SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects

David Ifeoluwa Adelani[1,2*], **Hannah Liu**[3*], **Xiaoyu Shen**[4*†], **Nikita Vassilyev**[3], **Jesujoba O. Alabi**[2,5], **Yanke Mao**[3], **Haonan Gao**[3], **and Annie En-Shiun Lee**[3,6].
[1]University College London, [2]Masakhane, [3]University of Toronto, [4]Amazon Alexa AI, [5]Saarland University, [6]Ontario Tech University
d.adelani@ucl.ac.uk, hannahhere.liu@mail.utoronto.ca, gyouu@amazon.com

Abstract

Despite the progress we have recorded in the last few years in multilingual natural language processing, evaluation is typically limited to a small set of languages with available datasets which excludes a large number of low-resource languages. In this paper, we created SIB-200—a large-scale open-sourced benchmark dataset for topic classification in 200 languages and dialects to address the lack of evaluation dataset for Natural Language Understanding (NLU). For many of the languages covered in SIB-200, this is the first publicly available evaluation dataset for NLU. The dataset is based on Flores-200 machine translation corpus. We annotated the English portion of the dataset and extended the sentence-level annotation to the remaining 203 languages covered in the corpus. Despite the simplicity of this task, our evaluation in full-supervised setting, cross-lingual transfer setting and prompting of large language model setting show that there is still a large gap between the performance of high-resource and low-resource languages when multilingual evaluation is scaled to numerous world languages. We found that languages unseen during the pre-training of multilingual language models, under-represented language families (like Nilotic and Altantic-Congo), and languages from the regions of Africa, Americas, Oceania and South East Asia, often have the lowest performance on our topic classification dataset. We hope our dataset will encourage a more inclusive evaluation of multilingual language models on a more diverse set of languages.[1]

1 Introduction

In the last few years, developing massively multilingual Pre-trained Language Models (PLMs) to scale to several written languages is a very active area of research—e.g. covering 100 languages (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021; He et al., 2023). However, evaluation is often limited to a few tens of languages with benchmark datasets (Conneau et al., 2018; Hu et al., 2020; Ruder et al., 2021; Zhang et al., 2022), limiting the large-scale evaluation of current multilingual language models on many languages especially the truly low-resourced languages.

While there is evidence from previous works that languages not covered during pre-training often lead to lower performance, such analysis is also limited to a small selection of languages with annotated datasets (Ponti et al., 2020; Pfeiffer et al., 2020; Adelani et al., 2022b; Lee et al., 2022).

Recently, there is a push to scale evaluation datasets to more than 100 languages, but this requires a very expensive annotation effort in terms of money and time. Often, this scaling is only carried out by a large community effort that spans many years like the Universal Dependency (UD) project (Nivre et al., 2017, 2020; de Marneffe et al., 2021) or financed by BigTech companies (Goyal et al., 2022; NLLB-Team et al., 2022; Federmann et al., 2022; Conneau et al., 2022; Pratap et al., 2023). However, the focus of these large-scale evaluation is on machine translation and speech recognition tasks—ideal for text generation tasks. In reality, there are only few benchmarks for NLU tasks that cover all the languages seen during the pre-training of multilingual PLMs (ImaniGooghari et al., 2023).

The large benchmark datasets that are available are UD, Taxi1500 (Ma et al., 2023) WikiANN (Pan et al., 2017), and Belebele (Bandarkar et al., 2023) for dependency parsing, text classification, named entity recognition, and reading comprehension respectively. The largest is Taxi-1500 for 1500 languages—based on the Bible but the dataset is not publicly available due to copyright issues of the dataset. WikiANN on the other hand, was automatically annotated and with few instances for low-resource languages. UD and Belebele were

*Equal contribution and corresponding authors
†Work done outside Amazon
[1]https://github.com/dadelani/sib-200

manually annotated and covered between 100 and 125 languages. However, many low-resource languages are still missing in the evaluation.

In this paper, we created SIB-200—a large-scale open-sourced benchmark dataset for topic classification to address the lack of evaluation datasets for natural language understanding. The dataset is based on Flores-200 dataset (Goyal et al., 2022; NLLB-Team et al., 2022)—a multi-way parallel corpus (i.e. same sentences are available in 204 languages). We annotated the English portion of the Flores-200 dataset and extend the sentence-level annotation to the remaining 203 languages covered in the machine translation corpus.

Our evaluation shows that there is still a large gap between the performance of high-resource and low-resource languages when multilingual evaluation is scaled to numerous world languages. Languages unseen during the pre-training of multilingual PLMs, under-represented language families (like Nilotic and Altantic-Congo languages), and languages from the regions of Africa, Americas, Oceania and South East Asia, often have the lowest performance on our text classification dataset. We also find that blindly scaling up the number of languages without scaling up the domains in pretraing is unhelpful (e.g., Glot-500 pre-trained on 500 languages largely under-performs XLM-R pretrained on 90 languages). It is crucial to mix text from various domains. For languages unseen during pre-training, we show the potential of multilingual language adaptive fine-tuning (MAFT)[2] (Tang et al., 2020; Alabi et al., 2022) in improving the performance of these languages by leveraging synthetic data for languages with tiny monolingual data. Evaluation of this approach on African languages results in significant improvement (up to +5% in accuracy on average) for the previously unseen languages.

Finally, we extend our evaluation to the zero-shot settings by training individually on English, French, Arabic and Chinese (Simplified) languages using XLM-R (Conneau et al., 2020), and performing zero-shot evaluation on other languages. We compared these results with prompting large language models (LLMs) like GPT-4. Our results show that LLMs perform poorly on over 63.6% (or 131 out of 204 languages) of the languages with less than 70% in accuracy while zero-shot adaptation from the English model only leads to performance less than 70% accuracy in 81 languages (or 39.3% of languages)[3]. This shows that leveraging cross-lingual transfer from high-resource languages is much better than prompting LLMs for many low-resourced languages in this task.

Label	TRAIN	DEV	TEST	TOTAL
science/technology	176	25	51	252
travel	138	20	40	198
politics	102	14	30	146
sports	85	12	25	122
health	77	11	22	110
entertainment	65	9	19	93
geography	58	8	17	83
Total	701	99	204	1,004

Table 1: **SIB-200 dataset**. We provided the data size of the annotated data by their SPLIT and category

2 SIB-200 dataset

2.1 Data source

We introduce our new dataset, SIB-200—a **S**imple **I**nclusive and **B**ig topic classification dataset for over 200 languages and dialects. We leveraged the multi-way parallel Flores-200 dataset (NLLB-Team et al., 2022) for the creation of the dataset. Flores-200 corpus is an extension of Flores-101 (Goyal et al., 2022)—for 101 languages. In both datasets, the source sentences were collected in English and translated by professional translators to several languages. In total, the corpus contains 3,001 sentences divided into `DEV` (997 sentences), `DEVTEST` (1,012 sentences) and `TEST` (992 sentences) sets. However, the authors did not release the `TEST` set.

Flores-200 released additional information to provide meta-data information about the domains and topics of the articles covered in the dataset. The domains are based on WikiNews, WikiJunior, and WikiVoyage with a total of 842 articles while the topics are based on "crime", "disasters", "entertainment", "geography", "health", "nature", "politics", "science", "sports", and "travel".[4] However, a quick review of the dataset revealed that at the sentence level, the article can belong to more than one topic. Therefore, we decided to add our topic categorization at the sentence level. Performing annotation at

[2]adaptation of an existing multilingual PLM to multiple or new sets of languages simultaneously.

[3]Performance of XLM-R on English is 92.1% in accuracy while prompting GPT-4 in English gave 76.6% in accuracy.

[4]We note that in the open-sourced dataset, there are more categories than the ten reported in the paper.

the sentence level also gives us the additional advantage of having more samples to annotate (2,009 rather than 562 instances[5]).

2.2 Data annotation

We recruited four annotators who are native speakers of English to label 2,001 sentences obtained from the `DEV` and `DEVTEST` sets of Flores-200[6]. We make use of an internal annotation tool for text classification. The annotation labelling scheme covers 15 categories, 10 are from the original Flores-200 categorization of articles (§2.1), and the others are "business", "religion", "technology", "education", and "uncategorized". We assigned sentences that do not fit any of the defined categories, and sentences lacking sufficient context about their topic to "uncategorized". An example of a sentence labelled as "uncategorized" is *"In Berlin, police estimated 6,500 protestors"*.

The annotators took about two weeks to complete the task, however on average it takes up to 60 seconds to annotate a sentence (approximately, 33 hours of annotation time).

2.3 Quality control

We report Fleiss Kappa score (Fleiss et al., 1971) to measure the agreement of annotation. The Fleiss Kappa score among the four annotators is **0.44**—which signifies a moderate level of agreement.

Choosing the final label per sentence We assigned the final label to a sentence by majority voting. Specifically, we assign a label to a sentence if at least two annotators agree on the category, but we excluded the situation, where any two annotators conflicted with the other two annotators. For example, for the sentence *"The major organ of the circulatory system is the heart, which pumps the blood."*, the first two annotators assigned "science" while the last two assigned "health". In total, we assigned a single label to 1,695 sentences, but there were 314 sentences with conflicts in the annotation. We asked the lead annotator to adjudicate the sentences with conflicting annotations and assigned a single label to each sentence. We later combined the fixed conflicting annotations with the others to give us back a total of 2009 annotated sentences.

Final classification dataset For the final dataset, we excluded sentences with the label of "uncategorized", we only selected label categories with more than 80 sentences, this removed categories such as "business" (80 sentences), "disasters" (73 sentences), "crime" (72 sentences), "education" (52 sentences), and "religion" (46 sentences). We note that having too many categories with few sentences makes building text classification models a bit difficult leading to a lower performance. Also, we combined "science" (138 sentences) and "technology" (114 sentences) category into a single category of "science/technology". Finally, we removed the "nature" category because there is a lot of conflict with "science" and "geography" categories. Our preliminary experiments show that adding "nature" significantly lowers the performance of our classifier. About half of the Flores-200 is part of the SIB-200 dataset (i.e. 1004 out of 2009 sentences).

Table 1 shows the number of sentences per label in each of the `TRAIN`, `DEV`, and `TEST` splits. We divided the sentences into the split using the 70%, 10%, 20% ratio.

3 Languages and their categorizations

Table 2 and Table 3 shows the grouping of languages in the SIB-200 dataset. We categorized them based on the following characteristics: (1) geographical regions, (2) language family, (3) coverage in multilingual PLMs, and (4) Joshi's classification (Joshi et al., 2020)—a categorization based on their labelled/unlabelled resources on the web—making it easy to analyze results.

Categorization by geographical regions Table 2 shows the grouping of languages into regions according to the United Nations Geoscheme[7]. The regions are: Africa, Americas, Asia 1 or Western & Central Asia, Asia 2 or Southern Asia, Asia 3 or South-Eastern & Eastern Asia, Europe 1 or Northern/Western/Southern Europe, Europe 2 or Eastern Europe, and Oceania. Asia, Europe, and Africa regions have the largest number of languages with 82, 57, and 56 languages respectively. The Oceania and the Americas regions have the lowest number of languages with four and five respectively.

Categorization by language family SIB-200 languages are grouped into 21 language families as shown in Table 3, the largest groups are: Indo-European (79 languages), Atlantic-Congo (34 languages), Afro-Asiatic (21 languages), Austronesian

[5] Although 842 articles are in Flores-200, only 562 articles are open-sourced as part of DEV and DEVTEST sets.

[6] All annotators are also authors of this paper.

[7] `https://en.wikipedia.org/wiki/United_Nations_geoscheme`

Region	Joshi's class 0	1	2	3	4	5	None	NLLB class LRL	HRL	Language Families
Africa	10	20	9	2	–	–	15	50	6	Atlantic-Congo (34), Afro-Asiatic (12), Nilotic (5), Indo-European (2), Mande (2), Austronesian (1)
Americas	1	3	–	–	–	–	1	5	–	Indo-European (2), Aymaran (1), Tupian (1), Quechuan (1)
Asia 1 (W & C)	2	8	–	4	2	1	7	18	6	Afro-Asiatic (8), Turkic (8), Indo-European (7), Kartvelian (1)
Asia 2 (S)	4	14	2	3	1	–	3	25	2	Indo-European (19), Dravidian (4), Sino-Tibetan (3), Austroasiatic (1)
Asia 3 (SE & E)	3	17	1	5	2	2	1	22	9	Austronesian (17), Sino-Tibetan (6), Tai-Kadai (3), Austroasiatic (2), Japonic (1), Mongolic-Khitan (1), Koreanic (1)
Europe 1 (N, W, S)	1	17	3	7	10	4	–	19	23	Indo-European (36), Uralic (3), Constructed (1), Basque (1), Afro-Asiatic (1)
Europe 2 (E)	–	6	–	6	3	–	–	7	8	Indo-European (12), Turkic (3)
Oceania	–	4	–	–	–	–	–	4	–	Austronesian (3), Indo-European (1)
Total	21	90	17	30	22	12	27	150	54	

Table 2: Language families covered in the **SIB-200 dataset** grouped by United Nations geoscheme **regions**, **Joshi's classes** (Joshi et al., 2020) (None – for languages not found in Joshi's dataset), and **NLLB classification** (NLLB-Team et al., 2022) of languages by the size of resources on the internet.

Language Family	#Lang FTC	XLM-R	Glot 500	Afri BERTa	Afro XLM-R	Indic BERTv2
Indo-European	79	50	72	1	4	14
Atlantic-Congo	34	2	32	6	10	-
Afro-Asiatic	21	6	14	4	5	-
Austronesian	21	5	17	-	1	-
Turkic	11	7	11	-	-	-
Sino-Tibetan	9	3	7	-	-	-
Nilotic	5	-	1	-	-	-
Dravidian	4	4	4	-	-	4
Tai-Kadai	3	2	2	-	-	-
Uralic	3	3	3	-	-	-
Austroasiatic	3	2	3	-	-	1
Mande	2	-	2	-	-	-
Japonic	1	1	1	-	-	-
Koreanic	1	1	1	-	-	-
Mongolic-Khitan	1	1	1	-	-	-
Constructed	1	1	1	-	-	-
Quechuan	1	-	1	-	-	-
Basque	1	1	1	-	-	-
Aymaran	1	-	1	-	-	-
Tupian	1	-	1	-	-	-
Kartvelian	1	1	1	-	-	-
Total	204	90	177	11	20	19

Table 3: **Languages covered in multilingual pre-trained language model** and their language families. We excluded MuRIL because it was trained in similar languages as IndicBERTv2 except for Santali in the Austroasiatic family.

(21 languages) and Turkic (11 languages).

Categorization by Joshi's classification Table 2 also shows the number of languages in each Joshi's class—a measure of the unlabelled or labelled resources available for each language on the web (Joshi et al., 2020). 128 languages can be categorized as low-resourced since they fall between class "0" and "2", 30 languages are mid-resourced in class "3", and the others are high-resourced (only 39 languages). This also corresponds to the NLLB classification for machine translation resources available on the web, but with only two categories—150 low-resourced languages and 54 high-resourced languages.

Categorization by availability in PLM Lastly, Table 3 shows our grouping of languages and language families by their inclusion in the training of multilingual PLMs. XLM-R (Conneau et al., 2020) covered 90 out of the 204 languages in our dataset while GLOT-500 (ImaniGooghari et al., 2023) covered 177. This is a good indication of performance in general since languages that are included during pre-training often have better performance (Ponti et al., 2020; Pfeiffer et al., 2020; Adelani et al., 2022b). Finally, we show the number of languages covered by region-specific PLMs such as AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022), MuRIL (Khanuja et al., 2021), and IndicBERTv2 (Doddapaneni et al., 2023).

4 Experimental setup

Here, we describe the different kinds of text classification models developed for the SIB-200 dataset and the experimental settings (i.e. full supervised setting and zero-shot transfer setting).

4.1 Text classification models

We trained a simple Multilayer Perceptron (MLP), fine-tuned several multilingual PLMs and prompted large language models for text classification.

Multi-Layer Perceptron For the input features, we make use of either n-gram features (n=1 up to 3 in our experiments) or XLM-R tokens obtained by first tokenizing the sentences using XLM-R tokenizer. We make use of the default setting on scikit-learn tool (Pedregosa et al., 2011)

Masked Language Models (MLM) Next, we fine-tune massively multilingual PLM such as XLM-R-base (270M parameters), XLM-R (550M)

Language Family	Count	Fully Supervised				Cross-Lingual Transfer (XLMR)				Zero-Shot Prompt	
		MLP	Glot-500	XLM-R (base)	XLM-R	English	French	Chinese	Arabic	GPT-3.5-Turbo	GPT-4
English	-	59.9	82.8	90.0	92.1	92.1	91.9	**92.5**	91.2	71.8	76.6
Indo-European	79	62.3	72.4	81.4	**86.2**	82.4	83.2	82.8	83.0	55.3	66.6
Atlantic-Congo	34	**61.3**	49.6	50.5	57.9	41.4	41.4	41.9	42.0	29.2	29.2
Afro-Asiatic	21	61.4	59.2	67.1	**72.6**	67.4	68.1	67.7	68.4	43.4	54.6
Austronesian	21	59.8	62.1	68.8	**73.9**	64.0	64.3	64.5	64.9	44.1	47.1
Turkic	11	64.8	74.2	79.8	**85.1**	80.2	80.9	80.4	80.9	50.2	59.2
Sino-Tibetan	9	**68.8**	66.2	62.2	65.4	57.9	58.3	57.1	57.1	30.7	40.6
Nilotic	5	**58.6**	35.0	48.2	53.7	34.8	33.0	34.0	34.0	16.1	10.1
Dravidian	4	64.7	76.1	84.4	**87.9**	87.8	88.1	88.2	88.0	57.2	69.6
Tai-Kadai	3	67.7	61.3	70.9	**76.8**	68.4	67.8	68.9	69.2	35.6	44.7
Uralic	3	62.1	74.1	86.5	**89.6**	89.1	90.4	90.2	89.6	62.4	74.8
Austroasiatic	3	66.5	65.5	66.2	**68.1**	67.5	66.8	67.2	66.2	34.8	48.7
Mande	2	**57.4**	36.1	42.7	48.7	32.5	32.4	32.3	32.1	18.0	13.3
Japonic	1	73.8	81.5	87.9	**89.9**	89.3	90.3	89.7	88.8	63.4	75.8
Koreanic	1	67.8	76.5	86.5	**88.5**	88.7	89.4	89.2	88.7	67.8	78.2
Mongolic-Khitan	1	66.2	74.8	82.9	**88.5**	86.1	85.8	85.5	86.2	57.7	67.6
Constructed	1	61.4	72.8	87.5	**89.4**	88.5	89.2	90.4	88.6	58.7	70.3
Quechuan	1	53.7	59.4	57.9	**64.1**	46.3	48.3	49.1	50.8	36.2	18.5
Basque	1	62.9	72.4	83.5	**89.2**	89.2	90.0	89.7	88.9	55.3	53.1
Aymaran	1	**55.7**	37.4	42.5	52.5	39.1	40.4	38.5	41.3	15.9	6.6
Tupian	1	57.7	63.7	69.6	**76.3**	61.3	61.7	61.7	61.1	32.3	28.2
Kartvelian	1	63.7	78.4	83.4	**88.5**	89.1	89.8	89.7	88.6	44.7	66.1
Average	-	62.8	64.2	71.0	**75.9**	69.1	69.5	69.5	69.5	43.3	48.7

Table 4: **Overall result of the performance of different text-classification models across different language families.** We compared different settings: fully-supervised, cross-lingual transfer and zero-shot prompting of LLMs. Cross-lingual transfer is based on the XLM-R model as it is the best-performing PLM. Performances from 4 source languages: English, French, Chinese and Arabic are reported.

Glot-500 (395M), which are trained on several languages: XLM-R and Glot-500 were trained on 100 and 500 languages respectively. We also fine-tune region-specific PLM trained on multiple country-level or continent-level languages: AfriBERTa (126M), AfroXLMR (550M), MuRIL (236M) and IndicBERTv2 (278M). We restrict region-level analysis to Africa and India because we only found these two regions with multilingual PLMs covering many languages.

MAFT with fewer data and synthetic data We explore how to improve over regional PLMs using MAFT—adaptation of an existing multilingual PLM to multiple or new set of languages simultaneously, this was effective for adapting XLM-R to 20 languages spoken in Africa (Alabi et al., 2022). To extend to more languages, we apply MAFT to 61 African languages with at least 10MB of monolingual data (AfroXLMR-61). The data was obtained from the concatenation of different web sources like AfroXLMR training corpus, MT560 (Gowda et al., 2021) (mostly religious articles), Flores-200 (multi-domain), and Wikipedia. In total, this results in 17GB of data. To further extend to more languages with less than 10MB of data, we generate machine-translated data using NLLB for 34 African languages (including 18 in AfroXLMR-61). The selected 34 languages are the ones with less than 10MB or only have MT560 (religious domain). We make use of the English news commentary dataset[8] (Kocmi et al., 2022) with over 600,000 sentences to translate to these 34 languages. We refer to the resulting model after adaptation as AfroXLMR-75 which has been pre-trained on 21GB of data.

Large Language Models Lastly, we also report results by prompting two popular large language models: GPT-3.5-Turbo (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0613). Compared with smaller language models from MLM and MAFT, they feature strong instruction-following capabilities without task-specific fine-tuning.

4.2 Training and evaluation scenarios

Fully-supervised In this setting, we trained on each language in SIB-200 and evaluated on the same language. We did this evaluation for 204 lan-

[8]we used version 16 of the data released for WMT.

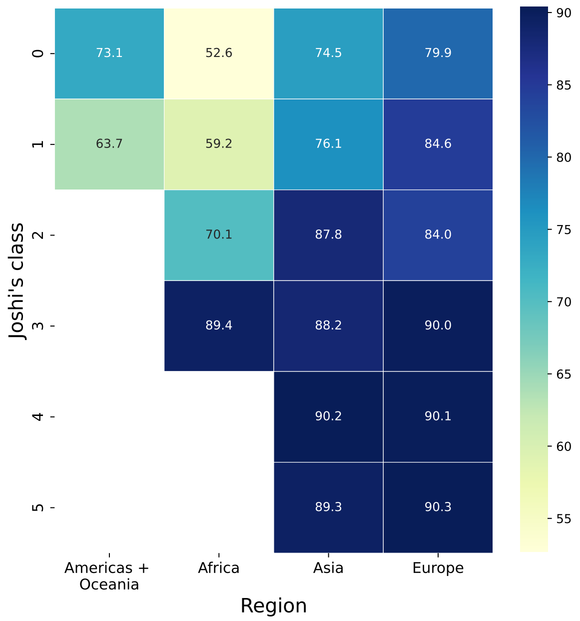

Figure 1: Heatmap of the performance by Region in each Joshi's class.

guages and compared the performance of different text classification models. The MLP models were trained for 300 iterations, and we used either word *ngram tokens* or *XLM-R tokens*. For the multilingual PLM, we fine-tune each language training data for 20 epochs, with a maximum sequence length of 164, batch size of 16, and learning rate of 0.00001 on a single Nvidia A10 GPU. **Here, we assume access to labelled data in the target language.**

Cross-lingual transfer For this setting, we **fine-tune** XLM-R on a language in Joshi's class 5 (we call it a "source" language), and **evaluate** on other languages. For this setting, we **fine-tune** XLM-R on a language in Joshi's class 5 (we call it a "source" language), and **evaluate** on other languages. We trained in four languages with three different scripts i.e. English, French, Arabic and Chinese (Simplified). **Here, we assume access to labelled data in a few high-resource languages.**

Zero-shot prompt We prompt GPT-3.5/4 for text classification for the 204 languages using an English template. We make use of a simple template from Sanh et al. (2022): *'Is this a piece of news regarding {{"science, technology, travel, politics, sports, health, entertainment, or geography"}}? {{INPUT}}'*. **Here, we assume no access to labelled data in any language**

5 Results

5.1 Baseline results

In order to demonstrate the effectiveness of our data set for multilingual evaluation, we benchmark the performance across various models and group the results by categorizations (Table 4). As XLM-R consistently outperforms Glot-500 across almost all language families, we use XLM-R as the baseline model in the cross-lingual transfer experiments[9]. Comparing English versus other languages, fine-tuning XLM-R on English achieved an accuracy of 92.1%, indicating that *the task itself is not difficult if given a properly pre-trained MLM and* $\sim$ 700 *training samples*. However, when fine-tuning the same model in other languages, the performance drops vastly to an average accuracy of 75.9%. Similarly, in the cross-lingual transfer and zero-shot prompt scenarios, the performance further drops.

Performances across language families The distribution of accuracy scores is imbalanced across language families. *Atlantic-Congo, Nilotic, Mande, Aymaran and Quechuan languages have the lowest accuracy scores*. Even under the fully supervised scenario, the best-performed model reaches <65% accuracy scores on these languages. There also tends to be a larger performance gap between fully-supervised and cross-lingual transfer scenarios, suggesting a poor semantic alignment (Conneau and Lample, 2019) for these languages. Surprisingly, Tupian is the only additional language family that has >10% drop from the fully supervised to cross-lingual transfer scenario. When moving further to the zero-shot prompt scenario, Basque shows the biggest performance drop (-36%), next come the above-mentioned languages. Interestingly, despite this large decrease, Basque scores exceptionally high ($\approx$90%) in the fully supervised and cross-lingual transfer scenarios.

Performances across Joshi's classes and geographical regions Figure 1 visualizes the performance of XLM-R[10] across different regions and Joshi's classes. We see a clear trend that languages with higher Joshi's classes perform better. Specifically, all languages with Joshi's class $\geq$3 have accuracy scores of $\approx$90%. *For languages in the same Joshi's class, African languages perform the worst, and European languages perform the best.*

[9] the full results are in Appendix A

[10] We omit other models and only show XLM-R as Table 4 has shown fine-tuning the XLM-R model performs the best

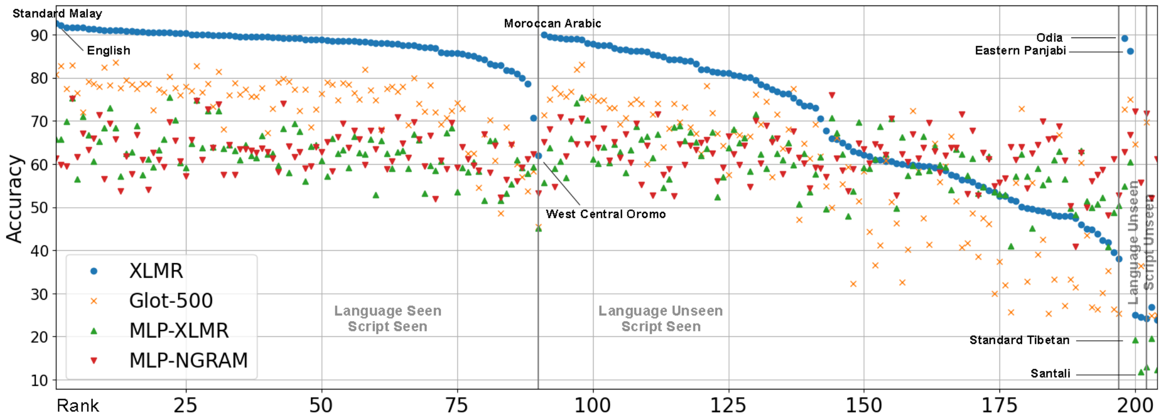

Figure 2: **Fully supervised Model Performance**. We group languages by whether they and their scripts are seen in the pre-training corpus of XLM-R. Languages are ordered by the XLM-R performance in every group.

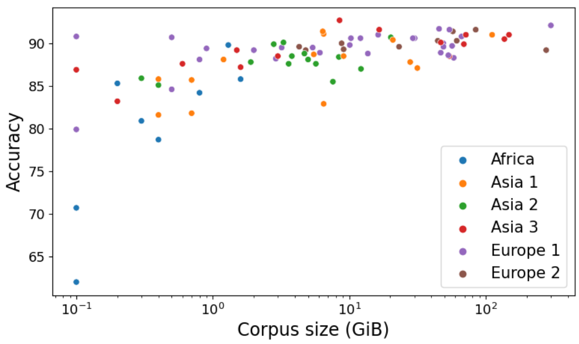

Figure 3: Accuracy of the XLM-R model vs Pre-Training corpus size in the fully supervised scenario. Bigger pre-training corpus in a target language generally improves the model performance.

On Joshi's class 0, African languages are even at least 20% worse than languages from other continents. Notably, there is no language with Joshi's class >3 in Africa and no American/Oceania languages have Joshi's class >1. *African and Oceania languages are also the only exceptions where MLP outperforms XLM-R*, implying a poorly learned representation of them. Future research should focus more on languages from these regions. Appendix B provides the evaluation across the eight sub-regions instead of four in Figure 1.

Performances across models In the fully supervised scenario, XLM-R performs the best on 16 out of the 22 language families. Among the remaining 6 language families, applying the simplest MLP classifier with n-gram input features outperforms more complex transformer-based MLMs (Glot-500 and XLM-R), suggesting they are not well adapted to these 6 language families. *Glot-500, despite being pre-trained with many more languages, outperforms XLM-R only on Sino-Tibetan languages*. Even on Sino-Tibetan languages, it fails to outperform the simplest MLP baseline. Cross-lingual transfer results are similar when using different source languages. On most language families, the results are comparable to fully supervised ones. Zero-shot prompting leads to a big drop due to the lack of supervised samples. The performance is good only for a few language families such as Indo-European, Uralic, Japonic and Koreanic.

5.2 Factors affecting performance

In order to determine the critical factor in this multilingual classification task, we conducted in-depth case studies on the model architecture choices and language categorizations.

Effect of language coverage in pre-training Figure 2 compares MLP, XLM-R and Glot-500 models based on language and script coverage in pre-training based on four groups: (1) language seen, script seen in XLM-R (2) language unseen, script seen in XLM-R (3) script unseen in XLM-R, language seen in Glot-500 (4) script unseen by both models. The results in each group are sorted by their performance on fine-tuned XLM-R model. Overall, *XLM-R performs the best all languages seen in its pre-training corpus without any exception*. Even for languages unseen in the pre-training corpus of XLM-R, it outperforms Glot-500 in most cases as long as the written scripts are seen. Glot-500 performs the best only for 3 out of all the 204 languages, implying their learned representations are far from sufficient. The reason could be that Glot-500 is pre-trained and evaluated on a religious corpus, which is quite different from the news domain in our task. In order to achieve a better generalization, we may have to mix text from various domains in the pre-training stage.

Effect of pre-training size Figure 3 shows the change of accuracy scores with increasing corpus included in the pre-training stage of XLM-R, where the corpus size is logarithmically scaled for better visualization. We can see that *with as little as 0.1GB pre-training corpus, the XLM-R model can already achieve >80% accuracy for almost all languages*, which further verified that this task itself is not difficult. Though the accuracy generally grows with increasing corpus size, and the model performance starts to saturate with $>$ 1GB pre-training corpus. Since African languages typically have tiny pre-training corpora or unavailable while European languages have large ones. This can explain the poor model performance on African languages.

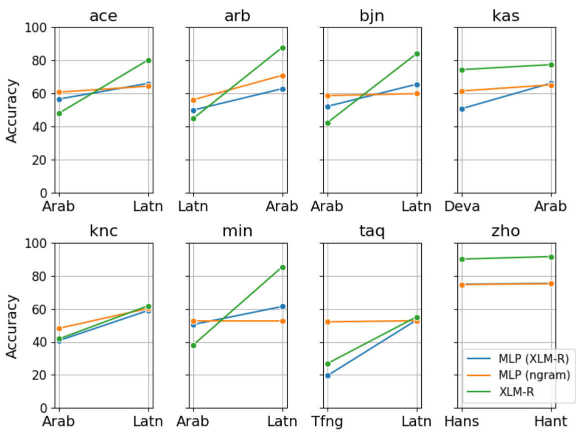

Figure 4: Script performance differences when one language has two different scripts. XLM-R and MLPs show the same trend. Using ngram features are more robust to script changes than using the XLM-R tokenizer.

Effect of script To see how the choice of scripts affects the model performance, we choose eight languages that can be written in different scripts, and visualize the performance of XLM-R, MLP with n-gram features (MLP-ngram), and MLP with words from the XLM-R tokenizer (MLP-XLM-R) in Figure 4. We can see that (1) The performance of MLP-XLMR usually correlates with that of XLM-R. This implies that under the XLM-R tokenizer, languages have their own preferred written scripts regardless of the effects from pre-training (because this preference stays the same even with the simplest MLP classifier); (2) The slope of XLM-R is often steeper than that of MLP-XLMR, implying the preferred script for a language also has better pre-trained representations; (3) The slope of MLP-n-gram is often less steep. This implies that n-gram features are more robust across different scripts compared with word features obtained from the XLM-R tokenizer; (4) The preferred script is often the more commonly used one for every language, suggesting *future work can focus on one single preferred script for every language*.

5.3 Comparison of different scenarios

Fine-tune vs. Prompted Out of all the 204 languages, GPT-4 outperforms GPT-3.5-turbo in 157 languages. Only on Buginese, Kabiyè, Mizo, Nuer and Ayacucho Quechua, GPT-3.5-Turbo outperforms GPT-4 for $>$ 10%. However, zero-shot prompting consistently underperforms fine-tuned methods. This might be due to the subjectivity of this task. It is hard to include extensive descriptions of the classification criteria in the prompt. Adding more examples to the prompt might improve the performance.

Cross-Lingual transfer vs Fully supervised Figure 5 takes a closer look at the comparison between cross-lingual transfer and fully-supervised methods. We can see that for all languages that are included in the pre-training corpus of XLM-R, the cross-lingual transfer performs similarly to fully supervised methods. The best source language for cross-lingual transfer is, surprisingly, French, rather than English, which has the largest amount of pre-training corpus, though the difference among various source languages is tiny. This suggests *languages included in the XLM-R pre-training corpus are pretty well aligned with all the four chosen high-resource languages*. The advantage of fully supervised methods over cross-lingual transfer becomes prominent mainly when the target language is not included in the pre-training corpus of XLM-R and its script is included. In this case, fully supervised methods can improve the performance by fine-tuning the model on the target languages, but cross-lingual transfer fails to capture the alignment with high-resource languages.

5.4 Region-specific pre-training

Evaluation of region-specific PLMs While our evaluation is primarily focused on multilingual PLMs trained on 100 languages or more, models pre-trained on a group of linguistically or geographically related languages often lead to better performance as observed for Indian languages (Table 5)

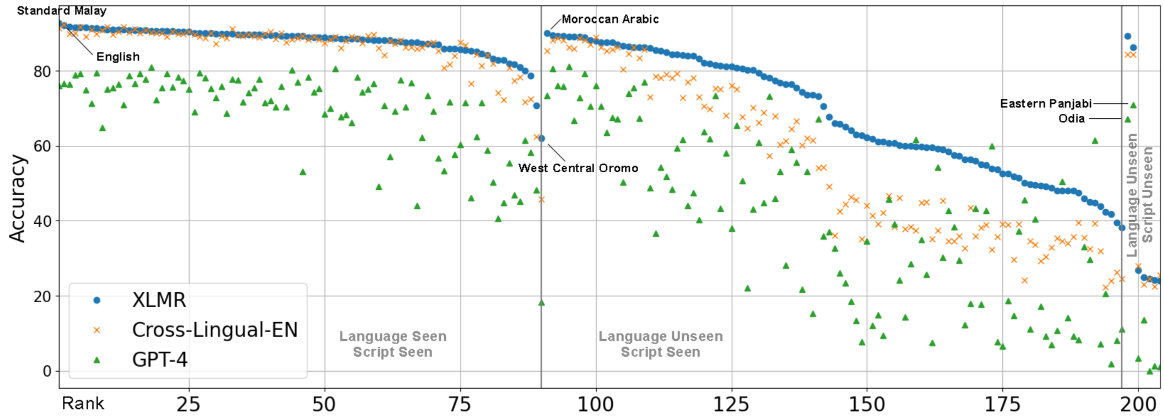

Figure 5: **Comparison of Various Scenarios**. We group languages by whether they and their scripts are seen in the pre-training corpus of XLM-R. Languages are ordered by the XLM-R fully-supervised performance in every group.

and African languages (Table 6). IndicBERTv2 and MurilBERT achieved better overall performance over XLM-R (550M parameters) despite their smaller capacity (236M-278M parameters), especially for Indian languages they both support, and better for languages not covered by XLM-R. Similarly for African languages, AfroXLMR—an adaptation of XLM-R through multilingual adaptive fine-tuning (MAFT) (Alabi et al., 2022) to 17 African languages gave roughly +9 improvement in performance. AfriBERTa on the other hand slightly gave worse result than XLM-R despite seeing the same number of African languages during pre-training (although not the exact languages) because it was pre-trained on less amount of data (1GB). Despite the improvement of AfroXLMR, it performs terribly for Nilotic, Mande and many Atlantic Congo families which shows that including more African languages in pre-training could improve performance.

Models	Language Family Indo-E (18)	Dravidian (4)	Austro-Asia (1)	Sino-Tib (1)	All (24)
XLM-R	86.5	87.9	24.6	48.7	82.6
IndicBERTv2	85.4	88.3	**65.5**	43.2	83.3
MurilBERT	**87.5**	**89.9**	23.5	**66.3**	**84.4**

Table 5: **Indic-centric Evaluation on SIB-200**.

Models	Language Family A.Congo (34)	Afro A. (12)	Nilo. (5)	Mande (2)	Aust. (1)	Indo-E (1)	All (55)
MLP	61.3	59.6	58.6	57.4	61.1	57.6	60.5
AfriBERTa	58.7	50.9	54.2	49.6	50.5	53.7	56.1
XLM-R	57.9	65.4	53.7	48.7	85.3	89.8	59.9
AfroXLMR	70.8	69.2	55.7	57.1	**88.4**	**90.4**	69.2
AfroXLMR-61	74.8	68.3	57.2	56.0	88.2	89.1	71.6
AfroXLMR-75	**77.1**	**69.5**	**62.3**	**64.0**	87.5	89.7	**74.1**

Table 6: **African-centric Evaluation on SIB-200**.

Performance of applying MAFT to more African languages We evaluated on two MAFT models described in (§4.1). Our evaluation of AfroXLMR-75 shows that MAFT with synthetic data was effective in improving the performance over AfroXLMR across many languages in Africa, especially for Nilotic (+6.6), Mande (+6.9) and Atlantic-Congo (+6.3) languages, similar to the findings of Urbizu et al. (2023). The performance improvement for AfroXLMR-61 was smaller on average (+2.4). There are few cases where it leads to a slight drop in performance on more-resourced languages due to curse-of-multilinguality (Conneau et al., 2020). The two newly developed PLMs are available on HuggugFace.[11] We provide the full result in Appendix C.

6 Related Work

Multilingual evaluation datasets: There have been several efforts to curate multilingual evaluation datasets, including various downstream tasks such as part-of-speech tagging (Nivre et al., 2016, 2020; Dione et al., 2023), named entity recognition (Pan et al., 2017; Adelani et al., 2022b; Mhaske et al., 2023), natural language inference (Conneau et al., 2018), text classification (Ma et al., 2023), machine translation (Adelani et al., 2022a; Goyal et al., 2022; NLLB-Team et al., 2022), and question answering (Lewis et al., 2020; Ogundepo et al., 2023; Shen et al., 2023; Doddapaneni et al., 2023; Bandarkar et al., 2023). All these initiatives have played a pivotal role in advancing the field of cross-

[11] https://huggingface.co/Davlan

lingual and multilingual NLP. Our work, which focuses on the creation of an extensive multilingual text classification dataset covering 200 languages, builds upon a line of related works that have significantly contributed to the expansion of the NLP community.

Specific to text classification, a few multilingual datasets are IndicNLP BBC news (Kunchukuttan et al., 2020), KINNEWS & KIRNEWS (Niyongabo et al., 2020), ANTC (Alabi et al., 2022), MasakhaNEWS (Adelani et al., 2023), and Taxi1500 (Ma et al., 2023). To the best of our knowledge, Taxi1500 is the most recent and largest of them all covering 1500 languages. However, this dataset is focused on the religious domain as the data comes from the Bible. Our work addresses a gap in multilingual text classification datasets by curating SIB-200 that covers a broader range of topics and domains.

Multilingual Large Language Models: In this work, we evaluated two categories of LLMs based on their pretraining objectives. These are Masked Language Models and Autoregressive Language Models (Yang et al., 2023; Zhao et al., 2023). Masked Language Model is the training paradigm for BERT-style[12]. On the other hand, Autoregressive Language Models, which are GPT-style decoder-only architectures, are trained by generating the next word in a sequence given the preceding words.

While it is possible to train these models on a single language (Conneau and Lample, 2019), such monolingual models typically demonstrate less cross-lingual capabilities compared to their multilingual variants. Hence, recent developments have seen the emergence of multilingual LLMs like mBERT (Devlin et al., 2019), XLM-Roberta (Conneau et al., 2020), Glot-500 (ImaniGooghari et al., 2023), XGLM (Lin et al., 2022), and GPT-3 (Brown et al., 2020). These models are trained on a diverse set of languages, primarily high-resource ones due to available corpora. This also includes the development of region-specific models like AfriBERTa (Ogueji et al., 2021), MuRIL (Khanuja et al., 2021), and IndicBERTv2 (Doddapaneni et al., 2023). However, these models tend to underperform on low-resource languages unseen during pretraining despite their cross-lingual capabilities (Philippy et al., 2023; Winata et al., 2022).

Consequently, there have been works to enhance LLMs for a broader range of languages, including vocabulary expansion (Wang et al., 2019), the use of lexicon (Wang et al., 2022), multilingual adaptive fine-tuning (MAFT) (Alabi et al., 2022; ImaniGooghari et al., 2023), and parameter efficient methods (Pfeiffer et al., 2021, 2022). All these techniques that have been shown to improve the performance of LLMs for low-resource languages and in cross-lingual transfer settings. In this work, we leveraged MAFT to adapt an existing LLM to 75 languages. In addition, we evaluated several LLMs including region-specific LLMs and we showed the capabilities of the different models on SIB-200.

7 Conclusion

In this paper, we created SIB-200—a large scale open-sourced benchmark dataset for topic classification in 200 languages and dialects to address the lack of evaluation datasets for natural language understanding especially for low-resource languages. We performed extensive evaluation across full-supervised setting, cross-lingual transfer setting and prompting of LLMs settings. Furthermore, we grouped the 200 languages in different categories based on language families, geographical regions, Joshi's class and coverage in multilingual pre-trained language models to provide insights into which group of languages have poor performance on this simple and inclusive benchmark.

Our findings are (1) There is a large performance gap between high-resource languages and low-resource ones, especially for under-represented languages concentrated in Africa, Americas, Oceania and south east Asia. (2) Written scripts can have big impact to the model performance. We suggest using one single preferred script for every language. (3) Including languages into the pre-training corpus is important for learning good representations and cross-lingual alignments. (4) It is crucial to mix text from various domains in the pre-training stage. Pre-training in a single domain can significantly deteriorate the performance on other domains. (5) Continued pre-training to cover more languages is an effective improving method, even if pre-training on synthetic data obtained from MT systems.

We hope our dataset will encourage a more inclusive evaluation of multilingual language models on a more diverse set of languages.

[12] BERT is one of the first neural language models that uses an with-encoder, masked pre-training, discriminative task.

8 Limitations

One of the main limitation of our work is that the labelled dataset created for other non-English languages are based on human translation and may suffer from translationese effect including a slight drop in performance. However, we believe this is an important contribution for many languages that often do not have news articles or Wikipedia articles that can be used for such annotation.

Another limitation is the choice of multilingual pre-trained language models, we note that XLM-R may not be the best multilingual encoder out there, there are other publicly available ones like InfoXLM (Chi et al., 2021), mDeBERTa (He et al., 2023) and others, however due to the scale of the experiments, we limited our evaluation to three multilingual models (XLM-R-base, XLM-R, and Glot-500). We believe our result may still be consistent with newer PLMs since they often cover similar set of languages as XLM-R.

9 Acknowledgement

David Adelani acknowledges the support of DeepMind Academic Fellowship programme. Jesujoba Alabi was partially funded by the BMBF project SLIK under the Federal Ministry of Education and Research grant 01IS22015C. This work was supported in part by Oracle Cloud credits and related resources provided by Oracle. We thank Google for providing GCP credits to train the AfroXLMR-61 model. Finally, we are grateful to OpenAI for providing API credits through their Researcher Access API programme to Masakhane for the evaluation of GPT-3.5 and GPT-4 large language models.

References

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. 2022a. 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*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. 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*, pages 3576–3588, Online. Association for Computational Linguistics.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. Muril: Multilingual representations for indian languages. *ArXiv*, abs/2103.10730.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Chunlan Ma, Ayyoob ImaniGooghari, Hao-Tong Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *ArXiv*, abs/2305.08487.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

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.

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen H. Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, Atnafu Lambebo Tonja, Iyanuoluwa Shode, Akari Asai, Tunde Oluwaseyi Ajayi, Clemencia Siro, Steven Arthur, Mofetoluwa Adeyemi, Orevaoghene Ahia, Aremu Anuoluwapo, Oyinkansola Awosan, Chiamaka Chukwuneke, Bernard Opoku, Awokoya Ayodele, Verrah Otiende, Christine Mwase, Boyd Sinkala, Andre Niyongabo Rubungo, Daniel A. Ajisafe, Emeka Felix Onwuegbuzia, Habib Mbow, Emile Niyomutabazi, Eunice Mukonde, Falalu Ibrahim Lawan, Ibrahim Said Ahmad, Jesujoba O. Alabi, Martin Namukombo, Mbonu Chinedu, Mofya Phiri, Neo Putini, Ndumiso Mngoma, Priscilla A. Amuok, Ruqayya Nasir Iro, and Sonia Adhiambo. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings* of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Mamdouh Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1, 000+ languages. *ArXiv*, abs/2305.13516.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. 2023. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada. Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. Mdia: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint arXiv:2208.13078*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Overall result

Table 7 shows the overall results for all languages.

B African languages result

Figure 6 shows the box plot for each region.

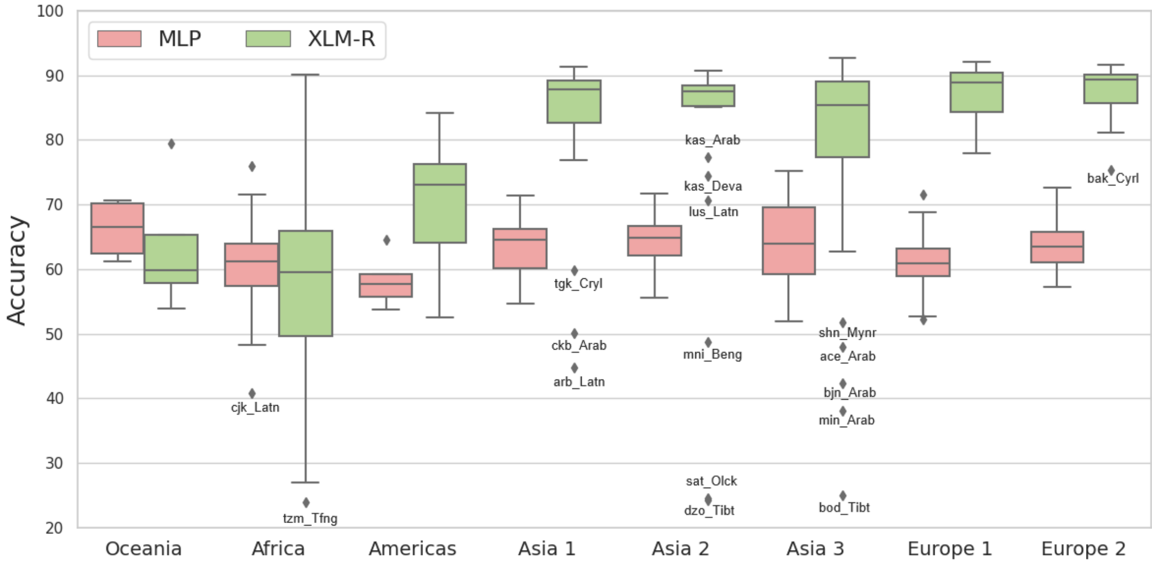

Figure 6: Box plot comparison of MLP and XLM-R model across regions

C African languages result

Table 8 shows the overall results for all languages.

						Fully Supervised			Cross-lingual XLM-R-based			Prompting	
Language name	Language code	Joshi's class	in XLM-R?	Language family	region	MLP	XLM-R base	XLM-R	eng	ara	zho	GPT-3.5	GPT-4
Acehnese (Arabic)	ace_Arab	1		Austronesian	Asia 3	**60.77**	35.7	48	34	32.5	33.2	9.1	14.2
Acehnese (Latin)	ace_Latn	1		Austronesian	Asia 3	64.35	74	**80.2**	64.7	67.4	67.4	31.4	22.1
Mesopotamian Arabic	acm_Arab			Afro-Asiatic	Asia 1	68.16	86.9	89.5	88.1	88.8	**89.7**	62.6	80.6
Ta'izzi-Adeni Arabic	acq_Arab			Afro-Asiatic	Asia 1	65.97	87.5	87.9	88.8	**89.4**	89.1	63	76
Tunisian Arabic	aeb_Arab			Afro-Asiatic	Africa	66.85	83.9	**86.5**	84.6	85.7	85.3	54.2	73.9
Afrikaans	afr_Latn	3	✓	Indo-European	Africa	57.55	87.8	**89.8**	88.7	88.6	**89.8**	63.6	68.6
South Levantine Arabic	ajp_Arab			Afro-Asiatic	Asia 1	60.28	83.1	86.3	86.2	**87.1**	**87.1**	63	75.4
Akan	aka_Latn	1		Atlantic-Congo	Africa	**61.33**	53.7	59.7	44.8	42.6	45	31.1	35
Tosk Albanian	als_Latn	1	✓	Indo-European	Europe 1	63.6	87.7	89.5	89.8	**90.5**	**90.5**	66.5	71.4
Amharic	amh_Ethi	2	✓	Afro-Asiatic	Africa	66.87	80.1	84.2	84.1	**85.7**	83.1	23.4	58.8
North Levantine Arabic	apc_Arab			Afro-Asiatic	Asia 1	54.66	85.1	89.3	88.6	**89.4**	89	59.9	76
MSA (Arabic)	arb_Arab	5	✓	Afro-Asiatic	Asia 1	70.93	88.2	87.8	88.1	**89.1**	88.9	68.4	77.6
MSA (Romanized)	arb_Latn			Afro-Asiatic	Asia 1	56.12	36.3	44.8	39.3	39.6	39.9	44.7	**61.5**
Najdi Arabic	ars_Arab			Afro-Asiatic	Asia 1	70.87	88.9	**89.2**	88.2	89.1	89.1	67	75.9
Moroccan Arabic	ary_Arab			Afro-Asiatic	Africa	65.09	83	**90.1**	85.3	85.8	85.6	47.3	73.3
Egyptian Arabic	arz_Arab	3		Afro-Asiatic	Africa	61.82	84.8	**89.1**	86.2	88	87.9	58	81.1
Assamese	asm_Beng	1	✓	Indo-European	Asia 2	63.4	75.5	**88.1**	86.1	85.2	85.6	51.8	49.2
Asturian	ast_Latn	1		Indo-European	Europe 1	63.71	85.8	**87.5**	86.3	86.1	86.1	58.5	70.6
Awadhi	awa_Deva	0		Indo-European	Asia 2	64.68	85	88.1	87	**88.8**	88.1	59.7	70.6
Central Aymara	ayr_Latn	1		Aymaran	Americas	**55.68**	42.5	52.5	39.1	41.3	38.5	15.9	6.6
South Azerbaijani	azb_Arab	1	✓	Turkic	Asia 1	64.27	79	**82.9**	74.1	73.6	74.1	32.9	40.6
North Azerbaijani	azj_Latn	1	✓	Turkic	Asia 1	71.4	85	**91.1**	90.8	90.3	**91.1**	57.2	64.8
Bashkir	bak_Cyrl	1		Turkic	Europe 2	67.51	72.2	**75.4**	66.5	70.2	68.6	45	55.6
Bambara	bam_Latn	1		Mande	Africa	**64.43**	42.1	49.3	29.2	29	29.5	23.6	17.2
Balinese	ban_Latn	0		Austronesian	Asia 3	64.3	79	**83.9**	78	79.7	76.9	51.7	47.4
Belarusian	bel_Cyrl	3	✓	Indo-European	Europe 2	63.48	86.4	**89.6**	**89.6**	89.3	89.3	54.3	74.2
Bemba	bem_Latn	0		Atlantic-Congo	Africa	**63.76**	52.9	59.5	44.9	45.9	45.3	21.8	25.6
Bengali	ben_Beng	3	✓	Indo-European	Asia 2	65.76	83	**88.4**	87.4	87.5	85.6	59.6	74.4
Bhojpuri	bho_Deva	1		Indo-European	Asia 2	69.37	82	**86.3**	83.3	83.2	84.1	49.4	67.4
Banjar (Arabic script)	bjn_Arab	1		Austronesian	Asia 3	**58.68**	36.6	42.3	22.2	21.8	22.5	12.1	20.5
Banjar (Latin script)	bjn_Latn	1		Austronesian	Asia 3	59.87	79.8	**84**	77	78.5	76.8	34.2	44.1
Standard Tibetan	bod_Tibt	1		Sino-Tibetan	Asia 3	**72.14**	24.8	25	23	20.9	20.7	5.8	13.6
Bosnian	bos_Latn	3	✓	Indo-European	Europe 1	57.68	87.8	90.8	91.1	90.4	**92.3**	65.7	76.6
Buginese	bug_Latn	1		Austronesian	Asia 3	54.38	72.6	**73.5**	61.5	63.1	64.1	26.9	15.2
Bulgarian	bul_Cyrl	3	✓	Indo-European	Europe 2	66.01	88.4	**91.4**	89.9	89.3	90.1	66.5	79.4
Catalan	cat_Latn	4	✓	Indo-European	Europe 1	64.49	88.6	89.8	91.1	91.4	**91.8**	55.4	77.7
Cebuano	ceb_Latn	3		Austronesian	Asia 3	62.5	77.9	**81.5**	75.3	77.2	76.6	62.3	73.4
Czech	ces_Latn	4	✓	Indo-European	Europe 1	53.67	88.7	91	91.8	91	**92.3**	62.4	70.9
Chokwe	cjk_Latn			Atlantic-Congo	Africa	40.83	43.8	**47.5**	39.5	40.8	39.9	14.3	8.2
Central Kurdish	ckb_Arab	0		Indo-European	Asia 1	**63.99**	37.7	50.1	24.1	25.7	24.9	45.6	45.6
Crimean Tatar	crh_Latn	1		Turkic	Europe 2	61.48	80.9	**86.6**	80.3	82.2	81.3	34.5	50.2
Welsh	cym_Latn	1	✓	Indo-European	Europe 1	67.76	81.3	**88.1**	84.2	84.5	83.8	59.4	70.7
Danish	dan_Latn	3	✓	Indo-European	Europe 1	59.42	88	**91.7**	89.8	90	91	69.6	76.3
German	deu_Latn	5	✓	Indo-European	Europe 1	61.81	88.4	90.8	90.2	91	**91.1**	70.7	78.6
Southwestern Dinka	dik_Latn	1		Nilotic	Africa	**64.58**	51.1	61	39.2	40	38.4	24	14.9
Dyula	dyu_Latn	0		Mande	Africa	**50.34**	43.3	48	35.8	35.1	35	12.3	9.3
Dzongkha	dzo_Tibt	1		Sino-Tibetan	Asia 2	**71.66**	26	24.2	22.4	20	20	0	1.2
Greek	ell_Grek	3	✓	Indo-European	Europe 1	59.04	85.5	88.9	88.4	89.1	**90.7**	65	78.3
English	eng_Latn	5	✓	Indo-European	Europe 1	59.91	90	92.1	92.2	91.2	**92.5**	71.8	76.6
Esperanto	epo_Latn	1	✓	Constructed	Europe 1	61.37	87.5	89.4	88.5	88.6	**90.4**	58.7	70.3
Estonian	est_Latn	3	✓	Uralic	Europe 1	59.56	84.4	88.9	89	90.2	**90.4**	61.9	74.4
Basque	eus_Latn	4	✓	Basque	Europe 1	62.88	83.5	89.2	89.2	88.9	**89.7**	55.3	53.1
Ewe	ewe_Latn	1		Atlantic-Congo	Africa	**71.54**	49.2	56.4	32.7	31.7	33.4	20.4	12.2
Faroese	fao_Latn	1		Indo-European	Europe 1	71.5	80.2	**85.3**	78.1	79	78.6	49.5	54.2
Fijian	fij_Latn	1		Austronesian	Oceania	**70.56**	54	60.7	38.4	38	38.4	40.1	39.1
Finnish	fin_Latn	4	✓	Uralic	Europe 1	67.01	89.2	**91.6**	89	90.1	90.1	65	74.9
Fon	fon_Latn			Atlantic-Congo	Africa	**65.94**	46.2	48.1	35.4	30.6	32.7	13.8	10.8
French	fra_Latn	5	✓	Indo-European	Europe 1	57.57	89.2	89.7	89.5	89.7	**90.1**	73.2	77.5
Friulian	fur_Latn	1		Indo-European	Europe 1	63.24	77.2	**83.2**	72.9	73.9	73	41.8	40.2
Nigerian Fulfulde	fuv_Latn	0		Atlantic-Congo	Africa	58.9	53.9	**63**	45.6	46.1	46.6	15.5	13.4
West Central Oromo	gaz_Latn	1	✓	Afro-Asiatic	Africa	53.21	34.2	**62**	45.8	48.7	43.3	27.2	18.4
Scottish Gaelic	gla_Latn	0	✓	Indo-European	Europe 1	59.21	60.7	**79.9**	71.7	73.9	73.2	49.7	61.5
Irish	gle_Latn	2	✓	Indo-European	Europe 1	58.59	72	**84.6**	81.3	82.5	82.1	52.3	71.4
Galician	glg_Latn	3	✓	Indo-European	Europe 1	67.7	89.2	88.2	89.6	90	**91.5**	59.5	76.5
Guarani	grn_Latn	1		Tupian	Americas	57.71	69.6	**76.3**	61.3	61.1	61.7	32.3	28.2
Gujarati	guj_Gujr	1	✓	Indo-European	Asia 2	60.59	83.8	**87.8**	87.1	87.5	86.9	65	69.2
Haitian Creole	hat_Latn	0		Indo-European	Americas	59.21	57.3	**73.1**	54	55.2	54	52.2	67.2
Hausa	hau_Latn	2	✓	Afro-Asiatic	Africa	55.99	72.5	**80.9**	78.3	80.4	77.2	38.2	45.1
Hebrew	heb_Hebr	3	✓	Afro-Asiatic	Asia 1	59.44	86.5	87.1	87.1	87.4	**88.2**	60.5	73.3
Hindi	hin_Deva	4	✓	Indo-European	Asia 2	66.71	83.9	**90.7**	90.1	90.4	89.9	63.1	79.3
Chhattisgarhi	hne_Deva			Indo-European	Asia 2	68.17	82.8	**87.5**	85.3	86.6	86	52.7	63.6
Croatian	hrv_Latn	4	✓	Indo-European	Europe 1	60.92	89.7	90.7	90.7	90.3	**91.5**	66.6	77.7
Hungarian	hun_Latn	4	✓	Uralic	Europe 1	59.86	86	88.3	89.3	88.6	**90**	60.3	75
Armenian	hye_Armn	1	✓	Indo-European	Asia 1	65.17	86.1	88.7	88.3	88.3	**89.2**	40.1	70
Igbo	ibo_Latn	1		Atlantic-Congo	Africa	**63.87**	46.3	57.5	34.6	35.7	33.2	31.3	38.4
Ilocano	ilo_Latn	1		Austronesian	Asia 3	63.11	72.3	**76.3**	64.4	66.9	66.6	59.6	58.9
Indonesian	ind_Latn	3	✓	Austronesian	Asia 3	56.54	88.9	91	91.7	**92**	91.6	67.3	75.1
Icelandic	isl_Latn	2	✓	Indo-European	Europe 1	62.83	85.1	89.5	90.4	90	**91.4**	62.5	72.1
Italian	ita_Latn	4	✓	Indo-European	Europe 1	54.05	89	90.6	90.2	**90.8**	90.6	62.8	81
Javanese	jav_Latn	1	✓	Austronesian	Asia 3	57.99	81.7	**83.2**	81.8	83	81.9	43	50.3
Japanese	jpn_Jpan	5	✓	Japonic	Asia 3	73.83	87.9	**89.9**	89.3	88.8	89.7	63.4	75.8
Kabyle	kab_Latn	1		Afro-Asiatic	Africa	**61.13**	36	39.5	26.3	24.5	26	15.2	8
Jingpho	kac_Latn	0		Sino-Tibetan	Asia 3	**64.75**	54.7	62.7	35.2	33.9	35.7	10	7.8

Kamba	kam_Latn	0		Atlantic-Congo	Africa	**56.67**	48.2	52.5	39	38.4	40.1	19.7	18.7
Kannada	kan_Knda	1	✓	Dravidian	Asia 2	65.56	86.5	**90.1**	89.6	89.7	89.9	60.1	69.1
Kashmiri (Arabic)	kas_Arab	1		Indo-European	Asia 2	65.16	68	**77.4**	67.8	70	69.1	33	46
Kashmiri (Devanagari)	kas_Deva	1		Indo-European	Asia 2	61.49	62.8	**74.4**	60.2	63.7	59.3	22.9	21.7
Georgian	kat_Geor	3	✓	Kartvelian	Asia 1	63.72	83.4	88.5	89.1	88.6	**89.7**	44.7	66.1
Kazakh	kaz_Cyrl	3	✓	Turkic	Asia 1	63.37	85.3	**91.4**	90.9	89.7	89	63.7	71.2
Kabiyè	kbp_Latn	1		Atlantic-Congo	Africa	**69.88**	37.9	49.1	30.4	29.8	30.5	23.7	9.3
Kabuverdianu	kea_Latn			Indo-European	Africa	64.26	78.9	**86.1**	73	75.7	72.3	47.5	48.8
Halh Mongolian	khk_Cyrl	1	✓	Mongolic-Khitan	Asia 3	66.21	82.9	**88.5**	86.1	86.2	85.5	57.7	67.6
Khmer	khm_Khmr	1	✓	Austroasiatic	Asia 3	74.08	85.8	**89.2**	87.5	86.8	87.7	39.2	70.4
Kikuyu	kik_Latn	1		Atlantic-Congo	Africa	**60.47**	50	59.9	38.3	39.4	38.5	28.5	28.6
Kinyarwanda	kin_Latn	1		Atlantic-Congo	Africa	**68.77**	45.3	48	34.5	35.1	35.7	47.1	50.5
Kyrgyz	kir_Cyrl		✓	Turkic	Asia 1	58.8	84.6	**88.1**	87.9	87.1	87.5	49.8	57
Kimbundu	kmb_Latn			Atlantic-Congo	Africa	**54.4**	40.9	49.7	34.6	36.4	35.7	13	11.2
Northern Kurdish	kmr_Latn	1	✓	Indo-European	Asia 1	64.38	75.8	**81.6**	77.3	79.4	78.3	42.5	46.8
Kanuri (Arabic)	knc_Arab	0		Nilotic	Africa	**48.2**	39	41.8	23.9	21.6	21.2	2.2	1.8
Kanuri (Latin)	knc_Latn	0		Nilotic	Africa	60.21	58.2	**61.8**	41.4	41.8	41.3	16.8	12
Kikongo	kon_Latn			Atlantic-Congo	Africa	58.59	58.5	**65**	44.5	46.8	46.3	18.4	23.5
Korean	kor_Hang	4	✓	Koreanic	Asia 3	67.81	86.5	88.5	88.7	88.7	**89.2**	67.8	78.2
Lao	lao_Laoo	2	✓	Tai-Kadai	Asia 3	69.66	**88.2**	87.6	85.9	86.8	86.9	26.3	44
Ligurian	lij_Latn	1		Indo-European	Europe 1	56.12	76.2	**81.3**	75.1	75.7	74.3	29.2	43.2
Limburgish	lim_Latn	1		Indo-European	Europe 1	58.8	79.1	**84.3**	78.1	77.6	76.7	43.1	48.4
Lingala	lin_Latn	1		Atlantic-Congo	Africa	61.39	60.6	**65.8**	42.6	43.8	45	29.1	26
Lithuanian	lit_Latn	3	✓	Indo-European	Europe 1	64.02	87.6	88.8	88.4	89.1	**89.6**	59.2	75.3
Lombard	lmo_Latn	1		Indo-European	Europe 1	54.75	72.6	**80.1**	70.1	71.9	71.1	36.2	43
Latgalian	ltg_Latn	1		Indo-European	Europe 2	66.5	79	**81.1**	76	76.1	75.8	29.4	38
Luxembourgish	ltz_Latn	1		Indo-European	Europe 1	64.95	76.1	**82**	70.6	71.9	70.9	62.4	63.8
Luba-Kasai	lua_Latn			Atlantic-Congo	Africa	53.05	52.9	**56.3**	43.5	43.6	44.5	25.6	17.9
Ganda	lug_Latn			Atlantic-Congo	Africa	**49.93**	41.8	45	32.5	33.8	33.4	36.8	29.6
Luo	luo_Latn			Nilotic	Africa	**62.3**	51.1	60	37.8	39.3	40.1	19.7	14.4
Mizo	lus_Latn	0		Sino-Tibetan	Asia 2	58.97	68.6	**70.6**	54.2	56.4	56.4	48.7	35.9
Standard Latvian	lvs_Latn	3	✓	Indo-European	Europe 2	72.63	89.9	90	90	90.1	**91.7**	63.8	75.2
Magahi	mag_Deva	0		Indo-European	Asia 2	65.07	82.7	86.9	86.3	**87**	86.1	54.2	67.1
Maithili	mai_Deva	1		Indo-European	Asia 2	69.94	82.5	**89.1**	85.8	87.4	87	49.5	66.7
Malayalam	mal_Mlym	1	✓	Dravidian	Asia 2	63.89	83.3	85.5	86.7	**88**	86.9	59.7	71.5
Marathi	mar_Deva	2	✓	Indo-European	Asia 2	63.77	83.1	**89.9**	87.2	88.4	87.3	57.8	72.8
Minangkabau (Arabic)	min_Arab	1		Austronesian	Asia 3	**52.7**	34.9	38.1	24.6	23.5	23.8	8.1	11.2
Minangkabau (Latin)	min_Latn	1		Austronesian	Asia 3	52.68	81	**85.4**	78.5	79.4	79.4	40.2	36.7
Macedonian	mkd_Cyrl	1	✓	Indo-European	Europe 2	61.69	86.7	**89.2**	88.6	88	88.8	62.2	76.9
Maltese	mlt_Latn	2		Afro-Asiatic	Europe 1	68.84	69.4	**78**	57.2	59.3	60	63.5	73.1
Meitei (Bengali script)	mni_Beng			Sino-Tibetan	Asia 2	**65.83**	37.6	48.7	32.8	31	28.5	4	6.9
Mossi	mos_Latn			Atlantic-Congo	Africa	**69.58**	52.2	59.5	35.1	36.9	35.3	14.3	7.5
Maori	mri_Latn	1		Austronesian	Oceania	**61.19**	44	53.9	32.4	32.1	32.4	45.8	60
Burmese	mya_Mymr	1	✓	Sino-Tibetan	Asia 3	65.73	81.1	**87.2**	86.2	85.9	85.3	19	62.2
Dutch	nld_Latn	4	✓	Indo-European	Europe 1	59.26	88.1	**90.6**	89.1	88.7	89.2	68.9	79.2
Norwegian Nynorsk	nno_Latn	1	✓	Indo-European	Europe 1	63.2	85.8	89.6	89.2	88.8	**90.5**	62.9	75.5
Norwegian Bokmål	nob_Latn	1	✓	Indo-European	Europe 1	61.04	86.4	**90**	88.7	88.5	**90**	64.6	78.1
Nepali	npi_Deva	1	✓	Indo-European	Asia 2	69.42	85.1	**88.5**	88	87.4	**88.5**	62.6	68.2
Northern Sotho	nso_Latn	1		Atlantic-Congo	Africa	**62.61**	50.2	54.8	38.9	38.6	38.3	35.6	42.7
Nuer	nus_Latn	0		Nilotic	Africa	**57.58**	41.4	43.9	31.9	27.1	29	18	7.2
Nyanja	nya_Latn			Atlantic-Congo	Africa	**62.16**	53	60.7	46.7	46.9	49.2	46.8	45.7
Occitan	oci_Latn	1		Indo-European	Europe 1	60.88	82.3	**87.5**	85.7	85.8	85.6	56.8	67.5
Odia	ory_Orya	1		Indo-European	Asia 2	62.83	82.1	**89.2**	84.3	84.6	83.8	56.2	67.1
Pangasinan	pag_Latn	1		Austronesian	Asia 3	60.17	74	**78.5**	68.8	71.4	70.8	54.6	44.8
Eastern Panjabi	pan_Guru			Indo-European	Asia 2	66.83	82.5	**86.3**	84.4	84.9	83.2	67.6	70.9
Papiamento	pap_Latn			Indo-European	Americas	64.55	77.5	**84.2**	72.8	74.2	71.6	56.3	61.7
Southern Pashto	pbt_Arab	1	✓	Indo-European	Asia 1	56.39	80.5	**81.8**	80.7	81.7	80.8	52.3	55.3
Western Persian	pes_Arab	4	✓	Indo-European	Asia 1	65.72	88.6	**91**	90.4	89.7	90.2	64.3	76.4
Plateau Malagasy	plt_Latn	1	✓	Austronesian	Africa	61.09	74.1	**85.3**	76.6	78.2	74.3	42.8	46.1
Polish	pol_Latn	4	✓	Indo-European	Europe 2	60.72	88	90.3	89.8	**90.7**	**90.7**	66.8	77.4
Portuguese	por_Latn	4	✓	Indo-European	Europe 1	62.62	89.6	89.6	88.9	88.7	**90.1**	61.7	77.8
Dari	prs_Arab	0		Indo-European	Asia 1	64.65	86	89.1	88.6	89.2	**89.7**	59.5	72.7
Ayacucho Quechua	quy_Latn	1		Quechuan	Americas	53.72	57.9	**64.1**	46.3	50.8	49.1	36.2	18.5
Romanian	ron_Latn	3	✓	Indo-European	Europe 2	57.25	87	90.3	90.3	89.9	**91**	66.5	75.1
Rundi	run_Latn	0		Atlantic-Congo	Africa	**62.66**	45.7	46	35.5	37.4	39	40.8	33.1
Russian	rus_Cyrl	4	✓	Indo-European	Europe 2	64.04	88.8	89.2	88.4	88.1	**89.7**	62.8	80.2
Sango	sag_Latn	1		Atlantic-Congo	Africa	**63.97**	54.3	61	42.2	42.5	43.1	15.3	9.4
Sanskrit	san_Deva	2	✓	Indo-European	Asia 2	60.87	81.6	**85.9**	80.8	82	82.5	37.8	53.3
Santali	sat_Olck	1		Austroasiatic	Asia 2	**55.63**	24.5	24.6	24.5	21.9	22.7	1.8	0
Sicilian	scn_Latn	1		Indo-European	Europe 1	59.22	75.8	**81.1**	68.1	72.1	70.6	43.9	58.1
Shan	shn_Mymr	0		Tai-Kadai	Asia 3	**63.92**	36.9	51.8	29.7	31	28.9	19.3	14.7
Sinhala	sin_Sinh	0	✓	Indo-European	Asia 2	64.79	85.8	**87.6**	86.2	87.4	87.2	22.5	70.3
Slovak	slk_Latn	3	✓	Indo-European	Europe 2	63.52	87.1	**89.6**	89.2	88.7	88.8	62.6	71.7
Slovenian	slv_Latn	3	✓	Indo-European	Europe 1	62.71	89.7	**90.6**	90.1	89.8	90	61.5	72.3
Samoan	smo_Latn	1		Austronesian	Oceania	**62.93**	50.8	59.1	37.5	36	37.5	50.8	54.3
Shona	sna_Latn	1		Atlantic-Congo	Africa	**60.97**	46.7	51.5	39.2	41	40.6	37.2	37.2
Sindhi	snd_Arab	1	✓	Indo-European	Asia 2	59.71	75.4	**85.1**	84.3	83.8	83.9	49.1	62.4
Somali	som_Latn	1	✓	Afro-Asiatic	Africa	61.2	65.1	**78.7**	72.5	74	72.8	48	58.3
Southern Sotho	sot_Latn	0		Atlantic-Congo	Africa	**67.74**	51.6	55.9	35.9	37.3	36.8	38.2	43.2
Spanish	spa_Latn	5	✓	Indo-European	Europe 1	58.28	87.4	88.6	**89.1**	88	87.9	67.5	80.6
Sardinian	srd_Latn	1		Indo-European	Europe 1	59.81	74.3	**80.5**	67.7	71.6	69.7	36.1	50.6
Serbian	srp_Cyrl	4	✓	Indo-European	Europe 2	58.19	89.2	89.3	**90.1**	**90.1**	**90.1**	55.4	75.9
Swati	ssw_Latn	1		Atlantic-Congo	Africa	58.03	41.9	**59**	45.4	46	43.6	26.1	30.3
Sundanese	sun_Latn	1	✓	Austronesian	Asia 3	51.97	80.3	**86.9**	85.2	85.9	85.6	50.8	56.7
Swedish	swe_Latn	4	✓	Indo-European	Europe 1	61	88.3	90.6	90.1	89.8	**90.8**	66	75.5
Swahili	swh_Latn	2	✓	Atlantic-Congo	Africa	58.91	77	**85.8**	81.7	83.7	82.8	63.8	71.6
Silesian	szl_Latn	1		Indo-European	Europe 2	57.48	79.4	**84.9**	79.1	79.9	79	38.4	51.7

Tamil	tam_Taml	3	✓	Dravidian	Asia 2	66.53	86.9	87	87.6	87.6	**88.8**	48.7	69.2
Tamasheq (Latin)	taq_Latn			Afro-Asiatic	Africa	52.82	52.9	**55.1**	38.1	38.4	37.9	22.8	17.7
Tamasheq (Tifinagh)	taq_Tfng			Afro-Asiatic	Africa	**52.1**	27.1	26.9	28	28	27.8	3.1	3.3
Tatar	tat_Cyrl	1		Turkic	Europe 2	65.64	75	**81.9**	69.8	71.6	71.1	46.1	61.9
Telugu	tel_Telu	1	✓	Dravidian	Asia 2	62.79	81	**88.8**	87.3	86.8	87.1	60.4	68.4
Tajik	tgk_Cyrl	1		Indo-European	Asia 1	**66.98**	46.9	59.8	37.5	38.8	37.9	49.1	61.6
Tagalog	tgl_Latn	3		Austronesian	Asia 3	55.19	84.7	86.2	85.6	**86.8**	86.7	67.9	76.9
Thai	tha_Thai	3	✓	Tai-Kadai	Asia 3	69.47	87.5	**91**	89.7	89.7	90.8	61.1	75.5
Tigrinya	tir_Ethi	2		Afro-Asiatic	Africa	57.04	53.9	**67.7**	49.1	52.9	50	19.3	37.1
Tok Pisin	tpi_Latn	1		Indo-European	Oceania	69.99	75.9	**79.4**	66.1	67.2	66.5	63.8	60.8
Tswana	tsn_Latn	2		Atlantic-Congo	Africa	**61.92**	47.8	58.5	34.6	33.7	35.6	31.8	42.8
Tsonga	tso_Latn	1		Atlantic-Congo	Africa	**63.8**	52.7	57.2	35.9	38	38.8	34.2	29.4
Turkmen	tuk_Latn	1		Turkic	Asia 1	65.9	64.4	**76.8**	60.3	61.2	59.6	52.3	53.1
Tumbuka	tum_Latn	1		Atlantic-Congo	Africa	**75.92**	53.8	66	36.1	39.9	38.1	28.5	32.6
Turkish	tur_Latn	4	✓	Turkic	Asia 1	65.4	87.4	90.4	89.8	89.9	**91.1**	66.1	78.2
Twi	twi_Latn	1		Atlantic-Congo	Africa	**63.43**	57.4	62.2	44	42.6	43.8	33	34.6
Tamazight	tzm_Tfng			Afro-Asiatic	Africa	**61.18**	24.4	23.9	25.5	25.2	24.3	4.1	1.1
Uyghur	uig_Arab	1	✓	Turkic	Asia 1	69.58	81.5	85.8	83.9	**85.9**	83.9	46.4	57.7
Ukrainian	ukr_Cyrl	3	✓	Indo-European	Europe 2	61.68	89.6	91.6	91.7	90.6	**92.2**	67.1	79.2
Umbundu	umb_Latn	0		Atlantic-Congo	Africa	**54.93**	48.8	53.6	35.8	36.8	35	15.2	7.7
Urdu	urd_Arab	3	✓	Indo-European	Asia 2	61.51	84.3	**87.6**	86.2	86.5	85.8	64.3	76.8
Northern Uzbek	uzn_Latn	3	✓	Turkic	Asia 1	59.2	82.1	85.7	87.5	**87.6**	87.3	57.6	60.3
Venetian	vec_Latn	1		Indo-European	Europe 1	52.64	79.6	**84.3**	79.3	79.7	78.9	44.1	59.4
Vietnamese	vie_Latn	4	✓	Austroasiatic	Asia 3	69.73	88.3	90.5	90.4	89.9	**91.2**	63.3	75.7
Waray	war_Latn	1		Austronesian	Asia 3	62.36	77.9	**80.7**	75.1	78.7	77.5	64.6	65.4
Wolof	wol_Latn	2		Atlantic-Congo	Africa	52.67	57.1	**60.1**	46.1	46.8	47.9	29.5	24.1
Xhosa	xho_Latn	2	✓	Atlantic-Congo	Africa	62.29	54	**70.7**	62.3	64.3	60.7	37.7	48.2
Eastern Yiddish	ydd_Hebr	1	✓	Indo-European	Europe 1	52.3	57.5	**82.9**	72.2	74.4	70.5	33.2	44.8
Yoruba	yor_Latn	2		Atlantic-Congo	Africa	**62.55**	42.4	49.6	33.6	32.5	30.6	33.8	40.5
Yue Chinese	yue_Hant			Sino-Tibetan	Asia 3	70.52	88.6	**88.9**	88.2	88.1	88	63.9	79.2
Chinese (Simplified)	zho_Hans	5	✓	Sino-Tibetan	Asia 3	74.67	**90.2**	90.1	89.1	88.8	89.9	60.6	79.5
Chinese (Traditional)	zho_Hant	1	✓	Sino-Tibetan	Asia 3	75.2	88.2	**91.6**	90.1	89.1	89.6	64.2	78.8
Standard Malay	zsm_Latn	3	✓	Austronesian	Asia 3	61.39	90	**92.7**	91.3	91.5	91.3	63.4	76.1
Zulu	zul_Latn	2		Atlantic-Congo	Africa	60.13	48.4	**73.5**	62	63.6	60.1	39.3	53.1
Average						62.3	70.9	**76.1**	68.8	69.3	69.1	45.1	52.6

Table 7: **Overall result of the performance of different text classification models across different languages.** We compared different settings: fully-supervised, cross-lingual transfer and zero-shot prompting of LLMs. We report cross-lingual transfer performances from 4 source languages: English, French, Chinese and Arabic.

Language	Language code	in XLM-R?	in Afri-BERTa?	in Afro-XLMR?	in Afro XLMR-61	in Afro XLMR-75	XLMR	Afri BERTa	Afro XLMR	Afro XLMR-61	Afro XLM-75
Tunisian Arabic	aeb_Arab	✓		✓	✓	✓	86.5	25.4	86.1	86.7	**87.3**
Moroccan Arabic	ary_Arab	✓		✓	✓	✓	**90.1**	26	87.2	87.3	87.4
Egyptian Arabic	arz_Arab	✓		✓	✓	✓	**89.1**	26.4	88.7	86.6	87.1
Afrikaans	afr_Latn	✓		✓	✓	✓	89.8	53.7	**90.4**	89.1	89.7
Akan	aka_Latn				✓	✓	59.7	52.6	59.4	74.9	**77.4**
Amharic	amh_Ethi	✓	✓	✓	✓	✓	84.2	80.2	**88.6**	87.2	85.9
Bambara	bam_Latn					✓	49.3	55.4	59.3	59.4	**71.3**
Bemba	bem_Latn				✓	✓	59.5	55.7	74.1	**80.8**	79.7
Chokwe	cjk_Latn					✓	47.5	40.9	48.9	56.6	**59.8**
Dinka	dik_Latn					✓	61	62.3	60.9	61.3	**63.7**
Dyula	dyu_Latn					✓	48	43.8	54.9	52.6	**56.7**
Ewe	ewe_Latn				✓	✓	56.4	61.6	59.5	71.4	**76.3**
Fon	fon_Latn				✓	✓	48.1	54.7	54.5	61.7	**68.3**
Nigerian Fulfulde	fuv_Latn				✓	✓	63	57.5	60.8	62.5	**64.2**
Oromo	gaz_Latn	✓	✓	✓	✓	✓	62	74.6	**82.6**	81.2	78.6
Hausa	hau_Latn	✓	✓	✓	✓	✓	80.9	80.4	**86.4**	85.6	83.3
Igbo	ibo_Latn		✓	✓	✓	✓	57.5	79.6	83.7	**83.9**	81.1
Kabyle	kab_Latn					✓	39.5	44	35.1	34.9	**51.2**
Kamba	kam_Latn					✓	52.5	53.7	59.6	59.4	**68.5**
Kabiyè	kbp_Latn					✓	49.1	55.2	58.8	59.2	**71.4**
Kanuri (Arabic)	knc_Arab					✓	41.8	36.1	47.7	**48.4**	46.9
Kanuri (Latin)	knc_Latn					✓	61.8	58.2	61.4	**62.7**	62.3
Kikuyu	kik_Latn				✓	✓	59.9	57.3	65.8	71.6	**79.6**
Kinyarwanda	kin_Latn		✓	✓	✓	✓	48	79.9	84.2	**85.3**	85.1
Kimbundu	kmb_Latn					✓	49.7	49.9	58.5	60.3	**65.1**
Kikongo	kon_Latn					✓	65	61.8	70.3	74.2	**78.7**
Lingala	lin_Latn				✓	✓	65.8	63.2	73.8	83.3	**84.1**
Luba-Kasai	lua_Latn				✓	✓	56.3	52	65.2	70.9	**72.7**
Ganda	lug_Latn				✓	✓	45	46.8	61.2	67.7	**69.3**
Luo	luo_Latn				✓	✓	60	59.5	61.2	67.4	**76.8**
Mossi	mos_Latn				✓	✓	59.5	52.1	61.9	63.8	**68.7**
Northern Sotho	nso_Latn				✓	✓	54.8	51.8	80.7	82.6	**82.7**
Nuer	nus_Latn					✓	43.9	54.7	47.5	46.2	**61.7**
Nyanja	nya_Latn			✓	✓	✓	60.7	58.1	83.3	**86.3**	83
Plateau Malagasy	plt_Latn	✓		✓	✓	✓	85.3	50.5	**88.4**	88.2	87.5
Rundi	run_Latn		✓	✓	✓	✓	46	77.9	82.4	83.5	**84.6**
Sango	sag_Latn					✓	61	61.4	62.1	65.4	**72.3**
Shona	sna_Latn			✓	✓	✓	51.5	55.2	81.3	80.3	**81.8**
Somali	som_Latn	✓	✓	✓	✓	✓	78.7	77.7	81.7	80	**82.4**
Southern Sotho	sot_Latn			✓	✓	✓	55.9	57.4	83.7	**84.0**	83.1
Swati	ssw_Latn				✓	✓	59	53.5	80.6	**81.8**	81.4
Swahili	swh_Latn	✓	✓	✓	✓	✓	85.8	85.8	**87.9**	87.2	86.5
Tamasheq (Latin)	taq_Latn					✓	55.1	53.4	58.1	57.7	**59.4**
Tamasheq (Tifinagh)	taq_Tfng					✓	26.9	26.4	**27.9**	26	26.4
Tigrinya	tir_Ethi		✓		✓	✓	67.7	70.3	**81.5**	81	79.2
Tswana	tsn_Latn				✓	✓	58.5	58	79.4	82.2	**83.6**
Tsonga	tso_Latn				✓	✓	57.2	58.8	68.5	**80.9**	80
Tumbuka	tum_Latn				✓	✓	66	67.3	82.7	87.2	**88.4**
Twi	twi_Latn				✓	✓	62.2	64.1	65.8	77.5	**78.9**
Tamazight	tzm_Tfng					✓	23.9	**26.5**	25.9	25	26
Umbundu	umb_Latn				✓	✓	53.6	51.8	59.9	62.3	**67.4**
Wolof	wol_Latn				✓	✓	60.1	50.1	64.3	66.6	**71.2**
Xhosa	xho_Latn	✓		✓	✓	✓	70.7	47.5	83.1	83.5	**84.5**
Yoruba	yor_Latn			✓	✓	✓	49.6	70.6	74.8	**80.5**	78.9
Zulu	zul_Latn			✓	✓	✓	73.5	53.6	84.9	84.2	**85.9**
Average							**59.9**	**56.1**	**69.2**	**71.6**	**74.1**

Table 8: **Evaluation result on different African languages pre-trained language models**