Sample-Size Scaling of the African Languages NLI Evaluation

Anuj Tiwari[1], Oluwapelumi Ogunremu[2], Terry Oko-odion[3], Jesujuwon Egbewale[4], Hannah Nwokocha[5]
Noida Institute of Engineering and Technology[1], ML Collective[1,2,3,4,5]
aj11anuj123@gmail.com, ogunremuoluwapelumi@gmail.com, terryokoodion@gmail.com, egbewalejesujuwon7@gmail.com, hannahsopuruchi@gmail.com

Abstract

African languages have very little labelled data, and it is unclear if augmenting the quantity of annotation data reliably enhances downstream performance. The study is a systematic sample-size scaling study of natural language inference (NLI) on 16 African languages based on the AfriXNLI benchmark. Under controlled conditions, two multilingual transformer models with roughly 0.6B parameters XLM-R Large fine-tuned on XNLI and AfroXLM-R Large are tested on sample sizes of between 50 and 500 labeled examples and average their results across random subsampling runs. As opposed to the usual belief of monotonic increase with increased data, we find a strongly language-sensitive and often non-monotonic scaling behavior. Some languages show early saturation or decrease in performance with sample size as well as high variance in low resource regimes. These results indicate that the volume of data is not enough to guarantee stable profits to African NLI, creating the necessity of language-sensitive datasets creation and stronger multilingual modelling strategies.

1 Introduction

The latest advancements in NLP have been fuelled by massive pretraining and access to large amounts of labeled data. The advances have however benefited more high resource languages unfairly and many African languages are still underexamined in training and evaluation standards. Therefore, the key problem of multilingual and low-resource NLP is to learn the performance of the models in relation to the amount of labeled data available. One of the most widely used assumptions in machine learning is that as more and more data is labelled, the better the performance will be downstream. Although this assumption is usually true in high-resource environments, it has not yet been carefully studied in the case of the low-resource languages especially the African languages with various typological and morphological characteristics. Practically, annotation is expensive, and expansion of dataset without a clear indication of the benefit can be inefficient or even counterproductive.

This paper provides an analysis of the behavior of natural language inference (NLI) with respect to the amount of labeled data in African language models by using the AfriXNLI benchmark. Rather than proposing new models or datasets, our objective is to empirically characterize scaling behavior, performance stability, and variance across languages and models under controlled experimental conditions. In particular, we answer the following research questions:

- **RQ1**: Does larger labeled data positively affect NLI performance when using African languages in AfriXNLI?
- **RQ2**: How does scaling behavior vary across languages and models?
- **RQ3**: To what extent are observed trends stable under random subsampling?

By answering these questions, we hope to offer empirical recommendations to dataset construction and evaluation practice in African NLP, as well as requirements on expectations of data scaling in low-resource semantic reasoning problems.

2 Related Work

Multilingual and African NLP Benchmarks The current research has contributed to the development of African-language NLP by extending the current standards and developing new assessment tools. AfriXNLI is a human-translated version of the XNLI benchmark of various African languages, allowing to evaluate the human natural language inference in low-resource conditions (Community, 2024) in a unified way.

MasakhaNER offers a named entity recognition system on a large scale on ten African languages, and this project proves that community-driven sets construction are effective in African NLP (Adelani et al., 2021). AfroLID presents neural language identification toolkit, spanning 517 languages in Africa, and greatly increasing the coverage of languages compared to the previous multi-lingual systems (Adebara et al., 2022). These combined efforts spell out long-term development towards determining assessment materials of the African languages in the context of multilingual NLP.

Scaling Data Laws and Efficiency In high-resource settings, language models exhibit predictable scaling behavior. (Kaplan et al., 2020) demonstrate that language modeling loss reduces according to power-law dependencies on both model size and dataset size. (Hoffmann et al., 2022) also show that compute-optimal training depends in proportionately more data, as the Chinchilla model can outperform much larger models trained on smaller data. (Muennighoff et al., 2023) however establish that performance improvements reduce quickly in data-constrained environments, and more compute or repeated data produces only small increases in performance. These results cast some doubts on the fact that classical scaling laws can be directly applied to low-resource and multilingual settings.

Data Scaling in Low-Resource NLP (Eiselen and Gaustad, 2023) look into the impact of the training data size on performance in African languages with particular attention to the morphologically diverse languages of South Africa. They demonstrate that although small data sets can be used to obtain useful models, languages with complex, conjunctive morphology need considerably more data to give similar performance. The importance of linguistic structure in relation to data efficiency is brought out in this work. Nonetheless, they only tested embedding-based models and problems like part-of-speech tagging and they pose the open question of behavior with data scaling in semantic reasoning problems and contemporary fine-tuned pretrained language models.

African Languages: Scarcity of Data and Benchmarking Systemic under-representation (Hussen et al., 2025) report that today only a tiny share of the 2000+ languages of Africa have been trained on modern large language models, and that the field of African languages has been far under-represented 15 compared to its representation across the world. (Adebara and Abdul-Mageed, 2022) attributes such scarcity to the fact that African languages are structurally unsupported by the current large language model development, and are significantly underrepresented relative to their global distribution. The latest benchmark projects like AfroBench (Ojo et al., 2025) and IrokoBench (Adelani et al., 2025) extend assessment to African languages and task categories, like reasoning and natural language understanding. Even with this extended coverage, it is evident that these benchmarks always indicate significant performance differences between African and high-resource languages, and that there are still continued issues in modeling and evaluation.

Multilingual Representation Models Multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), which have training based on pretraining on multiple languages, are commonly used as multilingual NLP baselines. (Conneau et al., 2020) demonstrate that multilingual pretraining significantly improves the cross-lingual test of XLM-R, especially on low-resource languages. Although these models can be shown to be effective in zero-shot transfer, their response to an incremental scale of data of a single African language has not been thoroughly explored. Our study supplements this literature by giving an empirical examination of sample-size scaling action of African-languages NLI.

3 Experimental Setup

3.1 Dataset

We use the AfriXNLI benchmark, which consists of the sentence pairs of NLI translated to various African languages. We use 16 languages in our experiments which adopt a variety of language families, scripts, and typological properties which are represented in AfriXNLI. All tests are performed on the test splits. Simulating the various labeled data regimes, we adjust the number of test examples to be evaluated by randomly subsampling, but we do not adjust the model parameters.

3.2 Models

To allow us to compare pretraining strategies, we assess two multilingual transformer models from similar architectures with about 0.6 billion parameters. The first model, XLM-R Large fine-

tuned on XNLI (Davison, 2020) is a powerful task-aligned multilingual baseline constructed on the XLM-R model (Conneau et al., 2020). The second model, AfroXLM-R Large (Alabi et al., 2022) is an African-based form of XLM-R trained with more focus on African-language-based data.

With the choice of similar scale and architecture model, we factor out the influence of pretraining data composition and language coverage and reduce the impact of model size.

3.3 Evaluation

For each language–model pair, we evaluate performance at sample sizes ranging from 50 to 500 examples. To control variance due to the selection of the data, we run several random subsampling runs to calculate the mean and standard deviation between the runs of a given sample size.

The most common metric that we report is accuracy, however, we also report precision and F1-score. Such an assessment plan allows us to differentiate between systematic scaling effects and those caused by sampling.

4 Results

4.1 Evaluation Variance between Sample Sizes

We initially analyse the patterns of the evaluation variance with the change in size of the sample. The standard deviation of the accuracy is reported in Figure 1 in the aggregate form over all the languages and models. The maximum variance is in low-resource samples (50-100 examples) and is sharply decreasing with increase in the sample size that reaches a point of an average of 300 samples after which the variance is constant.

This tendency shows that small sets of evaluation produce very unstable performance estimates which are highly dependent on the specific sets of samples that one is analyzing. With an increase in sample size, the variance decreases implying that the bigger the evaluation sets, the more true model performance is likely to have been estimated.

Figure 2 also further modifies this effect by model. Both XLM-R Large and AfroXLM-R Large have the following qualitative trend: large variance at the beginning of the sample size, and this is followed by a sharp rise in the value of the sample size. Although the absolute levels of different variance are slightly different, the overall trend is maintained in all models which implies that the instability of evaluation is not particular to a given pretraining strategy.

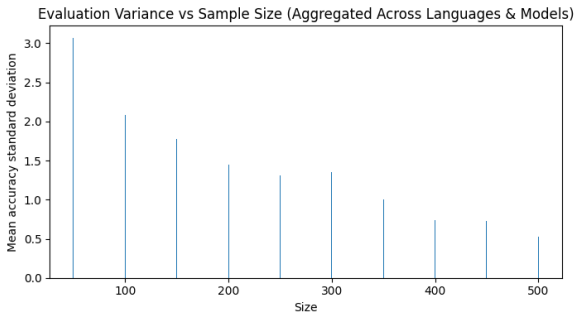

Figure 1: Evaluation variance (standard deviation of accuracy) as a function of sample size, aggregated across all languages and models. The variance decreases very rapidly with the size of the sample meaning that evaluation regimes with low resources are unstable.

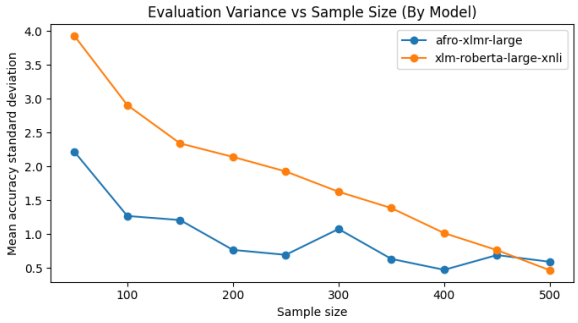

Figure 2: Comparison of evaluation variance with sample size, separately by XLM-R Large and AfroXLM-R Large. Both models show high variance in low-resource regimes and stabilize with larger evaluation sets.

4.2 Trends of Global Scaling across Languages and Models

As the sample size increases, the variance reduces, however, accuracy may not necessarily increase monotonically. Figure 3 provides a heatmap of scaling slopes of every language model pair, indicating whether the performance improves, stays constant or reduces with increased sample size.

The heatmap indicates that there is significant heterogeneity in languages. There are some nearly zero or slightly positive slopes, that is, weak gains or early saturation, and also negative slopes, that is, systematic degradation in performance with increasing evaluation sets becoming larger and more representative. Such patterns exist in both models, indicating that the scaling behaviour is highly language-specific, as opposed to being model-driven.

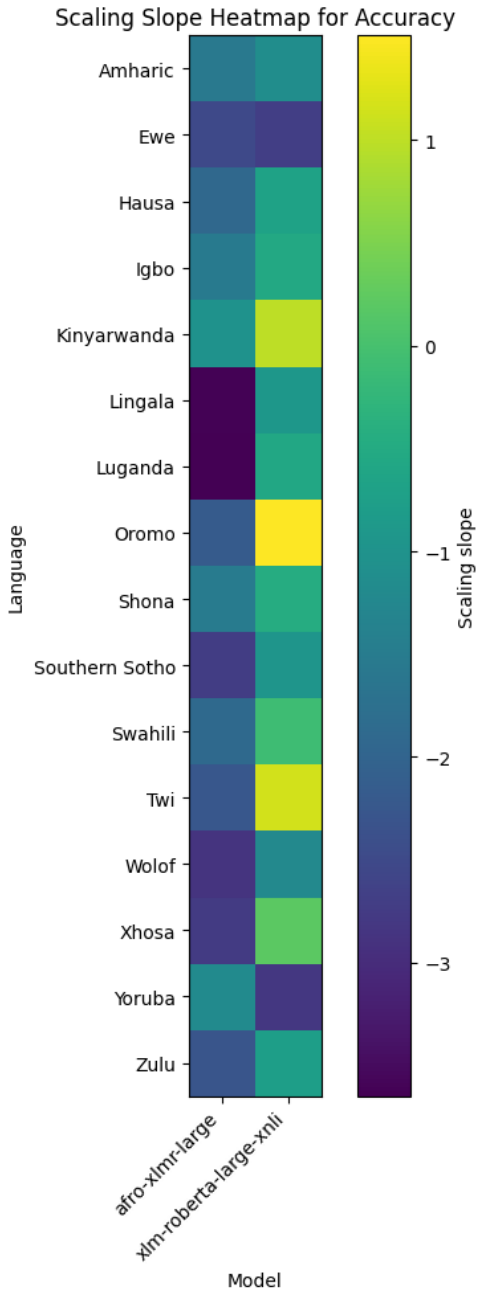

Figure 3: Scaling slope heatmap on accuracy between language-model pairs. Positive slopes mean that the performance improves as the sample size increases and negative slope means that the performance deteriorates.

4.3 Scaling Behavior Dependent on Language: Yoruba vs Kinyarwanda

To illustrate these trends concretely, we analyze scaling behavior for Yoruba and Kinyarwanda under each model. Figure 4 shows results for XLM-R Large. Yoruba exhibits pronounced small-sample optimism, with relatively high accuracy at 50 examples followed by a consistent decline as sample size increases. This monotonic degradation suggests

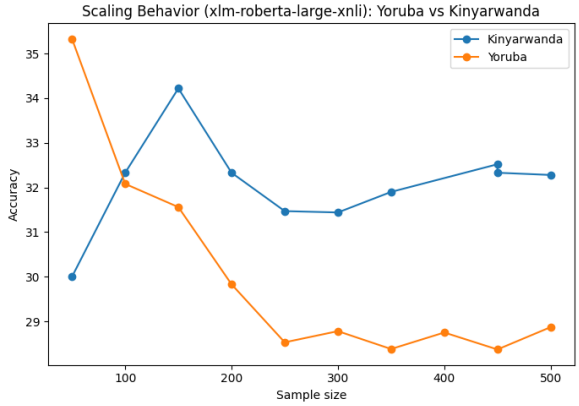

Figure 4: Yoruba and Kinyarwanda evaluation scaling behaviour with XLM-R Large. Yoruba experiences monotonic deterioration as the sample size increases and Kinyarwanda experiences initial improvement and afterwards saturation.

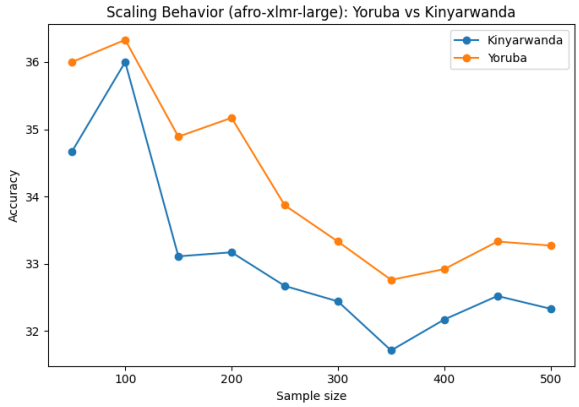

Figure 5: Yoruba and Kinyarwanda AfroXLM-R Large evaluation scaling behavior. The non-monotonic tendencies that are specific to language prevail within models.

that small evaluation subsets overestimate performance, masking systematic errors that emerge with broader coverage.

Conversely, there is a slight rise in performance of Kinyarwanda up to around 150 examples after which it starts to decrease and stabilize. At bigger sample sizes, variance collapses, meaning that it can no longer be measured with its actual performance level by the smaller subsets.

The same comparison is made in Figure 5 on AfroXLM-R Large. The qualitative trends are similar: Yoruba demonstrates again decreasing accuracy with the growing sample size at the same time Kinyarwanda demonstrates initial gains and then stabilities. The fact that such trends are maintained in the models supports the conclusion that the scaling behavior is more language-specific than model-specific.

5 Discussion

Our findings dispute the widely held belief that more uniformly augmented data is always better to use when applying African-language natural language inference. In languages, non-monotonic scaling behavior in evaluation accuracy is observed, where information other than data quantity influences it, e.g. the distribution of labels, ambiguity in translations and representativeness of evaluation subsets contribute significantly.

Notably, we come up with conclusions about AfriXNLI only and these need not be generalized as being applicable in all African languages. However, the similarity in trends observed at random subsampling runs and between two different multilingual models indicates that it is systematic, and not incidental.

Evaluation bias and small-sample optimism Among the main conclusions that this research made is that small sample evaluation is systematically biased in overestimating the performance of the models. The figure below (Figures 1 and 2) indicated that at sample sizes smaller than 200 evaluation variance is high resulting in unstable and optimistic bias accuracy estimates. With increase in sample size, both variance and accuracy collapse and this is not due to decreasing model quality but rather due to harder instances, neutral cases, and translation ambiguities displayed by larger evaluation sets that are underrepresented in small samples. This point is crucial: we are examining reliability of evaluation and not learning curves, and the apparent decline in performance indicates the lower estimation bias and not the alterations of the model behavior.

Language-specific non-monotonic scaling Figure 3, the scaling slope heatmap, indicates that there is a lot of heterogeneity between languages. Although in some languages, the slopes are weakly positive or near-zero, in others, the slopes are negative, which means that the performance of a language deteriorates as sets of evaluations increase. These patterns are present in both models and it is revealed that scaling behavior is more language dependent than model dependent. The differences between Yoruba and Kinyarwanda case studies, Figures 4 and 5, illustrate this contrast quite well: Yoruba has high small-sample optimism and decreases in a monotonic way, whereas Kinyarwanda has small gains at the start and stagnates. These variations indicate that an individual set evaluation size might provide inaccurate results when used across languages when used consistently.

Pretraining strategies Model effects Even though AfroXLM-R Large can achieve significantly better accuracy and reduced variance at very small sample sizes than XLM-R Large fine-tuned on XNLI, both models have similar qualitative scaling behaviors across languages. Africa-centric pretraining improves initial stability but does not eliminate non-monotonic scaling or language specific evaluation bias. This implies that data composition pretraining is not enough to consider the heterogeneity of African-language NLI assessment and that the choice of the model cannot influence not only stability in the evaluation but also absolute performance.

Benchmarking and evaluation implications These findings have direct methodological implications. Single scores on small test sets of accuracy can significantly exaggerate the ability of models to perform under low resource conditions. The larger sets of evaluation decrease variance and bias but can indicate lower real performance and therefore makes comparing studies more difficult. We also suggest that African language benchmarks should report variance between subsamples, should not over rely on small held-out sets, and should take into account language specific evaluation sizes instead of fixed-size test sets.

Overall, we find that the amount of data is not sufficient to ensure credible assessment on the part of African NLI. Rather, meaningful benchmarking in a low resource multilingual setting requires representative sampling, meticulous dataset construction, and stability analysis to contribute to it.

Evaluation stability and saturation In order to measure stability in evaluation, we approximate a saturation value of each language-model combination, which is the minimum size of the evaluation where the mean accuracy is varied by at most $\pm 0.5\%$. The smallest sample size at which additional increments in evaluation data do not produce significant performance variations is called the n^*.

$$n^* = \min \left\{ n \;\middle|\; \max_{m>n} |A(m) - A(n)| \leq \epsilon \right\}.$$

In case there is no $n \leq 500$, we declare saturation point as > 500. We are not training, but only testing. Here saturation refers to the amount of

Table 1: The saturation points are estimated at a sample size of ">500" at which the average error levels off within ±0.5% between the languages and models.

Language	xlm-roberta-large-xnli	afro-xlmr-large
Amharic	450	250
Ewe	250	400
Hausa	400	400
Igbo	300	250
Kinyarwanda	450	400
Lingala	300	350
Luganda	350	450
Oromo	400	400
Shona	200	400
Southern Sotho	300	350
Swahili	300	450
Twi	450	400
Wolof	>500	450
Xhosa	250	400
Yoruba	250	400
Zulu	300	400

evaluation data one only needs to achieve the stabilization of the estimated performance.

Table 1 provides the summary of these saturation points in all languages and models. We find very great language to language disparity. In some languages (e.g., Shona, Yoruba, Xhosa) as few as 200 or 250 evaluation samples can give the required stable performance estimates, whereas in others (e.g., Swahili, Kinyarwanda, Oromo) it can take 400 or 450 samples. It is worth noting that Wolof underestimates performance even in 500 samples of XLM-R Large, and this implies that Wolof remains unstable in performance estimation.

These variations are much the same across models indicating that language and dataset specific factors are the primary causes of saturation behaviour as opposed to model architecture itself. Overall, these findings indicate that model performance can be significantly misestimated using fixed-size evaluation benchmarks in the case of African languages, and that the required volume of evaluation data to make reliable estimates can differ significantly across languages.

6 Limitations

The limitations of our study are as follows:

- **Dataset scope** All experiments are conducted on AfriXNLI; thus, observed trends may reflect dataset-specific properties such as translation artifacts or label distribution biases.
- **Evaluation vs learning** We get the evaluation behavior as opposed to the dynamics of learning. Models do not optimize on successively large training sets hence results reflect the stability and bias of the performance estimates, not the improvement in performance with more training data.
- **Model scale** The scale of experiments is restricted to 2 models of multilingual size 0.6b. Relationships between data scaling and scales of models are not investigated.

Regardless of these constraints, the fact that trends were similar among languages, models, and random subsampling runs implies that we have been able to capture systematic elements of evaluation reliability to African NLI.

7 Future Work

This analysis can be developed in several ways in future work:

- **Broader tasks and datasets** The generalizability of the found evaluated scaling behaviour of the study should be tested by extending the study to other African NLP benchmark tasks, like sentiment analysis or named entity recognition.
- **Linguistic and dataset effects** Adding linguistic metadata, label distributions and tokenization statistics can be useful to explain language behavioral specifics of saturation and non-monotonic scaling.
- **Learning dynamics** Learning scaling behavior: Fine-tuning, but not evaluation, would help illuminate the effect of the addition of labeled data on real model learning with African languages.

8 Conclusion

Here we provide a detailed study of scale behavior in terms of sample-size on the AfriXNLI benchmark using African languages. Our results based on controlled evaluation in 16 languages and two multilingual models and a series of random subsampling runs demonstrate that growth in evaluation data does not come at uniform or monotonic benefits. Rather, scaling behavior is very language-specific, usually non monotonic, and hugely influenced by evaluation variation under low resource limitations.

We show that the performance estimates of small evaluation subsets are often optimistically biased, whereas the estimates of larger subsets are indicative of latent difficulty and their estimates are more stable. This demonstrates the difference between the evaluation reliability and model learning as one of the key issues in the African NLP.

On our results, we suggest: (i) do not use single-point assessment on very small test sets and report the mean $\pm$ standard deviation on more than one subsample, (ii) supplement aggregate measures with per-class measures and (iii) use at least 300 evaluation samples unless otherwise, and consider results below this scale as high-noise measures. In a broader sense, our paper warns on naive beliefs about the reliability of increased information as a means to have dependable evaluation and the need to have practices that are evaluation conscious in benchmarking of African languages.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, and 1 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, pages 1116–1131.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, and 1 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and 1 others. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning". In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Masakhane NLP Community. 2024. Afrixnli: Dataset.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Joe Davison. 2020. Xlm-roberta large fine-tuned on xnli. Hugging Face model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct south african languages. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, and 1 others. 2022. Training compute-optimal large language models. In *Proceedings of NeurIPS*.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The state of large language models for african languages: Progress and challenges. *ArXiv*, abs/2506.02280.

Jared Kaplan, Sam McCandlish, T. J. Henighan, and 1 others. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Niklas Muennighoff, Sébastien Candido, Jonathan Vamalké, and 1 others. 2023. Scaling data-constrained language models. *https://sl1nk.com/wyS64*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, and 1 others. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095. Association for Computational Linguistics.

A Appendix - Full Results for models

Table 2: Evaluation of Swahili with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 ± 1.89	0.339 ± 0.009	0.331 ± 0.014
100	34.08 ± 1.41	0.341 ± 0.014	0.340 ± 0.014
150	32.89 ± 0.31	0.332 ± 0.002	0.330 ± 0.008
200	33.33 ± 0.85	0.335 ± 0.004	0.334 ± 0.006
250	33.87 ± 1.61	0.341 ± 0.015	0.340 ± 0.016
300	32.89 ± 1.59	0.331 ± 0.015	0.330 ± 0.015
350	32.95 ± 1.52	0.332 ± 0.015	0.331 ± 0.015
400	33.25 ± 1.14	0.335 ± 0.011	0.334 ± 0.011
450	33.19 ± 0.93	0.333 ± 0.010	0.333 ± 0.009
500	33.13 ± 0.41	0.331 ± 0.004	0.331 ± 0.004

Table 3: Evaluation of Lingala with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	34.00 ± 0.00	0.346 ± 0.027	0.323 ± 0.010
100	29.67 ± 1.70	0.293 ± 0.026	0.281 ± 0.021
150	31.33 ± 0.94	0.308 ± 0.028	0.294 ± 0.014
200	30.83 ± 1.84	0.301 ± 0.030	0.284 ± 0.022
250	30.67 ± 1.91	0.302 ± 0.029	0.283 ± 0.022
300	31.89 ± 1.66	0.314 ± 0.018	0.296 ± 0.018
350	31.71 ± 1.46	0.313 ± 0.015	0.294 ± 0.015
400	31.67 ± 1.36	0.309 ± 0.015	0.293 ± 0.014
450	31.93 ± 0.90	0.311 ± 0.008	0.295 ± 0.008
500	31.87 ± 0.82	0.312 ± 0.008	0.294 ± 0.008

Table 4: Evaluation of Igbo with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	31.33 ± 1.89	0.424 ± 0.008	0.314 ± 0.026
100	32.00 ± 3.56	0.456 ± 0.011	0.314 ± 0.040
150	32.44 ± 2.06	0.458 ± 0.036	0.315 ± 0.024
200	31.33 ± 1.93	0.441 ± 0.020	0.306 ± 0.020
250	30.67 ± 1.80	0.424 ± 0.009	0.298 ± 0.019
300	29.89 ± 1.50	0.406 ± 0.015	0.299 ± 0.018
350	29.62 ± 1.28	0.408 ± 0.013	0.299 ± 0.014
400	30.08 ± 0.82	0.403 ± 0.009	0.296 ± 0.010
450	29.85 ± 0.46	0.406 ± 0.005	0.293 ± 0.005
500	30.08 ± 0.33	0.409 ± 0.005	0.294 ± 0.003

Table 5: Evaluation of Hausa with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.67 ± 2.49	0.317 ± 0.032	0.316 ± 0.026
100	32.67 ± 1.79	0.335 ± 0.007	0.327 ± 0.011
150	31.78 ± 1.37	0.317 ± 0.017	0.316 ± 0.015
200	32.80 ± 1.08	0.314 ± 0.006	0.316 ± 0.008
250	32.53 ± 2.93	0.325 ± 0.026	0.325 ± 0.027
300	32.56 ± 3.22	0.326 ± 0.029	0.325 ± 0.031
350	32.10 ± 1.94	0.321 ± 0.021	0.321 ± 0.020
400	31.42 ± 1.84	0.317 ± 0.016	0.315 ± 0.017
450	31.19 ± 1.18	0.317 ± 0.012	0.314 ± 0.012
500	31.13 ± 0.38	0.318 ± 0.006	0.314 ± 0.005

Table 6: Evaluation of Yoruba with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	35.33 ± 2.49	0.439 ± 0.060	0.342 ± 0.035
100	32.08 ± 2.83	0.421 ± 0.077	0.296 ± 0.026
150	31.56 ± 2.27	0.396 ± 0.046	0.291 ± 0.020
200	29.83 ± 1.43	0.391 ± 0.020	0.283 ± 0.021
250	28.53 ± 1.24	0.367 ± 0.010	0.268 ± 0.020
300	28.78 ± 0.68	0.359 ± 0.007	0.271 ± 0.013
350	28.38 ± 0.36	0.353 ± 0.005	0.268 ± 0.008
400	28.75 ± 0.89	0.369 ± 0.017	0.272 ± 0.011
450	28.37 ± 1.06	0.358 ± 0.018	0.269 ± 0.011
500	28.87 ± 0.52	0.358 ± 0.011	0.273 ± 0.005

Table 7: Evaluation of Kinyarwanda with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	30.00 ± 4.32	0.455 ± 0.064	0.305 ± 0.031
100	32.33 ± 3.09	0.430 ± 0.040	0.325 ± 0.023
150	34.22 ± 5.45	0.440 ± 0.040	0.337 ± 0.050
200	32.33 ± 4.50	0.420 ± 0.041	0.318 ± 0.043
250	31.47 ± 3.09	0.412 ± 0.019	0.311 ± 0.030
300	31.44 ± 1.75	0.402 ± 0.012	0.309 ± 0.018
350	31.90 ± 1.75	0.406 ± 0.015	0.312 ± 0.016
400	32.33 ± 1.45	0.415 ± 0.011	0.319 ± 0.015
450	32.52 ± 0.38	0.412 ± 0.004	0.320 ± 0.003
500	32.28 ± 0.16	0.403 ± 0.001	0.315 ± 0.003

Table 8: Evaluation of Zulu with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.08 ± 4.32	0.337 ± 0.057	0.319 ± 0.048
100	31.33 ± 2.49	0.349 ± 0.026	0.323 ± 0.029
150	33.33 ± 0.54	0.358 ± 0.011	0.342 ± 0.009
200	31.17 ± 2.09	0.359 ± 0.012	0.327 ± 0.016
250	30.88 ± 1.96	0.351 ± 0.014	0.326 ± 0.017
300	30.56 ± 2.01	0.341 ± 0.012	0.328 ± 0.017
350	30.67 ± 2.38	0.340 ± 0.018	0.321 ± 0.021
400	30.80 ± 1.47	0.338 ± 0.010	0.316 ± 0.013
450	30.07 ± 0.93	0.339 ± 0.006	0.317 ± 0.008
500	30.33 ± 0.77	0.341 ± 0.004	0.319 ± 0.006

Table 9: Evaluation of Amharic with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 ± 6.18	0.333 ± 0.058	0.332 ± 0.059
100	33.67 ± 4.11	0.333 ± 0.043	0.334 ± 0.041
150	32.08 ± 3.49	0.332 ± 0.022	0.325 ± 0.028
200	31.58 ± 2.27	0.333 ± 0.020	0.323 ± 0.021
250	32.13 ± 1.36	0.337 ± 0.086	0.329 ± 0.010
300	31.67 ± 0.82	0.327 ± 0.094	0.321 ± 0.006
350	31.85 ± 0.59	0.327 ± 0.092	0.318 ± 0.006
400	31.88 ± 0.82	0.329 ± 0.009	0.319 ± 0.008
450	30.96 ± 0.91	0.329 ± 0.010	0.319 ± 0.010
500	30.73 ± 0.52	0.327 ± 0.006	0.317 ± 0.004

Table 10: Evaluation of Southern sotho with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.00 ± 3.27	0.420 ± 0.040	0.308 ± 0.036
100	28.33 ± 4.11	0.397 ± 0.022	0.268 ± 0.027
150	29.78 ± 4.43	0.383 ± 0.093	0.278 ± 0.036
200	30.08 ± 4.14	0.399 ± 0.027	0.285 ± 0.037
250	29.73 ± 2.54	0.376 ± 0.021	0.281 ± 0.023
300	29.44 ± 2.20	0.355 ± 0.028	0.277 ± 0.021
350	29.14 ± 1.42	0.354 ± 0.028	0.275 ± 0.013
400	29.33 ± 1.25	0.357 ± 0.020	0.279 ± 0.011
450	29.48 ± 0.46	0.356 ± 0.011	0.280 ± 0.004
500	29.80 ± 0.49	0.357 ± 0.010	0.282 ± 0.004

Table 11: Evaluation of Oromo with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	24.00 ± 12.96	0.269 ± 0.118	0.246 ± 0.126
100	27.67 ± 4.64	0.333 ± 0.047	0.295 ± 0.046
150	28.44 ± 2.57	0.347 ± 0.022	0.305 ± 0.023
200	28.00 ± 2.04	0.341 ± 0.026	0.300 ± 0.022
250	28.80 ± 1.31	0.346 ± 0.011	0.308 ± 0.013
300	28.89 ± 0.83	0.342 ± 0.019	0.306 ± 0.011
350	28.76 ± 0.97	0.342 ± 0.019	0.306 ± 0.012
400	27.67 ± 0.82	0.337 ± 0.013	0.297 ± 0.010
450	28.15 ± 0.90	0.339 ± 0.011	0.301 ± 0.010
500	27.47 ± 0.34	0.333 ± 0.005	0.294 ± 0.004

Table 12: Evaluation of Twi with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	25.33 ± 4.71	0.276 ± 0.034	0.214 ± 0.034
100	27.00 ± 3.74	0.319 ± 0.064	0.225 ± 0.033
150	27.78 ± 3.94	0.349 ± 0.049	0.237 ± 0.040
200	27.00 ± 2.83	0.332 ± 0.053	0.231 ± 0.029
250	26.00 ± 2.47	0.316 ± 0.045	0.224 ± 0.023
300	26.78 ± 1.50	0.302 ± 0.031	0.232 ± 0.015
350	27.33 ± 1.66	0.322 ± 0.017	0.240 ± 0.015
400	27.08 ± 1.01	0.324 ± 0.022	0.237 ± 0.008
450	27.93 ± 0.21	0.334 ± 0.003	0.244 ± 0.001
500	28.00 ± 0.43	0.328 ± 0.014	0.245 ± 0.005

Table 13: Evaluation of Shona with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	26.00 ± 5.66	0.301 ± 0.122	0.240 ± 0.064
100	23.00 ± 4.55	0.256 ± 0.104	0.208 ± 0.050
150	25.33 ± 3.81	0.319 ± 0.063	0.237 ± 0.043
200	24.50 ± 2.27	0.315 ± 0.055	0.233 ± 0.029
250	24.53 ± 2.00	0.311 ± 0.032	0.233 ± 0.024
300	24.67 ± 1.52	0.307 ± 0.027	0.234 ± 0.018
350	24.18 ± 0.94	0.298 ± 0.013	0.228 ± 0.011
400	24.42 ± 1.30	0.314 ± 0.020	0.233 ± 0.014
450	24.81 ± 1.34	0.311 ± 0.014	0.235 ± 0.012
500	24.93 ± 0.77	0.310 ± 0.009	0.235 ± 0.008

Table 14: Evaluation of Xhosa with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	29.33 ± 2.49	0.369 ± 0.043	0.316 ± 0.031
100	29.33 ± 2.87	0.346 ± 0.027	0.311 ± 0.027
150	29.78 ± 3.62	0.354 ± 0.019	0.317 ± 0.031
200	30.33 ± 3.09	0.357 ± 0.017	0.324 ± 0.027
250	29.60 ± 2.36	0.354 ± 0.024	0.319 ± 0.024
300	29.33 ± 2.87	0.348 ± 0.028	0.315 ± 0.029
350	29.62 ± 2.58	0.345 ± 0.027	0.316 ± 0.026
400	29.92 ± 1.12	0.349 ± 0.013	0.320 ± 0.012
450	30.00 ± 0.54	0.353 ± 0.005	0.321 ± 0.003
500	29.80 ± 0.43	0.349 ± 0.007	0.319 ± 0.004

Table 15: Evaluation of Wolof with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 ± 5.25	0.396 ± 0.058	0.306 ± 0.024
100	32.67 ± 2.62	0.398 ± 0.040	0.311 ± 0.020
150	33.33 ± 0.54	0.392 ± 0.015	0.321 ± 0.008
200	32.67 ± 1.70	0.399 ± 0.021	0.317 ± 0.018
250	31.68 ± 1.50	0.389 ± 0.030	0.305 ± 0.016
300	32.80 ± 1.66	0.375 ± 0.034	0.306 ± 0.017
350	31.24 ± 1.28	0.373 ± 0.023	0.301 ± 0.013
400	31.58 ± 0.47	0.383 ± 0.008	0.304 ± 0.002
450	31.26 ± 0.38	0.379 ± 0.004	0.301 ± 0.003
500	30.53 ± 0.57	0.375 ± 0.007	0.294 ± 0.006

Table 16: Evaluation of Luganda with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.80 ± 1.63	0.470 ± 0.037	0.324 ± 0.020
100	30.08 ± 0.82	0.420 ± 0.020	0.294 ± 0.004
150	32.22 ± 0.83	0.428 ± 0.023	0.315 ± 0.005
200	32.58 ± 0.71	0.421 ± 0.013	0.325 ± 0.003
250	31.68 ± 0.86	0.402 ± 0.009	0.314 ± 0.009
300	32.22 ± 0.79	0.395 ± 0.007	0.318 ± 0.007
350	31.24 ± 0.49	0.384 ± 0.007	0.318 ± 0.005
400	31.25 ± 0.35	0.388 ± 0.006	0.311 ± 0.002
450	31.63 ± 0.28	0.395 ± 0.003	0.315 ± 0.004
500	31.53 ± 0.34	0.393 ± 0.002	0.313 ± 0.004

Table 17: Evaluation of Ewe with `xlm-roberta-large-xnli`. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 ± 3.27	0.424 ± 0.154	0.318 ± 0.049
100	33.00 ± 2.16	0.468 ± 0.042	0.284 ± 0.018
150	32.22 ± 1.26	0.417 ± 0.012	0.278 ± 0.010
200	30.08 ± 1.47	0.390 ± 0.010	0.260 ± 0.011
250	29.73 ± 1.86	0.385 ± 0.025	0.255 ± 0.015
300	30.22 ± 1.40	0.388 ± 0.020	0.260 ± 0.011
350	29.52 ± 1.52	0.392 ± 0.027	0.259 ± 0.016
400	30.00 ± 0.54	0.400 ± 0.012	0.265 ± 0.005
450	29.85 ± 0.64	0.394 ± 0.013	0.261 ± 0.004
500	29.80 ± 0.16	0.393 ± 0.009	0.257 ± 0.003

Table 18: Evaluation of Swahili with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	37.33 ± 0.94	0.184 ± 0.067	0.240 ± 0.057
100	35.33 ± 0.94	0.169 ± 0.052	0.223 ± 0.042
150	33.33 ± 0.94	0.210 ± 0.072	0.206 ± 0.040
200	34.67 ± 0.62	0.222 ± 0.082	0.219 ± 0.044
250	33.07 ± 0.19	0.202 ± 0.067	0.204 ± 0.045
300	32.89 ± 0.68	0.210 ± 0.074	0.202 ± 0.044
350	32.19 ± 0.13	0.204 ± 0.077	0.197 ± 0.045
400	32.42 ± 0.62	0.204 ± 0.077	0.198 ± 0.041
450	32.96 ± 0.55	0.207 ± 0.074	0.201 ± 0.038
500	33.00 ± 0.57	0.201 ± 0.068	0.201 ± 0.038

Table 19: Evaluation of Lingala with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	41.33 ± 3.40	0.246 ± 0.073	0.294 ± 0.060
100	37.67 ± 1.25	0.217 ± 0.061	0.257 ± 0.050
150	37.33 ± 2.72	0.216 ± 0.072	0.253 ± 0.065
200	35.83 ± 1.43	0.202 ± 0.061	0.238 ± 0.051
250	33.87 ± 0.68	0.187 ± 0.054	0.218 ± 0.039
300	33.56 ± 0.42	0.187 ± 0.053	0.215 ± 0.035
350	32.86 ± 0.62	0.183 ± 0.057	0.209 ± 0.042
400	32.42 ± 0.24	0.179 ± 0.053	0.207 ± 0.038
450	32.96 ± 0.55	0.207 ± 0.074	0.201 ± 0.038
500	33.00 ± 0.57	0.201 ± 0.068	0.201 ± 0.038

Table 20: Evaluation of Igbo with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 ± 0.94	0.164 ± 0.039	0.215 ± 0.023
100	35.67 ± 1.25	0.157 ± 0.034	0.208 ± 0.018
150	34.67 ± 0.54	0.271 ± 0.148	0.201 ± 0.031
200	34.50 ± 0.71	0.262 ± 0.137	0.201 ± 0.025
250	32.93 ± 0.19	0.254 ± 0.135	0.187 ± 0.025
300	32.89 ± 0.87	0.261 ± 0.135	0.186 ± 0.029
350	32.48 ± 0.49	0.259 ± 0.139	0.184 ± 0.031
400	32.75 ± 0.41	0.216 ± 0.084	0.187 ± 0.032
450	32.96 ± 0.42	0.217 ± 0.080	0.189 ± 0.027
500	33.13 ± 0.19	0.217 ± 0.082	0.189 ± 0.027

Table 21: Evaluation of Hausa with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 ± 0.94	0.238 ± 0.064	0.236 ± 0.030
100	35.67 ± 0.47	0.206 ± 0.054	0.223 ± 0.032
150	33.56 ± 0.31	0.192 ± 0.054	0.208 ± 0.033
200	33.67 ± 0.24	0.179 ± 0.050	0.201 ± 0.034
250	32.40 ± 0.86	0.171 ± 0.051	0.188 ± 0.037
300	31.09 ± 1.50	0.163 ± 0.046	0.185 ± 0.037
350	31.43 ± 0.62	0.157 ± 0.050	0.188 ± 0.037
400	31.83 ± 0.31	0.163 ± 0.048	0.185 ± 0.034
450	32.15 ± 1.00	0.169 ± 0.042	0.188 ± 0.028
500	32.27 ± 0.84	0.166 ± 0.040	0.188 ± 0.028

Table 22: Evaluation of Zulu with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	38.00 ± 1.63	0.148 ± 0.017	0.213 ± 0.020
100	37.00 ± 0.82	0.192 ± 0.083	0.205 ± 0.015
150	35.11 ± 1.57	0.196 ± 0.114	0.189 ± 0.023
200	34.17 ± 0.24	0.174 ± 0.081	0.181 ± 0.009
250	32.88 ± 0.86	0.166 ± 0.085	0.167 ± 0.012
300	32.22 ± 1.23	0.138 ± 0.047	0.162 ± 0.012
350	31.81 ± 0.49	0.136 ± 0.050	0.158 ± 0.008
400	32.33 ± 0.12	0.138 ± 0.048	0.162 ± 0.005
450	32.59 ± 0.69	0.159 ± 0.058	0.165 ± 0.004
500	32.73 ± 0.50	0.149 ± 0.057	0.165 ± 0.003

Table 23: Evaluation of Yoruba with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 ± 4.32	0.220 ± 0.069	0.239 ± 0.024
100	36.33 ± 1.25	0.210 ± 0.057	0.244 ± 0.041
150	34.89 ± 0.63	0.208 ± 0.066	0.228 ± 0.049
200	35.17 ± 1.03	0.201 ± 0.061	0.233 ± 0.051
250	33.87 ± 0.94	0.196 ± 0.061	0.221 ± 0.050
300	33.33 ± 0.47	0.193 ± 0.056	0.216 ± 0.042
350	32.76 ± 0.71	0.190 ± 0.062	0.214 ± 0.048
400	32.92 ± 0.51	0.188 ± 0.059	0.214 ± 0.046
450	33.33 ± 0.48	0.188 ± 0.053	0.217 ± 0.041
500	33.27 ± 0.34	0.191 ± 0.057	0.216 ± 0.038

Table 24: Evaluation of Kinyarwanda with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	34.67 ± 3.40	0.166 ± 0.041	0.219 ± 0.027
100	36.00 ± 1.63	0.280 ± 0.141	0.231 ± 0.038
150	33.11 ± 1.26	0.189 ± 0.052	0.205 ± 0.038
200	33.17 ± 1.55	0.185 ± 0.049	0.206 ± 0.033
250	32.67 ± 0.50	0.182 ± 0.051	0.200 ± 0.046
300	32.44 ± 1.50	0.172 ± 0.045	0.197 ± 0.049
350	31.71 ± 0.62	0.168 ± 0.048	0.191 ± 0.049
400	32.17 ± 0.12	0.165 ± 0.047	0.194 ± 0.046
450	32.52 ± 0.73	0.166 ± 0.041	0.196 ± 0.041
500	32.33 ± 0.82	0.162 ± 0.039	0.194 ± 0.038

Table 25: Evaluation of Amharic with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 ± 0.94	0.167 ± 0.043	0.223 ± 0.033
100	37.33 ± 1.89	0.186 ± 0.076	0.229 ± 0.050
150	34.67 ± 1.44	0.163 ± 0.068	0.203 ± 0.045
200	34.33 ± 0.85	0.161 ± 0.064	0.208 ± 0.038
250	32.80 ± 1.18	0.149 ± 0.062	0.183 ± 0.035
300	32.67 ± 1.66	0.152 ± 0.069	0.183 ± 0.038
350	32.57 ± 1.68	0.181 ± 0.064	0.183 ± 0.041
400	33.08 ± 1.24	0.192 ± 0.065	0.188 ± 0.037
450	33.04 ± 0.73	0.184 ± 0.053	0.188 ± 0.029
500	33.07 ± 0.47	0.182 ± 0.052	0.187 ± 0.027

Table 26: Evaluation of Southern sotho with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 ± 3.40	0.202 ± 0.092	0.238 ± 0.054
100	37.08 ± 0.82	0.174 ± 0.058	0.211 ± 0.022
150	34.67 ± 0.54	0.264 ± 0.150	0.191 ± 0.017
200	34.50 ± 0.71	0.261 ± 0.137	0.192 ± 0.014
250	33.33 ± 0.50	0.226 ± 0.089	0.182 ± 0.018
300	33.08 ± 0.94	0.226 ± 0.087	0.179 ± 0.019
350	32.57 ± 0.81	0.208 ± 0.075	0.174 ± 0.022
400	32.67 ± 0.42	0.184 ± 0.058	0.177 ± 0.018
450	32.89 ± 0.48	0.193 ± 0.057	0.179 ± 0.009
500	33.07 ± 0.25	0.192 ± 0.057	0.180 ± 0.011

Table 27: Evaluation of Oromo with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	38.00 ± 3.27	0.186 ± 0.038	0.230 ± 0.016
100	37.33 ± 1.25	0.287 ± 0.143	0.228 ± 0.030
150	35.56 ± 1.13	0.231 ± 0.082	0.212 ± 0.034
200	34.83 ± 0.62	0.216 ± 0.073	0.204 ± 0.030
250	33.33 ± 0.50	0.204 ± 0.073	0.190 ± 0.022
300	33.08 ± 0.54	0.186 ± 0.052	0.187 ± 0.021
350	32.19 ± 0.49	0.177 ± 0.053	0.179 ± 0.023
400	32.50 ± 0.35	0.179 ± 0.053	0.182 ± 0.025
450	32.81 ± 0.52	0.192 ± 0.058	0.185 ± 0.023
500	33.00 ± 0.33	0.193 ± 0.059	0.186 ± 0.025

Table 28: Evaluation of Twi with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	37.33 ± 0.94	0.179 ± 0.058	0.236 ± 0.049
100	35.33 ± 1.25	0.194 ± 0.048	0.238 ± 0.034
150	33.11 ± 0.83	0.188 ± 0.053	0.213 ± 0.033
200	33.33 ± 1.03	0.186 ± 0.051	0.213 ± 0.032
250	31.73 ± 1.00	0.175 ± 0.047	0.199 ± 0.030
300	31.67 ± 1.25	0.177 ± 0.046	0.197 ± 0.028
350	31.05 ± 0.75	0.171 ± 0.049	0.198 ± 0.031
400	31.83 ± 0.62	0.177 ± 0.052	0.196 ± 0.033
450	31.85 ± 1.21	0.173 ± 0.044	0.195 ± 0.025
500	32.13 ± 1.09	0.175 ± 0.046	0.196 ± 0.026

Table 29: Evaluation of Shona with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 ± 1.63	0.174 ± 0.053	0.229 ± 0.042
100	36.00 ± 0.00	0.173 ± 0.058	0.228 ± 0.050
150	33.78 ± 0.31	0.156 ± 0.058	0.208 ± 0.051
200	34.00 ± 0.00	0.155 ± 0.055	0.208 ± 0.048
250	32.67 ± 0.75	0.146 ± 0.057	0.196 ± 0.052
300	31.67 ± 1.44	0.139 ± 0.049	0.187 ± 0.044
350	31.52 ± 0.49	0.138 ± 0.054	0.186 ± 0.049
400	32.50 ± 0.54	0.144 ± 0.057	0.194 ± 0.052
450	32.67 ± 0.65	0.144 ± 0.051	0.195 ± 0.045
500	32.53 ± 0.74	0.180 ± 0.048	0.194 ± 0.049

Table 30: Evaluation of Xhosa with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 ± 2.49	0.258 ± 0.149	0.238 ± 0.028
100	37.67 ± 0.94	0.358 ± 0.156	0.214 ± 0.013
150	35.33 ± 1.44	0.347 ± 0.164	0.198 ± 0.016
200	35.00 ± 0.41	0.306 ± 0.193	0.189 ± 0.009
250	33.47 ± 1.05	0.307 ± 0.153	0.174 ± 0.014
300	32.89 ± 1.40	0.277 ± 0.146	0.169 ± 0.016
350	32.38 ± 0.71	0.260 ± 0.137	0.163 ± 0.018
400	32.83 ± 0.31	0.263 ± 0.135	0.167 ± 0.006
450	32.89 ± 0.63	0.219 ± 0.093	0.167 ± 0.007
500	33.00 ± 0.43	0.219 ± 0.092	0.168 ± 0.005

Table 31: Evaluation of Wolof with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 ± 1.89	0.236 ± 0.067	0.241 ± 0.031
100	38.08 ± 1.63	0.223 ± 0.076	0.226 ± 0.030
150	34.67 ± 1.44	0.172 ± 0.055	0.198 ± 0.028
200	34.17 ± 0.24	0.171 ± 0.040	0.194 ± 0.015
250	32.67 ± 0.50	0.166 ± 0.041	0.183 ± 0.013
300	32.11 ± 1.10	0.159 ± 0.034	0.177 ± 0.006
350	31.81 ± 0.13	0.160 ± 0.041	0.174 ± 0.014
400	32.17 ± 0.31	0.160 ± 0.040	0.177 ± 0.015
450	32.44 ± 1.01	0.163 ± 0.040	0.179 ± 0.013
500	32.73 ± 0.81	0.170 ± 0.044	0.182 ± 0.015

Table 32: Evaluation of Luganda with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	41.33 ± 3.40	0.298 ± 0.140	0.282 ± 0.052
100	35.67 ± 2.05	0.219 ± 0.072	0.235 ± 0.031
150	34.22 ± 2.45	0.210 ± 0.075	0.222 ± 0.035
200	33.33 ± 1.31	0.199 ± 0.065	0.213 ± 0.030
250	32.67 ± 1.05	0.196 ± 0.067	0.205 ± 0.030
300	32.08 ± 1.36	0.184 ± 0.052	0.198 ± 0.026
350	31.43 ± 0.62	0.176 ± 0.053	0.192 ± 0.032
400	31.67 ± 0.82	0.176 ± 0.052	0.194 ± 0.031
450	32.44 ± 0.79	0.181 ± 0.049	0.201 ± 0.033
500	32.93 ± 0.66	0.183 ± 0.051	0.205 ± 0.038

Table 33: Evaluation of Ewe with afro-xlmr-large. Measures are provided of means ± standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 ± 1.89	0.233 ± 0.070	0.260 ± 0.042
100	38.00 ± 2.83	0.226 ± 0.071	0.252 ± 0.051
150	36.44 ± 1.75	0.224 ± 0.078	0.241 ± 0.050
200	35.33 ± 1.25	0.203 ± 0.063	0.227 ± 0.041
250	33.20 ± 0.33	0.188 ± 0.055	0.207 ± 0.029
300	32.78 ± 0.83	0.185 ± 0.051	0.208 ± 0.022
350	32.67 ± 0.75	0.192 ± 0.064	0.208 ± 0.031
400	33.08 ± 0.59	0.192 ± 0.063	0.203 ± 0.033
450	33.26 ± 0.58	0.193 ± 0.057	0.204 ± 0.027
500	33.53 ± 0.82	0.195 ± 0.059	0.207 ± 0.030