KOYA: A RECOMMENDER SYSTEM FOR LARGE LANGUAGE MODEL SELECTION

Abraham Toluwase Owodunni[†,1] **and Chris Chinenye Emezue**[†,1,2,3,4]

[1] Masakhane, [2] Technical University of Munich, [3] Lanfrica, [4] Mila – Quebec Artificial Intelligence Institute
[†] These authors contributed equally to this work.

ABSTRACT

Pretrained large language models (LLMs) are widely used for various downstream tasks in different languages. However, selecting the best LLM (from a large set of potential LLMs) for a given downstream task and language is a challenging and computationally expensive task, making the efficient use of LLMs difficult for low-compute communities. To address this challenge, we present Koya, a recommender system built to assist researchers and practitioners in choosing the right LLM for their task and language, without ever having to finetune the LLMs. Koya is built with the Koya Pseudo-Perplexity (KPPPL), our adaptation of the pseudo perplexity, and ranks LLMs in order of compatibility with the language of interest, making it easier and cheaper to choose the most compatible LLM. By evaluating Koya using five pretrained LLMs and three African languages (Yoruba, Kinyarwanda, and Amharic), we show an average recommender accuracy of 95%, demonstrating its effectiveness. Koya aims to offer an easy to use (through a simple web interface accessible at `https://huggingface.co/spaces/koya-recommender/system`), cost-effective, fast and efficient tool to assist researchers and practitioners with low or limited compute access.

1 INTRODUCTION

Pre-training is a widely adopted strategy that enables very deep neural networks effectively learn from huge unlabeled text data (Qiu et al., 2020; Doddapaneni et al., 2021). While large-scale labeled datasets are expensive to annotate, there is an abundance of unlabeled, and sometimes noisy, text data and pre-training enables models to learn useful representations from them – representations that can improve performance on downstream natural language processing (NLP) tasks (Qiu et al., 2020).

The proliferation of large language models (LLMs) (Devlin et al., 2018; Sanh et al., 2019a; Tang et al., 2020; Ogueji et al., 2021; Chi et al., 2021; Dossou et al., 2022; Alabi et al., 2022; Adebara et al., 2022) has made the task of selecting the most suitable model for a given low-resource language and downstream task increasingly complex. The conventional approach to determine the best LLM involves finetuning multiple models and comparing their performance, which is computationally expensive and not feasible for low-resource communities (Bender et al., 2021; Ahia et al., 2021). The lack of time and computational resources makes it difficult for these communities to go through the process of finetuning a large number of models just to find and select the best-performing one (Amodei et al., 2018; Kaplan et al., 2020; Hoffmann et al., 2022).

To address this challenge, we present Koya, a powerful recommender system that simplifies the process of selecting a suitable model from a large number of possible pretrained models, without needing to finetune them. Koya is built with the Koya Pseudo-Perplexity (KPPPL), our adaptation of the pseudo perplexity, and ranks LMs in order of compatibility with the language of interest, making it easier and cheaper to choose the best-performing model.

2 Related Work

The objective of understanding an LLM's performance on a particular language without finetuning has been studied by Xia et al. (2020) and Ye et al. (2021). Xia et al. (2020) trained a regression model to predict the performance of a model on a supervised NLP task given the model, previous experimental setups, the training procedures, a test dataset and the language. In contrast, we propose a technique that is independent of training another model entirely and does not require the use of a model's experimental setup information which might not exist in an understudied domain or language. In addition the work uses language distance features from the URIRL database (Littell et al., 2017) which does not cover all existing languages. Similary, Ahuja et al. (2022) extends Microsoft Research's Project LITMUS [1] while training different machine learning models in order to predict another model's performance. This method depends on external tools and data points from the LITMUS project while our work is independent of these external variables that may not be present in a low-resource setting.

Evaluating LLMs has also been done by scoring them with their pseudo-log-likelihood (Salazar et al., 2019; Wang et al., 2019) which uses token masking. Our work builds on this with some key differences: we propose KPPPL which offers advantages like faster inference for Koya. We are also operating in a different context of constructing a recommender system with the masked language scores to recommend models for African languages. Finally, their work was directed toward evaluating ASR and NMT models while our initial work follows a model and task agnostic approach.

3 Motivation

Kunle is a masters student from a remote community in Rwanda who is interested in building an NLP model with the dataset he has been able to collect. Since his dataset is small, he will leverage transfer learning by finetuning a pretrained model on his downstream task. There are over 500 possible LMs on HuggingFace (HuggingFace) which Kunle can try. As Kunle does not work at Google, Meta or OpenAI, he does not have access to abundant compute resources. Luckily, he is able to get a Masakhane GCP compute but it has a limit of 1 day (the exact time it would take to finetune just one model) so Kunle cannot afford to finetune different models and test them in order to ascertain the most compatible for his downstream task.

All these make it a great difficulty for people like Kunle to find the optimal LLM for their use-case – until Koya.

4 Koya

Koya is a recommender system for choosing the most compatible pretrained LLM to use for a downstream task and language of interest. Koya aims to offer an easy to use (through a simple web interface), cost-effective, fast and efficient tool to assist researchers and practitioners with low or limited compute access.

4.1 How Koya Works

A pretrained LLM is not guaranteed to perform well in a given downstream task or language (Gururangan et al., 2020; Azunre, 2021; Adelani et al., 2022; Alabi et al., 2022). Therefore, after an LLM has been pretrained, it is important to estimate its performance on the downstream task and language of interest. Besides finetuning the LLM – which becomes computationally expensive as the set of LLMs in question increases – just to find the right one, one intrinsic method to evaluate a LLM relies on **perplexity**, which is defined as the inverse of the model's probability likelihood of the corpus (Jurafsky & Martin, 2009). The advantage of the perplexity is that as an intrinsic, task-agnostic measure of LLM performance, it does not require finetuning the LLM on any task, making the model evaluation much quicker (Gandhi). We give a formal definition of the perplexity of a language model below.

[1] https://www.microsoft.com/en-us/research/project/project-litmus/

Given a sequence S (e.g a sentence), of N tokens (e.g words) $(w_1, w_2, ..., w_N)$, the perplexity (PPL) of S is formally defined as the inverse of the model's probability likelihood of S. That is:

$$PPL(S) = \sqrt[N]{\frac{1}{P(w_1, w_2, ..., w_N)}}, \tag{1}$$

where $P(w_1, w_2, ..., w_N)$ is the likelihood of the S from the model (Jurafsky & Martin, 2009). For a causal autoregressive model (which models a word as the probability of the word given the sequence of words before it), we can decompose $P(w_1, w_2, ..., w_N)$ into

$$\prod_{i=1}^{N} P(w_i | w_1, ..., w_{i-1}) \tag{2}$$

and therefore Equation 1 becomes:

$$PPL(S) = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} P(w_i | w_1, ..., w_{i-1})}} \tag{3}$$

As a result, better language models will have a lower perplexity (or higher probability likelihood) for S. One can therefore interpret the perplexity as the model's ability to make sense of S (Wang et al., 2019). Koya is designed around this underlying concept: the more sense an LLM is able to make of a sentence in a given language, the easier and faster it will be for the model to learn the downstream task in that language of the task (Jurafsky & Martin, 2009; Wang et al., 2019). Therefore, the performance (on a downstream task) of an LLM depends on the extent to which the model is familiar with the language (Gonen et al., 2022).

Concretely, given a set M of n pretrained language models, $M = \{L_1, L_2, ...L_n\}$, and a dataset, D which will be used for a downstream task T, the user gives Koya one sentence S_T from D and Koya compares the perplexities from all the models in M for the sentence S_T. In theory, the perplexity can be computed over a corpus of many sentences, but in practice, we observe that with just a sentence, Koya is able to make a good recommendation, with the advantage of faster inference.

Adapting the perplexity to MLM-LLMs with Pseudo-Perplexity (PPPL) Masked language modelling (MLM) (Taylor, 1953) is the pretraining objective for bidirectional LLMs (which are ubiquitous) like BERT (Devlin et al., 2018) and its numerous variants like RoBERTa (Liu et al., 2019b), DistilBERT Sanh et al. (2019b), ALBERT (Lan et al., 2019), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), SpanBERT (Joshi et al., 2020). In MLM pretraining, we first randlomly mask out some tokens from the input sentence and then train the model to predict the masked tokens, using the rest of the unmasked tokens as context. MLM has been shown to be very ideal to learn useful linguistic representations of text (Devlin et al., 2018; Liu et al., 2019b), which is why it is used by many LMs, with extensions beyond pretraining (Davody et al., 2022).

In bidirectional MLMs, different from the autoregressive LMs, the conditional probability of a word/token w_i from a sequence $(w_1, w_2, ..., w_N)$ is

$$P(w_i | w_1, ...w_{i-1}, w_{i+1}, ..., w_N), \tag{4}$$

and due to the bidirectional nature, there isn't any known way, to the best of our knowledge, to compute the probability of the sequence itself (i.e $P(w_1, w_2, ..., w_N)$) that we need for perplexity (Equation 1). Salazar et al. (2019), building on past works (Chen et al., 2017; Wang & Cho, 2019; Shin et al., 2019), proposed the 'pseudo perplexity' for scoring MLM-LMs:

$$PPPL(S) = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} P(w_i | w_1, ...w_{i-1}, w_{i+1}, ..., w_N)}} \tag{5}$$

Koya Pseudo-Perplexity (KPPPL) For Koya, we adopt the pseudo-perplexity, and make a few key adjustments. Given a sequence, S, of N words/tokens $(w_1, w_2, ..., w_N)$, we calculate the Koya pseudo-perplexity (KPPPL for short) of S in the following steps:

1. We randomly select k tokens and *mask* a subset $W_R = \{w_{r_1}, ...w_{r_k}\}$ of S.
2. Then the KPPPL for S becomes:

$$KPPPL(S) \approx KPPPL(W_R) = \sqrt[k]{\frac{1}{\prod_{i=1}^{k} P(w_{r_i}|(w_1, ...w_N) \setminus w_{r_i})}} \tag{6}$$

3. Calculating $P(w_{r_i}|(w_1, ...w_N) \setminus w_{r_i})$: We will use a concrete example to explain how we get the likelihood of the token from the model. Let us assume we have a sequence $s = (s_1, ...s_n)$, and we have randomly selected one token s_k. After masking, the new sequence becomes $\hat{s} = (s_1, ...s_{k-i}, [MASK], ..., s_n)$. We then pass $\hat{s}$ to the model and through the logits of the masked token, we get a probability P_{mask} over the vocabulary size of the model. We then take $P_{mask}(s_k)$ which is interpreted as the likelihood of the model to choose s_k as the token that was masked.

5 Experimental Setting

We performed experiments to test the effectiveness of Koya, built on the KPPPL algorithm (proposed in Equation 6), as a recommender system. Concretely, we performed finetuning experiments on 2 downstream tasks and compared the ranked list of models, based on their downstream task, to the ranked list of models from Koya (using KPPPL). We chose news classification and named entity recognition (NER) as our downstream tasks and performed experiments involving 3 African languages: Amharic, Kinyarwanda and Yoruba. For this paper, we constrained ourselves to encoder-only LLMs which were trained with the MLM objective and considered five such models: AfriBERTa (Ogueji et al., 2021), Afro-XLMR (Alabi et al., 2022), mBERT (Devlin et al., 2018), InfoXLM (Chi et al., 2021), and XLM-RoBERTa Conneau et al. (2019). In the sections that follow, we will discuss the languages, models, training and recommendation setup.

5.1 Languages

Amharic: After Arabic, Amharic is the most widely spoken semantic language (Belay et al., 2022) belonging to the Afro-Asiatic language family. The language is made up of 34 characters and can be tokenized by whitespaces. In this work, we used the Amharic news classification dataset (Azime & Mohammed, 2021) which contains a total of 50706 articles that were collated from different new sources.

Kinyarwanda: Kinyarwanda belongs to the Niger-Congo and it is spoken by four different countries in Africa. It also belongs to a larger group of language families called *Rwanda-Rundi*. We used the KINNEWS dataset curated by Niyongabo et al. (2020) for this work. It contains 21,268 articles (3015 from MasakhaNER) having 14 classes. The data was collected from five newpapers and fifteen new websites all from Rwanda.

Yoruba: Yoruba also belongs to the Niger-Congo as Kinyarwanda . The dataset used to the Yoruba language section of this work is the Yoruba BBC News Topic Classification dataset which is hosted on HuggingFace[2]. It contained 1908 articles.

The MasakhaNER dataset Adelani et al. (2021), the first large publicly available high-quality dataset for named entity recognition (NER), was used for our NER experiments in our target languages. It has a total of 2500 Amharic samples, and 3035 Yoruba samples.

5.2 Models

mBERT mBert (Devlin et al., 2018) was pre-trained on 104 languages from the Wikipedia data using masked language modeling and next sentence prediction technique.

[2] `https://huggingface.co/datasets/yoruba_bbc_topics`

AfriBERTa: AfriBerta (Ogueji et al., 2021) was pretrained using masked language modeling technique but without next sentence prediction on 11 African languages, all constituting a total of 5,448,911 sentences in 1GB size of data. The languages belong to three language families which includes Afro-Asiatic, Niger-Congo and English Creole.

XLM-RoBERTa: XLM-RoBERTa (XLMR) (Conneau et al., 2019) is a multilingual variant of RoBERTa (Liu et al., 2019a). XLMR was pretrained on 100 languages with 2.5TB of Wikipedia data.

Afro-XLMR: Afro-XLMR (Alabi et al., 2022) is an African variant of XLM-RoBERTa that was pretrained on 17 African languages by including languages from 3 major language families in Africa and 3 high resource languages which are French, Arabic and English. It uses multilingual adaptive finetuning to extend XLMR to African languages (Alabi et al., 2022).

Info-XLM: InfoXLM (Chi et al., 2021) was pretrained using masked and translation language modeling (Lample & Conneau, 2019) and a contrastive learning objective where the encoded representation of two bilingual sentence pairs are presented to have almost the same meaning to the model. The model pretraining was done using several corpus by constructing the CC-100 dataset (Conneau et al., 2020) for masked language modeling objective which contains a total of 94 languages.

5.3 Finetuning Setup

All our finetuning experiments were implemented using the HuggingFace Transformers library (Wolf et al., 2020). We confined our finetuning to only the base versions of the models highlighted in Section 5.2. All finetuning was done for 3 epochs using the AdamW optimizer (Loshchilov & Hutter, 2017) and a batch size of 32. We used a learning rate or 5e-5 and a weight decay of 0.01.

5.4 Koya Recommendation Setup

For each language, we sampled a random sentence, S_r, from the test set of the respective task. Then, using the algorithm described in Section 4.1, we calculated KPPPL(S_r) for each of the focus models. While the perplexity was defined over the words of the sentence in Section 4.1, in practice they can also be taken over the tokens (i.e. after the sentence has been tokenized). Finally, we ranked the results for each language and compared with their ranking from the results of the finetuning experiments.

An important practical detail to note is that since the model outputs the conditional log-probabilities of the tokens, Equation 6 is revised to the log scale:

$$KPPPL(W_R) = exp[\frac{-1}{k}\sum_{i=1}^{k} log(P(w_{r_i}|(w_1, ...w_N) \setminus w_{r_i}))] (7)$$

5.4.1 Evaluating Koya: The RBO Score

Formally, the task of evaluating a recommender system is formulated as comparing ranked lists (Sarica et al., 2022) and there are broadly two different approaches to compare ranked lists: Rank Correlation and Set Based Measure (Ritesh, 2013). An in-depth analysis of each of these approaches is out of the scope of this paper and we refer the reader to the authors' work for that.

Concretely in our case, for a given language l, we have two ordered lists, R_s and R_f, of pretrained LLMs ranked from best to worst, R_s is got from our recommendation system using the KPPPL objective and R_f is got from the ranking of the performance of the LLMs in the finetuning experiments for l. The task is to find a function $F(R_f, R_s)$ that tells us how similar R_s is to R_f.

For our F we chose the rank-biased overlap (RBO) function (Webber et al., 2010). The RBO is a type of the set based measure, and is suited for our use case because 1) it is a bounded function – due to the use of the geometric series Webber et al. (2010), 2) it works even when the ranked lists are disjoint, i.e. some elements from one list are absent in the other list, and 3) during comparison, more weight is given to the elements in front than at the end of the list – which is more intuitive to a user's internal ranking system where the first few recommendations matter more than those at the end of the list (Webber et al., 2010; Sarica et al., 2022).

6 Results & Discussion

Table 1: F1 score on the news classification test set

	Yoruba	**Kinyarwanda**	**Amharic**
AfriBERTA	83.82	87.63	88.32
Afro-XLMR	55.71	85.73	85.66
mBERT	65.98	83.40	55.82
XLMR	29.85	79.33	86.78
Info-XLM	20.55	68.86	83.10

Table 2: F1 score on the NER validation set

	Yoruba	**Kinyarwanda**	**Amharic**
AfriBERTA	67.10	69.58	58.59
Afro-XLMR	54.93	78.38	49.09
mBERT	67.98	65.41	00.00
XLMR	51.97	68.50	38.08
Info-XLM	30.21	15.88	01.78

Table 1 and 2 shows our finetuning experiments on the news classification and NER tasks respectively. It can be seen that AfriBERTa leads in most of the experiments and Info-XLMR performs the worst in most of the experiments for news classification. The top performance of AfriBERTa can be attributed to the fact that AfriBERTa was pretrained on language data containing Yoruba, Amharic and Kinyarwanda (Ogueji et al., 2021). This attests to our earlier conjecture that LLMs perform better on a downstream task for a given language if they have some intrinsic understanding of that language already (often via pretraining) (Nekoto et al., 2020; Fan et al., 2021; Adelani et al., 2021; 2022; Babu et al., 2022; Team et al., 2022). The low performance of mBERT on Amharic NER was because mBERT's default tokenizer could not tokenize Ahmaric texts since it was not pre-trained on Amharic. This same result was obtained in the MasakhaNER dataset paper Adelani et al. (2021).

Table 3: RBO Score(%) of Koya for each language

Language	**News Classification**	**NER**
Amharic	96.38	95.39
Kinyarwanda	95.27	94.28
Yoruba	94.29	95.90

In Table 3, we report the similarity score for each language, between Koya's LLM recommendation (based on our KPPPL metric) and the ranking based on the LLM finetuning results (in Table 1 & 2). We observe that for all languages, Koya achieves very high RBO score, demonstrating its potential as a recommender system.

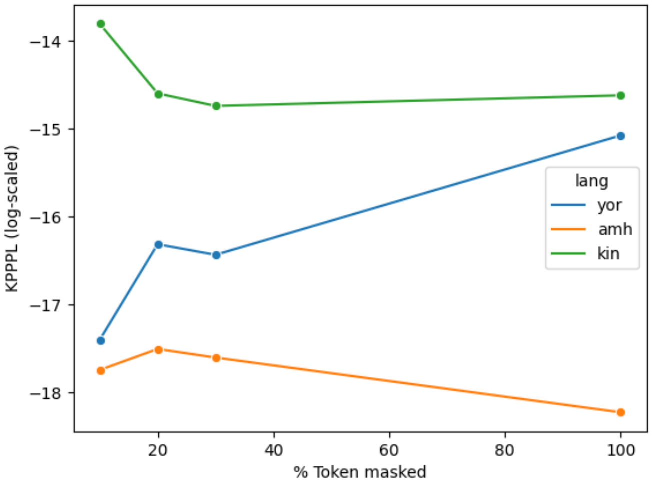

Figure 1: Effect of varying percentages of tokens masked for KPPPL. We use a fixed sequence length of 100. We show only results for AfriBERTa here due to space, and refer the reader to Figure 3 for an analysis on all models.

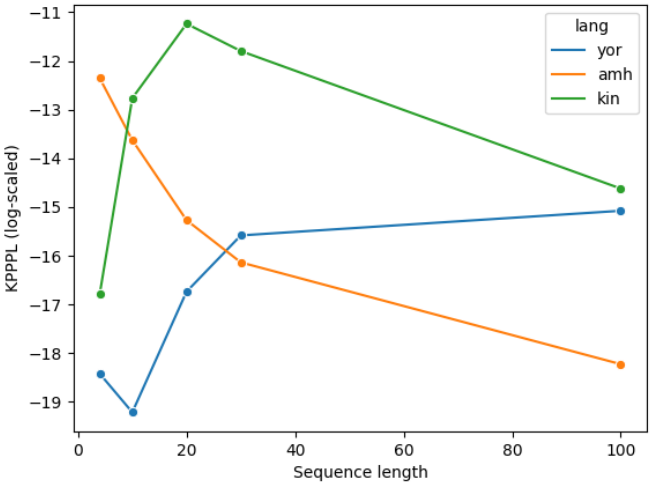

Figure 2: Effect of different sequence lengths on KPPPL for each language. Here all the tokens in the sequence are used to calculate KPPPL. We show only results for AfriBERTa here due to space, and refer the reader to Figure 4 for an analysis on all models.

KPPPL is a practical metric for ranking MLM-LLMs: Recall that in KPPPL (Equation 6) we considered a random subset of the sequence when calculating the perplexity. While this greatly improves inference time, we investigate its effectiveness by analyzing the effect of varying percentages of masked tokens (corresponding to different k in Equation 6). We evaluated the impact at 10%, 20%, 30%, and 100% masked tokens, where 100% means we are using the whole sequence. We do this analysis for all models and report our findings in Figure 3. For easy readability, we also show, in Figure 1, our findings for only the best-performing model from our finetuning experiments, AfriBERTa. The figures are log-scaled on the y-axis for an easier view of the behaviour (the perplexity scores have very low orders of magnitude).

We see that in both Figures 1 and 3, while the KPPPL slightly varies for the different percentages of masked tokens, the KPPPL ordering of the languages remains the same. In Figure 1, Amharic is always lower than Yoruba, which is in turn always lower than Kinyarwanda. Also, the KPPPL-based ordering (and hence the recommendation scoring) at 10% is the same at 100%, where the whole sequence length is used (similar to the standard pseudo perplexity scoring discussed in Equation 5), showing that the KPPPL method provides a reliable estimate of the pseudo perplexity (in terms of the ranking of LLMs). We believe that this is due to the randomization of the tokens used: since this random process mimics the actual MLM pretraining objective Devlin et al. (2018) of the LLM, KPPPL follows the same pattern of LLM training while estimating its performance. The biggest improvement to Koya with KPPPL is speed: we significantly reduce the time to query the LLM by considering a random subset (and not the full set) of the sequence.

In Figure 3, a holistic view of the impact of masking several percentage of the sequenced length can be seen. Yoruba achieves the least pseudo perplexity for all the three out of the five models (InfoXLM, Afro-XLMR and XLMR) while Amharic achieves the highest perplexity for three out of five models also (Info-XLM, XLMR and AfriBERTa).

Effect of the sequence length on ranking: Conversely in Figure 2 we see that the sequence length has an effect on the ordering. At a sentence length of 4, the ordering, based on ascending KPPPL order is (Yoruba, Kinyarwanda, Amharic). However this changes when we consider the same sentence, but its longer version. The new ordering becomes (Amharic, Yoruba, Kinyarwanda). It is important to note here that while the orderings are different, the KPPPL differences between each of the languages is infinitesimal. Altogether we infer that shorter sentences could give a slightly different ordering of the languages for a given LLM. When using Koya, customers are advised to input longer sequences (from a sentence to a paragraph) to get a more confident ranking.

7 Conclusion

Due to the combinatorial explosion of LLMs, datasets, and languages, the selection of a suitable pretrained LLM for a given task and language can be expensive and time-consuming, particularly for low-compute communities. To address this challenge, we introduce Koya, a recommender system that offers quick insight on the performance of LLMs for a given downstream task and language, without ever having to finetune the LLMs. The results of our experiments, using five pretrained LLMs and three African languages, show an average accuracy of 95%. Our results and analysis suggest that Koya is a promising solution for NLP practitioners looking to save time and computational resources when selecting an LLM.

8 FUTURE WORK

Future efforts will concentrate on expanding Koya to include additional model types, such as decoder-based models and models with different pretraining objectives. We also plan to investigate the applicability of our algorithm to models in different domains and the same language but different NLP tasks, since we only evaluated Koya with two downstream tasks. We are concluding the design of the web interface for Koya, which is hosted at `https://huggingface.co/spaces/koya-recommender/system`.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*, 2022.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, jul 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL `https://aclanthology.org/2022.naacl-main.223`.

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

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3316–3333, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.282. URL `https://aclanthology.org/2021.findings-emnlp.282`.

Kabir Ahuja, Antonios Anastasopoulos, Barun Patra, Graham Neubig, Monojit Choudhury, Sandipan Dandapat, Sunayana Sitaram, and Vishrav Chaudhary. Proceedings of the first workshop on scaling up multilingual evaluation. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, 2022.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.382`.

Dario Amodei, Danny Hernandez, Girish Sastry, Jack Clark, Greg Brockman, and Ilya Sutskever. Ai and compute. 2018.

Israel Abebe Azime and Nebil Mohammed. An amharic news text classification dataset, 2021. URL `https://arxiv.org/abs/2103.05639`.

P. Azunre. *Transfer Learning for Natural Language Processing*. Manning, 2021. ISBN 9781617297267. URL `https://books.google.de/books?id=WfkAzgEACAAJ`.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: self-supervised cross-lingual speech representation learning at scale. In Hanseok Ko and John H. L. Hansen (eds.), *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pp. 2278–2282. ISCA, 2022. doi: 10.21437/Interspeech.2022-143. URL `https://doi.org/10.21437/Interspeech.2022-143`.

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pp. 84–89. IEEE, 2022.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

X. Chen, Anton Ragni, X. Liu, and Mark John Francis Gales. Investigating bidirectional recurrent neural network language models for speech recognition. In *Interspeech*, 2017.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL `https://aclanthology.org/2021.naacl-main.280`.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference On Learning Representations*, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

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

Ali Davody, David Ifeoluwa Adelani, Thomas Kleinbauer, and Dietrich Klakow. TOKEN is a MASK: few-shot named entity recognition with pre-trained language models. In Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala (eds.), *Text, Speech, and Dialogue - 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings*, volume 13502 of *Lecture Notes in Computer Science*, pp. 138–150. Springer, 2022. doi: 10.1007/978-3-031-16270-1_12. URL `https://doi.org/10.1007/978-3-031-16270-1_12`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. A primer on pretrained multilingual language models. *arXiv preprint arXiv: Arxiv-2107.00676*, 2021.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages, 2022.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886, 2021.

Meet Gandhi. Evaluation of Language Models through Perplexity and Shannon Visualization Method — towardsdatascience.com. `https://towardsdatascience.com/evaluation-of-language-models-through-perplexity-and-shannon-visualization-method-9148fbe10bd0`. [Accessed 05-Feb-2023].

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv: Arxiv-2212.04037*, 2022.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740`.

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

HuggingFace. Huggingface models. URL https://huggingface.co/models?language=multilingual&amp;sort=downloads.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009. ISBN 9780131873216 0131873210. URL http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv: Arxiv-1909.11942*, 2019.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, 2017.

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

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

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017.

Wilhelmina Nekoto, V. Marivate, T. Matsila, Timi E. Fasubaa, T. Kolawole, T. Fagbohungbe, S. Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo KABENAMUALU, Salomey Osei, Sackey Freshia, Andre Niyongabo Rubungo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, L. Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, I. Ezeani, Idris Abdulkabir Dangana, H. Kamper, Hady ElSahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris C. Emezue, Bonaventure F. P. Dossou, Blessing K. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Oktem, Adewale Akinfaderin, and A. Bashir. Participatory research for low-resourced machine translation: A case study in african languages. *FINDINGS*, 2020. doi: 10.18653/v1/2020.findings-emnlp.195.

Rubungo Andre Niyongabo, Hong Qu, Julia Kreutzer, and Li Huang. Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi. *arXiv preprint arXiv:2010.12174*, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mrl-1.11.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897, 2020.

Agrawal Ritesh. Comparing Ranked List — ragrawal.wordpress.com. https://ragrawal.wordpress.com/2013/01/18/comparing-ranked-list/, 2013. [Accessed 04-Feb-2023].

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. *Annual Meeting Of The Association For Computational Linguistics*, 2019. doi: 10.18653/v1/2020.acl-main.240.

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

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

Alessia Sarica, Andrea Quattrone, and Aldo Quattrone. Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: A case study on parkinson's disease. In Mufti Mahmud, Jing He, Stefano Vassanelli, André van Zundert, and Ning Zhong (eds.), *Brain Informatics*, pp. 129–139, Cham, 2022. Springer International Publishing. ISBN 978-3-031-15037-1.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. Effective sentence scoring method using bert for speech recognition. In Wee Sun Lee and Taiji Suzuki (eds.), *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pp. 1081–1093. PMLR, 17-19 Nov 2019. URL https://proceedings.mlr.press/v101/shin19a.html.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. 2020.

Wilson L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953. doi: 10.1177/107769905303000401. URL https://doi.org/10.1177/107769905303000401.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *META*, 2022.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv: Arxiv-1902.04094*, 2019.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. Does it make sense? and why? a pilot study for sense making and explanation. *Annual Meeting Of The Association For Computational Linguistics*, 2019. doi: 10.18653/v1/P19-1393.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), nov 2010. ISSN 1046-8188. doi: 10.1145/1852102.1852106. URL https://doi.org/10.1145/1852102.1852106.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. *arXiv preprint arXiv:2005.00870*, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Zihuiwen Ye, Pengfei Liu, Jinlan Fu, and Graham Neubig. Towards more fine-grained and reliable nlp performance prediction. *arXiv preprint arXiv:2102.05486*, 2021.

A Appendix

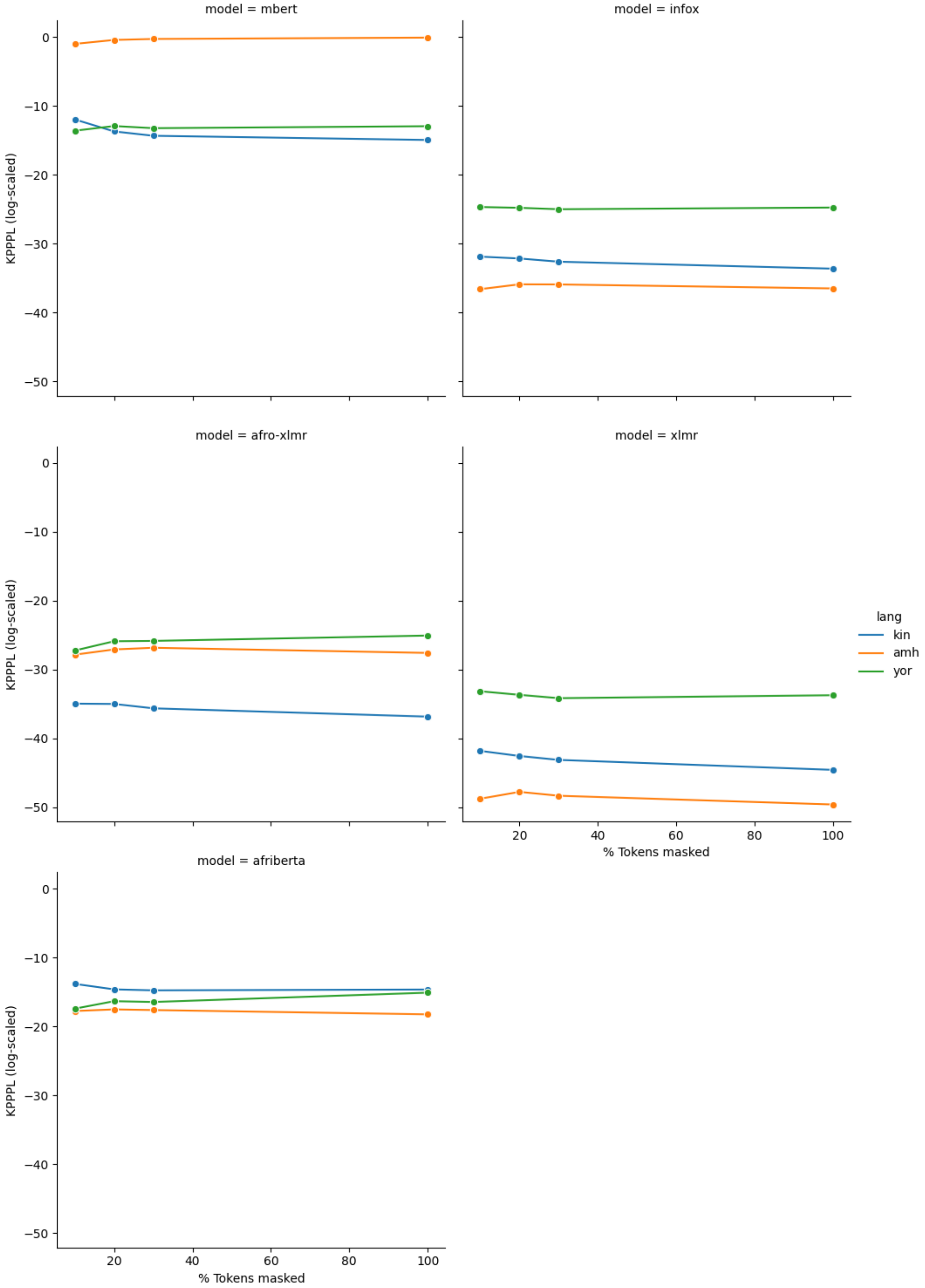

Figure 3: Effect of varying percentages of tokens masked and used for KPPL. We use a fixed sequence length of 100.

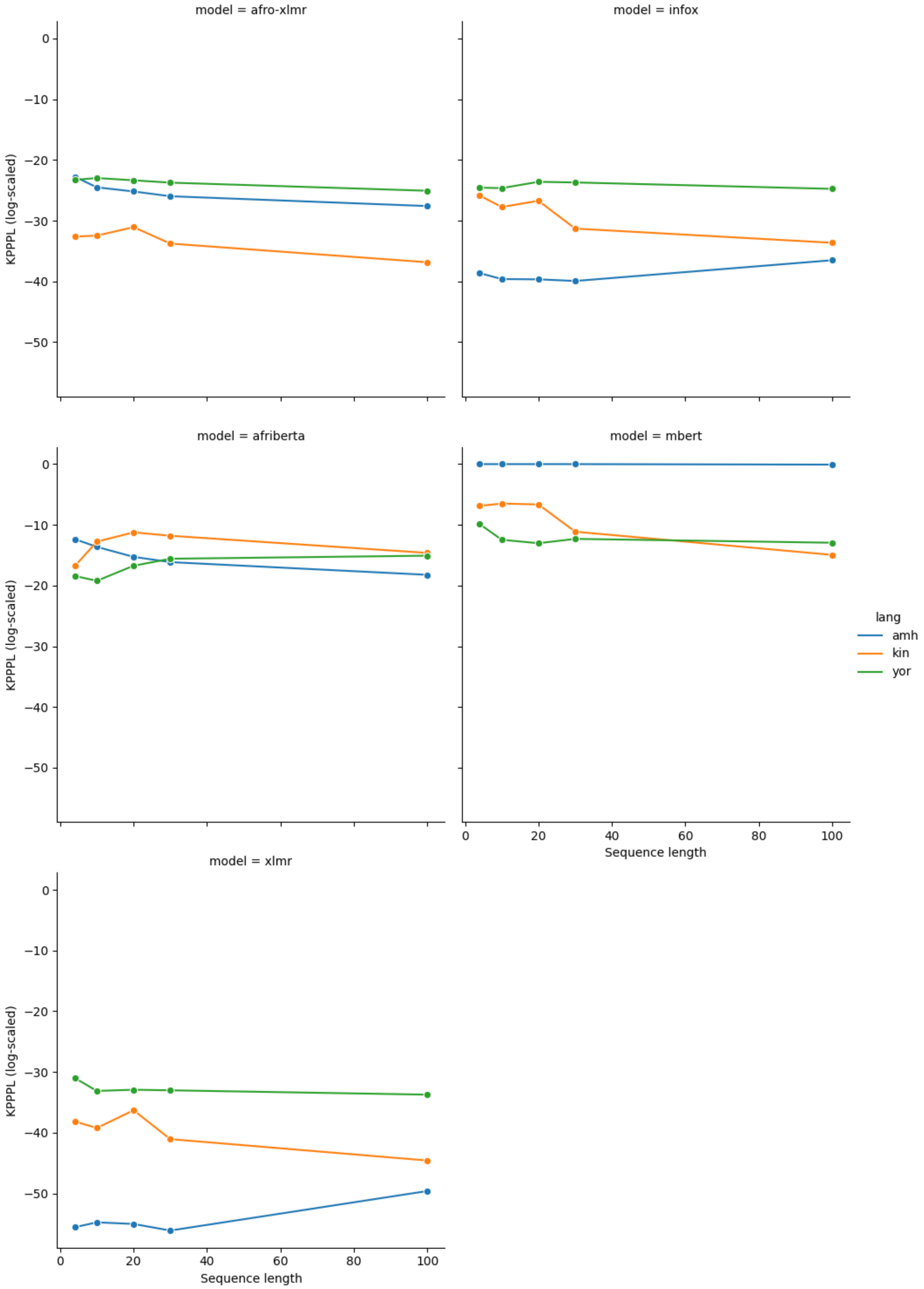

Figure 4: Effect of different sequence lengths on KPPPL for each language. Here all the tokens in the sequence are used to calculate KPPPL.