An Empirical Analysis of Machine Translation for Expanding Multilingual Benchmarks

Sara Rajaee[*] [◇] **Rochelle Choenni**[*] [♠] **Ekaterina Shutova**[♠] **Christof Monz**[◇]
[♠]ILLC, University of Amsterdam, the Netherlands
[◇]Language Technology Lab, University of Amsterdam, the Netherlands
{s.rajaee, r.m.v.k.choenni, e.shutov, c.monz}@uva.nl

Abstract

The rapid advancement of large language models (LLMs) has introduced new challenges in their evaluation, particularly for multilingual settings. The limited evaluation data are more pronounced in low-resource languages due to the scarcity of professional annotators, hindering fair progress across languages. In this work, we systematically investigate the viability of using machine translation (MT) as a proxy for evaluation in scenarios where human-annotated test sets are unavailable. Leveraging a state-of-the-art translation model, we translate datasets from four tasks into 198 languages and employ these translations to assess the quality and robustness of MT-based multilingual evaluation under different setups. We analyze task-specific error patterns, identifying when MT-based evaluation is reliable and when it produces misleading results. Our translated benchmark reveals that current language selections in multilingual datasets tend to overestimate LLM performance on low-resource languages. We conclude that although machine translation is not yet a fully reliable method for evaluating multilingual models, overlooking its potential means missing a valuable opportunity to track progress in non-English languages.

1 Introduction

Large-scale evaluation of multilingual language models (MLMs) across hundreds of languages has remained a persistent challenge as the existing benchmarks cover a subset, and mostly high-resource, of languages (Singh et al., 2024, 2025). Although human-translated (HT) evaluation sets offer accurate and reliable evaluation of MLMs, creating them for hundreds of languages is costly, time-consuming, and sometimes impossible. Consequently, tracking the progress of MLMs in most languages has lagged behind. Moreover, multilingual benchmarks often cover different subsets of languages, and such inconsistent evaluation setups create a fragmented picture of MLM capabilities (Liang et al., 2020; Hu et al., 2020; Ruder et al., 2021).

Recent advances in machine translation (MT) have emerged as a promising solution to these challenges of multilingual evaluation. Prior studies have shown that state-of-the-art MT systems achieve human-level translation quality in certain high-resource languages (Kocmi et al., 2023, 2024). While MT has been used to construct evaluation sets (Chen et al., 2024), to our knowledge, there has been no systematic study of the viability of MT-based evaluations across different evaluation setups and diverse downstream tasks in nearly 200 languages. Moreover, although previous work has shown the limitations of MT models, such as translation artifacts and stylistic shifts (Park et al., 2024; Wang et al., 2023), there has been no comprehensive analysis of the potential risks of using MT-based datasets for MLM evaluation.

In this work, and by considering recent MT advances, we investigate the viability of using MT for large-scale multilingual evaluation where human-annotated data is unavailable. Using four popular multilingual tasks, we translate their test sets into 198 languages using NLLB ("No Language Left Behind"), which officially supports 200 languages (NLLB Team et al., 2022). We then compare three MLMs' performance, i.e., XLM-R (both base and large versions) (Conneau et al., 2020), BLOOMz (Muennighoff et al., 2022), and AYA-101 (Üstün et al., 2024b), on these MT-based test sets against their HT equivalents, using multiple evaluation setups (zero-shot fine-tuning and zero-shot prompting) and both accuracy and rank correlation as metrics. Beyond overall performance comparisons, we analyze how translation quality relates to evaluation results, detecting MT-specific error types using an LLM-as-a-judge framework in the selected tasks. Finally, we assess whether

[*]Equal contribution.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 564–593
November 8–9, 2025.

current multilingual benchmarks misrepresent the performance of MLMs when limited to small language subsets. More specifically, we try to answer the following research questions:

- To what extent does the use of machine translation affect MLM performance estimates?
- To what extent is MLM performance on MT-based evaluations influenced by translation quality?
- What major translation error types occur in cases where MLMs succeed on human translations but fail on machine translations?
- To what extent do existing multilingual benchmarks misrepresent MLM performance?

Across tasks and models, MT-based evaluations yield results that are highly correlated with HT-based evaluations (average Spearman's 0.95), with only small average accuracy differences (<1.5 points), Table 3 and 4. Furthermore, translation quality shows a moderate positive correlation with performance differences, suggesting that MT reliability depends partly on the underlying MT system's accuracy and the sensitivity of the downstream task to the translation quality, Table 5. Our error analysis reveals that lower performance on MT data is often linked to lexical mistranslations and subtle semantic shifts, though major meaning-altering errors are relatively rare. We show that LLM-as-a-judge might serve as a proxy for filtering low-quality translation examples, thereby improving the reliability of MT-based evaluations. Finally, we find that restricting evaluation to the language subsets in the widely-used benchmarks underestimates the performance of MLMs on high-resource languages and overestimates their performance on low-resource languages by up to 4 accuracy points compared to evaluation across 198 languages.

2 Related work

Multilingual large language models (LLMs) are rapidly expanding their language coverage, reaching to hundreds of languages (Üstün et al., 2024a; Yang et al., 2025). Yet, benchmarks have struggled to keep pace, often covering only a fraction of these languages (e.g., Global MMLU covers 42 languages (Singh et al., 2025), whereas models like Gemma 3 support over 140 (Gemma Team et al., 2025)). This imbalance makes it difficult to measure progress fairly and consistently across languages. On the other hand, creating human-annotated benchmarks for every language is impractical due to high costs and limited expertise. In this regard, machine translation is an attractive potential alternative to expand existing multilingual datasets at scale. Although previous studies have shown MT systems may introduce challenges such as hallucinations and translationese artifacts (Artetxe et al., 2020; Wang and Sennrich, 2020; Zhang and Toral, 2019), recent improvements, especially for mid and low-resource languages, have considerably enhanced translation quality (Ranathunga et al., 2023). Thus, leveraging advanced MT systems is an efficient way to expand the language coverage of benchmarks more quickly. This approach opens the door to more comprehensive and equitable evaluations, better reflecting the multilingual capabilities of LLMs. While Thellmann et al. (2024) also investigate whether machine translated benchmarks can reliably assess model performance across languages, they limit the scope of their study to 20 European languages. Since European languages are already overrepresented in existing NLP resources and benchmarks (Asai et al., 2022), while many other languages remain severely underrepresented (Joshi et al., 2020), we instead scale our evaluation to 198 languages to provide a more balanced assessment.

3 Methods

In this paper, we evaluate multilingual LLMs on four tasks across 198 languages. To this end, we create MT test data using the NLLB model and assess performance under two evaluation paradigms: zero-shot testing and zero-shot prompting. In this section we provide a comprehensive overview of our methods used for large-scale multilingual evaluation. In Section 4, we will explain which methods were used to generate the MT test data itself.

3.1 Tasks and datasets

We massively scale the evaluation of LLMs on four datasets.

XNLI The Cross-Lingual Natural Language Inference (XNLI) dataset (Conneau et al., 2018) contains premise-hypothesis pairs labeled with: 'entailment', 'neutral', or 'contradiction' in 15 languages.

The original pairs come from English and the

test sets were human translated into the other languages.

PAWS-X The Cross-Lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) dataset (Yang et al., 2019) requires the model to determine whether two sentences are paraphrases of one another. The parallel test data has been provided in 7 languages. To create this dataset, a subset of the PAWS development and test sets (Zhang et al., 2019) was professionally translated from English to 6 other languages.

XCOPA The Cross-lingual Choice of Plausible Alternatives (Ponti et al., 2020) evaluates common-sense reasoning in 11 languages. The samples contain a premise and question paired with two answer choices from which the model can select. The dataset is manually translated from English into 11 other languages.

XStorycloze The Cross-lingual Storycloze (Lin et al., 2021) proposes a common-sense reasoning task in 11 languages, in which the model predicts which one of two story endings is the most likely to follow after a given short story. The Storycloze dataset was professionally translated from English into 10 other languages.

3.2 Multilingual language models

For the large-scale evaluation, we focus on the base and large version of XLM-R (Conneau et al., 2020) pre-trained on 100 languages as it is one of the most popular MLMs. In addition, we report scores from BLOOMz 7.1b (Scao et al., 2022) and AYA-101 13b (AYA)(Üstün et al., 2024b). BLOOMz is trained on 46 languages and AYA on 101 languages. Moreover, both models are further instruction-tuned on a mixture of prompts in different languages. Given that the PAWS-X dataset was included during instruction-tuning, we evaluate BLOOMz and AYA on the held out datasets only, i.e., XNLI, XCOPA and XStorycloze. Note that BLOOMz and AYA were selected over other MLMs, such as Llama (Touvron et al., 2023), as they are the largest publicly available and explicitly MLMs (i.e., statistics on the pretraining data distribution across languages is publicly available).

3.3 Selection of test languages

For our selection of test languages, there are two constraining factors: (1) the language has to be covered by the NLLB-200 translation model and FLORES-200 dataset, and (2) the script of the language needs to have been seen during the pretraining of the model. We then separately filter out the test languages by unseen scripts for each model. This leaves us with 196, 168, and 198 test languages for XLM-R, BLOOMz, and AYA, respectively, see Appendix J for the complete lists. Note that the number of compatible languages is lower for BLOOMz as it has seen fewer writing scripts during pretraining.

	Unseen	Low	Mid	High	Total
% of data	0	>0 and <0.1	$\geq$ 0.1 and <1	$\geq$1	
XLM-R	106	30	34	26	196
BLOOMz	131	21	7	9	168
AYA	103	57	25	13	198

Table 1: Number of languages categorized as high, mid, low, and unseen languages when looking at the percentage of seen pretraining data of the respective LMs.

Resource categorization by pretraining distribution Moreover, we categorize the test languages for each model separately based on the percentage of total data that they contributed during pretraining. In Table 1, we report the percentage thresholds used for our categorization and the resulting number of test languages for each category and model. As the pretraining data coverage is reported in numbers of GB for XLM-R, we convert these scores to percentages of the full pretraining data. For BLOOMz, we use the reported language distribution numbers [1], and for AYA, we consider mT5 pretraining data distribution as it is utilized as the base model for AYA.

3.4 Evaluation settings

Zero-shot testing We fine-tune XLM-R on the entire training set for each task in English. We then use our fine-tuned model for zero-shot testing in the test languages. We fine-tune our model using the HuggingFace Library, see Appendix A for details.

Zero-shot prompting BLOOMz and AYA are instruction-tuned on multiple classification tasks, thus we test these models out-of-the-box in a zero-shot prompting set up. This has the benefit that dataset artifacts, which are commonly known to be leveraged during fine-tuning, cannot be learned (McCoy et al., 2020). As BLOOMz and AYA fail to predict a third option for XNLI (neutral), we report results on a binarized version of the

[1] https://huggingface.co/bigscience/bloom

metric	High	Mid	Low	Unseen	Ave.	Med.
NLLB-Distil						
chrF++	49.34	46.92	43.22	37.12	44.15	45.07
NLLB-3.3B						
chrF++	52.5	50.8	46.1	40.3	44.5	45.5

Table 2: The quality of the MT system across high, mid, and low-resource languages using chrF++ based on the XLM-R model's categorization.

task by aggregating sentences with the 'neutral' and 'contradiction' labels into one class and making the model predict entailment or not. Moreover, the instructions are given in English. see Appendix A for the prompts per task.
Finally, note that our goal is not to compare performance of BLOOMz and AYA to XLM-R but rather to test the reliability of machine-translated test sets in two popular evaluation paradigms.

Evaluation metric As the automatic evaluation metric for testing our MT quality, we use the chrF++ (Popović, 2017). This metric calculates the character and word n-gram overlap between the machine and human reference translations. It is a tokenization-independent metric aligning better with human judgments for morphologically-rich languages compared to BLEU (Tan et al., 2015; Kocmi et al., 2021; Briakou et al., 2023). We also employ the LLMs-as-a-judge technique to evaluate translation quality and analyze the types of errors that occur in translation. For the evaluation, we use the FLORES-200 dataset (NLLB Team et al., 2022), which includes human-translated data for 200 languages, and select 100 sentences from each language.

4 Machine translating test data

For machine translation, we employ the NLLB model covering 202 languages (NLLB Team et al., 2022). Through extensive analysis, Zhu et al. (2024) have shown that NLLB performance surpasses other powerful LLMs such as ChatGPT (OpenAI., 2022) and GPT-4 (Achiam et al., 2023) when translating out of English ($En \Rightarrow$ `tgt` setting). NLLB also demonstrates a minimal difference to the closed source system, Google translate[2] on the Flores-101 dataset. To gain a better understanding of the role of the MT system on the translated data quality and, consequently, its performance on downstream tasks, we experiment with two NLLB versions. We choose the distill NLLB with 600M parameters and the 3.3B NLLB model using greedy sampling. For each example, we translate each sentence (e.g., SENTENCE1 and SENTENCE2 in PAWS-X) separately. In Appendix B, we provide some of the lessons learned from our MT experiments.

In Table 2, we report the average chrF++ scores for translations obtained with the NLLB-Distil and NLLB-3.3B models on the dev set of FLORES-200 dataset including human translation data in 200 languages (NLLB Team et al., 2022). We categorize languages by XLM-R resource ranking (high, mid, low, and unseen) and observe that NLLB-3.3B consistently performs better than the 600M version across all categories. Thus, while the smaller model has been added for analysis purposes, we use the NLLB-3.3B for translation throughout the paper unless stated otherwise. Moreover, we confirm that our scores for all languages are among the best performance of SOTA multilingual MT systems (Bapna et al., 2022; NLLB Team et al., 2022).

5 Evaluating the reliability of MT data

In this section, we study the opportunities and challenges of using MT to extend multilingual benchmarks and assess model performance across a broader range of languages from two perspectives, i.e. the performance gap when measuring on human versus machine translated data and the quality of the MT data itself. Thus, in Section 5.1, we first study the degree to which machine translation alters the reliability and validity of performance assessments for LLMs. Then, in Section 5.2, we evaluate the quality of the machine translated data itself through automatic metrics and LLM-based judgments, both to analyze error patterns and to serve as a proxy for filtering low-quality translated data for evaluation.

5.1 Evaluation gaps between MT and HT

5.1.1 Average performance changes

The key motivation for using machine translation in expanding multilingual benchmarks is to reduce costly and time-consuming human annotation. However, this raises an important question: *Does relying on machine-translated data compromise the validity of LLMs evaluations?* If model performance on MT data is substantially different from that on human translated data, evaluation outcomes

[2] https://translate.google.com/

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XLM-R																									
XCOPA	-	-	-	-	-	69/73	-	-	-	50/57	76/69	75/73	-	-	-	49/59	-	66/58	64/68	-	69/66	71/66	-	68/70	72/73
XStoryCloze	80/80	-	-	-	84/84	-	79/79	-	78/79	-	88/87	-	-	-	71/70	-	84/85	75/76	-	76/75	-	-	-	-	87/86
XNLI	78/78	82/82	82/82	81/80	83/78	-	-	82/83	75/80	-	-	-	-	-	-	-	79/82	70/74	-	-	76/76	77/79	71/78	78/81	79/74
PAWS-X	-	-	90/91	-	91/91	-	-	91/90	-	-	-	-	81/84	81/83	-	-	-	-	-	-	-	-	-	83/82	
BLOOMz																									
XCOPA	-	-	-	-	-	52/52	-	-	-	N/A	78/78	62/65	-	-	-	51/52	-	60/63	75/72	-	N/A	50/50	-	80/77	71/67
XStoryCloze	88/88	-	-	-	91/91	-	84/78	-	85/84	-	91/90	-	-	-	54/52	-	73/73	79/79	-	74/73	-	-	-	-	70/70
B-NLI	71/72	66/68	69/68	65/66	73/74	-	-	72/73	70/72	-	-	-	-	-	-	-	69/70	70/71	-	-	N/A	68/70	68/70	72/73	74/72
AYA																									
XCOPA	-	-	-	-	-	87/84	-	-	-	82/83	87/87	88/88	-	-	-	56/56	-	79/83	86/83	-	84/82	86/85	-	85/84	86/84
XStoryCloze	95/92	-	-	-	94/94	-	83/75	-	93/91	-	91/87	-	-	-	94/86	-	90/82	93/89	-	93/88	-	-	-	-	95/90
B-NLI	78/79	79/79	78/78	78/78	79/80	-	-	79/80	75/75	-	-	-	-	-	-	-	79/79	74/75	-	-	79/79	79/79	74/75	76/77	77/77

Table 3: The (%) accuracy of the models on the human translated (original)/our machine translated datasets. Darker colors indicate bigger gaps between the models' performance on human and machine-translated data.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
Pearson corr.	95.3	69.3	93.3	98.0	85.0	97.8	98.7	98.0	95.3	90.7
Spearman rank corr.	77.5	82.4	82.6	98.8	89.0	95.1	97.3	81.8	92.5	77.0

Table 4: Pearson and Spearman rank correlation between the performance on MLMs on human-translated (HT)(original data) and machine-translated (MT) data (ours). The high correlations indicate that the performance of MLMs on both machine and human-translated data is similar.

may become unreliable or biased.

To answer this question, we compare model performance on machine and human translated data. We evaluate all models using the same setups and report accuracy scores on machine and human translated data in Table 3.[3] Our results show that the difference in performance on machine and human-translated data is small (on average, about 2.6%). Moreover, this trend is consistent across all models. In some instances, the models achieve slightly higher performance on machine-translated data. We hypothesize that this may be due to the translationese effect. Conversely, in other cases, we observe a drop in performance, which might be caused by low-quality translations. We investigate this issue in Section 5.2, where we explore methods such as LLM-based filtering to detect and reduce the impact of poor quality MT outputs.

5.1.2 Ranking consistency

Building on the finding that accuracy scores between machine-translated and human-translated data are closely aligned, we further assess the reliability of MT-based evaluation by examining the consistency of model rankings, i.e. whether the relative performance of a model across different languages remains stable. Specifically, we want to see if the accuracy differences observed earlier affect which languages a model performs better or worse on. To measure this, we compute the average Pearson and Spearman rank correlations between model performances on human and machine-translated datasets (see Table 4).

The high correlation across all tasks shows that the rank order of the models' performance across different languages using human and machine-translated data is almost the same. This demonstrates that machine translated data could reliably be used to assess the relative differences in model performance across languages.

5.2 The impact of MT quality

Moreover, we analyze whether the quality of machine translation, as measured by chrF++, correlates with LLM performance. In other words, what matters in this experiment is the consistency of the correlations between human- vs. machine-translated data. Table 5 summarizes the numerical results. The results show no significant correlation in either case. That is, higher MT quality (as indicated by better chrF++ scores) does not systematically correspond to higher model performance. The pattern suggests that small variations in MT quality, as captured by the automatic metric, do not strongly influence LLM accuracy. This further supports the robustness of using MT for multilingual evaluation, although it also highlights that chrF++ may not fully capture the aspects of translation quality that affect downstream performance.

[3]Results of XLM-R base are reported in the appendix.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
Pearson Corr.										
chrF++ vs. HT	27.4	10.5	89.3	3.3	-10.4	-10.4	64.5	41.0	8.2	-18.5
chrF++ vs. MT	30.3	66.2	96.5	8.8	14.5	2.9	66.7	52.1	16.3	11.3
Spearman rank Corr.										
chrF++ vs. HT	29.6	24.4	65.7	16.4	5.7	5.7	68.1	22.7	8.1	-34.6
chrF++ vs. MT	22.6	48.1	82.7	22.4	18.3	14.7	75.8	55.5	26.1	22.4

Table 5: Average Pearson and Spearman rank correlation between chrf++ scores and human translated (HT) (original data) and machine-translated data (MT) (ours).

Source (en)	**Translation (es)**	**Error type (span)**	**Severity**
So I'm not really sure why. I am certain as to the reason why.	Así que no estoy muy seguro de por qué. Estoy seguro de la razón.	Accuracy/omission ("as to") Fluency/redundancy ("de por qué")	Major Minor
I'm covering the same stuff. I'm talking about the same things they did.	Estoy cubriendo las mismas cosas. Estoy hablando de las mismas cosas que ellos hicieron.	Accuracy/mistranslation ("que ellos hicieron") Fluency/grammar ("Estoy cubriendo")	Major Minor

Table 6: Examples of the MQM framework output for detected errors in translated XNLI test examples.

5.2.1 LLM-as-a-Judge Evaluation

In the earlier experiments, we observed that for certain tasks and languages, model performance on MT data was lower than on HT data. We hypothesize that these drops are often linked to lower translation quality. Automatic metrics such as chrF++, while useful, may be overly sensitive to minor deviations that are not critical for the downstream task, resulting in misleadingly low scores.

To better assess translation quality, we employ an LLM-as-a-judge approach. Specifically, we use it to (1) identify and categorize translation errors in examples where the model makes correct predictions on human translations but fails on machine-translated versions, and (2) evaluate its potential as a filtering mechanism to remove low-quality translations from translated evaluation sets.

To this end, we adopt the Multidimensional Quality Metrics (MQM) framework from Freitag et al. (2021). We prompt the LLM evaluator to follow the evaluation guidelines for the translation task to detect and classify translation errors into major and minor categories. The LLM-based evaluation provides the error span, error type, and error classification. Major errors (true errors) are generally easier to detect, whereas minor errors often stem from minor imperfections in translations, see Table 6 for an example.[4] As our LLM-as-a-judge, we use the Gemma 3 27B model (Gemma Team et al., 2025), covering over 140 languages, in an in-context learning setup to evaluate translation quality.[5]

In Table 7, we report the average number of major errors for all predictions (Ave. #Err.) and for the subset where predictions switch from correct on HT to wrong on MT (C→W). While the average error rates for these two categories are close, C→W shows slightly higher values across most languages, suggesting that these degraded translations can be systematically flagged. Moreover, consistently across all tasks and languages, the most frequent translation error type is *accuracy*, including a mistranslation error or omission from the source sentence. Using this observation, we filter out examples with more than 2 major errors, and evaluate the performance of the AYA model on the higher-quality examples. In Table 8, we present the percentage of gap reduction between the models' performance on HT and MT. Based on the results, in most cases, after removing low-quality MT data, the performance gap between MT and HT data is reduced up to 100%. Our results indicate that integrating LLM-based quality checks into MT-based evaluation pipelines can help reduce evaluation noise and improve the reliability of us-

[4] See Table 2 in Freitag et al. (2021) for an overview of the error categorizations.

[5] Please refer to the appendix for the experimental setups.

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XCOPA																									
Ave. #Err.	-	-	-	-	-	0.98	-	-	-	1.50	0.82	0.67	-	-	-	2.30	-	1.19	1.29	-	1.21	0.81	-	0.87	1.13
C→W	-	-	-	-	-	1.10	-	-	-	1.57	1.04	0.57	-	-	-	2.52	-	1.39	1.11	-	1.41	0.90	-	0.91	1.37
XStoryCloze																									
Ave. #Err.	1.84	-	-	-	1.03	-	2.19	-	1.38	-	1.18	-	-	-	1.69	-	1.22	1.57	-	2.04	-	-	-	-	1.77
C→W	1.92	-	-	-	1.15	-	2.17	-	1.42	-	1.17	-	-	-	1.73	-	1.23	1.83	-	2.04	-	-	-	-	1.95
B-NLI																									
Ave. #Err.	1.43	1.12	0.98	1.01	0.96	-	-	0.99	1.42	-	-	-	-	-	-	-	1.15	1.66	-	-	1.46	1.31	1.62	1.26	1.53
C→W	1.53	1.23	1.08	1.06	0.98	-	-	1.07	1.52	-	-	-	-	-	-	-	1.27	1.72	-	-	1.60	1.39	1.64	1.35	1.67

Table 7: Average number of major errors across languages on AYA model. The first row reports the number of machine translation errors of all predictions on MT data; the second row reports the number of errors when predictions switch from correct on HT to wrong on MT.

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XCOPA	-	-	-	-	-	0.0	-	-	-	100	0.0	0.0	-	-	-	0.0	-	25	33.3	-	0.0	0.0	-	0.0	-50
XStoryCloze	100	-	-	-	0.0	-	-112.5	-	50.0	-	-75.0	-	-	-	12.5	-	-62.5	125	-	60.0	-	-	-	-	40.0
B-NLI	100	0.0	0.0	0.0	100	-	-	0.0	0.0	-	-	-	-	-	-	-	0.0	0.0	-	-	0.0	0.0	0.0	100	0.0

Table 8: The percentage of gap reduction between HT and MT after filtering out low quality MT examples.

ing MT.

6 Large-scale evaluation results

Having confirmed in Section 5 that our translated test sets are of a reliable quality, we now move on to analyze how the MLMs perform on them. In Figure 1, we summarize the performance of XLM-R, BLOOMz, and AYA in 196, 168, and 198 languages, categorized by their data scarcity during the pre-training of each model as explained in Section 3.3. We find that the average performance for all models is similar for high- and mid-resource languages. Yet, while still above the random baseline, there is a notable drop in performance for low-resource and unseen languages. Moreover, we find that standard deviations across low-resource languages are larger than high- and mid-resource ones. This shows that performance across low-resource languages varies a lot, making the average score less reliable. Still, performance in unseen languages is relatively high; for XNLI and PAWS-X, on average, we obtain +18% and +29% above random performance, suggesting cross-lingual knowledge transfer to the unseen languages.

6.1 Representativeness of benchmark language sets

As each dataset contains a distinct selection of languages for testing, we study to what extent each of them provides a reliable estimate for how MLMs performance will generalize to more languages.

	High	Mid	Low	Ave.
XLM-R				
PAWS-X	89.5 / 89.6	- / 87.3	- / 85.6	89.5 / 85.3
XNLI	82.2 / 81.5	79.5 / 78.2	74.4 / 71.9	80.5 / 72.4
XCOPA	71.6 / 71.8	69.6 / 68.2	66.4 / 61.7	70.3 / 69.2
XStoryCloze	85.5 / 83.1	77.2 / 78.6	74.1 / 69.9	78.9 / 77.2
BLOOMz				
B-NLI	72.8 / 73.0	70.8 / 70.2	70.9 / 68.3	72.2 / 69.8
XCOPA	76.9 / 79.0	71.6 / 67.5	63.0 / 54.3	73.7 / 62.3
XStoryCloze	86.2 / 87.9	81.1 / 72.0	79.4 / 64.5	82.8 / 71.6
AYA				
B-NLI	78.4/77.8	77.7/77.6	73.5/75.5	77.4/76.4
XCOPA	83.65/86.3	84.8/84.6	82.8/79.8	83.7/82.0
XStoryCloze	85.8/89.7	91.0/89.6	85.5/86.4	87.0/87.7

Table 9: The average performance of high, mid, and low-resource languages covered by the original dataset/the languages covered by our machine-translated datasets. All results are computed on the machine-translated data.

While we do not cover all the world's languages, we compare the averages between the languages covered by the original datasets and those covered by our much larger translated datasets that contain 198 languages.

To this end, we split the languages from the original datasets based on our resource categorization reported in Table 1, and report the average performance for each category in Table 9. Importantly, all performance scores are computed on the translated data. From the results, we observe that for high and, to some extent, mid-resource languages, average performance on both language selections

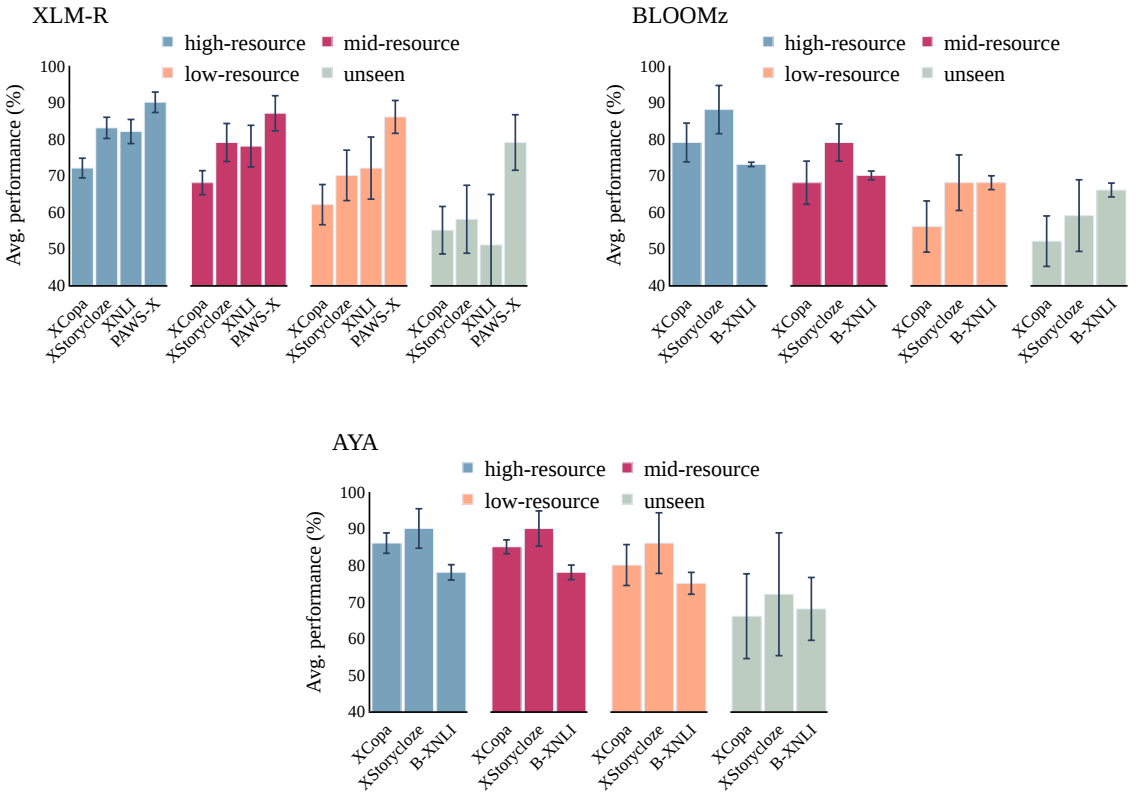

Figure 1: Average performance across test languages in a zero-shot fine-tuning setup for XLM-R and in a zero-shot prompting using BLOOMz. Results are categorized per task and data coverage during pretraining as reported in Table 1. Results across models are not directly comparable as their language categorizations differ.

is similar, making the language coverage from the original datasets sufficiently representative. Yet, for low-resource languages, we find a notable difference, which suggests that the datasets' language coverage is not representative of a wider range of low-resource languages. Specifically, across all tasks, we tend to overestimate performance, which can go up to 4.7% and 8.7% accuracy points (for XCOPA).

7 Conclusions

In this paper, we investigate the use of machine translation to create large-scale multilingual evaluation sets in scenarios where human-translated data is unavailable. Our experiments show that using SOTA MT models for evaluation yields results closely aligned with HT ones, with small average accuracy differences and high rank correlations, indicating stable relative performance across languages. Translation quality, measured by chrF++, does not strongly predict downstream performance differences, suggesting that minor variations in automatic scores are not critical for many tasks. However, LLM-as-a-judge analysis reveals that examples where models succeed on HT but fail on MT often contain more major errors, as such this method could be used to filter low-quality translations to reduce evaluation noise. Scaling evaluations to nearly 200 languages further reveals representativeness gaps in existing benchmarks. For high- and mid-resource languages, original language selections approximate broader trends well, but for low-resource languages, they overestimate performance—by up to 8.7 accuracy points in some cases. This demonstrates that MT-based evaluation can both expand coverage and uncover biases in benchmark design. Overall, our findings indicate that MT, when coupled with targeted quality control, enables broader, more representative, and more equitable multilingual evaluation, while highlighting the need to reassess how language selections in current benchmarks reflect true model capabilities.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Akari Asai, Trina Chatterjee, Junjie Hu, Eunsol Choi, et al. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *2022 Findings of the Association for Computational Linguistics: EMNLP 2022*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

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 generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings

of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2022. Chatgpt.

ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5193–5221, Bangkok, Thailand. Association for Computational Linguistics.

Edoardo M. 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)*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel

Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *CoRR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024a. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024b. Aya model: An instruction finetuned open-access multilingual language model. *arXiv e-prints*, pages arXiv–2402.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. 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 1298–1308.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Experimental setups

For the implementation of all models, we rely on the HuggingFace Library (Wolf et al., 2019). XLM-R large, BLOOMz, and AYA-101 (AYA) have 330M, 7.1B, and 13B parameters, respectively. Moreover, we have run all the BLOOMz and AYA experiments on an NVIDIA A100-SXM4 GPU with 40GB memory, and a single NVIDIA A6000 has been used for the MT and XLM-R experiments with 48GB memory.

XLM-R fine-tuning details For the NLI task, we have fine-tuned XLM-R with a learning rate of $2e$-5, AdamW optimizer, and a batch size of 32 for 3 epochs. For the PAWS-X task, we have considered a learning rate of $2e$-6, batch size 16 with a warm-up ratio of 0.01 for 3 epochs. For XCOPA and XStoryCloze tasks, first, we train the model on the training set of Social IQa (Sap et al., 2019) and then fine-tune it on the training set of XCOPA dataset (Gordon et al., 2012). We have selected a learning rate of $3e$-6, batch size of 16 for SIQa and 8 for XCOPA, a warm-up ratio of 0.1, and fine-tune the model for 3 epochs on each dataset.

BLOOMz and AYA zero-shot prompts For zero-shot prompting, we constructed the following prompts for XNLI, XCOPA and XStorycloze respectively:

Premise: `<premise>`
Hypothesis: `<hypothesis>`
Does the premise entail the hypothesis?
Pick between yes or no.

Premise: `<premise>`
Option A: `<choice1>`
Option B: `<choice2>`
Based on the premise, which `<cause/effect>` is more likely?
Pick between options A and B.
Answer:

Consider the following story:
`<story>`
Which ending to the story is most likely?
Pick between options A and B:
A: `<story_ending1>`
B: `<story_ending2>`
Answer:

B Lessons learned for Machine Translation

We now also share a few practical lessons learned from our MT experiments using NLLB to facilitate the translation of new datasets in future work:

- NLLB tends to skip sentences when translating paragraphs. Thus, it is important to translate the sentences one by one.
- NLLB has difficulty translating short phrases/names such as names, dates, locations, etc., because it tends to hallucinate additional content. This makes it challenging to translate the answers from QA datasets such as XQuAD.
- NLLB inconsistently chooses to code-switch to the target language. For instance, when translating the sentence 'Sara is asleep', it can choose to translate it either to 'Sara ? farsi' or 'fully farsi'. This can be particularly challenging for retrieval datasets where the answer does not tend to fully match the context.
- While translation quality tends to be similar for different NLLB model sizes, at least the 3.3B version should be used when translating to languages that were low-resource, considering NLLB's pretraining data.

C Correlation between Chrf++ scores and translations

See Table 10 and Table 11 for Spearman and Pearson correlation results between Chrf++ scores and performance on human and machine translated data.

D LLM-as-a-judge

Following previous work, we use 3-shot prompting for this experiment provided in Table 12.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
chrF++ vs. Human translated	29.6	24.4	65.7	16.4	5.7	5.7	68.1	22.7	8.1	-34.6
chrF++ vs. Machine translated	22.6	48.1	82.7	22.4	18.3	14.7	75.8	55.5	26.1	22.4

Table 10: Average Spearman rank correlation between chrf++ scores and human- (original data) and machine-translated data (ours).

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
chrF++ vs. Human translated	27.4	10.5	89.3	3.3	-10.4	-10.4	64.5	41.0	8.2	-18.5
chrF++ vs. Machine translated	30.3	66.2	96.5	8.8	14.5	2.9	66.7	52.1	16.3	11.3

Table 11: Average Pearson correlation between chrf++ scores and human- (original data) and machine-translated data (ours).

In-context-learning prompt:

Based on the given source, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.
Source: I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holder's permission.
Translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir `involvement`.
Errors:
Major: accuracy/mistranslation – `involvement`; accuracy/omission – `the account holder`
Minor: fluency/grammar – `wäre`; fluency/register – `dir`

Source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.
Translation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemže obě partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.
Errors:
Major: accuracy/addition – `ve Vídni`; accuracy/omission – `the stop-start`
Minor: terminology/inappropriate for context – `partaje`

Source: 大众点评乌鲁木齐家居商场频道为您提供高铁居然之家地址，电话，营业时间等最新商户信息，找装修公司，就上大众点评
Translation: Urumqi Home Furnishing Store Channel provides you with the latest business information such as the address, telephone number, business hours, etc., of high-speed rail, and find a decoration company, and go to the reviews.
Errors:
Major: accuracy/addition – `of high-speed rail`; accuracy/mistranslation – `go to the reviews`
Minor: style/awkward – `etc.,`

Source: *{source}*
Translation: *{translation}*
Errors:

Table 12: The prompt used for the llm-as-a-judge experiment.

E Full results XLM-R Large

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	71.05	fao_Latn	87.4	lij_Latn	86.2	slv_Latn	89.75
ace_Latn	85.05	fij_Latn	75.45	lim_Latn	87.55	smo_Latn	78.6
acm_Arab	88.5	fin_Latn	86.2	lin_Latn	77.25	sna_Latn	78.6
acq_Arab	87.85	fon_Latn	69.6	lit_Latn	89.2	snd_Arab	86.95
aeb_Arab	87.2	fra_Latn	91.7	lmo_Latn	86.9	som_Latn	86.15
afr_Latn	92.1	fur_Latn	86.9	ltg_Latn	81.4	sot_Latn	80.45
ajp_Arab	88.45	fuv_Latn	74.3	ltz_Latn	85.6	spa_Latn	92.2
aka_Latn	75.45	gaz_Latn	77.75	lua_Latn	75.65	srd_Latn	87.25
als_Latn	92.25	gla_Latn	88.15	lug_Latn	76.95	srp_Cyrl	90.65
amh_Ethi	82.85	gle_Latn	88.1	luo_Latn	73.95	ssw_Latn	80.15
apc_Arab	87.1	glg_Latn	91.75	lus_Latn	77.3	sun_Latn	89.6
arb_Arab	88.7	grn_Latn	79.75	lvs_Latn	87.5	swe_Latn	91.1
ars_Arab	89.15	guj_Gujr	87.75	mag_Deva	88.35	swh_Latn	89.65
ary_Arab	87.25	hat_Latn	86.5	mai_Deva	84.9	szl_Latn	87.45
arz_Arab	89.25	hau_Latn	86.15	mal_Mlym	82.9	tam_Taml	84.35
asm_Beng	83.2	heb_Hebr	89.05	mar_Deva	83.25	taq_Latn	73.25
ast_Latn	86.95	hin_Deva	88.4	min_Latn	88.4	tat_Cyrl	79.5
awa_Deva	84.95	hne_Deva	86.45	mkd_Cyrl	90.75	tel_Telu	85.65
ayr_Latn	71.45	hrv_Latn	90.6	mlt_Latn	82.65	tgk_Cyrl	77.65
azb_Arab	72.1	hun_Latn	88.65	mni_Beng	74.2	tgl_Latn	92.6
azj_Latn	86.0	hye_Armn	87.55	mos_Latn	69.95	tha_Thai	87.7
bak_Cyrl	78.45	ibo_Latn	77.8	mri_Latn	79.75	tir_Ethi	77.85
bam_Latn	78.35	ilo_Latn	84.4	mya_Mymr	81.45	tpi_Latn	74.4
ban_Latn	84.25	ind_Latn	92.6	nld_Latn	90.55	tsn_Latn	80.5
bel_Cyrl	88.5	isl_Latn	88.15	nno_Latn	84.15	tso_Latn	76.8
bem_Latn	74.0	ita_Latn	92.25	nob_Latn	92.7	tuk_Latn	74.25
ben_Beng	86.8	jav_Latn	89.4	npi_Deva	86.45	tum_Latn	73.0
bho_Deva	84.55	jpn_Jpan	83.25	nso_Latn	82.5	tur_Latn	85.8
bjn_Arab	70.4	kab_Latn	75.9	nus_Latn	71.5	twi_Latn	75.0
bjn_Latn	89.7	kac_Latn	72.05	nya_Latn	78.2	uig_Arab	80.25
bos_Latn	91.25	kam_Latn	64.25	oci_Latn	90.7	ukr_Cyrl	88.8
bug_Latn	81.25	kan_Knda	87.0	pag_Latn	82.15	umb_Latn	65.9
bul_Cyrl	90.25	kas_Arab	75.2	pan_Guru	87.4	urd_Arab	88.5
cat_Latn	92.1	kas_Deva	75.65	pap_Latn	87.45	uzn_Latn	84.0
ceb_Latn	87.75	kat_Geor	79.1	pbt_Arab	88.0	vec_Latn	89.95
ces_Latn	91.25	kaz_Cyrl	83.55	pes_Arab	87.85	vie_Latn	90.85
cjk_Latn	72.65	kbp_Latn	71.2	plt_Latn	87.2	war_Latn	85.1
ckb_Arab	76.5	kea_Latn	86.65	pol_Latn	90.2	wol_Latn	73.95
crh_Latn	83.8	khk_Cyrl	83.0	por_Latn	91.9	xho_Latn	85.55
cym_Latn	89.2	khm_Khmr	86.35	prs_Arab	89.2	ydd_Hebr	89.2
dan_Latn	92.25	kik_Latn	74.1	quy_Latn	68.7	yor_Latn	73.95
deu_Latn	90.6	kin_Latn	75.55	ron_Latn	91.2	yue_Hant	81.8
dik_Latn	72.85	kir_Cyrl	80.75	run_Latn	74.1	zho_Hans	82.4
dyu_Latn	69.45	kmb_Latn	67.7	rus_Cyrl	90.1	zho_Hant	67.15
ell_Grek	89.7	kmr_Latn	85.45	sag_Latn	74.05	zsm_Latn	92.6
eng_Latn	94.0	knc_Arab	77.7	san_Deva	75.35	zul_Latn	85.95
epo_Latn	91.3	knc_Latn	71.0	scn_Latn	87.8		
est_Latn	87.55	kon_Latn	75.0	shn_Mymr	68.1		
eus_Latn	78.65	kor_Hang	85.3	sin_Sinh	84.3		
ewe_Latn	75.9	lao_Laoo	89.25	slk_Latn	91.1		

Figure 2: The accuracy score of XLM-R model on PAWS-X task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	36.89	fao_Latn	58.26	lij_Latn	63.15	slv_Latn	80.4
ace_Latn	52.93	fij_Latn	41.34	lim_Latn	67.6	smo_Latn	40.68
acm_Arab	75.71	fin_Latn	80.02	lin_Latn	40.52	sna_Latn	40.82
acq_Arab	78.08	fon_Latn	40.2	lit_Latn	80.74	snd_Arab	77.05
aeb_Arab	65.63	fra_Latn	82.87	lmo_Latn	65.53	som_Latn	67.68
afr_Latn	83.53	fur_Latn	61.62	ltg_Latn	56.85	sot_Latn	40.74
ajp_Arab	75.63	fuv_Latn	39.9	ltz_Latn	50.56	spa_Latn	84.93
aka_Latn	40.14	gaz_Latn	56.53	lua_Latn	39.92	srd_Latn	65.19
als_Latn	81.22	gla_Latn	70.76	lug_Latn	42.32	srp_Cyrl	82.69
amh_Ethi	69.1	gle_Latn	75.05	luo_Latn	39.5	ssw_Latn	44.31
apc_Arab	72.87	glg_Latn	84.35	lus_Latn	42.38	sun_Latn	74.03
arb_Arab	80.16	grn_Latn	43.49	lvs_Latn	77.62	swe_Latn	82.99
ars_Arab	76.91	guj_Gujr	77.52	mag_Deva	70.74	swh_Latn	74.39
ary_Arab	69.24	hat_Latn	44.91	mai_Deva	67.05	szl_Latn	71.84
arz_Arab	77.92	hau_Latn	68.96	mal_Mlym	76.45	tam_Taml	76.37
asm_Beng	72.57	heb_Hebr	82.48	mar_Deva	78.26	taq_Latn	38.38
ast_Latn	70.66	hin_Deva	80.36	min_Latn	62.22	tat_Cyrl	48.24
awa_Deva	62.28	hne_Deva	69.6	mkd_Cyrl	82.02	tel_Telu	75.65
ayr_Latn	40.36	hrv_Latn	81.86	mlt_Latn	43.27	tgk_Cyrl	39.42
azb_Arab	40.42	hun_Latn	81.4	mni_Beng	36.39	tgl_Latn	79.34
azj_Latn	78.54	hye_Armn	78.28	mos_Latn	38.98	tha_Thai	76.93
bak_Cyrl	45.53	ibo_Latn	41.32	mri_Latn	40.12	tir_Ethi	45.55
bam_Latn	40.82	ilo_Latn	43.03	mya_Mymr	72.02	tpi_Latn	48.16
ban_Latn	49.54	ind_Latn	83.65	nld_Latn	83.09	tsn_Latn	41.36
bel_Cyrl	81.58	isl_Latn	77.35	nno_Latn	74.91	tso_Latn	41.4
bem_Latn	40.18	ita_Latn	84.11	nob_Latn	84.89	tuk_Latn	49.22
ben_Beng	77.86	jav_Latn	75.63	npi_Deva	74.71	tum_Latn	40.16
bho_Deva	71.76	jpn_Jpan	73.91	nso_Latn	41.28	tur_Latn	79.98
bjn_Arab	36.25	kab_Latn	37.39	nus_Latn	38.28	twi_Latn	40.3
bjn_Latn	65.79	kac_Latn	39.38	nya_Latn	42.04	uig_Arab	63.71
bos_Latn	82.36	kam_Latn	36.63	oci_Latn	71.14	ukr_Cyrl	81.64
bug_Latn	43.69	kan_Knda	77.47	pag_Latn	41.82	umb_Latn	37.09
bul_Cyrl	82.99	kas_Arab	48.54	pan_Guru	77.03	urd_Arab	78.12
cat_Latn	83.11	kas_Deva	50.78	pap_Latn	63.19	uzn_Latn	77.62
ceb_Latn	57.27	kat_Geor	56.23	pbt_Arab	78.32	vec_Latn	75.35
ces_Latn	82.26	kaz_Cyrl	76.25	pes_Arab	79.18	vie_Latn	80.72
cjk_Latn	40.02	kbp_Latn	37.37	plt_Latn	69.44	war_Latn	48.84
ckb_Arab	40.26	kea_Latn	51.64	pol_Latn	81.88	wol_Latn	40.58
crh_Latn	64.55	khk_Cyrl	74.59	por_Latn	84.23	xho_Latn	60.0
cym_Latn	79.04	khm_Khmr	73.87	prs_Arab	80.0	ydd_Hebr	74.81
dan_Latn	83.07	kik_Latn	38.54	quy_Latn	38.08	yor_Latn	38.88
deu_Latn	82.79	kin_Latn	39.68	ron_Latn	83.27	yue_Hant	70.16
dik_Latn	39.84	kir_Cyrl	73.99	run_Latn	40.42	zho_Hans	72.42
dyu_Latn	36.59	kmb_Latn	37.58	rus_Cyrl	81.76	zho_Hant	60.08
ell_Grek	82.12	kmr_Latn	72.59	sag_Latn	42.95	zsm_Latn	82.34
eng_Latn	87.25	knc_Arab	59.5	san_Deva	68.16	zul_Latn	60.0
epo_Latn	80.56	knc_Latn	39.34	scn_Latn	65.53		
est_Latn	80.92	kon_Latn	40.96	shn_Mymr	37.07		
eus_Latn	66.57	kor_Hang	77.86	sin_Sinh	75.39		
ewe_Latn	41.32	lao_Laoo	77.94	slk_Latn	82.1		

Figure 3: The accuracy score of XLM-R model on XNLI task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	49.44	fao_Latn	65.12	lij_Latn	59.7	slv_Latn	82.06
ace_Latn	58.24	fij_Latn	53.61	lim_Latn	72.73	smo_Latn	50.83
acm_Arab	72.27	fin_Latn	85.04	lin_Latn	49.77	sna_Latn	52.15
acq_Arab	79.29	fon_Latn	50.76	lit_Latn	78.56	snd_Arab	75.25
aeb_Arab	65.78	fra_Latn	81.47	lmo_Latn	63.86	som_Latn	64.2
afr_Latn	79.95	fur_Latn	60.89	ltg_Latn	60.42	sot_Latn	52.08
ajp_Arab	77.5	fuv_Latn	50.3	ltz_Latn	58.9	spa_Latn	84.65
aka_Latn	51.82	gaz_Latn	57.58	lua_Latn	52.42	srd_Latn	61.09
als_Latn	79.09	gla_Latn	63.73	lug_Latn	50.5	srp_Cyrl	78.76
amh_Ethi	66.05	gle_Latn	71.34	luo_Latn	53.61	ssw_Latn	57.05
apc_Arab	76.51	glg_Latn	84.18	lus_Latn	54.73	sun_Latn	72.87
arb_Arab	80.28	grn_Latn	54.8	lvs_Latn	77.1	swe_Latn	86.37
ars_Arab	80.28	guj_Gujr	73.46	mag_Deva	67.97	swh_Latn	75.71
ary_Arab	73.06	hat_Latn	54.53	mai_Deva	64.2	szl_Latn	70.35
arz_Arab	77.7	hau_Latn	68.03	mal_Mlym	77.9	tam_Taml	76.04
asm_Beng	64.2	heb_Hebr	82.79	mar_Deva	77.75	taq_Latn	50.96
ast_Latn	74.39	hin_Deva	78.29	min_Latn	62.94	tat_Cyrl	56.85
awa_Deva	64.2	hne_Deva	66.91	mkd_Cyrl	82.06	tel_Telu	76.31
ayr_Latn	51.29	hrv_Latn	83.52	mlt_Latn	50.43	tgk_Cyrl	52.55
azb_Arab	54.33	hun_Latn	82.26	mni_Beng	51.56	tgl_Latn	73.53
azj_Latn	80.15	hye_Armn	76.17	mos_Latn	50.36	tha_Thai	85.51
bak_Cyrl	58.5	ibo_Latn	51.95	mri_Latn	51.09	tir_Ethi	54.33
bam_Latn	51.95	ilo_Latn	54.47	mya_Mymr	70.2	tpi_Latn	49.9
ban_Latn	58.04	ind_Latn	88.75	nld_Latn	84.71	tsn_Latn	52.61
bel_Cyrl	78.49	isl_Latn	76.24	nno_Latn	79.95	tso_Latn	52.81
bem_Latn	54.2	ita_Latn	82.33	nob_Latn	85.9	tuk_Latn	59.96
ben_Beng	75.25	jav_Latn	73.46	npi_Deva	74.92	tum_Latn	51.75
bho_Deva	70.62	jpn_Jpan	79.95	nso_Latn	52.15	tur_Latn	84.25
bjn_Arab	50.63	kab_Latn	48.97	nus_Latn	52.68	twi_Latn	51.09
bjn_Latn	69.23	kac_Latn	52.35	nya_Latn	50.36	uig_Arab	70.22
bos_Latn	81.67	kam_Latn	50.69	oci_Latn	71.87	ukr_Cyrl	82.26
bug_Latn	53.34	kan_Knda	74.72	pag_Latn	53.28	umb_Latn	51.75
bul_Cyrl	82.46	kas_Arab	54.53	pan_Guru	70.46	urd_Arab	76.37
cat_Latn	81.4	kas_Deva	64.2	pap_Latn	59.3	uzn_Latn	73.4
ceb_Latn	60.36	kat_Geor	56.98	pbt_Arab	72.14	vec_Latn	74.52
ces_Latn	83.06	kaz_Cyrl	76.7	pes_Arab	79.75	vie_Latn	82.53
cjk_Latn	49.83	kbp_Latn	52.08	plt_Latn	64.99	war_Latn	58.44
ckb_Arab	51.69	kea_Latn	58.44	pol_Latn	83.26	wol_Latn	50.23
crh_Latn	69.82	khk_Cyrl	75.84	por_Latn	84.32	xho_Latn	56.98
cym_Latn	71.74	khm_Khmr	72.07	prs_Arab	81.14	ydd_Hebr	63.14
dan_Latn	85.9	kik_Latn	50.83	quy_Latn	47.39	yor_Latn	51.62
deu_Latn	82.53	kin_Latn	49.5	ron_Latn	80.94	yue_Hant	79.95
dik_Latn	52.68	kir_Cyrl	75.78	run_Latn	47.92	zho_Hans	86.9
dyu_Latn	50.43	kmb_Latn	50.63	rus_Cyrl	83.45	zho_Hant	79.62
ell_Grek	76.84	kmr_Latn	69.03	sag_Latn	52.28	zsm_Latn	87.29
eng_Latn	88.82	knc_Arab	58.9	san_Deva	65.59	zul_Latn	58.7
epo_Latn	78.89	knc_Latn	51.42	scn_Latn	66.91		
est_Latn	82.73	kon_Latn	53.61	shn_Mymr	46.39		
eus_Latn	76.37	kor_Hang	78.42	sin_Sinh	77.43		
ewe_Latn	52.68	lao_Laoo	76.37	slk_Latn	83.85		

Figure 4: The accuracy score of XLM-R model on XStoryCloze task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	50.0	fao_Latn	61.6	lij_Latn	62.0	slv_Latn	72.6
ace_Latn	54.8	fij_Latn	51.6	lim_Latn	64.2	smo_Latn	44.4
acm_Arab	67.0	fin_Latn	72.0	lin_Latn	51.4	sna_Latn	49.4
acq_Arab	66.6	fon_Latn	48.8	lit_Latn	69.6	snd_Arab	64.2
aeb_Arab	60.0	fra_Latn	73.6	lmo_Latn	60.6	som_Latn	57.0
afr_Latn	71.0	fur_Latn	56.2	ltg_Latn	59.6	sot_Latn	49.2
ajp_Arab	68.4	fuv_Latn	49.4	ltz_Latn	57.0	spa_Latn	73.4
aka_Latn	50.4	gaz_Latn	50.0	lua_Latn	48.8	srd_Latn	59.4
als_Latn	67.2	gla_Latn	57.6	lug_Latn	50.4	srp_Cyrl	66.8
amh_Ethi	61.8	gle_Latn	61.0	luo_Latn	50.0	ssw_Latn	52.6
apc_Arab	66.4	glg_Latn	72.4	lus_Latn	50.8	sun_Latn	60.4
arb_Arab	71.2	grn_Latn	48.8	lvs_Latn	68.6	swe_Latn	74.6
ars_Arab	69.6	guj_Gujr	64.6	mag_Deva	60.2	swh_Latn	66.4
ary_Arab	64.8	hat_Latn	50.4	mai_Deva	58.6	szl_Latn	63.4
arz_Arab	70.8	hau_Latn	54.4	mal_Mlym	66.6	tam_Taml	71.8
asm_Beng	62.2	heb_Hebr	67.0	mar_Deva	67.2	taq_Latn	48.4
ast_Latn	65.4	hin_Deva	66.4	min_Latn	53.8	tat_Cyrl	47.0
awa_Deva	64.2	hne_Deva	62.2	mkd_Cyrl	69.6	tel_Telu	65.8
ayr_Latn	48.2	hrv_Latn	71.0	mlt_Latn	51.4	tgk_Cyrl	50.2
azb_Arab	54.2	hun_Latn	68.6	mni_Beng	52.4	tgl_Latn	67.4
azj_Latn	67.2	hye_Armn	67.8	mos_Latn	48.4	tha_Thai	69.4
bak_Cyrl	50.4	ibo_Latn	49.0	mri_Latn	49.6	tir_Ethi	50.2
bam_Latn	51.4	ilo_Latn	50.8	mya_Mymr	65.4	tpi_Latn	50.0
ban_Latn	57.0	ind_Latn	74.6	nld_Latn	71.4	tsn_Latn	50.6
bel_Cyrl	64.8	isl_Latn	65.6	nno_Latn	70.6	tso_Latn	51.4
bem_Latn	53.0	ita_Latn	74.0	nob_Latn	70.6	tuk_Latn	51.6
ben_Beng	67.0	jav_Latn	59.8	npi_Deva	64.13	tum_Latn	50.6
bho_Deva	66.4	jpn_Jpan	72.4	nso_Latn	55.4	tur_Latn	68.6
bjn_Arab	49.8	kab_Latn	51.4	nus_Latn	49.0	twi_Latn	50.2
bjn_Latn	62.2	kac_Latn	50.0	nya_Latn	51.0	uig_Arab	60.4
bos_Latn	70.4	kam_Latn	52.4	oci_Latn	64.0	ukr_Cyrl	68.4
bug_Latn	51.2	kan_Knda	67.6	pag_Latn	51.6	umb_Latn	49.2
bul_Cyrl	70.8	kas_Arab	50.4	pan_Guru	61.8	urd_Arab	69.0
cat_Latn	72.0	kas_Deva	56.2	pap_Latn	55.6	uzn_Latn	64.2
ceb_Latn	59.0	kat_Geor	55.8	pbt_Arab	66.8	vec_Latn	62.4
ces_Latn	70.4	kaz_Cyrl	65.8	pes_Arab	66.4	vie_Latn	69.2
cjk_Latn	50.2	kbp_Latn	48.2	plt_Latn	58.2	war_Latn	54.0
ckb_Arab	47.2	kea_Latn	52.0	pol_Latn	70.8	wol_Latn	51.2
crh_Latn	60.2	khk_Cyrl	66.2	por_Latn	72.8	xho_Latn	52.8
cym_Latn	63.6	khm_Khmr	68.8	prs_Arab	69.6	ydd_Hebr	62.6
dan_Latn	74.8	kik_Latn	50.6	quy_Latn	50.0	yor_Latn	45.2
deu_Latn	74.2	kin_Latn	51.4	ron_Latn	73.8	yue_Hant	68.6
dik_Latn	49.2	kir_Cyrl	65.4	run_Latn	52.4	zho_Hans	70.6
dyu_Latn	52.0	kmb_Latn	49.6	rus_Cyrl	69.6	zho_Hant	72.0
ell_Grek	73.2	kmr_Latn	62.4	sag_Latn	52.2	zsm_Latn	73.0
eng_Latn	78.6	knc_Arab	54.4	san_Deva	58.8	zul_Latn	49.6
epo_Latn	67.0	knc_Latn	48.8	scn_Latn	63.8		
est_Latn	68.4	kon_Latn	51.0	shn_Mymr	50.0		
eus_Latn	59.6	kor_Hang	70.8	sin_Sinh	66.0		
ewe_Latn	51.8	lao_Laoo	67.2	slk_Latn	69.0		

Figure 5: The accuracy score of XLM-R model on XCOPA task across 196 languages.

F Full results BLOOMz

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.35	fur_Latn	65.45	mar_Deva	71.68	tum_Latn	66.49
ace_Latn	66.27	fuv_Latn	66.71	min_Latn	67.09	tur_Latn	61.68
acm_Arab	68.54	gaz_Latn	66.39	mlt_Latn	64.69	twi_Latn	66.73
acq_Arab	68.82	gla_Latn	66.35	mni_Beng	62.16	uig_Arab	59.8
aeb_Arab	67.15	gle_Latn	66.51	mos_Latn	66.75	umb_Latn	65.85
afr_Latn	65.87	glg_Latn	71.28	mri_Latn	66.83	urd_Arab	69.76
ajp_Arab	68.26	grn_Latn	66.33	nld_Latn	65.89	uzn_Latn	65.33
aka_Latn	67.03	guj_Gujr	70.44	nno_Latn	66.63	vec_Latn	66.59
als_Latn	62.89	hat_Latn	65.37	nob_Latn	65.95	vie_Latn	73.01
apc_Arab	67.66	hau_Latn	65.77	npi_Deva	70.86	war_Latn	64.97
arb_Arab	72.4	hin_Deva	71.9	nso_Latn	68.76	wol_Latn	65.43
ars_Arab	69.16	hne_Deva	65.85	nus_Latn	66.01	xho_Latn	66.83
ary_Arab	68.34	hrv_Latn	66.85	nya_Latn	67.86	yor_Latn	68.04
arz_Arab	68.82	hun_Latn	67.15	oci_Latn	65.05	yue_Hant	70.38
asm_Beng	68.16	ibo_Latn	68.32	pag_Latn	66.71	zho_Hans	72.18
ast_Latn	67.84	ilo_Latn	66.31	pap_Latn	65.53	zho_Hant	68.14
awa_Deva	67.33	ind_Latn	72.26	pbt_Arab	64.07	zsm_Latn	71.76
ayr_Latn	66.33	isl_Latn	66.43	pes_Arab	62.28	zul_Latn	67.01
azb_Arab	63.91	ita_Latn	70.36	plt_Latn	66.33		
azj_Latn	59.6	jav_Latn	64.27	pol_Latn	65.81		
bam_Latn	65.55	kab_Latn	66.01	por_Latn	74.09		
ban_Latn	65.97	kac_Latn	66.19	prs_Arab	62.0		
bem_Latn	67.09	kam_Latn	66.25	quy_Latn	66.73		
ben_Beng	70.46	kan_Knda	71.04	ron_Latn	65.17		
bho_Deva	68.62	kas_Arab	63.47	run_Latn	67.6		
bjn_Arab	64.57	kas_Deva	65.63	sag_Latn	66.79		
bjn_Latn	67.82	kbp_Latn	63.23	san_Deva	65.47		
bos_Latn	67.37	kea_Latn	64.15	scn_Latn	63.75		
bug_Latn	64.95	kik_Latn	65.95	slk_Latn	66.69		
cat_Latn	73.47	kin_Latn	68.02	slv_Latn	67.19		
ceb_Latn	66.05	kmb_Latn	65.99	smo_Latn	66.07		
ces_Latn	66.07	kmr_Latn	65.01	sna_Latn	67.92		
cjk_Latn	66.01	knc_Arab	65.87	snd_Arab	64.73		
ckb_Arab	61.42	knc_Latn	65.13	som_Latn	66.15		
crh_Latn	62.24	kon_Latn	66.91	sot_Latn	67.01		
cym_Latn	65.81	lij_Latn	64.27	spa_Latn	73.59		
dan_Latn	66.77	lim_Latn	64.89	srd_Latn	65.99		
deu_Latn	68.34	lin_Latn	68.62	ssw_Latn	66.37		
dik_Latn	64.59	lit_Latn	66.27	sun_Latn	64.77		
dyu_Latn	66.01	lmo_Latn	64.47	swe_Latn	66.27		
eng_Latn	72.63	ltg_Latn	65.83	swh_Latn	70.94		
epo_Latn	63.57	ltz_Latn	64.37	szl_Latn	65.13		
est_Latn	66.25	lua_Latn	67.33	tam_Taml	70.66		
eus_Latn	68.86	lug_Latn	67.52	taq_Latn	64.89		
ewe_Latn	66.53	luo_Latn	66.33	tel_Telu	70.64		
fao_Latn	67.01	lus_Latn	65.73	tgl_Latn	65.79		
fij_Latn	67.03	lvs_Latn	64.05	tpi_Latn	66.89		
fin_Latn	65.85	mag_Deva	67.82	tsn_Latn	67.78		
fon_Latn	66.57	mai_Deva	67.58	tso_Latn	67.27		
fra_Latn	73.13	mal_Mlym	71.36	tuk_Latn	62.34		

Figure 6: The accuracy score of BLOOMz model on B-NLI task across 168 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	11.18	fur_Latn	67.57	mar_Deva	73.2	tum_Latn	64.06
ace_Latn	60.29	fuv_Latn	51.69	min_Latn	64.46	tur_Latn	50.89
acm_Arab	81.07	gaz_Latn	54.07	mlt_Latn	56.12	twi_Latn	57.45
acq_Arab	85.31	gla_Latn	50.56	mni_Beng	53.67	uig_Arab	45.47
aeb_Arab	65.25	gle_Latn	54.27	mos_Latn	50.43	umb_Latn	48.25
afr_Latn	58.97	glg_Latn	86.83	mri_Latn	51.89	urd_Arab	69.49
ajp_Arab	83.85	grn_Latn	56.25	nld_Latn	64.06	uzn_Latn	51.29
aka_Latn	56.85	guj_Gujr	79.75	nno_Latn	60.62	vec_Latn	75.65
als_Latn	53.14	hat_Latn	60.49	nob_Latn	61.48	vie_Latn	89.81
apc_Arab	80.54	hau_Latn	53.87	npi_Deva	70.75	war_Latn	55.53
arb_Arab	88.02	hin_Deva	84.45	nso_Latn	65.65	wol_Latn	55.39
ars_Arab	86.1	hne_Deva	30.18	nus_Latn	49.17	xho_Latn	65.65
ary_Arab	73.86	hrv_Latn	55.86	nya_Latn	68.3	yor_Latn	71.54
arz_Arab	84.05	hun_Latn	51.16	oci_Latn	82.26	yue_Hant	42.22
asm_Beng	79.09	ibo_Latn	66.84	pag_Latn	53.47	zho_Hans	69.62
ast_Latn	82.26	ilo_Latn	56.06	pap_Latn	64.33	zho_Hant	31.5
awa_Deva	30.77	ind_Latn	89.54	pbt_Arab	21.91	zsm_Latn	88.48
ayr_Latn	56.85	isl_Latn	51.89	pes_Arab	39.44	zul_Latn	66.51
azb_Arab	1.13	ita_Latn	82.73	plt_Latn	55.26		
azj_Latn	52.08	jav_Latn	64.39	pol_Latn	56.59		
bam_Latn	56.92	kab_Latn	53.01	por_Latn	90.14		
ban_Latn	57.58	kac_Latn	53.87	prs_Arab	45.8		
bem_Latn	54.53	kam_Latn	52.15	quy_Latn	54.73		
ben_Beng	84.32	kan_Knda	76.7	ron_Latn	62.94		
bho_Deva	31.17	kas_Arab	2.65	run_Latn	61.02		
bjn_Arab	20.58	kas_Deva	32.3	sag_Latn	53.54		
bjn_Latn	72.14	kbp_Latn	49.31	san_Deva	34.81		
bos_Latn	57.38	kea_Latn	62.34	scn_Latn	66.71		
bug_Latn	53.74	kik_Latn	57.45	slk_Latn	55.33		
cat_Latn	89.94	kin_Latn	62.94	slv_Latn	54.6		
ceb_Latn	55.79	kmb_Latn	49.7	smo_Latn	52.55		
ces_Latn	56.52	kmr_Latn	52.68	sna_Latn	66.12		
cjk_Latn	52.15	knc_Arab	62.08	snd_Arab	6.42		
ckb_Arab	31.7	knc_Latn	53.01	som_Latn	51.95		
crh_Latn	51.62	kon_Latn	53.67	sot_Latn	65.19		
cym_Latn	51.95	lij_Latn	64.39	spa_Latn	91.0		
dan_Latn	62.01	lim_Latn	60.23	srd_Latn	63.4		
deu_Latn	71.94	lin_Latn	63.8	ssw_Latn	59.43		
dik_Latn	53.01	lit_Latn	54.73	sun_Latn	63.4		
dyu_Latn	54.4	lmo_Latn	70.55	swe_Latn	60.82		
eng_Latn	93.12	ltg_Latn	53.54	swh_Latn	79.09		
epo_Latn	64.99	ltz_Latn	58.57	szl_Latn	54.14		
est_Latn	51.29	lua_Latn	50.36	tam_Taml	76.64		
eus_Latn	77.83	lug_Latn	62.01	taq_Latn	53.01		
ewe_Latn	53.41	luo_Latn	53.34	tel_Telu	72.73		
fao_Latn	54.86	lus_Latn	54.86	tgl_Latn	54.8		
fij_Latn	52.02	lvs_Latn	53.08	tpi_Latn	56.78		
fin_Latn	53.28	mag_Deva	36.6	tsn_Latn	64.99		
fon_Latn	38.19	mai_Deva	41.69	tso_Latn	64.92		
fra_Latn	89.94	mal_Mlym	80.08	tuk_Latn	50.3		

Figure 7: The accuracy score of BLOOMz model on XStoryCloze task across 168 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.35	fur_Latn	65.45	mar_Deva	71.68	tum_Latn	66.49
ace_Latn	66.27	fuv_Latn	66.71	min_Latn	67.09	tur_Latn	61.68
acm_Arab	68.54	gaz_Latn	66.39	mlt_Latn	64.69	twi_Latn	66.73
acq_Arab	68.82	gla_Latn	66.35	mni_Beng	62.16	uig_Arab	59.8
aeb_Arab	67.15	gle_Latn	66.51	mos_Latn	66.75	umb_Latn	65.85
afr_Latn	65.87	glg_Latn	71.28	mri_Latn	66.83	urd_Arab	69.76
ajp_Arab	68.26	grn_Latn	66.33	nld_Latn	65.89	uzn_Latn	65.33
aka_Latn	67.03	guj_Gujr	70.44	nno_Latn	66.63	vec_Latn	66.59
als_Latn	62.89	hat_Latn	65.37	nob_Latn	65.95	vie_Latn	73.01
apc_Arab	67.66	hau_Latn	65.77	npi_Deva	70.86	war_Latn	64.97
arb_Arab	72.4	hin_Deva	71.9	nso_Latn	68.76	wol_Latn	65.43
ars_Arab	69.16	hne_Deva	65.85	nus_Latn	66.01	xho_Latn	66.83
ary_Arab	68.34	hrv_Latn	66.85	nya_Latn	67.86	yor_Latn	68.04
arz_Arab	68.82	hun_Latn	67.15	oci_Latn	65.05	yue_Hant	70.38
asm_Beng	68.16	ibo_Latn	68.32	pag_Latn	66.71	zho_Hans	72.18
ast_Latn	67.84	ilo_Latn	66.31	pap_Latn	65.53	zho_Hant	68.14
awa_Deva	67.33	ind_Latn	72.26	pbt_Arab	64.07	zsm_Latn	71.76
ayr_Latn	66.33	isl_Latn	66.43	pes_Arab	62.28	zul_Latn	67.01
azb_Arab	63.91	ita_Latn	70.36	plt_Latn	66.33		
azj_Latn	59.6	jav_Latn	64.27	pol_Latn	65.81		
bam_Latn	65.55	kab_Latn	66.01	por_Latn	74.09		
ban_Latn	65.97	kac_Latn	66.19	prs_Arab	62.0		
bem_Latn	67.09	kam_Latn	66.25	quy_Latn	66.73		
ben_Beng	70.46	kan_Knda	71.04	ron_Latn	65.17		
bho_Deva	68.62	kas_Arab	63.47	run_Latn	67.6		
bjn_Arab	64.57	kas_Deva	65.63	sag_Latn	66.79		
bjn_Latn	67.82	kbp_Latn	63.23	san_Deva	65.47		
bos_Latn	67.37	kea_Latn	64.15	scn_Latn	63.75		
bug_Latn	64.95	kik_Latn	65.95	slk_Latn	66.69		
cat_Latn	73.47	kin_Latn	68.02	slv_Latn	67.19		
ceb_Latn	66.05	kmb_Latn	65.99	smo_Latn	66.07		
ces_Latn	66.07	kmr_Latn	65.01	sna_Latn	67.92		
cjk_Latn	66.01	knc_Arab	65.87	snd_Arab	64.73		
ckb_Arab	61.42	knc_Latn	65.13	som_Latn	66.15		
crh_Latn	62.24	kon_Latn	66.91	sot_Latn	67.01		
cym_Latn	65.81	lij_Latn	64.27	spa_Latn	73.59		
dan_Latn	66.77	lim_Latn	64.89	srd_Latn	65.99		
deu_Latn	68.34	lin_Latn	68.62	ssw_Latn	66.37		
dik_Latn	64.59	lit_Latn	66.27	sun_Latn	64.77		
dyu_Latn	66.01	lmo_Latn	64.47	swe_Latn	66.27		
eng_Latn	72.63	ltg_Latn	65.83	swh_Latn	70.94		
epo_Latn	63.57	ltz_Latn	64.37	szl_Latn	65.13		
est_Latn	66.25	lua_Latn	67.33	tam_Taml	70.66		
eus_Latn	68.86	lug_Latn	67.52	taq_Latn	64.89		
ewe_Latn	66.53	luo_Latn	66.33	tel_Telu	70.64		
fao_Latn	67.01	lus_Latn	65.73	tgl_Latn	65.79		
fij_Latn	67.03	lvs_Latn	64.05	tpi_Latn	66.89		
fin_Latn	65.85	mag_Deva	67.82	tsn_Latn	67.78		
fon_Latn	66.57	mai_Deva	67.58	tso_Latn	67.27		
fra_Latn	73.13	mal_Mlym	71.36	tuk_Latn	62.34		

Figure 8: The accuracy score of BLOOMz model on XCOPA task across 168 languages.

G Full Results for AYA

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	55.0	fao_Latn	88.0	lij_Latn	84.0	slv_Latn	92.0
ace_Latn	81.0	fij_Latn	73.0	lim_Latn	89.0	smo_Latn	89.0
acm_Arab	88.0	fin_Latn	71.0	lin_Latn	72.0	sna_Latn	85.0
acq_Arab	90.0	fon_Latn	41.0	lit_Latn	90.0	snd_Arab	87.0
aeb_Arab	78.0	fra_Latn	93.0	lmo_Latn	86.0	som_Latn	83.0
afr_Latn	94.0	fur_Latn	86.0	ltg_Latn	77.0	sot_Latn	80.0
ajp_Arab	87.0	fuv_Latn	51.0	ltz_Latn	90.0	spa_Latn	94.0
aka_Latn	73.0	gaz_Latn	73.0	lua_Latn	67.0	srd_Latn	88.0
als_Latn	92.0	gla_Latn	86.0	lug_Latn	75.0	srp_Cyrl	93.0
amh_Ethi	80.0	gle_Latn	90.0	luo_Latn	51.0	ssw_Latn	77.0
apc_Arab	86.0	glg_Latn	90.0	lus_Latn	70.0	sun_Latn	90.0
arb_Arab	92.0	grn_Latn	55.0	lvs_Latn	89.0	swe_Latn	91.0
ars_Arab	90.0	guj_Gujr	89.0	mag_Deva	89.0	swh_Latn	89.0
ary_Arab	85.0	hat_Latn	91.0	mai_Deva	88.0	szl_Latn	87.0
arz_Arab	89.0	hau_Latn	86.0	mal_Mlym	90.0	tam_Taml	70.0
asm_Beng	87.0	heb_Hebr	85.0	mar_Deva	90.0	taq_Latn	51.0
ast_Latn	85.0	hin_Deva	91.0	min_Latn	83.0	tat_Cyrl	90.0
awa_Deva	78.0	hne_Deva	88.0	mkd_Cyrl	91.0	tel_Telu	88.0
ayr_Latn	51.0	hrv_Latn	91.0	mlt_Latn	88.0	tgk_Cyrl	90.0
azb_Arab	83.0	hun_Latn	92.0	mni_Beng	57.0	tgl_Latn	1.0
azj_Latn	90.0	hye_Armn	74.0	mos_Latn	49.0	tha_Thai	90.0
bak_Cyrl	89.0	ibo_Latn	86.0	mri_Latn	85.0	tir_Ethi	80.0
bam_Latn	59.0	ilo_Latn	65.0	mya_Mymr	86.0	tpi_Latn	81.0
ban_Latn	81.0	ind_Latn	87.0	nld_Latn	92.0	tsn_Latn	83.0
bel_Cyrl	91.0	isl_Latn	89.0	nno_Latn	91.0	tso_Latn	68.0
bem_Latn	60.0	ita_Latn	91.0	nob_Latn	94.0	tuk_Latn	77.0
ben_Beng	89.0	jav_Latn	81.0	npi_Deva	83.0	tum_Latn	80.0
bho_Deva	89.0	jpn_Jpan	90.0	nso_Latn	86.0	tur_Latn	91.0
bjn_Arab	57.0	kab_Latn	50.0	nus_Latn	44.0	twi_Latn	74.0
bjn_Latn	84.0	kac_Latn	56.0	nya_Latn	88.0	uig_Arab	77.0
bos_Latn	90.0	kam_Latn	51.0	oci_Latn	89.0	ukr_Cyrl	92.0
bug_Latn	57.0	kan_Knda	89.0	pag_Latn	58.0	umb_Latn	52.0
bul_Cyrl	85.0	kas_Arab	67.0	pan_Guru	91.0	urd_Arab	89.0
cat_Latn	93.0	kas_Deva	73.0	pap_Latn	87.0	uzn_Latn	90.0
ceb_Latn	36.0	kat_Geor	64.0	pbt_Arab	89.0	vec_Latn	89.0
ces_Latn	92.0	kaz_Cyrl	91.0	pes_Arab	89.0	vie_Latn	91.0
cjk_Latn	53.0	kbp_Latn	38.0	plt_Latn	41.0	war_Latn	65.0
ckb_Arab	87.0	kea_Latn	72.0	pol_Latn	93.0	wol_Latn	49.0
crh_Latn	86.0	khk_Cyrl	86.0	por_Latn	93.0	xho_Latn	86.0
cym_Latn	89.0	khm_Khmr	87.0	prs_Arab	91.0	ydd_Hebr	83.0
dan_Latn	93.0	kik_Latn	50.0	quy_Latn	62.0	yor_Latn	74.0
deu_Latn	93.0	kin_Latn	84.0	ron_Latn	93.0	yue_Hant	86.0
dik_Latn	46.0	kir_Cyrl	90.0	run_Latn	80.0	zho_Hans	90.0
dyu_Latn	56.0	kmb_Latn	59.0	rus_Cyrl	82.0	zho_Hant	81.0
ell_Grek	93.0	kmr_Latn	85.0	sag_Latn	74.0	zsm_Latn	90.0
eng_Latn	92.0	knc_Arab	80.0	san_Deva	75.0	zul_Latn	89.0
epo_Latn	93.0	knc_Latn	47.0	scn_Latn	87.0		
est_Latn	93.0	kon_Latn	62.0	shn_Mymr	70.0		
eus_Latn	75.0	kor_Hang	91.0	sin_Sinh	88.0		
ewe_Latn	65.0	lao_Laoo	88.0	slk_Latn	93.0		

Figure 9: The accuracy score of AYA on XStoryCloze task across 198 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	51.4	fao_Latn	76.0	lij_Latn	72.0	slv_Latn	86.2
ace_Latn	69.4	fij_Latn	56.8	lim_Latn	82.6	smo_Latn	76.8
acm_Arab	79.2	fin_Latn	85.0	lin_Latn	54.4	sna_Latn	72.8
acq_Arab	80.0	fon_Latn	50.6	lit_Latn	85.0	snd_Arab	77.2
aeb_Arab	70.8	fra_Latn	88.8	lmo_Latn	79.8	som_Latn	75.2
afr_Latn	86.4	fur_Latn	72.8	ltg_Latn	62.4	sot_Latn	79.4
ajp_Arab	78.4	fuv_Latn	51.0	ltz_Latn	81.2	spa_Latn	88.4
aka_Latn	58.6	gaz_Latn	56.8	lua_Latn	54.8	srd_Latn	78.2
als_Latn	84.0	gla_Latn	71.6	lug_Latn	61.0	srp_Cyrl	85.4
amh_Ethi	76.2	gle_Latn	74.0	luo_Latn	50.6	ssw_Latn	69.0
apc_Arab	75.6	glg_Latn	87.4	lus_Latn	58.4	sun_Latn	84.4
arb_Arab	82.0	grn_Latn	55.0	lvs_Latn	79.2	swe_Latn	87.0
ars_Arab	80.2	guj_Gujr	84.2	mag_Deva	78.2	swh_Latn	82.2
ary_Arab	75.6	hat_Latn	83.0	mai_Deva	79.4	szl_Latn	76.8
arz_Arab	80.4	hau_Latn	78.2	mal_Mlym	82.4	tam_Taml	82.8
asm_Beng	80.2	heb_Hebr	84.4	mar_Deva	84.6	taq_Latn	49.6
ast_Latn	80.0	hin_Deva	84.4	min_Latn	72.8	tat_Cyrl	81.8
awa_Deva	68.6	hne_Deva	81.6	mkd_Cyrl	85.4	tel_Telu	81.8
ayr_Latn	51.4	hrv_Latn	80.4	mlt_Latn	74.0	tgk_Cyrl	81.2
azb_Arab	70.8	hun_Latn	85.4	mni_Beng	53.2	tgl_Latn	79.4
azj_Latn	84.4	hye_Armn	83.8	mos_Latn	50.8	tha_Thai	81.6
bak_Cyrl	79.0	ibo_Latn	75.0	mri_Latn	71.2	tir_Ethi	70.2
bam_Latn	51.6	ilo_Latn	62.6	mya_Mymr	80.4	tpi_Latn	64.6
ban_Latn	67.6	ind_Latn	86.8	nld_Latn	86.8	tsn_Latn	73.2
bel_Cyrl	85.2	isl_Latn	80.2	nno_Latn	84.0	tso_Latn	56.8
bem_Latn	55.6	ita_Latn	88.0	nob_Latn	86.0	tuk_Latn	66.6
ben_Beng	84.2	jav_Latn	80.8	npi_Deva	79.2	tum_Latn	66.8
bho_Deva	81.4	jpn_Jpan	83.4	nso_Latn	76.6	tur_Latn	85.4
bjn_Arab	50.8	kab_Latn	50.4	nus_Latn	50.8	twi_Latn	60.2
bjn_Latn	77.4	kac_Latn	51.4	nya_Latn	78.2	uig_Arab	65.6
bos_Latn	82.2	kam_Latn	50.8	oci_Latn	81.6	ukr_Cyrl	86.6
bug_Latn	55.4	kan_Knda	82.6	pag_Latn	57.0	umb_Latn	51.4
bul_Cyrl	87.4	kas_Arab	56.2	pan_Guru	86.4	urd_Arab	86.0
cat_Latn	86.8	kas_Deva	65.0	pap_Latn	72.8	uzn_Latn	83.0
ceb_Latn	78.8	kat_Geor	61.0	pbt_Arab	82.0	vec_Latn	83.6
ces_Latn	85.8	kaz_Cyrl	82.0	pes_Arab	80.2	vie_Latn	84.4
cjk_Latn	52.8	kbp_Latn	51.0	plt_Latn	77.4	war_Latn	71.4
ckb_Arab	73.4	kea_Latn	62.2	pol_Latn	85.4	wol_Latn	50.2
crh_Latn	78.8	khk_Cyrl	71.8	por_Latn	88.4	xho_Latn	75.6
cym_Latn	77.4	khm_Khmr	78.2	prs_Arab	82.8	ydd_Hebr	75.6
dan_Latn	85.4	kik_Latn	50.4	quy_Latn	56.4	yor_Latn	61.8
deu_Latn	89.8	kin_Latn	69.8	ron_Latn	86.0	yue_Hant	78.2
dik_Latn	50.0	kir_Cyrl	82.8	run_Latn	68.4	zho_Hans	83.6
dyu_Latn	52.6	kmb_Latn	53.2	rus_Cyrl	85.0	zho_Hant	78.0
ell_Grek	85.0	kmr_Latn	74.0	sag_Latn	61.0	zsm_Latn	84.4
eng_Latn	89.6	knc_Arab	70.4	san_Deva	71.0	zul_Latn	75.6
epo_Latn	88.0	knc_Latn	50.4	scn_Latn	82.4		
est_Latn	84.0	kon_Latn	52.6	shn_Mymr	63.0		
eus_Latn	77.6	kor_Hang	82.4	sin_Sinh	78.8		
ewe_Latn	52.0	lao_Laoo	81.4	slk_Latn	84.8		

Figure 10: The accuracy score of AYA on XCOPA task across 198 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.37	fao_Latn	71.9	lij_Latn	71.66	slv_Latn	79.78
ace_Latn	77.41	fij_Latn	69.66	lim_Latn	73.69	smo_Latn	74.51
acm_Arab	75.45	fin_Latn	78.02	lin_Latn	68.12	sna_Latn	77.94
acq_Arab	77.03	fon_Latn	51.2	lit_Latn	78.46	snd_Arab	76.41
aeb_Arab	73.01	fra_Latn	79.58	lmo_Latn	73.33	som_Latn	75.27
afr_Latn	76.13	fur_Latn	73.95	ltg_Latn	69.92	sot_Latn	73.97
ajp_Arab	77.49	fuv_Latn	57.01	ltz_Latn	77.84	spa_Latn	80.0
aka_Latn	70.8	gaz_Latn	68.7	lua_Latn	66.41	srd_Latn	78.52
als_Latn	76.37	gla_Latn	76.45	lug_Latn	65.67	srp_Cyrl	76.99
amh_Ethi	69.16	gle_Latn	73.63	luo_Latn	58.8	ssw_Latn	69.32
apc_Arab	76.77	glg_Latn	77.01	lus_Latn	68.38	sun_Latn	81.28
arb_Arab	78.76	grn_Latn	55.41	lvs_Latn	77.6	swe_Latn	74.21
ars_Arab	76.83	guj_Gujr	73.77	mag_Deva	69.5	swh_Latn	75.41
ary_Arab	75.33	hat_Latn	75.83	mai_Deva	70.8	szl_Latn	71.04
arz_Arab	77.88	hau_Latn	79.44	mal_Mlym	70.12	tam_Taml	77.45
asm_Beng	74.97	heb_Hebr	75.19	mar_Deva	76.01	taq_Latn	62.0
ast_Latn	72.83	hin_Deva	74.79	min_Latn	74.93	tat_Cyrl	76.59
awa_Deva	68.46	hne_Deva	70.76	mkd_Cyrl	75.19	tel_Telu	77.74
ayr_Latn	53.23	hrv_Latn	75.29	mlt_Latn	71.64	tgk_Cyrl	76.63
azb_Arab	73.03	hun_Latn	76.83	mni_Beng	53.55	tgl_Latn	77.86
azj_Latn	80.34	hye_Armn	77.33	mos_Latn	51.52	tha_Thai	79.08
bak_Cyrl	74.97	ibo_Latn	75.15	mri_Latn	80.16	tir_Ethi	69.0
bam_Latn	65.55	ilo_Latn	68.52	mya_Mymr	73.81	tpi_Latn	73.93
ban_Latn	74.71	ind_Latn	78.46	nld_Latn	76.63	tsn_Latn	74.45
bel_Cyrl	78.66	isl_Latn	72.06	nno_Latn	77.56	tso_Latn	64.03
bem_Latn	62.42	ita_Latn	79.56	nob_Latn	75.13	tuk_Latn	67.45
ben_Beng	79.2	jav_Latn	78.08	npi_Deva	73.95	tum_Latn	67.74
bho_Deva	70.64	jpn_Jpan	77.66	nso_Latn	76.21	tur_Latn	79.3
bjn_Arab	64.33	kab_Latn	53.77	nus_Latn	50.92	twi_Latn	71.38
bjn_Latn	78.28	kac_Latn	67.78	nya_Latn	72.93	uig_Arab	60.78
bos_Latn	75.83	kam_Latn	52.38	oci_Latn	76.73	ukr_Cyrl	80.32
bug_Latn	59.42	kan_Knda	75.51	pag_Latn	63.61	umb_Latn	55.87
bul_Cyrl	79.42	kas_Arab	61.98	pan_Guru	71.96	urd_Arab	75.35
cat_Latn	80.04	kas_Deva	59.2	pap_Latn	74.35	uzn_Latn	78.14
ceb_Latn	76.55	kat_Geor	71.54	pbt_Arab	76.65	vec_Latn	77.03
ces_Latn	78.76	kaz_Cyrl	76.65	pes_Arab	80.98	vie_Latn	77.09
cjk_Latn	58.5	kbp_Latn	47.43	plt_Latn	77.98	war_Latn	74.45
ckb_Arab	73.79	kea_Latn	61.84	pol_Latn	78.28	wol_Latn	56.95
crh_Latn	71.24	khk_Cyrl	76.05	por_Latn	78.32	xho_Latn	69.42
cym_Latn	73.57	khm_Khmr	74.59	prs_Arab	81.06	ydd_Hebr	72.32
dan_Latn	77.07	kik_Latn	50.08	quy_Latn	54.55	yor_Latn	64.73
deu_Latn	78.42	kin_Latn	74.01	ron_Latn	73.77	yue_Hant	77.62
dik_Latn	53.43	kir_Cyrl	75.69	run_Latn	74.03	zho_Hans	76.67
dyu_Latn	63.41	kmb_Latn	57.31	rus_Cyrl	79.12	zho_Hant	72.93
ell_Grek	77.96	kmr_Latn	77.76	sag_Latn	70.78	zsm_Latn	77.29
eng_Latn	76.41	knc_Arab	69.26	san_Deva	70.44	zul_Latn	74.71
epo_Latn	76.39	knc_Latn	59.92	scn_Latn	79.42		
est_Latn	81.0	kon_Latn	62.44	shn_Mymr	67.35		
eus_Latn	76.11	kor_Hang	77.09	sin_Sinh	76.25		
ewe_Latn	58.66	lao_Laoo	73.67	slk_Latn	77.35		

Figure 11: The accuracy score of AYA on B-NLI task across 198 languages.

H Full Results for XLM-R base

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	37.29	fao_Latn	52.85	lij_Latn	56.95	slv_Latn	75.05
ace_Latn	49.72	fij_Latn	37.37	lim_Latn	58.78	smo_Latn	37.35
acm_Arab	66.37	fin_Latn	75.65	lin_Latn	37.84	sna_Latn	38.74
acq_Arab	68.5	fon_Latn	37.5	lit_Latn	75.31	snd_Arab	71.34
aeb_Arab	56.19	fra_Latn	78.54	lmo_Latn	57.58	som_Latn	61.86
afr_Latn	77.35	fur_Latn	56.33	ltg_Latn	53.75	sot_Latn	38.36
ajp_Arab	65.61	fuv_Latn	39.34	ltz_Latn	46.31	spa_Latn	79.92
aka_Latn	38.78	gaz_Latn	45.81	lua_Latn	37.07	srd_Latn	58.2
als_Latn	76.61	gla_Latn	62.53	lug_Latn	39.82	srp_Cyrl	77.96
amh_Ethi	64.83	gle_Latn	67.45	luo_Latn	37.66	ssw_Latn	39.3
apc_Arab	61.9	glg_Latn	79.5	lus_Latn	40.1	sun_Latn	64.45
arb_Arab	73.81	grn_Latn	40.62	lvs_Latn	74.11	swe_Latn	79.08
ars_Arab	68.6	guj_Gujr	71.96	mag_Deva	63.11	swh_Latn	67.96
ary_Arab	59.76	hat_Latn	41.56	mai_Deva	57.94	szl_Latn	59.16
arz_Arab	67.66	hau_Latn	61.72	mal_Mlym	72.97	tam_Taml	72.61
asm_Beng	64.89	heb_Hebr	76.55	mar_Deva	72.83	taq_Latn	38.46
ast_Latn	62.1	hin_Deva	75.57	min_Latn	51.36	tat_Cyrl	43.59
awa_Deva	58.0	hne_Deva	59.96	mkd_Cyrl	77.19	tel_Telu	71.82
ayr_Latn	38.8	hrv_Latn	77.01	mlt_Latn	39.66	tgk_Cyrl	37.76
azb_Arab	42.02	hun_Latn	76.75	mni_Beng	35.87	tgl_Latn	71.96
azj_Latn	74.09	hye_Armn	74.23	mos_Latn	37.01	tha_Thai	73.67
bak_Cyrl	43.55	ibo_Latn	38.22	mri_Latn	38.0	tir_Ethi	42.5
bam_Latn	41.24	ilo_Latn	39.24	mya_Mymr	66.05	tpi_Latn	44.81
ban_Latn	44.39	ind_Latn	79.22	nld_Latn	77.62	tsn_Latn	39.22
bel_Cyrl	75.45	isl_Latn	70.78	nno_Latn	70.86	tso_Latn	39.44
bem_Latn	38.06	ita_Latn	78.7	nob_Latn	80.38	tuk_Latn	47.78
ben_Beng	71.84	jav_Latn	67.6	npi_Deva	71.74	tum_Latn	37.21
bho_Deva	66.15	jpn_Jpan	69.46	nso_Latn	38.58	tur_Latn	75.43
bjn_Arab	36.83	kab_Latn	35.73	nus_Latn	38.12	twi_Latn	38.24
bjn_Latn	58.18	kac_Latn	38.26	nya_Latn	39.4	uig_Arab	60.24
bos_Latn	77.35	kam_Latn	34.35	oci_Latn	64.73	ukr_Cyrl	77.74
bug_Latn	41.9	kan_Knda	72.95	pag_Latn	39.66	umb_Latn	34.69
bul_Cyrl	79.02	kas_Arab	44.43	pan_Guru	71.92	urd_Arab	73.63
cat_Latn	78.34	kas_Deva	45.61	pap_Latn	58.26	uzn_Latn	70.42
ceb_Latn	52.18	kat_Geor	54.31	pbt_Arab	71.32	vec_Latn	70.06
ces_Latn	78.34	kaz_Cyrl	70.52	pes_Arab	74.57	vie_Latn	77.11
cjk_Latn	39.02	kbp_Latn	37.11	plt_Latn	61.18	war_Latn	43.19
ckb_Arab	41.48	kea_Latn	48.32	pol_Latn	76.67	wol_Latn	39.44
crh_Latn	58.58	khk_Cyrl	70.28	por_Latn	80.02	xho_Latn	46.83
cym_Latn	71.66	khm_Khmr	69.4	prs_Arab	76.47	ydd_Hebr	67.9
dan_Latn	79.36	kik_Latn	38.54	quy_Latn	37.94	yor_Latn	36.67
deu_Latn	77.39	kin_Latn	37.8	ron_Latn	78.64	yue_Hant	66.21
dik_Latn	39.8	kir_Cyrl	68.86	run_Latn	38.2	zho_Hans	69.64
dyu_Latn	36.63	kmb_Latn	36.15	rus_Cyrl	78.74	zho_Hant	58.74
ell_Grek	77.58	kmr_Latn	64.67	sag_Latn	38.46	zsm_Latn	78.08
eng_Latn	84.09	knc_Arab	45.87	san_Deva	62.79	zul_Latn	45.89
epo_Latn	75.07	knc_Latn	38.62	scn_Latn	54.29		
est_Latn	74.65	kon_Latn	37.49	shn_Mymr	35.15		
eus_Latn	63.75	kor_Hang	73.47	sin_Sinh	71.78		
ewe_Latn	38.66	lao_Laoo	73.15	slk_Latn	78.16		

Figure 12: The accuracy score of XLM-R base on XNLI task across 196 languages.

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	51.62	fao_Latn	58.7	lij_Latn	57.97	som_Latn	56.19
ace_Latn	56.45	fij_Latn	55.06	lim_Latn	62.21	sot_Latn	53.34
acm_Arab	61.68	fin_Latn	73.66	lin_Latn	54.4	spa_Latn	74.39
acq_Arab	67.44	fon_Latn	51.09	lit_Latn	70.62	srd_Latn	61.02
aeb_Arab	58.5	fra_Latn	72.27	lmo_Latn	60.42	srp_Cyrl	70.81
afr_Latn	72.14	fur_Latn	58.44	ltg_Latn	54.73	ssw_Latn	54.6
ajp_Arab	67.57	fuv_Latn	52.35	ltz_Latn	53.94	sun_Latn	62.14
aka_Latn	53.81	gaz_Latn	56.59	lua_Latn	49.9	swe_Latn	75.18
als_Latn	68.03	gla_Latn	59.83	lug_Latn	52.02	swh_Latn	63.73
amh_Ethi	60.56	gle_Latn	62.01	luo_Latn	56.39	szl_Latn	62.48
apc_Arab	65.12	glg_Latn	73.26	lus_Latn	54.73	tam_Taml	63.53
arb_Arab	70.62	grn_Latn	56.06	lvs_Latn	68.03	taq_Latn	51.42
ars_Arab	69.09	guj_Gujr	64.26	mag_Deva	62.67	tat_Cyrl	53.14
ary_Arab	63.34	hat_Latn	55.06	mai_Deva	59.03	tel_Telu	64.99
arz_Arab	65.06	hau_Latn	57.11	mal_Mlym	67.5	tgk_Cyrl	55.06
asm_Beng	58.24	heb_Hebr	71.01	min_Latn	57.18	tgl_Latn	64.13
ast_Latn	66.38	hin_Deva	69.16	mkd_Cyrl	71.94	tha_Thai	75.31
awa_Deva	58.9	hne_Deva	60.75	mlt_Latn	51.42	tir_Ethi	51.03
ayr_Latn	53.14	hrv_Latn	73.06	mni_Beng	50.23	tpi_Latn	57.97
azb_Arab	52.08	hun_Latn	73.26	mos_Latn	52.95	tsn_Latn	50.83
azj_Latn	69.82	hye_Armn	68.3	mri_Latn	50.96	tso_Latn	51.62
bak_Cyrl	55.72	ibo_Latn	49.97	nld_Latn	74.45	tuk_Latn	53.94
bam_Latn	52.95	ilo_Latn	54.8	nno_Latn	70.48	tum_Latn	52.42
ban_Latn	55.46	ind_Latn	76.77	nob_Latn	76.04	tur_Latn	73.46
bel_Cyrl	69.16	isl_Latn	66.84	nso_Latn	53.94	twi_Latn	52.88
bem_Latn	51.75	ita_Latn	73.79	nus_Latn	53.01	uig_Arab	60.23
ben_Beng	64.0	jav_Latn	62.41	nya_Latn	53.54	ukr_Cyrl	74.06
bho_Deva	63.86	jpn_Jpan	74.78	oci_Latn	63.67	umb_Latn	52.08
bjn_Arab	51.42	kab_Latn	52.02	pag_Latn	52.95	urd_Arab	66.78
bjn_Latn	59.1	kac_Latn	53.28	pap_Latn	59.23	uzn_Latn	63.27
bos_Latn	69.36	kam_Latn	50.83	pbt_Arab	64.39	vec_Latn	63.93
bug_Latn	53.21	kan_Knda	64.92	pes_Arab	69.76	vie_Latn	73.99
bul_Cyrl	74.59	kas_Arab	52.88	plt_Latn	55.72	war_Latn	54.93
cat_Latn	72.27	kas_Deva	58.44	pol_Latn	72.53	wol_Latn	51.49
ceb_Latn	57.58	kat_Geor	56.25	por_Latn	73.92	xho_Latn	54.07
ces_Latn	73.06	kaz_Cyrl	68.63	prs_Arab	70.88	ydd_Hebr	58.24
cjk_Latn	51.75	kbp_Latn	49.37	quy_Latn	50.36	yor_Latn	51.09
ckb_Arab	51.42	kea_Latn	56.98	ron_Latn	72.14	yue_Hant	71.21
crh_Latn	59.43	khk_Cyrl	69.69	run_Latn	47.85	zho_Hans	76.9
cym_Latn	61.35	khm_Khmr	65.06	rus_Cyrl	74.72	zho_Hant	72.47
dan_Latn	75.78	kik_Latn	53.28	sag_Latn	50.56	zsm_Latn	75.45
deu_Latn	72.87	kin_Latn	52.95	san_Deva	63.34	zul_Latn	55.53
dik_Latn	54.2	kir_Cyrl	65.98	scn_Latn	60.82		
dyu_Latn	52.68	kmb_Latn	52.35	shn_Mymr	46.33		
ell_Grek	67.44	kmr_Latn	61.48	sin_Sinh	67.04		
eng_Latn	80.34	knc_Arab	55.99	slk_Latn	73.4		
epo_Latn	69.76	knc_Latn	52.88	slv_Latn	71.34		
est_Latn	71.61	kon_Latn	53.94	smo_Latn	51.69		
eus_Latn	67.44	kor_Hang	69.89	sna_Latn	54.33		
ewe_Latn	51.95	lao_Laoo	64.53	snd_Arab	63.0		

Figure 13: The accuracy score of XLM-R base on XStoryCloze task across 196 languages.

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	67.3	fao_Latn	84.8	lij_Latn	82.95	slv_Latn	88.3
ace_Latn	82.6	fij_Latn	71.5	lim_Latn	85.5	smo_Latn	73.7
acm_Arab	85.9	fin_Latn	84.15	lin_Latn	74.8	sna_Latn	75.25
acq_Arab	85.9	fon_Latn	66.35	lit_Latn	87.45	snd_Arab	84.3
aeb_Arab	84.2	fra_Latn	89.9	lmo_Latn	84.9	som_Latn	83.0
afr_Latn	91.05	fur_Latn	83.6	ltg_Latn	80.8	sot_Latn	76.3
ajp_Arab	85.1	fuv_Latn	70.8	ltz_Latn	82.3	spa_Latn	89.4
aka_Latn	70.2	gaz_Latn	74.0	lua_Latn	71.55	srd_Latn	85.2
als_Latn	90.05	gla_Latn	84.6	lug_Latn	72.6	srp_Cyrl	89.65
amh_Ethi	81.5	gle_Latn	85.1	luo_Latn	71.3	ssw_Latn	74.15
apc_Arab	84.85	glg_Latn	89.8	lus_Latn	75.1	sun_Latn	87.35
arb_Arab	86.25	grn_Latn	75.0	lvs_Latn	86.2	swe_Latn	90.35
ars_Arab	85.25	guj_Gujr	83.95	mag_Deva	84.35	swh_Latn	86.95
ary_Arab	84.35	hat_Latn	84.4	mai_Deva	80.65	szl_Latn	85.25
arz_Arab	84.85	hau_Latn	83.8	mal_Mlym	78.75	tam_Taml	81.75
asm_Beng	80.35	heb_Hebr	86.95	mar_Deva	82.1	taq_Latn	70.2
ast_Latn	85.85	hin_Deva	86.1	min_Latn	85.5	tat_Cyrl	77.65
awa_Deva	78.9	hne_Deva	82.6	mkd_Cyrl	89.85	tel_Telu	83.65
ayr_Latn	68.95	hrv_Latn	89.5	mlt_Latn	79.9	tgk_Cyrl	75.5
azb_Arab	69.35	hun_Latn	87.3	mni_Beng	71.0	tgl_Latn	89.8
azj_Latn	83.8	hye_Armn	86.05	mos_Latn	66.6	tha_Thai	86.25
bak_Cyrl	75.8	ibo_Latn	72.55	mri_Latn	75.05	tir_Ethi	76.15
bam_Latn	72.8	ilo_Latn	78.9	mya_Mymr	78.85	tpi_Latn	73.7
ban_Latn	81.0	ind_Latn	91.0	nld_Latn	89.55	tsn_Latn	76.35
bel_Cyrl	87.55	isl_Latn	86.6	nno_Latn	84.2	tso_Latn	72.55
bem_Latn	70.75	ita_Latn	90.45	nob_Latn	91.05	tuk_Latn	71.65
ben_Beng	84.85	jav_Latn	87.6	npi_Deva	84.9	tum_Latn	70.75
bho_Deva	80.55	jpn_Jpan	81.35	nso_Latn	76.85	tur_Latn	83.35
bjn_Arab	66.65	kab_Latn	72.4	nus_Latn	67.3	twi_Latn	70.9
bjn_Latn	86.6	kac_Latn	71.25	nya_Latn	75.85	uig_Arab	77.95
bos_Latn	89.5	kam_Latn	62.85	oci_Latn	88.4	ukr_Cyrl	89.3
bug_Latn	78.2	kan_Knda	83.85	pag_Latn	78.45	umb_Latn	63.6
bul_Cyrl	89.95	kas_Arab	71.1	pan_Guru	84.95	urd_Arab	85.15
cat_Latn	90.65	kas_Deva	70.55	pap_Latn	84.95	uzn_Latn	82.35
ceb_Latn	84.85	kat_Geor	78.25	pbt_Arab	83.2	vec_Latn	88.05
ces_Latn	89.7	kaz_Cyrl	82.7	pes_Arab	85.55	vie_Latn	89.85
cjk_Latn	69.35	kbp_Latn	65.65	plt_Latn	82.4	war_Latn	80.7
ckb_Arab	72.35	kea_Latn	82.7	pol_Latn	88.2	wol_Latn	71.95
crh_Latn	81.8	khk_Cyrl	80.4	por_Latn	89.8	xho_Latn	81.1
cym_Latn	88.0	khm_Khmr	86.45	prs_Arab	86.9	ydd_Hebr	85.3
dan_Latn	90.0	kik_Latn	69.45	quy_Latn	67.55	yor_Latn	69.45
deu_Latn	89.55	kin_Latn	71.65	ron_Latn	89.95	yue_Hant	80.0
dik_Latn	69.3	kir_Cyrl	79.25	run_Latn	72.0	zho_Hans	79.65
dyu_Latn	66.7	kmb_Latn	65.0	rus_Cyrl	89.3	zho_Hant	69.5
ell_Grek	88.8	kmr_Latn	84.2	sag_Latn	70.65	zsm_Latn	91.25
eng_Latn	93.15	knc_Arab	72.8	san_Deva	73.65	zul_Latn	82.3
epo_Latn	89.5	knc_Latn	66.9	scn_Latn	84.8		
est_Latn	86.15	kon_Latn	71.25	shn_Mymr	66.6		
eus_Latn	78.35	kor_Hang	82.15	sin_Sinh	82.95		
ewe_Latn	71.45	lao_Laoo	86.6	slk_Latn	89.9		

Figure 14: The accuracy score of XLM-R base on PAWS-X task across 196 languages.

I ChrF++ scores per language

Language	Chrf++ score	Language	Chrf++ score	Language	Chrf++ score	Language	Chrf++ score
ace_Arab	18.38	fij_Latn	45.96	lim_Latn	44.76	smo_Latn	50.43
ace_Latn	42.42	fin_Latn	51.3	lin_Latn	50.46	sna_Latn	43.66
acm_Arab	39.32	fon_Latn	20.6	lit_Latn	51.6	snd_Arab	49.86
acq_Arab	42.23	fra_Latn	67.2	lmo_Latn	34.74	som_Latn	46.02
aeb_Arab	36.3	fur_Latn	55.68	ltg_Latn	44.87	sot_Latn	44.44
afr_Latn	65.22	fuv_Latn	24.62	ltz_Latn	54.1	spa_Latn	54.7
ajp_Arab	44.92	gaz_Latn	36.33	lua_Latn	36.78	srd_Latn	54.78
aka_Latn	36.34	gla_Latn	49.15	lug_Latn	39.41	srp_Cyrl	55.24
als_Latn	56.05	gle_Latn	55.47	luo_Latn	41.13	ssw_Latn	43.68
amh_Ethi	30.23	glg_Latn	58.52	lus_Latn	40.26	sun_Latn	48.42
apc_Arab	43.91	grn_Latn	34.0	lvs_Latn	46.98	swe_Latn	65.14
arb_Arab	52.77	guj_Gujr	48.15	mag_Deva	59.0	swh_Latn	62.33
ars_Arab	47.24	hat_Latn	52.06	mai_Deva	44.7	szl_Latn	49.19
ary_Arab	36.35	hau_Latn	53.77	mal_Mlym	43.29	tam_Taml	52.25
arz_Arab	43.39	heb_Hebr	52.48	mar_Deva	44.48	taq_Latn	27.69
asm_Beng	35.68	hin_Deva	56.72	min_Latn	55.33	tat_Cyrl	45.93
ast_Latn	53.02	hne_Deva	51.64	mkd_Cyrl	56.51	tel_Telu	51.88
awa_Deva	39.64	hrv_Latn	54.07	mlt_Latn	63.73	tgk_Cyrl	46.24
ayr_Latn	31.75	hun_Latn	54.43	mni_Beng	38.16	tgl_Latn	59.93
azb_Arab	23.53	hye_Armn	51.02	mos_Latn	27.92	tha_Thai	39.54
azj_Latn	42.08	ibo_Latn	41.94	mri_Latn	46.87	tir_Ethi	24.48
bak_Cyrl	46.46	ilo_Latn	54.64	mya_Mymr	32.35	tpi_Latn	44.67
bam_Latn	31.33	ind_Latn	65.11	nld_Latn	57.1	tsn_Latn	47.51
ban_Latn	43.27	isl_Latn	47.98	nno_Latn	49.49	tso_Latn	54.34
bel_Cyrl	40.34	ita_Latn	56.32	nob_Latn	58.69	tuk_Latn	36.45
bem_Latn	37.32	jav_Latn	53.07	npi_Deva	45.63	tum_Latn	38.62
ben_Beng	45.48	jpn_Jpan	24.78	nso_Latn	51.5	tur_Latn	57.18
bho_Deva	38.01	kab_Latn	31.66	nus_Latn	26.86	twi_Latn	40.66
bjn_Arab	21.12	kac_Latn	38.81	nya_Latn	46.42	uig_Arab	36.37
bjn_Latn	46.01	kam_Latn	24.22	oci_Latn	59.63	ukr_Cyrl	52.2
bos_Latn	56.14	kan_Knda	49.11	pag_Latn	49.66	umb_Latn	25.47
bug_Latn	35.41	kas_Arab	32.66	pan_Guru	49.79	urd_Arab	50.53
bul_Cyrl	61.75	kas_Deva	20.7	pap_Latn	55.38	uzn_Latn	47.26
cat_Latn	64.34	kat_Geor	42.11	pbt_Arab	34.14	vec_Latn	47.31
ceb_Latn	57.88	kaz_Cyrl	44.98	pes_Arab	44.16	vie_Latn	60.12
ces_Latn	54.49	kbp_Latn	33.21	plt_Latn	51.81	war_Latn	57.14
cjk_Latn	26.38	kea_Latn	49.26	pol_Latn	49.25	wol_Latn	30.78
ckb_Arab	44.65	khk_Cyrl	38.78	por_Latn	67.53	xho_Latn	43.99
crh_Latn	41.15	khm_Khmr	35.58	prs_Arab	44.15	ydd_Hebr	35.95
cym_Latn	65.55	kik_Latn	36.27	quy_Latn	28.58	yor_Latn	29.3
dan_Latn	64.09	kin_Latn	51.13	ron_Latn	60.1	yue_Hant	16.82
deu_Latn	61.69	kir_Cyrl	41.22	run_Latn	45.04	zho_Hans	17.0
dik_Latn	22.68	kmb_Latn	26.6	rus_Cyrl	53.87	zho_Hant	11.99
dyu_Latn	17.86	kmr_Latn	40.57	sag_Latn	34.33	zsm_Latn	65.86
ell_Grek	53.36	knc_Arab	11.39	san_Deva	22.74	zul_Latn	51.69
epo_Latn	60.17	knc_Latn	23.95	scn_Latn	43.16		
est_Latn	54.95	kon_Latn	44.35	shn_Mymr	36.29		
eus_Latn	43.48	kor_Hang	33.04	sin_Sinh	40.13		
ewe_Latn	38.58	lao_Laoo	49.11	slk_Latn	55.71		
fao_Latn	49.94	lij_Latn	46.65	slv_Latn	53.92		

Figure 15: Chrf++ scores for the selected languages.

J Language coverage

	Category	Languages
XLM-R	High	arb_Arab , bul_Cyrl , dan_Latn , deu_Latn , ell_Grek , eng_Latn , fin_Latn , fra_Latn , heb_Hebr , hun_Latn , ind_Latn , ita_Latn , jpn_Jpan , kor_Hang , nld_Latn , nob_Latn , pes_Arab , pol_Latn , por_Latn , ron_Latn , rus_Cyrl , spa_Latn , tha_Thai , ukr_Cyrl , vie_Latn , zho_Hans
	Mid	als_Latn , azj_Latn , bel_Cyrl , ben_Beng , cat_Latn , ces_Latn , est_Latn , glg_Latn , hin_Deva , hrv_Latn , hye_Armn , isl_Latn , kan_Knda , kat_Geor , kaz_Cyrl , khk_Cyrl , lit_Latn , lvs_Latn , mal_Mlym , mar_Deva , mkd_Cyrl , npi_Deva , sin_Sinh , slk_Latn , slv_Latn , srp_Cyrl , swe_Latn , tam_Taml , tel_Telu , tgl_Latn , tur_Latn , urd_Arab , zho_Hant , zsm_Latn
	Low	afr_Latn , amh_Ethi , asm_Beng , bos_Latn , ckb_Arab , cym_Latn , epo_Latn , eus_Latn , gaz_Latn , gla_Latn , gle_Latn , guj_Gujr , hau_Latn , jav_Latn , khm_Khmr , kir_Cyrl , lao_Laoo , mya_Mymr , pan_Guru , pbt_Arab , plt_Latn , san_Deva , snd_Arab , som_Latn , sun_Latn , swh_Latn , uig_Arab , uzn_Latn , xho_Latn , ydd_Hebr
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , ajp_Arab , aka_Latn , apc_Arab , ars_Arab , ary_Arab , arz_Arab , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , bak_Cyrl , bam_Latn , ban_Latn , bem_Latn , bho_Deva , bjn_Arab , bjn_Latn , bug_Latn , ceb_Latn , cjk_Latn , crh_Latn , dik_Latn , dyu_Latn , ewe_Latn , fao_Latn , fij_Latn , fon_Latn , fur_Latn , fuv_Latn , grn_Latn , hat_Latn , hne_Deva , ibo_Latn , ilo_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , kbp_Latn , kea_Latn , kik_Latn , kin_Latn , kmb_Latn , kmr_Latn , knc_Arab , knc_Latn , kon_Latn , lij_Latn , lim_Latn , lin_Latn , lmo_Latn , ltg_Latn , ltz_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , mag_Deva , mai_Deva , min_Latn , mlt_Latn , mni_Beng , mos_Latn , mri_Latn , nno_Latn , nso_Latn , nus_Latn , nya_Latn , oci_Latn , pag_Latn , pap_Latn , prs_Arab , quy_Latn , run_Latn , sag_Latn , scn_Latn , shn_Mymr , smo_Latn , sna_Latn , sot_Latn , srd_Latn , ssw_Latn , szl_Latn , taq_Latn , tat_Cyrl , tgk_Cyrl , tir_Ethi , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , twi_Latn , umb_Latn , vec_Latn , war_Latn , wol_Latn , yor_Latn , yue_Hant , zul_Latn
BLOOMz	High	arb_Arab , cat_Latn , eng_Latn , fra_Latn , ind_Latn , por_Latn , spa_Latn , vie_Latn , zho_Hans
	Mid	ben_Beng , eus_Latn , hin_Deva , mal_Mlym , tam_Taml , urd_Arab , zho_Hant’
	Low	aka_Latn , asm_Beng , bam_Latn , bho_Deva , fon_Latn , guj_Gujr , ibo_Latn , kan_Knda , kik_Latn , kin_Latn , lin_Latn , mar_Deva , npi_Deva , nso_Latn , sot_Latn , swh_Latn , tel_Telu , wol_Latn , xho_Latn , yor_Latn , zul_Latn
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , afr_Latn , ajp_Arab , apc_Arab , ars_Arab , ary_Arab , arz_Arab , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , azj_Latn , ban_Latn , bem_Latn , bjn_Arab , bjn_Latn , bos_Latn , bug_Latn , ceb_Latn , ces_Latn , cjk_Latn , ckb_Arab , crh_Latn , cym_Latn , dan_Latn , deu_Latn , dik_Latn , dyu_Latn , epo_Latn , est_Latn , ewe_Latn , fao_Latn , fij_Latn , fin_Latn , fur_Latn , fuv_Latn , gla_Latn , gle_Latn , glg_Latn , grn_Latn , hat_Latn , hau_Latn , hne_Deva , hrv_Latn , hun_Latn , ilo_Latn , isl_Latn , ita_Latn , jav_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , knc_Arab , knc_Latn , kbp_Latn , kea_Latn , kmb_Latn , kmr_Latn , kon_Latn , lij_Latn , lim_Latn , lit_Latn , lmo_Latn , ltg_Latn , ltz_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , lvs_Latn , mag_Deva , mai_Deva , min_Latn , plt_Latn , mlt_Latn , mni_Beng , mos_Latn , mri_Latn , nld_Latn , nno_Latn , nob_Latn , nus_Latn , nya_Latn , oci_Latn , gaz_Latn , pag_Latn , pap_Latn , pes_Arab , pol_Latn , prs_Arab , pbt_Arab , quy_Latn , ron_Latn , run_Latn , sag_Latn , san_Deva , scn_Latn , slk_Latn , slv_Latn , smo_Latn , sna_Latn , snd_Arab , som_Latn , als_Latn , srd_Latn , ssw_Latn , sun_Latn , swe_Latn , szl_Latn , tgl_Latn , taq_Latn , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , tur_Latn , twi_Latn , uig_Arab , umb_Latn , uzn_Latn , vec_Latn , war_Latn , yue_Hant , zsm_Latn

Table 13: The languages covered during pretraining of each of the MLMs categorized by the amount of data that was seen for them during pretraining.

	Category	Languages
AYA	High	hye_Armn , kan_Knda , tur_Latn , ita_Latn , nld_Latn , pol_Latn , por_Latn , isl_Latn , fra_Latn , deu_Latn , spa_Latn , rus_Cyrl , eng_Latn
	Mid	est_Latn , ben_Beng , mar_Deva , slv_Latn , lit_Latn , heb_Hebr , zsm_Latn , cat_Latn , tha_Thai , kor_Hang , slk_Latn , hin_Deva , bul_Cyrl , nob_Latn , fin_Latn , dan_Latn , hun_Latn , ukr_Cyrl , ell_Grek , ron_Latn , swe_Latn , arb_Arab , pes_Arab , zho_Hans , ces_Latn
	Low	hat_Latn , kor_Hang , xho_Latn , ibo_Latn , lao_Laoo , mri_Latn , smo_Latn , ckb_Arab , amh_Ethi , nya_Latn , hau_Latn , plt_Latn , pbt_Arab , gla_Latn , sun_Latn , jpn_Jpan , sot_Latn , ceb_Latn , pan_Guru , gle_Latn , kir_Cyrl , epo_Latn , sin_Sinh , guj_Gujr , yor_Latn , tgk_Cyrl , snd_Arab , mya_Mymr , kaz_Cyrl , khm_Khmr , som_Latn , swh_Latn , ydd_Hebr , uzn_Latn , hun_Latn , mlt_Latn , eus_Latn , bel_Cyrl , kat_Geor , mkd_Cyrl , mal_Mlym , khk_Cyrl , tha_Thai , afr_Latn , ukr_Cyrl , ltz_Latn , tel_Telu , urd_Arab , lit_Latn , npi_Deva , srp_Cyrl , tam_Taml , cym_Latn , als_Latn , glg_Latn , azj_Latn , lvs_Latn
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , ajp_Arab , aka_Latn , apc_Arab , ars_Arab , ary_Arab , arz_Arab , asm_Beng , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , bak_Cyrl , bam_Latn , ban_Latn , bem_Latn , bho_Deva , bjn_Arab , bjn_Latn , bos_Latn , bug_Latn , cjk_Latn , crh_Latn , dik_Latn , dyu_Latn , ewe_Latn , fao_Latn , fij_Latn , fon_Latn , fur_Latn , fuv_Latn , grn_Latn , hne_Deva , hrv_Latn , ilo_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , knc_Arab , knc_Latn , kbp_Latn , kea_Latn , kik_Latn , kin_Latn , kmb_Latn , kmr_Latn , kon_Latn , lij_Latn , lim_Latn , lin_Latn , lmo_Latn , ltg_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , mag_Deva , mai_Deva , min_Latn , mni_Beng , mos_Latn , nno_Latn , nso_Latn , nus_Latn , oci_Latn , gaz_Latn , pag_Latn , pap_Latn , prs_Arab , quy_Latn , run_Latn , sag_Latn , san_Deva , scn_Latn , shn_Mymr , srd_Latn , ssw_Latn , szl_Latn , tat_Cyrl , tgl_Latn , tir_Ethi , taq_Latn , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , twi_Latn , tzm_Tfng , uig_Arab , umb_Latn , vec_Latn , war_Latn , wol_Latn , yue_Hant , zho_Hant , dzo_Tibt

Table 14: The languages covered during AYA's pretraining categorized by the amount of data that was seen during pretraining.