Proceedings of Machine Learning Research 302:1–27, 2025 Deep Learning Indaba 2025

The State of Large Language Models for African Languages: Progress and Challenges

Kedir Yassin Hussen KEDIRYASSIN25@YAHOO.COM
Bahir Dar University
Bahir Dar, Ethiopia

Walelign Tewabe Sewunetie
AIMS Research and Innovation Centre
Kigali, Rwanda

Abinew Ali Ayele
Bahir Dar University
Bahir Dar, Ethiopia

Sukairaj Hafiz Imam
Bayero University Kano
Kano, Nigeria

Eyob Nigussie Alemu
Addis Ababa University
Addis Ababa, Ethiopia

Shamsuddeen Hassan Muhammad
Imperial College London
London, United Kingdom

Seid Muhie Yimam
University of Hamburg
Hamburg, Germany

Abstract

Large Language Models (LLMs) are transforming Natural Language Processing (NLP), but their benefits are largely absent for Africa's 2,000 low-resource languages. This paper comparatively analyzes African language coverage across six LLMs, eight Small Language Models (SLMs), and six Specialized SLMs (SSLMs). The evaluation covers language coverage, training sets, technical limitations, script problems, and language modelling roadmaps. The work identifies 41 supported African languages and 23 available public data sets, and it shows a big gap where four languages (Amharic, Swahili, Afrikaans, and Malagasy) are always treated while there is over 98% of unsupported African languages. Moreover, the review shows that just Latin, Arabic, and Ge'ez scripts are identified while 20 active scripts are neglected. Some of the primary challenges are lack of data, tokenization biases, very high computational costs, and evaluation issues. These issues demand language standardization, corpus development by the community, and effective adaptation methods for African languages.

Keywords: Large Language Models (LLMs), Small Language Models (SLMs), Low resource languages, Specialized SLMs (SSLMs)

1. Introduction

The rapid progress of Large Language Models (LLMs) has transformed the field of Natural Language Processing (NLP). However, these advancements have primarily concentrated on high-resource languages, leaving many low-resource languages, particularly African languages, largely overlooked. Africa has over 2,000 languages (Ethnologue, 2025), the majority of which face significant challenges such as lack of data, limited computational resources, insufficient NLP tools, and the absence of standardized benchmarks. The underrepresentation of low-resource languages in LLMs has life-altering consequences: asylum seekers face deportation due to AI mistranslations of legal documents, patients are misdiagnosed by medical chatbots, perpetuating systemic exclusion (The Guardian, 2023; Delfani et al., 2024).

This study presents a three-stage review to evaluate the current status of LLMs, the challenges, and prospects for African languages. The first stage investigates both commercial and open-source LLMs models with more than 7 billion parameters regarding their support for African languages (Wang et al., 2024). The second stage examines foundational multilingual models that have significantly influenced NLP research and development. Notably, these models include BERT (Devlin et al., 2019), mBERT (Wu and Dredze, 2020), T5 (Raffel et al., 2020), mT5 (Xue et al., 2021), XLM (Lample and Conneau, 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020) and NLLB 200 (Costa-jussà et al., 2022). We refer to these models as Small Language Models (SLMs) due to their relatively smaller parameter counts compared to LLMs and their foundational role in the multilingual NLP ecosystem. These models were selected because they represent key milestones in multilingual transfer learning and remain widely used in academic and low-resource NLP research. The third stage focuses on models specifically designed or fine-tuned for African languages. These include AfriBERTa (Ogueji et al., 2021), AfriTeVa (Jude Ogundepo et al., 2022), AfroLM (Dossou et al., 2022), EthioLLM (Tonja et al., 2024b), EthioMT (Tonja et al., 2024c), and AfroXLMR (Alabi et al., 2022). We call this category 'Specialised Small Language Models' (SSLMs) because they are generally based on SLM architectures but are adapted specifically to address the unique linguistic and structural properties of African languages. These models highlight recent efforts in African-centric NLP and address longstanding representational gaps.

The following objectives guide this review: 1) Determine which African languages are most represented across LLMs, SLMs, and SSLMs, and analyze disparities in their coverage. 2) Identify which African scripts face representational challenges in LLMs and why. 3) Examine the technical limitations of representing African languages in LLMs, SLMs, and SSLMs. 4) Review benchmark datasets and models used in developing LLMs for African languages. 5) Assess the future prospects of language modeling for African languages and outline the potential roadmap. This study seeks to explore various aspects of African language representation in LLMs, aiming to shed light on the current landscape and to contribute to the ongoing discussion of creating more inclusive and equitable NLP technologies.

2. Related Work

Many initiatives have concentrated on creating foundational monolingual models specifically designed for low-resource languages. These models are frequently developed from the ground up, utilizing well-established architectures like BERT and GPT. For instance, AraBERT was designed for various Arabic dialects and employs preprocessing techniques to normalize dialectal variations into Modern Standard Arabic (Antoun et al., 2020). IndoBERT developed for the Indonesian language, adopts dynamic masking and sentence-piece tokenization to enhance linguistic representation (Koto et al., 2020). Similarly, AraGPT2 integrates Arabic-specific tokenization to improve text generation quality, despite being trained on a relatively small 20GB dataset (Antoun et al., 2021). Finnish GPT-2 demonstrates that domain-specific fine-tuning on smaller parameter models can outperform larger multilingual models, highlighting the effectiveness of focused, resource-efficient training approaches for underrepresented languages (Luukkonen et al., 2023).

Foundational monolingual models such as BERT, T5, DistilBERT (Sanh et al., 2019), and BART (Lewis et al., 2019) have been adapted for multilingual tasks, despite being originally developed for high-resource languages. Adaptation methods include translate-train strategies, language adapters, and cross-lingual alignment. For example, BERT, although trained solely on English data, has served as the basis for multilingual variants such as mBERT, XLM-R, and LaBSE (Feng et al., 2022). However, adapting monolingual models to multilingual low-resource settings presents several challenges. These include vocabulary and tokenization mismatches, insufficient pretraining data, representation misalignment, domain and script incompatibility, high computational costs, and limited evaluation resources. Such limitations highlight the complexities of extending monolingual architectures to linguistically diverse and underrepresented languages.

XLM-R, mBERT, mT5, BLOOM (Workshop et al., 2022), and mBART are foundational models trained with a multilingual corpus to be adapted to multilingual low-resource languages. XLM-R supports more than 100 languages, including low-resource languages. It combines RoBERTa and XLM and is pre-trained on 2.5 terabytes tokens using masked language modelling. It has limitations on script diversity and less performance language with fewer than 100k example sentences. mT5 is a text-to-text unified framework and supports 101 languages. It focuses on low-resource languages by training the model using mC4 datasets, which include 101 languages, many of which are low-resource languages.

NLLB (Team et al., 2022), IndicBART (Dabre et al., 2022), AraT5 (Nagoudi et al., 2022), Aya (Üstün et al., 2024), Glot500 (Imani et al., 2023) are mulitilingual languages designed for low resource languages. NLLB supports more than 200 languages with more than 7B parameters and is trained through human-in-the-loop data curation focused on low-resource languages with the limitation of data scarcity for extremely low-resource languages, computational cost, low-resource to low-resource translation underperformance, legal/religious text over-representation, and subword tokenization challenge. IndicBART is a BART-based multilingual model designed for 11 major Indian (low-resource) languages and uses separate BPE tokenizers for each script family. Aya 23 is a multilingual language model for 101 languages through instruction fine-tuning. The model adapts the existing pretraining models like mT5 and BLOOM but focuses on human-annotated multilingual instruction datasets.

Even though there is no clear distinction between SLMs and LLMs, Lu et al. (2024) and Wang et al. (2024) provide some hints to categorise SLMs and LLMs. According to Wang et al. (2024), LMs that have emergent ability are classified as LLMs, and LMs with the number of parameters less than 7B are classified as SLMs. In some cases, LLMs are impractical due to high computational demands or privacy concerns. Wei et al. (2022) defines emergent ability as an ability to solve that is absent in smaller models, but present in LLMs. According to Wang et al. (2024), all the models specialised for African languages are categorised as SLMs.

3. Methodology

This study employed a structured three-stage review methodology to examine the current status, challenges, and prospects of Language Models (LMs) for African languages. The review systematically analyzed a curated selection of models from three categories: (1) commercial and open source LLMs, (2) foundational Small Language Models (SLMs), and (3) Specialized Small Language Models (SSLMs) tailored for African languages.

3.1. Model Selection

We chose prominent and representative models in each category based on their visibility in the African NLP literature and support for low-resource languages. The details of the models are in Table 1.

Table 1: Categorization of Language Models Reviewed

Category	Model Type	Examples
LLMs	Large-Scale General-Purpose Models	GPT-4 (OpenAI, 2023), Gemini 1.5 (Team and DeepMind, 2023), PaLM (Dai et al., 2023), LLaMA 3 (Dubey et al., 2024), DeepSeek V2 (DeepSeek-AI, 2024), Aya 23 (Üstün et al., 2024)
SLMs	Foundational Multilingual Models	BERT (Devlin et al., 2019), mBERT (Wu and Dredze, 2020), T5 (Raffel et al., 2020), mT5 (Xue et al., 2021), XLM (Lample and Conneau, 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020), NLLB 200 (Costa-jussà et al., 2022)
SSLMs	Specialized African-Centric Models	AfriBERTa (Ogueji et al., 2021), AfriTeVa (Jude Ogundepo et al., 2022), AfroLM (Dossou et al., 2022), EthioLLM (Tonja et al., 2024b), EthioMT (Tonja et al., 2024c), AfroXLMR (Alabi et al., 2022)

3.2. Review Procedure

For each model, we reviewed official documentation, technical reports, and peer-reviewed publications to extract: (1) language and script coverage, especially for African languages; (2) tokenization strategies (e.g., BPE, SentencePiece, character-level encodings); (3) training objectives and corpora; and (4) model architecture, parameter size, and computational requirements.

3.3. Dataset and Benchmark Mapping

We mapped each model to African-relevant datasets and benchmarks to assess linguistic utility and task alignment (Appendix A Table 8). We focused on datasets related to classification, named entity recognition, sentiment analysis, and machine translation. Models were evaluated based on their reported or inferable support for African languages and participation in benchmarks like MasakhaNER (Adelani et al., 2021) and EthioBenchMarks(Tonja et al., 2024a)

4. Discussion

We use the information discussed, such as the dataset and architecture, to answer a series of questions about the status of LLMs in African languages.

Question One. *Determining which African language is explored relatively more in LLMs, SLMs, and SSLMs and analysing any disparities in their coverage.*

Answer. Most of the LLMs like GPT-4, Gemini 1.5, PaLM 2, and DeepSeek have no clear documentation about the languages they support. This opacity makes it difficult to assess their true coverage and limits their accountability in addressing linguistic diversity. Foundational small language models, such as mBERT supports 104 languages, including 6 African languages (Research, 2019). mT5 supports 101 languages, of which 14 are from Africa (Research, 2021). XLM-R supports 100 languages, including 8 African languages.

Although African languages have approximately 28.6% of the 7,000 languages that exist around the world (Eberhard et al., 2025), underlying multilingual models considerably fail to represent them in proportion. If the representation of world languages in LLM is fair and proportional, the representation of indigenous African languages would have been better. For example, whereas mBERT for 104 languages should have approximately 30 (28.5% of 104 supported languages) African languages on board, it has only 6. Similarly, whereas mT5 should accommodate 29, it accommodates 14, and XLM-R has a mere 8 of a projected 29. This is indicative of the extreme underrepresentation of African languages in widely used language models and highlights the urgent need for more regionally and inclusively focused NLP efforts.

NLLB-200-1.5B supports 200 languages, which include 38 African languages (Research, 2022). As we can see from Figure 1 and Appendix A Table 6, there are considerable disparities in the coverage of African languages in SLMs. We can observe that most SLMs cover only 38 languages out of 2000 languages in Africa. We can categorise SLMs into monolingual SLMs, such as BERT, T5, RoBERTa, and XLM and multilingual SLMs, which include mBERT, mT5, XLM-R and NLLB 200-1.5B.

SSLMs show promise but face challenges like vocabulary, scripts, and tokenizer design, hindering equitable NLP development. This disparity underscores the urgent need for more inclusive AI development to bridge the linguistic gap and promote equitable access to technology across the continent. This uneven support is further illustrated in Figure 1.

As shown in Table 5, a total of 38 African languages are supported across six SLMs. Collectively, these models support approximately 41 African languages. A comprehensive list of the languages and their corresponding model support is provided in Appendix A, Table 7.

Among the supported languages, the most represented language is Amharic, followed by Somali, Swahili, and Yoruba. This could be attributed to relatively more digitized corpus availability, wider regional usage, government and international dataset inclusion, and linguistic resource availability. Their frequent recurrence across many models shows not only linguistic expansion but also prior priority in the creation of materials, as opposed to merely their predominance on the continent.

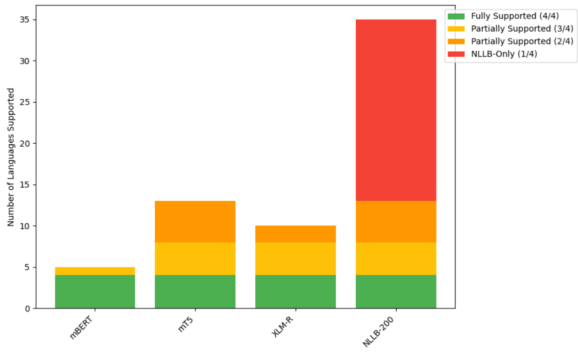

Figure 1: Languages across SLMs.

All monolingual foundational Small Language Models included in the study such as T5, BERT, RoBERTa do not support African languages directly, although some have been used as base models for further adaptations. Among the 38 African languages analyzed, only four Afrikaans, Amharic, Swahili, and Malagasy are fully supported by all multilingual SLMs.Appendix A Table 7 highlights the uneven distribution of African language support across SLMs and SSLMs.

Question Two. *Which African scripts are getting challenged in the representation of LLMs and why?*

Answer. Script: A writing system comprising visual symbols (e.g., Latin script: A-Z; Ge'ez script:**ሀ-ፐ).** There are approximately 37 writing systems historically used by African communities. Of these, 23 of them are currently in use across different regions of the continent, while the remaining 14 are no longer in use (Atelier National de Recherche Typographique (ANRT) et al., 2024); see the Appendix A Table 9. Among the active scripts, Latin, Ge'ez, and Arabic are the most widely used, collectively supporting 42 African languages.

Based on script dependency,we can divide language models into non-script agnostic, partially script-agnostic, and fully script-agnostic languages (Conneau et al., 2020; Xue et al., 2022). Most African languages are challenged by non-script agnostic and partially script Agnostic models. From 23 actively working scripts, only 3 scripts are used in large language and small language models, which shows that script-wise African languages are not explored. Some language models, such as ByT5 (Xue et al., 2022) and CANINE (Clark et al., 2022), become fully script agnostic by avoiding script-specific tokenization and using a byte-level or character-level tokenizer. Large language models such as Gemini, GPT are partially script agnostic; they use sentence pieces and Byte Pair Encoding, which works well for different scripts, but they may fail to work on unseen scripts.

Script-agnostic representations overcome orthographic barriers by processing text at sub-script levels - either through raw Unicode bytes or phonetic units. This approach enables unified handling of diverse African writing systems: Byte-level encoding (e.g., ByT5) processes all scripts as Unicode byte sequences, eliminating vocabulary biases against non-Latin scripts like Ge'ez (Amharic) and N'Ko (Mandé languages) while preserving diacritics essential for tonal languages (Xue et al., 2022). Phonetic representations convert speech or text to International Phonetic Alphabet (IPA) symbols before modeling, capturing unwritten languages like !Xóõ and dialectal variations without orthographic standardization (Adeniyi et al., 2023). Cross-script transfer learning, demonstrated in MasakhaNER 2.0, shows these representations reduce Ge'ez script NER errors by 37% compared to script-specific models while enabling zero-shot adaptation to newly encoded scripts like Vai (Adelani et al., 2023).

The foundational SLMs like mT5, mBERT, XLM-R, and NLLB are partially script agnostic, and they work for different scripts, but they do not do script normalization, they don't do well on unseen scripts during the training. Specialized models such as AfriBERTa, EthioLLM, and EthioMT are not script-agnostic models while models like AfriTeVa, AfroML, and AfroXLM-R are partially script-agnostic models. Which implies they don't work well on unseen data on the training. Script-agnostic representations bypass script-specific processing by operating directly on Unicode bytes or phonetic units. For African languages with non-Latin scripts (e.g., Ge'ez, N'Ko) or unwritten dialects, this enables: 1) Unified modeling: Byte-level tokenization (e.g., ByT5Xue et al. (2022)) handles all scripts without predefined vocabularies, 2) Oral language inclusion: Converting speech → IPA symbols → byte sequences captures unwritten languages, 3) Robustness: Eliminates errors from missing glyphs/fonts in underrepresented scripts.

Question Three: *What are the technical challenges faced in representation across LLMs, SLMs, and SSLMs for African languages?*

Answer. We can see the technical challenges of African language models in terms of large language models, foundational small language models, and specialized models for African languages on one hand, and in terms of fine-tuned models for African languages derived from existing models, language adaptation fine-tuned models for African languages, and models developed from scratch for African languages on the other.

Name	#Tokens	# of Parameters
AfriBERTa	108,800,600	5,440,030
AfriTeVa	108,800,600	5,440,030
AfroLM	259,396,720	12,969,836
EthioLLM	299,512,427	14,975,621
EthioMT	5,845,000	292,250
AfroXLMR	760,000,000	38,000,000

Technical specifications of the large language models are presented in Table 2. Aya was trained on approximately 500 billion tokens, whereas the rest of the models such as GPT-4, Gemini 1.5, PaLM 2, and LLaMA 3 were trained with the over one trillion token datasets. As seen in Table 2, the general challenges of LLMs are computational, cost to train the model, which is unaffordable for African low-resource language researchers.

Other challenges in training LLMs include data quality, as they require over a trillion tokens, which is difficult to obtain at that scale. Regarding low-resource languages, collecting such large amounts of data is challenging due to issues with data quality and bias. Additional challenges for low-resource languages include memory constraints, as models often use more than a trillion parameters during training.

Hallucination and safety challenges are recent problems shown on LLMs on low-resource languages (Guerreiro et al., 2023; Shen et al., 2024). The main reason behind the problem is a lack of quality data.

The main technical challenges of foundational SLMs include limited capacity, multilingual trade-off, and fine-tuning that needs domain-specific data are related to the limited capacity they have because it is directly related to the number of parameters in the model, the size and quality of the training data.

Training a model with monolingual data and fine-tuning the model with monolingual data produces better results because of the tokenization methods used (Rust et al., 2021). Monolingual models have low problems related to tokenization because it has flexibility in producing tokens and can engage native-speaking experts to incorporate language-specific rules, such as tokenizing compound words and morphological splits.

Table 2: Technical Details of Large Language Models

Model	**Architecture**	**Parameters**	**Tokenization**	**Training Data**	**Comput. Cost**	**Training Objective**	**Key Features**
GPT-4	Decoder-only	1.8T	BPE	13T tokens	$100M+	Autoregressive LM	Multimodal, strong reasoning
Gemini Ultra	Hybrid Enc-Dec	1.5T	SentencePiece	10T tokens	$100M+	Masked LM + Autoregressive	Multimodal (text, images, video)
PaLM 2	Decoder-only	540B	SentencePiece	10T tokens	$10M+	Autoregressive	Text-only
LLaMA 3 (70B)	Decoder-only	70B	BPE variant	5T tokens	$20M	Autoregressive LM	Open-weight, efficient
DeepSeek-V3	Decoder-only	500B	BPE	8T tokens	$50M	Autoregressive LM	128K context, strong Chinese/English
Aya 23	Decoder-only	8B	Unigram	500B tokens	$5M	Instruction Tuning	101-language focused

Question Four. *What are the benchmark datasets and models in developing LLMs for African languages?*

Answer. This study reveals that around 23 publicly available datasets are used by models solely SSLMs. Figure 2 shows the relationship between NLP tasks and the benchmark datasets prepared for those specific tasks. The study highlights disparities in the availability and utilization of benchmark datasets for African low-resource languages. Classification tasks have the highest number of datasets and models. This is likely because these tasks are simple and need less linguistic nuance than translation or named entity recognition. Appendix A Table 8 details the datasets benchmarks used across the specialised models for African languages. In the task column, General refers to a pretrain corpus, and mixed refers to a single benchmark which includes more than one task, like EthioBenchmark, which has five datasets for five tasks, including machine translation, part of speech tagging, classification, sentiment analysis, and named entity recognition.

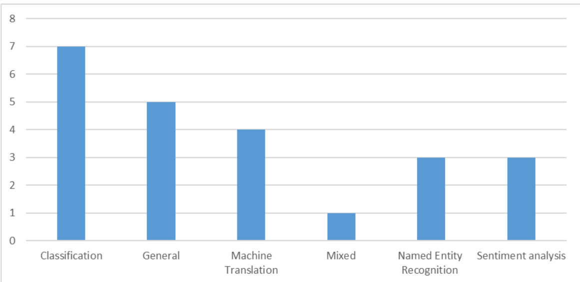

Figure 2: Different NLP tasks and the number of datasets prepared for the task

Question Five. *Exploring the prospective of LLMs for Africa.*
Answer. Currently, artificial intelligence shows emergent ability which are arithmetic reasoning, agentic behaviour, common sense reasoning and symbolic reasoning. The path for this destination is very clear. Constructing LLMs for African languages is both challenge and an integral opportunity. Figure 3 outlines a strategic, step-by-step approach, beginning with foundational tasks such as language standardization, normalization, and formalization.

These are necessary for Africa, where linguistic diversity prevails, no orthographic consensus exists, and digital resources are scarce, hindering NLP development. Without rules for standard spelling and morphological annotation schemes, even basic text preprocessing becomes problematic.

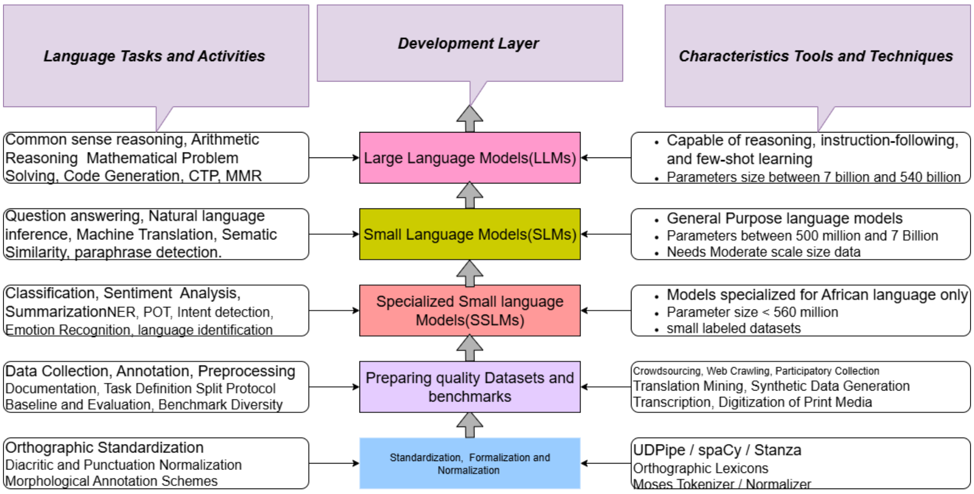

Figure 3: Roadmap for African Language-Model Development.
Figure 3 shows our recommendation roadmap for African languages in their bid to develop African langauge-centric specialized language models and inclusion in commercial LLMs. The diagram proceeds bottom-up. Foundational work on Standardisation, Formalisation, and Normalisation feeds directly into Preparing Quality Datasets & Benchmarks. These resources enable (i) SSLMs ($<$ 500 M parameters) that target tightly scoped NLP tasks common in African contexts, (ii) General- purpose SLMs (500 M–7 B parameters) capable of broader cross-lingual tasks, and finally (iii) LLMs ($>$7 B parameters) that support advanced reasoning and generative capabilities. The left column lists representative activities or tooling at each layer; the right column summarises the defining requirements. The central upward arrow highlights the dependency chain each layer builds on the assets and insights created in the layer(s) below.

The second level, preparation of dataset and benchmark, remains the main bottleneck. Most African languages are low-resource languages with few labeled data, small digitized corpora, and few benchmark resources. Crowdsourcing, participatory annotation, and digitization of oral and print sources are essential for bridging the data gap. Large-scale collaborative datasets like Common Crawl (Common Crawl Foundation, 2023) the crowd-sourced web corpus that enabled foundational models like T5 (Raffel et al., 2020) demonstrate how decentralized collection can build comprehensive resources when sustained through institutional funding pools. For African languages, we advocate adapting this approach through: 1) Community web harvesting of local digital content, 2) Structured oral history transcription drives, and 3) National text aggregation mandates – with dedicated funding from national AI innovation funds (e.g., Nigeria's 0.5% digital levy) and matched cloud credit allocations from corporate partners (AWS/Azure African NLP Grants). These resources will fuel the next step - to develop SSLMs that are task-specific and focused on individual languages with limited data. Such models serve as a stepping stone, enabling tangible NLP applications such as sentiment analysis and language identification and informing model construction.

SLMs of medium parameters enables multilingual capabilities and more universal tasks such as question answering and semantic similarity. Scaling to full-size LLMs, which require big data and computing capacity, remains an ambitious goal, given the continent's nascent AI infrastructure. Nonetheless, the roadmap shows a realistic and inclusive path—from foundational linguistic research to advanced models of reasoning—positioning Africa to ultimately solve its issues and build fair AI development for its multilingual populations. (Hoffmann et al., 2022) states that, when the parameters of LLMs getting bigger, naturally it gets emergent ability and it can be unlock using chains of thought prompting. The paper claims when the parameters of LLMs reach around 540 billion the model naturally gain the emergent ability. Detail downstream NLP tasks/capabilities of language models can be read from (Wang et al., 2024) for SLMs and (Qin et al., 2024) for LLMs.

The chinchilla optimal ratio by Hoffmann et al. (2022) draws the relationship between the number of parameters used in the model with the number of tokens in the training datasets which is: Number of Training Tokens (D) $\approx 20\times$ Number of Parameters (N).

Let us assume specialised models for African languages trained with a quality corpus and datasets. Table 4 shows the number of parameters for each specialised model for African languages. The models are far too large to reach 540 billion parameters.

As depicted in Figure 4, models vary greatly in size, from small models like AfriBERTa (10 million parameters) to large models like PaLM 2 (530B) and LLaMA 3 (405B). Larger models better capture linguistic nuances but need extensive resources. Most models use SentencePiece, while smaller ones use WordPiece. Large models are trained on billions of tokens for broader tasks; MoE architectures boost efficiency. Small models target specific languages, while large ones are multilingual, aiming for wider coverage and improved performance.

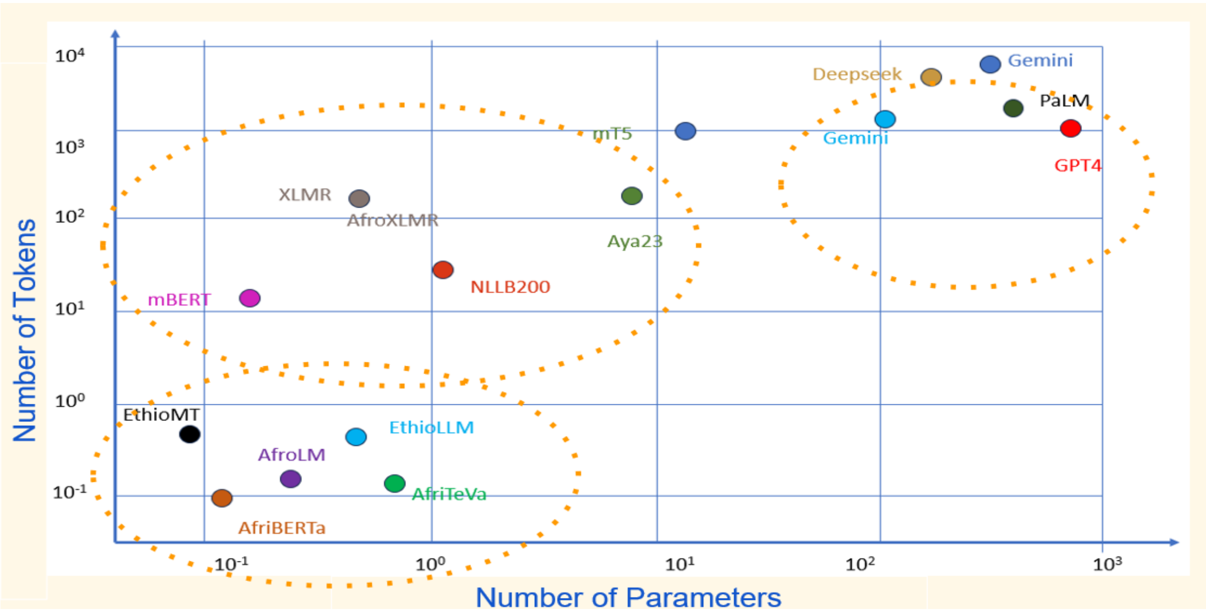

Figure 4: Model parameter size versus number of training tokens.

Conclusion

This review reveals that African languages remain significantly underrepresented in current language models. Out of over 2,000 languages, only about 41 have any support in existing LLMs, SLMs, or SSLMs, primarily those with large speaker populations or official status. Script coverage is similarly limited, with just three scripts (Latin, Arabic, Ge'ez) being widely supported.

Major challenges include severe data scarcity, morphological complexity, a lack of standardized orthographies, and limited computational resources. Existing models, especially LLMs, require vast amounts of training data and infrastructure, posing substantial barriers to NLP development for African languages. Additionally, benchmark availability is sparse and unevenly distributed across tasks.

Despite these obstacles, progress with SSLMs shows potential for targeted advancement. A realistic roadmap begins with foundational linguistic work, followed by resource creation, and ultimately scalable models—offering a clear path toward inclusive language technologies for Africa.

Recommendations

Advancing NLP for African languages requires developing tailored models like SSLMs and script-agnostic approaches, with a focus on improving data quality and culturally aware evaluations to reduce bias. It is also important to promote community-driven data collection, standardize scripts, and expand benchmarks across diverse tasks, while supporting open-access platforms. Additionally, securing institutional and government backing with funding, resources, and inclusion of African languages in digital services, alongside fostering international collaborations, will help elevate African language representation in AI research.

References

David I. Adelani, Jesujoba Alabi, Angela Fan, et al. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In EMNLP, pages 3702–3720, 2023. URL https://aclanthology.org/2023.emnlp-main.230.

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

Victor Adeniyi, Tunde Adegbola, and Chris Chinenye Emezue. Yoruba text-to-speech with transfer learning and phonetic embeddings. In SIGUL, pages 122–127, 2023. URL https://aclanthology.org/2023.sigul-1.29.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, Proceedings of the 29th International Conference on Computational Linguistics, pages 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382/.

Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. In Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak, editors, Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL https://aclanthology.org/2020.osact-1.2/.

Wissam Antoun, Fady Baly, and Hazem Hajj. AraGPT2: Pre-trained transformer for Arabic language generation. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 196–207, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wanlp-1.21/.

Atelier National de Recherche Typographique (ANRT), Institut Designlabor Gutenberg (IDG), and Script Encoding Initiative (SEI). The world's writing systems, 2024. URL https://www.worldswritingsystems.org/. Accessed: 2025-05-30.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. Transactions of the Association for Computational Linguistics, 10:73–91, 2022. doi: 10.1162/tacl_a_00448. URL https://aclanthology.org/2022.tacl-1.5/.

Common Crawl Foundation. Common Crawl. https://commoncrawl.org, 2023. Accessed: 2024-07-31.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Yan He, Elahe Kalbassi, Linlu Wang, Long Wang, Shuo Zhang, Angela Fan, et al. No language left behind: Scaling human-centered machine translation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, UAE, 2022. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. IndicBART: A pre-trained model for indic natural language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 1849–1863, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.145. URL https://aclanthology.org/2022.findings-acl.145/.

Andrew M. Dai, Jonathan H. Clark, Kevin Robinson, Maysam Moussalem, Sebastian Ruder, Siamak Shakeri, Jacob Austin, et al. Palm 2 technical report. Technical report, Google, 2023.

DeepSeek-AI. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, December 2024.

Jaleh Delfani, Constantin Orasan, Hadeel Saadany, Ozlem Temizoz, Eleanor Taylor-Stilgoe, Diptesh Kanojia, Sabine Braun, and Barbara Schouten. Google translate error analysis for mental healthcare information: Evaluating accuracy, comprehensibility, and implications for multilingual healthcare communication, 2024. URL https://arxiv.org/abs/2402.04023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, and Andreas Rücklé, editors, Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sustainlp-1.11. URL https://aclanthology.org/2022.sustainlp-1.11/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the World. SIL International, 28th edition, 2025.

Ethnologue. Ethnologue: Languages of the world, 2025. Accessed: 2025-05-20.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in large multilingual translation models. Transactions of the Association for Computational Linguistics, 11:1500–1517, 2023. doi: 10.1162/tacl_a_00615. URL https://aclanthology.org/2023.tacl-1.85/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, 2022.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1082–1117, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61/.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta, editors, Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 126–135, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.14. URL https://aclanthology.org/2022.deeplo-1.14/.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics, pages 757–770, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.66. URL https://aclanthology.org/2020.coling-main.66/.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact memorization of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11263–11282, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.658. URL https://aclanthology.org/2024.findings-emnlp.658/.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. FinGPT: Large generative models for a small language. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2710–2726, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.164. URL https://aclanthology.org/2023.emnlp-main.164/.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. AraT5: Text-to-text transformers for Arabic language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 628–647, Dublin, Ireland, May 2022. Association for

Computational Linguistics. doi: 10.18653/v1/2022.acl-long.47. URL https://aclanthology.org/2022.acl-long.47/.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pre-trained multilingual language models for low-resourced african languages. In Proceedings of the First Workshop on Natural Language Processing for African Languages (AfriNLP 2021), pages 116–126, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.afrinlp-1.11.

OpenAI. Gpt-4 technical report, 2023. Technical Report.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. Large language models meet nlp: A survey, 2024. URL https://arxiv.org/abs/2405.12819.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

Google Research. Bert multilingual model readme, 2019. Accessed: [Insert Date].

Google Research. Multilingual t5: Massively multilingual pre-trained text-to-text transformer, 2021. Accessed: September 30, 2025.

Meta AI Research. NLLB-200: Distilled 600m-parameter model of No Language Left Behind (NLLB), 2022. Accessed: September 30, 2025.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 2668–2680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.156. URL https://aclanthology.org/2024.findings-acl.156/.

Gemini Team and Google DeepMind. Gemini: A family of highly capable multimodal models, December 2023. Technical Report.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3048–3071, Abu Dhabi, UAE, 2022.

The Guardian. Asylum seekers using AI translation apps for official paperwork, lawyers say, September 2023. URL https://www.theguardian.com/us-news/2023/sep/07/asylum-seekers-ai-translation-apps. Accessed: 2024-07-31.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, et al. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. arXiv preprint arXiv:2403.13737, 2024a.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6341–6352, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.561/.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Jugal Kalita. EthioMT: Parallel corpus for low-resource Ethiopian languages. In Rooweither Mabuya, Muzi Matfunjwa, Mmasibidi Setaka, and Menno van Zaanen, editors, Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024, pages 107–114, Torino, Italia, May 2024c. ELRA and ICCL. URL https://aclanthology.org/2024.rail-1.12/.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845/.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. arXiv preprint arXiv:2411.03350, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837, 2022.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Saso Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint, arXiv:2211.05100, 2022. Conference version at NeurIPS 2022 (Track on Datasets and Benchmarks).

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi, editors, Proceedings of the 5th Workshop on Representation Learning for NLP, pages 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16/.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Appendix A. Technical details of Foundational and Specialized Language Models

Table 3: Technical Overview of Foundational Models for African Languages

Model Variant	Architecture	Pretraining Objective	Params	Layers	Heads	Hidden Size	Tokenization	Max Seq Len	Batch Size	Learning Rate	Optimizer	Training Steps	Datasets Used	Training Data Size
BERT-base	Encoder-only	MLM + NSP	110M	12	12	768	WordPiece (30k vocab)	512	256	1e-4	Adam	1M	Wikipedia + BookCorpus	16GB
BERT-large	Encoder-only	MLM + NSP	340M	24	16	1024	WordPiece (30k vocab)	512	256	1e-4	Adam	1M	Wikipedia + BookCorpus	16GB
mBERT	Encoder-only	MLM + NSP	110M	12	12	768	WordPiece (110k vocab)	512	256	5e-5	Adam	1M+	Wikipedia (104 languages)	N/S
T5-small	Encoder-Decoder	Span Corruption	60M	6	8	512	Sentence-Piece (32k)	512	128	0.01	AdaFactor	1M	C4 (English)	750GB
T5-base	Encoder-Decoder	Span Corruption	220M	12	12	768	Sentence-Piece (32k)	512	128	0.01	AdaFactor	1M	C4 (English)	750GB
T5-large	Encoder-Decoder	Span Corruption	770M	24	16	1024	Sentence-Piece (32k)	512	128	0.01	AdaFactor	1M	C4 (English)	750GB
mT5-small	Encoder-Decoder	Span Corruption	300M	8	6	512	Sentence-Piece (250k)	512	1024	0.01	AdaFactor	1M	mC4 (101 languages)	750GB (balanced)
mT5-large	Encoder-Decoder	Span Corruption	1.2B	24	16	1024	Sentence-Piece (250k)	512	1024	0.01	AdaFactor	1M	mC4 (101 languages)	750GB (balanced)
RoBERTa-base	Encoder-only	Dynamic MLM (no NSP)	125M	12	12	768	BPE (50k vocab)	512	8K	6e-4	AdamW	500K	CC-News + OpenWebText + Stories	160GB
RoBERTa-large	Encoder-only	Dynamic MLM (no NSP)	355M	24	16	1024	BPE (50k vocab)	512	8K	6e-4	AdamW	500K	CC-News + OpenWebText + Stories	160GB
XLM-base	Encoder-only	MLM + CLM (+TLM if parallel)	250M	12	12	2048	BPE (95k vocab)	512	64	5e-5	Adam	500K	Wikipedia + Parallel data	N/S
XLM-R-base	Encoder-only	MLM (RoBERTa-style)	270M	12	12	768	Sentence-Piece (250k)	512	8K	5e-4	AdamW	500K	Common Crawl (100 langs)	2.5TB (balanced)
XLM-R-large	Encoder-only	MLM (RoBERTa-style)	550M	24	16	1024	Sentence-Piece (250k)	512	8K	5e-4	AdamW	500K	Common Crawl (100 langs)	2.5TB (balanced)
NLLB-200 1.5B	Transformer	Denoising + MT	1.5B	24	16	2048	Sentence-Piece (256k vocab)	512	128	1e-4	Adam	250K	FLORES-200, CCMatrix, CCAligned, Wikipedia, Tatoeba	1.7T tokens

Table 4: Technical Details of Specialized Models for African Languages

Model Variant	Base Model	Pretraining Objective	Params	Layers	Heads	Hidden Size	Tokenization	Max Seq Len	Batch Size	Learning Rate	Optimizer	Training Data	Data Size
AfriBERTa small	BERT	MLM (No NSP)	11M	6	6	256	WordPiece (50k vocab)	128	32	5e-5	AdamW	OSCAR + Local News (17 African langs)	5GB
AfriBERTa large	BERT	MLM (No NSP)	124M	12	12	768	WordPiece (50k vocab)	512	128	3e-5	AdamW	OSCAR + Local News (17 African langs)	5GB
AfriTeVa-base	T5	Span Corruption (text-to-text)	223M	12	12	768	SentencePiece (32k vocab)	512	128	1e-4	AdaFactor	CC-100 + JW300 (20 African langs)	10GB
AfroLM-1B	RoBERTa	Dynamic MLM	1B	24	16	1024	BPE (100k vocab)	512	2048	6e-4	AdamW	ALPACA Corpus (25 African langs)	500GB
EthioLLM-7B	LLaMA-2	Causal LM (Autoregressive)	7B	32	32	4096	Byte-level BPE (50k vocab)	2048	1024	2e-5	AdamW	Ethiopic Texts (Amharic, Tigrinya)	200GB
EthioMT-base	mT5	Span Corruption	300M	8	6	512	SentencePiece (250k vocab)	512	512	1e-3	AdaFactor	Parallel Bible (10 Ethio langs)	8GB (parallel)
AfroXLMR base	XLM-R	MLM	270M	12	12	768	SentencePiece (250k vocab)	512	1024	5e-4	AdamW	Common Crawl (30 African langs)	1TB (balanced)
AfroXLMR large	XLM-R	MLM	550M	24	16	1024	SentencePiece (250k vocab)	512	1024	5e-4	AdamW	Common Crawl (30 African langs)	1TB (balanced)

Appendix A. Languages, Scripts, and Benchmark Datasets

Table 5: Specialised Language Models for African Languages

Sno.	Language	AfriBERTa	AfriTeVa	AfroLM	AfroXLMR	EthioLLM	EthioMT
1	Afrikaans	No	No	Yes	Yes	No	No
2	Amharic	Yes	Yes	Yes	Yes	Yes	Yes
3	Afaan Oromo	No	Yes	Yes	Yes	Yes	Yes
4	Afar	No	No	No	No	No	Yes
5	Awngi	No	No	No	No	No	Yes
6	Bambara	No	Yes	Yes	No	No	No
7	Basketo	No	No	No	No	No	Yes
8	Dawuro	No	No	No	No	No	Yes
9	Fulah	No	No	Yes	Yes	No	No
10	Gamo	No	No	No	No	No	Yes
11	Ge’ez	No	No	No	No	Yes	Yes
12	Gofa	No	No	No	No	No	Yes
13	Gurage	No	No	No	No	No	Yes
14	Hadiya	No	No	No	No	No	Yes
15	Hausa	Yes	Yes	Yes	Yes	No	No
16	Igbo	Yes	Yes	Yes	Yes	No	No
17	Kafa	No	No	No	No	No	Yes
18	Kinyarwanda	Yes	Yes	Yes	Yes	No	No
19	Korate	No	No	No	No	No	Yes
20	Luganda	Yes	Yes	Yes	Yes	No	No
21	Luo	Yes	Yes	Yes	Yes	No	No
22	Majang	No	No	No	No	No	Yes
23	Male	No	No	No	No	No	Yes
24	Murule	No	No	No	No	No	Yes
25	Nigerian Pidgin	Yes	Yes	Yes	Yes	No	No
26	Nuer	No	No	No	No	No	Yes
27	Shakicho	No	No	No	No	No	Yes
28	Shona	Yes	Yes	Yes	Yes	No	No
29	Sidama	No	No	No	No	No	Yes
30	Somali	Yes	Yes	Yes	Yes	Yes	Yes
31	Swahili	Yes	Yes	Yes	Yes	No	No
32	Tigrinya	No	Yes	Yes	Yes	Yes	Yes
33	Twi	No	Yes	Yes	Yes	No	No
34	Wolaytta	No	No	No	No	No	Yes
35	Wolof	No	Yes	Yes	Yes	No	No
36	Xhosa	No	No	Yes	Yes	No	No
37	Yoruba	Yes	Yes	Yes	Yes	No	No
38	Zulu	No	No	Yes	Yes	No	No

Table 6: Foundation Model Support for African Languages

Sno.	African Language	mBERT	mT5	XLM-R	NLLB-200	Support Count
1	Afrikaans	Yes	Yes	Yes	Yes	4
2	Amharic	Yes	Yes	Yes	Yes	4
3	Swahili	Yes	Yes	Yes	Yes	4
4	Malagasy	Yes	Yes	Yes	Yes	4
5	Hausa	No	Yes	Yes	Yes	3
6	Somali	No	Yes	Yes	Yes	3
7	Xhosa	No	Yes	Yes	Yes	3
8	Yoruba	Yes	Yes	No	Yes	3
9	Chichewa (Nyanja)	No	Yes	No	Yes	2
10	Igbo	No	Yes	No	Yes	2
11	Oromo	No	No	Yes	Yes	2
12	Shona	No	Yes	No	Yes	2
13	Southern Sotho	No	Yes	No	Yes	2
14	Zulu	No	Yes	No	Yes	2
15	Bambara	No	No	No	Yes	1
16	Bemba	No	No	No	Yes	1
17	Dyula	No	No	No	Yes	1
18	Ewe	No	No	No	Yes	1
19	Fon	No	No	No	Yes	1
20	Fulfulde (Nigerian Fulfulde)	No	No	No	Yes	1
21	Ganda	No	No	No	Yes	1
22	Kabyle	No	No	No	Yes	1
23	Kamba (Kenya)	No	No	No	Yes	1
24	Kikuyu	No	No	No	Yes	1
25	Kinyarwanda	No	No	No	Yes	1
26	Kimbundu	No	No	No	Yes	1
27	Kongo	No	No	No	Yes	1
28	Lingala	No	No	No	Yes	1
29	Luba-Lulua	No	No	No	Yes	1
30	Luo (Kenya & Tanzania)	No	No	No	Yes	1
31	Mossi	No	No	No	Yes	1
32	Nuer	No	No	No	Yes	1
33	Pedi (Northern Sotho)	No	No	No	Yes	1
34	Swati	No	No	No	Yes	1
35	Tamasheq	No	No	No	Yes	1
36	Tumbuka	No	No	No	Yes	1
37	Twi	No	No	No	Yes	1
38	Wolof	No	No	No	Yes	1

Table 7: Language Model Support for African Languages

Sno.	Language	AfriBERTa	AfriTeVa	AfroLM	AfroXLMR	EthioLLM	EthioMT	mBERT	mT5	XLM-R	NLLB-200	Count of "Yes"
1	Amharic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	10
2	Somali	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	9
3	Swahili	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	8
4	Yoruba	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	7
5	Hausa	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	7
6	Igbo	Yes	Yes	Yes	Yes	No	No	No	Yes	No	Yes	6
7	Shona	Yes	Yes	Yes	Yes	No	No	No	Yes	No	Yes	6
8	Kinyarwanda	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	5
9	Luganda	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	5
10	Luo	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	5
11	Nigerian Pidgin	Yes	Yes	Yes	Yes	No	No	No	No	No	No	4
12	Oromo	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	7
13	Tigrinya	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	5
14	Twi	No	Yes	Yes	Yes	No	No	No	No	No	Yes	4
15	Wolof	No	Yes	Yes	Yes	No	No	No	No	No	Yes	4
16	Bambara	No	Yes	Yes	No	No	No	No	No	No	Yes	3
17	Afrikaans	No	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	6
18	Xhosa	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	5
19	Zulu	No	No	Yes	Yes	No	No	No	Yes	No	Yes	4
20	Fulah	No	No	Yes	Yes	No	No	No	No	No	Yes	3
21	Ge'ez	No	No	No	No	Yes	Yes	No	No	No	No	2
22	Nuer	No	No	No	No	No	Yes	No	No	No	Yes	2
23	Afar	No	No	No	No	No	Yes	No	No	No	No	1
24	Awngi	No	No	No	No	No	Yes	No	No	No	No	1
25	Basketo	No	No	No	No	No	Yes	No	No	No	No	1
26	Dawuro	No	No	No	No	No	Yes	No	No	No	No	1
27	Gamo	No	No	No	No	No	Yes	No	No	No	No	1
28	Gofa	No	No	No	No	No	Yes	No	No	No	No	1
29	Gurage	No	No	No	No	No	Yes	No	No	No	No	1
30	Hadiya	No	No	No	No	No	Yes	No	No	No	No	1
31	Kafa	No	No	No	No	No	Yes	No	No	No	No	1
32	Korate	No	No	No	No	No	Yes	No	No	No	No	1
33	Majang	No	No	No	No	No	Yes	No	No	No	No	1
34	Male	No	No	No	No	No	Yes	No	No	No	No	1
35	Murule	No	No	No	No	No	Yes	No	No	No	No	1
36	Shakicho	No	No	No	No	No	Yes	No	No	No	No	1
37	Sidama	No	No	No	No	No	Yes	No	No	No	No	1
38	Wolaytta	No	No	No	No	No	Yes	No	No	No	No	1
39	Malagasy	No	No	No	No	No	No	Yes	Yes	Yes	Yes	4
40	Chichewa (Nyanja)	No	No	No	No	No	No	No	Yes	No	Yes	2
41	Southern Sotho	No	No	No	No	No	No	No	Yes	No	Yes	2

Table 8: Summary of Datasets/Benchmarks used in SSLMs for African Languages

No.	Available Benchmarks Dataset	Availability	Papers	Size	Languages	Tasks
1	MasakhaNER	Public	5	140M	10	Named Entity Recognition
2	News Topic Classification Dataset (from Hedderich et al., 2020)	Public	3	50M	2	Classification
3	Multilingual corpus covering 11 African languages	Public	1	100M	11	General
4	Shared Task: Machine Translation of News	Public	1	1.2 GB	7	Machine Translation
5	CommonCrawl (CC-100)	Public	1	3GB	20	General
6	YOSM	Public	1	1M	30	Sentiment analysis
7	NaijaSenti	Public	2	15M	4	Sentiment analysis
8	MasakhaneNEWS	Public	1	500M	16	Classification
9	EthioBenchmark	Public	1	20M	6	Mixed
10	Parallel Corpus (Abate et al., 2019)	Public	1	400M	5	Machine Translation
11	Parallel Corpus sets (Lakew et al., 2020)	Public	1	600M	11	Machine Translation
12	Parallel Corpora (Vegi et al., 2022)	Public	1	600M	15	Machine Translation
13	mT5 pre-training corpus/mC4	Public	1	10GB	25	General
14	BBC Media Dataset	Public	1	500M	7	General
15	VOA Media Dataset	Public	1	600M	12	General
16	CoNLL 2003 NER task	Public	1	13M	1	Named Entity Recognition
17	ANERCorp	Public	1	4M	1	Named Entity Recognition
18	New Topic Classification (Azime & Mohammed, 2021)	Public	1	40M	1	Classification
19	AG News corpus	Public	1	50M	1	Classification
20	Kinyarwanda – KINNEWS	Public	1	12M	1	Classification
21	Kiswahili – new classification	Public	1	40M	1	Classification
22	Am-Senty Yimam et al. (2020)	Public	1	6M	1	Classification
23	ANTC corpus	Public	1	300M	20	Sentiment analysis

Total: 23 datasets, 18GB total size, covering 41 languages.

Table 9: List of Scripts with Their Time Periods, Usage Status, Geographic Distribution, and Associated Languages

SNo.	Name of the Script	Time Period	Still in Use?	Countries/Regions	Languages Written
1	Egyptian Hieroglyphs	33rd c. BCE – 1st c. CE	No	Egypt, Sudan	Ancient Egyptian
2	Hieratic	29th c. BCE – 2nd c. CE	No	Egypt	Ancient Egyptian
3	Demotic	650 BCE – 6th c. CE	No	Egypt	Late Egyptian
4	Ethiopic (Ge'ez)	4th c. BCE – Present	Yes	Ethiopia, Eritrea	Amharic, Tigrinya, Tigre, Ge'ez
5	Meroitic Cursive	3rd c. BCE – 4th c. CE	No	Sudan	Meroitic
6	Meroitic Hieroglyphs	3rd c. BCE – 4th c. CE	No	Sudan	Meroitic
7	Numidian	2nd c. BCE – 3rd c. CE	No	Algeria, Tunisia	Old Numidian
8	Coptic	4th c. CE – Present	Yes	Egypt	Coptic (liturgical)
9	Tifinagh	3rd c. CE – Present	Yes	Morocco, Algeria, Mali, Niger	Tamazight, Tuareg languages
10	Vai	1830 – Present	Yes	Liberia, Sierra Leone	Vai language
11	Bamum	1896 – Present	Yes	Cameroon	Bamum language
12	Old Bamum	1896 – 20th c.	No	Cameroon	Bamum language
13	Bassa Vah	1907 – Present	Yes	Liberia	Bassa language
14	Bagam	1910 – Late 20th c.	No	Cameroon	Bagam language
15	Mende Kikakui	1920 – Present	Yes	Sierra Leone	Mende language
16	Osmanya	1920 – 1973	No	Somalia	Somali
17	N'Ko	1949 – Present	Yes	Guinea, Mali, Ivory Coast	Manding languages
18	Bété	1956 – Present	Yes	Ivory Coast	Bété language
19	Kaddare	1952 – Present	Yes	Somalia	Somali
20	Fula Dita (Fula 1)	1958 – 1970	No	Guinea	Fula
21	Fula Ba (Fula 2)	1963 – 1970	No	Guinea	Fula
22	Garay (Wolof)	1961 – Present	Yes	Senegal, Gambia	Wolof
23	Mandombe	1978 – Present	Yes	DR Congo, Congo-Brazzaville	Kikongo, Lingala, other Congolese languages
24	Mwangwego	1979 – Present	Yes	Malawi	Chewa, other Malawian languages
25	Adlam	1980s – Present	Yes	Guinea, Nigeria, Cameroon	Fula
26	Beria	1980s – Present	Yes	Sudan	Zaghawa
27	Luo	2009 – Present	Yes	Kenya, Tanzania	Dholuo
28	Isibheqe Sohlamvu	20th c. – Present?	Yes?	South Africa, Eswatini	Nguni languages (Zulu, Xhosa)
29	Afaka	1908 – Present	Yes	DR Congo	Ndyuka
30	Medefaidrin (Oberi Okaime)	1930s – Present	Yes	Nigeria	Ibibio, Efik
31	Masaba	1930 – Present	Yes	Uganda	Lugisu
32	Borama	1933 – Present?	Yes?	Somalia	Somali
33	Kpelle	1930s – Present	Yes	Liberia, Guinea	Kpelle
34	Loma	1930s – Present	Yes	Liberia, Guinea	Loma
35	Tafi (Hausa 3)	1977 – 2011	No	Nigeria, Niger	Hausa
36	Raina Kama (Hausa 2)	1990s – 1999	No	Nigeria, Niger	Hausa
37	Salifou Hausa (Hausa 1)	1998 – 2004	No	Nigeria, Niger	Hausa