A Recipe of Parallel Corpora Exploitation for Multilingual Large Language Models

Anonymous ACL submission

Abstract

Recent studies have highlighted the potential of exploiting parallel corpora to enhance multilingual large language models in both bilingual tasks, e.g., machine translation, and general-purpose tasks, e.g., text classification. Building upon these findings, our comprehensive study aims to identify the most effective strategies for leveraging parallel corpora. We investigate the impact of parallel corpus quality and quantity, training objectives, and model size on the performance of multilingual large language models enhanced with parallel corpora across diverse languages and tasks. Our analysis reveals several key insights: (i) filtering noisy translations is essential for exploiting parallel corpora, while language identification and short sentence filtering have little effect; (ii) even a corpus with just 10K parallel sentences can yield results comparable to those obtained from larger datasets; (iii) employing only the machine translation objective yields the best results among various training objectives and their combinations; (iv) larger multilingual language models benefit more from parallel corpora than smaller models Our study offers valuable insights into the optimal utilization of parallel corpora to enhance multilingual large language models, extending the generalizability of previous findings from limited languages and tasks to a broader range of scenarios.

1 Introduction

Recent multilingual large language models (mLLMs), represented by BLOOM (Scao et al., 2022), MaLA500 (Lin et al., 2024b), and Aya (Üstün et al., 2024), have shown impressive capacity on diverse tasks across languages. Parallel corpora have emerged as crucial resources for enhancing mLLMs, both for specific tasks, e.g., machine translation (Xu et al., 2023; Alves et al., 2024), and for general-purpose tasks (Cahyawijaya et al., 2023; Zhu et al., 2023; Li et al., 2023).

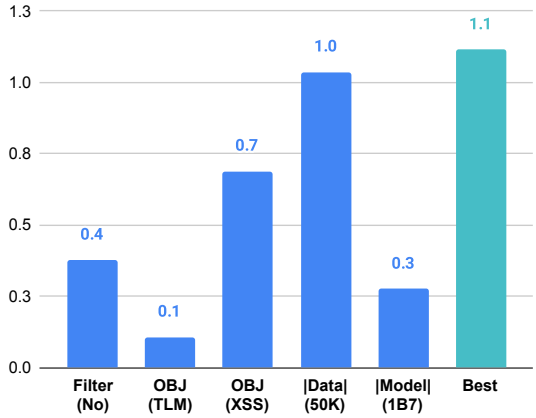

Figure 1: Average performance improvements (y-axis) achieved by mLLMs enhanced with parallel corpora compared to their base models. **Best: Instruction tuning of BLOOM-7B1 with the machine translation objective (MT) using 10K high-quality parallel sentences yields the best results**. Main variations explored include: **Filter (No)** (using the original data); **OBJ (TLM)** (translation language modeling objective); **OBJ (XSS)** (cross-lingual semantic similarity objective); **|Data| (50K)** (a larger 50K-sentence dataset); **|Model| (1B7)** (BLOOM-1B7 model).

However, existing studies often overlook a comprehensive exploration of methodologies for harnessing parallel corpora. The quality and quantity of parallel corpora remain inadequately explored, inhibiting the full potential of such resources. Moreover, the influence of different training objectives and mLLM sizes across diverse languages and tasks remains under-investigated. This limitation impedes the generalization of parallel corpus exploitation methods across varied linguistic landscapes and task domains. Therefore, this paper aims to address these gaps by presenting a comprehensive recipe for exploiting parallel corpora for mLLMs. We focus on four key factors, with some main results shown in Figure 1.

Quality: We explore three dimensions of paral-

lel corpus quality: translation accuracy, sentence length, and language identification. Our results show that translation quality is vital for exploiting parallel corpora, while sentence length filtering and language identification have minimal impact.

Quantity: Acquiring large amounts of high-quality parallel corpora is challenging, especially for low-resource languages. Our study examines the minimum corpus size necessary to achieve performance improvements across diverse tasks. We find that even a corpus of just 10K sentences can yield results comparable to those obtained from much larger datasets.

Objective: Previous studies (Cahyawijaya et al., 2023) have investigated the effectiveness of different training objectives and their combinations on classification tasks of Indonesian local languages, using smaller-sized mLLMs up to 1B7. We extend this investigation by examining the impact of various training objectives and their combinations on larger mLLMs across a range of languages and tasks. Our experiments demonstrate that employing the machine translation objective produces the most promising results.

Model Size: The size of mLLMs can greatly impact their ability to comprehend instructions derived from parallel corpora. Our findings indicate that larger mLLMs exhibit superior comprehension and cross-task transferability compared to their smaller counterparts. Consequently, they achieve more substantial improvements across a broader spectrum of tasks.

In light of the critical role parallel corpora play in mLLMs, our study provides a comprehensive recipe for effectively exploiting parallel corpora. We have identified four primary factors: quality (§4), quantity (§5), objective (§6), and model size (§7). Our detailed analysis of these factors reveals their great impact on mLLM performance across diverse languages and tasks. By delving into these aspects, we offer actionable insights that can inform the development and optimization of strategies for parallel corpus exploitation, ultimately contributing to the advancement of mLLMs in both bilingual and general-purpose tasks.

2 Related Work

2.1 Parallel Data for Multilingual Language Models

Over the years, multilingual language models have evolved from earlier, smaller models, such as XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), and Glot500 (Imani et al., 2023), to more recent, larger models, including BLOOM (Scao et al., 2022), MaLA500 (Lin et al., 2024b), and Aya (Üstün et al., 2024). These models consistently demonstrate strong performance across various downstream tasks (Ahuja et al., 2023; Lin et al., 2024c).

Parallel corpora have played a pivotal role in both the analysis (Piqueras and Søgaard, 2022; Lin et al., 2024a) and enhancement (Conneau and Lample, 2019; Ouyang et al., 2020; Yang et al., 2020; Huang et al., 2019; Chi et al., 2021a; Wei et al., 2021; Hu et al., 2021; Chi et al., 2021b; Reid and Artetxe, 2022b; Liu et al., 2023) of small multilingual language models.

In the era of mLLMs, parallel corpora are constructed as instruction data and used to enhance mLLMs through supervised fine-tuning (Cahyawijaya et al., 2023; Zhu et al., 2023; Li et al., 2023). Specifically, Cahyawijaya et al. (2023) propose three methods of incorporating parallel corpora as instruction tuning data: Machine Translation (MT), Translation Language Modeling (TLM), and Cross-Lingual Semantic Similarity (XSS) (see §3.3). However, their evaluation is limited to small models with up to 1.7 billion parameters and focuses solely on classification tasks within Indonesian local languages. Both Zhu et al. (2023) and Li et al. (2023) propose using machine-translation-style instruction data to improve mLLMs but do not explore different training objectives. While these studies yield promising results, their scope is limited. Firstly, they do not explore critical factors such as the quality and quantity of parallel corpora, considering the high cost of collecting high-quality and massive parallel corpora, especially for low-resource languages. Secondly, their investigations do not encompass an in-depth analysis of training objectives and mLLMs with varied model sizes across diverse languages and tasks.

2.2 Key Elements for Language Modeling

Previous research has extensively examined critical factors essential for the pretraining and enhancement of language models.

Quality: Kreutzer et al. (2022) conducted manual audits of prevalent monolingual and parallel corpora, revealing significant portions of low-quality data, particularly in corpora for low-resource languages. Follow-up studies have investigated the impact of data quality on model performance. Artetxe

et al. (2022) observed that similar results on downstream tasks can be achieved regardless of the degree of quality of the corpus used for pretraining, while other studies found that the quality of parallel corpora matters for machine translation (Ranathunga et al., 2024) and general-purpose tasks (Reid and Artetxe, 2022a).

Quantity: Recent works (Chen et al., 2023; Zhou et al., 2023; Gupta et al., 2023) have focused on the impact of fine-tuning with small amounts of high-quality instruction data, such as one or a few thousand instances, showing promising performance gains in evaluation tasks. Xu et al. (2023) demonstrate that as few as 10K high-quality parallel sentences can significantly enhance machine translation performance.

Objective: Different training objectives based on parallel corpora for enhancing mLLMs can be viewed as distinct instructions. Wang et al. (2023) explore the impact of various types of instruction tuning data and find that their combination can be optimal in certain scenarios.

Model Size: Recent studies indicate that scaling up language models enhances their capability to excel in diverse and complex reasoning tasks (Wei et al., 2022, 2023; Lu et al., 2023). Follow-up studies (Wei et al., 2023) further illustrate distinct behavioral differences between larger and smaller models.

However, these factors have not yet been comprehensively explored in the context of leveraging parallel corpora to enhance mLLMs across diverse languages and tasks.

3 Setup

3.1 Language

We use two criteria for language selection. Firstly, we select languages well covered by mLLMs and evaluation benchmarks, allowing for robust evaluation across diverse downstream tasks. Secondly, we select typologically diverse languages, enabling our investigation to generalize to a wide range of low-resource languages. Thus, we select five languages: Arabic (ar), Spanish (es), Hindi (hi), Vietnamese (vi) and Chinese (zh).

3.2 Data

We utilize the OPUS100 dataset (Zhang et al., 2020), an English-centric multilingual corpus, to gather parallel sentences between English (en) and each target language. The quality of OPUS100 is assessed across three dimensions:

Translation Quality Manual quality assessment of the vast amount of parallel corpora is impractical. Instead, we employ COMETKIWI (Rei et al., 2022)[1], a tool for estimating the quality of machine translation outputs across multiple languages. We set a COMETKIWI score threshold τ_c, retaining parallel corpora with scores not lower than τ_c.

Sentence Length Given the variation in character length across languages, we avoid using it as a metric for consistency. Instead, we measure sentence length by the number of tokens, as determined by the tokenizer of our chosen mLLM, BLOOM-7B1. We establish a length threshold τ_l, retaining parallel corpora where both source and target sentences contain no fewer than τ_l tokens.

Language Identification To identify sentences potentially not in the correct language, we employ GlotLID (Kargaran et al., 2023), an open-source language identification model. This language identification filter is applied to both the source and target sentences.

3.3 Training

We select the BLOOM series (Scao et al., 2022) for our investigation due to its offering of different sizes of mLLMs which well cover the five target languages under consideration.[2] The training data size for the five target languages ranges from 23GB (Hindi) to 452GB (English). We explore BLOOM models of various sizes, including 7B1, 3B, and 1B7. Due to limited computational resources, we use LoRA (Hu et al., 2022), which is known for its competitive performance compared to full-parameter training (Alves et al., 2023), for instruction tuning of BLOOM. We configure the learning rate to 1×10^{-4}, weight decay to 0.1, and set the rank of LoRA to 16 based on preliminary experiment in §B. The maximum sequence length for both source and target sentences is set to 128. To maintain consistency across experiments with different quantities of parallel corpora, we ensure a uniform training budget of 50K parallel sentences. Specifically, we calculate the number of epochs as 50K divided by the number of sentences consid-

[1] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

[2] Additional experiments with XGLM, as presented in §A, yield conclusions consistent with those of BLOOM.

Objective	Template
MT	`Translate the following text from [SOURCE_LANG] to [TARGET_LANG].\nText: [SOURCE_TEXT]\nTranslation: [TARGET_TEXT]`
TLM	`[INPUT_TEXT]. Denoise the previous [TARGET_LANG] text to its equivalent sentence in [SOURCE_LANG]: [SOURCE_TEXT]\n[TARGET_TEXT]`
XSS	`[SOURCE_LANG] sentence: [SOURCE_TEXT]\n[TARGET_LANG] sentence: [TARGET_TEXT]\nDo the two sentences have the same meaning? [LABEL]`

Table 1: Templates of MT, TLM, and XSS for instruction data construction based on parallel corpora.

ered from the OPUS100 dataset. The batch size is 128, and we save the checkpoints every 20 steps.

Following Cahyawijaya et al. (2023), we construct the data for instruction tuning based on the parallel corpora by three distinct patterns: Machine Translation (MT), Translation Language Modeling (TLM), and Cross-Lingual Semantic Similarity (XSS). Table 1 presents the templates for these three objectives. Here, `[SOURCE_LANG]` and `[TARGET_LANG]` represent the language names of the source and target languages, respectively. In our study, we consider both English-to-target-language and target-language-to-English directions, where `[SOURCE_LANG]` represents English or `[TARGET_LANG]` represents English. For MT, `[SOURCE_TEXT]` and `[TARGET_TEXT]` refer to the parallel sentences in the source and target languages, respectively. For TLM, a portion of tokens in `[TARGET_TEXT]` are masked to generate `[INPUT_TEXT]`. For XSS, our objective is to predict whether parallel sentences `[SOURCE_TEXT]` and `[TARGET_TEXT]` are semantically similar, with `[LABEL]` being "Yes" or "No". Specifically, we utilize the parallel corpora as positive examples and introduce perturbations to `[TARGET_TEXT]` to construct negative examples. We consider applying the objectives both individually and in combination. We tune the model with the objective of casual language modeling without loss mask.

3.4 Evaluation

We conduct evaluation across five diverse benchmarks: FLORES (Costa-jussà et al., 2022), MUSE (Lample et al., 2018), MLQA (Lewis et al., 2020), XQUAD (Artetxe et al., 2020), and SIB (Adelani et al., 2023). A comprehensive overview of these benchmarks is available in Table 2. Our evaluation spans both classification tasks (SIB) and generation tasks (FLORES, MUSE, MLQA, and XQuAD), covering a spectrum of cross-language (FLORES, MUSE, and MLQA) and in-language tasks (XQuAD and SIB).

Dataset	Task	|Data|	Metric	I/C	C/G
FLORES	Machine Translation	1012	COMETKIWI	C	G
MUSE	Word Translation	1500	F1	C	G
MLQA	Question Answering	4918 - 5495	F1	C	G
XQuAD	Question Answering	1190	F1	I	G
SIB	Text Classification	204	Acc	I	C

Table 2: Details of evaluation benchmarks. |Data|: Number of samples for evaluation. I/C: In-language/Cross-language. C/G: Classification/Generation.

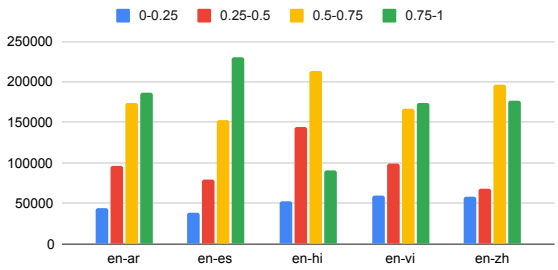

Figure 2: Translation quality measured by COMETWIKI of 500K parallel sentences from OPUS100 for our five language pairs. The COMETWIKI scores are segmented into four ranges: 0-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1. Higher scores represent better translation quality.

For translation tasks within FLORES and MUSE, we explore bidirectional translation: from English to other languages (en-xx) and from other languages to English (xx-en). Additionally, for MLQA, we evaluate scenarios where questions are in English and the passages and answers are in other languages (en-xx), as well as situations where questions are in other languages and the passages and answers are in English (xx-en).

To provide a thorough understanding of our evaluation procedures, we offer detailed prompts for each task in §C. In all experiments, we employ a 2-shot in-context learning approach, where the model is given two examples appended to the input query to aid in making predictions.

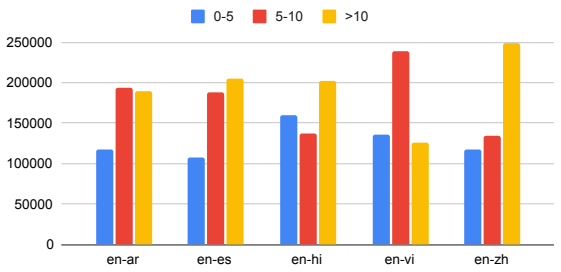

Figure 3: Sentence length of 500K parallel sentences from OPUS100 for our five language pairs. The three categories are 0-5, 5-10, greater than 10 tokens.

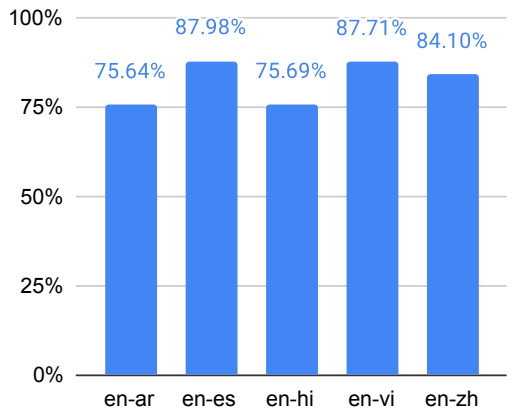

Figure 4: Percentage of sentences retained after language identification filtering of 500K parallel sentences from OPUS100 for our five language pairs.

4 Quality

4.1 Quality of OPUS100

We measure the quality of 500K parallel sentences from OPUS100 for our five language pairs using three key metrics: translation quality, sentence length, and language identification accuracy, as illustrated in Figures 2–4.

A considerable portion of OPUS100 is of low-quality. All quality measures indicate that a large portion of OPUS100 contains low-quality data. Approximately 10% of the data has COMETKIWI scores below 0.25, indicating very poor translation quality. Additionally, between 10% to 30% of the data falls within the 0.25-0.5 score range, which is still considered sub-optimal. Regarding sentence length, we find that over 20% of the OPUS100 data consists of very short sentences, with a length of no more than five tokens. For language identification, 13% to 25% of the data is removed due to incorrect language identification results in one of the two parallel sentences.

Low-resource languages suffer more from low-quality issues. For low-resource languages like Hindi, there are fewer high-quality parallel sentences compared to high-resource languages such as Spanish. Analysis of translation quality indicates that the English-Hindi pair has less than 20% of parallel sentences with high COMETWIKI scores (0.75-1), whereas the English-Spanish pair has around 45%. For sentence length, the English-Hindi pair contains 10% more short sentences (0-5 tokens) compared to high-resource language pairs. Moreover, both the English-Arabic and English-Hindi pairs exhibit about 10% more parallel sentences that may be in the wrong languages.

These comprehensive findings underscore the critical importance of data quality when exploiting parallel corpora for mLLM training.

4.2 Effect of Quality

Table 3 presents the performance of BLOOM-7B1 after instruction tuning with the machine translation objective, using 10K parallel corpora with various quality filtering strategies.

Parallel corpora containing noisy translations still improve results. Comparing the results of the experiment with $\tau_c = 0$ (ID 1) to the original model (ID 0), there's an average improvement of 0.4% for all tasks. The most notable improvements are observed in both bilingual tasks (en-xx) and in-language tasks. However, generating English for bilingual tasks yields degraded or marginally improved results. Experiment 0 exhibits 0.7% and 1.1% decrements in FLORES and MUSE respectively, with only a 0.3% improvement in MLQA.

Filtering out noisy translations leads to notable improvements. When $\tau_c = 0.5$, the average performance rises from 53.2% to 53.7%. Further refinement to $\tau_c = 0.75$ achieves an additional 0.3% improvement. These improvements are consistently observed across all evaluated tasks. In the optimal setting (ID 5), there's a 1.2% improvement compared to BLOOM-7B1 (ID 0). The improvements corroborate the reliability of COMETKIWI as a metric for filtering low-quality translations.

Filtering short sentences yields slightly worse results than using unfiltered data. The experiment with filtering short sentences (ID 3) achieves comparable or slightly worse results compared to that without filtering short sentences (ID 5). This suggests that short sentences, whether at the word or phrase level, may offer some benefits for sentence-level tasks.

ID	Model			FLORES		MUSE		MLQA		XQuAD	SIB	AVG
				EN-XX	XX-EN	EN-XX	XX-EN	EN-XX	XX-EN			
0	BLOOM-7B1			69.1	**72.4**	43.1	53.7	36.4	42.7	47.2	58.1	52.8
	τ_c	τ_l	LID									
1	0	0	✓	69.7	71.7	44.4	52.6	37.8	43.0	47.7	58.8	53.2
2	0.5	0	✓	69.9	72.1	45.0	53.0	**38.1**	43.7	48.1	**59.8**	53.7
3	0.75	5	✓	70.3	72.1	**45.7**	53.6	**38.1**	43.5	47.8	59.2	53.8
4	0.75	0	✗	**70.5**	72.1	44.9	53.7	37.7	**44.0**	**48.3**	59.6	53.9
5	0.75	0	✓	70.3	72.3	45.5	**53.9**	38.0	43.9	**48.3**	59.5	**54.0**

Table 3: Performance (%) of BLOOM-7B1 after instruction tuning with the machine translation objective using 10K parallel corpora with various quality filtering strategies. Parameters include τ_c for COMETWIKI score threshold, τ_l for sentence length threshold, and LID indicating the adoption of language identification filtering.

Using data with language identification filtering results in only a 0.1% improvement on average. A comparison of experimental outcomes with and without language identification filtering (ID 4 and 5) reveals that using data with language identification filtering yields merely a 0.1% improvement on average. The most notable performance difference is observed in the MUSE task, where using data with language identification filtering leads to improvements of 0.6% (en-xx) and 0.2% (xx-en). This marginal enhancement may be attributed to the presence of sentences in similar languages within OPUS100, which exhibit minor linguistic variations compared to the true language. These variations could potentially have a slight negative impact on word-level translations while having little impact on sentence-level tasks.

5 Quantity

5.1 Effect of Quantity Across Tasks

Based on Table 4, which shows the performance of BLOOM-7B1 after instruction tuning with the machine translation objective using different amounts of parallel sentences, we can derive the following key findings:

Adding merely 1K parallel sentences helps. Exploiting 1K parallel sentences for instruction tuning improves the overall average score by 1%. This increase is observed across most tasks, with notable improvements in FLORES (en-xx), MUSE (en-xx), and SIB.

Using 10K parallel sentences leads to the optimal performance. The best overall performance is achieved with 10K parallel sentences, resulting in an average score of 54.0%. This setting yields the highest scores in MUSE and SIB. This aligns with the findings in Xu et al. (2023).

More data achieves comparable results. Increasing the number of parallel sentences beyond 10K results in comparable performance. Specifically, using 25K or 50K parallel sentences yields average scores of 53.9%, which are very close to the score obtained with 10K sentences.

The analysis suggests that instruction tuning with a moderate amount of parallel sentences (around 10K) yields the best overall improvement in performance for the BLOOM-7B1 model across various tasks.

5.2 Effect of Quantity Across Languages

We delve deeper into the influence of parallel corpora quantity, as depicted in Table 5.

Using 10K parallel sentences achieves optimal performance across most languages. For the majority of languages, except Vietnamese (vi) and Chinese (zh), the highest performance is obtained with 10K parallel sentences. Even for Vietnamese and Chinese, leveraging 10K parallel sentences can yield comparable results. These observations align with the findings in §5.1.

Different languages exhibit varying appetites for parallel corpora. Across most languages, increasing the number of parallel sentences used for instruction tuning generally leads to incremental improvements in performance. However, for Hindi (hi) and Chinese (zh), transitioning from 1K to 10K parallel sentences does not yield improvement. This phenomenon may be attributed to BLOOM-7B1's limited proficiency in these languages compared to others, as reflected in the results of the original BLOOM-7B1 model (|Sent|=0).

|SENT|	FLORES		MUSE		MLQA		XQUAD	SIB	AVG
	EN-XX	XX-EN	EN-XX	XX-EN	EN-XX	XX-EN			
0	69.1	**72.4**	43.1	53.7	36.4	42.7	47.2	58.1	52.8
1K	70.0	72.2	45.3	53.6	**38.2**	43.6	47.9	59.2	53.8
5K	70.3	72.2	45.4	53.5	**38.2**	43.8	48.2	**59.5**	53.9
10K	70.3	72.3	**45.5**	**53.9**	38.0	43.9	48.3	**59.5**	**54.0**
25K	70.3	72.2	45.1	53.8	38.0	**44.0**	**48.4**	**59.5**	53.9
50K	**70.4**	72.2	45.1	53.8	38.1	43.7	48.3	**59.5**	53.9

Table 4: Performance (%) of BLOOM-7B1 after instruction tuning with the machine translation objective using varying amounts of parallel sentences, obtained with the best filtering strategy (ID 5) as shown in Table 3. |SENT| indicates the number of parallel sentences used for instruction tuning, with |SENT|=0 representing the original BLOOM-7B1 model.

	ar	es	hi	vi	zh
0	49.5	57.7	46.5	63.8	46.7
1K	50.8	58.1	**47.7**	64.3	47.8
5K	51.2	58.2	47.6	64.6	**47.9**
10K	**51.3**	**58.4**	**47.7**	64.6	47.8
25K	51.2	58.2	**47.7**	**64.7**	47.7
50K	51.2	58.2	**47.7**	**64.7**	47.6

Table 5: Performance (%) of BLOOM-7B1 after instruction tuning with the machine translation objective using varying amounts of parallel sentences, obtained with the best filtering strategy (ID 5) as shown in Table 3. |SENT| indicates the number of parallel sentences used for instruction tuning, with |SENT|=0 representing the original BLOOM-7B1 model.

6 Objective

We explore the effectiveness of different objectives and their combinations, with results in Table 6.

BLOOM-7B1 performs well on English generation tasks. The baseline BLOOM-7B1 model exhibits robust performance across a spectrum of evaluation tasks, notably excelling in English generation tasks such as FLORES (xx-en) and MUSE (xx-en). Further exploitation of parallel corpora fails to yield any discernible improvement.

MT emerges as the top performer. The MT objective consistently outperforms the baseline BLOOM-7B1 model, showcasing an average improvement of 1.2%. Moreover, MT achieves the highest performance in 5 out of 8 evaluated tasks.

TLM exhibits limited effectiveness. While TLM shows slight improvements on average (0.2%), primarily driven by enhancements in tasks like MUSE (en-xx), MLQA (xx-en), XQuAD, and SIB, it also leads to degradation in tasks including FLORES and MUSE (xx-en).

XSS achieves strong performance for classification. Using the XSS objective improves BLOOM-7B1 by 0.7%, though it performs 0.5% worse than MT. The major decrease is observed in translation tasks, especially from English to other languages. However, XSS can still slightly improve translation tasks compared to BLOOM-7B1. Notably, XSS achieves 0.3% better performance on SIB, highlighting its effectiveness for classification.

Combining training objectives does not provide large benefits. While combinations of different objectives can improve BLOOM-7B1 by 0.2% to 1.0%, none surpass the performance of using the MT objective alone. The combination of MT and XSS is the best among the combinations, slightly worse than MT by 0.2%, but better than all other objectives. Notably, MT +XSS achieves the best results on SIB, and TLM +XSS yields the best results on MLQA (xx-en). These observations indicate that no single objective excels across all tasks.

7 Model Size

We explore the impact of parallel corpora on various sizes of BLOOM models, detailed in Table 7.

Smaller models exhibit more pronounced improvements in FLORES. Notably, BLOOM-1B7 demonstrates larger improvements compared to its larger counterparts in the FLORES task, where the prompt is the same as the one used during instruction tuning with the MT objective. This is attributed to the smaller models' less developed in-context learning capabilities before instruction tuning, allowing for more substantial improvements when supplemented with parallel corpora.

Larger models excel in diverse tasks. Con-

MODEL	FLORES		MUSE		MLQA		XQUAD	SIB	AVG
	EN-XX	XX-EN	EN-XX	XX-EN	EN-XX	XX-EN			
BLOOM-7B1	69.1	**72.4**	43.1	53.7	36.4	42.7	47.2	58.1	52.8
MT	**70.3**	72.3	**45.5**	**53.9**	**38.0**	43.9	**48.3**	59.5	**54.0**
TLM	67.2	72.2	44.3	53.0	36.3	44.4	47.6	58.7	53.0
XSS	69.4	72.2	43.7	53.5	37.0	44.2	**48.3**	60.0	53.5
MT +TLM	69.3	72.1	44.1	53.2	36.8	43.8	47.2	59.5	53.2
MT +XSS	70.3	72.1	44.9	53.3	37.4	44.5	47.9	**60.4**	53.8
TLM +XSS	67.7	72.2	43.0	52.5	34.9	**45.6**	48.2	60.0	53.0
MT +TLM +XSS	69.5	72.1	44.2	53.2	36.1	45.1	47.7	59.0	53.4

Table 6: Performance (%) of BLOOM-7B1 after instruction tuning with different objectives and their combinations using 10K parallel corpora, obtained with the best filtering strategy (ID 5) as shown in Table 3.

MODEL	FLORES		MUSE		MLQA		XQUAD	SIB	AVG
	EN-XX	XX-EN	EN-XX	XX-EN	EN-XX	XX-EN			
BLOOM-7B1	69.1	72.4	43.1	53.7	36.4	42.7	47.2	58.1	52.8
+ Parallel Data	70.3	72.3	45.5	53.9	38.0	43.9	48.3	59.5	54.0
Δ	01.2	-00.1	02.4	00.2	01.6	01.2	01.0	01.4	01.2
BLOOM-3B	64.0	68.9	39.7	50.9	29.4	26.2	32.7	54.5	45.8
+ Parallel Data	65.0	69.1	41.4	51.6	30.9	26.7	34.5	56.9	47.0
Δ	01.0	00.2	01.8	00.7	01.5	00.5	01.8	02.4	01.2
BLOOM-1B7	59.0	65.8	37.2	48.5	20.0	22.2	24.8	53.0	41.3
+ Parallel Data	61.1	65.7	38.9	48.0	20.8	20.9	24.4	53.0	41.6
Δ	02.0	-00.1	01.6	-00.6	00.8	-01.3	-00.3	00.0	00.3

Table 7: Effect of parallel corpora on BLOOM models of different sizes across various tasks. '+ Parallel Data' indicates instruction tuning of the given mLLM with the MT objective, using 10K parallel corpora obtained with the best filtering strategy (ID 5) as shown in Table 3.

versely, larger models generally demonstrate greater enhancements in tasks beyond machine translation. Both BLOOM-7B1 and BLOOM-3B exhibit a 1.2% improvement compared to their original mLLMs, while BLOOM-1B7 shows a slight 0.3% improvement. Specifically, BLOOM-7B1 and BLOOM-3B display notable improvements in tasks except for FLORES, while BLOOM-1B7 achieves comparable or even worse results.

These findings demonstrate that when leveraging parallel corpora to enhance mLLMs, larger models not only exhibit improvements in direct tasks, such as machine translation, but also demonstrate a more substantial overall enhancement across a variety of tasks. In contrast, smaller models primarily show benefits in direct tasks. This difference can be attributed to the superior cross-task transferability of larger mLLMs, where insights gained from parallel corpora in one task contribute to improved performance in others.

8 Conclusion

This paper investigates the impact of four critical factors – data quality, data quantity, objectives, and mLLM sizes – on leveraging parallel corpora to enhance mLLMs across diverse languages and tasks. Our findings underscore the crucial importance of filtering out noisy translations to procure high-quality training data for improving mLLMs. Surprisingly, even a relatively modest dataset of 10K samples can yield promising results. Furthermore, our analysis shows that employing the machine translation objective leads to optimal outcomes. Importantly, larger models exhibit a greater capacity to benefit from parallel corpora, achieving more substantial improvements. This study provides a comprehensive recipe for effectively leveraging parallel corpora to enhance mLLMs. These insights significantly contribute to advancing the understanding and optimization of mLLMs across different languages and tasks.

Limitations

Due to limited computational resources, we opted not to explore full-parameter training for leveraging parallel corpora. Instead, we focused on LoRA, drawing on insights from previous studies. Additionally, our investigation is restricted to the BLOOM series, and we did not extend our analysis to other mLLMs. Furthermore, we did not also explore mLLMs larger than 7B1.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2023. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *CoRR*, abs/2309.07445.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

Duarte M. Alves, Nuno Miguel Guerreiro, João Alves, José Pombal, Ricardo Rei, José Guilherme Camargo de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11127–11148. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11*, pages 7383–7390. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instruct-align: Teaching novel languages with to llms through alignment-based cross-lingual instruction. *CoRR*, abs/2305.13627.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. 2023. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *CoRR*, abs/2305.09246.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3418–3430. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc 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. *CoRR*, abs/2207.04672.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction tuned models are quick learners. *CoRR*, abs/2306.05539.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3633–3643. Association for Computational Linguistics.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2485–2494. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6155–6218. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7315–7330. Association for Computational Linguistics.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023. Align after pre-train: Improving multilingual generative models with cross-lingual alignment. *CoRR*, abs/2311.08089.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2024a. mplm-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 276–310. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024b. Mala-500: Massive language adaptation of large language models. *CoRR*, abs/2401.13303.

Peiqin Lin, André F. T. Martins, and Hinrich Schütze. 2024c. XAMPLER: learning to retrieve cross-lingual in-context examples. *CoRR*, abs/2405.05116.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. 2023. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. *CoRR*, abs/2311.08849.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *CoRR*, abs/2309.01809.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-M: enhanced multilingual representation by

aligning cross-lingual semantics with monolingual corpora. *CoRR*, abs/2012.15674.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3597–3605. International Committee on Computational Linguistics.

Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. *CoRR*, abs/2402.07446.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2022a. On the role of parallel data in cross-lingual transfer learning. *CoRR*, abs/2212.10173.

Machel Reid and Mikel Artetxe. 2022b. PARADISE: exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 800–810. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin 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. 2024. Aya model: An instruction finetuned open-access multilingual language model. *CoRR*, abs/2402.07827.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *CoRR*, abs/2306.04751.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9386–9393. AAAI Press.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1628–1639. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language

models to non-english by aligning languages. *CoRR*, abs/2308.04948.

A Additional Experiments on XGLM

We explore the effect of parallel corpora quality on XGLM-7.5B, and the results are shown in Table 8. As shown, experiments on XGLM-7.5B exhibits consistent findings as those on BLOOM-7B1.

ID	MODEL			Accuracy
0	XGLM-7.5B			52.8
	τ_c	τ_l	LID	
1	0	0	✓	54.6
2	0.5	0	✓	55.0
3	0.75	5	✓	54.4
4	0.75	0	✗	55.0
5	0.75	0	✓	**55.5**

Table 8: Performance (%) of XGLM-7.5B on SIB after instruction tuning with the machine translation objective using 10K parallel corpora with various quality filtering strategies. Parameters include τ_c for COMETWIKI score threshold, τ_l for sentence length threshold, and LID indicating the adoption of language identification filtering.

B Effect of LoRA Rank

We conduct initial experiments on Arabic to explore the effect of setting the rank of LoRA, and the results are shown in Table 9. As shown, setting the rank of LoRA as 16 lead to best performance. Therefore, we use LoRA rank as 16 across all the experiments.

Rank	Accuracy
8	59.0
16	**62.0**
32	60.0
128	61.0
256	60.5
512	59.5

Table 9: Effect of LoRA rank: Accuracy on SIB using English-Arabic parallel data to improve BLOOM-7B1.

C Prompt

The prompts of FLORES, MUSE, MLQA, XQuAD, and SIB are shown as follows:

FLORES/MUSE

```
Translate the following
text from [SOURCE_LANG]
to [TARGET_LANG].\nText:
[SOURCE_TEXT]\nTranslation:
[TARGET_TEXT]
```

MLQA/XQuAD

```
[Passage] \nQ:
[Question]\nA: [Answer]
```

SIB

```
The topic of the news
[Passage] is [Label]
```