Language inequality in artificial intelligence: Nigerian languages, AI underrepresentation, and digital inclusion

Daniel Nwonye

Abstract

Artificial intelligence increasingly shapes access to information, Education, public services, civic communication, and economic opportunity. However, AI systems do not support all languages equally. This paper examines the underrepresentation of Nigerian languages in artificial intelligence and its implications for digital inclusion. It has three objectives: to examine how selected Nigerian languages and language varieties are represented in AI and NLP systems; to assess how selected public-facing AI tools respond to English, Nigerian-language, and mixed-language examples; and to discuss the implications of language inequality for digital inclusion, public services, Education, civic participation, and economic opportunity. Focusing on Yorùbá, Igbo, Hausa, and Nigerian Pidgin as illustrative cases, the paper employs conceptual analysis, secondary evidence from African natural language processing research, and a small exploratory assessment of AI tools, including ChatGPT, Google Gemini, and Microsoft 365 Copilot Chat. The findings suggest that language inequality in AI operates across data availability, model representation, and user-facing performance. Although selected Nigerian languages appear in emerging African NLP resources, their presence does not guarantee robust support for public-facing use. The AI-tool assessment showed that English responses were more stable and complete. In contrast, Nigerian-language and mixed-language outputs exhibited greater variation in meaning preservation, register, cultural context, information completeness, and public usefulness. The paper argues that Nigerian-language underrepresentation is both a technical problem and a governance issue shaped by low-resource conditions, limited datasets, market incentives, infrastructure constraints, and postcolonial language politics. It concludes that linguistically inclusive AI in Nigeria requires stronger datasets, community-led evaluation, policy support, sustainable funding, and AI governance frameworks that treat Nigerian languages as part of digital infrastructure.

Keywords: language inequality, natural language processing, AI underrepresentation; digital inclusion; AI governance;

Daniel Nwonye, danynnocent@gmail.com, *Independent Researcher, Port Harcourt, Nigeria*

The Nigerian Journal of Business and Social Sciences, Volume 13, 2026 (Special Issue)
A Journal of the Faculty of Social Sciences, University of Lagos, Akoka, Lagos, Nigeria

1. Introduction

Artificial intelligence is increasingly shaping how people access information, communicate, learn, work, and participate in public life. Through natural language processing (NLP), AI systems support machine translation, search engines, chatbots, voice assistants, educational platforms, sentiment analysis, and public information systems. However, these systems do not serve all languages equally. Their performance depends on the availability of digitized text, speech data, annotated datasets, computational resources, and research communities. As a result, languages with large digital footprints, especially English and other globally dominant languages, are usually better supported than languages with limited digital resources (Joshi et al., 2020; Ogueji et al., 2021). This imbalance is important for Nigeria because the country is highly multilingual. English remains the official language of government, education, law, and many digital systems. However, everyday communication also depends on Nigerian languages and language varieties such as Yorùbá, Igbo, Hausa, Nigerian Pidgin, and many others. These languages carry local knowledge, cultural meaning, public trust, social identity, and everyday civic communication. However, they remain underrepresented in AI and NLP systems, creating a gap between the languages people use in daily life and those most strongly supported by digital technologies. African NLP research shows that this gap is not accidental. Major multilingual systems have historically given limited representation to African languages, while English and other high-resource languages receive stronger data coverage, evaluation, and model support (Joshi et al., 2020; Ogueji et al., 2021). Although initiatives such as Masakhane, MasakhaNER, NaijaSenti, AfriBERTa, and No Language Left Behind show that progress is possible, African and Nigerian language support remains uneven, task-specific, and insufficient for many public-facing AI uses (Nekoto et al., 2020; Adelani et al., 2021; Muhammad et al., 2022; NLLB Team et al., 2022). This paper examines the underrepresentation of Nigerian languages in AI and its implications for digital inclusion. It focuses on Yorùbá, Igbo, Hausa, and Nigerian Pidgin as illustrative case-study languages because they are widely used in Nigeria and appear in emerging African NLP research and datasets. The paper has three objectives: to examine how selected Nigerian languages and language varieties are represented in AI and NLP systems; to assess how selected public-facing AI tools respond to English and selected Nigerian-language or mixed-language examples; and to discuss the implications of language inequality for digital inclusion, public services, Education, civic participation, and economic opportunity. The paper contributes to the existing literature by linking the underrepresentation of African-language NLP to digital inclusion in Nigeria. It argues that linguistic inclusion should not be measured only by whether a language appears in a dataset, model, or AI tool, but by whether AI systems can process that language accurately, fluently, culturally, and usefully in real-world tasks. Although the issue has broader relevance across Africa, this paper focuses on selected Nigerian languages and language varieties. The paper proceeds by presenting the conceptual framework, explaining the methodology, examining Nigerian-language underrepresentation, analysing selected AI-tool outputs, discussing digital inclusion implications, and proposing pathways toward linguistically inclusive AI.

2. Conceptual Framework and Key Concepts

This section clarifies the main concepts used in the paper and presents the conceptual framework that guides the analysis. The paper argues that language inequality in artificial intelligence is not only a technical issue but also a digital inclusion issue. To make this argument clear, it is necessary to define the key concepts and explain how they are connected.

2.1 Key Concepts

Natural Language Processing (NLP) refers to AI systems that process, analyse, generate, translate, or respond to human language. In this paper, NLP is important because AI performance depends on the availability and quality of language data (Joshi et al., 2020; Ogueji et al., 2021). Language inequality refers to unequal visibility, support, and performance across languages in social, political, educational, and technological systems. In AI, it appears that some languages receive stronger datasets, model support, evaluation, and research attention than others (Joshi et al., 2020; Blasi et al., 2022). Low-resource languages are languages with limited machine-readable text, speech data, annotated corpora, translation datasets, benchmarks, and computational tools. Here, low-resource status is treated as a structural condition shaped by data availability, investment, standardisation, evaluation, and AI-system integration (Joshi et al., 2020; Kreutzer et al., 2022; Ogueji et al., 2021). AI underrepresentation refers to the weak or uneven inclusion of a language in AI datasets, models, tools, and evaluation systems. It may involve absence from major systems, low data volume, poor task support, or low-quality representation (Ogueji et al., 2021; Blasi et al., 2022). Digital inclusion means meaningful access to digital technologies, services, and opportunities. This paper includes linguistic accessibility: the ability to use AI tools and public information systems in languages that are clear, familiar, and socially meaningful (Warschauer, 2003; Bird, 2020). AI governance refers to the policies, institutions, ethical principles, and accountability systems that shape AI development and use. In relation to language, it includes decisions about language prioritisation, data collection, dataset documentation, evaluation, and public benefit (Bender & Friedman, 2018; UNESCO, 2021). Postcolonial language politics refers to the persistence of colonial languages' dominance in Education, governance, media, administration, and digital systems. In Nigeria, English dominance creates a gap between official digital systems and everyday communication in Yorùbá, Igbo, Hausa, Nigerian Pidgin, and other local language forms (Bamgbose, 1991; Bird, 2020).

2.2 Conceptual Framework

The conceptual framework of this paper links language data inequality to digital exclusion through AI underrepresentation and uneven AI performance. The framework assumes that AI systems depend heavily on the availability, quality, and diversity of language data. Where a language has limited digital text, speech resources, annotated datasets, and evaluation benchmarks, it is less likely to be strongly represented in AI and NLP systems. Weak representation can then affect how well AI tools perform in translation, information extraction, public communication, Education, civic participation, and other digital tasks.

The framework can be expressed as follows:

Language data inequality → AI underrepresentation → uneven AI performance → linguistic digital exclusion

This framework does not suggest that language is the only factor shaping digital exclusion. Other factors such as poverty, infrastructure, literacy, affordability, digital skills, and access to computing resources also matter. However, it argues that language is a central and often neglected dimension of digital inclusion. A person may have internet access and a mobile device but still be excluded from the full benefits of AI if the available tools do not work effectively in the language that person understands and uses. The framework guides the analysis in three ways. First, it explains why Nigerian and other African languages remain underrepresented in AI systems despite their social importance. Second, it provides a basis for examining evidence from NLP datasets, multilingual models, and selected AI-tool outputs. Third, it supports the paper's argument that linguistic inclusion

should be treated as part of AI governance and digital development policy, rather than merely a cultural or technical afterthought.

3. Methodology

3.1 Research Design

This study adopts a qualitative exploratory research design. It combines conceptual analysis, a review of the secondary literature, and a small exploratory assessment of an AI tool. The conceptual analysis clarifies the relationship between natural language processing, language inequality, AI underrepresentation, and digital inclusion. The literature review examines existing evidence on the representation of African and Nigerian languages in AI and NLP systems. At the same time, the AI-tool assessment illustrates how uneven language support may manifest in everyday AI use. The study is not designed as a large-scale benchmark or statistically generalisable NLP evaluation. Rather, it provides exploratory evidence linking technical underrepresentation in AI systems to broader concerns about access to information, Education, civic participation, public services, and economic opportunity.

3.2 Sources of Evidence

The study draws on three main sources of evidence. First, it draws on peer-reviewed research in natural language processing, African-language NLP, multilingual language models, dataset quality, low-resource language modelling, and language technology evaluation. Second, it draws on policy and practitioner literature on digital inclusion, AI governance, civic technology, and African-language technology development to connect language underrepresentation with broader questions of access, participation, public services, governance, and development. Third, it uses a small AI-tool assessment involving selected public-facing AI systems to provide illustrative evidence of how uneven language support may appear in everyday use.

3.3 Exploratory AI-Tool Assessment

The assessment examined ChatGPT GPT-5.5 Thinking, Google Gemini 3.5 Flash, and Microsoft 365 Copilot Chat GPT-5.3 Instant through their public chat interfaces in May 2026. These model versions were recorded because public-facing AI systems are frequently updated. The languages and varieties assessed were English, Yorùbá, Igbo, Hausa, Nigerian Pidgin, and mixed Nigerian Pidgin/English. English served as the high-resource baseline, while the Nigerian languages and varieties were selected because they are widely used in Nigeria and are emerging in African NLP research. The assessment covered six task areas: public communication clarity, public health information extraction, public safety/security understanding, code-switching interpretation, cultural-context recognition, and informal register recognition. The same prompts were entered into each system and evaluated against expected meaning elements. Typed prompts were used for the main scoring exercise. Voice-input trials, where available, were treated only as supplementary observations and were not included in Table 1.

3.4 Evaluation Criteria and Scoring Procedure

AI outputs were assessed through checklist-based scoring, qualitative error analysis, and native-speaker review. The evaluation considered meaning preservation, information completeness, fluency, register, cultural context, and public usefulness. Automatic metrics such as BLEU, chrF, COMET, and BERTScore were not used as the main evaluation method because they require reference outputs and are limited in assessing tone, code-switching, cultural meaning, and public usefulness (Papineni et al., 2002; Popović, 2015; Rei et al., 2020; Zhang et al., 2020). The scores were interpreted on a four-point scale. A score of 3 indicated strong performance, where the output

preserved the intended meaning and was contextually appropriate. A score of 2 indicated generally adequate performance with minor weaknesses. A score of 1 indicated partial understanding or incomplete output, while a score of 0 indicated failure, serious misunderstanding, or unsupported response. Three native-speaker reviewers were consulted for each Nigerian language or variety. They rated outputs independently using the rubric and expected meaning elements. Final scores were assigned by majority vote; when ratings differed among all three reviewers, the more conservative score was used. Because the study is exploratory, it does not claim full statistical validation; these methodological limitations are addressed further in Section 8.

4. Nigerian Languages and the Problem of NLP Underrepresentation

This section examines Nigerian-language underrepresentation through low-resource status, data gaps, evidence from datasets and models, and technical, economic, and institutional constraints. Yorùbá, Igbo, Hausa, and Nigerian Pidgin are used as illustrative cases because they are widely used in Nigeria and appear in emerging African NLP research.

4.1 Low-Resource Status and Digital Data Gaps

Languages with limited text, speech, annotated corpora, translation datasets, benchmarks, and evaluation systems usually receive weaker NLP support and are often classified as low-resource (Joshi et al., 2020; Ogueji et al., 2021). In this paper, low-resource status refers to limited digitisation, weak institutional support, insufficient funding, uneven standardisation, limited evaluation, and poor integration into AI development pipelines, not to the social importance of a language. Many Nigerian languages are widely used but computationally weak. They often lack labelled datasets, large corpora, pre-trained models, speech resources, dictionaries, spell-checkers, morphological analysers, and orthographic keyboards (Adelani, 2023; Martinus & Abbott, 2019). As a result, AI systems may recognise a language in name while still performing poorly in translation, fluency, meaning preservation, cultural interpretation, or information extraction. Data quality is also a problem. Kreutzer et al. (2022) show that web-crawled multilingual datasets often contain serious problems for lower-resource languages, including unusable text, misidentified languages, and noisy corpora. For Nigerian languages, such data-quality problems can weaken AI performance on dialects, orthographies, code-switching, public communication, and culturally specific meanings.

4.2 Dataset and Model Evidence from Yorùbá, Igbo, Hausa, and Nigerian Pidgin

The gap between the speaker population and digital visibility is evident in the available evidence. Nekoto et al. (2020) show that English had over six million Wikipedia articles, while Yorùbá had 32,572 and Igbo had only 1,487, despite both having millions of speakers. This gap is important because the selected Nigerian languages and language varieties have large speaker communities: Hausa is estimated at 80–100 million speakers, with some HausaNLP research giving higher estimates (Adjovi et al., 2026; Muhammad et al., 2025); Yorùbá has roughly 47 million speakers (Ahia et al., 2024); Igbo is described in NLP research as having more than 40–50 million speakers globally, depending on the source used (Ezeani et al., 2020; Ekle & Das, 2025); and Nigerian Pidgin is estimated at 40 million first-language speakers and 80 million second-language speakers (Muhammad et al., 2022). These figures show that the problem is not a lack of social use or speaker population, but rather limited machine-readable content and weak AI resource development.

African NLP initiatives show progress, but mostly in selected tasks. MasakhaNER provides named entity recognition datasets for African languages, including Hausa, Igbo, Nigerian Pidgin, and Yorùbá (Adelani et al., 2021). NaijaSenti provides sentiment analysis data for Hausa, Igbo, Yorùbá, and Nigerian Pidgin (Muhammad et al., 2022). AfriBERTa shows that useful multilingual language

models can be trained with low-resource African languages and relatively small datasets (Ogueji et al., 2021). Masakhane also demonstrates the value of participatory African-language machine translation research (Nekoto et al., 2020). Overall, the evidence shows growing but uneven support. Yorùbá, Igbo, Hausa, and Nigerian Pidgin appear in some NLP resources, yet still lack strong support across speech recognition, question answering, summarisation, educational AI, civic information, public safety communication, and health communication.

4.3 Technical, Economic, and Institutional Constraints

Nigerian-language underrepresentation is shaped by technical constraints such as dialect variation, tonal complexity, inconsistent orthographies, limited corpora, code-switching, and scarce annotated data. Nigerian Pidgin adds further complexity because it is flexible, informal, mixed with English and indigenous languages, and less institutionally standardised.

Economic and computational constraints also matter. Large-scale NLP development requires funding, skilled labour, data collection, annotation, storage, evaluation, and computing resources. Global technology companies often prioritise languages with larger commercial markets, stronger digital footprints, and clearer monetisation opportunities, which reinforces the wider concentration of NLP research and model development around high-resource languages (Joshi et al., 2020; Blasi et al., 2022). Leong et al. (2023) describe African NLP as facing a "low-resource double-bind," where data scarcity is combined with limited access to computing resources. These constraints are also institutional and political. Policy choices, research funding, colonial language hierarchies, market incentives, and the priorities of global technology companies shape them. Addressing Nigerian-language underrepresentation therefore requires sustained investment in language resources, community-led research, low-compute methods, evaluation systems, and AI governance frameworks that treat Nigerian languages as part of digital infrastructure.

5. Evidence of Language Inequality in AI Systems

Language inequality in AI is evident at both the model and user-facing levels. This section therefore examines representation in multilingual models and the results of the exploratory assessment of AI tools.

5.1 Representation in Multilingual AI Models

Uneven representation in multilingual AI models offers One way to understand language inequality in AI. Although many systems are described as multilingual, language coverage does not necessarily mean comparable levels of support across languages. A model may include many languages while still relying far more heavily on data from high-resource languages such as English than on data from lower-resource African languages. Studies on linguistic diversity in NLP show that global NLP development remains concentrated around a relatively small group of high-resource languages, while many African languages receive less data coverage, weaker model support, and limited evaluation attention (Joshi et al., 2020; Blasi et al., 2022). Evidence from multilingual language models illustrates this imbalance. Ogueji et al. (2021) show that mBERT included only three African languages out of 104, while XLM-R included only eight African languages out of 100. They also show that African languages accounted for only 4.80 GB of XLM-R's training data, representing about 0.2% of the dataset. Such limited representation matters because the volume and quality of training data can influence how well models learn vocabulary, syntax, orthography, dialectal variation, and contextual meaning. For Nigerian languages such as Yorùbá, Igbo, Hausa, and Nigerian Pidgin, this may affect performance in translation, information extraction, speech recognition, civic communication, educational support, and cultural-context interpretation. Recent

African-language benchmarks show progress while also indicating the need for broader evaluation. AfriSenti provides a sentiment analysis benchmark for 14 African languages, including Hausa, Igbo, Nigerian Pidgin, and Yorùbá (Muhammad et al., 2023). IrokoBench extends African-language evaluation into natural language inference, mathematical reasoning, and knowledge-based question answering (Adelani et al., 2025), while AfroBench evaluates large language models across multiple African languages, tasks, and datasets (Ojo et al., 2025). These studies suggest that African-language evaluation is expanding, but that broader coverage is still needed across tasks, domains, and public-facing AI applications.

In this paper, model-level evidence is therefore treated as part of a broader pattern encompassing dataset quality, benchmark coverage, and everyday AI tool performance.

5.2 Results of the Exploratory AI-Tool Assessment

The second form of evidence comes from the exploratory assessment of the AI tool. It examined how ChatGPT, Google Gemini, and Microsoft 365 Copilot Chat responded to English, Yorùbá, Igbo, Hausa, Nigerian Pidgin, and mixed Nigerian Pidgin/English tasks linked to public communication, information extraction, civic understanding, code-switching, cultural context, and informal register. The prompts used for the assessment are listed in Appendix A, while Table 1 presents the condensed scoring results.

Table 1: Summary of AI Performance Across English and Selected Nigerian-Language Tasks

Test item	Language	AI tool	Error type	Score
Election notice	English	ChatGPT	None observed	3
Election notice	Yorùbá	ChatGPT	Minor ambiguity	2
Election notice	Yorùbá	Gemini	Minor fluency/context issue	2
Election notice	Yorùbá	Microsoft Copilot	Fluency issue	1
Public health warning	English	ChatGPT	None observed	3
Public health warning	Hausa	ChatGPT	Formal wording	2
Public health warning	Hausa	Gemini	Minor register issue	2
Public health warning	Hausa	Microsoft Copilot	Meaning loss	1
Security advisory	English	ChatGPT	None observed	3
Security advisory	Igbo	ChatGPT	Clarity issue	2
Security advisory	Igbo	Gemini	Minor information/clarity issue	2
Security advisory	Igbo	Microsoft Copilot	Information omission	1
Code-switched election message	Nigerian Pidgin/English	ChatGPT	Register loss	2
Code-switched election message	Nigerian Pidgin/English	Gemini	Minor tone/register loss	2
Code-switched election message	Nigerian Pidgin/English	Microsoft Copilot	Context loss	1
Cultural expression	Igbo	ChatGPT	Cultural-context loss	1
Cultural expression	Igbo	Gemini	Minor cultural-context loss	2
Cultural expression	Igbo	Microsoft Copilot	Cultural loss	1
Informal greeting	Nigerian Pidgin	ChatGPT	Register issue	2
Informal greeting	Nigerian Pidgin	Gemini	Minor register issue	2
Informal greeting	Nigerian Pidgin	Microsoft Copilot	Context loss	1

Scoring scale: *3 = accurate and contextually appropriate; 2 = mostly accurate with minor issues; 1 = partly inaccurate or incomplete; 0 = failed or unsupported.*

The results show that the English baseline examples received the strongest scores because the outputs were stable, complete, and contextually clear. By contrast, the selected Nigerian-language and mixed-language examples showed greater variation, particularly in ambiguity, register loss, information omission, cultural context loss, and reduced clarity in public communication. Gemini and ChatGPT generally preserved the main meaning, although both still showed minor issues in tone, register, or cultural nuance. Microsoft Copilot produced the weakest results in this sample,

especially where the task required cultural interpretation, local expression, or full extraction of public-safety information. These patterns support the paper's broader argument. Nigerian-language underrepresentation in AI is not only evident in datasets and models but also in everyday AI outputs. The assessment suggests that linguistic inclusion should be evaluated through real-world usability, including meaning preservation, information completeness, fluency, register, cultural interpretation, and public usefulness.

6. Discussion: From Language Underrepresentation to Digital Exclusion

The evidence presented in this paper shows that Nigerian-language underrepresentation in AI operates across connected levels of data visibility, model representation, user-facing performance, and development access. At the data level, many Nigerian languages have limited machine-readable resources, annotated datasets, speech corpora, and evaluation benchmarks. At the model level, multilingual systems may include some African languages but still provide stronger data coverage and performance for English and other high-resource languages. At the user level, the AI-tool assessment illustrates how these inequalities can appear in public-facing tools.

The central discussion, therefore, is not simply that Nigerian languages are underrepresented in AI. That concern is already well established in African NLP research. The contribution of this paper is to connect that technical underrepresentation to digital inclusion. The findings suggest that language inequality in AI can affect who benefits from AI-enabled information, Education, civic communication, public services, and economic opportunity. However, this relationship should not be understood as a simple direct cause. Language is One dimension of digital exclusion, interacting with other factors such as poverty, literacy, internet access, infrastructure, digital skills, institutional responsiveness, and computing resources.

6.1 Language Data Inequality and AI Performance

The evidence in the previous sections shows that language data inequality affects not only whether Nigerian languages are included in AI systems but also how reliably those systems work for real users. Limited data, poor data quality, uneven model representation, and weak evaluation coverage create conditions in which an AI system may recognise a language but still be poorly supported in practice (Joshi et al., 2020; Kreutzer et al., 2022; Ogueji et al., 2021). This distinction is important because surface-level inclusion does not guarantee meaningful usability. A model may identify Yorùbá, Igbo, Hausa, or Nigerian Pidgin yet still struggle with register, dialect variation, tone, code-switching, informal expression, cultural meaning, or public service communication.

The AI-tool assessment in this paper reflects this problem: the weakest outputs were not always total failures, but partial responses that lost tone, omitted relevant information, or reduced culturally specific meaning. Language data inequality therefore produces uneven AI performance in practical ways, making outputs less accurate, less fluent, less locally meaningful, and less useful for ordinary users. For Nigerian languages, the issue is not simply whether more data exists, but whether AI systems are trained, evaluated, and improved in ways that reflect how these languages are actually used in public, educational, civic, and everyday communication (Nekoto et al., 2020; Adelani et al., 2021; Muhammad et al., 2022).

6.2 Digital Inclusion Implications

The second implication is that language inequality in AI can contribute to digital exclusion. Digital inclusion is often measured by internet connectivity, mobile phone access, affordability, and digital skills. These are important, but they do not fully capture the realities of multilingual societies. A person may have internet access and a smartphone but still face exclusion if AI-enabled systems

operate mainly in English or perform poorly in the language they understand best. This concern is especially important because, as shown in Section 4.2, the selected Nigerian languages and language varieties are used by very large communities. Hausa, Yorùbá, Igbo, and Nigerian Pidgin are not marginal in terms of speaker populations. When languages used by tens of millions of people are only weakly supported in AI systems, the result is not simply poor model performance; it is a large-scale inclusion problem that affects who can access, understand, and benefit from AI-enabled services.

Recent digital inclusion data for Nigeria strengthens this point. Kemp's Digital 2026: Nigeria report for DataReportal estimated that Nigeria had 109 million internet users at the end of 2025, representing an internet penetration rate of 45.5%, while about 130 million people, or 54.5% of the population, remained offline (Kemp, 2025). These figures show that gaps in access and connectivity already constrain digital inclusion in Nigeria. Language inequality adds another layer, as even online users may not benefit equally if public-facing AI systems perform better in English than in Nigerian languages and language varieties.

These implications are visible in public communication, Education, civic participation, and economic activity. Health warnings, election notices, security advisories, agricultural information, and government announcements require clear, trustworthy language. Learners, small business owners, farmers, traders, and citizens may also receive weaker support if AI tools work better in English than in Nigerian languages or Nigerian Pidgin. The exploratory AI-tool assessment showed that even when Nigerian-language outputs preserved the main meaning, they sometimes lost clarity, tone, urgency, or culturally relevant meaning.

The paper does not claim that language inequality alone explains digital exclusion. Infrastructure, income, literacy, gender, disability, access to electricity and devices, and institutional trust also shape digital participation. The argument is more specific: linguistic accessibility is a necessary but often neglected part of digital inclusion. Without it, AI systems may expand technologically while remaining socially unequal in their benefits.

6.3 AI Governance and Policy Implications

The third implication is that the inclusion of Nigerian languages should be treated as an AI governance issue. AI governance is not only about data protection, privacy, algorithmic accountability, or safety. It also involves decisions about which languages are prioritised, whose data is collected, who evaluates AI systems, and which communities benefit from technological development. If Nigerian languages are weakly represented in these decisions, AI systems may reproduce existing linguistic inequalities.

This governance issue is partly about data. Nigerian-language data should not be treated merely as raw material for model training. It carries social meaning, cultural knowledge, community identity, and public value. Therefore, questions of consent, documentation, ownership, dialect diversity, and community benefit matter. Bender and Friedman's (2018) argument for data statements is relevant here, as language datasets need clear information on source, domain, language variety, speaker community, and limitations.

It is also a question of participation. African-language NLP research has shown the importance of involving speakers, translators, curators, evaluators, and local language experts in dataset Creation and evaluation (Nekoto et al., 2020). This is important because cultural meaning, tone, idioms, code-switching, and dialect variation cannot be fully captured by external developers alone. Without local

participation, AI systems may recognise Nigerian languages at a surface level while still misrepresenting how they are used in practice.

Finally, language inclusion raises questions about institutional responsibility. Public-facing AI systems may claim multilingual support while performing unevenly across languages. Governments, universities, technology companies, and research funders therefore have a role in ensuring that language inclusion is evaluated, documented, and supported. This is especially important in Nigeria, where English remains dominant in official and digital systems despite the everyday importance of Yorùbá, Igbo, Hausa, Nigerian Pidgin, and other local languages.

In summary, the underrepresentation of Nigerian languages in AI is not only a technical limitation. It is also a governance concern involving data, participation, accountability, and public benefit. The next section builds on this discussion by outlining practical pathways toward linguistically inclusive AI in Nigeria.

7. Toward Linguistically Inclusive AI in Nigeria

Achieving linguistically inclusive AI in Nigeria requires practical action across dataset development, community participation, policy, funding, and regional cooperation.

7.1 African-Language Dataset Development

The priority is the deliberate development of datasets for Nigerian languages and language varieties. This should include text corpora, speech datasets, dictionaries, annotated datasets, translation data, benchmarks, and evaluation resources across domains such as public health, Education, agriculture, civic information, public safety, market communication, and cultural expression. These datasets should be clearly documented with respect to source, domain, language variety, dialect, orthography, collection method, and limitations, especially since Nigerian languages often involve dialect variation, tonal features, diacritics, code-switching, and spelling variation (Bender & Friedman, 2018; Kreutzer et al., 2022). Dataset development should also address consent, attribution, community benefit, and protection against extractive data practices.

7.2 Community-Led and Participatory NLP Research

The second priority is community-led NLP research. Native speakers, linguists, translators, educators, civic organisations, and local technologists should be involved in dataset development, model evaluation, and AI governance. Participatory African NLP initiatives, such as Masakhane, demonstrate the value of collaboration among speakers, translators, curators, researchers, and evaluators (Nekoto et al., 2020). This is important because automatic metrics may not fully capture cultural meaning, register, tone, public usefulness, and local appropriateness. Community participation can therefore improve technical accuracy, social relevance, and protection against extractive AI practices (Bird, 2020).

7.3 Policy, Funding, and Regional Collaboration

The third priority is stronger policy, funding, and regional collaboration. Linguistically inclusive AI cannot depend only on volunteers, individual researchers, or short-term projects. Nigeria's digital transformation and AI policy frameworks should treat language inclusion as part of digital infrastructure by evaluating public-facing AI tools, e-government platforms, civic technologies, educational technologies, and public information services for linguistic accessibility. Funding is also needed for data collection, annotation, translation, speech recording, model development, evaluation, storage, computing infrastructure, training programmes, and low-compute methods (Leong et al., 2023). Regional collaboration among Nigerian institutions, ECOWAS, the African

Union, universities, civic technology networks, and African NLP communities can support shared benchmarks, interoperable datasets, translation resources, and evaluation standards.

8. Limitations and Future Research

This study has some limitations. First, it focuses on Yorùbá, Igbo, Hausa, and Nigerian Pidgin as illustrative cases. These languages are widely used and appear in emerging African NLP research, but they do not represent Nigeria's full linguistic diversity. Future studies should include more Nigerian languages, especially minority and less institutionally supported languages.

Second, the exploratory AI tool assessment used a small set of prompts and selected public-facing AI tools. It was intended to illustrate how uneven language support may appear in everyday AI use, not to produce a statistically generalisable benchmark. Future research should use larger prompt sets, more task categories, repeated testing across time, and wider platform coverage.

Third, the study used checklist-based scoring, qualitative error analysis, and native-speaker review rather than formal inter-rater reliability testing or broader user-based evaluation. Future studies should involve larger groups of evaluators, calculate inter-rater reliability, and combine human evaluation with task-specific NLP metrics when reference outputs are available.

Fourth, the main scoring table was based on typed prompt entries, while voice-input trials were treated only as supplementary observations. Future research should examine voice-based Nigerian-language AI interactions more directly, including speech recognition, accent interpretation, pronunciation variation, and the effects of background noise.

Finally, the paper relies partly on secondary evidence from African NLP research, policy literature, and digital inclusion reports. Future research should include interviews, surveys, or user-based studies with users, developers, educators, civic technology organisations, and public communication officers to examine how language inequality in AI is experienced in real educational, civic, health, and economic settings.

9. Conclusion

This paper has examined the underrepresentation of Nigerian languages in artificial intelligence and its implications for digital inclusion. Drawing on conceptual analysis, secondary evidence from African NLP research, and a small exploratory assessment of an AI tool, it suggests that language inequality in AI operates across data availability, model representation, and user-facing performance. The analysis indicates that language inclusion should be assessed not only by whether a language appears in an AI system or dataset, but also by how well that language is supported in real-world communication tasks such as information extraction, public communication, cultural interpretation, code-switching, Education, and civic participation.

The findings suggest that Yorùbá, Igbo, Hausa, Nigerian Pidgin, and other Nigerian language varieties remain unevenly supported in AI systems. Although these languages appear in important African NLP resources, their presence does not necessarily translate into strong support across public-facing uses. The exploratory AI-tool assessment indicated that English responses were generally more stable and complete than the selected Nigerian-language and mixed-language outputs, which showed more variation in meaning preservation, register, cultural context, information completeness, and public usefulness.

The paper contributes by narrowing the broader discussion of African-language underrepresentation to selected Nigerian cases, linking technical underrepresentation in NLP to digital inclusion, and proposing a usability-oriented understanding of linguistic inclusion. It also suggests that Nigerian-language underrepresentation is shaped by both technical constraints and governance factors,

including dialect variation, tonal complexity, code-switching, limited corpora, annotation costs, infrastructure, access to compute, policy choices, funding priorities, market incentives, and the dominance of English in official and digital systems.

The evidence presented here does not support broad statistical claims about all Nigerian languages or all AI systems. Rather, it provides exploratory evidence of how language inequality can appear in model representation, dataset availability, and everyday AI outputs. Overall, the paper suggests that linguistically inclusive AI in Nigeria requires stronger datasets, community-led evaluation, policy support, sustainable funding, and regional collaboration. Nigerian languages should therefore be treated not only as cultural heritage but also as part of digital infrastructure.

References

Adelani, D. I. (2023). *Natural language processing for African languages* [Doctoral dissertation, Saarland University]. https://universaar.unisaarland.de/bitstream/20.500.11880/36297/1/DavidAdelani_Thesis_10_08_2023.pdf

Adelani, D. I., Abbott, J., Neubig, G., D'Souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., ... Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics, 9*, 1116–1131. https://doi.org/10.1162/tacl_a_00416

Adelani, D. I., Ojo, J., Azime, I. A., Zhuang, J. Y., Alabi, J. O., He, X., Ochieng, M., Hooker, S., Bukula, A., Lee, E.-S. A., Chukwuneke, C. I., Buzaaba, H., Sibanda, B. K., Kalipe, G. K., Mukiibi, J., Kabongo Kabenamualu, S., Yuehgoh, F., Setaka, M., Ndolela, L., ... Stenetorp, P. (2025). IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2732–2757). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.139/

Adjovi, M. P., Olufemi, V., Eiselen, R., & Mitra, P. (2026). A survey of text and speech resources for Hausa and Fongbe: Availability, quality, and gaps for NLP development. *arXiv*. https://doi.org/10.48550/arXiv.2605.22828

Ahia, O., Aremu, A., Abagyan, D., Gonen, H., Adelani, D. I., Abolade, D., Smith, N. A., & Tsvetkov, Y. (2024). Voices unheard: NLP resources and models for Yorùbá regional dialects. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4392–4409). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.251

Bamgbose, A. (1991). *Language and the nation: The language question in Sub-Saharan Africa*. Edinburgh University Press for the International African Institute.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). International

Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5486–5505). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Ekle, O. A., & Das, B. (2025). Low-resource neural machine translation using recurrent neural networks and transfer learning: A case study on English-to-Igbo. *arXiv*. https://doi.org/10.48550/arXiv.2504.17252

Ezeani, I., Rayson, P., Onyenwe, I., Uchechukwu, C., & Hepple, M. (2020). Igbo-English machine translation: An evaluation benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2004.00648

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Kemp, S. (2025, November 8). *Digital 2026: Nigeria*. DataReportal. https://datareportal.com/reports/digital-2026-nigeria

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Ortiz Suárez, P., ... Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics, 10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Leong, C., Shandilya, H., Dossou, B. F. P., Tonja, A. L., Mathew, J., Omotayo, A.-H., Yousuf, O., Akinjobi, Z., Emezue, C. C., Muhammad, S., Kolawole, S., Choi, Y., & Adewumi, T. (2023). Adapting to the low-resource double-bind: Investigating low-compute methods on low-resource African languages. *arXiv*. https://arxiv.org/abs/2303.16985

Martinus, L., & Abbott, J. Z. (2019). A focus on neural machine translation for African languages. *arXiv*. https://arxiv.org/abs/1906.05685

Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jorge, A., & Brazdil, P. (2022). NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 590–602). European Language Resources Association. https://aclanthology.org/2022.lrec-1.63/

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Ousidhoum, N., Adelani, D. I., Yimam, S. M., Ahmad, I. S., Beloucif, M., Mohammad, S. M., Ruder, S., Hourrane, O., Brazdil, P., Jorge, A., Ali, F. D. M. A., David, D., Osei, S., Bello, B. S., Ibrahim, F., Gwadabe, T., ... Arthur, S. (2023). AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13968–13981). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.862

Muhammad, S. H., Ahmad, I. S., Abdulmumin, I., Lawan, F. I., Sani, B., Imam, S. H., Aliyu, Y., Sani, S. A., Umar, A. U., Church, K., & Marivate, V. (2025). HausaNLP: Current status,

challenges and future directions for Hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP, 2025)* (pp. 176–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.africanlp-1.27

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., ... Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144–2160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., ... Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *arXiv*. https://arxiv.org/abs/2207.04672

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resource languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 116–126). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Ojo, J., Ogundepo, O., Oladipo, A., Ogueji, K., Lin, J., Stenetorp, P., & Adelani, D. I. (2025). AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 19048–19095). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.976

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Warschauer, M. (2003). *Technology and social inclusion: Rethinking the digital divide*. MIT Press.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr