SabiYarn: Advancing Low Resource Languages with Multitask NLP Pretraining

Jeffrey Otoibhi
drjeffreypaul@gmail.com

Oduguwa Damilola
University of Lagos
oduguwadamilola40@gmail.com

Okpare David
dave@datached.com

Abstract

The rapid advancement of large language models (LLMs) has revolutionized natural language processing, yet a significant challenge persists: the under representation of low-resource languages. This paper introduces SABIYARN, a novel 125M parameter decoder-only language model specifically designed to address this gap for Nigerian languages. Our research demonstrates that a relatively small language model can achieve remarkable performance across multiple languages even in a low-resource setting when trained on carefully curated task-specific datasets. We introduce a multitask learning framework designed for computational efficiency, leveraging techniques such as sequence packing to maximize token throughput per batch. This allows SABIYARN to make the most of a limited compute budget while achieving strong performance across multiple NLP tasks.

This paper not only highlights the effectiveness of our approach but also challenges the notion that only massive models can achieve high performance in diverse linguistic contexts, outperforming models over 100 times its parameter size on specific tasks such as translation (in both directions), Named Entity Recognition, Text Diacritization, and Sentiment Analysis in the low-resource languages it was trained on. SabiYarn-125M represents a significant step towards democratizing NLP technologies for low-resource languages, offering a blueprint for developing efficient, high-performing models tailored to specific linguistic regions. Our work paves the way for more inclusive and culturally sensitive AI systems, potentially transforming how language technologies are developed and deployed in linguistically diverse areas like Nigeria and beyond.

1 Introduction

The field of natural language processing (NLP) has witnessed remarkable advancements in recent years, driven by the development of large-scale, pre-trained language models. These powerful models have demonstrated impressive capabilities in handling a variety of language-related tasks, from text generation to language understanding, and emergent reasoning abilities as they scale to ever-increasing model sizes (Wei et al., 2022). Despite the remarkable progress in NLP, the performance of large language models (LLM) in African languages remains suboptimal. Recent studies, such as the analysis by (Ojo et al., 2023), highlight the significant performance gap between African languages and high-resource languages such as English in the state-of-the-art large language models, including LLaMa 2 (Touvron et al., 2023), and GPT-4 (Achiam et al., 2023). Their findings reveal that while GPT-4 achieves average or impressive results on classification tasks, it performs poorly on generative tasks like machine translation, while LLaMa 2 recorded the worst performance due to its English-centric pretraining and limited multilingual capabilities. These results underscore the urgent need to address the under representation of African languages in LLMs, ensuring they are not left behind as these technologies continue to evolve.

To address this gap, we present SABIYARN-125M, a decoder-only foundational (pre-trained) language model specifically designed to support the major languages spoken in Nigeria. Our model tackles two main challenges in developing NLP solutions for Nigerian languages: limited computational resources and a scarcity of high-quality data sources. Using a diverse training dataset and a multitask learning approach, this model aims to provide versatile and inclusive language technology that can empower Nigerian communities and contribute to the global NLP landscape. Our model is pre-trained on a diverse dataset covering nine Nigerian languages: Yoruba, Igbo, Hausa, Pidgin English, Fulani, Fulah, Fulfulde, Uhrobo, and Efik. Previ-

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 95–107

ous models have predominantly focused on the four major Nigerian Languages, Yoruba, Igbo, Pidgin, and Hausa. Our work builds on this foundation by extending further language coverage beyond the four major Nigerian Languages, to include several underrepresented languages, increasing language diversity and enabling our model SABIYARN to perform various NLP tasks while preserving cultural and linguistic nuances.

We adopt a mixture of training strategies, including a technique called Sequence Packing (Krell et al., 2022)) for the efficient processing of sequences to speed up pretraining and minimize wasted attention computation, task-conditioning prompts inspired by (Raffel et al., 2020), a multi-task learning objective (Zhang and Yang, 2021) and a custom loss computation strategy that leverages sequence packing, ensuring the model learns precisely from the task-relevant information. This hybrid approach allows us to maximize the potential of each parameter given the limited resources, achieving impressive results across a range of NLP tasks, including Named Entity Recognition, Topic classification, Translation, Diacritization, and Sentiment Analysis, even in zero-shot settings.

In the following sections, we detail our methodology, present our results, and discuss the implications of our findings for the future of NLP in Nigeria and potentially other linguistically diverse regions. Our work contributes to the democratization of NLP technologies but also paves the way for more inclusive AI solutions that respect and preserve linguistic diversity.

2 Related Work

The rapid advancement of large language models (LLMs) has revolutionized natural language processing (NLP), with models like GPT (Radford and Narasimhan, 2018) demonstrating the power of scaling decoder-only architectures. These models, pre-trained with multi-task instructions, have achieved human-level performance in zero-shot and few-shot settings (Brown et al., 2020), setting a new standard for NLP. However, a critical limitation persists: the underrepresentation of low-resource languages, particularly African languages, in these advancements. This gap has motivated research into developing specialized models that address the unique challenges of low-resource linguistic contexts.

Early efforts to address this gap, such as AFRIBERTA (Ogueji et al., 2021), marked a significant step forward. AfriBERTa, a 126M-parameter encoder-only model, was pre-trained on 11 African languages and outperformed larger multilingual models like XLM-R (Conneau et al., 2020) and MBert (Devlin et al., 2019) on African language benchmarks. This success highlighted the potential of smaller, high-quality models tailored to low-resource languages, challenging the assumption that larger models are always superior. However, AfriBERTa's encoder-only architecture limited its applicability to generative tasks, leaving a gap for decoder-based models that could better handle tasks like text generation and diacritization.

Further advancements by (Hedderich et al., 2020) and (Alabi et al., 2022) explored fine-tuning and adaptation techniques for African languages. While (Hedderich et al., 2020) focused on single-language adaptation, (Alabi et al., 2022) introduced Multi-Language Adaptation Fine-Tuning (MAFT), which extended adaptation to multiple languages. Their work resulted in Afro-XLM-R[1], a model that outperformed AfriBERTa by leveraging techniques like non-African language token removal. Despite these improvements, these models remained encoder-based and relied on large-scale multilingual pretraining, which often dilutes the representation of low-resource languages. Recent successes in Large Language Models (LLMs) have highlighted the superiority of decoder-only architectures in various NLP tasks, necessitating re-evaluating approaches to modeling Nigerian languages. Efforts such as (Buzaaba* et al., 2024) and (Mwongela et al., 2024) have explored the decoder-only architectures for low-resourced African languages. However these models were fine-tuned or adapted from pretrained base models. Our approach considers pretraining the model entirely from scratch.

We argue that decoder-only models offer unique advantages, such as multi-task learning and emergent abilities that arise with scaling, (Wei et al., 2022). These capabilities are reflected in our model, SABIYARN, which excels at tasks it was not necessarily pre-trained on, such as inter-language translation between Nigerian languages. This underscores the potential of decoder-only architectures to better capture the linguistic intricacies and practical utility of these languages. The trend of scaling LLMs

[1] https://huggingface.co/Davlan/afro-xlmr-large

to larger parameter sizes has dominated NLP research, with larger models demonstrating improved reasoning and zero-shot capabilities. However, (Hoffmann et al., 2022) revealed that many models are under-trained relative to their compute budgets, emphasizing the need for efficient training strategies. This finding is particularly relevant for low-resource languages, where data scarcity and computational constraints make large-scale training impractical. Recent work has also shown that smaller models, when trained on carefully curated datasets, can achieve competitive performance (Abdin et al., 2024), challenging the necessity of massive models for low-resource settings. Notable data collection efforts like WURA (Oladipo et al., 2023), a publicly available high-quality dataset for African languages, that builds on mC4[2] and amounts to 19GB of African texts on various tasks, aim to tackle the problem of high-quality African data.

Despite these advancements, Nigerian languages remain severely underrepresented in NLP research. Existing models often fail to capture the linguistic and cultural nuances of these languages, limiting their practical applicability. This gap underscores the need for a targeted, resource-efficient approach that prioritizes high-quality data curation and efficient parameter utilization. Our work, SABIYARN, addresses this need by introducing a 125M-parameter decoder-only model specifically trained for Nigerian languages. By leveraging a multi-task learning framework (Zhang and Yang, 2021) and adhering to Chinchilla scaling laws, SABIYARN demonstrates that smaller, meticulously trained models can achieve remarkable performance in low-resource settings, offering a viable alternative to the prevailing trend of massive, indiscriminate scaling.

3 Methodology

This section details the development of SABIYARN-125M, including the dataset collation, processing, model architecture, and training.

3.1 Dataset Curation and Cleaning

The preparation of our datasets involved a meticulous process of collation, deduplication, task-specific tagging, and tokenization. This section outlines our methodology for ensuring the datasets were optimally structured for our multi-task learning approach.

[2]urlhttps://paperswithcode.com/dataset/mc4

The training dataset for SabiYarn was curated through a comprehensive effort that involved manually aggregating relevant data sets from sources such as Hugging face and the BBC Africa news website. The resulting dataset comprised approximately 114.7 million samples, representing 10.1 billion tokens (see Table 7 and Table 8 for data distribution), encompassing a diverse range of text data in various Nigerian languages, including the bible, news articles, social media posts, literary works, and educational resources for different NLP tasks. These tasks include: text generation, translation, sentiment and topic classification, text summarization, headline generation, text diacritization, text cleaning, instruction following and reasoning.

The text diacritization and cleaning datasets were generated by introducing random noise into a portion of the already collated data. For each character in the original data, there was a 15% probability of applying a random modification. This modification involved either inserting a random character or deleting the original character.

To ensure dataset quality and relevance, a rigorous cleaning and filtering process was applied to all collected datasets. This involved the following techniques:

- **Manual Scrutiny**: Duplicates, unwanted samples, and unreadable characters were manually identified and removed.
- **Normalization**: Text formats were standardized for consistency, including the conversion of Unicode characters to their language equivalents.
- **Quality Refinement**: Data integrity issues were addressed. This included removing data exhibiting social, gender, and sexual biases (identified during manual selection), filtering out repeated nonsensical characters using regular expressions, and excluding poor-quality samples. All sentence lengths and single-word translations were considered, while empty strings were discarded. This was a time-intensive but crucial step.

The resulting dataset is a rich and diverse corpus that captures linguistic nuances and incorporates cultural contexts specific to the target (9) Nigerian languages including English. However, the complete dataset has not yet been made publicly available.

3.2 Dataset Task Assignment

For each dataset described in the previous section, we undertook a manual review process to determine its suitability for specific NLP tasks. This critical step ensured that each dataset was appropriately matched to tasks such as translation, sentiment classification, named entity recognition, topic classification, instruction-following and so on.

3.2.1 Task-Specific Tagging

Upon establishing the task relevance of each dataset, we implemented a unique tagging system. This system involves the use of task-specific tag pairs, designed to clearly demarcate the input and output segments of each data sample. The tagging process follows this structure:

- A unique start tag is prepended to the input text segment.
- A corresponding end tag is appended after the input text, followed by the output text.

For instance, in a sentiment classification task:

`<classify>I love rice!<sentiment> positive`

Here, `<classify>` and `<sentiment>` are the task-specific tags, "I love rice!" is the input text, and "positive" is the output text. Other tags can be seen in Table 9

3.2.2 Rationale for Tagging

This tagging approach serves several crucial purposes.

1. **Task Identification**: It allows the model to identify the specific NLP task associated with each input during training and inference.
2. **Input-Output Demarcation**: It clearly separates the input text from the expected output, facilitating more effective learning of the input-output relationship through focused loss computation.
3. **Multi-Task Learning**: Using consistent tagging for different tasks, we enable the model to learn multiple tasks within a unified framework.

3.3 Tokenization

SabiYarn-125M utilizes the Bloom tokenizer, a BPE tokenizer pretrained on a curated dataset to effectively handle the linguistic nuances and diacritics of 9 Nigerian languages. Informed by the vocabulary sizes of GPT-2 and Mistral v3 tokenizers, and considering the training corpus's linguistic diversity, we established a vocabulary size of 52,050 tokens. A vocabulary size of 52k was chosen to achieve a compromise between adequate coverage across 9 languages and practical compute/memory limitations. This decision is supported by the findings of (Dagan et al., 2024), who suggest that increasing vocabulary size, and consequently decreasing sequence length, may lead to diminished performance as a result of reduced FLOPS efficiency during training. Task-specific tags were incorporated as special tokens during tokenizer training.

The trained tokenizer was subsequently used to tokenize the cleaned training data into a stream of token ID sequences, which were stored in a binary file in `uint8` format. During this process, a validation set comprising approximately 6 million tokens was generated by random sampling and stored in a separate binary file.

3.4 Model Architecture

SabiYarn-125M is a 125-million-parameter language model based on the Generative Pre-trained Transformer J (GPT-J) architecture. To enhance generalization, particularly in low-resource settings, we extend the attention module's output vectors with additional information via a feedforward network in each transformer block following the design used in GPT-J[3] (see comparison in Fig 1). However, we employed a trainable positional embedding layer unlike the rotary embedding layer seen in GPT-J's architecture. This choice was motivated by the hypothesis that trainable embeddings could offer greater flexibility in learning positional relationships within a smaller parameter space, potentially leading to faster convergence and improved performance compared to fixed rotary embeddings at this scale. We believe that this design enables the model to handle a wide range of NLP tasks with limited data. See Table 1 for specific details.

The model features 12 layers, 12 attention heads, an embedding size of 768, and a context length of 1024, and employs learned positional embeddings, optimizing its learning capacity. These specifications align with the GPT-2 medium model.

[3] https://www.eleuther.ai/artifacts/gpt-j

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Context Length	Learning Rate
SabiYarn-125M	125M	12	768	12	64	1024	6.0×10^{-5}

Table 1: SabiYarn Model Specifications

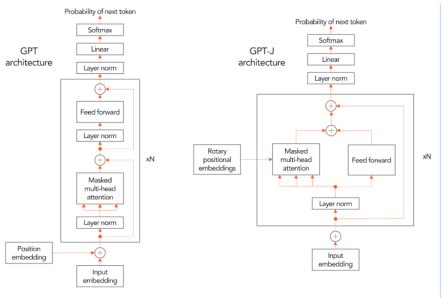

Figure 1: GPT-J architecture

3.5 Pretraining

SabiYarn-125M was pre-trained using causal language modeling with a multitask objective on a diverse, multilingual Nigerian corpus. This joint training enriches shared linguistic representations, improving next-word prediction and generalizing across tasks and languages. By increasing effective training data size and diversity (crucial for underrepresented languages), the model develops stronger token representations, enhancing language understanding and prediction. This multitask framework yields transferable and effective representations for various NLP applications, boosting performance and versatility. Table 8 presents the token distribution per language and task.

During model pretraining, we implemented a comprehensive masking strategy to prevent information leakage and ensure robust learning. Our approach consisted of two key components:

1. **Task dependent, token-level masking:** During training, when processing packed token sequences, a custom mask is applied for next-token prediction. If a sequence includes task-specific tags (e.g., for translation or NER, detailed in table 9), all tokens between these paired tags (representing the typical input) are masked out when calculating the cross-entropy loss on the shifted target sequence. This strategy trains the model to predict not only subsequent tokens generally but also to generate the correct output sequence conditioned on the presence of a downstream task and its corresponding input. This masking mechanism is illustrated in Fig 2).

2. **Sequence Packing:** We isolated attention calculations to prevent information leakage between different data samples in a batch, ensuring that each sample's processing remained independent.

As visualized in Figure 2, this dual masking approach created a more challenging learning environment that encourages the model to develop a genuine understanding of linguistic patterns rather than relying on shortcuts or memorization. Using this technique, we significantly improved the model's ability to learn task-specific features and generalize to unseen data.

Figure 2: Masking during loss computation

The model was trained on a single 24GB GPU, token ID sequences of length 1,024 (block size) were randomly sampled from the binary file to form batches of size 12. A gradient accumulation step of 40 was used, resulting in an effective token batch size of 406000 tokens, in conjunction with a cosine learning rate scheduler with a maximum learning rate of 6×10^{-4} and a minimum learning rate of 6×10^{-5}. Training was carried out with precision bfloat16 to optimize memory usage and accelerate training without compromising quality.

4 Evaluation and Results

4.1 Evaluation Methodology

The performance of SabiYarn-125M was evaluated across a spectrum of NLP tasks relevant to the Nigerian linguistic landscape. To ensure a comprehensive and reproducible assessment, we adopted the benchmark datasets and tasks used by (Ojo et al., 2023), including Translation, News Classification, Named Entity Recognition (NER), Sentiment Analysis, Text Diacritization, and Text Cleaning. These datasets, MASAKHANEWS (Adelani et al., 2023) for news classification, AFRISENTI (Muhammad et al., 2023) for sentiment analysis,

and MASAKHANER(Adelani et al., 2021) for named entity recognition, provide a robust framework for assessing the model's capabilities across diverse African languages. By adhering to these established benchmarks, we facilitate a fair and meaningful comparison between SabiYarn-125M and existing state-of-the-art language models.

4.2 Fine-tuning

In addition to evaluating the base pre-trained model, we fine-tuned SabiYarn-125M on the training sets of the benchmark datasets mentioned above. This process yielded several specialized models, each designed to excel in specific NLP tasks:

- **SabiYarn-finetune**: Fine-tuned on the aggregated training sets of all benchmark datasets, encompassing all four Nigerian languages (Yoruba, Igbo, Hausa, and Pidgin) and including back-translation data.
- **SabiYarn-translate**: Optimized for translation tasks, fine-tuned on the benchmark translation dataset and its corresponding back-translations across all languages.
- **SabiYarn-topic**: Optimized for topic classification, fine-tuned on the combined multilingual topic classification dataset.
- **SabiYarn-sentiment**: Optimized for sentiment analysis, fine-tuned on the aggregated sentiment classification dataset across all languages.
- **SabiYarn-NER**: Optimized for Named Entity Recognition, fine-tuned on the combined NER dataset spanning all languages.
- **SabiYarn-diacritics-cleaner**: Optimized for text diacritization and cleaning across all languages.

It should be noted that our approach diverges from that of M2M-100, which employed separate fine-tuning processes for each language and translation direction. We adopted a unified fine-tuning strategy across languages, a method employed in various multilingual models. To create the dataset for text diacritization and cleaning fine tuning tasks, we utilized pre-existing datasets and applied custom transformations. For diacritization, we selectively removed diacritical marks with a 50% - 100% probability, creating pairs of original and diacritic-free text. For text cleaning, we introduced controlled noise to the text, simulating common errors and inconsistencies found in real-world data. The resulting datasets were split into train, validation, and test sets, with 15,000, 1,000, and 5,000 samples respectively for each language and task.

5 Results and Discussion

The subsequent sections provide a detailed analysis of the performance of SabiYarn-125M across the evaluated tasks. We present comparative results against existing models and discuss the implications of our findings for low-resource language processing in the African context.

5.1 Task-specific Performance

Translation: SabiYarn, despite its significantly smaller size (125M parameters), demonstrates competitive performance in machine translation tasks, particularly excelling in forward translation for Igbo and pidgin and backward translation for Yoruba. While larger models like mT0-MT (13B) and M2M-100 (418M) achieve higher scores in several categories, SabiYarn's performance is remarkable considering its parameter efficiency. The model's strong performance in Nigerian Pidgin (Pcm) translation, outperforming many larger models, highlights its effectiveness in handling this unique linguistic context. However, the reliability of the evaluation is somewhat constrained by the benchmark dataset's use of only a single reference translation per source sentence. This is particularly limiting for Nigerian languages such as Yoruba, where multiple valid translations are often possible, potentially underestimating the models' true capabilities. Additionally, SabiYarn's tendency to avoid verbosity and its occasional struggle with coherence during translation present areas for future improvement, suggesting that refining the model's ability to balance conciseness with contextual understanding could further enhance its performance.

Sentiment Analysis: SabiYarn, with only 125M parameters, demonstrates impressive performance in sentiment analysis across Nigerian languages, achieving average accuracies of 66.0% (SabiYarn-Sentiment) and 65.3% (SabiYarn-Finetune). While AfroXLMR-Large (550M parameters) leads in most categories as seen in Table 4, SabiYarn consistently outperforms larger models like GPT4 and

Task	avg	Yor	Hau	Ibo	Pcm
Text Diacritization	96.9	**100.0**	-	93.7	-
Text Cleaning	71.3	77.83	54.67	**81.54**	71.17

Table 2: **Text Diacritization and Cleaning Results**. We show the **BLEU** score of **SabiYarn-diacritics-cleaner**.

Model Name	Size	avg	Yor	Hau	Ibo	Pcm
xx-en						
SabiYarn-Translate	125M	40.9	31.2	32.3	46.4	54.9
SabiYarn-Finetune	125M	41.1	29.1	34.4	46.0	54.9
M2M-100	418M	38.3	35.1	35.1	46.1	36.7
mT0	13B	36.0	35.7	32.0	31.2	44.9
mT0-MT	13B	**45.7**	**40.8**	**38.1**	**46.8**	56.9
GPT4	-	27.2	13.6	14.7	21.8	**58.8**
Llama2	13B	29.0	20.8	17.4	23.1	54.8
en-xx						
SabiYarn-Translate	125M	41.3	34.8	31.6	43.3	55.4
SabiYarn-Finetune	125M	41.4	34.4	30.72	42.3	58.0
M2M-100	418M	**48.3**	**35.9**	**43.3**	**50.0**	**64.0**
mT0	13B	19.9	6.3	15.4	23.5	34.2
mT0-MT	13B	31.3	15.2	23.11	38.5	48.3
GPT4	-	35.8	18.1	36.1	35.7	53.4
Llama2	13B	15.7	10.4	14.7	16.3	21.4

Table 3: **Machine Translation Results**: Comparison of ChrF score of SABIYARN and results obtained from Jessica et al. (2023)

Llama2 (13B parameters) across all languages. Notably, SabiYarn-Finetune surpasses AfroXLMR-Large in Nigerian Pidgin (Pcm), highlighting its effectiveness in low-resource languages. The consistent performance of the model in various Nigerian languages (63.6% to 66.8%) emphasizes its robustness and efficiency in handling multilingual sentiment analysis tasks with significantly fewer parameters.

News Classification: In news classification (Table 5), SabiYarn-Topic showcases remarkable performance with an average F1 score of 87. 03%. This is particularly impressive when compared to much larger models like mT0 (41.6%) and GPT4 (55.45%). SabiYarn even outperforms the larger AfroXLMR-Large model in Nigerian Pidgin (pcm) with a score of 96.3%. This demonstrates SabiYarn's strong capability in understanding and categorizing news content in Nigerian languages, despite its smaller size.

Named Entity Recognition: SabiYarn-Finetune achieves the highest F1 score of 93.4, outperforming all other models, including the larger AfroXLMR-Large (550M) and prompting-based LLMs like GPT-4 and Llama2 (Table 6). In contrast, larger models like mT0 and mT0-MT fail to perform well in this task, scoring 0.0, while GPT-4 and Llama2 achieve modest results of 55.6 and 17.8, respectively. This may underscore the limitations of prompting-based methods for NER tasks compared to specialized fine-tuned models such as SabiYarn.

Text Diacritization: The results for text diacritization, as shown in Table 2, demonstrate the model's strong performance in this task. SabiYarn-diacritics-cleaner model achieved a perfect BLEU score of 100.0 for Yoruba and a high score of 93.7 for Igbo. These results indicate the model's exceptional ability to accurately restore diacritical marks, particularly in Yoruba text, and its strong performance in Igbo, suggesting its potential for improving text processing in these languages.

Text Cleaning: As seen in Table 2, The model achieves the highest BLEU score of 77.83 for Yoruba, indicating strong performance in this language. However, performance varies significantly between languages, with Hausa scoring the lowest at 54.67, probably due to the lack of diacritics in this language, suggesting room for improvement in handling linguistic diversity and complexity.

6 Conclusion

Although originally trained in Nigerian languages, SabiYarn-125M represents a significant advancement in the field of natural language processing (NLP) for languages with limited data. By encompassing a diverse range of languages and offering a comprehensive suite of NLP functionalities, this model establishes a robust foundation for the potential transformation of language technology not only in Nigeria but across the African continent, thus making a substantial contribution to the global NLP community.

The development of SabiYarn-125M is driven by several key objectives:

1. **Empowering Researchers:** This model serves as a versatile foundation for future research and development, facilitating the creation of more culturally relevant and impactful language technologies.

2. **Addressing Linguistic Diversity:** By supporting multiple Nigerian languages, SabiYarn-125M tackles the unique challenges posed by Africa's rich linguistic landscape.

Model Name	Size	avg	Yor	Hau	Ibo	Pcm
SabiYarn-Sentiment	125M	66.0	65.0	66.1	66.0	66.0
SabiYarn-Finetune	125M	65.3	64.8	66.0	63.6	**66.8**
AfroXLMR-Large	550M	**75.0**	**74.1**	**80.7**	**79.5**	**68.7**
Prompting of LLMs						
mT0	13B	41.6	35.6	40.5	26.7	63.6
mT0-MT	13B	34.4	23.7	36.1	27.2	50.7
GPT4	-	55.0	55.6	41.8	66.7	57.7
Llama2	13B	27.8	24.0	25.5	35.1	24.3

Table 4: **Sentiment Analysis Results**: Comparison of Accuracy score of SABIYARN and results obtained from (Ojo et al., 2023)

Model Name	Size	avg	Yor	Hau	Ibo	Pcm
SabiYarn-Topic	125M	90.9	89.0	**90.2**	87.7	96.7
SabiYarn-Finetune	125M	87.03	84.4	82.1	85.3	67.8
AfroXLMR-Large	550M	**92.95**	**94.0**	92.2	**93.4**	92.1
AfriTeVa-V2	428M	91.2	92.3	89.4	86.1	**96.8**
Prompting of LLMs						
mT0	13B	41.6	35.6	40.5	26.7	63.6
mT0-MT	13B	34.4	23.7	36.1	27.2	50.7
GPT4	-	55.45	55.6	41.8	66.7	57.7
Llama2	13B	27.22	24.0	25.5	35.1	24.3

Table 5: **News Classification Results** We compare the F1-score of **SabiYarn** with that of the current **SOTA** models.

Model Name	Size	avg
SabiYarn-NER	125M	93.2
SabiYarn-Finetune	125M	**93.4**
AfroXLMR-Large	550M	84.6
Prompting of LLMs		
mT0	13B	0.0
mT0-MT	13B	0.0
GPT4	-	55.6
Llama2	13B	17.8

Table 6: **Named Entity Recognition Results**: We compare the F1 score of SABIYARN with results obtained from (Ojo et al., 2023).

3. **Enhancing NLP Capabilities:** The model's wide array of functionalities paves the way for advanced applications in machine translation, sentiment analysis, named entity recognition, and beyond.

Looking ahead, SabiYarn-125M opens up numerous avenues for future research:

- **Expansion to Additional Languages:** Future iterations could incorporate more African languages, further enhancing the model's versatility and impact.
- **Domain-Specific Adaptations:** Researchers could fine-tune newer versions of the model for specific domains such as healthcare, education, or legal applications, tailoring it to address sector-specific challenges.
- **Cross-Lingual Transfer Learning:** Investigating the model's capacity for cross-language fine-tuning across related African languages could yield valuable insights for low-resource language processing.

In conclusion, SabiYarn-125M represents a significant step towards bridging the gap in NLP research and technology for underrepresented languages. By showcasing the model's capabilities and potential applications, we hope to inspire and encourage further advancements in this field, ultimately contributing to the preservation and empowerment of Africa's rich linguistic heritage in this digital age and a more inclusive and equitable global language technology ecosystem.

Limitations

The scope of our evaluation was necessarily limited to the aforementioned Nigerian languages due to two critical constraints: the acute scarcity of high-quality, diverse datasets for African languages, and the limited availability of substantial computational resources. These limitations not only underscore the challenges inherent in low-resource language research but also highlight a systemic issue in the field of artificial intelligence as it pertains to linguistically diverse regions. The paucity of com-

prehensive datasets and the computational divide present significant barriers to advancing NLP capabilities across the African continent. This situation urgently calls for a multi-faceted approach: increased investment in data collection and curation for African languages, enhanced allocation of computational resources for research in these areas, and a concerted effort to build local AI research capacity. Addressing these challenges is crucial not only for advancing NLP technologies in the region but also for ensuring that the benefits of AI are equitably distributed across diverse linguistic communities. Future research must prioritize these areas to foster a more inclusive and representative landscape in global NLP development.

Acknowledgments

We extend our sincere gratitude to all who contributed to this research. First, we thank our academic advisors for their invaluable guidance and feedback, and the Nigerian/African language experts for their expertise in curating and validating the benchmark datasets. We also acknowledge the open-source community, particularly the developers of tools like PyTorch and Accelerate, which were instrumental to our work. Additionally, we are grateful to our colleagues in Natural Language Processing and African language research for laying the groundwork that inspired this study.

Finally, we express profound appreciation to our families, friends, and loved ones for their unwavering support and encouragement throughout this journey. While many have contributed to this work, any errors or oversights remain our responsibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan

Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages. *Preprint*, arXiv:2103.11811.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages. *Preprint*, arXiv:2304.09972.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Happy Buzaaba*, Alexander Wettig*, David Ifeoluwa Adelani, and Christiane Fellbaum. 2024. Model card for afrollama. `https://huggingface.co/Masakhane/afro-llama`. Accessed: 2025-03-06.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. *Preprint*, arXiv:2402.01035.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. *Preprint*, arXiv:2010.03179.

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. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Mario Michael Krell, Matej Kosec, Sergio P. Perez, Mrinal Iyer, and Andrew W Fitzgibbon. 2022. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id

Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *Preprint*, arXiv:2302.08956.

Stanslaus Mwongela, Jay Patel, Sathy Rajasekharan, Lyvia Lusiji, Francesco Piccino, Mfoniso Ukwak, and Ellen Sebastian. 2024. Afrollama 3. Accessed: 2025-03-06.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

A Appendix

A.1 HuggingFace Datasets Used (Non-exhaustive)

- ayymen/Pontoon-Translations
- igbo_ner
- chymaks/Igbo_ner
- yoruba_wordsim353
- yoruba_gv_ner
- yoruba_bbc_topics
- HausaNLP/HausaVG
- hausa_voa_ner
- Arnold/hausa_common_voice
- moro23/hausa_ng
- vpetukhov/bible_tts_hausa
- mangaphd/hausaBERTdatatrain
- masakhane/masakhaner2
- masakhane/afriqa
- sbmaruf/forai_ml_masakhane_mafand
- masakhane/afriqa-gold-passages
- HausaNLP/HausaVQA
- masakhane/mafand
- mozilla-foundation/common_voice_12_0

Table 7: Detailed breakdown of number of samples per language per task

Language	Trans	Class	Sum	Headline	Instruct	Monolingual	Diac	Clean	Total
English	0	32,536	0	53	2,243,235	11,097,016	0	6,869,858	20,242,698
Yoruba	729,878	21,572	15,572	56,564	5,016,319	11,673,886	3,602,690	381,418	21,497,899
Hausa	2,619,081	29,171	17,721	320,945	3,435,696	11,797,952	0	2,579,220	20,799,786
Igbo	6,377,666	30,265	41,303	162,148	4,967,183	14,376,298	3,393,839	4,056,681	33,405,383
Pidgin	8,988,159	12,087	145,815	289,865	512,816	4,564,139	0	3,650,887	18,163,768
Urhobo	129,668	0	162,970	0	198	32,711	0	0	325,547
Fulfulde	0	0	0	0	0	126,000	0	0	126,000
Fulah	4,018	0	751	0	2,526	134,968	0	0	142,263
Efik	0	0	0	0	0	9,567	0	0	9,567
Total	18,848,470	125,631	384,132	829,575	16,177,973	53,812,537	6,996,529	17,538,064	114,712,911

Table 8: Detailed breakdown of number of tokens per language per task

Language	Trans	Class	Sum	Headline	Instruct	Monolingual	Diac	Clean	Total
English	-	295,542	-	100,612	290,387,169	493,068,409	-	650,222,243	1,434,073,975
Yoruba	115,188,816	603,945	9,030,269	14,666,102	646,467,834	834,162,538	242,229,137	135,178, 107	1,997,526,748
Hausa	250,989,822	617,583	8,761,900	63,429,798	386,761,152	1,186,571,221	-	483,250,638	2,380,382,114
Igbo	609,811,051	530,117	18,338,768	24,977,727	646,811,000	751,549,672	161,485,201	533,371,421	2,746,874,957
Pidgin	298,282,535	276,284	95,768,701	52,113,111	112,421,167	308,031,286	-	580,618,346	1,447,511,430
Urhobo	6,451,518	-	97,198,864	-	56,358	893,162	-	-	104,599,902
Fulfulde	-	-	-	-	-	3,677,103	-	-	3,677,103
Fulah	286,795	-	436,133	-	1,069,200	9,953,441	-	-	11,745,569
Efik	-	-	-	-	-	139,740	-	-	139,740
Total	1,281,010,537	2,323,471	229,534,635	155,287,350	2,083,973,880	3,588,046,572	403,714,338	2,382,640,755	10,126,531,538

- mc4
- cyanic-selkie/wikianc
- HausaNLP/afrisenti-lid-data
- HausaNLP/NaijaSenti-Twitter
- HausaNLP/Naija-Lex
- wikimedia/wikipedia
- gsarti/flores_101
- opus100
- castorini/africlirmatrix
- severo/flores_101
- graelo/wikipedia
- igbo_english_machine_translation
- iamwille/igbo-translation
- castorini/wura
- csebuetnlp/xlsum
- cis-lmu/GlotStoryBook
- wili_2018
- cis-lmu/Glot500
- google/fleurs
- google/xtreme_s
- masakhane/masakhanews
- masakhane/afriqa_wiki_en_fr_100
- bigscience/xP3all
- CohereForAI/aya_collection_language_split
- udhr
- mtek2000/yoruba_newsclass_topic
- mxronga/cultura-x-deduped-yoruba
- mxronga/yoruba-proverbs-parallel-corpora
- ayymen/Weblate-Translations

article booktabs caption

B Previous Section

Some text before the tables. This is to demonstrate the spacing.

Table 9: Task-specific tags used for multi-task training

Task	**Start Tag**	**End Tag**
Translation	<translate>	<lang>:
Sentiment Classification	<classify>	<sentiment>:
Topic Classification	<classify>	<topic>:
Instruction Following	<prompt>	<response>:
Headline Generation	<title>	<headline>:
Text Diacritization	<diacritize>	<lang>:
Question Generation	<prompt>	<response>:
Question-Answering	<prompt>	<response>:
Text Summarization	<summarize>	<summary>:
Text Cleaning	<clean>	<lang>:

Table 10: Language tags used for multi-lingual training

Language	**Tag**
Yoruba	<yor>
Hausa	<hau>
Igbo	<ibo>
English	<eng>
Urhobo	<urh>
Fulah	<ful>
Efik	<efi>
Nigerian Pidgin	<pcm>