Metadata Analysis of the Generative AI Usefulness for African Languages

Yohanna Joseph Waliya * and Margaret Mary Okon

Posted Date: 6 April 2026

doi: 10.20944/preprints202604.0360.v1

Keywords: natural language processing (NLP); large language models (LLMs); Cheetah; Ibibio; Margi; Nigeria; Africa

 doi:10.20944/preprints202604.0360.v1

Article

Metadata Analysis of the Generative AI Usefulness for African Languages

Yohanna Joseph Waliya [1,*] and Margaret Mary Okon [2]

[1] The Nigeria French Language Village, Badagry, Lagos, Nigeria

[2] University of Calabar, Calabar, Nigeria

* Correspondence: yohannawaliya@frenchvillage.edu.ng

Abstract

In the contemporary landscape, natural language processing (NLP) stands as a vital force, empowering computers to comprehend and engage with human languages, thereby enhancing the realm of human-computer interaction (HCI) through the utilisation of large language models (LLMs) and multilingual pre-trained language models (mPLMs). The widespread adoption of these LLMs on a global scale is obvious. However, a critical observation reveals a significant gap in their capacity to effectively recognize some low-resource African languages, a concern observed by numerous researchers. This paper endeavours to contribute to the discourse by conducting a comprehensive metadata analysis of existing African language models. Through this investigation, the aim is to outline the importance, strengths, and weaknesses inherent in these models. By shedding light on these aspects, the paper seeks to not only underscore the current limitations but also to provide valuable insights and recommendations for future research endeavours in the domain of language recognition, particularly focusing on African languages. In doing so, the paper aspires to catalyse advancements that promote inclusivity and a more nuanced understanding of linguistic diversity within the realm of natural language processing. Multilingual Testing shall be used on Cheetah to evaluate the model's proficiency strength in multiple languages, including those that are less widely spoken such as Margi and Ibibio as well as identify any language-specific weaknesses or limitations of the LLMs, especially in recognizing and understanding languages like Margi spoken in the North-East geo-political zone of Nigeria and Ibibio spoken in the South-South geo-political zone of Nigeria.

Keywords: natural language processing (NLP); large language models (LLMs); Cheetah; Ibibio; Margi; Nigeria; Africa

1. Introduction

From time immemorial in Ancient Egypt, Mesopotamia, Maya, and Chinese civilisations, homo sapiens have been interacting with machines though analogues to leverage their usefulness to lessen their daily duties (Angelakıs et al., 2020), and giving the names of the machines in their languages. That interaction was never symbiotic. It was just like a slave-and-master relationship where the machines just obeyed instructions without interactive feedback. However, with the advent of digital computers, human-computer interaction (HCI) has become animated and synthesised through the intervention of natural language processing (Nash, 2024). This computational turn has led to the building of linguistic models such as Transformers and large language models (LLMs) which feed themselves from the Big Data through the help of the Internet of Things (IoT) to empower linguistic interactivity for possible automatic translation, distant writing, distant reading, sentiment analysis, cultural analytics, stylometry, Generative Artificial Intelligence creativity, etc. (Alabi et al., 2022; Waliya, 2022; Nash, 2024). These models are Anglo-centric inclined inventions (L. Xue et al., 2021). They recognised almost all African languages as low-resource except 41 African languages- Afrikaans, Akan/Twi, Amharic, Bambara, Basaa, Bemba, Chichewa, Chitumbuka, Éwé, Fon, Gahuza, Gikuyu, Ghomala, Hausa, Igbo, Kinyarwanda, Kirundi, Kikuyu, Lingala, Luganda, Luo, Malagasy,

 doi:10.20944/preprints202604.0360.v1

Mooré, Mossi, Naija, Ndebele, Northern Sotho, Oromo(Afan), Shona, Sesotho, Somali, Swahili, the Sepedi, Tigrinya, Tshivenda, Tswana/Setswana, Wolof, Xhosa, Xitsonga, Yoruba, isiZulu—AfriBERTa, AfriByT5, AfriMT5, AfriMBART, AfriVeTa, AfroLM, Afro-XMLR, Bloom, MARGE, M2M-100, mBERT, mT5, SERENGETI, XMLR(Adebara et al., 2023, 2024; Alabi et al., 2022b, 2022a; Awobade et al., 2024; Diandaru et al., 2024; Dossou et al., 2022; Homskiy & Maloyan, 2023; Ogueji et al., 2021; Scao et al., 2022; University of Waterloo, 2021; B. Xue et al., 2024; L. Xue et al., 2021). Compared to more than 2, 000 living languages in Africa, 41 languages are still very insignificant for even Nigeria, Cameroon and the Democratic Republic of Congo alone have 530, 277, and 214 living languages respectively (Statista Research Department, 2024, Eberhard et al. 2025) summing up to 1,021 languages used daily in just only three countries out of 54 African countries. Eberhard et al., (2021) precisely counted African languages to be 2,144 languages. Among the 41 languages mentioned above only four are used in Nigeria. African languages are not only low-resource languages but also non-resource languages due to the lack of data online and scholarly attention from computational linguistic scholars (Adebara & Abdul-Mageed, 2022; Alabi et al., 2022b; Ogueji et al., 2021; University of Waterloo, 2021). Fortunately, in recent years, the field of natural language processing (NLP) has been closing this gap and making substantial advancements with the development of pre-trained multilingual models adapted to low-resource languages, including numerous African languages through the intervention of Mozilla Common Voice language (cf. Ngué Um E et al., 2025). This development is a deliberate attempt to rescue African languages from being suffocated by English in the digital linguistic ecosystem (Okon & Noah, 2021).Therefore, there is hope for Ibibio and Margi which are also low-resource languages spoken mostly in West Africa and Central Africa, Nigeria, Cameroon, Ghana, Equatorial Guinea, and Chad.

Margi (alternate orthography Marghi) is ethnoglossonym—the name by which both the language and the people are called. It is a tone language and belongs to the Afro-Asiatic Chadic family of languages spoken by more than 509, 000 people in the Southeast Borno and the Northwest Adamawa of Nigeria (Birdling, 2009, 2013; Modu & Jawur, 2021; Joshua Project, 2024), although this very recent statistics is only for Margi Babal (Margi Central) and Margi Ti-duntum (Margi South) not including the other six dialect groups such as Margi Udzurngu, Margi Putai, Margi Gwara, Margi Wurga, and Margi Wandala in Cameroon which are larger in number vis-à-vis the two mentioned (Adzu, 2014). Margi people are very tolerant of their worldview and culture (Vaughan Jr, 2000) because of their strong traditional justice system under the leadership of the Ptil (Adzu, 2014). To this fact, a Nigerian Fulani traveller who had lived among the Kwang people of South-eastern Chad and many other tribes in Sahara Desert, in the course of his expedition from Nigeria to Ethiopia, (S. Muhammad personal communication, 2020) argues that Kwang is also a Margi dialect because they share the same language and culture with Margi people in Nigeria and Cameroon. Kwang lives not far from the District of Mokolo formerly known as the Margi-Wandala District of Colonial Cameroon (Archives Nationales d'outre-mer, 2017). Consequently, Margi can be tagged as an international minority language with very minimal digital standardised corpora.

As for Ibibio, it is also ethnoglossonym, that is, people and their language are called by that name. Ibibio is a language used in four African countries-Nigeria, Ghana, Cameroon, and Equatorial Guinea. Though Akwa Ibom State is their major home, they spread to Cross River and Eastern Nigeria, precisely Abia State. They constitute the fourth largest ethnic group in Nigeria with a growing population of almost 7 million (Joshua Project, 2024a). They were originally known as Afaha people who migrated from the village of Usak Edet in Cameroon and settled in their present abode around 7000BC (Noah, 1988, p. 5) According to Garry et al. (2001) Ibibio is one of the 200 major languages of the world. Both Ibibio and Margi constitute the active languages of about 8 million people in Africa which is why our study is indispensable. However, both of them suffer divergent treatment within NLP systems.

 doi:10.20944/preprints202604.0360.v1

Objective of the Study

This metadata analysis evaluates 15 prominent models— AfriBERTa, AfriByT5, AfriMT5, AfriMBART, AfriVeTa, AfroLM, Afro-XMLR, Bloom, Cheetah, MARGE, M2M-100, mBERT, mT5, SERENGETI, and XMLR. It likewise focuses on their design, language coverage, strengths, and weaknesses to identify the most effective model for enhancing NLP tasks in African languages.

2. Methodology

This metadata analysis integrates findings from 15 multilingual pre-trained language models from 2018 to 2024 sourcing from conference papers, published articles, and preprints to provide a comprehensive comparison of the capabilities and limitations of each model as established by experts considering language documentation coverage, linguistic architectural design and evaluation scope. We employed a Digital Humanities methodology called network graph visualisation to visualise the interactions between language models and African languages they cover over the years particularly on high-resource African languages vis-à-vis the low-resources in a comparative approach. This helps us to identify central and peripheral languages, revealing patterns of concentrations and marginalisation. In other words, DH methodology is adopted to foreground metadata as a site of critical inquiry. Rather than evaluating language models solely through performance metrics, we examine the fundamental situation in which African languages input into or fail to prompt generative AI systems. In the end, to buttress the metadata analysis, we conducted prompt-based qualitative testing of the Cheetah model on Ibibio and Margi across tasks such as language identification, translation, paraphrasing and text continuation. Prompts were culturally grounded and manually validated to avoid English dominance bias (Miller & Òkôn, 2026).

2.1. Limitation

This study is limited to multilingual pre-trained language models fine-tuned for African languages that cover more than three languages. It means we will not discuss mBART and KinyaBERT(Adebara et al., 2024; Dossou et al., 2022) which cover one to three African languages. More than 400+ new languages used to train SERENGETI and Cheetah models will not be discussed due to their newness and because they never supported NLP tasks before.

2.2. Models Overview

In recent times, African Natural Language Processing received much attention from computational African linguistic scholars in the diaspora who are savvy in computational linguistics and machine learning. Nevertheless, they are confronted with a lack of textual data (Adebara et al., 2023; Adebara & Abdul-Mageed, 2022; Dossou et al., 2022) and graphics processing unit and its ilk-dGPU, iGPU, TPU computes for speedy compilation of codes (Alabi et al., 2022b; Dossou et al., 2022) then coupled with building standard Unicode Transfer Format configuration for glyphs like Amharic, Tigrinya diacritics. The first model to be optimized for African languages from scratch was AfriBERTa in 2021. It improved performance in 11 low-resource languages not with Southern African languages through careful fine-tuning and emphasising the importance of multilingual data for model robustness. It likewise outperforms larger models like mBERT and XLM-R by up to 10 F1 accuracy matric points in terms of text classification (Ogueji et al., 2021). In the same vein, SERENGETI was released in 2023. It is a robust African multilingual pre-trained language model which claims to cover 517 African languages by improving natural language understanding, and generation capabilities. It addresses the challenges of limited resources and data scarcity in African language processing. SERENGETI is declared as the first robust African multilingual model that developed NLP tasks for more than 400 languages from scratch to outperform AfriBERTta, XMLR, mBERT, and BERT (Adebara, Adelani, & Alabi, 2022). A competitive model to SERENGETI by the name of Cheetah was developed by the same Adebara and his team at the University of British Columbia. Cheetah outperforms mBART, mBERT, mT5, MT0, AfriByT5, AfriMT5, AfriMBART,

 doi:10.20944/preprints202604.0360.v1

AfriTeVa, XLM-R, AfriBERTa, AfroLM Afro-XLM-R KINYaBERT, and SERENGETI while tested on the same datasets on which each model is trained upon (Adebara et al., 2023, 2024; Adebara & Abdul-Mageed, 2022; Dossou et al., 2022; B. Xue et al., 2024; L. Xue et al., 2021)

2.3. Metadata Analysis of Multilingual Afrocentric Models from 2018 to 2024

SERENGETI and Cheetah are models developed in 2023 and 2024 respectively and trained on 517 African languages and language varieties across 14 language families established in 50 of 54 African countries and are written in six different scripts for text summarisation, title generation, paraphrase, machine translation, question answering and cloze (Adebara et al., 2023, 2024). However, SERENGETI is tested on 26 languages though trained in 517 languages whereas Cheetah was tested on 88 languages comparing it to the other models existing before its training (Adebara et al., 2024). In the network graph below (Figure 1), 15 multilingual pre-trained language models cover 41 languages. Multitask fine-tuned mBERT was the first general multilingual representation model to have matching 9 African languages in the NLP task in 2018 (Devlin et al., 2019; Groeneveld et al., 2024). In 2021, this similar variant of BERT was optimised as AfriBERTa for 11 languages whilst XMLR, AfriMBART, and M2M-100 cover 8 languages each. Next, the multilingual Afrocentric models of the text-to-text transfer transformer model-mT5 tasks in 13 African languages whereas AfriByT5, and AfriMT5 support 16 languages each not 17 languages as reported by Adebara et al., (2024) and L. Xue et al., (2021). The specific variant of the VErsatile TrAnsformer (VeTa) model tailored for African languages also known as AfriVeTa fathoms 10 African languages whilst AfroLM, (a multilingual version of XLM-RoBERTa (XLM-R)) fine-tuned for African languages on large multilingual corpora tailored towards 23 and 17 African languages respectively feeding from the datasets containing almost similar languages. Lastly, the BigScience Open-science Open-access Multilingual model (Bloom) supports 22 African languages. Cheetah and SERENGETI have 41 and 26 African languages concurrently. Moreover, the golden nodes are 41 languages as well as 15 diverse coloured nodes—red, blue, magenta, orange, cyan, brown, pink, silver, purple, yellow, yellow-green, grey, dark-green, grey, and light blue represent African multilingual pre-trained language models connecting between themselves using green edges symbolising networked interactions between similar languages from 2018 to 2024 explicating the frequency of relationship between languages covered and the models. The languages spread at the extreme edges of the network graph are the low-resource because they only support two to five models [Tchivenda, Ghomala, Bemba, Mossi, Northern Sotho, Ndebele, Sepedi, Mooré, Gikuyu, Xitsonga, Kikuyu, Chitumbuka, Éwé, Basaa, Kirundi, Tigrinya, Lingala,] whereas those that support three 6-15 models are classified as African high-resource languages apart from French, English, Portuguese and Arabic. These languages are displayed in the centre of the graph. They are as follows: Hausa, Swahili, isiZulu, Wolof, Kinyarwanda, Igbo, Amharic, Xhosa, Chichewa, Yoruba, Shona, Naija, Oromo, Akan., Luganda, Fon, Sesotho, Tswana, Somali, Luo, Malagasy, Bambara and Afrikaans. Among them all, Swahili and Hausa are the most resourced African languages. It is the reason that they are too close to each other on the graph for both of supported by 14 and 15 mPLMs respectively.

To sum up, the network graph analysis exposes an elaborate stratification among African languages. Only a few clusters—Swahili, Hausa, Yoruba, Amharic and others centre the network to establish that they are supported by most of the models. These languages mentioned above work as algorithmic hubs profiting from the continuous inclusion across architectures and training regimes. On the other hand, many African languages are displayed at the edges of the graph especially those ones supported by 1-5 models which proved them as a low-resource language, historical neglect, limited datafication and digital infrastructural exclusion. The graph network visualisation displays the hidden patterns in technical documentation letting the systems reproduce internal hierarchies even within Africa itself.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2026 doi:10.20944/preprints202604.0360.v1

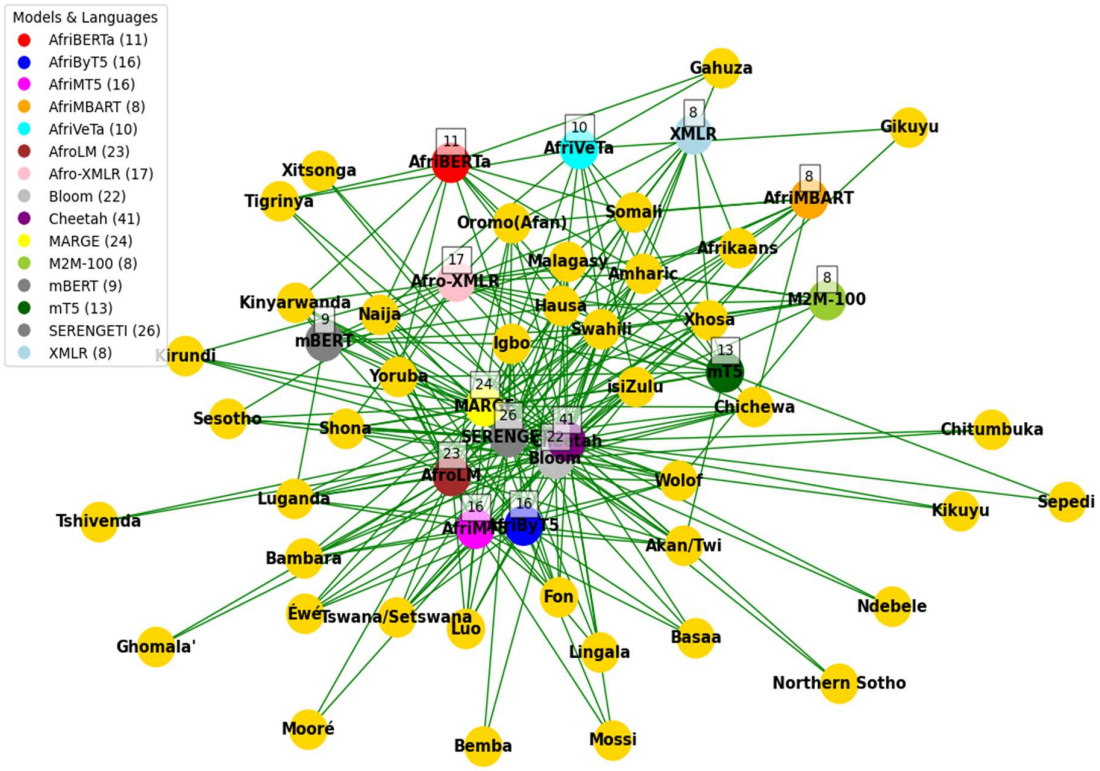

Figure 1. Network graph of African Languages and Models with the total number of languages covered by each model from 2018 to 2024.

3. Multilingual Testing Proficiency of Cheetah on Ibibio and Margi: A Big Data Analysis of Linguistic Stratification

The initial phase of this research aimed to evaluate the zero-shot cross-lingual capabilities of the Cheetah model on two low-resource Nigerian languages—Ibibio (Niger-Congo, Lower Cross subgroup) and Margi (Afro-Asiatic, Chadic family). The selection of Cheetah was predicated on its reported superior performance in African language generation and understanding (Adebara et al., 2024), positioning it as a state-of-the-art multilingual model. However, our engagement with the model revealed significant accessibility and performance barriers, prompting a methodological pivot from quantitative benchmarking to a qualitative, prompt-based evaluation. This shift was necessary to move beyond aggregate performance metrics and instead document the *lived consequences* of linguistic stratification embedded within large language models (LLMs).

This study analyses a critical research question at the intersection of big data and computational linguistics: How does the composition of pre-training data specifically, the volume, quality, and linguistic diversity of corpora translate into functional linguistic competence for African languages? Our findings reveal a critical insight: inclusion in a multilingual training dataset does not guarantee meaningful machine understanding. We demonstrate that for a language like Margi, which lacks standardised orthography and a robust digital footprint, the model's performance collapses, while for Ibibio, which has marginally more representation, the model achieves only partial, often structurally compromised, competence.

3.1. A. Prompt-Based Testing and Linguistic Tasks

To assess functional proficiency, we designed a prompt-based evaluation framework covering five core NLP tasks that are indicative of LLM utility:

1. Language Identification (LID): The model's ability to correctly classify the input language.

 doi:10.20944/preprints202604.0360.v1

2. Basic Sentence Translation (L1 → English): Assessing cross-lingual transfer and semantic preservation.
3. Text Continuation: Evaluating the model's grasp of discourse coherence and syntactic structure.
4. Paraphrase Generation: Measuring semantic flexibility and lexical diversity.
5. Question Answering (Closed-domain): Testing grounded understanding and information retrieval.

To mitigate Anglocentric bias, all prompts were manually curated and validated by native speakers. For Margi, prompts were designed to account for tonal ambiguity and dialectal variation; for Ibibio, they reflected common morphosyntactic constructions and pragmatic expressions. The evaluation was conducted using the T5-small variant of the Cheetah model.

3.2. B. Empirical Results

The results reveal a stark asymmetry in performance, which we attribute directly to disparities in pre-training data representation.

I. Ibibio: Partial Competence and Indirect Exposure

Cheetah demonstrated *moderate* competence with Ibibio. While it could often identify the language and produce approximate translations, the outputs were consistently marked by lexical borrowing and syntactic interference from English and Nigerian Pidgin.

a. Language Identification: The model failed to provide a clear identification for a complex Ibibio sentence.

b. Prompt: Usen oyop ekpo ama adat ekpo aya atop ekeka idiọñ mfuut ufọk ekarika edi ete buuñ ifia. Model Output: Identify the language: Usen oy' → *This output is fragmented and fails to produce a coherent language label.*

This suggests that the model's training data for Ibibio is likely derived from indirect sources—such as code-switched text or documents where Ibibio appears alongside English—rather than from a dedicated, high-quality Ibibio corpus. The semantic preservation in over half of the test cases, despite structural interference, indicates a form of *statistical footprint* rather than robust, generative language modelling. The model exhibits what we term sporadic semantic competence, where core meaning is partially accessible but grammatical structure is compromised.

II. Margi: Semantic Collapse and Representational Void

The situation for Margi is markedly different and points to a profound representational void in the training data. The model's performance is characterised by semantic instability, structural incoherence, and frequent misclassification.

a. Language Identification: The model consistently failed to identify Margi, often misclassifying it as the more demographically dominant Hausa or Kanuri.

b. Prompt : Abar pidar nyi mji Margi. Nyai lapya?

c. Model Output: Identify the language: 'Abar pidar ' → *The output is a nonsensical repetition of the input, indicating an inability to process the language.*

This pattern of misclassification is a critical data point. It suggests that in the training corpus, Margi sentences are either entirely absent or are erroneously labelled, leading the model to map the linguistic features of Margi onto the feature space of higher-resource Chadic or Nilo-Saharan languages present in the data. The generated outputs exhibit semantic drift, where the model produces text that is lexically and syntactically unrelated to the prompt, and structural incoherence, where the output fails to adhere to any recognisable grammatical pattern. Question-answering tasks were largely unsuccessful, confirming a lack of grounded language representation.

3.3. C. From Inclusion to Functional Usability

The performance disparity between Ibibio and Margi serves as a case study in the limitations of current approaches to multilingual NLP. The key findings are as follows:

 doi:10.20944/preprints202604.0360.v1

Language	Training Data Profile	Model Performance	Key Linguistic Artifacts
Ibibio	Indirect, code-switched, fragmented corpora	Moderate competence, partial semantic understanding	syntactic interference
Margi	Minimal or non-existent; mislabelled data	Semantic collapse; structural incoherence	semantic drift; task failure

These findings underscore a crucial insight: training coverage claims by Adebara et al. does not equate to functional usability. For Margi, the model's failure is not merely a matter of lower accuracy but a total breakdown of linguistic functionality.

This confirms that languages lacking standardised orthography, digitised corpora, or a consistent online presence are effectively excluded from the functional benefits of current LLM architectures.

The model's behaviour is not a reflection of the languages' inherent complexity but a direct consequence of the composition of the massive, often opaque, datasets used for pre-training. This research demonstrates that to achieve genuine multilingual equity, the field must move beyond the mere inclusion of languages in datasets to a focus on *representational adequacy*, ensuring that the data is of sufficient quality, volume, and structural diversity to enable robust, functionally competent models.

4. Conclusions

Finally, it is obvious that Cheetah emerges as a highly specialised model for African languages, with significant strengths in low-resource settings likewise SERENGETI offers the broadest language support, making it suitable for extensive applications, though it requires substantial computational resources, and may overly fit languages with more data. Afro-XLM-R strikes a balance with a strong performance in specific tasks but covers fewer languages. General multilingual models provide broad coverage but may lack the tailored performance for African languages seen in specialized models.

For practical applications across the diverse linguistic landscape of Africa, Cheetah and SERENGETI are single models covering 24.11% of African languages. They offer the most comprehensive language support, while AfriBERTa and Afro-XLM-R provide specialised, robust performance for the languages they support. Combining the strengths of these models could offer the most effective solution for advancing NLP in African languages.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adebara, I., & Abdul-Mageed, M. (2022). *Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go* (arXiv:2203.08351). arXiv. http://arxiv.org/abs/2203.08351
2. Adebara, I., Elmadany, A., & Abdul-Mageed, M. (2024). *Cheetah: Natural Language Generation for 517 African Languages* (arXiv:2401.01053). arXiv. http://arxiv.org/abs/2401.01053
3. Adebara, I., Elmadany, A., Abdul-Mageed, M., & Inciarte, A. A. (2023). *SERENGETI: Massively Multilingual Language Models for Africa* (arXiv:2212.10785). arXiv. http://arxiv.org/abs/2212.10785
4. Adzu, I. S. (2014). *The Margi and their culture* (First edition). Paraclete Publishers.
5. *Afriberta Large*. (2023, January 13). https://huggingface.co/castorini/afriberta_large/blob/main/README.md
6. Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022a). *Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning* (arXiv:2204.06487). arXiv. http://arxiv.org/abs/2204.06487
7. Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022b). Multilingual Language Model Adaptive Fine-tuning: A Study on African Languages. *AfricaNLP Workshop*. ICLR2022.

 doi:10.20944/preprints202604.0360.v1

8. Angelakıs, A. N., Zaccaria, D., Krasilnikoff, J., Salgot, M., Bazza, M., Roccaro, P., Jimenez, B., Kumar, A., Yinghua, W., Baba, A., Harrison, J. A., Garduno-Jimenez, A., & Fereres, E. (2020). Irrigation of World Agricultural Lands: Evolution through the Millennia. *Water*, *12*(5), 1285. https://doi.org/10.3390/w12051285
9. Archives nationales d'outre-mer. (2017). *Recherche géographique* [Web]. IREL. http://anom.archivesnationales.culture.gouv.fr/geo.php?lieu=Margui-Wandala%2C+Circonscription+%28Cameroun%29
10. Awobade, B., Oduwole, M., & Kolawole, S. (2024). What Happens When Small Is Made Smaller? Exploring the Impact of Compression on Small Data Pretrained Language Models (arXiv:2404.04759). arXiv. http://arxiv.org/abs/2404.04759
11. Birdling, E. A. (2009). Rethinking the Effects of the Foreign Missionaries' Mission to Africa, Focusing on the Church of the Brethren Missionaries Among the Margi Udzirngu in Northern Nigeria [Master]. University of Kansas.
12. Birdling, E. A. (2013). The Evolution of the Built Environment of the Margi Ethnic Group of Northeastern Nigeria [PhD Dissertation]. University of Kansas.
13. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805
14. Diandaru, R., Susanto, L., Tang, Z., Purwarianti, A., & Wijaya, D. (2024). *Could We Have Had Better Multilingual LLMs If English Was Not the Central Language?* (arXiv:2402.13917). arXiv. http://arxiv.org/abs/2402.13917
15. Dossou, B. F. P., Tonja, A. L., Yousuf, O., Osei, S., Oppong, A., Shode, I., Awoyomi, O. O., & Emezue, C. C. (2022). *AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages* (arXiv:2211.03263). arXiv. http://arxiv.org/abs/2211.03263
16. Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2021). Ethnologue: Languages of the world. *Online Version: Http://Www. Ethnologue. Com*.
17. Garry, J., Rubino, C. R. G., & Bodomo, A. B. (Eds). (2001). Facts about the world's languages: An encyclopedia of the world's major languages, past and present. H.W. Wilson Co.
18. Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., … Hajishirzi, H. (2024). *OLMo: Accelerating the Science of Language Models* (arXiv:2402.00838). arXiv. http://arxiv.org/abs/2402.00838
19. Homskiy, D., & Maloyan, N. (2023). DN at SemEval-2023 Task 12: Low-Resource Language Text Classification via Multilingual Pretrained Language Model Fine-tuning. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1537–1541. https://doi.org/10.18653/v1/2023.semeval-1.212
20. Joshua Project. (2024a). *Ibibio in Nigeria* [Web]. https://joshuaproject.net/people_groups/12171/NI
21. Joshua Project. (2024b). *Nigeria people groups, languages and religions | Joshua Project* [Web]. Joshua Project. https://joshuaproject.net/countries/NI
22. Miller, I., & Òkôn, M. M. P. (2026). 0.2.4 Case study. In D. Proctor, *Practicing Digital Ethnography* (1st edn, pp. 33–41). Routledge. https://doi.org/10.4324/9781032672663-7
23. Modu, A., & Jawur, J. I. (2021). Domains of Kanuri Loanwords in Margi. *Crossings*, *12*, 203–219.
24. Nash, B. L. (2024). Love and Learning in the Age of Algorithms: How Intimate Relationships with Artificial Intelligence May Shape Epistemology, Sociality, and Linguistic Justice. *Reading Research Quarterly*, rrq.549. https://doi.org/10.1002/rrq.549
25. Noah, M. E. (Ed.). (1988). *Proceedings of the Ibibio Union, 1928-1937*. Modern Business Press Ltd.
26. Ogueji, K., Zhu, Y., & Lin, J. (2021). Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, & G. G. Sahin (Eds), *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
27. Okon, M. M., & Noah, P. (2021). Cultural Dominance and Language Endangerment: The case of Efut in Cross River State, Nigeria. *Macrolinguistics*, *9*(14), 134–150. https://doi.org/10.26478/ja2021.9.14.8

 doi:10.20944/preprints202604.0360.v1

28. Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2022). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2211.05100
29. Statista Research Department. (2024, June 30). *Africa: Number of living languages by country 2022*. Statista. https://www.statista.com/statistics/1280625/number-of-living-languages-in-africa-by-country/
30. University of Waterloo. (2021, November 9). *New AI brings the power of natural language processing to African languages* [Web]. TechExplore. https://techxplore.com/news/2021-11-ai-power-natural-language-african.html
31. Vaughan Jr, J. H. (2000). *The Margi of the Mandaras: A Society on the Verge*. Indiana University Press. http://www.indiana.edu/~margi/
32. Waliya, Y. J. (2022). Twittérature: Lecture symétrique du Twitterbot-théâtre. In *Exploring Contemporary Digital Poetics.* (pp. 217–245). Laboratoire de Langue, Littérature, Imaginaire et Esthétique.
33. Xue, B., Wang, H., Wang, W., Wang, R., Wang, S., Liu, Z., & Wong, K.-F. (2024). *A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models* (arXiv:2402.13606). arXiv. http://arxiv.org/abs/2402.13606
34. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer* (arXiv:2010.11934). arXiv. https://doi.org/10.48550/arXiv.2010.11934