

An Empirical Analysis of Machine Translation for Expanding Multilingual Benchmarks

Sara Rajae^{*◇} 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.

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

	Unseen	Low	Mid	High	Total
% of data	0	>0 and <0.1	≥ 0.1 and <1	≥ 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.

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.

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 numbers¹, 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 ($En \Rightarrow \text{tgt}$ setting). NLLB also demonstrates a minimal difference to the closed source system, Google translate² 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 perfor-

mance 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.

³Results 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-

⁴error categorizations.

⁵Please 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

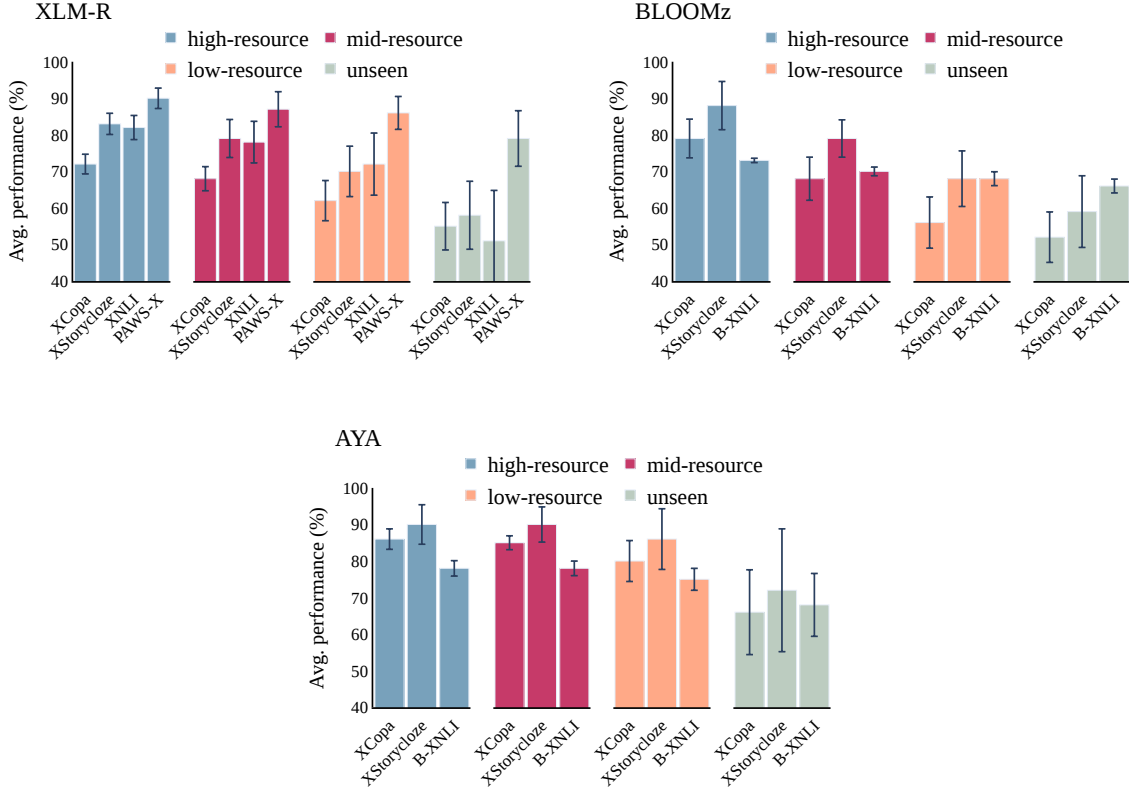


Figure 1: Average performance across test languages in a zero-shot fine-tuning setup for XLM-R and in a zero-shot prompting using BLOOMz. Results are categorized per task and data coverage during pretraining as reported in Table 1. Results across models are not directly comparable as their language categorizations 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 Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine 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, Isabella 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, Johannes 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 Kopic, 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 Nee-lakantan, 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 Pokorný, 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, Qiming 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 Maudiff 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, Rutu 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 Kennealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Naveen 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 Pettrini, 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, Naveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivan, 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 dataset for 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 Hefernan, 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, Elie 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 Limkonchotiawat, 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, Prajwal 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, přičemž 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	swl 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	uiq 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
ckj 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	sno 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 Cyril	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	gui Gujr	77.52	mag Deva	70.74	swl 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 Cyril	48.24
awa Deva	62.28	hne Deva	69.6	mkd Cyril	82.02	tel Telu	75.65
ayr Latn	40.36	hrv Latn	81.86	mlt Latn	43.27	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	81.64
bug Latn	43.69	kan Knda	77.47	pag Latn	41.82	umb Latn	37.09
bul Cyril	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 Cyril	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 Cyril	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 Cyril	73.99	run Latn	40.42	zho Hans	72.42
dyu Latn	36.59	kmb Latn	37.58	rus Cyril	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	sno 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 Cyril	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	gui Gujr	73.46	mag Deva	67.97	swb 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 Cyril	56.85
awa Deva	64.2	hne Deva	66.91	mkd Cyril	82.06	tel Telu	76.31
ayr Latn	51.29	hrv Latn	83.52	mlt Latn	50.43	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	82.26
bug Latn	53.34	kan Knda	74.72	pag Latn	53.28	umb Latn	51.75
bul Cyril	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 Cyril	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 Cyril	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 Cyril	75.78	run Latn	47.92	zho Hans	86.9
dyu Latn	50.43	kmb Latn	50.63	rus Cyril	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	sno 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 Cyril	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	gui Gujr	64.6	mag Deva	60.2	swb 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 Cyril	47.0
awa Deva	64.2	hne Deva	62.2	mkd Cyril	69.6	tel Telu	65.8
ayr Latn	48.2	hrv Latn	71.0	mlt Latn	51.4	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	68.4
bug Latn	51.2	kan Knda	67.6	pag Latn	51.6	umb Latn	49.2
bul Cyril	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 Cyril	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 Cyril	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 Cyril	65.4	run Latn	52.4	zho Hans	70.6
dyu Latn	52.0	kmb Latn	49.6	rus Cyril	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	gui 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		
azi 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		
ckj 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	swl 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	gui 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	swl 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	gui 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	swl 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	swl 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
ckj 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	sno 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 Cyril	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	gui Gujr	84.2	mag Deva	78.2	swb 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 Cyril	81.8
awa Deva	68.6	hne Deva	81.6	mkd Cyril	85.4	tel Telu	81.8
ayr Latn	51.4	hrv Latn	80.4	mlt Latn	74.0	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	86.6
bug Latn	55.4	kan Knda	82.6	pag Latn	57.0	umb Latn	51.4
bul Cyril	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 Cyril	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 Cyril	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 Cyril	82.8	run Latn	68.4	zho Hans	83.6
dyu Latn	52.6	kmb Latn	53.2	rus Cyril	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	sno 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 Cyril	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	gui Gujr	73.77	mag Deva	69.5	swb 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 Cyril	76.59
awa Deva	68.46	hne Deva	70.76	mkd Cyril	75.19	tel Telu	77.74
ayr Latn	53.23	hrv Latn	75.29	mlt Latn	71.64	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	80.32
bug Latn	59.42	kan Knda	75.51	pag Latn	63.61	umb Latn	55.87
bul Cyril	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 Cyril	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 Cyril	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 Cyril	75.69	run Latn	74.03	zho Hans	76.67
dyu Latn	63.41	kmb Latn	57.31	rus Cyril	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	swl 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
azi 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
ckj 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 Cyr	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	swl 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	gui Guj	64.26	mag Deva	62.67	tat Cyr	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 Cyr	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 Cyr	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 Cyr	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 Cyr	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 Cyr	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 Cyr	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 Cyr	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 Cyr	69.69	run Latn	47.85	zho Hans	76.9
cym Latn	61.35	khm Khmr	65.06	rus Cyr	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 Cyr	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	sno 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	sno 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 Cyril	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	gui Gujr	83.95	mag Deva	84.35	swb 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 Tamil	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 Cyril	77.65
awa Deva	78.9	hne Deva	82.6	mkd Cyril	89.85	tel Telu	83.65
ayr Latn	68.95	hrv Latn	89.5	mlt Latn	79.9	tgk Cyril	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 Cyril	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 Cyril	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 Cyril	89.3
bug Latn	78.2	kan Knda	83.85	pag Latn	78.45	umb Latn	63.6
bul Cyril	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 Cyril	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 Cyril	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 Cyril	79.25	run Latn	72.0	zho Hans	79.65
dyu Latn	66.7	kmb Latn	65.0	rus Cyril	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	gui Gujr	48.15	mag Deva	59.0	swl 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
ckj 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_Lao , mya_Mymr , pan_Guru , pbt_Arab , plt_Latn , san_Deva , snd_Arab , som_Latn , sun_Latn , swl_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 , swl_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 , hin_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_Lao , 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 , swl_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.