Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown continuously improving multilingual capabilities, and even small-scale open-source models have demonstrated rapid performance enhancement. In this paper, we systematically explore the abilities of open LLMs with less than ten billion parameters to handle multilingual machine translation (MT) tasks. We conduct comprehensive evaluations on six popular LLMs and find that models like Gemma2-9B exhibit impressive multilingual translation capabilities. We then introduce the **P**arallel-**F**irst **M**onolingual-**S**econd (PFMS) data mixing strategy in the continual pretraining stage to further enhance the MT performance and present GemmaX2-28, a 9B model achieving top-tier multilingual translation performance across 28 languages[1]. Specifically, GemmaX2-28 consistently outperforms the state-of-the-art (SOTA) models such as TowerInstruct (Alves et al., 2024) and X-ALMA (Xu et al., 2024b) and achieves competitive performance with Google Translate and GPT-4-turbo[2].

1 Introduction

Large language models (LLMs) such as GPT models (OpenAI, 2023a,b), PaLM models (Chowdhery et al., 2022; Anil et al., 2023), and others have demonstrated remarkable capabilities in multilingual translation (Jiao et al., 2023; Vilar et al., 2023). Recently, significant efforts have been directed towards bolstering the multilingual competencies of open LLMs across a diverse linguistic spectrum, beyond English and Chinese. For example, over 5% of the LLaMA3 (Dubey et al., 2024) pretraining datasets consist of high-quality non-English data covering over 30 languages; Qwen2/2.5 (Yang et al., 2024) claim to have multilingual support for over 29 languages; Üstün et al. (2024) introduce Aya-101, a massively multilingual generative language model supporting 101 languages.

Zhu et al. (2024b) evaluate popular LLMs on multilingual MT tasks and find that the translation capability of LLMs is continually evolving. However, the models assessed (XGLM (Lin et al., 2022), OPT (Zhang et al., 2022), Falcon (Almazrouei et al., 2023), BLOOMZ (Workshop et al., 2023), LLaMA2 (Touvron et al., 2023b)) are outdated due to the rapid development of open LLMs, and the multilingual translation ability of the latest models remains unclear. One natural question arises: 1) *How do the latest open LLMs with practical scale perform multilingual MT tasks?*

To further boost LLMs' translation capability, existing approaches usually adopt multilingual corpora during continual pretraining on open LLMs. Specifically, Xu et al. (2024a) utilize continual pretraining on monolingual datasets. Guo et al. (2024) apply continual pretraining on monolingual datasets followed by parallel datasets. Alves et al. (2024) perform continual pretraining on a multilingual mixture of monolingual (two-thirds) and parallel (one-third) datasets. However, the optimal approach to leverage monolingual and parallel datasets for multilingual MT remains underexplored. Another natural question arises: 2) *What are the best practices for leveraging monolingual and parallel corpora to enhance the multilingual MT performance of LLMs?*

In this paper, we first evaluate the in-context capabilities of the latest open-source LLMs including Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2/2.5-7B (Yang et al., 2024), LLaMA3/3.1-8B (Dubey et al., 2024), and Gemma2-9B (Team et al., 2024) across 28 languages on the multilingual MT tasks.

[1]Arabic (`ar`), Bengali (`bn`), Czech (`cs`), German (`de`), English (`en`), Spanish (`es`), Persian (`fa`), French (`fr`), Hebrew (`he`), Hindi (`hi`), Indonesian (`id`), Italian (`it`), Japanese (`ja`), Khmer (`km`), Korean (`ko`), Lao (`lo`), Malay (`ms`), Burmese (`my`), Dutch (`nl`), Polish (`pl`), Portuguese (`pt`), Russian (`ru`), Thai (`th`), Tagalog (`tl`), Turkish (`tr`), Urdu (`ur`), Vietnamese (`vi`), and Chinese (`zh`).

[2]We will release our models upon publication.

We find that open LLMs like Gemma2-9B exhibit remarkable multilingual translation capabilities but still fall behind strong closed-source models. We then systematically investigate the most effective methods to leverage monolingual and parallel data for further boosting Gemma2-9B's capabilities and propose the Parallel-First Monolingual-Second (PFMS) data mixing strategy. Leveraging PFMS strategy when continually pretraining on Gemma2-9B followed by instruction finetuning on a small set of high-quality translation pairs, we learn GemmaX2-28-9B, a many-to-many multilingual MT model that performs competitively with Google Translate and GPT-4-turbo across 28 widely spoken languages. The contributions of this paper can be summarized as follows:

- We benchmark the latest open-source LLMs on multilingual MT tasks in 28 widely spoken languages, covering English-centric and Chinese-centric translation directions.
- We systematically explore the optimal training recipe for boosting multilingual MT with LLMs and introduce the Parallel-First Monolingual-Second (PFMS) data mixing strategy for the continual pretraining stage.
- We publicly release GemmaX2-28-9B, the many-to-many multilingual translation model supporting 28 languages, which consistently outperforms open-source alternates and is competitive with closed-source models such as Google Translate and GPT-4-turbo.

2 Related Work

2.1 Multilinguality in LLMs

Large language models have gained substantial attention from the research community due to their outstanding performance on various tasks (Winata et al., 2021; Wei et al., 2022; Touvron et al., 2023a). However, most advancements in LLMs have primarily focused on English or Chinese, resulting in inadequate performance in low-resource languages (Ebrahimi et al., 2022; Asai et al., 2024). The reason is that the language distribution of the training data for LLMs is highly imbalanced and the quality varies across languages (Ding et al., 2024). Numerous efforts have been made to enhance the multilingual capabilities of LLMs. Some studies seek to enhance the performance of LLMs on low-resource languages through continual pretraining (Cui et al., 2024b) or supervised finetuning (Üstün et al., 2024) using data from these languages. Additionally, some researchers (Li et al., 2024a) leverage contrastive learning to align the internal representations of different languages, allowing the model to improve its cross-lingual capabilities with minimal training data. In our study, we assessed the latest open-source LLMs on translation tasks across 28 languages. We find that these latest models still show significant performance disparities across different languages, indicating that low-resource languages continue to pose a significant challenge for these models. Therefore, we expand the Gemma2 model to cover 28 commonly used languages, enabling it to demonstrate strong generative and translation capabilities across these languages.

2.2 Multilingual MT with LLMs

The use of LLMs has shown significant progress in MT tasks (Lu et al., 2024; Xu et al., 2024c; Li et al., 2024b; Gao et al., 2024), leading to different approaches on LLM-based translation. One line of work focuses on the in-context translation capabilities, and LLMs are provided with parallel sentences to guide the model in generating the target sentence. Some studies (Agrawal et al., 2023; Zhu et al., 2024a; Cui et al., 2024a) show that MT performance can be improved by using semantically related parallel sentences as examples, exhibiting promising results in scenarios with limited computational resources and insufficient parallel data.

Another line involves finetuning with translation instructions. Xu et al. (2024a) initially pretrained the base model on monolingual data, followed by finetuning on small human-written parallel datasets, which resulted in strong translation performance. Guo et al. (2024) challenge the perspective that the impact of parallel data is reduced in the LLM era and demonstrate the effectiveness of parallel data during continual pretraining on enhancing multilingual translation. However, the distribution of the parallel datasets could be highly imbalanced across different languages for massively multilingual translation scenarios. Alves et al. (2024) incorporates a fixed multilingual mixture of monolingual (two-thirds) and parallel (one-third) data, further enhancing LLMs' translation capabilities. Our work differs in that we primarily focus on exploring the optimal mixing strategy of monolingual and parallel data during continual pretraining to achieve the best translation performance.

3 Datasets and Baseline Settings

In this section, we describe the evaluation datasets used in Sections 4 and 5 as well as the model configurations.

3.1 Datasets

We conduct experiments on 28 languages across a broad linguistic spectrum. The detailed information of all languages is summarized in Table 3. We evaluate the multilingual translation performance on the FLORES-200 (Team et al., 2022) benchmark. To avoid data leakage issues, we also consider the WMT-24 test sets for evaluating multilingual translation capabilities. Only sources are provided in the WMT-24 benchmark, and we adopt English sentences for reference-free evaluation.

3.2 Models

We evaluate the translation performance of six open-source LLMs: Mistral-7B-v0.3 (Jiang et al., 2023), Qwen2/2.5-7B (Yang et al., 2024), LLaMA3/3.1-8B (Dubey et al., 2024) and Gemma2-9B (Team et al., 2024). We also summarize the results of several SOTA models as follows:

- Google Translate: The translation performance by leveraging the Google Translate API[3]. We include it to represent the commercial MT system.
- GPT-3.5/4 Turbo: The translation performance by leveraging the OpenAI API[4] with greedy decoding for inference. We use the prompt strategy demonstrated in Table 4.
- NLLB-54.5B: The largest and best multilingual NMT model with encoder-decoder architecture released by the No Language Left Behind (NLLB) project (Team et al., 2022).

3.3 Evaluation

For the FLORES-200 benchmark, we evaluate MT performance by spBLEU[5] (Goyal et al., 2022) and COMET[6] (Rei et al., 2020). For the WMT-24 benchmark, we adopt two reference-free models, XCOMET[7] (Guerreiro et al., 2023) and COMETKiwi[8] (Rei et al., 2023), each of which has 10B parameters and demonstrates high correlation with human judgments (Freitag et al., 2023).

4 Benchmarking Open LLMs for Multilingual Machine Translation

In this section, we report empirical results on tokenizer efficiency and multilingual MT performance across various open-source LLMs.

4.1 Tokenizer Efficiency

The open-source LLMs are typically pretrained on one or a few dominant languages, and tokenization on low-resource languages usually results in undesirable long sequences of subwords which leads to excessive GPU memory consumption during training and slow decoding speed during inference.

Following the line of Liao et al. (2024), we compare the length differences between tokenized English sentences and their corresponding non-English counterparts to evaluate the tokenizer efficiency across different LLMs. Specifically, we define the length ratio as follows:

$$\text{length ratio} = \frac{\text{length}(\text{tokenizer}(y))}{\text{length}(\text{tokenizer}(x))}, \quad (1)$$

where x represents the English sentence, and y denotes the corresponding non-English sentence. A smaller length ratio is preferred, meaning that the tokenizer encodes the non-English sentences as efficiently as the English sentence.

We conduct our experiments on the FLORES-200 devtest dataset, which includes 1012 sentences for each language. Note that LLaMA3 and LLaMA3.1 share the same tokenizer, as do Qwen2 and Qwen2.5. We also include the tokenizer of the NLLB-54.5B model as a strong baseline for comparison. The averaged length ratio for each language is illustrated in Figure 1. By checking the length ratio for various languages, we have the following observations: 1) Due to its extensive support and optimization for more than 200 languages, NLLB-54.5B has the most balanced tokenizer and achieves low length ratios across different languages. 2) Open-source LLMs demonstrate notably high length ratios for low-resource languages such as Khmer, Lao, and Burmese. Gemma2-9B consistently achieves better length ratios for almost all languages compared with other LLMs.

[3] https://translate.google.com/
[4] https://api.openai.com/v1/chat/completions
[5] We calculate the spBLEU scores via sacreBLEU (Post, 2018) with the `flores200` tokenizer.
[6] https://huggingface.co/Unbabel/wmt22-comet-da
[7] https://huggingface.co/Unbabel/XCOMET-XXL
[8] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

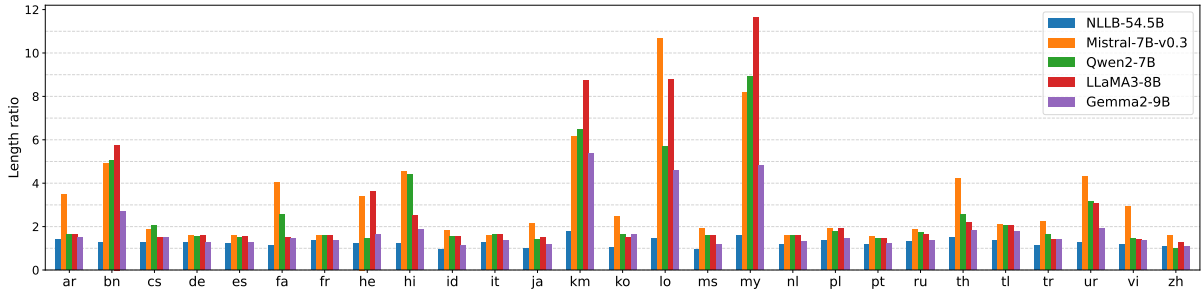

Figure 1: The tokenizer efficiency of open-source LLMs for each non-English language. The smaller the length ratio is, the more efficient the tokenizer is.

Model	WMT-24	FLORES-200			
	en → xx	en → xx	xx → en	zh → xx	xx → zh
Google Translate	77.64 / 73.00	41.52 / 88.51	45.35 / 88.83	29.64 / 85.69	34.86 / 87.20
GPT-4-turbo	79.35 / 75.40	37.41 / 87.61	41.00 / 88.15	24.75 / 84.84	27.91 / 85.67
GPT-3.5-turbo	72.47 / 67.39	33.16 / 83.15	35.66 / 85.21	20.89 / 80.13	23.64 / 81.82
NLLB-54.5B	-	37.34 / 87.05	43.63 / 88.32	25.41 / 84.42	20.73 / 80.72
Gemma2-9B	72.06 / 67.05	33.05 / 84.65	42.00 / 87.94	20.54 / 80.54	27.77 / 85.34
Llama3.1-8B	65.23 / 58.89	28.52 / 81.60	37.72 / 86.74	17.66 / 78.09	25.53 / 83.49
Llama3-8B	65.02 / 58.10	27.93 / 81.28	37.39 / 86.59	16.83 / 77.40	22.81 / 83.11
Qwen2.5-7B	59.36 / 51.59	22.65 / 75.20	34.64 / 84.94	14.97 / 72.64	25.30 / 83.16
Qwen2-7B	58.12 / 49.68	21.69 / 75.68	34.22 / 85.15	14.44 / 73.19	25.45 / 83.55
Mistral-7B-v0.3	51.21 / 40.48	17.04 / 67.61	29.44 / 81.35	8.18 / 62.31	12.19 / 72.62

Table 1: Performance of different models on WMT-24 (XCOMET / COMETKiwi) and FLORES-200 (spBLEU / COMET) benchmarks. The detailed experimental results are summarized in Tables 5, 6, and 7.

4.2 In-context Multilingual Translation Performance with Open LLMs

We here investigate the multilingual translation performance of different LLMs with in-context learning strategies on the FLORES-200 and WMT-24 benchmarks. For each open LLM, we report its translation performance across 28 languages with five randomly selected translation pairs[9] from the FLORES-200 development dataset as the in-context exemplars. Note that we also adopt the FLORES-200 development dataset for the WMT-24 benchmark. Following the line of Zhu et al. (2024b), we adopt the format "<X>=<Y>" as the in-context template, where <X> and <Y> denote the source and target sentences of the select parallel sentence pairs. All experiments are conducted based on OpenICL[10] (Wu et al., 2023).

We report the averaged multilingual translation performance on the WMT-24 and FLORES-200 benchmarks in Tables 1. By checking the translation performance for various languages and models, we have the following observations:

- Open LLMs demonstrate impressive multilingual translation capability. Specifically, Gemma2-9B performs the best among the open LLMs evaluated and outperforms GPT-3.5-turbo on average.
- Open LLMs still lag behind the strong supervised NMT models such as NLLB-54.5B, especially in low-resource languages.
- GPT-4-turbo exhibits remarkable multilingual translation performance and even outperforms Google Translate in the WMT-24 benchmark.
- We do not observe serious data leakage issues on public datasets (FLORES-200) for the open LLMs we evaluated, where the translation performance on both WMT-24 and FLORES-200 benchmarks share a similar trend.

We then illustrate the translation performance on the FLORES-200 benchmark according to varying numbers of in-context exemplars in Figure 2. We can see that the COMET scores increase dramatically with the number of exemplars increasing from 1 to 5, and the translation performance plateaus afterward except for Mistral-7B-v0.3 which has relatively weak multilingual capability.

[9] We apply the same five randomly selected translation pairs as exemplars for each direction during evaluation.

[10] https://github.com/Shark-NLP/OpenICL

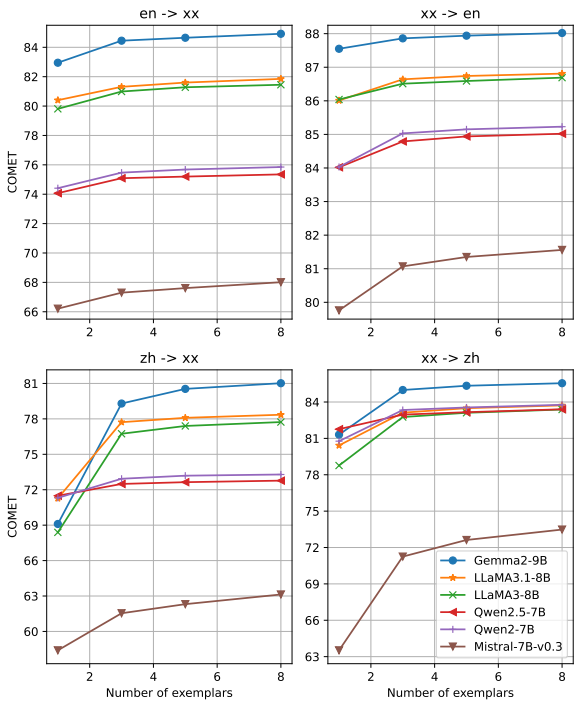

Figure 2: MT performance on the FLORES-200 benchmark with different numbers of in-context exemplars.

5 GemmaX: Boosting Multilingual Translation with Gemma Models

Given its impressive multilingual capabilities discussed in Section 4, we select Gemma2-9B as our backbone model for learning many-to-many multilingual machine translation across 28 languages. We continue the pretraining of Gemma2-9B with multilingual corpora and finetune the continual pretrained model using a small but high-quality parallel dataset with the translation prompt as follows: Translate this from [*source language*] to [*target language*]:\n[*source language*]: <*source sentence*>\n[*target language*]:<*target sentence*>.

5.1 Pretraining Data

Monolingual Data We collect monolingual data from CulturaX (Nguyen et al., 2024) and MADLAD-400 (Kudugunta et al., 2023). CulturaX contains 6.3 trillion tokens in 167 languages, which undergo meticulous cleaning and deduplication through a comprehensive pipeline. MADLAD-400 is a multilingual dataset with 3 trillion tokens in 419 languages, on which the authors perform a self-audit to guarantee the data quality.

Parallel Data We collect all Chinese-centric and English-centric parallel datasets from the OPUS collection[11] (Tiedemann, 2012) up to August 2024, which is comprised of multiple corpora, ranging from movie subtitles (Tiedemann, 2016) to Bible (Christodouloupoulos and Steedman, 2015) to web crawled datasets (El-Kishky et al., 2020; Schwenk et al., 2021). We download all available corpora and concatenate them without curating the datasets or trying to balance the representation of different domains. After collecting all parallel datasets, we adopt the data-cleaning process as follows: 1) We remove duplicate sentence pairs and discard sentence pairs by utilizing some heuristic approaches. 2) Language identification filtering is applied by utilizing the fastText toolkit (Joulin et al., 2016, 2017). 3) Semantic similarity filtering is performed based on LaBSE (Feng et al., 2022) and MuSR (Gao et al., 2023). Specifically, we only keep the sentence pairs that have similarity scores between 0.75 and 0.99. After the cleaning process, we have about 3.4 billion cleaned Chinese-centric and English-centric sentence pairs covering 28 languages. The distribution of our parallel datasets for each language is illustrated in Figure 3.

5.2 Supervised Finetuning Data

Inspired by Xu et al. (2024a) that a small amount of high-quality data could dramatically boost the translation performance, we construct our finetuning dataset mostly from human-annotated records: 1) We extract the sentence pairs covering ten languages from the general translation task in the TowerBlock dataset[12]. 2) For the languages not covered in TowerBlock, we randomly sample 1000 English-centric sentence pairs for each language either from the NTREX-128 (Federmann et al., 2022) and FLORES-200 dev datasets or the OPUS dataset bidirectionally filtered by *wmt23-cometkiwi-da-xxl* model with quality scores above 0.85. 3) To mitigate the off-target issue and enhance the zero-shot translation performance, inspired by Wu et al. (2024), we also randomly select 25 non-English-centric translation directions and sample 100 sentence pairs for each direction from the NTREX-128 and FLORES-200 dev datasets. In summary, the total number of sentence pairs is around 196K, where there are about 189K in the English-centric directions and 7K in other directions.

[11] http://www.opus.nlpl.eu

[12] https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.2

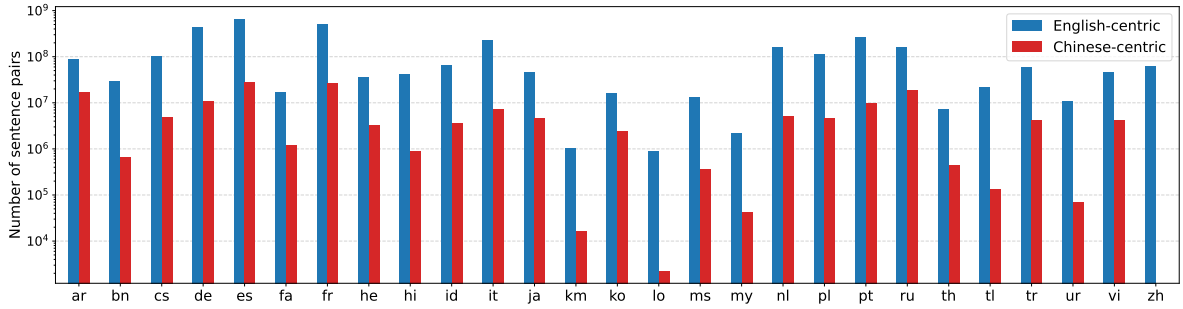

Figure 3: Number of sentences in different languages for Chinese-centric and English-centric parallel dataset.

5.3 Training Configuration

We train all models with the LLaMA-Factory (Zheng et al., 2024) framework for one epoch on 32 and 8 NVIDIA H800 GPUs for the pretraining and finetuning stages respectively.

In the pretraining stage, we utilize an effective batch size of 1.57 million tokens, corresponding to a batch size of 4 per GPU with a gradient accumulation of 6 and a maximum sequence length of 2048. We use the AdamW optimizer with a learning rate of 2e-5. We adopt full-weight pretraining with bf16 precision by employing a cosine learning rate scheduler. In the finetuning stage, we adopt a batch size of 4 per GPU with a gradient accumulation of 8 and a maximum sequence length of 2048. Additionally, we employ an inverse square root learning rate scheduler. Detailed training configurations can be found in Tables 8 and 9. We use greedy decoding to generate translations.

5.4 Exploring the Best Data Recipe for Multilingual Translation with LLMs

Xu et al. (2024a) conduct continual pretraining solely on monolingual data followed by instruction finetuning on a small number of high-quality translation pairs, achieving impressive translation results. They conclude that the reliance on parallel data is diminished in the era of LLMs. Subsequently, Guo et al. (2024) reemphasize the importance of parallel data and propose a three-stage training strategy: 1) continual pretraining with monolingual data; 2) continual pretraining with parallel data; 3) finetuning with source language consistent translation instructions. However, as illustrated in Figure 3, the parallel data adopted in the second stage could be highly imbalanced for massively multilingual settings, which would weaken the model's understanding and generative capabilities on low-resource languages learned in the first stage. Alves et al. (2024) utilize a multilingual mixture of monolingual (two-thirds) and parallel (one-third) data during continual pretraining, which significantly improves the translation performance of their model. In our experiments, we follow the idea of a multilingual mixture in the continual pretraining stage and raise two questions: 1) *whether a monolingual to parallel ratio of 2:1 is the best data recipe for multilingual machine translation?* 2) *If not, what is the best data recipe?*

Therefore, we consider the following configurations when preparing the pretraining dataset: 1) monolingual data only, 2) 2 : 1 ratio of monolingual to parallel data, 3) 1 : 1 ratio, 4) 1 : 2 ratio, and 5) parallel data only. For each language, we collect two billion tokens, allocating them between monolingual and parallel sentences based on the specified ratios mentioned above. For each parallel sentence, we construct a new sentence in either <*source sentence*>\n<*target sentence*> or <*target sentence*>\n<*source sentence*> manner, with the order of the source and target sentences determined randomly. We randomly intersperse the monolingual and parallel data. Due to the inherent characteristics of low-resource languages, the corresponding monolingual or parallel data may be insufficient, resulting in the total number of tokens in the pretraining dataset falling short of 56 billion. As a result, we construct five pretraining datasets corresponding to these configurations.

We continually pretrain the Gemma2-9B model using five datasets containing varying proportions of monolingual and parallel data, followed by translation instruction finetuning on our high-quality datasets. Based on Table 3, we classify the 28 languages into high-resource (18), mid-resource (7), and low-resource (3) languages according to their resource availability. We then report the translation performance during continual pretraining on these three language groups in Figures 4 and 5.

By checking model performance under different

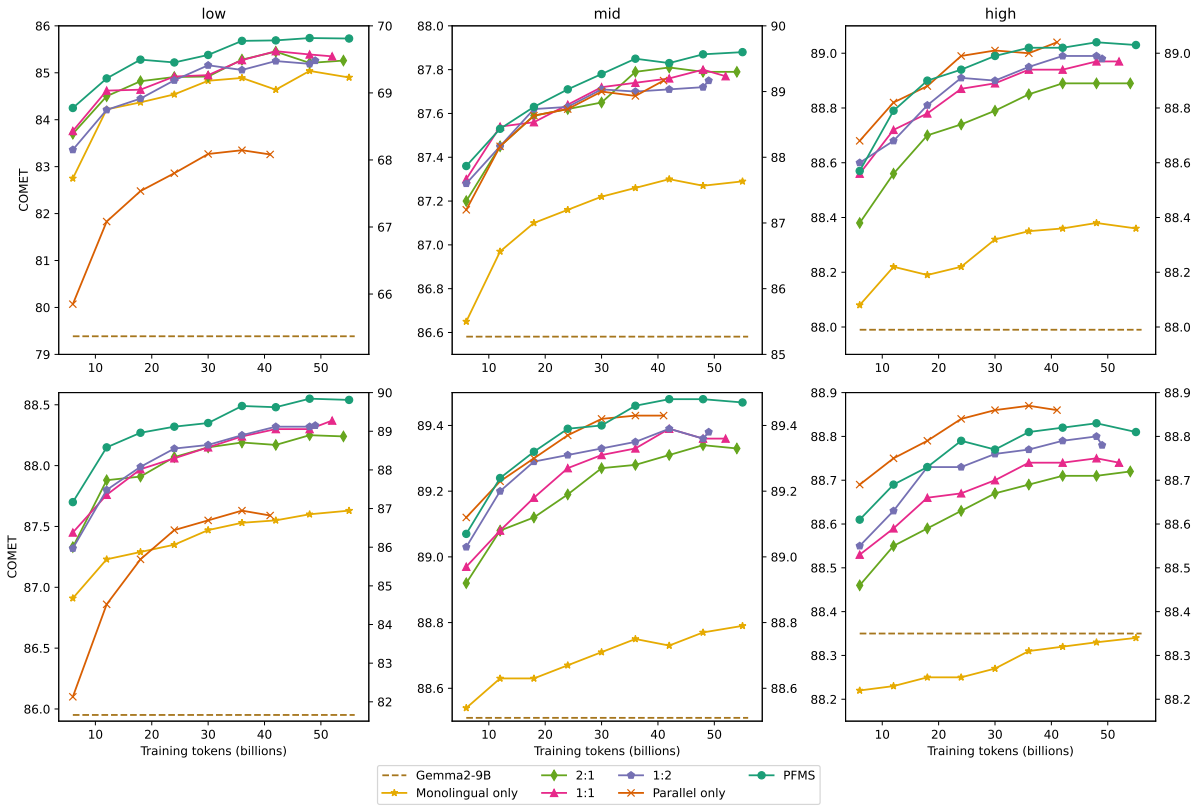

Figure 4: The translation performance (COMET) of models trained with different data recipes during continual pretraining on low-resource (left), mid-resource (middle), and high-resource (right) languages. The upper subfigures illustrate the en→xx translation performance, while the lower subfigures depict the xx→en translation performance. Note that "Gemma2-9B" refers to the direct finetuning of the model without continual pretraining, and its performance is reflected in the right-hand y-axis. The translation performance in BLEU scores is illustrated in Figure 5.

combinations of data recipes and language groups, we have the following observations:

- The current prevalent training paradigm, which involves continual pretraining on large-scale monolingual data, proves to be suboptimal. While this approach improves the overall performance across various languages, it does so at the cost of translation quality for high-resource languages, particularly in the xx→en directions.

- Incorporating parallel data into the pretraining stage, regardless of the volume, significantly enhances the model's overall translation capabilities. Notably, the translation performance for high-resource language directions improves as the quantity of parallel data increases. However, such improvement was not evident for low-resource and mid-resource languages.

- We believe that the base model has sufficient understanding and generation capabilities for high-resource languages. It mainly needs parallel data during pretraining to align the internal representations across different languages. In contrast, the model's generative and comprehension abilities for low-resource and mid-resource languages are not up to par. To improve its performance with these languages, it's crucial to include more data in the pretraining process to enhance its generative capacities. Unfortunately, due to the nature of low-resource and mid-resource languages, we are unable to obtain enough parallel data. Therefore, we think it's essential to incorporate monolingual data into the pretraining dataset for low-resource and mid-resource languages.

5.5 Main Result

Based on the experimental results in Section 5.4, we propose a Parallel-First Monolingual-Second (PFMS) data mixing strategy, where we give higher priority to parallel data than monolingual data when preparing the continual pretraining dataset. Specifically, for 2 billion tokens per language, we utilize parallel data as much as possible and supplement it

Model	WMT-24	FLORES-200			
	en → xx	en → xx	xx → en	zh → xx	xx → zh
10 **languages**					
TowerInstruct-7B-v0.2	84.48 / 75.02	38.91 / 88.46	42.20 / 88.29	23.36 / 85.21	26.87 / 85.69
TowerInstruct-13B-v0.1	86.10 / 76.74	40.60 / 88.89	43.14 / 88.50	25.28 / 86.13	29.32 / 86.59
GemmaX2-28-9B	**86.59 / 77.25**	**42.58 / 89.15**	**43.95 / 88.62**	**28.55 / 86.84**	**34.10 / 87.70**
23 **languages**					
X-ALMA	81.67 / 74.73	39.31 / 88.69	42.09 / 88.54	-	-
GemmaX2-28-9B	**82.05 / 75.31**	**41.98 / 88.97**	**45.27 / 88.97**	28.64 / 86.36	34.26 / 87.56
28 **languages**					
Aya-101	71.67 / 63.88	27.70 / 84.80	33.84 / 86.21	17.36 / 81.56	18.70 / 81.72
LLaMAX3-Alpaca-8B	67.29 / 60.21	28.11 / 83.12	35.81 / 87.05	18.17 / 80.78	21.08 / 83.38
GPT-3.5-turbo	72.47 / 67.39	33.16 / 83.15	35.66 / 85.21	20.89 / 80.13	23.64 / 81.82
NLLB-54.5B	-	37.34 / 87.05	43.63 / 88.32	25.41 / 84.42	20.73 / 80.72
GPT-4-turbo	79.35 / **75.40**	37.41 / 87.61	41.00 / 88.15	24.75 / 84.84	27.91 / 85.67
Google Translate	77.64 / 73.00	**41.52 / 88.51**	**45.35** / 88.83	**29.64 / 85.69**	**34.86** / 87.20
GemmaX2-28-9B	**79.37** / 74.41	39.72 / 88.35	45.07 / **88.95**	27.48 / **85.69**	33.74 / **87.38**

Table 2: Translation performance on WMT24 (XCOMET / COMETKiwi) and FLORES-200 (spBLEU / COMET) benchmarks. The detailed results are summarized in Tables 10, 11 and 12.

with monolingual data if needed. The translation performance by leveraging the PFMS strategy is also illustrated in Figures 4 and 5. We can see that the PFMS strategy consistently outperforms other data mixing strategies except for high-resource languages to English directions. Such a phenomenon might be due to the fact that the dominant part of the parallel-only dataset is the English-centric corpora with high-resource languages.

We continually pretrain Gemma2-9B with PFMS strategy and learn GemmaX2-28-9B by finetuning the pretrained model with high-quality translation pairs covering 28 languages. The experimental results are summarized in Table 2. Besides the models discussed in Section 3.2, We include several strongest open-source massively multilingual MT models as baselines, including:

- TowerInstruct-7/13B (Alves et al., 2024): 7B and 13B LLaMA2-based models for translation-related tasks supporting 10 languages.
- X-ALMA (Xu et al., 2024b): a LLaMA2-based multilingual MT model covering 50 languages, which consists of a 13B dense model and multiple language-specific modules, with a total of approximately 29B parameters. Note that X-ALMA only supports English-centric directions, and we report the translation performance for the overlapping directions.
- Aya-101 (Üstün et al., 2024): a 13B massively multilingual generative language model that follows instructions in 101 languages.
- LLaMAX3-Alpaca-8B (Lu et al., 2024): an 8B LLaMA3-based model supporting 102 languages with powerful multilingual capabilities without loss instruction-following capabilities.

Despite having fewer parameters and supporting a large number of languages, GemmaX2-28 consistently outperforms the current SOTA open-source models, demonstrating the effectiveness of our proposed PFMS data mixing strategy and the powerful translation capabilities of our model. In addition, our model achieves comparable translation performance to GPT-4-turbo and Google Translate, indicating that our model has achieved translation capabilities on par with industry standards.

6 Conclusion

In this paper, we systematically evaluate the multilingual in-context translation of the latest open LLMs at a practical scale. While these models demonstrate strong translation capabilities, they still fall short compared to closed-source models. Furthermore, we explore the best data recipe for multilingual translation with LLMs and propose a Parallel-First Monolingual-Second (PFMS) data mixing strategy. Leveraging the PFMS strategy during continual pretraining on Gemma2-9B, we achieve significant improvements, bringing its translation performance to a level comparable with GPT-4-turbo and Google Translate. To the best of our knowledge, GemmaX2-28-9B is the open model with the highest translation quality. We aim to develop models that support a broader range of languages and possess enhanced translation capabilities as part of our future work.

Limitations

Due to limited computational resources, we only conduct multilingual in-context translation evaluation and explore the optimal data mixing strategy on open LLMs with parameter sizes below ten billion. The translation performance and optimal data recipe for larger models remain unclear.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

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. *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 *First Conference on Language Modeling*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024a. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10885–10897, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024b. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian

Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models. *Preprint*, arXiv:2401.05861.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Learning multilingual sentence representations with cross-lingual consistency regularization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 243–262, Singapore. Association for Computational Linguistics.

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.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *Preprint*, arXiv:1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024a. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. 2024. Ikun for wmt24 general mt task: Llms are here for multilingual machine translation. *Preprint*, arXiv:2408.11512.

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 the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *Preprint*, arXiv:2407.05975.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

OpenAI. 2023a. Chatgpt [large language model]. https://chat.openai.com/.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. 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.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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. *Preprint*, arXiv:2302.13971.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, 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, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhat-

tacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15092–15108, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023. OpenICL: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 489–498, Toronto, Canada. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024b. X-alma: Plug play modules and adaptive rejection for quality translation at scale. *Preprint*, arXiv:2410.03115.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024c. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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. *Preprint*, arXiv:2205.01068.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

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

Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

ISO Code	Language	Script	Family	Subgrouping	Resource
ar	Arabic	Arabic	Afro-Asiatic	Semitic	High
bn	Bengali	Bengali	Indo-European	Indo-Aryan	Mid
cs	Czech	Latin	Indo-European	Balto-Slavic	High
de	German	Latin	Indo-European	Germanic	High
en	English	Latin	Indo-European	Germanic	High
es	Spanish	Latin	Indo-European	Italic	High
fa	Persian	Arabic	Indo-European	Iranian	High
fr	French	Latin	Indo-European	Italic	High
he	Hebrew	Hebrew	Afro-Asiatic	Semitic	Mid
hi	Hindi	Devanagari	Indo-European	Indo-Aryan	High
id	Indonesian	Latin	Austronesian	Malayo-Polynesian	Mid
it	Italian	Latin	Indo-European	Italic	High
ja	Japanese	Japanese	Japonic	Japanesic	High
km	Khmer	Khmer	Austroasiatic	Khmeric	Low
ko	Korean	Hangul	Koreanic	Korean	High
lo	Lao	Lao	Tai-Kadai	Kam-Tai	Low
ms	Malay	Latin	Austronesian	Malayo-Polynesian	Mid
my	Burmese	Myanmar	Sino-Tibetan	Burmo-Qiangic	Low
nl	Dutch	Latin	Indo-European	Germanic	High
pl	Polish	Latin	Indo-European	Balto-Slavic	High
pt	Portuguese	Latin	Indo-European	Italic	High
ru	Russian	Cyrillic	Indo-European	Balto-Slavic	High
th	Thai	Thai	Tai-Kadai	Kam-Tai	Mid
tl	Tagalog	Latin	Austronesian	Malayo-Polynesian	Mid
tr	Turkish	Latin	Turkic	Common Turkic	High
ur	Urdu	Arabic	Indo-European	Indo-Aryan	Mid
vi	Vietnamese	Latin	Austroasiatic	Vietic	High
zh	Chinese	Han	Sino-Tibetan	Sinitic	High

Table 3: 28 languages supported by our model. The resource of each language is determined according to the taxonomy classes by Joshi et al. (2020).

Prompt
Translate this from [src] into [tgt]. Do not provide any explanations or text apart from the translation. \n[src]: <SRC>\n[tgt]:

Table 4: Prompt strategy adopted when using GPTs, where [src] and [tgt] denote the source and target languages, and <SRC> denotes the source sentence.

Direction	Mistral-7B	Qwen2-7B	Qwen2.5-7B	Llama3-8B	Llama3.1-8B	Gemma2-9B	GPT3.5-turbo	GPT4-turbo	Google
en→ar	41.11 / 31.29	59.12 / 52.88	68.34 / 62.36	65.78 / 58.86	66.15 / 60.33	73.16 / 68.69	77.18 / 72.84	81.45 / 77.78	77.23 / 72.53
en→bn	28.37 / 18.93	39.30 / 36.79	38.51 / 35.22	59.02 / 57.20	56.73 / 54.98	63.46 / 62.29	60.54 / 59.89	74.93 / 72.25	67.36 / 66.82
en→cs	59.94 / 47.11	53.64 / 40.42	57.05 / 45.94	67.00 / 55.98	67.24 / 56.96	76.29 / 68.19	78.69 / 70.6	82.58 / 76.33	77.69 / 69.08
en→de	85.70 / 62.46	84.15 / 60.77	86.52 / 65.85	87.59 / 68.15	87.94 / 68.69	90.17 / 73.60	92.13 / 77.03	93.05 / 79.16	93.63 / 79.26
en→es	79.67 / 66.42	81.22 / 68.50	82.19 / 69.88	83.19 / 70.74	83.07 / 70.57	85.25 / 74.24	86.95 / 76.08	86.75 / 77.13	86.32 / 75.20
en→fa	28.82 / 18.50	47.61 / 42.86	44.90 / 39.93	65.23 / 61.14	65.46 / 61.87	75.12 / 72.24	72.88 / 71.01	79.45 / 77.95	73.06 / 69.65
en→fr	74.59 / 63.03	76.15 / 65.68	77.95 / 67.55	77.14 / 66.72	76.47 / 66.50	81.13 / 72.24	83.06 / 74.37	84.16 / 76.99	81.47 / 73.22
en→he	29.06 / 14.47	48.81 / 38.16	48.24 / 37.52	58.44 / 50.15	60.01 / 52.30	70.53 / 63.78	69.98 / 63.37	77.07 / 71.88	74.62 / 68.29
en→hi	32.57 / 26.45	37.75 / 34.44	37.76 / 35.11	57.23 / 55.32	58.60 / 56.92	62.72 / 62.67	62.27 / 61.69	72.81 / 67.88	77.30 / 71.63
en→id	64.89 / 49.71	74.99 / 61.66	75.48 / 65.49	76.96 / 66.28	77.13 / 66.55	81.35 / 71.79	84.08 / 75.75	86.05 / 79.16	83.11 / 76.00
en→it	76.55 / 63.93	77.36 / 64.60	79.43 / 67.36	79.90 / 68.23	80.62 / 69.44	84.41 / 74.70	85.06 / 74.95	85.29 / 77.93	83.82 / 74.54
en→ja	59.27 / 61.91	71.76 / 74.69	73.05 / 76.46	72.84 / 74.77	72.16 / 74.95	80.98 / 81.40	82.50 / 82.05	83.92 / 84.03	86.64 / 86.18
en→km	10.94 / 2.85	18.58 / 20.69	17.36 / 18.69	25.18 / 29.73	27.12 / 32.56	39.10 / 49.53	31.89 / 36.23	62.80 / 69.46	65.88 / 72.85
en→ko	59.43 / 57.74	67.71 / 65.51	71.72 / 70.04	73.74 / 69.83	72.80 / 70.55	76.24 / 71.73	81.94 / 78.78	85.85 / 82.59	80.49 / 78.07
en→lo	13.64 / 2.65	14.30 / 5.72	17.22 / 7.76	16.58 / 9.72	15.85 / 9.23	33.34 / 32.53	24.18 / 21.16	47.73 / 50.05	61.67 / 63.59
en→ms	60.97 / 46.86	70.17 / 57.46	70.32 / 60.10	72.30 / 62.10	72.17 / 62.45	76.49 / 66.87	81.17 / 73.19	83.00 / 76.39	76.25 / 67.71
en→my	11.72 / 1.03	14.43 / 10.11	11.48 / 6.31	23.52 / 25.33	24.49 / 28.24	30.28 / 37.82	22.90 / 22.33	56.69 / 63.26	59.83 / 70.76
en→nl	78.58 / 62.17	78.46 / 63.04	79.07 / 64.96	83.81 / 71.15	84.39 / 71.31	87.27 / 74.96	89.88 / 79.84	90.52 / 81.30	88.51 / 78.86
en→pl	66.30 / 52.06	63.52 / 48.74	65.01 / 53.03	72.70 / 60.18	73.06 / 60.44	81.96 / 71.46	83.43 / 74.34	85.69 / 77.81	82.59 / 74.22
en→pt	78.42 / 65.55	80.69 / 68.02	82.57 / 70.61	81.07 / 68.80	80.81 / 68.49	84.82 / 74.23	87.17 / 77.15	88.18 / 79.70	86.42 / 75.78
en→ru	71.72 / 61.82	75.28 / 67.42	76.32 / 68.88	75.63 / 67.59	76.34 / 67.87	80.52 / 74.01	81.18 / 74.67	84.47 / 78.66	81.05 / 74.40
en→th	34.48 / 25.08	66.46 / 63.25	71.25 / 67.34	67.95 / 62.92	64.96 / 59.28	76.44 / 72.16	68.97 / 64.97	80.55 / 77.17	75.48 / 71.67
en→tl	46.74 / 36.71	43.39 / 29.25	38.95 / 25.96	57.58 / 50.65	57.01 / 51.54	67.46 / 63.14	75.02 / 72.58	75.69 / 75.55	71.46 / 64.89
en→tr	40.88 / 31.49	47.60 / 42.15	51.90 / 49.52	60.78 / 57.13	63.12 / 60.30	72.00 / 69.17	75.66 / 73.42	77.72 / 76.69	77.22 / 75.74
en→ur	24.00 / 17.08	25.92 / 21.18	25.30 / 18.12	44.41 / 43.99	49.00 / 51.48	57.71 / 60.42	59.83 / 63.48	71.68 / 73.64	65.71 / 67.00
en→vi	54.44 / 42.18	72.71 / 64.65	76.98 / 70.68	75.09 / 67.74	74.17 / 68.09	78.98 / 73.84	78.92 / 74.25	83.17 / 79.10	79.35 / 75.31
en→zh	69.74 / 63.57	78.18 / 72.73	77.92 / 72.35	74.85 / 68.23	74.32 / 68.26	78.34 / 72.52	79.12 / 73.59	81.20 / 77.05	82.04 / 77.64

Table 5: Evaluation results (XCOMET / COMETKiwi) on the WMT-24 benchmark. Note that the translation performance of the open LLMs is based on the 5-shot in-context learning strategy.

Direction	Mistral-7B	Qwen2-7B	Qwen2.5-7B	Llama3-8B	Llama3.1-8B	Gemma2-9B	NLLB-54B	GPT3.5-turbo	GPT4-turbo	Google
ar→en	29.99 / 82.55	38.92 / 86.46	40.60 / 86.94	41.42 / 87.01	41.94 / 87.01	46.04 / 87.94	48.26 / 88.13	41.36 / 87.23	45.94 / 88.29	50.59 / 88.73
ar←en	8.66 / 63.97	21.76 / 78.71	27.11 / 83.13	27.65 / 82.86	28.63 / 83.4	35.87 / 86.02	43.03 / 87.22	37.80 / 86.70	40.56 / 87.86	47.80 / 88.72
bn→en	17.28 / 78.07	28.01 / 85.6	27.96 / 85.75	32.74 / 87.25	32.42 / 87.15	38.52 / 88.61	42.22 / 89.31	28.26 / 85.52	39.00 / 89.16	43.52 / 89.62
bn←en	3.06 / 44.88	9.1 / 66.48	8.14 / 64.31	21.85 / 80.71	21.58 / 80.24	23.24 / 82.58	36.04 / 87.08	22.00 / 88.02	31.80 / 86.98	37.49 / 86.60
cs→en	41.61 / 87.91	40.78 / 87.52	42.18 / 87.87	43.39 / 88.23	43.87 / 88.39	46.53 / 88.96	45.31 / 88.67	44.63 / 88.73	45.87 / 88.97	48.68 / 89.34
cs←en	28.17 / 85.32	24.17 / 81.43	24.87 / 82.34	32.41 / 87.57	33.35 / 87.97	36.75 / 90.45	42.38 / 91.54	40.66 / 91.36	42.49 / 92.13	45.95 / 91.51
de→en	46.23 / 88.99	45.96 / 88.95	45.09 / 88.92	47.29 / 88.99	47.72 / 89.18	50.72 / 89.62	49.80 / 89.36	48.88 / 89.61	49.84 / 89.80	52.80 / 90.19
de←en	34.35 / 84.6	33.12 / 83.78	34.51 / 84.81	38.54 / 86.06	39.39 / 86.2	45.01 / 87.88	46.62 / 88.10	47.30 / 88.46	48.93 / 88.90	48.27 / 89.29
es→en	34.94 / 86.96	35.24 / 87.06	36.05 / 87.15	36.95 / 87.17	37.34 / 87.24	39.02 / 87.47	38.80 / 87.49	34.73 / 87.22	35.86 / 87.55	37.35 / 87.53
es←en	28.8 / 84.81	29.26 / 85.24	30.14 / 85.56	30.55 / 85.92	30.92 / 85.96	33.45 / 86.67	33.05 / 86.39	33.35 / 86.87	33.89 / 87.22	35.23 / 87.24
fa→en	24.54 / 81.42	33.47 / 86.21	33.80 / 86.12	37.92 / 87.55	38.19 / 87.58	43.07 / 88.65	44.19 / 88.70	36.84 / 87.34	41.32 / 88.78	44.68 / 88.90
fa←en	4.49 / 48.09	13.49 / 72.72	13.41 / 70.40	25.68 / 83.71	26.18 / 84.11	32.19 / 87.31	36.13 / 87.75	29.39 / 86.16	34.51 / 88.27	39.72 / 88.34
fr→en	47.38 / 89.01	47.41 / 89.02	47.79 / 89.15	49.17 / 89.01	49.57 / 89.19	51.93 / 89.63	51.54 / 89.44	49.90 / 89.50	50.63 / 89.71	53.44 / 89.89
fr←en	45.66 / 85.95	46.32 / 86.59	47.23 / 86.80	49.43 / 87.13	50.12 / 87.2	55.12 / 88.36	56.16 / 88.15	56.18 / 88.69	56.12 / 89.00	58.98 / 89.23
he→en	24.07 / 76.46	40.28 / 86.57	41.28 / 86.70	44.98 / 87.81	45.02 / 87.84	49.33 / 88.80	49.02 / 88.62	41.53 / 86.75	47.43 / 88.77	51.80 / 89.19
he←en	3.99 / 46.57	17.12 / 74.23	15.95 / 71.86	30.32 / 83.23	30.67 / 83.55	34.09 / 85.84	46.82 / 88.71	34.48 / 84.88	42.42 / 88.49	48.80 / 88.88
hi→en	23.34 / 82.32	33.57 / 87.3	33.87 / 87.12	39.15 / 88.97	39.25 / 88.97	45.56 / 89.79	47.26 / 90.27	36.75 / 88.14	44.53 / 89.91	51.52 / 90.99
hi←en	6.32 / 49.27	11.79 / 61.07	12.18 / 61.34	25.77 / 75	27.56 / 75.92	35.33 / 79.07	40.55 / 81.17	29.09 / 77.46	36.63 / 81.47	38.94 / 83.13
id→en	41.26 / 87.9	44.26 / 88.94	45.78 / 89.16	46.41 / 89.16	46.72 / 89.29	50.87 / 90.00	49.88 / 89.37	46.64 / 89.46	50.29 / 90.10	53.25 / 90.29
id←en	24.9 / 82.61	34.75 / 88.5	37.40 / 89.19	41.02 / 90.06	41.4 / 90.15	48.53 / 91.55	49.18 / 91.18	48.93 / 91.88	49.21 / 92.34	52.76 / 92.61
it→en	38.18 / 87.76	38.32 / 87.82	38.89 / 87.90	39.96 / 87.93	40.1 / 88.1	41.9 / 88.52	41.56 / 88.13	37.52 / 88.01	38.54 / 88.28	41.31 / 88.54
it←en	30.19 / 86.13	28.35 / 85.71	29.08 / 86.33	32.98 / 87.39	33.63 / 87.53	36.39 / 88.56	38.27 / 88.58	37.27 / 88.68	38.45 / 89.23	39.86 / 89.38
ja→en	26.03 / 85.79	30.11 / 87.43	31.01 / 87.64	30.89 / 87.59	30.81 / 87.5	34.51 / 88.24	34.54 / 87.92	30.68 / 87.84	32.88 / 88.49	38.14 / 89.36
ja←en	13.25 / 82.79	21.12 / 88.66	24.30 / 89.34	24.45 / 89.04	24.54 / 89.01	30.4 / 91.03	20.07 / 89.08	31.20 / 90.90	32.17 / 91.52	37.08 / 92.75
km→en	7.82 / 60.54	16.21 / 77.29	14.79 / 74.12	23.86 / 82.01	23.98 / 82.56	30.3 / 84.77	38.64 / 87.02	14.11 / 72.53	32.73 / 86.70	35.10 / 85.82
km←en	0.19 / 30.09	1.9 / 45.18	1.59 / 41.26	4.82 / 56.06	5.44 / 57.08	7.15 / 66.08	22.97 / 79.86	5.76 / 54.80	20.50 / 80.96	27.36 / 83.37
ko→en	26.42 / 85.78	30.3 / 87.18	31.78 / 87.59	31.27 / 87.24	31.75 / 87.34	35.96 / 88.47	35.42 / 88.03	31.74 / 87.78	35.49 / 88.83	39.31 / 89.32
ko←en	10.1 / 79.37	15.49 / 85.14	17.71 / 86.23	19.31 / 86.46	20.21 / 86.98	23.24 / 88.1	26.72 / 89.47	25.56 / 89.09	29.09 / 90.16	31.31 / 90.24
lo→en	4.92 / 53.28	9.64 / 64.26	11.03 / 64.81	14.57 / 71.85	15.17 / 72.33	27.13 / 80.54	42.40 / 87.67	10.50 / 64.18	24.87 / 79.58	44.05 / 88.53
lo←en	0.1 / 31.6	0.86 / 35.42	0.92 / 33.22	1.49 / 36.92	1.47 / 37.9	5.9 / 55.2	29.59 / 84.10	3.72 / 45.35	13.01 / 67.80	29.59 / 83.19
ms→en	40.32 / 87.03	42.12 / 87.84	42.88 / 87.87	46.05 / 88.71	46.37 / 88.66	51.42 / 89.57	51.20 / 89.21	46.61 / 88.92	50.27 / 89.51	53.29 / 89.76
ms←en	20.16 / 79.33	25.95 / 84.98	25.08 / 85.03	35.99 / 87.29	36.34 / 87.29	41.95 / 89.15	45.54 / 89.01	41.77 / 89.50	43.88 / 90.31	47.41 / 89.91
my→en	2.22 / 54.38	6.12 / 69.18	4.63 / 64.70	17.69 / 81.08	18.44 / 81.62	24.75 / 84.00	34.73 / 87.50	3.68 / 60.27	25.38 / 85.60	34.97 / 87.23
my←en	0.09 / 31.99	1.39 / 42.81	0.74 / 36.87	6.63 / 64	7.13 / 65.03	6.05 / 69.63	17.74 / 84.58	2.28 / 51.78	16.82 / 85.53	24.45 / 87.50
nl→en	35.02 / 86.93	35.43 / 87.06	35.24 / 87.15	35.87 / 86.67	36.32 / 87.1	38.14 / 87.65	38.94 / 87.62	37.01 / 87.69	37.06 / 87.05	39.68 / 87.97
nl←en	27.82 / 84.88	26.23 / 84.2	26.11 / 84.45	30.49 / 86.64	31 / 86.63	33.48 / 87.67	35.61 / 87.74	36.28 / 88.39	37.02 / 88.71	37.58 / 88.72
pl→en	33.71 / 85.8	33.58 / 85.71	33.86 / 85.75	34.75 / 85.85	34.82 / 85.9	37.1 / 86.76	36.68 / 86.35	34.18 / 86.29	35.37 / 86.54	38.05 / 86.93
pl←en	23.58 / 84.69	19.87 / 81.34	20.61 / 82.11	26.18 / 86.96	26.14 / 87.1	29.78 / 89.07	32.54 / 89.38	32.50 / 89.62	33.78 / 90.43	36.69 / 90.28
pt→en	50.5 / 89.19	51.66 / 89.29	52.42 / 89.43	52.86 / 89.34	53.35 / 89.46	56.49 / 90.06	55.19 / 89.57	54.10 / 89.86	55.24 / 90.05	57.44 / 90.30
pt←en	45.43 / 87.36	45.59 / 88.17	47.61 / 88.57	49.34 / 88.67	49.46 / 88.71	54.04 / 89.78	52.87 / 89.26	56.57 / 90.27	56.42 / 90.52	56.98 / 90.58
ru→en	38.65 / 86.25	38.25 / 86.33	39.07 / 86.63	39.68 / 86.37	39.9 / 86.48	42.51 / 87.06	42.21 / 86.98	38.98 / 86.65	40.68 / 87.13	43.95 / 87.52
ru←en	31.17 / 86.29	31.25 / 86.93	32.42 / 87.61	34.36 / 87.82	34.88 / 87.77	38.61 / 89.46	41.03 / 89.45	38.76 / 89.47	41.05 / 90.33	44.35 / 90.60
th→en	20.03 / 80.62	31.91 / 87.25	33.40 / 87.76	33.91 / 87.68	34.36 / 87.71	37.92 / 88.59	36.87 / 87.88	26.94 / 85.87	35.23 / 88.71	36.96 / 88.07
th←en	6.01 / 55.55	27.29 / 83.73	31.36 / 85.11	31.13 / 84.93	30.69 / 83.82	37.71 / 87.72	35.08 / 85.71	32.67 / 84.78	41.70 / 88.88	46.31 / 88.77
tl→en	36.99 / 82.95	40.03 / 84.37	38.55 / 83.14	45.97 / 86.24	46.36 / 86.45	52.44 / 88.01	54.58 / 88.38	46.47 / 87.16	52.59 / 88.38	57.18 / 88.78
tl←en	11.21 / 66.19	11.08 / 67.32	8.10 / 61.47	25.45 / 79.77	26.19 / 79.67	33.09 / 83.87	38.34 / 84.79	34.06 / 84.88	39.89 / 86.10	39.76 / 84.79
tr→en	28.22 / 84.38	33.75 / 86.91	34.72 / 87.38	38.42 / 88.64	38.73 / 88.6	43.68 / 89.68	45.82 / 89.84	41.72 / 89.40	44.34 / 90.08	48.79 / 90.48
tr←en	8.35 / 65.91	13.06 / 77.01	18.33 / 80.25	24.86 / 84.88	25.52 / 85.26	34.7 / 88.63	41.52 / 89.84	38.60 / 89.72	39.66 / 90.52	45.38 / 91.25
ur→en	14.07 / 73.73	25.92 / 83.11	24.44 / 82.00	31.89 / 85.57	32.82 / 85.98	38.73 / 87.45	43.08 / 88.30	29.58 / 84.60	39.97 / 88.37	43.26 / 88.39
ur←en	1.83 / 39.69	3.59 / 53.02	2.30 / 46.69	13.05 / 71.02	16.64 / 74.04	20.58 / 78.4	30.52 / 81.75	21.95 / 78.45	28.99 / 83.37	32.46 / 83.05
vi→en	31.87 / 84.55	38.52 / 87.11	39.39 / 87.43	39.79 / 87.14	40.49 / 87.37	43.55 / 88.04	43.75 / 87.79	37.72 / 87.09	41.70 / 88.06	45.40 / 88.40
vi←en	18.83 / 74.06	34.62 / 86.56	37.17 / 87.84	38.24 / 87.46	38.37 / 87.66	42.31 / 89.01	43.30 / 88.19	40.97 / 88.77	43.53 / 89.82	47.08 / 89.90
zh→en	29.33 / 85.88	34.2 / 87.22	34.80 / 87.23	32.66 / 86.77	32.66 / 86.86	35.96 / 87.52	36.13 / 87.16	31.75 / 87.07	33.83 / 87.68	39.82 / 88.46
zh←en	23.28 / 83.37	37.11 / 88.43	37.06 / 88.36	32.2 / 86.99	32.5 / 87.04	37.28 / 88.42	26.61 / 82.25	36.15 / 88.22	37.42 / 88.65	43.36 / 89.94

Table 6: English-centric evaluation results (spBLEU / COMET) on the FLORES-200 benchmark. Note that the translation performance of the open LLMs is based on the 5-shot in-context learning strategy.

Direction	Mistral-7B	Qwen2-7B	Qwen2.5-7B	Llama3-8B	Llama3.1-8B	Gemma2-9B	NLLB-54B	GPT3.5-turbo	GPT4-turbo	Google
ar→zh	6.78 / 67.75	26.64 / 84.24	27.49 / 84.64	22.01 / 82.54	22.84 / 82.66	28.54 / 84.87	21.25 / 80.25	25.89 / 83.51	29.23 / 85.37	36.21 / 86.50
ar←zh	2.91 / 56.15	14.7 / 77.21	16.53 / 80.10	14.09 / 78.08	15.96 / 79.07	22.66 / 82.16	27.53 / 84.08	22.93 / 83.18	25.99 / 84.65	32.32 / 85.30
bn→zh	5.16 / 65.41	21.86 / 82.77	21.24 / 82.69	20.77 / 82.3	21.35 / 82.68	25.39 / 84.94	20.01 / 81.33	18.63 / 79.78	26.17 / 85.40	33.55 / 87.07
bn←zh	0.74 / 37.07	5.54 / 61.44	4.85 / 58.94	12.81 / 74.3	12.71 / 73.98	13.27 / 76.97	23.67 / 81.96	14.39 / 75.60	21.26 / 82.51	27.64 / 82.38
cs→zh	17.4 / 80.41	28.93 / 85.82	28.40 / 85.64	25.2 / 84.51	25.62 / 84.74	30.42 / 86.25	19.60 / 79.15	27.77 / 85.60	29.31 / 86.31	36.19 / 87.51
cs←zh	11.15 / 75.53	14.45 / 79.42	14.07 / 80.08	18.13 / 85.07	19.24 / 86.15	22.17 / 87.39	25.63 / 88.81	24.02 / 88.19	26.48 / 90.22	30.65 / 89.81
de→zh	19.32 / 81.31	30.96 / 86.8	30.68 / 86.81	27.33 / 85.5	27.42 / 85.51	31.82 / 86.93	21.48 / 80.87	30.08 / 86.60	31.27 / 87.23	37.03 / 87.97
de←zh	16.15 / 77.83	20.08 / 80.72	21.06 / 81.75	21.2 / 82.14	22.08 / 82.5	24.65 / 83.03	27.15 / 84.27	27.27 / 85.03	29.29 / 85.87	32.97 / 86.27
en→zh	23.28 / 83.37	37.11 / 88.43	37.06 / 88.36	32.2 / 86.99	32.5 / 87.04	37.28 / 88.42	26.61 / 82.25	36.15 / 88.22	37.42 / 88.65	43.36 / 89.94
en←zh	29.33 / 85.88	34.2 / 87.22	34.80 / 87.23	32.66 / 86.77	32.66 / 86.86	35.96 / 87.52	36.13 / 87.16	31.75 / 87.07	33.83 / 87.68	39.82 / 88.46
es→zh	16.91 / 81.54	27.93 / 86.63	28.40 / 86.76	24.26 / 85.33	24.77 / 85.61	28.02 / 86.81	19.03 / 79.85	26.70 / 86.20	27.73 / 86.82	32.74 / 87.24
es←zh	15.06 / 80.17	20.56 / 83.72	20.93 / 83.94	20.09 / 83.44	20.65 / 83.64	23.26 / 84.64	23.69 / 84.46	22.80 / 84.86	24.05 / 85.54	25.59 / 85.47
fa→zh	9.14 / 71.62	24.98 / 84.44	24.60 / 84.16	23.02 / 83.81	23.15 / 84.16	27.38 / 85.66	21.89 / 81.66	23.84 / 83.91	27.75 / 85.86	34.43 / 87.16
fa←zh	1.67 / 43.92	9.43 / 71	8.21 / 68.32	14.84 / 79.92	16.61 / 81.01	21.16 / 83.97	23.10 / 84.43	18.41 / 83.31	22.65 / 85.52	27.68 / 85.42
fr→zh	19.01 / 81.3	31.89 / 87	30.97 / 86.91	26.6 / 85.4	27.07 / 85.6	31.56 / 86.88	21.19 / 80.44	30.39 / 86.51	31.38 / 87.14	37.71 / 87.90
fr←zh	22.32 / 79.96	28.14 / 83.33	28.62 / 83.24	26.7 / 82.68	27.68 / 83.04	28.42 / 81.98	34.16 / 84.28	32.18 / 84.73	34.13 / 85.59	38.24 / 85.89
he→zh	7.65 / 66.82	26.08 / 83.58	25.99 / 83.70	23.36 / 82.85	24.38 / 83.12	29.31 / 85.17	19.77 / 78.50	24.33 / 82.12	29.37 / 85.37	36.89 / 86.80
he←zh	1.09 / 41.27	8.92 / 71.81	7.65 / 69.60	16.25 / 78.76	16.4 / 79.35	18.61 / 81.91	27.04 / 84.87	18.24 / 80.93	24.39 / 84.79	31.91 / 85.77
hi→zh	7.18 / 69.31	24.34 / 83.71	23.66 / 83.41	23.35 / 83.75	23.58 / 84.09	27.86 / 85.47	21.12 / 81.60	22.73 / 82.72	28.36 / 85.82	35.64 / 87.51
hi←zh	1.93 / 40.23	6.12 / 52.79	7.02 / 54.20	14.35 / 67.57	14.97 / 67.84	17.84 / 70.81	24.20 / W73.81	17.03 / 70.19	22.55 / 74.74	28.80 / 75.99
id→zh	16.48 / 79.47	30.22 / 86.13	30.57 / 86.24	26.25 / 84.81	26.79 / 85.12	31.15 / 86.49	21.38 / 79.92	28.67 / 85.57	30.64 / 86.48	37.24 / 87.66
id←zh	7.72 / 75.08	20.73 / 85.59	22.22 / 86.64	21.6 / 86.24	22.62 / 86.77	26.44 / 87.2	29.55 / 87.99	26.22 / 88.31	27.44 / 89.01	33.18 / 89.40
it→zh	17.95 / 81.48	29.17 / 87.01	29.20 / 86.70	24.87 / 85.33	25.51 / 85.47	29.14 / 86.83	19.33 / 79.68	27.86 / 86.35	29.11 / 87.17	34.59 / 87.65
it←zh	14.88 / 80.28	19.26 / 84.04	19.42 / 84.48	20.1 / 84.58	21.04 / 84.9	23.9 / 85.94	25.77 / 85.73	24.90 / 86.49	26.92 / 87.35	28.98 / 87.31
ja→zh	15.36 / 81.36	25.74 / 87.51	26.62 / 87.54	21.79 / 85.61	22.32 / 85.9	26.06 / 87.4	19.64 / 81.81	24.35 / 87.09	26.58 / 87.87	33.54 / 88.71
ja←zh	8.58 / 81.35	16.65 / 87.72	18.59 / 88.68	17.42 / 87.32	18.15 / 87.69	21.51 / 89.08	16.78 / 87.83	22.16 / 89.86	23.46 / 90.37	28.21 / 90.97
km→zh	1.26 / 49.34	13.74 / 76.05	10.52 / 71.66	14.78 / 78.29	16.12 / 78.85	20.82 / 82.03	21.33 / 83.00	7.22 / 65.87	22.29 / 83.54	29.39 / 85.41
km←zh	0.14 / 28.15	1.47 / 42.66	1.38 / 39.96	3.25 / 51.52	3.81 / 52.31	4.51 / 60.19	17.55 / 79.83	4.66 / 52.10	16.83 / 79.29	22.26 / 80.68
ko→zh	14.79 / 79.61	26.7 / 86.41	27.12 / 86.50	23.4 / 84.71	24.23 / 85.17	26.74 / 86.19	21.31 / 83.01	24.84 / 85.43	27.72 / 86.92	34.03 / 87.87
ko←zh	4.39 / 73.84	11.77 / 83.31	13.79 / 85.00	14.25 / 84.34	15.65 / 85.14	16.98 / 85.8	20.39 / 86.86	17.98 / 86.90	20.93 / 87.98	24.80 / 88.12
lo→zh	0.94 / 46.14	6.6 / 62.99	8.62 / 63.95	8.02 / 68.08	9.25 / 69.03	16.84 / 77.22	23.44 / 83.71	5.29 / 58.29	15.29 / 75.63	33.45 / 86.67
lo←zh	0.09 / 30.12	0.51 / 34.27	0.58 / 32.13	0.74 / 34.22	0.89 / 34.62	2.67 / 47.97	24.30 / 82.08	2.36 / 43.35	9.97 / 65.14	22.51 / 80.22
ms→zh	15.32 / 78.42	27.99 / 84.8	28.33 / 84.74	24.98 / 83.76	26.02 / 84.16	30.16 / 85.66	23.16 / 80.89	26.20 / 84.05	29.61 / 85.65	36.75 / 87.02
ms←zh	6.28 / 72	14.14 / 82.04	14.67 / 82.71	18.48 / 83.34	19.16 / 83.82	21.86 / 84.93	26.33 / 85.77	22.15 / 85.79	23.69 / 86.57	30.23 / 86.50
my→zh	0.21 / 43.3	5.8 / 69.34	3.99 / 63.66	12.51 / 77.49	13.76 / 78.56	16.98 / 80.35	16.77 / 80.42	1.18 / 53.81	17.64 / 81.68	27.72 / 85.02
my←zh	0.02 / 29.98	1.2 / 40.85	0.68 / 34.90	4.72 / 56.48	5.16 / 58.1	3.88 / 62.3	15.79 / 82.04	1.93 / 48.87	13.94 / 82.94	20.33 / 84.40
nl→zh	15.9 / 80.08	27.38 / 85.72	27.21 / 85.61	23.79 / 84.21	24.22 / 84.54	27.29 / 85.8	16.61 / 77.64	26.20 / 85.37	27.16 / 85.72	32.10 / 86.66
nl←zh	13.09 / 77.7	17.57 / 81.87	17.77 / 82.17	19.2 / 82.99	19.78 / 83.7	21.48 / 84.03	24.74 / 85.31	23.76 / 85.94	25.36 / 86.53	28.00 / 86.50
pl→zh	16.04 / 79.7	26.73 / 85.3	26.58 / 85.14	22.87 / 83.61	23.81 / 84.04	27.42 / 85.44	17.37 / 77.75	25.34 / 84.86	27.34 / 85.72	32.27 / 86.52
pl←zh	13.15 / 78.46	15.14 / 81.76	14.70 / 81.59	17.6 / 85.21	18.1 / 85.73	21 / 86.86	23.19 / 87.82	22.68 / 88.13	24.72 / 89.37	27.95 / 89.29
pt→zh	18.81 / 81.45	31.29 / 87.27	31.02 / 87.15	26.51 / 85.63	27.47 / 85.94	31.31 / 87.25	20.54 / 80.02	29.87 / 86.66	31.29 / 87.31	38.01 / 88.43
pt←zh	19.15 / 81.99	25.5 / 85.6	27.19 / 85.78	25.14 / 85.13	25.98 / 85.54	29.62 / 86.11	31.17 / 86.39	30.29 / 86.83	31.58 / 87.47	34.84 / 87.67
ru→zh	15.22 / 78.65	29.07 / 85.74	29.43 / 85.88	24.57 / 83.81	25.1 / 84.1	29.61 / 85.43	21.22 / 80.67	27.41 / 85.03	29.65 / 86.07	35.13 / 86.64
ru←zh	13.36 / 75.88	20.64 / 85.83	22.11 / 86.42	20.91 / 85.79	21.69 / 86.52	24.22 / 86.68	28.21 / 87.99	24.98 / 87.63	27.34 / 88.72	31.52 / 88.99
th→zh	4.48 / 65.73	26.43 / 86.56	27.00 / 86.69	22.56 / 84.69	23.2 / 85.15	27.49 / 86.46	20.07 / 82.01	20.40 / 83.01	27.39 / 87.12	31.78 / 87.03
th←zh	2.15 / 55.72	21.29 / 81.02	24.81 / 83.01	24.55 / 82.23	22.19 / 80.98	29.39 / 85	30.20 / 84.27	26.41 / 82.90	33.84 / 86.54	38.26 / 86.75
tl→zh	13.26 / 73.48	25.35 / 81.22	23.90 / 80.34	22.66 / 80.99	23.82 / 81.47	29.7 / 84.14	21.61 / 79.12	25.25 / 82.25	29.70 / 84.57	37.18 / 86.02
tl←zh	3.51 / 60.23	4.86 / 62.35	3.51 / 57.10	10.59 / 73.37	12.24 / 74.61	19.13 / 80.13	22.81 / 80.87	19.48 / 81.19	23.70 / 82.43	25.30 / 81.40
tr→zh	13.63 / 76.11	26.32 / 83.84	26.51 / 84.42	23.87 / 83.59	24.54 / 83.95	28.63 / 85.39	20.27 / 79.47	26.88 / 84.63	29.06 / 85.85	36.06 / 87.42
tr←zh	3.58 / 59.25	10.44 / 74.55	10.88 / 75.85	13.37 / 79.38	16.72 / 80.6	20.64 / 82.8	24.82 / 84.47	23.21 / 85.08	23.97 / 86.14	30.19 / 86.85
ur→zh	2.77 / 58.04	19.76 / 80	18.69 / 79.35	20.07 / 81.06	21.24 / 81.99	24.59 / 83.9	21.79 / 82.21	19.59 / 79.80	26.77 / 84.92	33.02 / 86.07
ur←zh	0.59 / 34.17	2.14 / 48.52	1.17 / 40.56	6.72 / 63.72	8.66 / 68.09	11.31 / 73.18	20.39 / 78.26	13.85 / 73.79	19.96 / 79.25	23.44 / 78.96
vi→zh	14.77 / 79.62	28.11 / 86.63	29.17 / 86.72	24.38 / 85.22	25.15 / 85.62	28.37 / 86.69	21.79 / 82.34	25.24 / 85.79	28.47 / 86.83	35.14 / 87.95
vi←zh	7.88 / 70.03	24.52 / 85.44	26.99 / 87.00	24.78 / 85.28	25.94 / 85.99	28.15 / 86.12	31.75 / 87.63	28.04 / 87.21	29.93 / 88.36	34.69 / 88.74

Table 7: Chinese-centric evaluation results (spBLEU / COMET) on the FLORES-200 benchmark. Note that the translation performance of the open LLMs is based on the 5-shot in-context learning strategy.

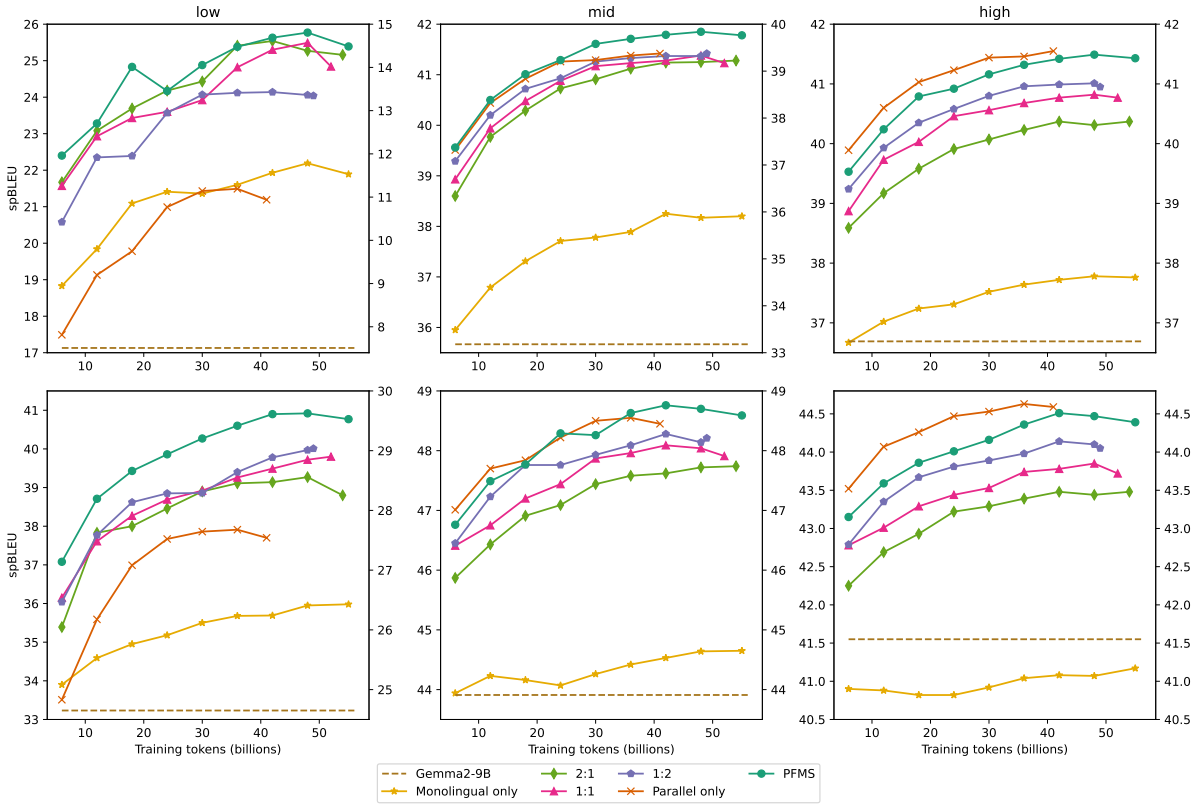

Figure 5: The translation performance (BLEU) of models trained with different data recipes during continual pretraining on low-resource (left), mid-resource (middle), and high-resource (right) languages. The upper subfigures illustrate the en→xx translation performance, while the lower subfigures depict the xx→en translation performance. Note that "Gemma2-9B" refers to the direct finetuning of the model without continual pretraining, and its performance is reflected in the right-hand y-axis.

Training batch size on each device	4
Number of GPUs	32
Gradient Accumulation Steps	6
Maximum Sequence Length	2048
Number of Epochs	1
Learning Rate	2e-5
LR Scheduler	cosine
Finetuning Type	full
bf16	true
Template	empty
Maximum Gradient Norm	1.0
Warmup	0.01
Weight Decay	0.01
Optimizer	AdamW
DeepSpeed	ZeRO2

Table 8: Hyperparameter settings for the pretraining experiments.

Training batch size on each device	4
Number of GPUs	8
Gradient Accumulation Steps	8
Maximum Sequence Length	2048
Number of Epochs	1
Learning Rate	2e-5
LR Scheduler	inverse sqrt
Finetuning Type	full
bf16	true
Template	empty
Maximum Gradient Norm	1.0
Warmup	0.01
Weight Decay	0.01
Optimizer	AdamW
DeepSpeed	ZeRO2

Table 9: Hyperparameter settings for the finetuing experiments.

Direction	TowerInstruct-7B	TowerInstruct-13B	X-ALMA	Aya-101	LLaMAX3-Alpaca-8B	GemmaX2-28
en→ar	-	-	78.72 / 73.56	67.77 / 62.85	65.41 / 58.95	78.16 / 73.06
en→bn	-	-	-	64.62 / 60.78	58.3 / 56.21	70.53 / 70.86
en→cs	-	-	80.1 / 72.76	69.62 / 61.35	66.97 / 57.09	81.48 / 73.48
en→de	91.05 / 74.12	91.99 / 76.28	92.64 / 76.6	86.69 / 65.64	86.34 / 65.96	92.42 / 76.63
en→es	86.46 / 75.03	87.44 / 75.97	86.93 / 76.16	80.22 / 67.42	81.36 / 67.9	88.05 / 76.96
en→fa	-	-	78.59 / 74.49	68.46 / 65.64	65.79 / 62.31	79.84 / 75.87
en→fr	83.01 / 73.74	84.02 / 75.33	83.05 / 75.1	74.04 / 62.09	74.28 / 63.05	84.46 / 75.66
en→he	-	-	79.68 / 72.82	66.34 / 58.6	63.9 / 56.43	81.12 / 74.65
en→hi	-	-	70.12 / 64.37	61.31 / 54.32	56.35 / 55.55	66.18 / 66.03
en→id	-	-	84.53 / 76.62	77.79 / 67.8	76.17 / 63.08	83.58 / 74.95
en→it	85.87 / 75.57	86.91 / 76.71	85.53 / 76.12	77.5 / 63.45	77.78 / 66.23	87.1 / 76.89
en→ja	-	-	82.54 / 81.23	73.32 / 73.39	71.6 / 74.13	81.81 / 80.82
en→km	-	-	-	67.42 / 72.55	45.71 / 52.32	68.21 / 74.38
en→ko	80.87 / 77.51	83.83 / 80.01	82.03 / 77.65	68.96 / 65.74	74.22 / 69.81	84.73 / 80.8
en→lo	-	-	-	69.68 / 67.24	43.31 / 32.9	64.35 / 66.5
en→ms	-	-	81.13 / 73.29	76.69 / 61.22	72.73 / 61.18	78.29 / 69.64
en→my	-	-	-	58.72 / 67.7	37.44 / 41.92	67.37 / 75.74
en→nl	89.81 / 78.88	90.75 / 79.7	90.2 / 79.38	83.16 / 68.3	81.09 / 68.13	91.07 / 80.18
en→pl	-	-	83.93 / 74.21	74.44 / 62.37	70.87 / 59.21	84.64 / 74.18
en→pt	86.16 / 76.24	87.98 / 77.31	86.72 / 77.18	81.39 / 64.57	80.78 / 67.59	88.13 / 77.46
en→ru	80.39 / 72.45	82.6 / 75.19	83.09 / 75.77	74.54 / 65.96	73.71 / 65.9	82.54 / 75.36
en→th	-	-	76.73 / 71.07	72.5 / 67.44	67.83 / 62.15	77.55 / 73.14
en→tl	-	-	-	70.4 / 49.57	62.96 / 53.84	67.41 / 64.79
en→tr	-	-	76.26 / 73.11	66.91 / 63.41	63.17 / 58.22	77.8 / 75.51
en→ur	-	-	71.94 / 70.89	59.45 / 57.74	52.63 / 54.42	73.29 / 73.38
en→vi	-	-	82.48 / 78.01	71.57 / 62.53	72.89 / 66.38	82.03 / 76.85
en→zh	76.74 / 71.62	79.42 / 74.13	79.82 / 73.73	71.67 / 65.01	73.35 / 64.68	80.85 / 75.29

Table 10: Translation performance (XCOMET / COMETKiwi) of different models on the WMT-24 benchmark.

Direction	TowerInstruct-7B	TowerInstruct-13B	X-ALMA	Aya-101	LLaMAX3-Alpaca-8B	GemmaX2-28
ar→en	-	-	43.71 / 87.8	35.23 / 85.7	37.55 / 86.37	49.35 / 88.61
ar←en	-	-	40.73 / 87.74	25.07 / 82.98	27.14 / 82.49	42.9 / 87.76
bn→en	-	-	-	30.06 / 86.45	32.52 / 87.65	42.76 / 89.71
bn←en	-	-	-	20.03 / 82.27	23.08 / 81.3	34.83 / 86.59
cs→en	-	-	46.05 / 89.06	38.22 / 87.37	41 / 87.91	47.18 / 89.2
cs←en	-	-	41.93 / 91.71	31.28 / 88.51	32.18 / 87.39	42.69 / 91.64
de→en	49.18 / 89.49	50.02 / 89.63	49.7 / 89.73	42.01 / 88.13	44.21 / 88.6	49.89 / 89.61
de←en	43.68 / 87.82	45.4 / 88.16	45.95 / 88.45	34.74 / 85.17	36.33 / 85.07	47.13 / 88.5
es→en	37.67 / 87.46	37.56 / 87.68	38.1 / 87.89	32.47 / 86.34	33.91 / 86.79	37.88 / 87.71
es←en	33.09 / 86.79	33.98 / 87.05	33.67 / 87	28.2 / 84.96	28.58 / 85.13	35.45 / 87.23
fa→en	-	-	41.12 / 88.5	33.13 / 86.67	35.27 / 87.32	44.95 / 89.18
fa←en	-	-	35.46 / 88.35	26.39 / 84.98	26.63 / 84.18	38.66 / 88.88
fr→en	50.98 / 89.55	51.32 / 89.68	49.89 / 89.56	43.78 / 88.27	45.2 / 88.47	51.18 / 89.64
fr←en	53.85 / 88.42	55.65 / 88.94	55.39 / 88.73	43.15 / 85.26	45 / 85.79	57.69 / 89.11
he→en	-	-	46.34 / 88.6	38.56 / 86.67	41.22 / 87.23	51.58 / 89.37
he←en	-	-	44.45 / 89.26	26.74 / 83.65	31.36 / 83.35	46.26 / 89.26
hi→en	-	-	44.63 / 89.81	35.65 / 87.84	37.05 / 88.51	49.14 / 90.53
hi←en	-	-	38.46 / 81.45	24.29 / 76.16	27.68 / 75.99	41.27 / 81.1
id→en	-	-	48.7 / 89.85	40.92 / 88.37	43.07 / 88.83	53.27 / 90.37
id←en	-	-	49.58 / 92.07	38.06 / 89.85	38.62 / 89.18	52.61 / 92.24
it→en	40.54 / 88.4	41.17 / 88.46	40.61 / 88.46	35.28 / 87.12	35.77 / 87.39	42.06 / 88.59
it←en	36.5 / 88.79	38.17 / 89.18	37.72 / 89.16	29.4 / 86.25	31.09 / 86.46	38.83 / 89.25
ja→en	-	-	32.5 / 88.22	23.99 / 85.6	27.73 / 87.18	36.5 / 88.78
ja←en	-	-	29.58 / 91.19	21 / 88.79	22.01 / 88.94	33.26 / 91.29
km→en	-	-	-	28.08 / 83.98	28.74 / 85.69	41.83 / 88.56
km←en	-	-	-	19.48 / 81.34	14.41 / 72.92	24.13 / 84.22
ko→en	34.51 / 88.19	36.01 / 88.51	34.31 / 88.31	27.38 / 86.53	29.07 / 87.13	38.76 / 89.04
ko←en	26.39 / 89.4	28.39 / 89.87	25.74 / 89.27	16.34 / 85.94	19.62 / 86.79	30.37 / 90.13
lo→en	-	-	-	33.44 / 85.25	25.96 / 83.19	44.22 / 88.82
lo←en	-	-	-	28.85 / 83.25	10.17 / 62.95	31.53 / 84.66
ms→en	-	-	48.33 / 89.31	41.62 / 87.89	43.67 / 88.38	53.12 / 90.03
ms←en	-	-	43.23 / 89.9	34.21 / 87.11	35.7 / 87.39	47.03 / 89.89
my→en	-	-	-	20.43 / 80.38	22.75 / 84.73	36.14 / 88.21
my←en	-	-	-	15.24 / 84.99	9.9 / 73.82	20.14 / 88.15
nl→en	37.61 / 87.66	38.72 / 87.92	38.78 / 88.01	33.25 / 86.56	34.49 / 86.98	39.15 / 87.94
nl←en	35.24 / 88.31	36.41 / 88.66	35.97 / 88.59	28.72 / 86.09	29.51 / 85.66	37.54 / 88.82
pl→en	-	-	37.07 / 86.79	29.93 / 84.96	32.51 / 85.43	38.49 / 86.91
pl←en	-	-	31.89 / 89.79	24.96 / 86.52	25.35 / 85.92	33.5 / 89.88
pt→en	55.26 / 89.91	56.17 / 90.09	53.48 / 89.93	47.06 / 88.58	49.4 / 88.93	57.09 / 90.16
pt←en	49.19 / 89.42	50.76 / 89.73	53.42 / 90.1	43.34 / 87.69	45.28 / 87.88	53.17 / 90.04
ru→en	40.74 / 86.93	42.76 / 87.23	40.69 / 87.11	33.68 / 85.43	36.32 / 86.13	43.17 / 87.28
ru←en	38.37 / 89.4	40.27 / 89.93	39.9 / 89.96	31.3 / 86.85	33.42 / 87.23	41.4 / 90.13
th→en	-	-	34.65 / 88.24	27.69 / 86.2	30.73 / 87.33	40.33 / 89.25
th←en	-	-	35.87 / 87.47	32.45 / 86.12	30.79 / 84.72	42.53 / 88.7
tl→en	-	-	-	37.04 / 84.2	43.35 / 86.34	55.97 / 88.92
tl←en	-	-	-	23.43 / 80.55	27.08 / 81.19	38.15 / 84.52
tr→en	-	-	43.63 / 89.73	34.67 / 87.76	36.62 / 88.31	47.97 / 90.47
tr←en	-	-	37.61 / 89.9	28.61 / 86.93	24.55 / 85.32	42.17 / 90.52
ur→en	-	-	38.35 / 87.59	30.23 / 84.98	32.21 / 86.08	43.15 / 88.61
ur←en	-	-	29.56 / 83.38	17.78 / 77.56	20 / 75.32	30.89 / 83.93
vi→en	-	-	41.5 / 88.03	35.02 / 86.09	36.71 / 86.95	45.45 / 88.53
vi←en	-	-	44.21 / 89.64	30.71 / 84.93	36.53 / 86.59	46.65 / 89.96
zh→en	33.27 / 86.99	34.56 / 87.32	33.94 / 87.32	24.85 / 84.38	29.9 / 86.4	36.37 / 87.62
zh←en	33.9 / 87.77	36.35 / 88.45	34.43 / 88.04	24.24 / 84.81	26.96 / 85.4	41.6 / 89.18

Table 11: English-centric translation performance (spBLEU / COMET) of different models on the FLORES-200 benchmark.

Direction	TowerInstruct-7B	TowerInstruct-13B	Aya-101	LLaMAX3-Alpaca-8B	GemmaX2-28
ar→zh	-	-	19.62 / 81.35	21.34 / 82.33	34.11 / 86.37
ar←zh	-	-	16.31 / 79.75	18.35 / 80.54	27.62 / 84.39
bn→zh	-	-	15.91 / 79.83	19.3 / 83.08	31.43 / 87.14
bn←zh	-	-	11.43 / 76.82	14.78 / 77.21	24.57 / 82.68
cs→zh	-	-	19.73 / 82.54	22.82 / 83.86	35.03 / 87.46
cs←zh	-	-	17.93 / 85.82	18.86 / 86.4	28.17 / 89.81
de→zh	28.43 / 85.96	30.43 / 86.71	21.55 / 83.32	23.39 / 84.54	35.49 / 87.78
de←zh	23.93 / 83.48	26.53 / 84.66	18.52 / 81.28	21.08 / 82.65	29.29 / 85.39
es→zh	25.82 / 85.83	27.93 / 86.68	24.24 / 84.81	26.96 / 85.4	32.84 / 87.9
es←zh	21.34 / 84.21	22.84 / 84.76	24.85 / 84.38	29.9 / 86.4	25.45 / 85.54
en→zh	33.9 / 87.77	36.35 / 88.45	19.96 / 83.76	21.3 / 84.33	41.64 / 89.19
en←zh	33.27 / 86.99	34.56 / 87.32	17.63 / 82.42	19.42 / 83.81	36.37 / 87.63
fa→zh	-	-	18.66 / 82.13	20.83 / 83.74	33.12 / 87.17
fa←zh	-	-	15.92 / 81.43	17.85 / 82.72	26.12 / 86.01
fr→zh	28.84 / 86.2	31.18 / 86.94	21.77 / 83.83	23.81 / 84.59	35.68 / 87.94
fr←zh	30.07 / 83.92	31.55 / 84.96	24.29 / 80.98	26.03 / 83.06	35.25 / 85.63
he→zh	-	-	19.48 / 81.98	21.91 / 82.77	35.14 / 86.95
he←zh	-	-	13.92 / 79.22	18.36 / 81.24	28.53 / 86.03
hi→zh	-	-	19.77 / 82.8	20.59 / 83.04	33.27 / 87.18
hi←zh	-	-	13.86 / 68.98	16.89 / 70.61	25.41 / 74.49
id→zh	-	-	19.99 / 82.24	23.17 / 83.98	35.79 / 87.7
id←zh	-	-	19.72 / 86.1	20.77 / 86.26	32.34 / 89.21
it→zh	26.61 / 86.14	29.15 / 86.9	19.82 / 83.5	22.21 / 84.43	33.93 / 87.89
it←zh	22.33 / 86.17	23.68 / 86.71	17.38 / 83.25	19.88 / 84.92	27.14 / 87.36
ja→zh	-	-	18.68 / 85.02	19.6 / 85.48	32.02 / 88.6
ja←zh	-	-	16.86 / 87.91	15.73 / 88.04	26.1 / 90.55
km→zh	-	-	7.71 / 69.75	18.11 / 81.95	30.82 / 86.78
km←zh	-	-	15.84 / 78.6	9.15 / 67.7	21.99 / 81.93
ko→zh	24.11 / 85.1	27.83 / 86.64	19.58 / 83.73	21.32 / 84.95	32.62 / 88
ko←zh	17.69 / 86.5	20.11 / 87.41	11.92 / 83.3	14.37 / 85.51	22.77 / 88.13
lo→zh	-	-	17.09 / 80.82	13.85 / 79.06	31.63 / 86.91
lo←zh	-	-	18.84 / 80.25	6.28 / 58.96	25.93 / 82.88
ms→zh	-	-	18.84 / 81.08	22.45 / 82.88	35.42 / 87.41
ms←zh	-	-	17.94 / 82.92	18.16 / 83.91	28.57 / 86.72
my→zh	-	-	9.01 / 75.51	12.78 / 80.9	27.63 / 86.06
my←zh	-	-	11.44 / 81.88	7.3 / 67.8	15.21 / 84.47
nl→zh	25.34 / 84.97	27.25 / 85.99	18.37 / 81.77	20.97 / 83.19	31.65 / 86.93
nl←zh	22.46 / 85.32	23.31 / 85.84	17.88 / 82.86	19.76 / 84.14	26.92 / 86.43
pl→zh	-	-	19.48 / 82.3	20.77 / 83.18	32.2 / 86.77
pl←zh	-	-	16.36 / 84.62	17.9 / 85.49	25.72 / 89.03
pt→zh	29.02 / 86.37	31.32 / 87.19	21.02 / 83.84	23.65 / 84.7	36.88 / 88.35
pt←zh	26.18 / 85.91	28.02 / 86.73	22.4 / 83.97	25.19 / 85.53	33.01 / 87.38
ru→zh	26.81 / 84.94	29.46 / 85.66	20.65 / 82.63	22.75 / 83.3	33.73 / 86.84
ru←zh	22.88 / 86.18	26.17 / 87.99	19.18 / 85.34	22.02 / 86.89	28.57 / 88.82
th→zh	-	-	19.08 / 83.36	21.09 / 85.05	32.53 / 88.16
th←zh	-	-	25.67 / 83.94	24.4 / 82.97	35.18 / 86.75
tl→zh	-	-	18.78 / 78.78	21.36 / 81.07	35.63 / 86.02
tl←zh	-	-	13.76 / 77.03	15.74 / 77.96	24.19 / 81.87
tr→zh	-	-	19.02 / 81.57	21.59 / 83.16	34.79 / 87.29
tr←zh	-	-	16.5 / 81.75	13.94 / 81.37	26.43 / 86.05
ur→zh	-	-	16.77 / 79.86	19.15 / 81.54	31.67 / 86.48
ur←zh	-	-	10.98 / 73.4	13.57 / 73.03	21.42 / 79.73
vi→zh	-	-	20.43 / 84.24	22 / 84.89	34.26 / 87.95
vi←zh	-	-	21.31 / 83.98	24.81 / 85.98	33.79 / 88.8

Table 12: Chinese-centric translation performance (spBLEU / COMET) of different models on the FLORES-200 benchmark.