhibit significant domain concentration, with heavy reliance on religious texts, news, and institutional translations (Alabi et al., 2025; Yan and Xu, 2024). This skew narrows lexical diversity and reinforces standardized orthographies over community-specific variants. Transliteration-based data augmentation compounds the problem by mediating script diversity through Latin representations, reducing the structural distinctiveness of original writing systems in training data (Xhelili et al., 2024).

4.3. Representational Inequity

Representational capacity in multilingual LLMs is implicitly distributed across scripts through vocabulary size, token frequency, and segmentation granularity. This creates a structural imbalance analogous to bandwidth allocation in communication system. Non-Latin scripts are over-segmented, requiring up to seven times more tokens for equivalent semantic content (Teklehaymanot and Nejdl, 2025; Kanjirangat et al., 2025; Petrov et al., 2023).

This imbalance has measurable functional consequences. Token inflation degrades reasoning and numeracy performance in scripts like N'Ko and ADLaM (Reddy et al., 2026), and reduces effective context windows for African-scripted text (Ahia et al., 2024). These disparities trace back to data allocation and subword segmentation design rather than to properties of the languages themselves (Shani et al., 2026; Emezue, 2026).

4.4. Quantifying Segmentation Imbalance

Several empirical indicators have been proposed to quantify script-level tokenization disparities, including characters-per-token ratios, tokens-per-word ratios, and fragmentation rates across scripts (Teklehaymanot and Nejdl, 2025).

Table 2 presents a cross-study synthesis of tokenization disparities across script families. Reported metrics include Tokens per Sentence (TPS), Characters per Token (CPT), and relative Token Premium compared to English. Color coding denotes efficiency tiers ranging from optimal to severely disadvantaged, with gray indicating unmeasured cases.

TPS values are computed using the cl100kbase tokenizer on FLORES-200 Teklehaymanot and Nejdl, 2025. CPT captures average character-to-token ratios, while Token Premium reflects relative token inflation across tokenizer types compared to English (Petrov et al., 2023; Asprovska and Hunter, 2024). Segmentation behavior summarizes documented BPE granularity patterns (Ahia et al., 2024; Kanjirangat et al., 2025). Prior studies report inflation ratios of up to 13 times across 108 languages, with disparities persisting even under byte-level models (Asprovska and Hunter, 2024; Petrov et al., 2023).

Synthesizing these metrics across script families reveals a marked efficiency gradient, from relatively compressed Latin-based scripts to over-segmentation in Ethiopic and Devanagari. Notably, Indigenous African scripts such as ADLaM, N'Ko, and Tifinagh remain absent from comparative benchmarks. This omission constitutes more than incomplete coverage: it reflects a diagnostic blind spot, where the scripts most in need of systematic evaluation are excluded from the measurement frameworks used to assess inequity.

These distortions are unevenly distributed: scripts with complex grapheme structures or extensive diacritic systems are particularly prone to fragmentation under subword tokenization schemes trained on Latin-dominant corpora (Teklehaymanot and Nejdl, 2025). In digraphic contexts, such as Ajami, orthographic variation and parallel script usage further destabilize segmentation (Le et al., 2025). Tailored tokenizer designs for specific scripts have demonstrated potential in mitigating these effects, enhancing both tokenization efficiency and evaluation stability in morphologically rich languages like Tigrinya (Teklehaymanot et al., 2025).

4.5. Modeling Consequences

Frequency-driven vocabulary allocation further amplifies the representational gaps described above. Subword merges optimized on Latin-dominant corpora consistently under-allocate capacity to low-frequency scripts, resulting in persistent over-segmentation that inflates sequence length and raises per-token attention costs (Kanjirangat et al., 2025; Asprovska and Hunter, 2024; Petrov et al., 2023). Scripts with dense grapheme inventories and rich morphological structures, such as Ethiopic, are especially vulnerable: generic segmentation schemes fail to capture their structural patterns, and current models support only around 42 African languages despite over 2,000 being spoken (Hussen et al., 2025). Incorporating script-specific metadata into multilingual pretraining pipelines has shown measurable downstream benefits, underscoring the importance of treating script properties as explicit training signals (Liu et al., 2025).

Script Family	TPS	CPT	Prem.	Segmentation Behavior
Latin	50.2	2.61	1.0×	Word-level; optimal compression across all benchmarks.
Han (Simplified)	56.8	—	0.9–1.1×	Near-parity; single-token characters compensate for multi-byte encoding.
Cyrillic	—	1.58	1.2–2.5×	Moderate fragmentation; stable TP in encoder models.
Arabic	—	1.28	1.1–1.4×	Alphabet-family consistency (SD = 0.51); reduced compression.
Ethiopic (Ge'ez)	—	<1.0	>2.0×	Over-segmentation; dense grapheme inventory and diacritics exacerbate fragmentation.
Devanagari	—	0.99	—	Subword-level under BPE; character-level in byte models. MAGNET reduces to word-level.
Myanmar	357.2	—	4.4–12×	Character/byte-level fragmentation; highest tokenization cost measured.
Tibetan	—	0.49	3.7–6.7×	Severe fragmentation; UTF-8 multi-byte overhead compounds disparity.
ADLaM, N'Ko, Tifinagh, Osmanya, Vai, and Bamum are absent from all major tokenization benchmarks (FLORES-200, Common Crawl evaluations). No TPS, CPT, or segmentation data exists for these African scripts.				

Table 2: Cross-study synthesis of tokenization disparities across script families. Metrics include Tokens per Sentence (TPS), Characters per Token (CPT), and relative Token Premium compared to English. Color coding indicates efficiency tiers (optimal to severely disadvantaged), with gray denoting unmeasured cases.

5. Digital Infrastructure and Standardization

The digital vitality of African writing systems depends on a complete technical ecosystem, spanning encoding standards, input tools, and accessible corpora. Evidence indicates that this infrastructure shapes which scripts achieve functional usability in digital environments and which remain marginalized (Zaugg et al., 2022; Yan and Xu, 2024). Simpson (2025) further shows that Unicode operates as a gatekeeping mechanism, conferring technological legitimacy through processes that are as institutional and political as they are technical.

5.1. Unicode Integration and Encoding Constraints

Unicode inclusion marks a critical first step toward digital recognition, but achieving encoding is a prolonged process. For instance, ADLaM was invented in 1989 but not included in Unicode until version 9.0 in 2016, a trajectory requiring decades of grassroots advocacy (Waddell, 2016; Simpson, 2025). Subsequent platform adoption further illustrates the scale of post-encoding effort: Microsoft invested in font development and keyboard integration to support ADLaM alongside several other African scripts, framing digital infrastructure as a vehicle for script revitalization and cultural preservation (Bach, 2019). Nevertheless, formal inclusion alone does not ensure practical usability, as digital inequities persist when platform-level implementation lags behind standardization (Zaugg, 2020; Zaugg et al., 2022).

Beyond encoding, orthographic inconsistency presents an additional challenge. Even languages with established Unicode support, such as Icibemba, experience unstandardized grapheme-to-phoneme mappings that complicate automated processing (Kasonde, 2025). Ajami traditions face an amplified version of this problem, where the absence of unified orthographic conventions across regions introduces variation that destabilizes corpus creation (Le et al., 2025). More broadly, Graaf (2025) argues that current text encoding models, including Unicode, rely on assumptions—such as linearity, plain text, and formatting independence—that fail to capture the structural complexity of many writing systems.

5.2. Input Systems and Digital Mediation

Even after scripts are encoded, daily writing depends on accessible input tools. The prevalence of ASCII-compatible systems and QWERTY keyboards makes native-script typing cumbersome, prompting many users to adopt Latin transliteration (Zaugg et al., 2022; Yan and Xu, 2024). In Ethiopia, limited Ethiopic input tool usability has driven frequent Latin transliteration, affecting the representation of the script in digital corpora (Zaugg, 2020). Researchers working with African manuscripts often develop custom keyboards and specialized tools to accommodate extended diacritics and script-specific characters, particularly for Ajami texts (Yousuf et al., 2026). The development of ADLaM support in Microsoft Windows—including the Ebrima font and dedicated keyboard layouts—illustrates the scale of platform integration required to bridge the gap between Unicode encoding and everyday usability (Bach, 2019). This dependence on sustained advocacy and ad hoc solutions highlights the persistent distance between formal encoding and practical accessibility, reinforcing corpus imbalances.

6. NLP Tasks Across African Scripts

NLP tasks serve as diagnostic lenses, revealing script-specific constraints. Empirical studies indicate that performance differences across African languages are frequently driven by orthographic structure and tokenization design rather than intrinsic linguistic complexity (Reddy et al., 2026; Shani et al., 2026). Large-scale benchmarks further show that current LLMs consistently underperform on African languages across multiple tasks, and that simply increasing model size does not eliminate these disparities (Ojo et al., 2025).

6.1. Morphological Processing and Machine Translation

Morphologically rich languages are especially sensitive to tokenizer design. Syllable-based tokenization improves representation quality in syllable-rich languages like Swahili, outperforming statistically derived subword methods that fail to capture agglutinative structures (Atuhurra et al., 2024). In Southern African Bantu languages, BPE vocabulary size and tokenizer implementation strongly influence translation quality, with SentencePiece outperforming subword-nmt in agglutinative contexts (Rajab, 2022). Language-specific tokenizers for Swahili, Hausa, and Yoruba further demonstrate that monolingual or regional tokenizers outperform global multilingual alternatives (Erasmo Ndomba et al., 2025). For Nguni languages, learning subword segmentation during training rather than relying on fixed preprocessing yields stronger results under low-resource conditions (Meyer, 2025).

This finding suggests that current "one-size-fits-all" multilingual models suffer from a **representational bottleneck**, where global vocabularies fail to capture the morphological density of African scripts. It implies that for a multilingual system to be truly equitable, it must incorporate decentralized, script-specific representational layers rather than relying on a single shared vocabulary.

In machine translation, script mismatch represents an additional modeling barrier beyond lexical divergence. Transliteration-based post-training improves cross-script alignment (Xhelili et al., 2024), while script-aware tokenization and domain-adaptive fine-tuning are critical for morphologically complex targets such as Tigrinya (Teklehaymanot et al., 2025; Gaim and Park, 2025).

6.2. Classification, NER, and Sentiment Analysis

Sentiment analysis and emotion recognition benchmarks reveal persistent performance gaps for African languages. The AfriSenti dataset, covering 14 languages, shows that language-specific pretraining substantially improves classification accuracy compared to generic multilingual models (Muhammad et al., 2023). Similarly, the BRIGHTER dataset demonstrates that current LLMs struggle with emotion recognition across 28 languages, highlighting significant limitations for low-resource contexts (Muhammad et al., 2025). In healthcare applications, MT and NER errors in African languages pose safety-relevant risks, with tokenization inefficiency and dataset imbalance identified as key contributors (Okafor, 2025).

Small, targeted language models trained on curated regional corpora can outperform much larger general-purpose LLMs on classification and generation tasks for low-resource African languages, indicating that data quality and architectural alignment are more influential than model scale (Otoibhi et al., 2025).

6.3. Script-Level Evaluation

Evaluation benchmarks are increasingly incorporating script-level diagnostics. The EXECUTE benchmark assesses character- and word-level token manipulations across diverse scripts, showing that task difficulty is shaped more by writing system structure and tokenization segmentation than by character count alone (Edman et al., 2025). For digraphic and manuscript traditions, OCR and handwritten text recognition require specialized datasets and preprocessing pipelines tailored to regional orthographic conventions (Yousuf et al., 2026).

Table 3 formalizes what Reddy et al. (2026) describe as the "script tax": a systematic performance penalty arising from script representation rather than task complexity, with converging evidence from numeracy tasks, multi-task benchmarks, performance decomposition, and token manipulation evaluations.

7. Toward Native Multi-Script Language Models

The findings of this survey highlight three interdependent strategies for achieving script-level equity in multilingual NLP: tokenization reform, balanced pretraining, and evaluation redesign.

7.1. Script-Aware Tokenization

Mitigating segmentation imbalances requires moving beyond frequency-driven subword methods. Ahia et al. (2024) introduce script-specific adaptive compression that equalizes segmentation rates across scripts without compromising model quality, showing that fairness can be explicitly incorporated as a tokenization parameter. Similarly, learning segmentation during training rather than relying

Task Domain	Baseline	Script-Level Impact	Δ	Primary Disparity Source
Arithmetic reasoning	HA numerals: ≈100% accuracy (4 LLMs)	ADLaM, N'Ko, Osmanya: ≈0% (excluded from regression). All non-HA scripts show sig. negative coefficients ($p < 10^{-4}$).	66–100%	Tokens-per-digit ($\beta = -0.198$, $p < 10^{-8}$); script under-representation in pretraining corpora.
Multi-task NLU (7 tasks, 64 langs)	English avg: 70.0% (Gemma 2 9B)	African language avg: 39.6%. Largest gaps on knowledge QA and math.	−30.4 pp	Resource availability; tokenization inefficiency; gaps persist with model scaling.
Knowledge & reasoning	English MMLU: 69.8%; Math: 68.8%	African MMLU: 36.1%; Math: 20.7% (same model).	−33.7 / −48.1 pp	Reasoning-intensive tasks amplify script-mediated disparities beyond surface-level NLU.
Token manipulation (char/word level)	English: 64.8% avg (Gemma 2 9B)	Arabic: 51.6%; Hindi: 57.9%. Amharic/Ge'ez: ≈98% (inverse effect).	Variable	CWT statistics predict difficulty; low-resource scripts may outperform due to absence of learned linguistic bias.
Disparity decomposition: Performance gaps decompose into orthographic (UTF-8 byte asymmetries, shared-vocabulary bias toward Latin), morphological (disappears under morphology-aware tokenization), lexical (amplified by subword fragmentation), and data exposure factors. When these design choices are controlled for, much of the apparent difficulty diminishes, suggesting the gaps are artifacts of pipeline construction.				

Table 3: The "script tax": task-level performance degradation attributable to script representation. Arithmetic data from Reddy et al. (2026), who coined the term; multi-task and knowledge benchmarks from AfroBench (Ojo et al., 2025); token manipulation from EXECUTE (Edman et al., 2025); disparity decomposition from Shani et al. (2026). Δ = performance gap relative to baseline. pp = percentage points. HA = Hindu-Arabic. CWT = character-word-token ratio. The convergence across independent studies confirms that these performance penalties are traceable to pipeline design (tokenization, data allocation, and vocabulary construction) rather than to inherent task difficulty.

on fixed preprocessing improves performance for morphologically complex languages (Meyer, 2025), and region-specific tokenizers consistently outperform global multilingual alternatives.

7.2. Balanced Multiscript Pretraining

Equitable modeling necessitates careful control over vocabulary allocation and sampling distributions during pretraining (Emezue, 2026; Teklehaymanot and Nejdl, 2025). Region-specific foundation models trained on curated corpora provide a practical alternative to large-scale parameter expansion, with evidence that architectural alignment and data quality can outweigh sheer model size for African language tasks (Otoibhi et al., 2025). More broadly, calls for infrastructural reform highlight the importance of diversified corpora, robust tooling ecosystems, and community-led data governance as essential foundations for sustainable progress (Yan and Xu, 2024; Adebara, 2024; Ògúnrmí et al., 2023).

7.3. Evaluation Reform and Epistemic Reframing

Current evaluation frameworks can reinforce script bias by assuming Latin-compatible orthographic norms or relying on benchmarks insensitive to graphemic variation and segmentation instability (Reddy et al., 2026; Edman et al., 2025). Persistent use of the "low-resource" label risks normalizing scarcity that stems from infrastructural inequality, historical standardization, and policy neglect (Zaugg et al., 2022; Ògúnrmí et al., 2023; Minhas and Salawu, 2024). Script-robust evaluation should integrate metrics sensitive to formatting variation and segmentation stability, along with native-script benchmarks for manuscript traditions and non-Latin writing systems (Yousuf et al., 2026; Ojo et al., 2025). Addressing these challenges requires not only technical innovation but a reorientation of evaluation design toward the linguistic and orthographic realities of African communities (Adebara, 2024; Emezue, 2026).

The "low-resource" label normalizes scarcity by framing performance gaps as an inevitable consequence of data volume, which obscures the mechanical role of engineering choices—such as the $13\times$ **token inflation** observed in some African scripts. This framing discourages the development of script-aware architectures by treating technical failure as a data limitation.

8. Research Gaps and Open Problems

Despite growing awareness of script-level disparities, several structural gaps remain. First, script-aware interventions—such as custom tokenizers, adaptive segmentation, and script-informed pretraining—are largely isolated experiments rather than integrated into general-purpose multilingual models (Teklehaymanot et al., 2025; Liu et al., 2025; Ahia et al., 2024). The interaction between segmentation design and morphological richness has yet to be systematically examined across script families (Teklehaymanot and Nejdl, 2025; Meyer, 2025).

Second, script-sensitive scaling behavior is

poorly understood. While representational variation clearly affects performance, comprehensive scaling laws that account for script diversity have not been established (Reddy et al., 2026; Shani et al., 2026). It remains an open question whether increasing model or data scale reduces or exacerbates script-level disparities (Ojo et al., 2025; Xuan et al., 2025).

Third, digitally disadvantaged languages often enter NLP pipelines through transliteration or partial tooling rather than fully native-script pathways (Yan and Xu, 2024; Zaugg et al., 2022). Coordinated efforts to develop multiscript corpora, segmentation-aware architectures, and script-sensitive evaluation frameworks are essential for achieving structural progress.

9. Conclusion

This survey has argued that writing systems, rather than languages alone, represent a key axis of inequity in multilingual NLP. Across the four analytical layers considered—infrastructural, representational, functional, and epistemic—a clear pattern emerges: African scripts are systematically disadvantaged by choices embedded in encoding standards, tokenization algorithms, corpus construction, and evaluation frameworks—rather than by the inherent properties of the languages or scripts themselves.

Framing African languages as "low-resource" obscures the infrastructural and historical conditions—such as encoding exclusion, corpus imbalance, and evaluation bias—that generate scarcity (Ògúnrmí et al., 2023; Zaugg et al., 2022). Sustainable progress requires rethinking governance, data ownership, and modeling assumptions away from extractive paradigms (Adebara, 2024; Yan and Xu, 2024), a principle that directly extends to digital environments where script exclusion perpetuates historical marginalization.

Addressing these disparities calls for coordinated structural reform: integrating script-aware tokenization into general-purpose models, imposing deliberate constraints on vocabulary allocation during pretraining, and developing evaluation frameworks capable of detecting script-level asymmetries. Without treating writing systems as explicit computational variables, multilingual models risk reproducing the very digital hierarchies they inherit.

The future of multilingual NLP depends not only on adding more languages, but on rethinking how writing systems are represented, allocated, and evaluated within foundational architectures. Achieving multiscript equity is not an optional extension of multilingual NLP—it is its foundational precondition.

Ultimately, reclaiming African voices in the digital age requires treating writing systems as explicit computational variables; this is not merely a technical adjustment, but a structural precondition for a truly inclusive and decolonized multilingual NLP.

10. Ethics Statement

This survey examines structural inequities in how NLP systems represent African writing systems. Several ethical considerations warrant explicit acknowledgment, as highlighted below.

Positionality and Framing. We adopt a decolonial framework that foregrounds power asymmetries in language technology. While we argue that the "low-resource" label may reflect structural marginalization as much as a technical condition, we acknowledge that framing choices carry epistemic consequences. Our analysis is conducted from an academic institutional perspective, with efforts to center the priorities and perspectives articulated by African NLP researchers and communities throughout the surveyed literature.

Community Agency and Data Sovereignty. African writing systems are disadvantaged by design choices embedded in encoding standards, tokenization algorithms, and corpus construction practices. Addressing these disparities must involve affected communities as decision-makers, not merely as data providers. Corpus development, orthographic standardization, and digital infrastructure design for Indigenous scripts should be guided by the communities themselves, in line with principles of data sovereignty and self-determination emphasized in the reviewed literature.

Risks of Reductive Framing. Survey-level analysis requires generalization across diverse scripts, languages, and communities. We caution against treating African writing systems as monolithic; the scripts examined (ADLaM, N'Ko, Vai, Ge'ez, Tifinagh, Ajami, and others) differ in typology, history, digital maturity, and community context. Our four-layer analytical framework is intended as a diagnostic tool, not as a prescriptive or uniform solution.

Potential for Misuse. Although this work aims to promote equity in NLP, documenting script-level vulnerabilities—such as tokenization disparities or gaps in digital infrastructure—could theoretically be misused to justify exclusion or neglect. We encourage the NLP community to interpret these findings as motivation for structural investment rather than as evidence of inherent technical intractability.

No Human Subjects. This study is a literature survey and does not involve human participants,

personal data collection, or experimentation on individuals or communities.

11. Bibliographical References

Ifeoluwanimi Adebara. 2024. *Towards Afrocentric natural language processing*. Ph.D. thesis, University of British Columbia.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A. Smith. 2024. MAGNET: Improving the Multilingual Fairness of Language Models with Adaptive Gradient-Based Tokenization. *Advances in Neural Information Processing Systems*, 37:47790–47814.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the Landscape of African NLP: Mapping Progress and Shaping the Road Ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Marijana Asprovska and Nathan Hunter. 2024. The Tokenization Problem: Understanding Generative AI's Computational Language Bias. *Ubiquity Proceedings*, 4(1).

Jesse Atuhurra, Hiroyuki Shindo, Hidetaka Kamigaito, and Taro Watanabe. 2024. Introducing Syllable Tokenization for Low-resource Languages: A Case Study with Swahili. ArXiv:2406.15358 [cs].

Deborah Bach. 2019. Ibrahima & Abdoulaye Barry: How a new alphabet is helping an ancient people write its own future. Microsoft New Zealand News Centre.

Lukas Edman, Helmut Schmid, and Alexander Fraser. 2025. EXECUTE: A Multilingual Benchmark for LLM Token Understanding. ArXiv:2505.17784 [cs] version: 1.

Chris Chinenye Emezue. 2026. *Improving language models for underserved languages and communities*. Ph.D. thesis, Université de Montréal.

Goodwill Erasmo Ndomba, Medard Edmund Mswahili, and Young-Seob Jeong. 2025. Tokenizers for African Languages. *IEEE Access*, 13:1046–1054.

Fitsum Gaim and Jong C. Park. 2025. Natural Language Processing for Tigrinya: Current State and Future Directions. ArXiv:2507.17974 [cs].

Kevin Graaf. 2025. Carving Text at Its Joints: A New Perspective on Writing and Computers. In *Proceedings of the 2025 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, Onward! '25, pages 194–203, New York, NY, USA. Association for Computing Machinery.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The State of Large Language Models for African Languages: Progress and Challenges. ArXiv:2506.02280 [cs].

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and Representation Biases in Multilingual Models on Dialectal NLP Tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23992–24010, Suzhou, China. Association for Computational Linguistics.

Alex Kasonde. 2025. The long march to Unicode: a digital approach to variability in Icibemba orthography. *Cogent Arts & Humanities*, 12(1):2477347. _eprint: https://doi.org/10.1080/23311983.2025.2477347.

Piers Kelly. 2018. The invention, transmission and evolution of writing: Insights from the new scripts of West Africa. *Paths into Script Formation in the Ancient Mediterranean*. ISBN: 9788871408989.

Ngoc Tan Le, Ali Mijiyawa, Abdoulahat Leye, and Fatiha Sadat. 2025. The Best of Both Worlds: Exploring Wolofal in the Context of NLP. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 1–6, Abu Dhabi, UAE. Association for Computational Linguistics.

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, and Hinrich Schuetze. 2025. LangSAMP: Language-Script Aware Multilingual Pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1743–1770, Vienna, Austria. Association for Computational Linguistics.

Francois Rolihlahla Meyer. 2025. *Subword segmental neural language generation for Nguni languages*. Ph.D. thesis, University of Cape Town.

Shahid Minhas and Abiodun Salawu. 2024. Strategic Frameworks for the Empowerment of African Languages: Policy, Practice and Prospects. *Forum for Linguistic Studies*, 6(6):753–766.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Ugochi Okafor. 2025. Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 221–229, Vienna, Austria. Association for Computational Linguistics.

Jeffrey Otoibhi, Oduguwa Damilola, and Okpare David. 2025. SabiYarn: Advancing Low Resource Languages with Multitask NLP Pretraining. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 95–107, Vienna, Austria. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *Advances in Neural Information Processing Systems*, 36:36963–36990.

Jenalea Rajab. 2022. Effect of Tokenisation Strategies for Low-Resourced Southern African Languages. In *3rd Workshop on African Natural Language Processing*.

Varshini Reddy, Craig W. Schmidt, Seth Ebner, Adam Wiemerslage, Yuval Pinter, and Chris Tanner. 2026. The Effect of Scripts and Formats on LLM Numeracy. ArXiv:2601.15251 [cs].

Mamadou Youry Sall. 2020. African Ajami: The Case of Senegal. In Jamaine M. Abidogun and Toyin Falola, editors, *The Palgrave Handbook of African Education and Indigenous Knowledge*, pages 545–557. Springer International Publishing, Cham.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The Roots of Performance Disparity in Multilingual Language Models: Intrinsic Modeling Difficulty or Design Choices? ArXiv:2601.07220 [cs].

Logan David Simpson. 2025. *Modern Indigenous writing systems: From inception to Unicode*. Ph.D. thesis, Queen Mary University of London.

Hailay Kidu Teklehaymanot, Gebrearegawi Gidey, and Wolfgang Nejdl. 2025. Low-Resource English-Tigrinya MT: Leveraging Multilingual Models, Custom Tokenizers, and Clean Evaluation Benchmarks. ArXiv:2509.20209 [cs].

Hailay Kidu Teklehaymanot and Wolfgang Nejdl. 2025. Tokenization Disparities as Infrastructure Bias: How Subword Systems Create Inequities in LLM Access and Efficiency. ArXiv:2510.12389 [cs].

Alex de Voogt. 2014. The Cultural Transmission of Script in Africa: the presence of syllabaries. *International Journal of Writing Systems: SCRIPTA*.

Kaveh Waddell. 2016. The Alphabet That Will Save a People From Disappearing. *The Atlantic*. Section: Technology.

Orgest Xhelili, Yihong Liu, and Hinrich Schuetze. 2024. Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11283–11296, Miami, Florida, USA. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Nan Yan and Cheng Xu. 2024. Decolonizing African NLP: A Survey on Power Dynamics and Data Colonialism in Tech Development. In *5th Workshop on African Natural Language Processing*.

Oreen Yousuf, Abdulmalik Aminu, Musa Salih Muhammad, Bashir Usman, Mustapha Kurfi Hashim, Joakim Nivre, Beáta Megyesi, and Christian Høgel. 2026. A Handwritten Text Recognition Dataset for Ajami Manuscripts in Fulfulde and Hausa. In *Document Analysis and Recognition – ICDAR 2025*, pages 620–637, Cham. Springer Nature Switzerland.

Isabelle A. Zaugg. 2020. Digital Inequality and Language Diversity: An Ethiopic Case Study. In Massimo Ragnedda and Anna Gladkova, editors, *Digital Inequalities in the Global South*, pages 247–267. Springer International Publishing, Cham.

Isabelle A. Zaugg, Anushah Hossain, and Brendan Molloy. 2022. Digitally-disadvantaged languages. *Internet Policy Review*, 11(2):1–11.

Tolúlp Ògúnrmí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing NLP for "Low-resource Languages": Applying Abebe Birhane's Relational Ethics. *GRACE: Global Review of AI Community Ethics*, 1(1).