XLLaMA2: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate remarkable translation capabilities in high-resource language tasks, yet their performance in low-resource languages is hindered by insufficient multilingual data during pre-training. To mitigate this, we continued pre-train LLaMA2-7B to support translation across more than 100 languages. Following a thorough analysis of training strategies, including vocabulary expansion and data augmentation, we apply extensive multilingual continued pre-training to the LLaMA series model, resulting in XLLaMA2. Without loss of the generality ability, the translation performance of XLLaMA2 significantly surpassed existing LLMs and is on par with that of a specialized translation model (M2M-100-12B) on the Flores-101 benchmark. Specifically, XLLaMA2 achieves an average spBLEU score improvement of over 10 points compared to the original LLaMA2 model. Further testing XLLaMA2 on Flores-200, XLLaMA2 exhibited notable performance gains even for languages not included in the training set. We will make the code and model publicly available.

1 Introduction

Large Language Models (LLMs; OpenAI, 2023; Zhang et al., 2022; Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a,b) exhibit excellent in translation tasks involving high-resource languages (Vilar et al., 2023; Zhu et al., 2023), yet their effectiveness in low-resource translation is suboptimal (Hendy et al., 2023; Zhu et al., 2023; Bang et al., 2023). As illustrated in Figure 1, which presents the number of translation directions with performance exceeding 10 spBLEU (Goyal et al., 2022) scores on Flores-101 (Goyal et al., 2022), it is evident the majority of models are clustered around the origin for ar-centric translations, demonstrating a significant disparity when compared to their en-centric performance.

This discrepancy is primarily due to the lack of pre-training data for these languages (Wei et al., 2023; Alves et al., 2024; Yuan et al., 2023b). Many researchers are actively working to address this issue. Guo et al. (2024) enhance the LLMs' ability by translating low-resource languages after learning textbooks. Zhu et al. (2023) find cross-lingual examples that can provide better task guidance for low-resource translation. In addition to the efforts focus on the fine-tuning stage, some studies (Wei et al., 2023) have attempted to train a multilingual LLM from scratch, or to train a language-specific LLM (Faysse et al., 2024; Alves et al., 2024). However, the languages covered by these works are not extensive (Wei et al., 2023; Alves et al., 2024; Luo et al., 2023), and in some cases, the translation performance is still insufficient (Wei et al., 2023; Alves et al., 2024; Luo et al., 2023). Moreover, these efforts represent promising strides toward improving the translation performance of LLMs, especially in low-resource languages.

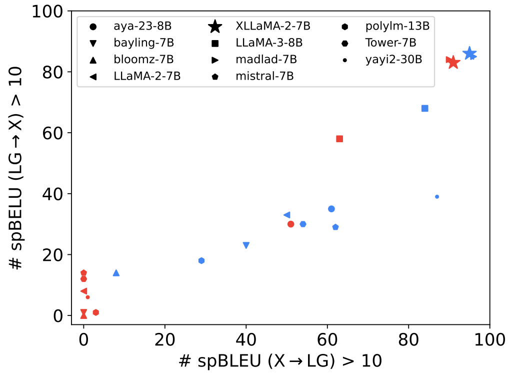

Figure 1: We evaluate both X→LG and LG→X translations on various models using Flores-101 test. The results are visualized in a figure where different markers represent various models, a red marker signifies LG=ar, and a blue marker denotes LG=en. We count the number of translation directions that achieve a spBLEU score higher than 10. The findings indicate that modest LLMs demonstrate strong support for en-centric translation, but underperform in ar-centric translation.

Meanwhile, for LLMs, translation is a critical capability (Zhu et al., 2024a, 2023), representing a complex multilingual challenge. Effective translation necessitates a profound comprehension of both the source and target languages, encompassing their syntax, semantics, and pragmatics. This process extends beyond word-to-word mapping, requiring the capture of subtle nuances, idiomatic expressions, and cultural references inherent in different languages.

In the quest to enhance the translation performance of LLMs, we have comprehensively covered the 102 languages supported by Flores-101, utilizing a continued pre-training strategy on parallel and monolingual data. Firstly, we delve into critical technical issues in the training. This technical framework lays the groundwork for the training procedure, influencing its efficacy and, ultimately, the performance of the LLMs. It involves the vocabulary extension and the data augmentation.

Deciding which vocabulary to use is the foremost crucial issue for expanding language support. We conducted a quantitative analysis of the impact of adding various language-specific new tokens, evaluating from the perspectives of tokenization granularity, embedding quality, and the influence of the model's inner distribution. Interestingly, we find that adhering to the original vocabulary of LLMs is the most cost-effective approach for expanding the LLMs to 102 languages. The introduction of a small number of new tokens can significantly impact the performance of the existing LLMs, while a large number of tokens increase the difficulty of training and also require more data.

Another well-recognized challenge for low-resource languages is data scarcity. Firstly, engage in a comprehensive discourse on varied dictionary-based data augmentation strategies applicable to both monolingual and parallel datasets, investigating their optimal implementation on monolingual data, parallel data, or a combination thereof. Our findings indicate that for LLMs, dictionary-based data augmentation is more effective when applied to parallel data. Furthermore, we also delve into the usage of different dictionaries and find that the performance of augmentation is correlated with the number of entities covered.

By incorporating these techniques, we execute large-scale, multilingual continued pre-training on LLaMA2-7B, significantly enhancing its translation capabilities. XLLaMA2 demonstrated comparable translation performance to M2M-100-12B on Flores-101, showing an average improvement of over 10 spBLEU compared to LLaMA2 model on low-resource languages. We extended our testing to Flores-200 and observed substantial performance improvements even for languages not included in the training set. Importantly, these enhancements did not compromise the performance of general tasks. The multilingual model derived from XLLaMA2 outperformed the zero-shot capabilities of the original LLaMA2. Furthermore, by applying supervised fine-tuning to this multilingual model, we achieved a performance increase of more than 4 points on multilingual tasks compared to the previous LLaMA2 model. Our main contributions:

- An open-sourced XLLaMA2 extends LLaMA2 to support more than 100 languages.
- Comprehensive analysis of the key techniques, including vocabulary sharing and data augmentation, in continued per-training to LLMs.
- Extensive experiments on key technique design, translation benchmark, and general tasks, prove the superiority of XLLaMA2.

2 Related Work

Multilingual Large Language Models. Large Language Model (LLMs; OpenAI, 2023; Zhang et al., 2022; Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a,b) trained with English-centric data can also solve various non-English tasks (**?**Srivastava et al., 2022; Kwiatkowski et al., 2019; Hendrycks et al., 2021c), but the performance between non-English and English is significantly large (Yuan et al., 2023b). Efforts to develop more multilingual LLMs in two different ways: retraining LLMs with diverse multilingual data from scratch (Wei et al., 2023); or continuous training of pre-trained models using language-specific data with the option to expand the vocabulary(Zhao et al., 2024a; Cui et al., 2024; Faysse et al., 2024; Alves et al., 2024). Instead of training from scratch, continued pre-training aims at updating pre-trained models with new data, making the process more efficient and cost-effective (Gupta et al., 2023; Alves et al., 2024; Xie et al., 2023).

Multilinguality in LLMs. Recent research has shed light on the multilingual capabilities of LLMs. A comprehensive survey by Huang et al. (2024a) discusses various aspects of multilingualism in LLMs, including training and inference methods, model security, multi-domain with languages culture, and emphasizes the need for language-fai technology. Yuan et al. (2023b) analysis multilin-

Algorithm 1: Illustration of the Training Data Construction Process During a Single Training Epoch

```
Input: $A$: all language list. $\mathcal{D}^A_{\text{mono}}$: monolingual data for all languages. $\mathcal{D}_{\text{En}}$: an English monolingual data. $\mathcal{D}^A_{\text{para}}$: a
        parallel data for all translation directions. Notably, $\mathcal{D}^A_{\text{mono}} \bigcap \mathcal{D}_{\text{En}} = \varnothing$. $\boldsymbol{x}$: a single data point. $g(\boldsymbol{x}; \boldsymbol{\varphi})$: A
        translation model with parameter $\boldsymbol{\varphi}$. $f(\boldsymbol{x}; \boldsymbol{\theta})$: a large language model with parameter $\boldsymbol{\theta}$.
Output: $\mathcal{D}_{\text{train}}$: a training dataset for current training epoch.
$\mathcal{D}_{\text{train}} = \{\}$
for $s \in A$ do
    $\mathcal{D}^s_{\text{mono}} \subset \mathcal{D}^A_{\text{mono}}$ // Extract a $s$-specific monolingual subset
    for $t \in A$ do
        $\mathcal{D}_{\text{para}} \leftarrow \mathcal{D}^{s\rightarrow t}_{\text{para}} \cup \mathcal{D}^{t\rightarrow s}_{\text{para}}$
        $\mathcal{D}^s_{\text{para}} \subset \mathcal{D}_{\text{para}}$ // Extract the $s$-centric parallel subset
        if $|\mathcal{D}^s_{\text{para}}| < 25,000$ then
            // The quantity of 25,000 determined by the machine's memory capacity
            $\mathcal{D}^s_{\text{En}} \subset \mathcal{D}_{\text{En}}$, s.t. $|\mathcal{D}^s_{\text{En}}| = 25,000 - |\mathcal{D}^s_{\text{para}}|$ // Extract an English subset for $s$ language
            $\mathcal{D}^{s\rightarrow t}_{\text{En}} \leftarrow g(\boldsymbol{x}; \boldsymbol{\varphi})$ or $\mathcal{D}^{t\rightarrow s}_{\text{En}} \leftarrow g(\boldsymbol{x}; \boldsymbol{\varphi})$, where $\boldsymbol{x} \in \mathcal{D}^s_{\text{En}}$
            $\mathcal{D}^{s\rightarrow t}_{\text{aug}}, \mathcal{D}^{t\rightarrow s}_{\text{aug}}$ // using dictionary to augment
            $\mathcal{D}^s_{\text{aug}} \leftarrow \mathcal{D}^{s\rightarrow t}_{\text{aug}} \cup \mathcal{D}^{t\rightarrow s}_{\text{aug}}$
    end
    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \mathcal{D}^s_{\text{mono}} \cup \mathcal{D}^s_{\text{para}} \cup \mathcal{D}^s_{\text{aug}}$
end
```

gualism of LLMs from the vocabulary sharing aspect. Zhao et al. (2024b) delve into the architecture of LLMs to find how LLMs handle multilingualism. Li et al. (2024) quantify the multilingual performance of LLMs. These studies provide valuable insights into the multilingual capabilities of LLMs, and the key technical design of continued pre-training for XLLaMA2.

3 Training Data Construction

To build a powerful translation model based on LLMs that supports translation across a hundred languages, it is crucial to collect and construct a sufficient amount of data.

3.1 Components of Training Data

During the continued pertaining stage, the collected training data covering 102 languages (refer to A, which are all languages supported by Flores-101), mainly consists of two parts: monolingual ($\mathcal{D}^A_{\text{mono}}$) and parallel ($\mathcal{D}^A_{\text{para}}$) data. For languages with limited data availability, we generated a pseudo-parallel dataset ($\mathcal{D}_{\text{aug}}$) with multilingual dictionaries: MUSE (Lample et al., 2018) and PanLex (Wang et al., 2022). Details regarding the supported languages, dataset description, and data statistics can be found in Appendix B.

Monolingual Data ($\mathcal{D}^A_{\text{mono}}$). Our monolingual training data includes 94 languages supported by Flores-101 from MC4 (Xue et al., 2021) and MADLAD (Kudugunta et al., 2024), totaling 40,000,000 sentences. To ensure efficient handling and processing of the data, we take a strategy where each piece of monolingual data is split into multiple entries, with a block size of 512.

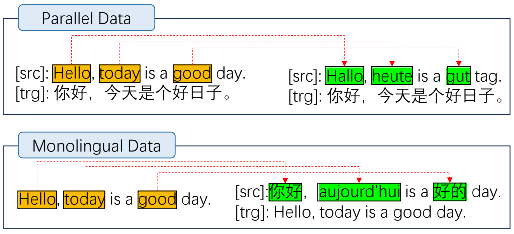
Parallel Data
[src]: Hello, today is a good day.
[trg]: 你好，今天是个好日子。
[src]: Hallo, heute is a gut tag.
[trg]: 你好，今天是个好日子。
Monolingual Data
Hello, today is a good day.
[src]:你好, aujourd'hui is a 好的 day.
[trg]: Hello, today is a good day.

Figure 2: A case illustrating the detailed process of constructing pseudo-parallel data using multilingual dictionary from monolingual or parallel data sources.

Parallel Data ($\mathcal{D}^A_{\text{para}}$). Our parallel data from Lego-MT (Yuan et al., 2023a) encompasses 102 languages, forming a total 4737 language pairs and 9474 translation directions. For each translation direction, denoted as source language (s) to target language (t), we concatenate each translation set, merely using a space as a delimiter, to form a single entry for training data. For each language pair, the probability of occurrence for each translation direction, for example, $s \rightarrow t$ and $t \rightarrow s$ is set as 50%. During the training stage, the gradient is computed for the entire data entry, rather than solely for the target sentence. For language pairs that have fewer than 25,000 (bound by machine resources) sentence pairs, we replicate the original data thrice (Muennighoff et al., 2023).

Data Generated Through Augmentation ($\mathcal{D}_{\text{aug}}$). The way to obtain code-switch data consists of two steps: 1) build multilingual lexicons; 2) construct

# New Token				ro						bn		
	fertility	cosine	R@1	shift distance	# shift token	spBLEU	fertility	cosine	R@1	shift distance	# shift token	spBLEU
0	2.25	0.39	0.37	0.4708	112	**32.50**	8.62	0.17	0.01	0.4689	112	**20.12**
100	2.19	0.36	0.34	0.4720	112	28.75	4.96	0.14	0.02	0.4680	113	14.02
800	2.02	0.35	0.36	0.4682	113	27.78	3.21	0.13	0.02	0.4706	113	10.18
1600	1.93	0.34	0.34	0.4690	113	26.40	2.78	0.13	0.02	0.4695	113	1.82
6,400	1.74	0.31	0.31	0.4694	113	22.66	2.15	0.12	0.02	0.4712	113	1.96
12,800	1.63	0.29	0.29	0.0205	1	21.95	1.95	0.12	0.02	-	0	1.84
25,600	1.53	0.27	0.28	-	0	19.72	1.80	0.12	0.02	-	0	2.58
51,200	1.45	0.26	0.25	0.0203	1	17.79	1.70	0.12	0.03	-	0	1.14

Table 1: Building upon LLaMA2, we add varying numbers of languages-specific new tokens, fully fine-tune LLaMA2, and test the translation performance of en→ro (bn) using Flores-101 test. Furthermore, we assess the effect of new tokens using several metrics: fertility, the cosine similarity with English sentence embeddings, the performance in the English language retrieval translation task (R@1), and the distribution shift of the original embedding vector. Our experiments demonstrate that the inclusion of new words significantly complicates the learning process, underscoring that the integration of new words is a complex task.

pseudo-parallel data. We show the data augmentation process in Figure 2.

Step 1: Building multilingual lexicons. The existing multilingual dictionaries, MUSE and PanLex, encompass multiple bilingual dictionaries, such as en-fr, en-de, en-zh bilingual dictionaries. A dictionary comprises numerous entries, each being a word or a term defined, usage, and provided with other relevant information. We iterate through each entry in the bilingual dictionary, reformat all entries, and create entries in the format of *{entity}_{language}*. For instance, the English word "hello" as translation in three bilingual dictionaries (en-fr, en-de, en-zh), leading us to construct a multilingual lexicons entry as "hello_en, Bonjour_fr, Hallo_de, 你好_zh".

Step 2: Constructing pseudo-parallel data. The foundational data for construction can be based on either parallel or monolingual data, as shown in Figure 2. For each sentence, we convert it to lowercase and subsequently divide it into multiple words using spaces (for Chinese sentences, the Jieba tokenizer is utilized). In parallel data processing, words in a source sentence are randomly replaced with synonyms from a different language using the multilingual dictionary created in Step 1. During the training, the loss is computed solely on the target sentence. In monolingual data processing, each word is individually replaced with a randomly chosen word from the multilingual dictionary. If no suitable replacement word in another language is found, the original word remains unchanged. Consequently, the modified sentence and the original sentence can form pseudo-parallel data. During the training, the loss is computed solely on both the source and the target sentence.

We further conduct an experimental analysis in Section 4, and find the augmentation based on parallel data outperforms that on monolingual data. Therefore, the data augmentation is merely based on parallel data during continued pretraining.

3.2 Training Algorithm.

Given an LLM $f(\boldsymbol{x};\boldsymbol{\theta})$ on a collected training data $\{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$, where $\boldsymbol{\theta}$ is the pre-trained parameters, our objective is to obtain an LLM through continue pre-training, denoted as $f(\boldsymbol{x};\boldsymbol{\theta}')$. Here, $\boldsymbol{\theta}'$ signifies the updated parameters. The target of $f(\boldsymbol{x};\boldsymbol{\theta}')$ is to preserve the model's general capabilities on high-resource languages, while simultaneously enhancing the translation performance across all translation directions among 102 languages. The process of constructing training data is outlined in Algorithm 1. We gather monolingual data for each of the languages and parallel data for every translation direction. Notably, there is no augmentation for translations involving high-resource languages. Instead, we solely augment the translation data that is insufficient by utilizing a trained translation model, Lego-MT (Yuan et al., 2023a). Then we train the $f(\boldsymbol{x};\boldsymbol{\theta})$, the loss function is calculated as:

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{t=1}^{T_i} \log f_{[x_t^{(i)}]}(\boldsymbol{x}_{<t}^{(i)};\boldsymbol{\theta}) \qquad (1)$$

where T is the total decoding time step.

4 Key Technique Design

In this section, we primarily analyze two key challenges related to extending language support: which vocabulary to use (in Section 4.1) and how to perform data augmentation (in Section 4.2). For more detailed analysis results, for discussions on

Setting	spBLEU			# entity				similarity		
	MUSE	PanLex	Δ	MUSE	PanLex	Δ	ratio	MUSE	PanLex	Δ
en→ta	3.74	3.45	-0.29	139,134	91,652	-47,482	0.66	0.08	0.04	-0.04
en→th	5.45	6.14	0.69	21,567	297,573	276,006	13.80	0.20	0.06	-0.14
en→fr	44.03	43.85	-0.18	139,134	568,428	429,294	4.09	0.31	0.35	0.04
en→zh	14.65	16.64	1.99	139134	1,333,762	1,194,628	9.59	0.14	0.09	-0.05
en→es	26.98	27.36	0.38	142,780	433,468	290,688	3.04	0.28	0.32	0.04

Table 2: Evaluate a specific data augmentation technique with different dictionaries. We measure translation performance (spBLEU), the number of target language entities in the dictionary (# entity), and average cosine similarity of entities (similarity), revealing a strong correlation between performance and "# entity".

the selection of multi-hop translation in the lexicon (in Appendix F) and the format of parallel data during the continued pre-training (in Appendix G).

4.1 Preserving the Original Vocabulary.

Setting We conduct a series of analytical experiments on the LLaMA2 vocabulary. Our initial focus is on examining the correlation between fertility and the quality of token representation. Here, fertility refers to the ratio of the length of the token sequence produced by the LLaMA2 tokenizer to the length of the input sentence when split by spaces. Furthermore, we carry out experiments using 10,000 en→ro and en→bn bilingual sentence pairs from Lego-MT. In each experiment, we introduce a varying number of language-specific new tokens and evaluate each model on the Flores-101.

Setting	Aug	en-centric		ta-centric		th-centric		zh-centric	
		en→X	X→en	ta→X	X→ta	th→X	X→th	zh→X	X→zh
LLaMA2	✗	18.31	23.61	0.99	0.49	4.83	1.15	10.02	7.35
$\mathcal{D}_{\mathrm{P}_1}$	✗	19.06	25.98	3.20	0.91	7.66	3.13	11.32	7.83
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}_{\mathrm{P}_2}$	✗	19.46	26.40	4.17	1.76	7.28	3.02	11.65	8.82
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}_{\mathrm{M}}$	✗	19.22	25.91	3.51	1.34	7.64	2.83	11.56	7.99
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}_{\mathrm{P}_2}+\mathcal{D}_{\mathrm{M}}$	✗	19.36	26.47	4.35	1.82	7.78	3.49	11.44	9.14
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}'_{\mathrm{P}_2}$	✓	19.47	26.65	4.54	1.83	7.66	3.13	11.89	9.17
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}'_{\mathrm{M}}$	✓	18.59	25.98	3.61	1.36	6.72	2.35	10.81	6.45
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}'_{\mathrm{P}_2}+\mathcal{D}_{\mathrm{M}}$	✓	**19.70**	**26.71**	4.68	1.82	**8.21**	**3.65**	**12.05**	**9.28**
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}_{\mathrm{P}_2}+\mathcal{D}'_{\mathrm{M}}$	✓	19.17	26.58	4.57	**1.95**	7.12	3.12	11.52	7.73
$\mathcal{D}_{\mathrm{P}_1}+\mathcal{D}'_{\mathrm{P}_2}+\mathcal{D}'_{\mathrm{M}}$	✓	18.80	26.56	**4.78**	1.79	7.31	3.18	11.35	7.28

Table 3: A comprehensive analysis of data augmentation sources reveals that using a dictionary to augment parallel data alone improves performance. "Aug" refers to whether or not a dictionary is used for augmentation.

High fertility lowers the representation quality. We assess the quality of LLaMA's multilingual representation by en→x translation task. This task identifies the translated result that best aligns with the corresponding English sentence within an extensive target dataset, and evaluates with R@1, which is commonly employed in information retrieval. A higher R@1 value signifies a more robust quality of the representation. Concurrently, we present the cosine similarity of representations generated by LLaMA2 for identical sentences in English and other languages. On experiments across 102 languages, more details in Appendix D, there exists a strong correlation between fertility and the quality of representation, evidenced by a Spearman correlation coefficient of approximately **-0.88** for each assessed quality metric.

Adding new tokens to reduce fertility does not yield immediate performance improvements. Extending vocabulary is a common method to support more languages. However, this strategy may not yield the desired results on LLMs. Simply adding new tokens with semantic average initialization (Dobler and de Melo, 2023) can lead to an increase in the input dimension of attention layers without necessarily improving its ability to capture and generalize linguistic patterns across multiple languages. As shown in Table 1, the more new tokens added, the worse the translation performance.

New tokens have a significant impact on model performance. As demonstrated in Table 1, even the addition of a small number (100) of new language-specific tokens can have a significant impact on the multilingual performance of LLMs. In addition, we conduct a further analysis on the original tokens (32k) embedding distribution and the token number before and after adding new tokens by KS-Lottery (Yuan et al., 2024). For more details on KS-Lottery, refer to Appendix E. As the experimental result of "shift distance" and "# shift token" in Tabel 1, fine-tuning the entire model with limited new tokens follows a similar pattern to that with the original vocabulary. However, an excessive number of new tokens can shift the model's training focus. This holds true regardless of whether the language (ro) is well-supported by the model or not (bn). The influence of these additional tokens is substantial, indicating that the process of enhancing the multilingual capabilities of LLMs is not as straightforward as simply expanding the vocabulary and training with more multilingual data.

Maintaining the vocabulary suffices to boost

the multilingualism of LLMs. The LLaMA tokenizer, which utilizes the Byte-level Byte Pair Encoding (BBPE; Wang et al., 2019) algorithm, is the foundation for multilingual language processing tasks. Its universal compatibility across all languages, in conjunction with the absence of the requirement for an "unknown" token, optimizes vocabulary sharing and improves its robustness. Its allows the model to understand/generate responses in various languages using the same vocabulary. Meanwhile, studies have shown that LLMs trained on unbalanced English-centric datasets, often use English as an internal pivot language. This helps LLMs to map the inputs closer to English in internal space before generating the output (Yoon et al., 2024; Huang et al., 2024b; Zhu et al., 2024b). Maintaining the original vocabulary helps to preserve this behavior, which also benefits for improving the multilingual capability.

4.2 Data Augmentation

Setting Given a parallel dataset subset ($\mathcal{D}_{\mathrm{P}}$) from $\mathcal{D}^{A}_{\mathrm{para}}$ that contains translations in all directions for 6 languages (en,fr,es,zh,ta,th) and a monolingual subset ($\mathcal{D}_{\mathrm{M}}$) from $\mathcal{D}^{A}_{\mathrm{mono}}$ for the same 6 languages. We then perform non-repetitive sampling 12,500 sentence pairs from $\mathcal{D}_{\mathrm{P}}$ in each direction to generate two subsets of parallel corpus data $\mathcal{D}_{\mathrm{P}_1}$ and $\mathcal{D}_{\mathrm{P}_2}$, respectively. Consequently, we preserve $\mathcal{D}_{\mathrm{P}_1}$ and evaluate the effect of augmentation on parallel data $\mathcal{D}_{\mathrm{P}_2}$ or monolingual data $\mathcal{D}_{\mathrm{M}}$, resulting in two new dataset, $\mathcal{D}'_{\mathrm{P}_2}$ and $\mathcal{D}'_{\mathrm{M}}$, post-augmentation. To assess both the in-domain and out-of-domain capabilities of the model, we perform inference on it using 10 languages (en, fr, es, pt, de, zh, ta, th, is, zu), utilizing the Flores-101.

The choice of dictionary is related to the number of entities for the language in the dictionary. As shown in Table 2, there is no clear dictionary preference is observed for en/ta/th/zh-centric translation, with optimal performance randomly distributed across the two dictionaries. Furthermore, we conduct an in-depth analysis of the MUSE and PanLex dictionary for translation from en to another 5 languages. We compare the end-to-end translation performance (spBLEU), the number of target language entities in the dictionary (# entity), and the similarity of entities embedding (simple average with entity token embeddings) extracted from the trained model. And find a clear correlation between the translation performance and #entity.

5 Benchmarking Results

In this section, we present multilingual benchmarking results to comprehensively demonstrate the potential of XLLaMA2. We evaluate translation quality with spBLEU (Goyal et al., 2022) and COMET-22 (Rei et al., 2020) for both LLMs and translation models. See Appendix C for training details on XLLaMA2 and description of baseline models.

We significantly enhances the multilingual translation capabilities of the base LLaMA2 model through massive multilingual continued pre-training. First, we demonstrate the benefits of our continued pre-training in enhancing the base LLM's multilingual translation capabilities. Evaluation results on Flores-101 benchmark are shown in Table 4. By comparing our multilingual-enhanced XLLaMA2 model with the base LLaMA2 model in instruction-tuned versions (XLLaMA2-Alpaca vs. LLaMA2-Alpaca), we consistently observe a significant performance improvement on both English-centric and non-English-centric translation. In addition to Flores-101, we also make evaluation on a range of diverse translation benchmarks (Table 5). The performance enhancement brought by our multilingual continued pre-training is consistent across these benchmarks.

Our constructed model outperforms other open-source decoder-only LLMs on multilingual translation by a large margin. Next, we compare our language-extended XLLaMA2-Alpaca model with other open-source decoder-only LLMs built for multilingual purposes (Table 4, Table 5). Compared to other from-scratch trained LLMs, such as PolyLM, Yayi2, XLLaMA2 consistently shows better performance across various multilingual translation benchmarks, indicating that the LLaMA2 base model provides a strong foundation for language extension. Furthermore, when compared to other LLaMA-based continued pre-trained models, such as TowerInstruct, ChineseLLaMA2-Alpaca, XLLaMA2 also achieves superior performance, demonstrating the effectiveness of our optimized continued pre-training pipeline.

Our performed multilingual continued pre-training benefits unseen long-tail low-resource languages as well. A significant challenge in multilingual enhancement is that the substantial cost of collecting scarce multilingual resources makes it prohibitive to cover massive languages.

System	Size	en-X		zh-X		de-X		ne-X		ar-X		az-X		ceb-X	
		spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET
LLaMA2 (Touvron et al., 2023b)	7B	4.21	43.95	0.91	44.62	2.14	45.26	0.39	38.22	0.54	39.43	0.68	47.43	1.49	33.50
PolyLM (Wei et al., 2023)	13B	5.72	45.16	1.42	52.41	3.59	47.89	0.45	38.00	1.04	45.82	0.57	38.65	0.77	29.74
Yayi2 (Luo et al., 2023)	30B	7.80	54.13	4.38	55.23	4.72	56.48	0.92	47.88	1.73	49.45	1.23	53.06	1.87	36.75
TowerInstruct (Alves et al., 2024)	7B	9.41	58.69	4.15	57.75	6.79	58.31	2.07	51.42	3.35	50.76	1.79	48.01	3.36	41.69
Aya-23 (Aryabumi et al., 2024)	8B	11.18	57.91	7.20	56.65	9.30	55.69	3.50	51.78	8.00	55.49	3.27	51.45	4.24	44.14
ChineseLLaMA2-Alpaca (Cui et al., 2024)	7B	-	-	2.31	49.72	-	-	-	-	-	-	-	-	-	-
LLaMA2-Alpaca (Taori et al., 2023)	7B	9.44	52.83	3.80	51.29	6.82	51.47	1.31	46.59	2.84	46.76	1.36	48.63	2.69	41.02
XLLaMA2-Alpaca	7B	23.17	76.66	14.17	73.54	18.96	73.82	14.49	74.64	15.82	72.00	11.34	70.91	15.53	68.67

System	Size	X-en		X-zh		X-de		X-ne		X-ar		X-az		X-ceb	
		spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET
LLaMA2 (Touvron et al., 2023b)	7B	11.80	55.46	0.55	43.50	3.22	43.10	0.42	34.41	0.25	39.13	0.59	43.98	1.16	41.64
PolyLM (Wei et al., 2023)	13B	7.75	50.98	1.20	42.60	3.69	43.95	0.36	33.69	1.67	42.27	0.44	40.24	0.96	39.29
Yayi2 (Luo et al., 2023)	30B	19.37	68.06	6.07	57.81	5.62	53.82	0.48	40.95	0.52	46.61	0.71	49.29	1.71	45.50
TowerInstruct (Alves et al., 2024)	7B	18.87	65.37	10.37	64.26	12.81	60.73	0.62	38.80	0.39	44.72	0.71	47.17	2.24	47.15
Aya-23 (Aryabumi et al., 2024)	8B	20.57	67.53	11.20	66.11	14.09	63.09	2.69	44.33	11.84	63.59	1.19	46.97	2.29	45.17
ChineseLLaMA2-Alpaca (Cui et al., 2024)	7B	-	-	6.15	55.06	-	-	-	-	-	-	-	-	-	-
LLaMA2-Alpaca (Taori et al., 2023)	7B	16.44	65.85	4.46	56.53	9.01	56.76	1.03	34.96	2.18	44.10	0.63	40.67	1.73	45.69
XLLaMA2-Alpaca	7B	30.63	80.55	13.53	75.52	19.26	74.47	15.47	67.36	15.32	75.40	10.27	72.03	16.11	65.05

Table 4: Benchmarking results on Flores-101 dataset, where X refers to all another 101 languages. This table compares our instruction-aligned XLLaMA2 model (XLLaMA2-Alpaca) with the instruction-aligned LLaMA2 model (LLaMA2-Alpaca) to demonstrate the benefits of our multilingual continued pre-training. Additionally, we compare XLLaMA2 with other open-source multilingual-focus LLMs to highlight the impressive multilingual capabilities of XLLaMA2.

System	Size	TED (en-X)		TED (X-en)		TICO (en-X)		WMT23 (en-X)		WMT23 (X-en)	
		spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET
LLaMA2 (Touvron et al., 2023b)	7B	3.34	52.15	8.66	61.54	3.45	39.63	2.96	51.55	14.87	65.68
PolyLM (Wei et al., 2023)	13B	5.53	50.18	7.28	55.16	7.17	40.36	10.62	62.67	19.09	69.15
Yayi2 (Luo et al., 2023)	30B	8.54	61.53	14.09	70.92	7.91	47.02	10.76	65.69	20.47	75.60
TowerInstruct (Alves et al., 2024)	7B	8.22	64.83	15.29	70.91	10.14	50.48	18.42	74.03	30.03	80.08
Aya-23 (Aryabumi et al., 2024)	8B	10.69	68.06	16.44	72.87	12.98	52.44	27.15	83.29	31.21	82.00
LLaMA2-Alpaca (Taori et al., 2023)	7B	9.15	62.04	12.67	68.62	8.60	44.73	17.23	73.17	24.97	75.82
XLLaMA2-Alpaca	7B	16.12	75.58	17.81	76.18	19.79	68.33	23.91	80.17	30.30	79.55

Table 5: Benchmarking results on WMT23, TED and TICO dataset. X denotes various languages across different translation benchmarks; detailed information is available in Appendix B. Evaluation results across these benchmarks further validate the strong multilingual translation capabilities of XLLaMA2.

System	Size	en-X		zh-X		de-X		ne-X		ar-X		az-X		ceb-X	
		spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET
M2M-100 (Fan et al., 2021)	418M	17.26	63.76	10.13	61.41	14.10	61.62	4.03	46.98	11.52	59.97	4.17	45.75	6.13	44.23
M2M-100 (Fan et al., 2021)	1.2B	21.54	70.00	13.13	67.29	17.73	67.62	7.14	56.04	12.57	62.62	6.06	52.39	9.46	52.79
M2M-100 (Fan et al., 2021)	12B	24.74	74.19	14.91	71.56	20.34	72.07	9.68	62.19	16.36	68.91	6.24	54.78	12.48	60.09
Lego-MT (Yuan et al., 2023a)	1.2B	24.96	69.49	16.28	68.23	21.42	69.20	16.98	68.37	18.38	65.57	13.51	65.69	16.83	58.21
MADLAD-400 (Kudugunta et al., 2024)	7B	31.26	80.62	19.47	76.73	25.05	77.72	18.67	74.32	23.70	77.11	10.70	63.15	16.40	66.39
Aya-101 (Üstün et al., 2024)	13B	26.19	80.66	16.57	78.34	22.44	79.49	19.97	80.91	19.79	77.84	14.05	78.32	20.03	74.47
XLLaMA2-Alpaca	7B	24.81	79.41	15.08	76.07	20.31	76.64	15.52	77.06	16.92	74.43	12.14	73.44	16.59	70.79

System	Size	X-en		X-zh		X-de		X-ne		X-ar		X-az		X-ceb	
		spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET	spBLEU	COMET
M2M-100 (Fan et al., 2021)	418M	21.19	68.47	10.34	62.15	14.25	60.19	1.30	40.43	11.53	63.33	2.44	49.74	4.85	47.80
M2M-100 (Fan et al., 2021)	1.2B	26.26	73.06	12.94	67.91	19.33	67.78	1.40	42.60	8.57	60.28	4.58	55.86	6.83	55.87
M2M-100 (Fan et al., 2021)	12B	28.01	74.45	13.35	69.27	21.31	70.17	2.85	45.50	15.15	69.94	6.44	61.36	8.77	57.07
Lego-MT (Yuan et al., 2023a)	1.2B	30.71	75.44	16.42	71.41	23.75	70.75	15.02	59.66	18.21	70.73	11.88	66.73	15.06	59.28
MADLAD-400 (Kudugunta et al., 2024)	7B	39.98	84.97	21.71	80.35	28.43	79.64	14.37	62.78	23.48	79.66	14.66	77.33	4.30	51.37
Aya-101 (Üstün et al., 2024)	13B	33.64	82.85	16.29	79.14	23.53	80.28	17.86	71.07	17.23	79.37	14.05	80.22	22.34	68.41
XLLaMA2-Alpaca	7B	32.41	82.64	14.25	77.27	20.53	76.66	16.31	68.95	16.13	77.05	10.91	73.92	17.07	65.88

Table 6: Benchmarking results on Flores-101 dataset. Given that the M2M-100 baselines cover only 86 languages (Goyal et al., 2022; Yuan et al., 2023a) from Flores-101, we restrict our model comparisons to 85 languages, denoted as X = 85. This table compares our instruction-aligned XLLaMA2 model (XLLaMA2-Alpaca) with other multilingual translation model with encoder-decoder architecture to demonstrate we are closing the gap between decoder-only LLM and traditional encoder-decoder translation systems.

	Knowledge			Commonsense Reasoning		Math Reasoning		Code		Avg.
	MMLU	BBH	NQ	HellaSwag	Winogrande	GSM8K	Math	HumanEval	MBPP	
LLaMA2-Alpaca	44.22	37.95	24.32	31.12	61.09	14.03	3.82	14.63	27.63	28.76
XLLaMA2-Alpaca	44.60	38.25	23.21	33.75	61.48	12.21	3.74	12.20	25.29	28.30

Table 7: Evaluation results on monolingual general benchmarks.

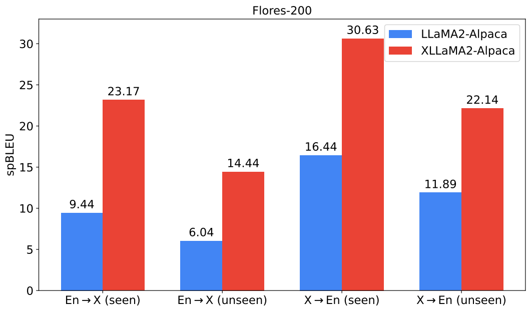

Figure 3: Comparison results between XLLaMA2-Alpaca and LLaMA2-Alpaca on Flores-200. Although multilingual continued pre-training does not cover all non-English languages in Flores-200, but it also boosts model's translation performance on these languages.

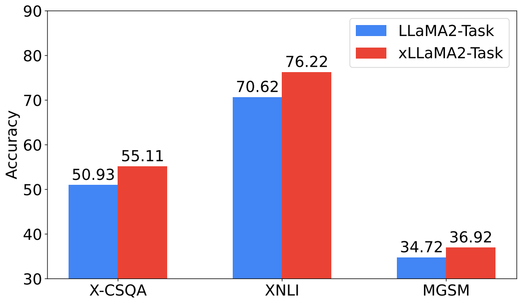

Figure 4: Comparison results between instruction-tuning our multilingual enhanced model and the base model with specialized instruction data. We take X-CSQA, XNLI, MGSM as three examples tasks.

While our multilingual pre-training corpus already covers 102 languages, we acknowledge that there remains a large group of long-tail, low-resource languages that are not well covered. To assess the generalization capability of XLLaMA2, we evaluate it on Flores-200 dataset and observe its performance on these unseen languages (Figure 3). We find that for languages not encountered during training, XLLaMA2 still achieves significant improvements, demonstrating the generalization capability of our massive continued pre-training.

Our multilingual-enhanced model is closing the performance gap between open-source LLM translator and specialized encoder-decoder translation systems. While XLLaMA2 has achieved the state-of-the-art translation performance among open-source decoder-only LLMs, the next critical question is whether we can close the gap between LLMs and specialized encoder-decoder translation systems. Table 6 provides a comprehensive comparison, reveals XLLaMA2 has reached the level of the M2M100-12B model. Future work will be needed to optimize the language extension framework to match the performance of advanced translation systems, e.g., MADLAD-400.
Our continued pre-trained XLLaMA2 model provides a better starting point for specialized instruction-tuning In the end, we demonstrate the usage of our continued pre-trained model (XLLaMA2) on tasks beyond translation. While in previous experiments we use basic Alpaca instruction data to teach LLM to follow translation instructions, we now show that our released checkpoint can be enpowered to handle more multilingual tasks beyond translation. Figure 4 presents three example tasks where we use specialized instruction data to unlock XLLaMA2's abilities on specific tasks, such as math reasoning and common sense reasoning. We find that the instruction-tuned XLLaMA2 model outperforms its LLaMA2 model counterpart in non-English performance across all three tasks, demonstrating that provides a better starting point for specialized instruction-tuning.

Our performed multilingual continued pre-training does not cause catastrophic forgetting issue. A common concern with continued pre-training on additional multilingual corpus is that the process might disturb the parametric knowledge and working pattern of the original model, a phenomenon known as catastrophic forgetting (Goodfellow et al., 2013). Furthermore, we compare XLLaMA2 with LLaMA2 on popular English benchmarks that measure a diverse set of core capabilities of LLMs. Experiment results in Table 7 show that the two models achieve very similar performance on these benchmarks overall, demonstrating that our continued pre-training does not compromise the English capability of the base model.

6 Conclusion

In this work, we enhance LLaMA2's translation performance for 102 languages through continued pre-training, creating XLLaMA2. We compare XLLaMA2 's translation capabilities with other decoder-only LLMs and encoder-decoder models across multiple benchmarks. XLLaMA2 is also assessed on general tasks and fine-tuned with task-specific instructions. Our results indicate that XLLaMA2 improves translation quality while maintaining general capabilities, indicating XLLaMA2 is an ideal foundation for downstream tasks.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

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

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

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

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model?

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024a. A survey on large language models with multilingualism: Recent advances and new frontiers.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024b. Mindmerger: Efficient boosting llm reasoning in non-english languages.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. 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 1274–1287. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021b. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Yin Luo, Qingchao Kong, Nan Xu, Jia Cao, Bao Hao, Baoyu Qu, Bo Chen, Chao Zhu, Chenyang Zhao, Donglei Zhang, et al. 2023. Yayi 2: Multilingual open-source large language models. *arXiv preprint arXiv:2312.14862*.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

OpenAI. 2023. Gpt-4 technical report.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

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. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords.

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

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models.

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

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023a. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Fei Yuan, Chang Ma, Shuai Yuan, Qiushi Sun, and Lei Li. 2024. Ks-lottery: Finding certified lottery tickets for multilingual language models.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2023b. How multilingual is multilingual llm?

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism?

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024a. A preference-driven paradigm for enhanced translation with large language models.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv*, abs/2304.04675.

Outline

- Section A: Discussion about the generalizability and reliability of the work.
- Section B: The comprehensive details of the training data, including monolingual and parallel data, and the evaluation benchmark (Table 8).
- Section C: The detailed information of different models, including open-source Large Language Models (Section C.1) and well-trained translation models (Section C.2).
- Section D: Analysis the correlation between embedding quality of LLaMA2 and fertility using Flores-101 test (Figure 5).
- Section E: A detailed introduction to the KS-Lottery method.
- Section F: Selection about multi-hop translation (Table 9).
- Section G: The selection of the appropriate format for parallel data during training (Table 10).
- Section H: The comparison of translation performance across all seven languages between Lego-MT and GPT-4 (Figure 6).
- Section I: Information about use of AI assistants.

A Limitations

This work focuses on the discussion of some key technologies, such as the use of vocabulary lists and the determination of data augmentation schemes. However, it does not delve into further processing of the quality of open-source data. We acknowledge a gap in the literature regarding the thorough evaluation of open-source data quality, suggesting an opportunity for future research to improve data preprocessing methods for better model training outcomes.

B Collected Dataset Information

In this section, we will introduce the sources of our training data (Section B.1), the evaluation benchmarks (Section B.2). For translation tasks, we apply beam search to each model with beam size=4.

B.1 Training Dataset

The dataset was compiled from three distinct open-source datasets, with details on data statistics and supported languages presented in the Table 8.

MC4 (Xue et al., 2021) is a multilingual variant of the C4 dataset, comprising natural text in 101 languages sourced from the Common Crawl web scrape. It was introduced to support the training of massively multilingual pre-trained text-to-text transformers like mT5.

MADLAD-400 (Kudugunta et al., 2024) is a manually audited, general domain monolingual dataset based on CommonCrawl, encompassing 419 languages and designed for document-level analysis. It is notable for its extensive language coverage and the rigorous auditing process involved in its creation.

Lego-MT (Yuan et al., 2023a) is a benchmark for massively multilingual machine translation, featuring a detachable model built upon an efficient training recipe. It includes a comprehensive translation benchmark with data from OPUS, covering 433 languages and 1.3 billion parallel data points.

B.2 Evaluation Benchmark

Flores-101 (Goyal et al., 2022) is a benchmark for machine translation evaluation, comprising a multi-way dataset derived from English Wikipedia and produced by professional translators.

Flores-200 (Team et al., 2022) is an extension of Flores-101 dataset and also serves as a benchmark for machine translation. This dataset contains parallel sentences for 200 languages, with each language identified by its ISO 639-3 code ((e.g. eng)) and an additional code (e.g., "eng_Latn",) that describes the script.

WMT-23 (Kocmi and Federmann, 2023) is also a comprehensive translation evaluation benchmark, proposed in 2023. We incorporate this dataset into our evaluation to mitigate the risk of data leakage in LLMs. Based on benchmark, we evaluate the English-centric translation task performance, including de→en, en→cs, en→de, en→he, en→ja, en→ru, en→uk, en→zh, he→en, ja→en, ru→en, uk→en, zh→en.

TICO (Anastasopoulos et al., 2020) dataset represents a joint translation effort targeting COVID-19 materials, developed in collaboration with academic, industry stakeholders, and Translators without Borders. It comprises translation memories, a glossary of translated COVID-19 terms, and functions as a benchmark for translation-related evaluations. The all evaluated translation is en→{am, bn,

Family	ISO	Language	# Mono.	# Para.	# Direct.	Family	ISO	Language	# Mono.	# Para.	# Direct.
Afro-Asiatic	ha	Hausa	420,964	3,147,704	96	Indo-European	ne	Nepali	702,334	8,907,527	97
	om	Oromo	18,895	191,319	96		or	Odia	100,530	812,235	97
	so	Somali	697,864	3,804,551	97		pa	Punjabi	513,987	3,737,780	97
	am	Amharic	269,171	4,031,552	97		sd	Sindhi	472,217	821,996	95
	ar	Arabic	716,063	9,940,756	97		ur	Urdu	711,354	4,137,619	97
	he	Hebrew	300,000	3,928,938	96		fa	Persian	721,307	4,111,536	97
	mt	Maltese	671,716	1,518,533	94		ku	Kurdish	517,239	3,597,863	97
Austroasiatic	km	Khmer	687,690	4,044,652	97		ps	Pashto	588,340	3,717,480	97
	vi	Vietnamese	760,472	4,112,089	97		tg	Tajik	700,237	4,131,709	97
Austronesian	jv	Javanese	505,619	2,799,761	97		ast	Asturian	0	1,535,714	96
	id	Indonesian	707,962	4,243,235	97		ca	Catalan	724,597	4,145,004	97
	ms	Malay	711,895	4,121,713	97		es	Spanish	706,307	4,258,477	98
	mi	Maori	180,678	3,437702	97		fr	French	787,316	4,290,003	99
	ceb	Cebuano	418,058	2,217,926	91		gl	Galician	726,512	3,131,730	96
	tl	Tagalog	0	3,927,576	97		it	Italian	846,107	4,233,108	96
Dravidian	te	Telugu	708,459	4,219,702	97		oc	Occitan	36,379	1,752,951	95
	kn	Kannada	712,832	3,592,636	97		pt	Portuguese	795,818	4,258,604	97
	ml	Malayalam	715,387	4,516,012	97		ro	Romanian	702,002	4,219,414	97
	ta	Tamil	711,863	4,444,734	97	Japonic	ja	Japanese	726,455	4,207,728	97
Indo-European	hy	Armenian	712,835	3,677,780	97	Kartvelian	ka	Georgian	703,515	4,182,651	97
	lt	Lithuanian	718,382	3,946,735	96	Koreanic	ko	Korean	711,406	4,234,653	97
	lv	Latvian	700,889	4,011,628	97	Kra–Dai	lo	Lao	357,758	2,642,799	97
	be	Belarusian	708,288	4,169,719	95		th	Thai	707,719	4,437,476	97
	bg	Bulgarian	711,500	4,131,053	97	Mongolic	mn	Mongolian	701,304	3,894,353	97
	bs	Bosnian	300,000	2,953,912	97	Niger–Congo	wo	Wolof	871	802,521	97
	cs	Czech	711,179	4,135,944	97		ln	Lingala	3,325	159,684	96
	hr	Croatian	300,000	4,106,335	97		ns	Northern Sotho	0	96,288	88
	mk	Macedonian	702,035	4,009,787	97		lg	Luganda	13,030	216,135	95
	pl	Polish	792,829	4,200,001	98		ny	Nyanja	226,940	3,104,349	92
	ru	Russian	853,407	4,204,365	97		sn	Shona	386,588	3,140,063	97
	sk	Slovak	715,540	4,100,272	98		sw	Swahili	700,422	3,775,394	97
	sl	Slovenian	731,613	4,073,213	97		umb	Umbundu	0	54	2
	sr	Serbian	711,535	4,033,130	97		xh	Xhosa	122,720	3,955,426	97
	uk	Ukrainian	714,181	4,070,250	97		yo	Yoruba	98,281	3,364,040	96
	cy	Welsh	703,507	3,777,953	97		zu	Zulu	470,403	2,899,738	97
	ga	Irish	693,460	2,814,912	96		ig	Igbo	147,319	3,314,731	96
	is	Icelandic	704,159	4,088,886	97		kam	Kamba	0	8	1
	sv	Swedish	726,893	4,213,939	97		ff	Fulani	26	313,870	97
	da	Danish	721,543	4,194,587	97	Nilo-Saharan	luo	Dholuo	0	91	6
	no	Norwegian	721,715	4,045,571	97	Portuguese	kea	Kabuverdianu	0	0	0
	af	Afrikaans	703,546	4,143,358	98	Sino-Tibetan	zh	Chinese	726,112	14,215,583	96
	de	German	881,553	10,273,597	97		zhtrad	Chinese	0	3,747,297	96
	en	English	846,712	19,548,583	100		my	Burmese	579,160	3,887,841	97
	lb	Luxembourgish	574,166	1,035,619	94	Turkic	uz	Uzbek	723,096	2,344,375	95
	nl	Dutch	769,778	4,199,773	96		kk	Kazakh	701,849	3,836,259	97
	el	Greek	707,751	4,081,607	97		ky	Kyrgyz	704,438	3,725,583	97
	bn	Bengali	707,099	4,560,978	97		az	Azerbaijani	712,947	8,080,151	97
	as	Assamese	33,825	1,656,861	97		tr	Turkish	727,711	4,169,259	97
	gu	Gujarati	704,619	3,761,401	97	Uralic	et	Estonian	706,720	4,056,200	97
	hi	Hindi	715,691	4,186,127	97		fi	Finnish	719,416	40,76,885	97
	mr	Marathi	702,382	4,295,708	97		hu	Hungarian	731,479	4,154,132	97

Table 8: The detailed information of the collected monolingual and parallel datasets includes the translation directions for each supported language. Specifically, the "# Para." represents the count of language-centric sentence pairs, while "# Mono" denotes the number of individual monolingual sentences.

din, fa, fuv, hi, km, ku, ln, ms, ne, om, ps, ru, so, ta, ti_ER, tl, zh, ar, ckb, es_LA, fr, ha, id, kr, lg, mr, my, nus, prs, pt_BR, rw, sw, ti, ti_ET, ur, zu}.

TED (Cettolo et al., 2012) is a massively multilingual dataset derived from TED Talk transcripts, covering 60 languages with parallel arrays of language and text. It is designed for natural language processing tasks and filters out missing or incomplete translations. We also evaluate the English-centric translation performance. The translation direction covers all 60 languages, including en↔{af, am, ar, arq, art-x-bork, as, ast, az, be, bg, bi, bn, bo, bs, ca, ceb, cnh, cs, da, de, el, eo, es, et, eu, fa, fi, fil, fr, fr-ca, ga, gl, gu, ha, he, hi, hr, ht, hu, hup, hy, id, ig, inh, is, it, ja, ka, kk, km, kn, ko, ku, ky, la, lb, lo, lt, ltg, lv, mg, mk, ml, mn, mr, ms, mt, my, nb, ne, nl, nn, oc, pa, pl, ps, pt, pt-br, ro, ru, rup, sh, si, sk, sl, so, sq, sr, srp, sv, sw, szl, ta, te, tg, th, tl, tlh, tr, tt, ug, uk, ur, uz, vi, zh, zh-cn, zh-tw}

X-CSQA (Lin et al., 2021a) is a multilingual extension of the Commonsense Question Answering (CSQA) dataset, designed for commonsense reasoning research. It facilitates the evaluation and improvement of multilingual language models in commonsense reasoning tasks.

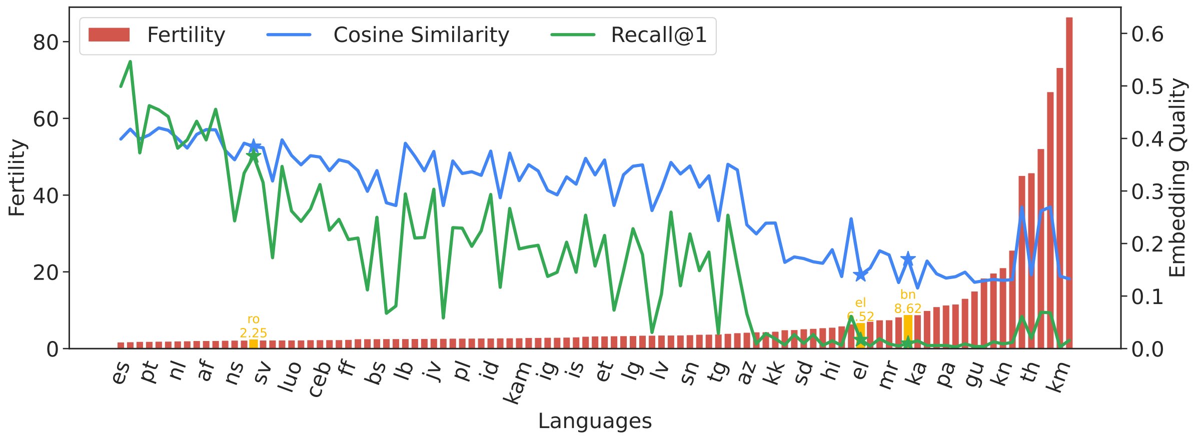

Figure 5: Correlation between embedding quality and fertility. The embedding quality of LLaMA2 is measured by cosine similarity and Recall@1 on Flores-101 test. Fertility refers to the ratio of the length of a sentence after tokenization compared to its length before tokenization. A high fertility may result in a poor quality of embedding.

XStoryCloze (Lin et al., 2021b) is a benchmark dataset that comprises the professionally translated English StoryCloze dataset (Spring 2016 version) into 10 non-English languages. It is designed to evaluate the zero- and few-shot learning capabilities of multilingual language models.

XCOPA (Ponti et al., 2020) is a benchmark dataset that assesses machine learning models' ability to transfer commonsense reasoning across languages. It is an extension of the English COPA dataset and includes 11 languages from diverse language families and geographical regions.

XWinograd (Muennighoff et al., 2022; Tikhonov and Ryabinin, 2021) s a benchmark dataset that consists of a multilingual collection of Winograd Schemas, designed for the evaluation of cross-lingual commonsense reasoning capabilities covering six languages.

XNLI (Conneau et al., 2018) is a cross-lingual extension of the SNLI (Bowman et al., 2015)/MultiNLI (Williams et al., 2018), consisting of a subset of English examples translated into 14 different languages. It is used for evaluating textual entailment and classification tasks, where the goal is to determine if one sentence implies, contradicts, or is neutral to another sentence

MGSM (Shi et al., 2023) a dataset of grade-school math problems, each translated into 10 languages by human annotators. It is derived from the GSM8K (Cobbe et al., 2021) dataset and is designed to support question answering on basic mathematical problems that require multi-step reasoning.

MMLU (Hendrycks et al., 2021a,b) is a benchmark for evaluating language models' capabilities in language comprehension and reasoning across diverse domains. It consists of about 16,000 multiple-choice questions spanning 57 academic subjects, designed to measure knowledge acquired during pretraining in zero-shot and few-shot settings.

BBH (Srivastava et al., 2022) is a subset of the BIG-Bench, focusing on 23 challenging tasks that current language models struggle to perform, where they do not outperform the average human-rater. It serves as a rigorous evaluation suite to test the limits of language models' capabilities.

HellaSwag (Zellers et al., 2019) s a dataset designed to evaluate advanced natural language understanding and common sense reasoning, which introduces more complexity and diversity, challenging AI models to predict the ending of incomplete narratives.

WinoG (Sakaguchi et al., 2021) is a large-scale dataset containing 44k problems inspired by the Winograd Schema Challenge, designed to improve the scale and hardness of coreference resolution tasks. It presents fill-in-the-blank questions with binary options, testing the model's ability to understand nuanced human language.

NQ (Kwiatkowski et al., 2019) is a dataset for question answering research, containing over

Setting	Dictionary	en-centric		ta-centric		th-centric		zh-centric	
		en→x	x→en	ta→x	x→ta	th→x	x→th	zh→x	x→zh
$\mathcal{D}_{\mathrm{P}_1}$+$\mathcal{D}'_{\mathrm{P}_2}$+$\mathcal{D}'_{\mathrm{M}}$	MUSE: 1-hop	18.80	26.56	4.78	1.79	7.31	3.18	11.35	7.28
$\mathcal{D}_{\mathrm{P}_1}$+$\mathcal{D}'_{\mathrm{P}_2}$+$\mathcal{D}'_{\mathrm{M}}$	MUSE: 2-hop	18.70	26.50	4.47	1.83	7.08	3.26	10.74	6.68
$\mathcal{D}_{\mathrm{P}_1}$+$\mathcal{D}'_{\mathrm{P}_2}$+$\mathcal{D}'_{\mathrm{M}}$	PanLex: 1-hop	19.33	26.54	4.40	1.83	7.57	3.31	10.86	8.08

Table 9: Select a specific data augmentation technique and evaluate various dictionary configurations, including 1-hop and 2-hop, as well as different dictionaries.

300,000 examples each consisting of a real user query and a corresponding Wikipedia page. It is designed to train and evaluate automatic question answering systems by emulating how people search for information.

HumanEval (Chen et al., 2021) is designed to evaluate the code generation capabilities of large language models, featuring 164 hand-crafted programming challenges that include function signatures, docstrings, bodies, and unit tests. On average, each problem is accompanied by 7.7 tests to assess functional correctness.

MBPP (Austin et al., 2021) comprises approximately 1,000 crowd-sourced Python programming problems, aimed at entry-level programmers and covering programming fundamentals and standard library functionality. Each problem includes a task description, code solution, and three automated test cases.

GSM8K (Cobbe et al., 2021) consists of 8.5K high-quality, linguistically diverse grade school math word problems created by human problem writers. It is designed to support question answering on basic mathematical problems that require multi-step reasoning.

Math (Hendrycks et al., 2021c) is a collection of 12,500 intricate problems derived from competition mathematics. Every problem within the Math dataset includes a comprehensive solution with step-by-step guidance, which serves as a resource for training models to produce detailed answer justifications and explanations.

C Detailed Information of Used Models

Model details about the baseline models for comparison, including decode-only large language models (LLMs) in Section C.1 as well as translation models in Section C.2 with an encoder-decoder structure.

C.1 Large Language Models

XLLaMA2 follows the model architecture of LLaMA2 without vocabulary extension. We utilized 20 A100 80GB GPUs and extended the pretraining on the amassed data for over 60 days. We set per device training batch size to 32, learning rate to 2e-5, and the epoch number to 1.0.

LLaMA2 (?) is a decoder-only language model that predicts the next token based on the input sequence of ordered tokens, with a collection of pre-trained and fine-tuned models ranging from 7 billion to 70 billion parameters. The LLaMA2 7B model serves as our foundational model. Unless otherwise specified, any reference to LLaMA or LLaMA2 is the LLaMA2 7B model. The model leverages a Byte-level Byte Pair Encoding (BBPE; (Wang et al., 2019)) tokenizer, an efficient subword tokenizer that tokenizes at the byte level, allowing it to handle any language and be robust to noise in the data. The BBPE tokenizer is particularly useful for languages with large vocabularies and many rare words.

PolyLM (Wei et al., 2023) is an open-source multilingual Large Language Model (LLM) trained on 640 billion tokens, available in two model sizes: 1.7B and 13B. It boasts proficiency in 15 major non-English languages, employing advanced training techniques to enhance its language processing capabilities.

Yayi2 (Luo et al., 2023) is a multilingual open-source Large Language Model pre-trained from scratch on a corpus containing 2.65 trillion tokens. It is aligned with human values through supervised fine-tuning and reinforcement learning from human feedback.

TowerInstruct (Alves et al., 2024) is a 7B parameter language model fine-tuned on translation-related tasks, supporting multiple languages including English, Portuguese, Spanish, French, and

Setting	Translation Tasks		General Tasks			Multilingual Tasks		
	ceb→x	**x→ceb**	**QNLI**	**QQP**	**MRPC**	**XStoryCloze**	**XCOPA**	**XWinograd**
splited-parallel + mono	3.36	2.74	49.46	36.82	68.38	59.20	56.82	73.72
connected-parallel + mono	4.45	3.68	49.46	36.82	68.38	59.10	56.80	74.07
Setting	**ceb→ca**	**ceb→de**	**ceb→en**	**ceb→es**	**ceb→fr**	**ceb→it**	**ceb→pt**	**ceb→ru**
splited-parallel + mono	10.32	8.94	23.19	13.30	15.96	10.01	12.66	8.05
connected-parallel + mono	10.97	11.37	27.06	14.91	18.04	12.03	15.55	10.26
Setting	**ca→ceb**	**de→ceb**	**en→ceb**	**es→ceb**	**fr→ceb**	**it→ceb**	**pt→ceb**	**ru→ceb**
splited-parallel + mono	5.90	4.91	7.44	5.14	6.02	5.54	6.12	4.24
connected-parallel + mono	7.62	6.92	9.88	6.41	7.39	6.91	7.62	6.54

Table 10: Design for the utilization of parallel data, we take ceb-centric data as an example, apply two distict approaches, and discover that treating parallel data as two independent monolingual datasets harms to translation performance.

others. It is designed for tasks such as machine translation, automatic post-editing, and paraphrase generation.

Aya-23 (Aryabumi et al., 2024) is an open weights research release of an instruction fine-tuned decoder-only model with advanced multilingual capabilities, serving 23 languages. It pairs a performant pre-trained Command family of models with the Aya Collection for robust language processing tasks.

ChineseLLaMA2-Alpaca (Cui et al., 2024) is founded on LLaMA2 and enhanced with an extensive Chinese vocabulary that concentrates on Chinese languages. This is a fine-tuned version of ChineseLLaMA2 using Alpaca (Taori et al., 2023) data.

LLaMA2-SFT (Taori et al., 2023) is a fine-tuned version of LLaMA2 model, leveraging a set of 52,000 diverse instructions in Alpaca (Taori et al., 2023) to enhance the instruction-following capabilities of the model.

C.2 Translation Models

M2M-100 (Fan et al., 2021) encompasses multilingual machine translation models designed to translate between any pair of 100 languages directly, without the need for English as an intermediary. The M2M-100 series includes models of varying sizes, specifically 418M, 1.2B, and 12B parameters. These models are part of a groundbreaking approach in the field of machine translation, aiming to enhance direct translation efficiency across a wide array of languages.

Lego-MT (Yuan et al., 2023a) is a novel approach to massively multilingual machine translation, featuring detachable models with individual branches for each language or group of languages. This design supports plug-and-play training and inference, enhancing flexibility and efficiency in language processing tasks.

MADLAD-400 (Kudugunta et al., 2024) is a multilingual machine translation model that leverages the T5 architecture and has been trained on a vast corpus of 250 billion tokens, covering over 450 languages.

Aya-101 (Aryabumi et al., 2024) is an open-source, massively multilingual generative language model that operates on the mT5 (Xue et al., 2021) architecture, covering 101 languages and designed to bridge the performance gap in non-dominant languages. It incorporates a 13B parameter base and has undergone instruction-finetuning to achieve high performance across its extensive language range.

D The correlation between fertility and representation quality.

We conduct experiments on Flores-101. Fertility is defined as the ratio of the L_s to the L_T, where L_s is the number of words for space-separated languages and characters for others and L_T is the number of tokens after applying LLaMA2 tokenizer. The quality estimation of LLaMA on Flores-101 test. Cosine similarity focuses on the similarity in the expressions of LLaMA across sentence representation of the same sentence in English and other languages. Recall@1 is often used in the context of information retrieval, which measures the quality of representation. The experimental results, as shown in Figure 5, indicate fertility has a high correlation with the representation quality.

E Introduction to KS-Lottery.

KS-Lottery is a technique designed to identify a small, highly effective subset of parameters within LLMs for multilingual capability transfer. The core concept of this method involves utilizing the Kolmogorov-Smirnov Test to examine the distribution shift of parameters before and after fine-tuning. This approach helps in pinpointing the “winning tickets” or the most impactful parameters that contribute significantly to the model’s performance in multilingual tasks.

F 1-hop translation in data augmentation is enough.

Given a parallel dataset subset ($\mathcal{D}_{\mathrm{P}}$) from $\mathcal{D}^{A}_{\mathrm{para}}$ that contains translations in all directions for 6 languages (en,fr,es,zh,ta,th) and a monolingual subset ($\mathcal{D}_{\mathrm{M}}$) from $\mathcal{D}^{A}_{\mathrm{mono}}$ for the same 6 languages. We then perform non-repetitive sampling 12,500 sentence pairs from $\mathcal{D}_{\mathrm{P}}$ in each direction to generate two subsets of parallel corpus data $\mathcal{D}_{\mathrm{P}_1}$ and $\mathcal{D}_{\mathrm{P}_2}$, respectively. Consequently, we preserve $\mathcal{D}_{\mathrm{P}_1}$ and evaluate the effect of augmentation on parallel data $\mathcal{D}_{\mathrm{P}_2}$ or monolingual data $\mathcal{D}_{\mathrm{M}}$, resulting in two new dataset, $\mathcal{D}'_{\mathrm{P}_2}$ and $\mathcal{D}'_{\mathrm{M}}$, post-augmentation. To assess both the in-domain and out-of-domain capabilities of the model, we perform inference on it using 10 languages (en, fr, es, pt, de, zh, ta, th, is, zu), utilizing the Flores-101.

We use two different multilingual dictionaries MUSE provided by Lample et al. (2018) [1], and PanLex (Wang et al., 2022). In the context of a multilingual dictionary, we can use “1-hop” and “2-hop” to characterize the translation relationship among different languages, an example shown in Table 9.

1-hop translation		2-hop translation	
Direction	Example	Direction	Example
en→fr fr→de	dog → chien chien → Hund	en→fr→de	dog → chien → Hund

Table 11: Case of 1-hop and 2-hop translations.

We use the MUSE dictionary to perform data augmentation on both parallel $\mathcal{D}_{\mathrm{P}_2}$ and monolingual $\mathcal{D}_{\mathrm{M}}$ data, utilizing 1-hop and 2-hop translations. As shown in Table 9, using different hop translation for augmentation does not significantly impact the final translation performance. Multi-hop transaltion sometimes can even result in poorer performance.

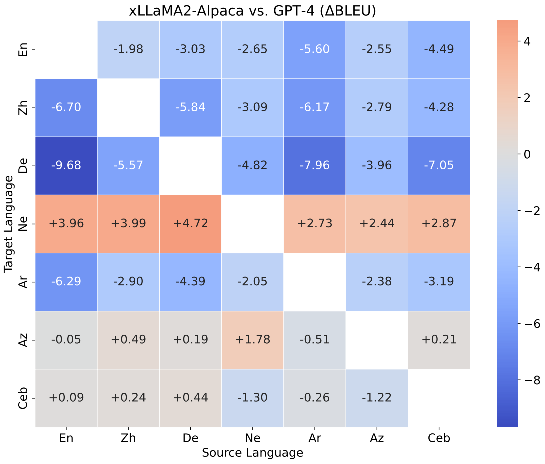

Figure 6: The spBLEU gap between XLLaMA2 and GPT-4. Positive scores mean the result of XLLaMA2 is better than GPT-4. Empirical evidence demonstrates that while XLLaMA2 trails GPT-4 in high-resource translation scenarios, it outperforms in low-resource translation contexts.

G Design of parallel format

The Usage of Parallel Data. Parallel data can be utilized in two distinct ways: split-parallel or connected-parallel. **Split-Parallel**: Consider the source language data and target language data involved in parallel data as two distinct monolingual datasets, which are randomly shuffled throughout the entire training set. **Connected-Parallel**: In the training process, we treat each pair of source and target language sentences from the parallel dataset as a single data point by concatenating them.

Based on different forms of parallel data, supervised fine-tuning (SFT) is conducted separately on ceb-centric using both parallel and monolingual datasets. As indicated in Table 10, we observed that the form of parallel data primarily impacts translation performance, with no significant difference in general tasks and cross-lingual general tasks; however, the disparity in translation is pronounced. We specifically highlighted some high-resource translation directions and found that such gaps are quite significant.

H Comparison Results Between Our Model and GPT-4

In Figure 6, we compare the performance gap between our model and GPT-4. Considering the API cost of evaluating GPT-4, we only evaluate the

[1] https://github.com/facebookresearch/MUSE.

mutual translation performance among seven languages (En, Zh, De, Ne, Ar, Az, Ceb). Experiment results show that while our model lags behind in high-resource translation directions, it achieves on-par or even superior performance in low-resource translation.

I Information about use Of AI assistants

AI assistants are utilized to refine sentence-level writing.