

Just Go Parallel: Improving the Multilingual Capabilities of Large Language Models

Muhammad Reza Qorib, Junyi Li, and Hwee Tou Ng

Department of Computer Science, National University of Singapore
mrqorib@u.nus.edu, junyi_cs@nus.edu.sg, nght@comp.nus.edu.sg

Abstract

Large language models (LLMs) have demonstrated impressive translation capabilities even without being explicitly trained on parallel data. This remarkable property has led some to believe that parallel data is no longer necessary for building multilingual language models. While some attribute this to the emergent abilities of LLMs due to scale, recent work suggests that it is actually caused by incidental bilingual signals present in the training data. Various methods have been proposed to maximize the utility of parallel data to enhance the multilingual capabilities of multilingual encoder-based and encoder-decoder language models. However, some decoder-based LLMs opt to ignore parallel data instead. In this work, we conduct a systematic study on the impact of adding parallel data on LLMs’ multilingual capabilities, focusing specifically on translation and multilingual common-sense reasoning. Through controlled experiments, we demonstrate that parallel data can significantly improve LLMs’ multilingual capabilities.¹

1 Introduction

To democratize the benefits of large language models (LLMs) for the whole world, many initiatives have been undertaken to build LLMs that possess multilingual capabilities (Scao et al., 2022; Sen-gupta et al., 2023). Multilingual capabilities enhance the accessibility and inclusivity of the model and help reduce its inherent bias (Zhu et al., 2024a; Navigli et al., 2023).

Even without being explicitly trained with parallel data, LLMs are reported to have impressive translation capabilities (Radford et al., 2019; Lin et al., 2022). Briakou et al. (2023) reported that the translation capabilities of LLMs are highly correlated with the bilingual signals in their training

data, particularly translation pairs. The presence of bilingual signals is often incidental, meaning they are not deliberately added to the training data. When these bilingual signals are removed, LLMs lose the capability to translate between English and languages with non-Latin scripts.

While some multilingual LLMs are trained with parallel data (Alves et al., 2024), some are not (Scao et al., 2022), despite having parallel data sources such as OPUS-100 (Zhang et al., 2020) or EuroParl (Koehn, 2005) in their training data. While parallel texts are not always available especially for extremely low-resource languages, the situation today has improved dramatically compared to a decade ago. Due to efforts such as NLLB (No Language Left Behind; Costa-jussà et al. 2022), many more languages – even for languages that were previously considered low-resource like Indonesian – now have enough publicly available parallel texts to make a difference in building multilingual large language models, as we demonstrate in this paper.

Many encoder-based (Conneau and Lample, 2019; Ouyang et al., 2021) and encoder-decoder (Liu et al., 2020; Chi et al., 2021a) language models aim to improve their multilingual capabilities through parallel corpora. Much work has been proposed to specifically enhance cross-lingual alignments (Cao et al., 2020; Luo et al., 2021), even when parallel sentences are unavailable (Lu et al., 2023). On the other hand, decoder language models often deliberately ignore the parallel data available in their training corpus, such as by not including the English portion of the parallel data or randomly mixing all the training data together. This situation warrants a systematic investigation of the effect of parallel data on large language models’ multilingual capabilities.

In this research, we conduct a systematic study to investigate whether adding parallel data helps in enhancing a large language model’s multilingual

¹Source code, checkpoints, and data are available at: <http://github.com/nusnlp/just-go-parallel>

capabilities and what the best strategy is for incorporating them into a language model’s training data.

Our main contributions are as follows.

1. We report that adding parallel data to the training data is more effective than adding unrelated monolingual corpora of other languages for enhancing the multilingual capabilities of decoder-based large language models.
2. We find that training the model with parallel data at the end of the training process is the most effective approach for improving multilingual performance, while training at the beginning of the process leads to serious catastrophic forgetting and wastes the parallel data.
3. We report that LLMs do not have the ability to perform translation in the opposite direction of what they were trained on.
4. We show that the amount of bilingual signal in the training data affects LLMs’ translation capability.
5. We release the code, checkpoints, and training data of our models to facilitate further study.

2 Related Work

2.1 Multilingual Language Models

Benefiting from the complete open-sourcing of data and code from many LLMs, such as BLOOM (Scao et al., 2022), LLM360 (Liu et al., 2024), and Falcon (Almazrouei et al., 2023), many studies have begun developing multilingual language models based on these LLMs by incorporating multilingual data for continual pre-training (Conneau and Lample, 2019; Xue et al., 2021; Lin et al., 2022). Although recent instruction-following LLMs are primarily pre-trained and fine-tuned on a limited number of resource-rich languages, they have demonstrated substantial multilingual comprehension and generation capabilities (Touvron et al., 2023; Wang et al., 2024; Niklaus et al., 2023). However, due to the limitation of imbalanced training data distribution (Yang et al., 2023), these multilingual language models still fall short in those languages with scarce resources (Qin et al., 2024). To understand how language models acquire multilingual capabilities, Nezhad and Agrawal (2024) investigate factors influencing the performance of multilingual language models and reveal that script type and

language family are critical for unseen languages, highlighting the importance of cross-lingual transfer learning. Additionally, research by Tang et al. (2024) aims to explain the underlying mechanisms by which LLMs process multilingual texts. Their research indicates that the proficiency of LLMs in processing languages is predominantly due to a small subset of neurons, primarily situated in the models’ top and bottom layers.

2.2 Enhancing Multilingual Capabilities

To enhance the multilingual capabilities of LLMs, one line of work is cross-lingual transfer, where the capability of one language can be transferred to other languages (Huang et al., 2023; Etxaniz et al., 2024; Ranaldi et al., 2024). By designing cross-lingual alignment prompting that instructs LLMs to self-translate a question into another language (Qin et al., 2023) or utilize an external machine translation system (Zhao et al., 2024), the capabilities of language generation and instruction-following can be transferred to a non-English language. Besides, several efforts have been devoted to knowledge distillation on synthetic data from high-resource languages to low-resource ones (Chai et al., 2025; Al-Maamari et al., 2024; Zhang et al., 2024b). Another line of work is cross-lingual alignment, which involves constructing alignment data and loss functions to align mid- and low-resource languages with those that are resource-rich (Schuster et al., 2019; Wen-Yi and Mimno, 2023; Zhu et al., 2024b). For example, Chi et al. (2021b) introduce a denoising word alignment pre-training task that predicts masked tokens in another language. Mao and Yu (2024) leverage the capabilities of LLMs to translate previously unsupported languages for building aligned data, overcoming the weak cross-lingual signals caused by data scarcity.

3 Method

Large language models are often trained on internet-sourced corpora predominantly in English (Biderman et al., 2023; Groeneveld et al., 2024). We investigate whether incorporating parallel data enhances multilingual capabilities by comparing it to training without parallel data or with monolingual corpora from other languages. Our objective is to empirically examine how the inclusion of parallel data in the training set affects multilingual LLMs’ performance across multiple languages.

To ensure a controlled comparison, we maintain the order and quantity of training data across all experiments. When parallel data are added to the training set, an equivalent amount of non-parallel data from the last portion is removed to preserve a consistent training set size and order.

We also explore the optimal placement of parallel data within the language model’s training process. Specifically, we experiment with inserting parallel data at the beginning of training, distributing it throughout the training data, and adding it at the end. We define seven experimental settings: NO PARALLEL, MULTILINGUAL, PARALLEL NON-ADJACENT, PARALLEL FIRST, PARALLEL DISTRIBUTED, PARALLEL LAST (ALL), and PARALLEL LAST (UNI), as detailed below.

3.1 No Parallel

In this setting, no parallel data are added to the training data. This approach mirrors the typical strategy used for constructing English-centric LLMs, such as Pythia (Biderman et al., 2023) and TinyLlama (Zhang et al., 2024a). Although no parallel data are intentionally included, the training data may incidentally contain some parallel texts sourced from the internet. However, according to our language detection analysis (Table 8), the amount of such data is minimal.

3.2 Multilingual

In this setting, monolingual data from other languages are distributed uniformly throughout the non-parallel data. This approach resembles common strategies for building multilingual language models by incorporating monolingual corpora from multiple languages (Lu et al., 2023; Scao et al., 2022). To control for the choice of text, the monolingual data added are derived from the non-English half of the parallel data. That is, in our experiments, monolingual data in Indonesian and Chinese are added in this setting, and the equivalent amount of English data from the last portion of the NO PARALLEL setting is removed. Note that the English half of the parallel data are not included in this setting. This setup evaluates whether a language model can learn cross-lingual mappings independently without explicitly aligning semantically equivalent sentences across languages.

3.3 Parallel Non-Adjacent

In this setting, parallel data for all language pairs are uniformly distributed throughout the non-

parallel data. However, an English sentence and its translation are not placed next to each other. Instead, the English sentences in the parallel data are randomly shuffled, such that each non-English sentence from the parallel data is followed by a random English sentence. This setup evaluates whether the presence of semantically equivalent sentences in the training data, but without explicitly placing a sentence next to its translation, helps in learning cross-lingual mappings.

3.4 Parallel First

In this setting, parallel data for all translation directions are introduced at the beginning of training. The parallel data are presented as adjacent sentence pairs, where a sentence and its translation are placed next to each other. The rationale behind this setup is that early exposure to parallel data may help the model establish cross-lingual mappings, enabling it to better leverage incidental bilingual signals present in the non-parallel training data.

3.5 Parallel Distributed

In this setting, parallel data for all translation directions, presented as adjacent sentence pairs, are distributed throughout the non-parallel data. Since large language models are believed to acquire multilingual capabilities from bilingual signals (Briakou et al., 2023), particularly translation pairs, this setup aims to amplify such signals.

3.6 Parallel Last (all)

In this setting, parallel data for all translation directions, presented as adjacent sentence pairs, are added at the end of training. This approach can also be interpreted as second-stage training, where the model receives bilingual exposure after being pre-trained primarily on English data.

3.7 Parallel Last (uni)

This setting is similar to PARALLEL LAST (ALL), except that the model is trained on only one translation direction (e.g., English to Chinese). For each translation direction, a separate model is trained, resulting in specialized models rather than a general multilingual LLM.

4 Experiments

We build our model based on TinyLlama, a 22-layer transformer decoder LLM with 1.1B parameters. We use a subset of SlimPajama (Soboleva et al.,

Corpus	# sents	# tokens	
		Chinese	English
ParaCrawl	14.2M	620.6M	357.9M
NewsComm	0.3M	19.8M	10.3M
Wiki Titles	0.9M	6.1M	10.9M
UN Parallel	15.9M	999.4M	579.9M
WikiMatrix	2.6M	150.0M	83.5M
Total	33.9M	1,795.9M	1,042.5M

Table 1: Statistics of Chinese-English training parallel data.

Corpus	# sents	# tokens	
		Indonesian	English
Total	54.1M	1,222.0M	883.5M

Table 2: Statistics of Indonesian-English training parallel data.

2023) as our non-parallel training data. When referring to non-parallel data, we mean SlimPajama data, which are predominantly in English and do not deliberately include parallel data. The subset of SlimPajama used consists of 82.35% English, 0.19% Indonesian, 0.12% Chinese, and 17.34% other languages (See Table 8 in the Appendix).

We measure the language model’s multilingual capabilities by evaluating its translation and common-sense reasoning performance. In our experiments, we focus on English (EN), Simplified Chinese (ZH), and Indonesian (ID) as case studies. We selected Chinese for its script diversity and Indonesian for its mid-resource status. To incorporate parallel data into our training set, we use widely adopted parallel corpora for machine translation tasks. Specifically, we use the training data from the WMT-2022 general (news) machine translation task (Kocmi et al., 2022) for Chinese-English translation pairs, excluding the CCMT corpus² (Table 1). For Indonesian-English translation pairs, we use the training data from the WMT-2021 large-scale multilingual machine translation task (Wenzek et al., 2021) (Table 2).

For each translation pair, we format the parallel data as plain text using the template: “{source language}: {source sentence}\n{target language}: {target sentence}”. This approach is inspired by the format of incidental parallel data found in PALM’s (Chowdhery et al., 2023) training set, as reported by Briakou et al. (2023).

²The CCMT corpus requires registration, but we did not receive a response after submitting our registration request.

By default, we alternate the translation direction of the language pairs (e.g., EN \rightarrow ID, ID \rightarrow EN, ZH \rightarrow EN, EN \rightarrow ZH). Text sequences are concatenated using the end-of-sentence pseudo-token `<\s>` and split into chunks, each with a size equal to eight times the context window. Leftover text segments that are shorter than this are discarded. After pre-processing, the parallel data amount to 4.5B tokens³.

We train the model on up to 167B tokens⁴ using NVIDIA H100 GPUs. Due to the high computational cost, each experiment is conducted only once. We save checkpoints every 5,000 steps (~5.2B tokens), but for the first 5,000 steps of the PARALLEL FIRST setting and the last 5,000 steps of the PARALLEL LAST settings, we save checkpoints every 1,000 steps to analyze the effects of parallel data in a more fine-grained manner.

After training, we select the checkpoint with the highest average BLEU score (Papineni et al., 2002) across all translation directions on the development set. We use the WMT-2022 test set for Chinese-to-English and English-to-Chinese translation and the Flores-200 dev set (Costa-jussà et al., 2022) for Indonesian-to-English and English-to-Indonesian translation as the development set.

4.1 Evaluation

We assess the model’s translation performance using the BLEU score metric from SacreBLEU (Post, 2018). Machine translation performance is evaluated on the Chinese-to-English and English-to-Chinese test sets of WMT-2023⁵ (Semenov et al., 2023) and on the devtest set of Flores-200 for Indonesian-to-English and English-to-Indonesian. Since Flores-200 does not provide test sets for both directions, we reverse the translation directions during evaluation. We have verified that none of the evaluation data are included in the training data, ruling out the possibility of data leakage.

³Throughout this paper, the number of tokens is consistently measured using the TinyLlama tokenizer, which shares the same vocabulary as the Llama 2 models.

⁴All experimental settings are trained on 167B tokens, except the PARALLEL LAST (ALL) (166B tokens) and PARALLEL LAST (UNI) (164B tokens) settings. The PARALLEL LAST (ALL) setting uses 162B tokens of non-parallel data and 4B tokens of parallel data (due to saving the checkpoints every 1B tokens, not all parallel data are consumed), while the PARALLEL LAST (UNI) setting uses 162B tokens of non-parallel data and 2B tokens of parallel data (EN \leftrightarrow ZH or EN \leftrightarrow ID parallel data only, instead of all available parallel data).

⁵<https://www.statmt.org/wmt23/translation-task.html>

Model	Param	EN → ID	ID → EN	EN → ZH	ZH → EN
Zero-Shot					
No Parallel	1.1B	2.49	1.52	0.80	1.30
Multilingual	1.1B	2.38	5.92	0.81	3.72
Parallel Non-Adjacent	1.1B	1.98	14.69	1.01	4.50
Parallel First	1.1B	7.42	5.57	9.64	2.71
Parallel Distributed	1.1B	21.95	27.48	12.08	7.40
Parallel Last (all)	1.1B	35.91	35.36	9.62	10.73
Parallel Last (uni)	1.1B	44.19	41.91	28.51	16.10
BLOOM	1.1B	2.19	18.39	2.27	4.58
NLLB	1.3B	44.64	43.06	27.58	19.25
Few-Shot (5 examples)					
No Parallel	1.1B	2.60	0.97	0.79	0.83
Multilingual	1.1B	2.49	9.75	1.73	4.25
Parallel Non-Adjacent	1.1B	3.31	13.90	3.76	2.69
Parallel First	1.1B	25.41	21.02	18.61	7.09
Parallel Distributed	1.1B	31.80	33.66	23.21	12.51
Parallel Last (all)	1.1B	41.51	39.08	32.61	15.20
Parallel Last (uni)	1.1B	44.32	41.60	33.31	16.87
BLOOM	1.1B	21.92	27.29	18.10	10.72

Table 3: Translation performance (BLEU scores) of each experimental setting, along with the number of parameters in the model (PARAM). NLLB is specialized for machine translation, so we do not evaluate its few-shot performance.

We perform zero-shot and few-shot evaluations using 5 examples. The evaluation follows the code and prompt style of ALMA (Xu et al., 2024) (Table 11 in the Appendix). Statistical significance is measured using the paired approximate randomization method with 10,000 trials and a significance threshold (p -value) of 0.05.

We also evaluate the model’s common-sense reasoning performance in English, Chinese, and Indonesian using a zero-shot approach across several benchmarks. For English, we test on ARC (Easy and Challenge) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), BoolQ (Clark et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SciQ (Welbl et al., 2017), and WinoGrande (Sakaguchi et al., 2020). For Chinese, we use the Chinese subsets of XWinograd (Tikhonov and Ryabinin, 2021), XStoryCloze (Lin et al., 2022), XNLI (Conneau et al., 2018), and XCOPA (Ponti et al., 2020). Since XWinograd and XNLI do not have Indonesian subsets, Indonesian common-sense reasoning is evaluated using the Indonesian subsets of XStoryCloze and XCOPA. We utilize the Language Model Evaluation Harness (LM-Eval) framework⁶ for common-

sense reasoning evaluation.

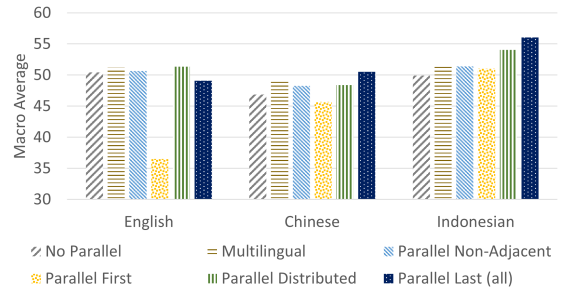


Figure 1: Macro-average of the common-sense reasoning benchmarks.

5 Results

We find that training the model with parallel data at the end yields the best translation performance, especially when the model is trained in only one language direction (Table 3). Adding parallel data at the beginning provides only a minor improvement over the NO PARALLEL strategy in zero-shot evaluation, but the difference becomes much more pronounced in few-shot evaluation. Few-shot inference also greatly improves the scores of PARALLEL

⁶<https://github.com/EleutherAI/lm-evaluation-harness/>

n-harness/

Model	ARC _C	ARC _E	BoolQ	HS	OBQA	PIQA	SciQ	WG	Avg
No Parallel	25.51	45.08	60.12	46.30	32.00	68.66	72.50	53.35	50.44
Multilingual	27.30	46.93	57.31	48.25	32.40	69.15	74.00	53.91	51.16
Parallel Non-Adjacent	25.43	46.21	60.15	47.67	30.20	68.23	73.20	54.06	50.64
Parallel First	21.84	30.56	41.07	25.73	25.00	52.88	44.60	50.28	36.50
Parallel Distributed	26.45	47.05	60.61	46.49	31.60	68.88	76.30	53.28	51.33
Parallel Last (all)	24.40	41.79	60.21	42.21	30.20	66.10	73.80	53.91	49.08
BLOOM	25.60	45.41	59.11	42.97	29.40	67.25	74.60	55.01	49.92

Table 4: English common-sense reasoning performance in each experimental setting, measured based on the model’s accuracy on ARC Challenge, ARC Easy, BoolQ, HellaSwag, OpenBookQA, PIQA, SciQ, and WinoGrande.

DISTRIBUTED and PARALLEL LAST (ALL), narrowing the gap between PARALLEL LAST (ALL) and PARALLEL LAST (UNI). Notably, the PARALLEL DISTRIBUTED setting significantly outperforms MULTILINGUAL and PARALLEL NON-ADJACENT. The PARALLEL DISTRIBUTED and PARALLEL NON-ADJACENT settings are trained on the same data for the same duration, differing only in how the parallel data are placed in the training data. The PARALLEL DISTRIBUTED setting even outperforms BLOOM, despite BLOOM being trained with more Indonesian and Chinese data – 20 GB of Indonesian and 261 GB of Chinese texts, compared to less than 5 GB of texts in each language in our setup. This highlights the effectiveness of parallel data in enhancing LLMs’ translation capabilities.

Beyond translation, the PARALLEL DISTRIBUTED setting also outperforms the MULTILINGUAL and PARALLEL NON-ADJACENT settings on English and Indonesian common-sense reasoning benchmarks while maintaining comparable performance on Chinese common-sense reasoning benchmarks (Figure 1). The PARALLEL LAST (ALL) setting significantly outperforms the other experimental settings in Chinese and Indonesian, although it shows a slight decline in English performance, as shown in Table 4. Detailed scores for Chinese and Indonesian common-sense reasoning are provided in Table 12 and Table 13, respectively, in the Appendix. These findings indicate that parallel data not only improves translation performance but also enhances LLMs’ common-sense reasoning abilities in non-English languages.

6 Discussions

6.1 All Directions vs. Unidirectional Translation

As shown in Table 3, training with unidirectional data leads to higher translation performance compared to training with data in all translation directions, although the unidirectional approach creates specialized models instead of a single multilingual model. These models cannot translate languages they are not trained on, and surprisingly, they also fail to translate in the opposite direction of the language pair they are trained on (Table 5). Their performance is very poor, even worse than models trained without parallel data. We found that they often produce nonsensical outputs due to this limitation. This phenomenon may be related to the issue of autoregressive LLMs struggling with inverse relationships, dubbed the *reversal curse* (Berglund et al., 2024).

Adding few-shot examples does not help unidirectional models to translate other languages or the opposite direction of the same language pair. Even in the same translation direction as their training data, the performance gain is modest. In contrast, the few-shot performance of the PARALLEL LAST (ALL) model is quite close to the best performance of each unidirectional model while retaining the ability to translate between all language pairs.

On the common-sense reasoning task, models trained with unidirectional parallel data directed *into* the evaluated language perform slightly worse than the model trained with parallel data in all directions (PARALLEL LAST (ALL)). For English, models trained only on ID → EN or ZH → EN parallel data achieve lower scores than the PARALLEL LAST (ALL) model (Table 6). A similar trend holds for Chinese and Indonesian: training on only EN → ZH and EN → ID parallel data results in worse

performance on Chinese and Indonesian reasoning tasks, respectively.

While unidirectional training may not be suitable for building multilingual large language models with general capabilities, it is highly effective for developing machine translation models. By fine-tuning a large language model (pretrained primarily on English data) with just one epoch of parallel data, we can achieve very high translation performance—exceeding a BLEU score of 41—especially for languages that use the same script as English, such as Indonesian.

Parallel Data	Translation Evaluation			
	EN → ID	ID → EN	EN → ZH	ZH → EN
Zero-Shot				
All directions	35.91	35.36	9.62	10.73
EN → ID	44.19	0.07	0.77	0.21
ID → EN	0.02	41.91	0.25	0.03
EN → ZH	0.09	0.59	28.51	0.01
ZH → EN	0.00	2.73	0.13	16.10
None	2.49	1.52	0.80	1.30
Few-Shot (5 examples)				
All directions	41.51	39.08	32.61	15.20
EN → ID	44.32	0.14	0.72	0.11
ID → EN	0.04	41.60	0.43	0.01
EN → ZH	0.04	2.51	33.31	0.00
ZH → EN	2.58	2.71	0.06	16.87
None	2.60	0.97	0.79	0.83

Table 5: Translation performance (BLEU scores) of adding parallel data at the end of training. Training with parallel data of all directions is synonymous with the PARALLEL LAST (ALL) training setup, while using no parallel data (NONE) is synonymous with the NO PARALLEL training setup. BLEU scores on the diagonal correspond to the PARALLEL LAST (UNI) training setup.

6.2 Quality of Translation Pairs

Next, we investigate the effect of translation pair quality on LLMs’ translation capability. Parallel data are sometimes collected automatically, leading to varying degrees of quality. In traditional machine translation systems, noisy translation pairs are often filtered out from the training data (Lowphansirikul et al., 2020). However, LLMs are typically trained on massive datasets that are inherently noisy. In this analysis, we examine whether filtering the parallel data could further enhance the multilingual performance of LLMs.

Parallel Data	Common-Sense Reasoning		
	English	Indonesian	Chinese
All directions	49.08	56.06	50.55
EN → ID	48.77	55.34	45.87
ID → EN	48.73	55.87	45.47
EN → ZH	49.80	49.96	50.49
ZH → EN	48.49	50.02	50.45

Table 6: Macro-average scores on common-sense reasoning benchmarks for different translation directions of parallel training data. Training with parallel data of all directions is synonymous with the PARALLEL LAST (ALL) training setup.

Parallel Data	Translation Evaluation			
	EN → ID	ID → EN	EN → ZH	ZH → EN
Zero-Shot				
PARALLEL LAST (ALL)	35.91	35.36	9.62	10.73
with filtering	36.20	35.17	24.20	10.60
Few-Shot (5 examples)				
PARALLEL LAST (ALL)	41.51	39.08	32.61	15.20
with filtering	40.50	37.71	32.28	14.83

Table 7: Translation performance (BLEU scores) of LLMs in relation to the quality of the parallel data.

We assess the quality of the parallel data using CometKiwi-2022 (Rei et al., 2022), a state-of-the-art quality estimation model for machine translation. CometKiwi-2022 produces a score between 0 and 1, indicating the quality of each translation pair. Based on manual observation, we set a threshold of 0.42 for Chinese-English pairs and 0.58 for Indonesian-English pairs. Applying this filter reduces the number of parallel sentences from 33.9M to 25M for Chinese and from 54.1M to 15.6M for Indonesian. In this setting, we keep the amount of non-parallel data constant, resulting in the filtered model being trained on slightly less total data.

We conduct this experiment using the PARALLEL LAST (ALL) strategy. In zero-shot evaluation, we observe a significant improvement in English-Chinese translation performance and a slight improvement in English-Indonesian translation performance, but not in the other language pairs (Table 7). In few-shot evaluation, the performance of the filtered model is slightly worse than that of the unfiltered model. This suggests that at this scale, LLMs are quite resilient to noisy data.

6.3 Catastrophic Forgetting

Adding parallel data at the beginning (PARALLEL FIRST) improves LLMs’ translation capability, but few-shot examples are needed to achieve considerable performance (Table 3). However, this performance is measured using the best checkpoint on the development set. When examining the performance at later checkpoints, we observe that the translation capability completely disappears once the parallel data are exhausted (Figure 2). The only difference between the PARALLEL FIRST, PARALLEL DISTRIBUTED, and PARALLEL LAST strategies is the position of the parallel data within the training set, yet placing it in the wrong position can completely negate its benefits. We attribute this to catastrophic forgetting (McCloskey and Cohen, 1989). Similar patterns are observed in other translation directions.

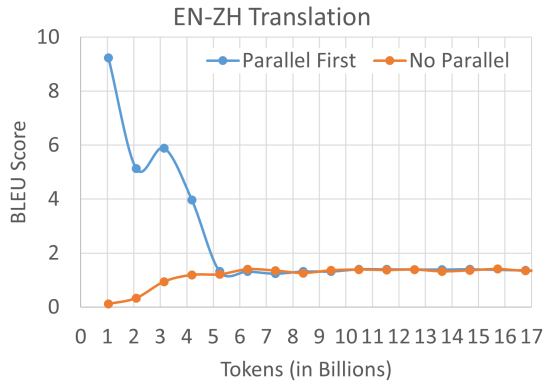


Figure 2: Progression of the BLEU score on English-Chinese translation between NO PARALLEL and PARALLEL FIRST experimental settings.

6.4 Impact of Incidental Bilingual Signals

In this section, we investigate how incidental bilingual signals in the training data affect the LLMs’ translation performance. Briakou et al. (2023) reported that incidental bilingual signals, especially parallel text, in predominantly English training data contribute to LLMs’ translation abilities. However, they did not examine the relationship between the frequency of these bilingual signals and translation performance. In our analysis, we aim to extend their findings by exploring how the frequency of bilingual signals in our subset of SlimPajama influences the translation performance of models trained without parallel data.

We detect bilingual signals by first building a word translation dictionary by aligning words in the

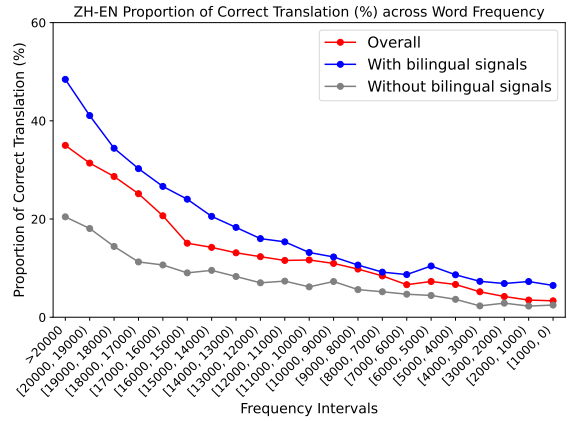


Figure 3: Chinese-to-English translation performance across word frequency in the 167B training data.

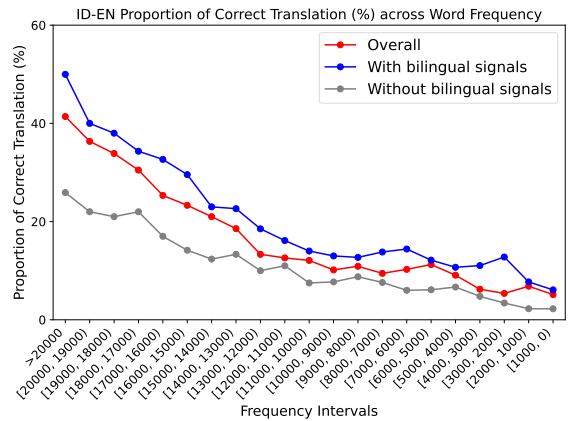


Figure 4: Indonesian-to-English translation performance across word frequency in the 167B training data.

source and target sentences of the WMT-2022 test set (for Chinese-English) and the Flores-200 dev set (for Indonesian-English) using SimAlign (Sabet et al., 2020). Before processing with SimAlign, we perform word segmentation using jieba⁷ for Chinese and NLTK (Bird and Loper, 2004) for Indonesian and English. For each word in these test sets, we examine its context and frequency in the training data using Elasticsearch. We then categorize the words into two groups: words with bilingual signals and words without bilingual signals. A word is considered to have bilingual signals if it appears together with its translation in the same 2048-token context.

Next, we measure the translation accuracy of each word by checking whether its translation appears in the model’s predictions. We then analyze the relationship between the frequency of Chinese or Indonesian words and their correct translation

⁷<https://github.com/fxsjy/jieba>

ratio, as shown in Figure 3 and Figure 4. In these figures, the x-axis represents the word frequency in the 167B tokens of training data. We group the words into 21 frequency intervals and calculate the average correct translation ratio for each interval.

The results indicate that words with bilingual signals achieve a higher correct translation ratio compared to words without bilingual signals. This demonstrates that the amount of bilingual signals in the training data significantly impacts the LLM’s translation performance, which motivates our main experiment.

6.5 Recommendations

In this section, we briefly discuss our key findings for the future development of LLMs.

1. Leverage parallel data

When building multilingual large language models, parallel data should be fully utilized. We believe that reducing parallel data to merely monolingual data in other languages by scattering it throughout the training data—as done by BLOOM—is wasteful. The information contained in parallel data should be maximally leveraged by preserving its parallel format.

2. Second-stage training

In addition to maintaining the format of parallel data, the timing of its introduction to the model also plays a significant role. [Gururangan et al. \(2020\)](#) reported that second-phase pre-training on task-specific domains can improve a model’s performance on those tasks. From our systematic study, we found that training a model with parallel data after non-parallel data not only improves translation performance but also enhances multilingual common-sense reasoning.

This strategy has been adopted by recent multilingual LLMs such as Tower ([Alves et al., 2024](#)) and Pangea ([Yue et al., 2025](#)), to some extent, in the form of instruction tuning with parallel data from machine translation tasks or by augmenting multimodal instruction tuning data with their translations. We expect to see broader utilization of parallel data in future multilingual LLMs.

3. Specialized translation systems

Fine-tuning a large language model using parallel data, even though it is pre-trained primar-

ily on English data, offers a quick and effective method for developing high-quality machine translation systems. This approach has gained popularity recently ([Xu et al., 2025](#); [Zeng et al., 2024](#)). Furthermore, we found that the model can achieve even higher performance by formatting the parallel data solely in the desired translation direction.

7 Conclusion

In this work, we report a systematic study of the effect of including parallel data in the training data on large language models’ multilingual capabilities, specifically focusing on translation and multilingual common-sense reasoning. Using English, Chinese, and Indonesian as case studies, we conduct controlled experiments to compare training large language models with mainly English data, with monolingual corpora of other languages, and with parallel data. We found that training LLMs with parallel data significantly enhances LLMs’ multilingual capabilities.

Furthermore, we investigate how the location of parallel data affects the multilingual capabilities. We found that training the model with parallel data at the beginning of the training process is the least effective and leads to serious catastrophic forgetting. Conversely, training the model with parallel data at the end of the training process is the most effective in enhancing the model’s multilingual capabilities. It significantly improves translation scores and also enhances the model’s common-sense reasoning in other languages. Therefore, it is more effective to incorporate parallel data in second-stage training rather than randomly mixing them with non-parallel data.

Training the model in one translation direction can improve translation performance more than training in all translation directions simultaneously. However, this comes with a caveat: the model becomes totally unable to perform translation in other directions, including the opposite of what it was trained on.

Our source code, models, and data are publicly available to support further research into the multilingual capabilities of large language models. We hope that parallel data will be effectively leveraged in developing future open multilingual LLMs.

Limitations

Due to resource constraints, we experiment with a language model containing 1.1B parameters, trained on 167B tokens, focusing on multilingual capabilities in English, Chinese, and Indonesian. Exploring additional languages and larger-scale models is left for future work.

Our experiments only use publicly available data. We believe it does not pose any direct societal risks.

Acknowledgments

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG3-RP-2022-030).

References

- Mohammed Al-Maamari, Mehdi Ben Amor, and Michael Granitzer. 2024. Mixture of modular experts: Distilling knowledge from a multilingual teacher into specialized modular language models. *arXiv preprint arXiv:2407.19610*.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*.
- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *Proceedings of COLM*.
- Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. [The reversal curse: LLMs trained on “A is B” fail to learn “B is A”](#). In *Proceedings of ICLR*.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*, pages 2397–2430.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of AAAI*, pages 7432–7439.
- Eleftheria Briakou, Colin Cherry, and George Foster. 2023. [Searching for needles in a haystack: On the role of incidental bilingualism in PaLM’s translation capability](#). In *Proceedings of ACL*, pages 9432–9452.
- Steven Cao, Nikita Kitaev, and Dan Klein. 2020. [Multilingual alignment of contextual word representations](#). In *Proceedings of ICLR*.
- Linzhang Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. xCoT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of AAAI*.
- Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021a. [mT6: Multilingual pretrained text-to-text transformer with translation pairs](#). In *Proceedings of EMNLP*, pages 1671–1683.
- 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 ACL*, pages 3418–3430.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and et al. 2023. PaLM: Scaling language modeling with pathways. *JMLR*, 24(1).
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL*, pages 2924–2936.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*.
- 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 EMNLP*, pages 2475–2485.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

- Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of NAACL*, pages 550–564.
- Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muenighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. *OLMo: Accelerating the science of language models*. In *Proceedings of ACL*, pages 15789–15809.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. *Don’t stop pretraining: Adapt language models to domains and tasks*. In *Proceedings of ACL*, pages 8342–8360.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of EMNLP*, pages 12365–12394.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H  rve J  gou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL*, pages 427–431.
- Tom Kocmi, Rachel Bawden, Ond  rej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Nov  k, Martin Popel, and Maja Popovi  . 2022. *Findings of the 2022 Conference on Machine Translation (WMT22)*. In *Proceedings of WMT*, pages 1–45.
- Philipp Koehn. 2005. *Europarl: A parallel corpus for statistical machine translation*. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. *Few-shot learning with multilingual generative language models*. In *Proceedings of EMNLP*, pages 9019–9052.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. *Multilingual denoising pre-training for neural machine translation*. *TACL*, 8:726–742.
- Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2024. *LLM360: Towards fully transparent open-source LLMs*. In *Proceedings of COLM*.
- Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2020. *A large English–Thai parallel corpus from the web and machine-generated text*. *Language Resources and Evaluation*, 56:477 – 499.
- Jinliang Lu, Yu Lu, and Jiajun Zhang. 2023. *Take a closer look at multilinguality! Improve multilingual pre-training using monolingual corpora only*. In *Findings of EMNLP*, pages 2891–2907.
- Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. *VECO: Variable and flexible cross-lingual pre-training for language understanding and generation*. In *Proceedings of ACL*, pages 3980–3994.
- Zhuoyuan Mao and Yen Yu. 2024. *Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages*. In *Proceedings of LoResMT*, pages 1–25.
- Michael McCloskey and Neal J. Cohen. 1989. *Catastrophic interference in connectionist networks: The sequential learning problem*. *Psychology of Learning and Motivation*, 24:109–165.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of EMNLP*, pages 2381–2391.
- Roberto Navigli, Simone Conia, and Bj  rn Ross. 2023. *Biases in large language models: Origins, inventory, and discussion*. *J. Data and Information Quality*, 15(2):1–21.
- Sina Bagheri Nezhad and Ameeta Agrawal. 2024. *What drives performance in multilingual language models?* In *Proceedings of VarDial*, pages 16–27.
- Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias St  rmer, and Ilias Chalkidis. 2023. *LEXTREME: A multi-lingual and multi-task benchmark for the legal domain*. In *Findings of EMNLP*, pages 3016–3054.

- Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. [ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora](#). In *Proceedings of EMNLP*, pages 27–38.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of ACL*, pages 311–318.
- Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal common-sense reasoning. In *Proceedings of EMNLP*, pages 2362–2376.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of EMNLP*, pages 2695–2709.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. 2024. [Empowering multi-step reasoning across languages via program-aided language models](#). In *Proceedings of EMNLP*, pages 12171–12187.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of WMT*, pages 634–645.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of EMNLP*, pages 1627–1643.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of AAAI*, pages 8732–8740.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie 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. *arXiv preprint arXiv:2211.05100*.
- Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL*, pages 1599–1613.
- Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. [Findings of the WMT 2023 shared task on machine translation with terminologies](#). In *Proceedings of WMT*, pages 663–671.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. [SlimPajama: A 627B token cleaned and deduplicated version of RedPajama](#).
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of ACL*, pages 5701–5715.
- 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](#). In *Findings of ACL*, pages 3534–3546.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,

- Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Guan Wang, Sijie Cheng, Xianyu Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. OpenChat: Advancing open-source language models with mixed-quality data. In *Proceedings of ICLR*.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of W-NUT*, pages 94–106.
- Andrea W. Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of EMNLP*, pages 1124–1131.
- Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of WMT*, pages 89–99.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *Proceedings of ICLR*.
- Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *Proceedings of ICLR*.
- 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 NAACL*, pages 483–498.
- Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.
- Xiang Yue, Yueqi Song, Akari Asai, Simran Khanuja, Anjali Kantharuban, Seungone Kim, Jean de Dieu Nyandwi, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *Proceedings of ICLR*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, pages 4791–4800.
- Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. In *Proceedings of AAAI*, pages 19488–19496.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.
- Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of ACL*, pages 11189–11204.
- Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond English: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.
- Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024a. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.
- Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. In *Findings of ACL*, pages 8411–8423.

A Appendix

A.1 Language Detection

To measure the proportion of different languages in the training data, we use fastText⁸ (Joulin et al., 2016, 2017) to detect each word’s language. For each word, fastText outputs a probability distribution over languages, and we assign its language based on the highest predicted probability. If a word is tokenized into multiple subwords by the tokenizer (e.g., BPE), we attribute the number of subwords to the original word’s detected language when computing language proportions. Table 8 presents the proportions of thirty languages in the 167B training data (each denoted by its ISO language code) with a proportion higher than 0.1%.

⁸<https://fasttext.cc/docs/en/language-identification.html>

Language	Percentage	Language	Percentage	Language	Percentage
EN	82.35	PL	0.52	CS	0.18
FR	2.46	CA	0.51	TR	0.18
DE	2.13	CEB	0.34	DA	0.18
ES	1.74	FI	0.29	JA	0.18
IT	1.54	HU	0.23	FA	0.16
SV	0.92	AR	0.22	ALS	0.15
PT	0.86	ID	0.19	ZH	0.12
RU	0.74	EO	0.19	SR	0.12
NL	0.66	NO	0.18	HR	0.11
UK	0.60	KN	0.18	LA	0.11

Table 8: Percentage (%) of different languages (larger than 0.1%) in our non-parallel training corpus.

A.2 Hyper-Parameters Setting

Our experiments use the TinyLlama 1.1B model, with architectural details provided in Table 9. The hyper-parameters used for training are reported in Table 10.

Hyper-parameter	Value
Number of Layers	22
Embedding Dimension	2048
Intermediate Dimension	5632
Attention Heads	32
Query Groups	4
Context Window	2048
Vocabulary Size	32000

Table 9: Architecture of our model.

Hyper-parameter	Value
Number of GPUs	8
Global Batch Size	512
Micro Batch Size	16
Learning Rate	4e-4
Warmup Steps	2000
Weight Decay	1e-1
Optimizer	AdamW
(β_1, β_2)	(0.9, 0.95)
Gradient Clip	1.0
Minimal Learning Rate	4e-5

Table 10: Hyper-parameters of our experiments.

A.3 Common-Sense Reasoning Performance

We provide the details of the models’ performance on Chinese and Indonesian common-sense reasoning tasks in Table 12 and Table 13 respectively.

Translate this from English to Indonesian
English: The pilot was identified as Squadron Leader Dilokrit Pattavee.
Indonesian:

Table 11: An example input prompt for English to Indonesian zero-shot translation evaluation.

Model	XCOPA	XNLI	XStoryCloze	XWG	Avg
No Parallel	52.20	33.33	48.64	53.37	46.88
Multilingual	53.40	34.10	49.04	59.13	48.92
Parallel Non-Adjacent	51.00	33.57	49.37	59.13	48.27
Parallel First	48.40	34.74	48.38	51.19	45.68
Parallel Distributed	52.00	34.90	49.17	57.54	48.40
Parallel Last (all)	54.40	33.73	52.15	61.90	50.55
BLOOM	59.40	36.67	58.04	69.05	55.79

Table 12: Chinese common-sense reasoning performance (accuracy) in each experimental setting, measured based on the model’s accuracy on XCOPA, XNLI, XStoryCloze, and XWinograd.

Model	XCOPA	XStoryCloze	Avg
No Parallel	50.40	49.44	49.92
Multilingual	52.60	50.36	51.48
Parallel Non-Adjacent	52.40	50.43	51.41
Parallel First	52.20	49.90	51.05
Parallel Distributed	55.20	52.88	54.04
Parallel Last (all)	57.40	54.73	56.06
BLOOM	64.60	57.78	61.19

Table 13: Indonesian common-sense reasoning performance (accuracy) in each experimental setting, measured based on the model’s accuracy on XCOPA and XStoryCloze.