LMT: Toward Inclusive and Scalable Multilingual Machine Translation with LLMs

Anonymous ACL submission

Abstract

Large language models have significantly advanced Multilingual Machine Translation (MMT), yet scaling to many languages while keeping quality robust across directions remains challenging. In this paper, we identify a failure mode of multilingual supervised fine-tuning (SFT) on multi-way parallel data: when such data are reused symmetrically around a pivot language (e.g., English), performance on reverse directions (X $\rightarrow$ pivot) can drop substantially. We term this phenomenon Directional Degeneration and attribute it to excessive many-to-one mappings, which encourage shortcut learning. We propose **Strategic Down-sampling**, a simple yet effective method to mitigate this degeneration. In addition, we introduce **Parallel Multilingual Prompting (PMP)**, which augments translation instructions with an auxiliary parallel sentence to promote cross-lingual transfer during training and enables optional test-time enhancement when auxiliary translations are available. We further develop **LMT**, a Chinese–English-centric suite of **L**arge-scale **M**ultilingual **T**ranslation models spanning four sizes (0.6B/1.7B/4B/8B) covering 60 languages and 234 directions. Comprehensive evaluations show that LMT is competitive among open-source MMT systems, and that our 4B LMT model performs on par with or better than substantially larger baselines. We release our models to support inclusive and scalable MMT.

1 Introduction

Large language models (LLMs) have reshaped the way we build machine translation (MT) systems. Instead of training a dedicated neural MT model from scratch, a widely used approach is to adapt a foundation LLM to translation through post-training (Yang et al., 2023; Alves et al., 2024; Xu et al., 2025; Cui et al., 2025; Luo et al., 2025; Zheng et al., 2025b). This shift has substantially improved translation quality and expanded the capability frontier of MT systems. However, it also raises a key problem for multilingual MT (MMT): how can we adapt foundation LLMs for massively multilingual scale while maintaining robust performance across all translation directions?

LLM for MT	#CPT	#Langs	Zh-centric	Base Model
BigTranslate (Yang et al., 2023)	90B	102	✗	LLaMA
ALMA (Xu et al., 2024)	20B	6	✗	LLaMA-2
TowerInstruct (Alves et al., 2024)	20B	10	✗	LLaMA-2
Guo et al. (2024)	120G	3	✗	LLaMA-2
X-ALMA (Xu et al., 2025)	40B	50	✗	LLaMA-2
GemmaX2 (Cui et al., 2025)	54B	28	✓	Gemma-2
Hunyuan-MT (Zheng et al., 2025a)	-	33	✓	Hunyuan-7B
Seed-X (Cheng et al., 2025)	200B	28	✓	-
LMT (Ours)	90B	60	✓	Qwen-3

Table 1: Comparison of typical LLM-based MMT models. We summarize their total number of continued pretraining (CPT) data, supported languages, support for Zh-centric translation, and the base models used.

Most recent LLM-based MMT systems (representative examples are summarized in Table 1) follow a multi-stage training recipe. To bridge the low-resource gap evident in foundation models (see Figure 1, top), the first stage, Continued Pre-training (CPT), serves as the primary step to enhance multilingual competence through large-scale training. The second stage is supervised fine-tuning (SFT), which aligns the model to high-quality instruction-style translation. Since SFT benefits most from clean and diverse supervision (Zhou et al., 2023; Zhu et al., 2024), it is typically built from human-translated corpora. However, for many low-resource directions, such supervision is limited. As a result, multi-way human-translated corpora, such as FLORES-200 (Costa-jussà et al., 2022) and NTREX-128 (Federmann et al., 2022) -like datasets, have become an important source for scaling SFT coverage, since their multi-way structure can theoretically support any directions from a relatively small amount of annotation.

In this work, our findings indicate that current multi-way SFT practices fail when scaling to many languages. When multi-way corpora are reused symmetrically around a pivot language (e.g., English), we observe an asymmetric outcome: while pivot → X improves as expected, the reverse X → pivot directions drop substantially, producing fluent but less faithful translations. We term this phenomenon Directional Degeneration. We analyze its cause as a data-usage issue. In multi-way corpora, the same pivot-language sentence may repeatedly appear as the target for many different sources, which increases target-side repetition and encourages shortcut learning. To mitigate this, we propose **Strategic Downsampling (SD)**, a simple data-level strategy that retains full supervision for pivot → X directions while keeping only a small fraction of reverse-direction instances. This change reduces excessive many-to-one mapping and stabilizes reverse-direction translation.

In addition to using multilingual supervision properly, we also study how to elicit cross-lingual transfer more explicitly in MMT. We introduce **Parallel Multilingual Prompting (PMP)**, which augments the translation instruction with an auxiliary parallel sentence as in-context guidance, to facilitate cross-lingual transfer in MMT. PMP is applied during training, and it can be activated at inference time as a lightweight enhancement when auxiliary translations are available, including those generated by the model itself.

We instantiate these ideas in **LMT**, a Chinese–English-centric suite of **L**arge-scale **M**ultilingual machine **T**ranslation models covering 60 languages and 234 translation directions, with four model sizes (0.6B/1.7B/4B/8B). We first perform large-scale CPT on about 90B tokens to strengthen the multilingual base. To address the scarcity of Chinese-centric resources (as shown in Figure 1, bottom), we expand coverage through broad data collection and curation. We then perform SFT, where SD and PMP are integrated to improve directional robustness and cross-lingual transfer. Finally, we apply preference optimization with GRPO (Shao et al., 2024), reusing the same supervised pairs from SFT to further refine translation quality. Comprehensive evaluation shows that LMT is competitive among open-source MMT systems with comparable language coverage. In particular, LMT-60-4B is on par with or better than substantially larger baselines such as X-ALMA-13B (Xu et al., 2025), Aya-101-13B (Üstün et al., 2024), and NLLB-54B (Costa-jussà et al., 2022) on the overlapping directions.

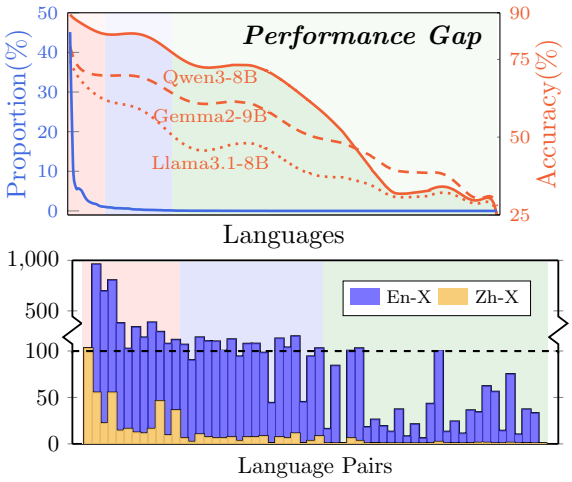


Figure 1: **Top:** Performance of base LLMs (orange) on the Belebele benchmark across 108 languages, plotted against their data ratios in the CulturaX (blue). **Bottom:** Bilingual data volume (million sentence pairs) from the OPUS corpus for 60 languages in our study, covering English-centric (blue) and Chinese-centric (orange) directions. Languages are grouped into high-, medium-, and low-resource tiers.

In summary, our contributions are threefold:

- We identify Directional Degeneration as a failure mode in large-scale multilingual SFT with symmetric multi-way data reuse, analyze its cause, and propose Strategic Downsampling as an effective data-level mitigation.
- We introduce Parallel Multilingual Prompting (PMP), which strengthens cross-lingual transfer during training and can be leveraged for optional test-time enhancement.
- We present and release LMT, a suite of Chinese–English-centric multilingual translation models in four sizes, providing broad language coverage and strong performance.

2 The Pitfall of Directional Degeneration

SFT Data A common takeaway in instruction tuning is that high-quality, diverse supervision matters more than sheer scale (Zhou et al., 2023). However, for many low-resource languages, such high-quality parallel data is extremely scarce, often limited to a few multi-way benchmarks. Consequently, broad-coverage corpora like the FLORES-200 devset (Costa-jussà et al., 2022) and NTREX-128

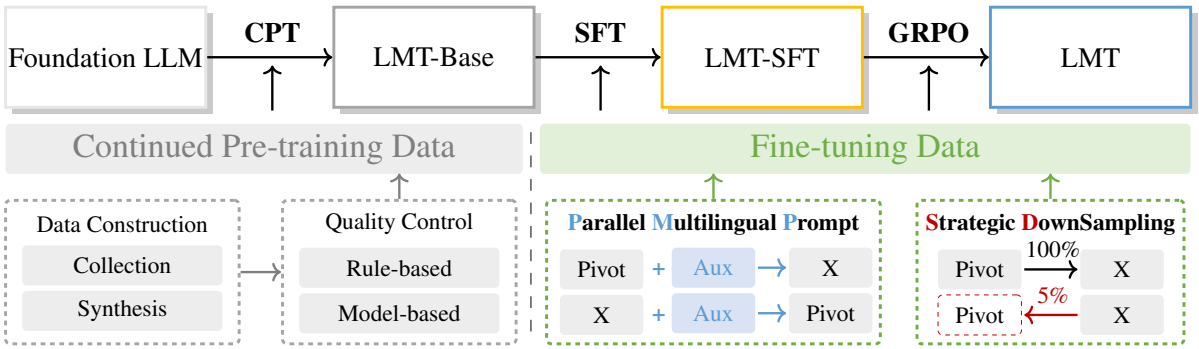


Figure 2: Overview of the LMT framework. The pipeline consists of three stages: Continued Pre-training (CPT) on mixed monolingual and bilingual corpora, Supervised Fine-Tuning (SFT) enhanced by Parallel Multilingual Prompting (PMP) and Strategic Downsampling (SD), and Group Relative Policy Optimization (GRPO) to further refine model outputs.

(Federmann et al., 2022) have become indispensable, as they represent the few reliable sources of clean, human-annotated translations for these long-tail directions. Following this widely used recipe, we take Flores-200 devset and NTREX-128 as the primary backbone to ensure broad linguistic coverage, augment them with SMol (Caswell et al., 2025) for additional underrepresented pairs, and further add WMT14–23 and IWSLT17–24 test sets to increase style and domain diversity. In total, the resulting SFT dataset contains ≈567K high-quality parallel pairs, allocating 3K–20K examples per direction across 117 English- and Chinese-centric directions spanning 60 languages.

The Phenomenon and Hypothesis We begin by applying standard SFT to the Qwen3-4B-Base (Yang et al., 2025a) model, utilizing parallel pairs in both directions (En/Zh↔X) as per common practice. While this standard approach yielded expected improvements in the En/Zh→X directions, it unexpectedly resulted in a significant performance drop in the reverse X→En/Zh directions. Qualitative analysis shows that the model falls into a state of fluent hallucination, producing grammatically correct but factually unfaithful outputs. Table 4 shows a representative case. We term this phenomenon ***Directional Degeneration***. We hypothesize that this pathology stems from a ***Shallow Mapping Trap***. The symmetric usage of multi-way data inherently creates excessive many-to-one mappings, where a single English/Chinese target is paired with dozens of distinct source languages. This structural imbalance incentivizes the model to learn a shortcut: bypassing source semantics to simply overfit the high-frequency target patterns, thereby sacrificing faithfulness.

Experimental Design To test this hypothesis and rule out model-specific factors, we design a systematic suite of experiments along three axes: **data usage**, **model**, and **multilingual scale**. On the data-usage axis, we (i) *Break Symmetry* by replacing the reverse X→En/Zh portion of the multi-way SFT data with a completely disjoint subset sampled from our bilingual CPT corpus (symmetry-breaking replacement), and (ii) perform *Gradual Symmetry Injection* by training on the original multi-way data while increasing the reverse retention rate p from 0% to 100% (fully symmetric). On the model axis, we repeat this protocol on Qwen3 models ranging from 0.6B to 8B parameters, and further on Llama-3.1-8B (Dubey et al., 2024) and Gemma-2-9B (Rivière et al., 2024), to verify whether the phenomenon persists across different model sizes and families. On the multilingual axis, we vary the number of languages involved in SFT from 10 to 50 (using Qwen3-4B-Base) to assess how the density of many-to-one mappings influences the severity of degeneration.

Results The results visualized in Figure 3 provide converging evidence for our hypothesis. First, the sharp contrast between the symmetry-breaking replacement setting (dashed lines) and fully symmetric multi-way reuse (solid curves at 100%) in Qwen3-4B-Base shows that the collapse is driven by how the multi-way data are reused, rather than by the intrinsic difficulty of X→En/Zh directions. Furthermore, the gradual injection curves reveal an "inverted-V" trajectory: performance peaks rapidly at a low retention rate ($p \approx 5\%$) but suffers a severe decline as p increases toward 100%, confirming that excessive target repetition triggers the degeneration. Second, the same asymmetric pattern consis-

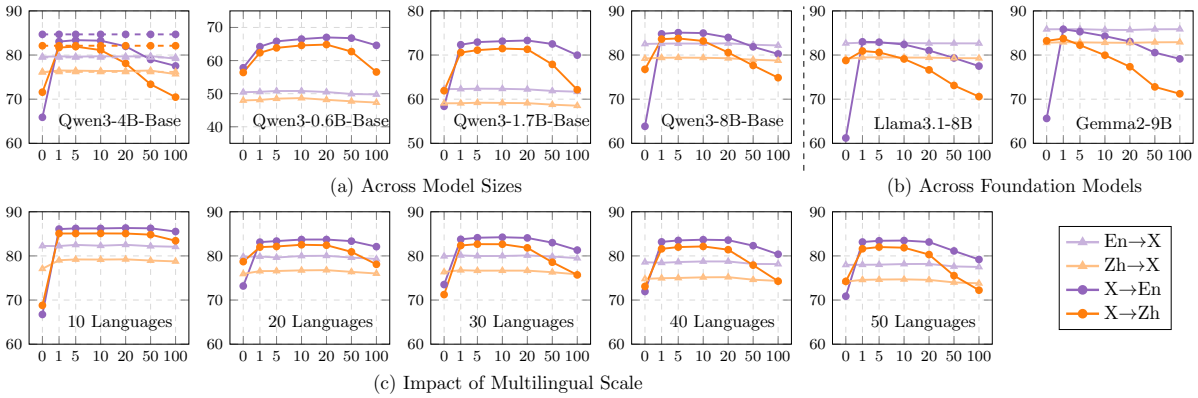


Figure 3: Comprehensive analysis of Directional Degeneration across different dimensions. The plots illustrate the COMET performance trend as a function of the strategic downsampling proportion (p). Dashed lines in Qwen3-4B-Base represent the use of disjoint data for the X→En/Zh directions.

tently appears across all evaluated model sizes and backbone families, indicating that directional degeneration is a general failure mode of large-scale multilingual SFT. Finally, the degradation becomes more pronounced as we scale the number of languages, supporting the interpretation that stronger many-to-one target repetition at larger multilingual scales amplifies shortcut learning and undermines faithfulness.

2.1 Mitigation via Strategic Downsampling

While replacing reverse-direction data with disjoint data proves effective in our analysis, relying on external data sources for SFT is not always viable in practice. We therefore seek a self-contained solution that resolves the degeneration solely within the existing multi-way SFT corpus. The *Gradual Symmetry Injection* experiment suggests a simple yet effective mitigation strategy. As shown by the solid curves in Figure 3: while full reuse (p=100%) leads to a clear collapse, retaining a small fraction of reverse examples is sufficient to maintain alignment without triggering the collapse. Motivated by this, we propose Strategic Downsampling (SD): during SFT, we retain all En/Zh→X data, and for the multi-way portion of the corpus we independently subsample each X→En/Zh instance with probability p, using p=5% as the default setting in our models.

It is worth noting that recent work (Zheng et al., 2025b) also reports this asymmetric degradation, attributes it to the curse of multilinguality, and addresses it with model-level interventions such as direction-aware training and group-wise model merging. Compared with this line of work, our analysis supports a more specific data-usage explanation, showing that symmetric reuse of multi-way corpora induces the degradation. We further find that a simple data-level strategy is sufficient to prevent it in practice.

3 Parallel Multilingual Prompting

Recent studies have highlighted the utility of multi-way parallel data in enhancing multilingual LLMs. For instance, Shen et al. (2025) demonstrate that utilizing multi-way corpora during CPT effectively promotes cross-lingual alignment and transfer. Parallelly, investigations into inference-time strategies (Mu et al., 2024; Yang et al., 2025b) indicate that incorporating auxiliary translations into the prompt improves performance, with Mu et al. (2024) further identifying specific neuronal mechanisms that support such parallel processing. Despite these advancements in pre-training and inference, the integration of such multi-source signals directly into the SFT for large-scale machine translation remains unexplored. To address this gap, we introduce Parallel Multilingual Prompting (PMP), a training strategy that incorporates auxiliary parallel context into the input prompts during SFT. By explicitly conditioning on both the source and an auxiliary anchor, PMP enables more direct cross-lingual transfer.

Formulation Let S be a source sentence in language L_S and T the target in L_T. Under a standard translation prompt (STP), the model is trained to model $P_\theta(T \mid S;\ \tau_{L_S \to L_T})$, where $\tau_{L_S \to L_T}$ denotes the translation direction. PMP augments the input with an auxiliary sentence A in an auxiliary language L_A, which is itself a translation of S, and

instead models $P_\theta(T \mid S, A; \tau_{L_S \to L_A \to L_T})$, so that (S, A) provide two parallel views of the same semantic content. Here, the auxiliary context A acts as a semantic anchor that aids the model in interpreting the source S and guiding the generation of the target T. Figure 5 compares these two prompting strategies.

Auxiliary Language Selection To select the most effective auxiliary language, our strategy considers two primary factors: linguistic affinity and model proficiency. Intuitively, an anchor is most beneficial when it shares close linguistic ties with the specific language being supported while being well-mastered by the model. Following this rationale: For En ↔ X, we look for a high-proficiency neighbor: we select a language that shares close typological ties with X (e.g., via script, phylogeny, or contact) and on which the model demonstrates robust competence. For Zh↔X, we use English (L_A=En), which the model typically handles most reliably and thus provides a stable semantic anchor. This choice also facilitates test-time PMP, since self-generating an English anchor is typically easier and more robust than generating anchors in other languages. Detailed mapping information are provided in Appendix Table 12.

Training and Inference We integrate PMP into SFT by probabilistically mixing PMP samples with standard STP samples. This explicitly trains the model to exploit auxiliary parallel context while preserving performance under STP. At inference, the model translates using the standard STP prompt by default. When an auxiliary sentence A is provided (e.g., from a high-quality external MT system or self-generated), we can use the PMP prompt to activate the learned behavior, which can yield additional improvements in translation quality.

4 The LMT Framework

LMT is a suite of Chinese–English-centric MMT models scaled across four distinct sizes (0.6B, 1.7B, 4B, and 8B). The suite covers 60 languages spanning diverse families and scripts, supporting a total of 234 translation directions (including English ↔ 59 languages and Chinese ↔ 58 languages). The complete language list is available in the Appendix. Guided by our preliminary comparison in Figure 1 (top), which shows that Qwen3 offers a more comprehensive multilingual capability than contemporary open models with similar scale, we adopt Qwen3 as the backbone for all LMT variants.

4.1 Adaptation Pipeline

We adapt the base models to the MMT task through a three-stage pipeline, applied consistently across all four model scales to produce the LMT-60 suite:

- **Stage 1: Continued Pre-training (CPT).** We first train the Qwen3 base models on a large-scale mixture of monolingual and parallel corpora to strengthen the model's broad multilingual and translation knowledge.
- **Stage 2: Supervised Fine-tuning (SFT).** We instruction-tune the model on the curated STF data (Section 2). During SFT, we apply SD to the multi-way portion to prevent directional degeneration, and incorporate PMP to train the model to leverage auxiliary parallel context.
- **Stage 3: Preference Optimization (PO).** We further refine the model with GRPO, reusing the SFT prompts to sample candidate translations and scoring them with COMET-22 as a reference-based reward. This setup improves quality without introducing extra preference data.

4.2 CPT Data Curation

To support effective CPT, we construct a multi-stage data pipeline comprising large-scale collection, pseudo-parallel synthesis, and systematic filtering. For monolingual data, we aggregate text for the 60 target languages from a broad range of public, curated multilingual sources, and apply standard cleaning and de-duplication. For parallel data, we start from OPUS and examine the available parallel volume for our 117 directions. As illustrated in Figure 1 (bottom), we observe a significant imbalance: while English-centric directions are relatively well-covered, Chinese-centric directions suffer from severe data scarcity. To bridge this gap, we extensively augment the authentic parallel pool with synthetic data generated by high-performing MT systems. Following a systematic quality filtering process, we obtain a total of approximately 2.1B English-centric and 2.9B Chinese-centric sentence pairs across 117 language pairs, which lay a solid data foundation for subsequent adaptation. Detailed corpus composition, quality estimation, and statistics are provided in Appendix A.

5 Results and Analyses

5.1 Setup

Training Details For CPT, we use 90B tokens balanced at a 1:1:1 ratio across monolingual, Chinese-

Model	High Resource				Medium Resource				Low Resource			
	En→X	Zh→X	X→En	X→Zh	En→X	Zh→X	X→En	X→Zh	En→X	Zh→X	X→En	X→Zh
Qwen3-4B-Base	84.63	80.77	85.87	85.44	79.68	76.21	86.09	84.55	56.81	53.33	75.36	75.35
SFT	$87.72_{+3.09}$	$85.11_{+4.34}$	$83.82_{-2.05}$	$73.60_{-11.84}$	$86.71_{+7.03}$	$83.58_{+7.37}$	$80.00_{-6.09}$	$72.18_{-12.37}$	$77.51_{+20.70}$	$73.68_{+20.35}$	$73.78_{-1.58}$	$67.94_{-7.41}$
+ SD	$87.80_{+0.08}$	$85.32_{+0.21}$	$87.49_{+3.67}$	$86.55_{+12.95}$	$86.72_{+0.01}$	$83.67_{+0.09}$	$87.67_{+7.67}$	$85.87_{+13.69}$	$78.68_{+1.17}$	$75.15_{+1.47}$	$80.72_{+6.94}$	$79.13_{+11.19}$
+ CPT	$89.03_{+1.23}$	$86.72_{+1.40}$	$87.97_{+0.48}$	$87.39_{+0.84}$	$89.77_{+3.05}$	$86.96_{+3.29}$	$88.56_{+0.89}$	$87.06_{+1.19}$	$87.14_{+8.46}$	$84.17_{+9.02}$	$86.02_{+5.30}$	$84.74_{+5.61}$
+ PMP	$88.98_{-0.05}$	$86.74_{+0.02}$	$88.00_{+0.03}$	$87.53_{+0.14}$	$89.73_{-0.04}$	$86.92_{-0.04}$	$88.62_{+0.06}$	$87.20_{+0.14}$	$87.06_{-0.08}$	$84.08_{-0.09}$	$86.07_{+0.05}$	$84.90_{+0.16}$
+ GRPO	**$89.43_{+0.45}$**	**$87.20_{+0.46}$**	**$88.46_{+0.46}$**	**$88.19_{+0.66}$**	**$90.23_{+0.50}$**	**$87.52_{+0.60}$**	**$89.10_{+0.48}$**	**$87.97_{+0.77}$**	**$87.85_{+0.79}$**	**$84.92_{+0.84}$**	**$86.60_{+0.53}$**	**$85.81_{+0.91}$**
Qwen3-8B-Base	87.02	83.55	86.65	86.05	84.69	81.98	87.29	85.60	66.43	63.32	78.50	78.59
SFT	$88.53_{+1.51}$	$85.96_{+2.41}$	$84.86_{-1.79}$	$77.83_{-8.22}$	$88.21_{+3.52}$	$85.22_{+3.24}$	$82.27_{-5.02}$	$76.55_{-9.05}$	$80.72_{+14.29}$	$76.95_{+13.63}$	$77.61_{-0.89}$	$73.08_{-5.51}$
+ SD	$88.57_{+0.04}$	$86.13_{+0.17}$	$87.69_{+2.83}$	$86.84_{+9.01}$	$88.28_{+0.07}$	$85.37_{+0.15}$	$87.99_{+5.72}$	$86.33_{+9.78}$	$82.49_{+1.77}$	$79.12_{+2.17}$	$82.83_{+5.22}$	$81.52_{+8.44}$
+ CPT	$89.31_{+0.74}$	$87.07_{+0.94}$	$88.02_{+0.33}$	$87.46_{+0.62}$	$90.06_{+1.78}$	$87.35_{+1.98}$	$88.57_{+0.58}$	$87.17_{+0.84}$	$87.42_{+4.93}$	$84.51_{+5.39}$	$86.32_{+3.49}$	$85.18_{+3.66}$
+ PMP	$89.29_{-0.02}$	$87.10_{+0.03}$	$88.06_{+0.04}$	$87.60_{+0.14}$	$90.06_{+0.00}$	$87.28_{-0.07}$	$88.63_{+0.06}$	$87.39_{+0.22}$	$87.38_{-0.04}$	$84.50_{-0.01}$	$86.41_{+0.09}$	$85.41_{+0.23}$
+ GRPO	**$89.60_{+0.31}$**	**$87.41_{+0.31}$**	**$88.50_{+0.44}$**	**$88.22_{+0.62}$**	**$90.39_{+0.33}$**	**$87.70_{+0.42}$**	**$89.10_{+0.47}$**	**$87.95_{+0.56}$**	**$87.93_{+0.55}$**	**$85.04_{+0.54}$**	**$86.91_{+0.50}$**	**$86.08_{+0.67}$**

Table 2: COMET-22 scores of 4B and 8B models as we progressively enable components of the LMT training pipeline: supervised fine-tuning (SFT), Strategic Downsampling (SD), continued pre-training (CPT), Parallel Multilingual Prompting (PMP), and preference optimization (PO). **Bold** numbers indicate the best score within each backbone block. Subscripts denote the score difference compared to the previous row. We use red to mark entries affected by *directional degeneration*, and green to highlight substantial improvements (>1.0).

centric bilingual, and English-centric bilingual data. The bilingual samples are used equally in both directions (En/Zh→X and X→En/Zh; 50/50), and adopt an **Informative Formatting** with explicit direction tags and a target-language separator (as shown in Figure 5), which we find performs better slightly than naïve newline source–target concatenation (Guo et al., 2024; Iyer et al., 2024). During SFT, for forward directions (En/Zh→X), STP and PMP each account for 50%. For reverse directions (X → En/Zh), we apply strategic downsampling with a total retention of 5%, split evenly between formats (STP 2.5%, PMP 2.5%). Finally, we perform preference optimization with GRPO by reusing SFT prompts to generate rollouts and scoring them with COMET-22 (Rei et al., 2022a) as a reference-based reward. We train four model sizes (0.6B/1.7B/4B/8B) on 16 NVIDIA H200 GPUs, and the detailed hyperparameters are provided in the Appendix Table 5.

Evaluation Data and Metrics We evaluate on FLORES-200 *devtest* (Costa-jussà et al., 2022). To address the lack of a Mongolian (traditional script) testset, we translated the Chinese side of FLORES into Mongolian with native annotators [1]. We adopt COMET-22 as our primary evaluation metric, and report SacreBLEU (Post, 2018) in the Appendix. For brevity, we present LMT-60-4/8B in the main text, with full results for all four sizes provided in the Appendix. In addition, we also provide a comparison against the WMT24++ (Deutsch et al., 2025) in the Appendix.

[1]We will release this dataset to fill a gap in MT Benchmark.

5.2 Main Results

Table 2 summarizes the evolution of translation quality as we progressively integrate components of the LMT training pipeline. We report COMET-22 scores averaged over all 60 languages, categorized by resource tiers and translation direction. Starting from the 3-shot base models, SFT improves En/Zh→X on both 4B and 8B settings, but all X→En/Zh drop significantly below the corresponding base scores, as indicated by the red subscripts. This empirically confirms that SFT under symmetric multi-way reuse can trigger severe directional degeneration in reverse directions. Integrating SD effectively reverses this effect: X→En/Zh directions typically recover by approximately 2–13 COMET points relative to the SFT baseline, surpassing base performance while maintaining strong results on En/Zh→X.

Compared to the SD+SFT baseline, CPT yields substantial improvements across the board. It contributes 1–3 COMET points in high- and medium-resource scenarios and a remarkable 5–9 points in low-resource directions, underscoring its critical role in strengthening the model's broad translation ability. [2] The introduction of PMP yields gains on X→En/Zh directions, likely by enriching source-side diversity to alleviate the directional degeneration. Finally, GRPO reuses the same SFT examples but still brings a further gain of about 0.3–0.8 points across resource tiers. This indicates that preference optimization can extract additional benefit from the supervised data by exploring alternative genera-

[2]Detailed analyses are provided in Appendix B.

#Langs	Model	En→X	X→En	Zh→X	X→Zh	Avg.
10	TowerInstruct-13B	88.91	88.51	86.29	86.81	87.63
	LMT-60-4B	89.29	88.58	87.09	**88.40**	88.34
	LMT-60-8B	**89.42**	**88.59**	**87.30**	**88.40**	**88.43**
23	Aya-expanse-8B	88.66	88.31	86.26	86.20	87.36
	LMT-60-4B	89.44	88.62	87.00	**87.99**	88.26
	LMT-60-8B	**89.63**	**88.65**	**87.18**	87.98	**88.36**
27	Seed-X-PPO-7B	**90.78**	**89.05**	**88.48**	87.96	**89.07**
	LMT-60-4B	90.35	88.87	88.02	**88.19**	88.86
	LMT-60-8B	90.49	88.88	88.23	88.16	88.94
28	GemmaX2-28-9B	88.39	**88.95**	85.56	87.37	87.57
	LMT-60-4B	88.72	88.52	85.97	87.71	87.73
	LMT-60-8B	**88.83**	88.62	**86.12**	**87.76**	**87.83**
35	Hunyuan-MT-7B	86.78	86.42	83.88	85.74	85.71
	LMT-60-4B	88.72	88.00	86.03	87.26	87.50
	LMT-60-8B	**88.84**	**88.12**	**86.18**	**87.36**	**87.63**
40	X-ALMA-13B	89.17	88.68	-	-	88.92
	LMT-60-4B	89.24	88.67	-	-	88.96
	LMT-60-8B	**89.38**	**88.73**	-	-	**89.06**
54	Aya-101-13B	84.87	86.45	81.53	82.54	83.85
	LMT-60-4B	88.61	88.23	85.67	87.15	87.42
	LMT-60-8B	**88.75**	**88.36**	**85.84**	**87.24**	**87.55**
55	LLaMAX3-Alpaca	81.29	86.27	77.24	81.02	81.45
	LMT-60-4B	88.66	88.20	85.74	87.16	87.44
	LMT-60-8B	**88.78**	**88.32**	**85.91**	**87.23**	**87.56**
59	NLLB-54B	86.89	87.72	84.06	80.50	84.79
	LMT-60-4B	88.78	87.93	86.00	87.00	87.43
	LMT-60-8B	**88.91**	**88.07**	**86.16**	**87.11**	**87.56**

Table 3: COMET-22 scores of our LMT models compared with a range of general-purpose multilingual LLMs and dedicated MMT models, averaged over all overlapping languages for each system in four directions. "#Langs" denotes the number of languages shared between the baseline and LMT. **Bold** numbers indicate the best score in each comparison group, and underlined numbers the second best. The symbol "-" indicates directions not supported by the baseline model.

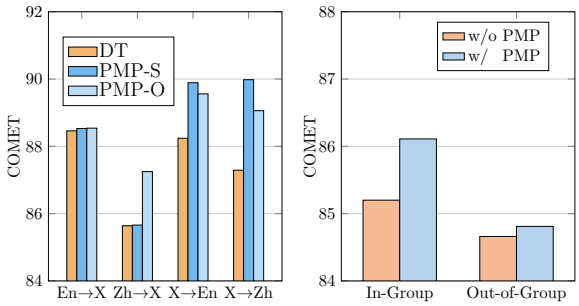


Figure 4: Analysis of the Parallel Multilingual Prompting (PMP). **Left:** Comparison of different inference-time strategies. **Right:** Comparison of zero-shot performance with and without PMP training.

tions and reinforcing better candidates, even when no new training examples are introduced.

5.3 Comparison with Existing MMT Systems

To benchmark LMT against the current SOTA, Table 7 presents a comprehensive comparison with a diverse range of systems, categorized into two groups: (1) *General-purpose Multilingual LLMs* capable of instruction-following translation, including Aya-Expanse-8B (Dang et al., 2024), Aya-101-13B (Üstün et al., 2024), and LLaMAX3-8B-Alpaca (Lu et al., 2024); (2) *Dedicated MMT Models*, including TowerInstruct-13B (Alves et al., 2024), GemmaX2-28-9B (Cui et al., 2025), X-ALMA-13B (Xu et al., 2025), Hunyuan-MT-7B (Zheng et al., 2025a), Seed-X-PPO-7B (Cheng et al., 2025), and NLLB-54B (Costa-jussà et al., 2022). To ensure a fair evaluation, we calculate metrics solely on the intersection of languages supported by each baseline, reporting the averaged COMET-22 scores over overlapping directions. We report a resource-tier breakdown on FLORES and additional results in Appendix C.

The results demonstrate that LMT models deliver robust and consistent performance across all comparison groups. Against general-purpose multilingual LLMs, LMT achieves substantial improvements, surpassing models like Aya-101-13B and LLaMAX3-Alpaca by a margin of approximately 3–6 COMET points on average. Compared to dedicated MMT systems, LMT remains highly competitive. It outperforms strong baselines such as NLLB-54B and GemmaX2-28-9B, and performs on par with top-tier systems like Seed-X-PPO-7B. Most notably, LMT-60-4B demonstrates exceptional parameter efficiency, matching or exceeding the performance of significantly larger models (e.g., the Aya-101-13B and NLLB-54B). Overall, these results position LMT as a competitive MMT baseline: it covers a broader set of languages than most existing LLM-based MT systems, and achieves comparable or better quality than much larger baselines on the overlapping directions.

5.4 Understanding the Effect of PMP

Test-time Enhancement We first analyze the effect of PMP at inference time on the PMP-enabled languages (see Table 12). We compare three modes: Direct Translation (DT), which uses the standard translation prompt (STP), PMP with a self-generated auxiliary sentence (PMP-S), and PMP with an oracle auxiliary sentence (PMP-O). As illustrated in Figure 4 (left), the effectiveness of PMP varies by direction. On X→En/Zh, PMP-S performs comparably to or better than PMP-O, sug-

gesting that for high-resource targets, the model is resilient to imperfect anchors; self-generated context is sufficient to boost performance. Conversely, on Zh→X, we find that only PMP-O yields clear gains, while PMP-S does not. This implies that translating into X is sensitive to anchor quality: while gold-standard context helps, noisy self-generated anchors may fail to provide positive guidance. Practically, these findings highlight PMP-S as a robust inference-time enhancement strategy for X→En/Zh directions, enabling the model to self-boost performance without relying on external references.

PMP Improves Zero-shot Transfer Beyond inference benefits, we examine whether PMP training enhances the model's cross-lingual transfer in zero-shot directions. We compare models trained with and without PMP on two distinct subsets of directions among the participating languages (Figure 4 (right)). In the *In-Group* setting (translations between auxiliary languages and their associated targets, A↔X), PMP training substantially boosts the average COMET-22 score from 85.20 to 86.11. Furthermore, in the *Out-of-Group* setting, where we evaluate directions between high-resource languages never explicitly used as PMP anchors (e.g., {Es, Ja}↔X), we still observe a positive performance uplift. These results indicate that PMP not only strengthens the anchored language pairs but also promotes zero-shot transfer to related directions that were never explicitly used as auxiliary anchors during training.

6 Related Work

LLM Adaptation for MMT While dedicated encoder-decoder models like NLLB (Costa-jussà et al., 2022) and M2M-100 (Fan et al., 2021) have long served as the backbone of multilingual translation, recent trends have shifted towards adapting decoder-only LLMs for this task (Yang et al., 2023; Alves et al., 2024; Xu et al., 2025; Cui et al., 2025). This transition brings not only stronger translation quality but also advanced capabilities such as context awareness (Wang et al., 2024, 2025b) and reasoning (Wang et al., 2025a; Chen et al., 2025). Despite this progress, current LLM-based MMT systems still face several limitations. First, regarding language coverage, most existing adaptations remain restricted to a narrow set of dominant languages, struggling to match the extensive linguistic breadth required for truly universal translation. Second, regarding translation directions, the research landscape remains heavily English-centric. Consequently, non-English directions are largely underserved, resulting in inferior performance compared to English-centric tasks, with only a few recent works extending capabilities to Chinese-centric translation(Cui et al., 2025; Zheng et al., 2025a; Cheng et al., 2025). Finally, low-resource translation remains a persistent challenge for adapted LLMs. Under severe data scarcity and the dominance of high-resource languages in pre-training, these models often struggle to maintain robustness on resource-poor languages.

Utilization of Multi-way Data Multi-way parallel data, where aligned sentences exist across three or more languages, provides richer semantic constraints than bilingual counterparts. In traditional NMT, this concept was explored through multi-source translation to improve disambiguation and robustness (Nishimura et al., 2020; Xu et al., 2021). In the era of LLMs, recent analysis has begun to reveal the parallel multilingual learning mechanisms within these models, suggesting that LLMs inherently benefit from cross-lingual alignment (Mu et al., 2024; Shen et al., 2025; Yang et al., 2025b). However, while these works demonstrate the value of multi-way signals, how to properly leverage them for consistent gains in large-scale multilingual machine translation, and the potential pitfalls of widely used symmetric multi-way mixtures in SFT remain under-explored.

7 Conclusion

In this work, we introduce LMT, a Chinese-English-centric MMT model, covering 60 languages across 234 translation directions and achieving competitive performance among models with similar language coverage. Our adaptation pipeline begins with the construction of a large, curated multilingual corpus, followed by extensive continued pre-training to integrate broad translation knowledge into the model. We then identify directional degeneration, a salient yet previously overlooked issue in SFT with multi-way data, and mitigate it by introducing a simple strategic downsampling method. Furthermore, we propose Parallel Multilingual Prompting, a simple but effective technique to enhance cross-lingual transfer. We release LMT models as publicly available baselines to facilitate future research on inclusive and high-quality multilingual machine translation.

Limitations

While LMT shows promising performance, we note several limitations that also point to natural directions for future work. First, although our evaluation covers a broad set of standard academic benchmarks and primarily relies on COMET, these settings may not fully reflect the diversity of real-world translation use cases. Future work could extend this evaluation to a wider range of real-world scenarios to further assess generalization and capture more nuanced aspects of translation quality. Second, LMT adopts a Chinese–English-centric design as a step toward moving beyond an English-only focus. This bi-centric setting is still a simplifying choice and may not be optimal for other regions or language communities. It would be useful to explore tri-centric or more general multi-centric configurations, and to study how they affect scalability, interference, and cross-lingual transfer. Third, LMT currently supports 60 languages, which remains a small subset of global linguistic diversity. Expanding coverage is constrained not only by training cost but also by the availability and quality of text and parallel data for many underrepresented languages, especially those with limited written resources. Future work could prioritize extending support to additional languages and improving adaptation under extreme data scarcity through more effective data collection, filtering, and transfer strategies.

References

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

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Erik Henriksson, and 1 others. 2025. An expanded massive multilingual dataset for high-performance language technologies. *arXiv preprint arXiv:2503.10267*.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, Solo Farabado Cissé, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. *Preprint*, arXiv:2502.12301.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *CoRR*, abs/2502.11544.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation LLM with 7b parameters. *CoRR*, abs/2507.13618.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *NAACL (Long Papers)*, pages 5420–5443. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language

Coverage of WMT24 to 55 Languages & Dialects. *Preprint*, arXiv:2502.12404.

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, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

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.

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 *NAACL-HLT (Findings)*, pages 639–649. Association for Computational Linguistics.

HuggingFaceFW. 2024. fineweb (revision af075be).

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *WMT*, pages 1393–1409. 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.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Yingfeng Luo, Tong Zheng, Yongyu Mu, Bei Li, Qinghong Zhang, Yongqi Gao, Ziqiang Xu, Peinan Feng, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2025. Beyond decoder-only: Large language models can be good encoders for machine translation. *Preprint*, arXiv:2503.06594.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6976–6997.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

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.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2020. Multi-source neural machine translation with missing data. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:569–580.

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, Belgium, Brussels. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022a. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In *WMT*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *WMT*, pages 634–645. Association for Computational Linguistics.

Morgane Rivière, 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, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Yingli Shen, Wen Lai, Shuo Wang, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. From unaligned to aligned: Scaling multilingual llms with multi-way parallel corpora. *CoRR*, abs/2505.14045.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. `https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama`.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. DRT: deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6770–6782. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7175–7187. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025b. Delta: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, and 11 others. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *ICLR*. OpenReview.net.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: plug & play modules and adaptive rejection for quality translation at scale. In *ICLR*. OpenReview.net.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3029–3041.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Sen Yang, Yu Bao, Yu Lu, Jiajun Chen, Shujian Huang, and Shanbo Cheng. 2025b. Enanchored-x2x: English-anchored optimization for many-to-many translation. *CoRR*, abs/2509.19770.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Jia Yu, Fei Yuan, Rui Min, Jing Yu, Pei Chu, Jiayang Li, Wei Li, Ruijie Zhang, Zhenxiang Li, Zhifei Ren, Dong Zheng, Wenjian Zhang, Yan Teng, Lingyu Meng, ZhenJiang Jin, Jiantao Qiu, ShaSha Wang, Zhongying Tu, Dahua Lin, and 4 others. 2025. Wanjuansilu: A high-quality open-source webtext dataset for low-resource languages. *Preprint*, arXiv:2501.14506.

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

Mao Zheng, Jiaqiang Wang, Haoran Xiong, Hongda Liu, Canrong He, Ziyue Jiang, and Dingkai Long. 2025a. Hunyuan-MT Technical Report. *Preprint*, arXiv:2509.05209.

Tong Zheng, Yan Wen, Huiwen Bao, Junfeng Guo, and Heng Huang. 2025b. Asymmetric conflict and synergy in post-training for llm-based multilingual machine translation. *CoRR*, abs/2502.11223.

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

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *EMNLP*, pages 388–409. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, 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. *arXiv preprint arXiv:2402.07827*.

A Continued Pre-training Data Curation

The foundation of CPT relies on a high-quality, large-scale, and diverse corpus. Recognizing that constructing such a corpus is a major challenge, we designed a systematic, multi-stage data curation pipeline to ensure both linguistic breadth and quality consistency.

CPT Monolingual Data To achieve comprehensive coverage and diversity, our monolingual CPT corpus aggregates several sources. For English and Chinese, we utilized the well-curated SlimPajama (Soboleva et al., 2023) and Skywork (Wei et al., 2023) corpora, respectively. For the remaining languages, we collected from CulturaX (Nguyen et al., 2024), OpenDataLab (Yu et al., 2025), and Wikimedia, with full source details in Table 10.

CPT Bilingual Data The foundation of the bilingual corpus is curated from OPUS [3] sub-corpora. To significantly scale this up, we employed pseudo-parallel synthesis using open-source models in two ways: (1) direct synthesis, creating synthetic En/Zh $\rightarrow$ X data by translating monolingual corpora, and (2) pivoted synthesis via English, leveraging typically higher-quality En$\leftrightarrow$X and En$\rightarrow$Zh models to obtain Zh$\leftrightarrow$X data. Finally, all data was unified and subjected to a rigorous quality control pipeline. We used OpusFilter (Aulamo et al., 2020) for rule-based cleaning, including length-based sanity checks and misalignment removal; in particular, we filtered out sentence pairs with a maximum length ratio greater than 3.0. For language identification, we employed the standard FastText LID model integrated in OpusFilter, with an adaptive strategy to accommodate different resource levels: for high-resource languages we enforced a confidence threshold of 0.5 to ensure data purity, while for low-resource languages (where LID probabilities are often poorly calibrated) we prioritized recall by accepting samples as long as the Top-1 predicted language matched the target tag, regardless of its probability. We further applied CometKiwi (Rei et al., 2022b) for quality-based scoring and selection. This yields approximately 2.1B sentence pairs for English-centric and 2.9B for Chinese-centric directions, with the vast majority comprising over 10M high-quality pairs across the 117 targeted directions.

[3] https://opus.nlpl.eu/

Quality Analysis of the Bilingual Corpus Figures 7 and 8 present the COMETKiwi score (Rei et al., 2022b) distributions for the English-centric (En $\leftrightarrow$ X) and Chinese-centric (Zh $\leftrightarrow$ X) portions of our curated CPT bilingual corpus, respectively. A primary observation is that the score distributions for low-resource languages are noticeably skewed to the left compared to the high-resource counterparts. This skew is particularly pronounced for the Chinese-centric low-resource pairs (Figure 8), highlighting the challenge of sourcing non-English-centric data. We attribute this phenomenon to two factors. First, it likely reflects the real scarcity of clean, high-quality bilingual data for many low-resource languages, especially those paired with non-English. Second, it may reveal a bias in the quality estimation (QE) model itself—models like COMETKiwi may yield less reliable scores for underrepresented or non-English language pairs due to limited exposure during training. This could lead to systematically lower scores for some language pairs, irrespective of their true quality.

Overall, these findings point to a dual challenge in multilingual MT: the limited availability of clean bilingual data for many non-English-centric pairs, and the potential calibration limitations of current QE models in such settings. They also motivate future work on QE methods that are better calibrated and more robust across diverse, non-English-centric language pairs.

B CPT Gains Across Languages

This section details the language-specific impact of Continued Pre-training (CPT), extending the aggregated ablation results in Section 5.2. Figure 6 presents COMET score comparisons between Base+SFT+SD (without CPT) and Base+CPT+SFT+SD (with CPT) across four translation directions per language.

Across nearly all languages and directions, incorporating CPT yields consistent performance improvements, underscoring its role as a fundamental adaptation step. However, the magnitude of this gain is not uniform, and is more pronounced for medium- and low-resource languages. This trend aligns with the expectation that CPT is particularly

beneficial for strengthening foundational linguistic knowledge where the base model lacks sufficient prior exposure. Additionally, for many high- and medium-resource languages, the baseline model exhibits lower performance in the En/Zh → X directions than in X → En/Zh, consistent with prior observations (Arivazhagan et al., 2019) that generating into diverse target languages is inherently more challenging. CPT substantially improves performance in these directions, highlighting its effectiveness in strengthening multilingual generation capabilities.

C Additional Evaluation

This appendix complements Section 5.3 by (i) reporting FLORES results stratified by resource tier, and (ii) evaluating on WMT24++ (Deutsch et al., 2025) to test robustness beyond FLORES.

C.1 FLORES Breakdown by Resource Tier

Table 6 provides a resource-tier breakdown that complements the direction-level averages reported in the main text, and helps assess whether improvements are concentrated in a particular subset of languages. In the high-resource tier, where most baselines already perform strongly, LMT-60-4B/8B remains consistently competitive across all four directions and is frequently among the top systems, suggesting that the overall gains are not achieved by sacrificing performance on well-resourced pairs. In the medium- and low-resource tiers, the relative advantage of LMT becomes more pronounced and more stable, particularly on Chinese-centric directions where several baselines either lag behind or do not support the full direction set. For example, on the 54-language overlap with Aya-101-13B, LMT-60-4B improves low-resource En→X from 81.68 to 86.92 and low-resource X→Zh from 81.58 to 85.98. On the 59-language overlap with NLLB-54B, LMT-60-4B yields a clear margin on low-resource X→Zh (80.56→85.88) and improves low-resource Zh→X (82.33→84.52), while remaining strong on En-centric directions.

Overall, the tiered results indicate that LMT's performance is broadly competitive on high-resource pairs and delivers larger, more consistent gains on the long tail, rather than being driven by a narrow subset of directions.

C.2 Results on WMT24++

Table 8 reports results on WMT24++, a document-level benchmark that emphasizes cross-sentence coherence and discourse-level adequacy. In this setting, LMT remains competitive on many directions but trails Hunyuan-MT on several subsets. More broadly, these results are consistent with a general limitation of current LLM-based MT systems: document-level translation is still challenging, likely because most models are post-trained primarily with sentence-level supervision and receive limited exposure to discourse-level signals. For LMT, our SFT focuses on sentence pairs and includes little document-context training, which can weaken performance on benchmarks that require cross-sentence consistency. In future work, we will incorporate targeted document-level parallel data and context-aware SFT to strengthen discourse-level behaviors.

Continued Pre-training Prompt

<en2ky> Fellow wrestlers also paid tribute to Luna.
<ky> Балбандар да Лунага сый көрсөтүштү.

Standard Translation Prompt (STP)

Translate this from English to Kirghiz:
English: Fellow wrestlers also paid tribute to Luna.
Kirghiz: Балбандар да Лунага сый көрсөтүштү.

Parallel Multilingual Prompt (PMP)

Translate this from English to Russian and Kirghiz:
English: Fellow wrestlers also paid tribute to Luna.
Russian: Борцы тоже почтили Луну.
Kirghiz: Балбандар да Лунага сый көрсөтүштү.

Figure 5: Examples of the three prompt formats for the CPT and SFT stages of LMT adaptation. The underlined text indicates the part used for loss computation during training.

Category	Content
Source	他补充道：“我们现在有4 个月大没有糖尿病的老鼠，但它们曾经得过该病。”
Reference	"We now have 4-month-old mice that are non-diabetic that **used to be diabetic**," he added.
Prediction	He added: "We now have four-month-old mice that **have never had diabetes but were given the disease**."

Table 4: Translation error cases: hallucinations and fabricated outputs

Hyperparameter	CPT Stage	SFT Stage	GRPO Stage
Learning Rate	2e-5	2e-5	5e-7
Adam β	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
LR Scheduler	cosine	cosine	cosine
Number of Epochs	1	1	1
Global Batch Size	1536	1024	1024
Max Length	2048	1024	1024
Train Steps	40,000	500	500
Warmup Ratio	0.05	0.01	0.01
Weight Decay	0.01	0.01	0.01
Rollout	-	-	8
Temperature	-	-	1.0
KL β	-	-	0.001

Table 5: Hyperparameter configuration during training.

# Langs (Hig./Med./Low)	Model	High Resource				Medium Resource				Low Resource			
		En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh
10 (9/1/0)	TowerInstruct-13B	88.78	88.51	86.15	86.83	89.92	88.53	87.44	86.66	-	-	-	-
	LMT-60-4B	89.14	88.56	86.91	**88.42**	**90.47**	**88.68**	88.49	88.28	-	-	-	-
	LMT-60-8B	**89.28**	**88.58**	**87.14**	88.39	**90.47**	88.61	**88.56**	**88.43**	-	-	-	-
23 (13/9/1)	Aya-expanse-8B	88.60	88.16	86.44	86.32	88.82	88.50	86.19	86.13	**87.82**	88.41	**84.76**	85.28
	LMT-60-4B	89.43	88.46	87.20	**88.19**	89.67	**88.85**	86.98	**87.81**	87.55	88.58	84.69	87.27
	LMT-60-8B	**89.60**	**88.50**	**87.41**	88.17	**89.87**	**88.85**	**87.18**	87.80	87.80	**88.65**	84.41	**87.39**
27 (13/12/2)	Seed-X-PPO-7B	**89.91**	**88.59**	**87.73**	87.94	**91.58**	**89.48**	**89.25**	88.07	**91.27**	**89.22**	**88.78**	87.39
	LMT-60-4B	89.43	88.46	87.20	**88.19**	91.14	89.26	88.84	**88.26**	91.09	89.00	88.46	87.84
	LMT-60-8B	89.60	88.50	87.41	88.17	91.26	89.24	89.04	88.20	91.14	89.01	88.67	**87.88**
28 (13/8/7)	GemmaX2-28-9B	89.34	**88.62**	86.99	87.66	88.71	**89.46**	85.82	87.59	86.42	**88.93**	82.81	86.63
	LMT-60-4B	89.43	88.46	87.20	**88.19**	89.11	89.04	86.16	87.99	87.06	88.04	83.63	86.56
	LMT-60-8B	**89.60**	88.50	**87.41**	88.17	**89.21**	89.06	**86.29**	**88.03**	**87.09**	**88.30**	**83.73**	**86.77**
35 (13/9/13)	Hunyuan-MT-7B	89.43	87.56	87.08	87.38	89.10	87.76	86.29	86.77	82.74	84.43	79.27	83.50
	LMT-60-4B	89.43	88.46	87.20	**88.19**	89.28	88.89	86.49	87.91	87.67	86.96	84.63	85.96
	LMT-60-8B	**89.60**	**88.50**	**87.41**	88.17	**89.41**	**88.90**	**86.64**	**87.93**	**87.75**	**87.22**	**84.74**	**86.22**
40 (13/16/11)	X-ALMA-13B	89.41	**88.51**	-	-	90.37	**89.29**	-	-	87.15	87.98	-	-
	LMT-60-4B	89.43	88.46	-	-	90.31	89.13	-	-	87.47	88.24	-	-
	LMT-60-8B	**89.60**	88.50	-	-	**90.47**	89.13	-	-	**87.55**	**88.40**	-	-
54 (13/18/23)	Aya-101-13B	87.00	86.55	84.34	83.29	87.54	87.32	84.26	83.26	81.68	85.71	77.92	81.58
	LMT-60-4B	89.43	88.46	87.20	**88.19**	90.23	**89.10**	87.52	**87.97**	86.92	87.43	83.42	85.98
	LMT-60-8B	**89.60**	**88.50**	**87.41**	88.17	**90.39**	**89.10**	**87.70**	87.95	**87.01**	**87.72**	**83.56**	**86.20**
55 (13/18/24)	LLaMAX3-Alpaca	85.22	87.19	82.28	82.49	84.80	87.87	80.90	82.23	76.69	84.60	71.97	79.38
	LMT-60-4B	89.43	88.46	87.20	**88.19**	90.23	**89.10**	87.52	**87.97**	87.10	87.40	83.69	86.03
	LMT-60-8B	**89.60**	**88.50**	**87.41**	88.17	**90.39**	**89.10**	**87.70**	87.95	**87.17**	**87.65**	**83.82**	**86.22**
59 (13/18/28)	NLLB-54B	87.95	88.17	85.82	80.06	88.95	88.85	85.58	80.69	85.12	86.81	82.33	80.56
	LMT-60-4B	89.43	88.46	87.20	**88.19**	90.23	**89.10**	87.52	**87.97**	87.57	86.96	84.52	85.88
	LMT-60-8B	**89.60**	**88.50**	**87.41**	88.17	**90.39**	**89.10**	**87.70**	87.95	**87.65**	**87.23**	**84.64**	**86.12**

Table 6: COMET-22 scores of our LMT models compared with a range of general-purpose multilingual LLMs and dedicated MMT models. Evaluation is conducted only on the intersection of language pairs supported by each baseline and our models. The first column (# Langs) denotes the number of overlapping languages, followed by their distribution across resource tier (high/medium/low). **Bold** numbers indicate the best in each group, and the underlined numbers the second best. The symbol '-' indicates directions not supported by the baseline model.

# Langs (Hig./Med./Low)	Model	High Resource				Medium Resource				Low Resource			
		En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh
10 (9/1/0)	TowerInstruct-13B	37.46	**39.55**	21.56	36.72	30.54	**31.99**	21.34	33.86	-	-	-	-
	LMT-60-4B	37.35	39.15	23.03	41.93	29.38	30.95	23.20	38.59	-	-	-	-
	LMT-60-8B	**38.34**	39.28	**24.25**	**42.09**	**31.36**	31.66	**24.20**	**39.34**	-	-	-	-
23 (13/9/1)	Aya-expanse-8B	31.74	36.64	19.70	34.58	31.14	37.67	19.61	34.53	**27.45**	42.71	13.53	35.12
	LMT-60-4B	33.79	37.97	21.42	41.10	32.61	39.69	21.10	40.48	24.68	44.13	14.13	41.32
	LMT-60-8B	**35.00**	**38.16**	**22.48**	**41.33**	**34.01**	**39.89**	**22.00**	**40.80**	27.15	**44.58**	**14.53**	**41.76**
27 (13/12/2)	Seed-X-PPO-7B	**37.88**	37.14	**23.13**	34.03	**38.33**	40.63	**23.12**	34.95	**39.32**	42.53	**21.84**	33.86
	LMT-60-4B	33.79	37.97	21.42	41.10	32.83	40.77	21.38	**41.23**	35.02	42.70	20.64	41.25
	LMT-60-8B	35.00	**38.16**	22.48	**41.33**	33.91	**40.79**	22.32	41.21	35.79	**42.72**	21.52	**41.26**
28 (13/8/7)	GemmaX2-28-9B	**36.23**	**39.78**	**22.60**	40.42	**32.21**	**41.59**	**20.40**	39.21	22.06	**43.09**	**13.74**	38.21
	LMT-60-4B	33.79	37.97	21.42	41.10	28.96	37.87	19.21	39.66	21.51	38.61	13.01	37.71
	LMT-60-8B	35.00	38.16	22.48	**41.33**	30.22	38.22	20.12	**40.17**	**22.28**	39.51	13.60	**38.42**
35 (13/9/13)	Hunyuan-MT-7B	27.72	28.65	17.55	28.49	21.24	27.38	14.27	26.48	9.99	21.81	6.59	22.04
	LMT-60-4B	33.79	37.97	21.42	41.10	28.92	38.12	18.95	39.84	18.00	34.70	10.97	36.93
	LMT-60-8B	**35.00**	**38.16**	**22.48**	**41.33**	**30.22**	**38.47**	**19.88**	**40.27**	**19.01**	**35.74**	**11.80**	**37.45**
40 (13/16/11)	X-ALMA-13B	**35.76**	**38.22**	-	-	**35.12**	**41.53**	-	-	**22.16**	35.74	-	-
	LMT-60-4B	33.79	37.97	-	-	31.87	40.42	-	-	21.17	35.38	-	-
	LMT-60-8B	35.00	38.16	-	-	33.01	40.49	-	-	21.96	**36.23**	-	-
54 (13/18/23)	Aya-101-13B	22.75	30.67	12.86	24.10	20.02	32.61	10.33	24.01	7.70	29.01	3.71	21.98
	LMT-60-4B	33.79	37.97	21.42	41.10	30.96	40.11	19.13	40.61	20.05	36.17	11.98	37.02
	LMT-60-8B	**35.00**	**38.16**	**22.48**	**41.33**	**32.11**	**40.22**	**20.08**	**40.77**	**20.85**	**37.26**	**12.53**	**37.77**
55 (13/18/24)	LLaMAX3-Alpaca	24.51	32.28	12.94	26.18	22.53	34.46	9.41	25.61	10.96	27.51	4.57	20.67
	LMT-60-4B	33.79	37.97	21.42	41.10	30.96	40.11	19.13	40.61	20.38	36.01	12.27	37.10
	LMT-60-8B	**35.00**	**38.16**	**22.48**	**41.33**	**32.11**	**40.22**	**20.08**	**40.77**	**21.11**	**37.01**	**12.82**	**37.80**
59 (13/18/28)	NLLB-54B	33.02	**39.15**	20.81	24.51	31.67	**41.60**	18.40	24.96	19.20	**37.06**	10.79	25.04
	LMT-60-4B	33.79	37.97	21.42	41.10	30.96	40.11	19.13	40.61	18.93	34.77	11.56	36.79
	LMT-60-8B	**35.00**	38.16	**22.48**	**41.33**	**32.11**	40.22	**20.08**	**40.77**	**19.75**	35.81	**12.21**	**37.47**

Table 7: BLEU scores of our LMT models compared with a range of general-purpose multilingual LLMs and dedicated MMT models. Evaluation is conducted only on the intersection of language pairs supported by each baseline and our models. The first column (# Langs) denotes the number of overlapping languages, followed by their distribution across resource tier (high/medium/low). **Bold** numbers indicate the best in each group, and the underlined numbers the second best. The symbol '-' indicates directions not supported by the baseline model.

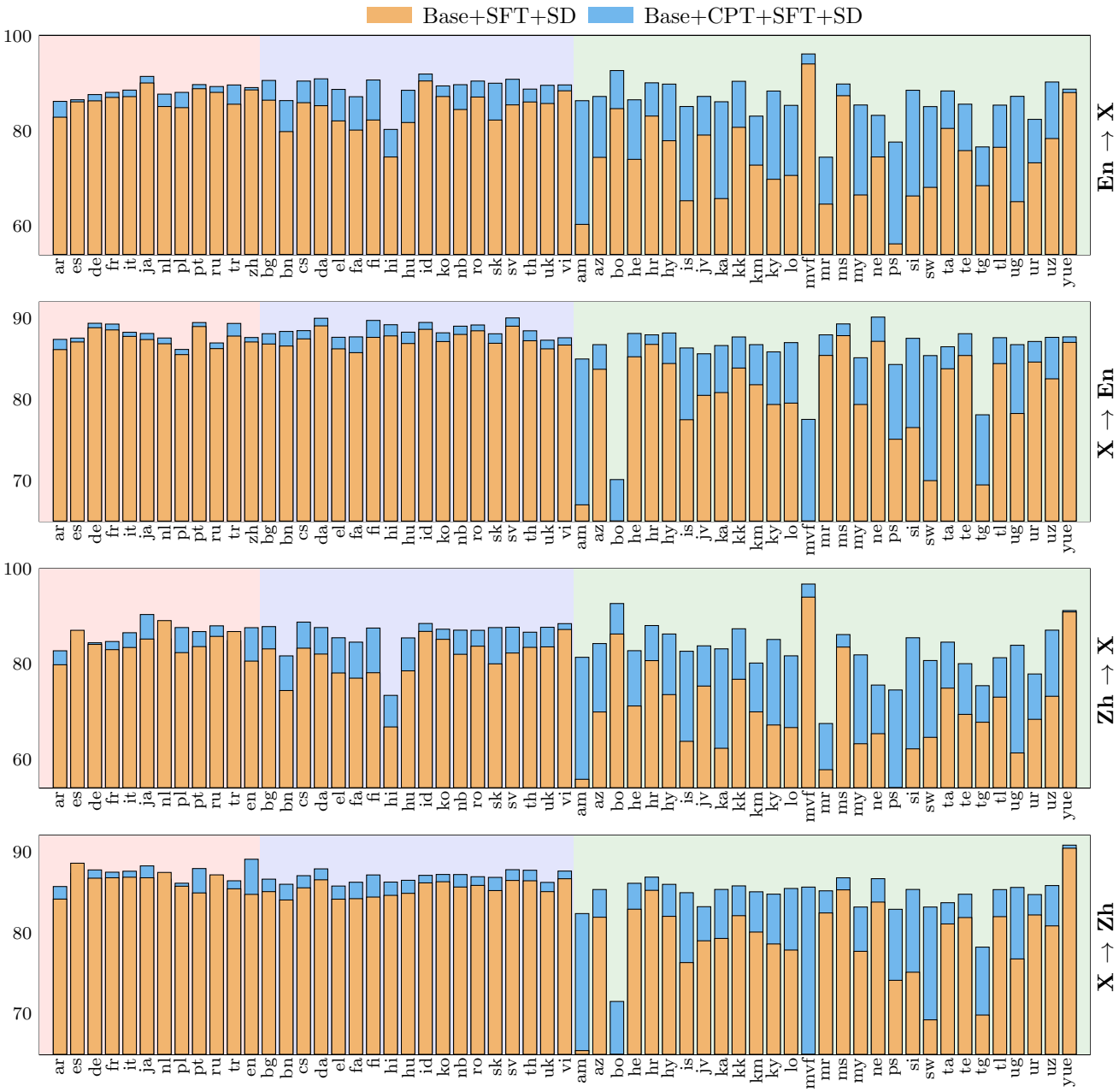


Figure 6: Performance improvements brought by Continued Pre-training (CPT). Languages are grouped by resource level: high, medium, and low. The orange portion of the bar shows the performance of the model without CPT (Base+SFT+SD), while the total height of the bar represents the performance after including the CPT stage (Base+CPT+SFT+SD). The blue portion visually represents the performance gain (ΔCOMET) contributed by CPT.

# Langs (Hig./Med./Low)	Model	High Resource				Medium Resource				Low Resource			
		En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh
10 (9/1/0)	TowerInstruct-13B	82.27	83.33	81.78	81.06	84.28	81.29	82.14	78.12	-	-	-	-
	LMT-60-4B	82.64	83.37	82.66	85.09	86.92	84.35	85.82	84.96	-	-	-	-
	LMT-60-8B	83.04	83.54	83.22	85.01	87.39	84.97	86.73	85.29	-	-	-	-
23 (13/9/1)	Aya-expanse-8B	82.50	83.03	82.85	82.08	83.38	83.78	83.38	82.10	82.92	83.46	81.59	80.98
	LMT-60-4B	83.11	82.93	82.79	84.46	83.97	82.95	83.06	83.57	80.95	83.14	77.72	83.39
	LMT-60-8B	83.41	83.15	83.32	84.59	84.12	83.84	83.72	84.36	81.34	83.47	79.18	83.80
24 (13/8/3)	GemmaX2-28-9B	75.49	76.42	74.40	76.26	74.92	68.62	75.34	75.97	75.04	58.88	73.21	76.92
	LMT-60-4B	83.11	82.93	82.79	84.46	83.64	83.40	82.68	83.67	81.30	82.92	78.80	82.78
	LMT-60-8B	83.41	83.15	83.32	84.59	83.77	84.18	83.02	84.76	81.27	83.60	79.49	83.65
28 (13/9/6)	Hunyuan-MT-7B	85.35	83.32	84.21	85.30	86.59	84.40	84.55	85.11	81.29	83.58	77.83	83.23
	LMT-60-4B	83.11	82.93	82.79	84.46	83.74	83.14	82.77	83.54	79.74	81.92	76.81	80.70
	LMT-60-8B	83.41	83.15	83.32	84.59	83.96	83.98	83.22	84.60	80.83	83.56	78.24	82.93
33 (13/16/4)	X-ALMA-13B	80.15	79.31	-	-	81.60	78.87	-	-	75.35	75.98	-	-
	LMT-60-4B	83.11	82.93	-	-	84.69	83.30	-	-	78.41	82.82	-	-
	LMT-60-8B	83.41	83.15	-	-	84.96	84.10	-	-	78.78	83.72	-	-
39 (13/18/8)	Aya-101-13B	75.14	77.68	77.85	75.00	76.64	78.69	77.94	75.06	71.80	77.87	70.82	74.04
	LMT-60-4B	83.11	82.93	82.79	84.46	84.62	83.25	83.50	83.76	80.49	81.76	77.99	81.04
	LMT-60-8B	83.41	83.15	83.32	84.59	84.88	84.04	84.07	84.40	81.36	83.23	79.11	82.94

Table 8: COMET-22 results on the WMT24++ benchmark (document-level). Only models without WMT24++ training are included. Evaluation is conducted only on the intersection of language pairs supported by each baseline and our models. The first column (# Langs) denotes the number of overlapping languages, followed by their distribution across resource tier (high/medium/low). The symbol '-' indicates directions not supported by the baseline model.

# Langs (Hig./Med./Low)	Model	High Resource				Medium Resource				Low Resource			
		En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh	En→X	X→En	Zh→X	X→Zh
10 (9/1/0)	TowerInstruct-13B	32.33	35.15	15.54	25.74	16.07	16.92	10.81	15.26	-	-	-	-
	LMT-60-4B	28.89	28.59	17.41	34.83	15.39	17.10	18.43	30.19	-	-	-	-
	LMT-60-8B	32.45	34.24	19.88	29.31	25.28	21.57	22.41	24.05	-	-	-	-
23 (13/9/1)	Aya-expanse-8B	28.46	32.20	17.15	31.30	27.15	31.95	16.36	30.07	24.48	34.64	13.12	29.58
	LMT-60-4B	25.00	26.40	15.46	33.16	19.33	24.47	14.12	27.48	15.28	30.02	8.12	30.96
	LMT-60-8B	28.18	29.74	17.48	28.18	26.58	21.72	17.17	26.58	15.78	35.22	8.12	29.73
24 (13/8/3)	GemmaX2-28-9B	16.84	22.03	10.56	21.98	12.39	13.60	8.06	16.88	12.60	9.60	7.19	19.33
	LMT-60-4B	25.00	26.40	15.46	33.16	15.27	21.77	12.03	25.79	20.41	25.37	11.43	27.58
	LMT-60-8B	28.18	29.74	17.48	28.18	20.92	18.76	14.53	26.57	20.81	27.36	12.06	29.42
28 (13/9/6)	Hunyuan-MT-7B	19.28	26.60	14.40	28.44	14.66	23.38	11.12	25.21	7.47	22.09	5.51	20.84
	LMT-60-4B	25.00	26.40	15.46	33.16	15.72	22.22	11.96	25.81	14.13	20.09	7.80	21.27
	LMT-60-8B	28.18	29.74	17.48	28.18	21.40	18.74	14.56	26.30	16.29	18.54	9.33	23.59
33 (13/16/4)	X-ALMA-13B	11.98	12.58	-	-	11.99	11.55	-	-	6.72	9.07	-	-
	LMT-60-4B	25.00	26.40	-	-	21.24	26.08	-	-	14.84	23.11	-	-
	LMT-60-8B	28.18	29.74	-	-	28.10	26.45	-	-	17.00	22.03	-	-
39 (13/18/8)	Aya-101-13B	10.26	15.01	5.91	7.53	11.39	15.75	5.11	8.96	6.63	15.39	2.40	8.61
	LMT-60-4B	25.00	26.40	15.46	33.16	20.08	25.17	13.41	28.36	15.73	19.76	8.90	22.39
	LMT-60-8B	28.18	29.74	17.48	28.18	26.78	25.62	15.91	26.68	17.19	19.33	9.87	22.59

Table 9: BLEU results on the WMT24++ benchmark (document-level). Only models without WMT24++ training are included. Evaluation is conducted only on the intersection of language pairs supported by each baseline and our models. The first column (# Langs) denotes the number of overlapping languages, followed by their distribution across resource tier (high/medium/low). The symbol '-' indicates directions not supported by the baseline model.

Languages	Data Source
en	SlimPajama-6B (Soboleva et al., 2023)
zh	Skywork (Wei et al., 2023)
ar, th, vi	WanJuanSiLu (Yu et al., 2025)
cs, fa, hi, de, fr, it, pt, es, el, uk, fi sv, nb, da, ro, hu, sk, bg, nl, pl, tr	CulturaX(Nguyen et al., 2023)
hr, sw	MADLAD-400 (Kudugunta et al., 2023)
ne	CulturaX, MADLAD-400, MrBinit/Nepali [a]
bn, he, id, ms, az, kk, ps, ta, ur, uz, am, jv	CulturaX, MADLAD-400, Opus-Corpora(Zhang et al., 2020), fineweb-2(HuggingFaceFW, 2024)
km, lo, my, tl	CulturaX, MADLAD-400, fineweb-2, Opus-Corpora, C4 (Raffel et al., 2020), OSCAR [b]
ja, ko, ru	WanJuanSiLu (Yu et al., 2025)
ug, bo, mvf	In-house dataset
yue	AlienKevin/yue_and_zh_sentences [c]
si, te, mr, is, tg, ky, ka, hy	HPLT (Burchell et al., 2025), OSCAR, Wikipedia

Table 10: Monolingual data sources information for 60 languages.

[a] https://huggingface.co/datasets/MrBinit/nepali_dataset_text_cleaned
[b] https://oscar-project.github.io/documentation/versions/oscar-2301/
[c] https://huggingface.co/datasets/AlienKevin/yue_and_zh_sentences

ISO Code	Language	Script	Family	Resource
13 High-resource Languages				
en	English	Latin	Indo-European	High
ar	Arabic	Arabic	Afro-Asiatic	High
es	Spanish	Latin	Indo-European	High
de	German	Latin	Indo-European	High
fr	French	Latin	Indo-European	High
it	Italian	Latin	Indo-European	High
ja	Japanese	Japanese	Japonic	High
nl	Dutch	Latin	Indo-European	High
pl	Polish	Latin	Indo-European	High
pt	Portuguese	Latin	Indo-European	High
ru	Russian	Cyrillic	Indo-European	High
tr	Turkish	Latin	Turkic	High
zh	Chinese	Han	Sino-Tibetan	High
18 Medium-resource Languages				
bg	Bulgarian	Cyrillic	Indo-European	Mid
bn	Bengali	Bengali	Indo-European	Mid
cs	Czech	Latin	Indo-European	Mid
da	Danish	Latin	Indo-European	Mid
el	Modern Greek	Greek	Indo-European	Mid
fa	Persian	Arabic	Indo-European	Mid
fi	Finnish	Latin	Uralic	Mid
hi	Hindi	Devanagari	Indo-European	Mid
hu	Hungarian	Latin	Uralic	Mid
id	Indonesian	Latin	Austronesian	Mid
ko	Korean	Hangul	Koreanic	Mid
nb	Norwegian	Latin	Indo-European	Mid
ro	Romanian	Latin	Indo-European	Mid
sk	Slovak	Latin	Indo-European	Mid
sv	Swedish	Latin	Indo-European	Mid
th	Thai	Thai	Tai-Kadai	Mid
uk	Ukrainian	Cyrillic	Indo-European	Mid
vi	Vietnamese	Latin	Austroasiatic	Mid
29 Low-resource Languages				
am	Amharic	Ge'ez	Afro-Asiatic	Low
az	Azerbaijani	Latin	Turkic	Low
bo	Tibetan	Tibetan	Sino-Tibetan	Low
he	Modern Hebrew	Hebrew	Afro-Asiatic	Low
hr	Croatian	Latin	Indo-European	Low
hy	Armenian	Armenian	Indo-European	Low
is	Icelandic	Latin	Indo-European	Low
jv	Javanese	Latin	Austronesian	Low
ka	Georgian	Georgian	Kartvelian	Low
kk	Kazakh	Cyrillic	Turkic	Low
km	Central Khmer	Khmer	Austroasiatic	Low
ky	Kirghiz	Cyrillic	Turkic	Low
lo	Lao	Lao	Tai-Kadai	Low

ISO Code	Language	Script	Family	Resource
mvf	Inner Mongolian	Mongolian	Mongolic	Low
mr	Marathi	Devanagari	Indo-European	Low
ms	Malay	Latin	Austronesian	Low
my	Burmese	Myanmar	Sino-Tibetan	Low
ne	Nepali	Devanagari	Indo-European	Low
ps	Pashto	Arabic	Indo-European	Low
si	Sinhala	Sinhala	Indo-European	Low
sw	Swahili	Latin	Atlantic-Congo	Low
ta	Tamil	Tamil	Dravidian	Low
te	Telugu	Telugu	Dravidian	Low
tg	Tajik	Cyrillic	Indo-European	Low
tl	Tagalog	Latin	Austronesian	Low
ug	Uighur	Arabic	Turkic	Low
ur	Urdu	Arabic	Indo-European	Low
uz	Uzbek	Latin	Turkic	Low
yue	Yue Chinese	Han	Sino-Tibetan	Low

Table 11: Detailed information of 60 languages. The languages are grouped into categories based on their data ratios in the CulturaX (Nguyen et al., 2024): High Resource (>1%), Medium Resource (0.1%–1%], and Low Resource ($\leqslant$0.1%).

ISO Code	Language	Script	Family	Auxiliary Language	ISO Code
es	Spanish	Latin	Indo-European	Portuguese	pt
de	German	Latin	Indo-European	Dutch	nl
fr	French	Latin	Indo-European	Italian	it
it	Italian	Latin	Indo-European	French	fr
nl	Dutch	Latin	Indo-European	German	de
pl	Polish	Latin	Indo-European	Czech	cs
bg	Bulgarian	Cyrillic	Indo-European	Russian	ru
cs	Czech	Latin	Indo-European	Polish	pl
da	Danish	Latin	Indo-European	Norwegian	nb
fa	Persian	Arabic	Indo-European	Arabic	ar
fi	Finnish	Latin	Uralic	Hungarian	hu
hi	Hindi	Devanagari	Indo-European	Bengali	bn
hu	Hungarian	Latin	Uralic	Finnish	fi
id	Indonesian	Latin	Austronesian	Dutch	nl
nb	Norwegian	Latin	Indo-European	Danish	da
ro	Romanian	Latin	Indo-European	Italian	it
sk	Slovak	Latin	Indo-European	Czech	cs
sv	Swedish	Latin	Indo-European	Norwegian	nb
uk	Ukrainian	Cyrillic	Indo-European	Russian	ru
vi	Vietnamese	Latin	Austroasiatic	French	fr
az	Azerbaijani	Latin	Turkic	Turkish	tr
hr	Croatian	Latin	Indo-European	Czech	cs
is	Icelandic	Latin	Indo-European	Danish	da
jv	Javanese	Latin	Austronesian	Indonesian	id
kk	Kazakh	Cyrillic	Turkic	Russian	ru
ky	Kirghiz	Cyrillic	Turkic	Russian	ru
lo	Lao	Lao	Tai-Kadai	Thai	th
mr	Marathi	Devanagari	Indo-European	Hindi	hi
ms	Malay	Latin	Austronesian	Indonesian	id
ne	Nepali	Devanagari	Indo-European	Hindi	hi
ps	Pashto	Arabic	Indo-European	Persian	fa
tg	Tajik	Cyrillic	Indo-European	Russian	ru
tl	Tagalog	Latin	Austronesian	Indonesian	id
ug	Uighur	Arabic	Turkic	Persian	fa
ur	Urdu	Arabic	Indo-European	Persian	fa
uz	Uzbek	Latin	Turkic	French	fr

Table 12: The set of languages that utilize the Parallel Multilingual Prompting (PMP) method and their corresponding auxiliary languages.

Direction	COMET				BLEU			
	0.6B	1.7B	4B	8B	0.6B	1.7B	4B	8B
13 High-resource Languages								
en→ar	82.97	85.51	87.01	87.14	17.64	19.45	22.54	25.09
en→es	84.77	86.67	87.32	87.40	24.82	27.14	29.04	29.39
en→de	83.58	87.66	88.54	88.87	28.92	35.39	37.74	38.79
en→fr	85.70	88.10	88.70	89.01	40.79	45.41	47.42	49.50
en→it	86.17	88.58	89.40	89.49	25.70	27.99	30.22	31.21
en→ja	90.09	91.61	92.26	92.38	30.69	35.35	37.24	38.35
en→nl	84.75	87.72	88.65	88.78	22.32	24.82	26.11	27.42
en→pl	84.48	88.86	90.04	90.54	15.58	19.31	21.37	22.33
en→pt	88.03	89.78	90.32	90.41	42.78	46.27	48.32	49.37
en→ru	86.49	89.22	90.37	90.47	24.92	28.20	30.73	31.66
en→tr	86.72	89.74	90.78	90.90	20.36	23.86	25.59	27.55
en→zh	87.92	89.49	89.82	89.84	44.12	48.03	49.21	49.34
18 Medium-resource Languages								
en→bg	87.90	90.74	91.39	91.58	31.84	35.90	37.52	38.85
en→bn	83.20	86.81	87.85	87.60	13.47	15.01	15.91	16.99
en→cs	87.19	91.04	92.05	92.18	25.48	28.95	31.15	32.30
en→da	88.27	90.60	91.65	91.76	38.27	41.34	44.11	45.04
en→el	84.59	88.68	89.71	90.08	19.39	22.54	24.99	26.10
en→fa	82.58	86.70	88.31	88.51	19.56	22.53	24.64	25.85
en→fi	85.85	90.98	92.67	92.83	15.25	19.18	22.85	23.77
en→hi	76.81	80.45	81.85	82.28	24.11	27.00	28.85	30.45
en→hu	83.81	88.80	90.26	90.28	17.04	20.48	23.59	24.14
en→id	90.62	91.93	92.38	92.47	43.23	45.00	45.71	47.10
en→ko	87.55	89.38	90.47	90.47	26.61	28.17	29.38	31.36
en→nb	87.89	89.91	90.64	90.77	28.57	30.39	31.46	31.87
en→ro	86.68	90.20	91.36	91.55	32.33	35.74	38.28	40.00
en→sk	85.90	90.22	91.44	91.86	25.60	28.56	31.41	32.84
en→sv	87.40	90.52	91.51	91.63	36.37	39.99	42.74	43.44
en→th	85.83	88.63	89.79	89.86	11.56	11.82	14.14	14.94
en→uk	85.08	89.50	90.64	91.00	20.86	25.40	28.64	30.21
en→vi	88.25	89.65	90.22	90.31	39.10	40.82	41.88	42.74
29 Low-resource Languages								
en→am	79.76	86.81	88.40	88.67	5.38	8.57	10.77	11.57
en→az	84.31	88.17	89.25	89.15	11.80	13.04	14.23	14.38
en→bo	87.37	91.51	92.82	93.09	1.71	3.63	3.26	2.84
en→he	81.79	83.31	87.55	87.80	17.16	20.37	24.68	27.15
en→hr	86.02	90.06	91.56	91.58	22.48	25.60	28.98	29.62
en→hy	86.42	89.90	91.10	91.11	17.23	20.13	22.33	23.10
en→is	81.43	85.65	86.91	87.38	18.77	21.95	23.75	24.53
en→jv	85.89	87.89	88.37	88.41	23.92	27.22	28.74	29.58
en→ka	81.89	86.89	89.02	89.17	10.89	12.22	14.84	14.90
en→kk	88.46	91.05	91.53	91.69	18.53	21.52	23.06	23.76
en→km	79.32	83.38	85.04	85.11	6.52	7.72	7.50	7.49
en→ky	85.84	89.11	89.98	89.84	11.35	13.58	14.45	14.73
en→lo	80.52	85.09	86.81	86.78	11.57	13.64	14.31	14.42
en→mvf	95.25	95.22	95.48	95.55	15.12	11.74	15.17	16.67
en→mr	70.80	75.00	76.45	76.67	11.20	14.13	15.54	16.72
en→ms	88.98	90.23	90.62	90.70	39.25	40.67	41.05	41.95
en→my	77.47	86.78	88.79	88.97	2.72	3.36	4.28	4.51
en→ne	81.45	84.14	85.23	85.22	15.79	18.21	19.45	20.49
en→ps	75.67	79.49	80.25	80.15	9.36	11.06	12.04	12.03
en→si	82.90	88.20	90.29	90.64	10.33	12.89	15.50	17.16
en→sw	82.88	85.47	86.08	86.35	30.52	33.45	35.56	36.72
en→ta	85.42	89.05	90.27	90.35	10.79	13.07	14.29	15.33
en→te	83.10	86.59	87.49	87.59	15.33	17.57	19.29	20.54
en→tg	75.64	77.25	77.68	77.81	15.19	18.64	20.73	21.87
en→tl	83.16	85.42	86.37	86.18	32.19	35.54	37.34	38.30
en→ug	84.05	87.82	88.93	88.68	12.13	14.52	16.70	18.01
en→ur	78.67	82.87	84.23	84.08	17.40	19.49	21.38	22.16
en→uz	88.52	90.74	91.45	91.32	15.18	18.21	20.46	20.81
en→yue	85.88	88.85	89.63	89.85	5.02	9.32	5.58	8.32

Table 13: COMET-22 and SacreBLEU scores of LMT on the FLORES-200 devtest set (En → X).

Direction	COMET				BLEU			
	0.6B	1.7B	4B	8B	0.6B	1.7B	4B	8B
13 High-resource Languages								
ar→en	84.53	87.15	87.78	87.97	33.56	38.83	40.93	41.40
es→en	86.38	87.41	87.85	87.94	29.17	31.27	33.64	33.57
de→en	88.14	89.38	89.59	89.52	39.99	43.33	44.45	43.99
fr→en	88.43	89.34	89.48	89.49	41.93	44.98	45.34	45.71
it→en	87.09	88.14	88.60	88.59	31.12	33.97	35.88	35.95
ja→en	86.85	88.30	88.71	88.82	25.51	28.35	30.40	31.09
nl→en	86.27	87.70	87.84	87.94	29.92	32.36	33.05	33.56
pl→en	84.47	86.36	86.74	86.72	27.50	30.83	31.93	31.94
pt→en	88.54	89.53	89.98	89.96	45.49	48.63	50.56	50.51
ru→en	85.50	86.81	87.21	87.36	31.91	35.56	37.70	38.03
tr→en	87.12	89.28	89.76	89.83	32.26	36.79	39.18	39.31
zh→en	86.73	87.79	87.96	87.87	28.62	30.95	32.59	32.88
18 Medium-resource Languages								
bg→en	86.72	88.05	88.47	88.40	36.89	40.82	41.85	41.68
bn→en	85.21	88.04	88.98	89.25	25.12	32.25	34.61	35.73
cs→en	86.78	88.48	88.98	88.88	35.38	39.52	41.44	41.28
da→en	88.86	90.15	90.36	90.27	44.61	48.59	48.66	47.73
el→en	85.09	87.49	88.12	88.14	30.94	35.74	37.86	38.17
fa→en	85.32	87.87	88.48	88.65	30.48	35.40	37.34	38.10
fi→en	86.46	89.36	90.09	90.23	27.84	33.16	34.76	35.45
hi→en	86.97	89.16	89.94	89.99	31.72	38.53	40.53	40.44
hu→en	86.30	88.36	88.91	88.93	30.87	35.52	37.16	37.17
id→en	88.61	89.72	89.93	89.95	41.19	44.60	45.69	45.56
ko→en	86.62	88.31	88.68	88.61	25.94	30.12	30.95	31.66
nb→en	87.57	88.96	89.24	89.29	41.13	43.75	44.30	44.40
ro→en	88.00	89.42	89.59	89.66	39.91	43.51	44.88	44.81
sk→en	86.32	88.24	88.64	88.51	34.59	39.32	40.64	40.35
sv→en	88.52	90.04	90.27	90.26	44.68	48.14	48.82	48.00
th→en	86.78	88.62	89.13	89.01	28.22	32.63	34.02	34.46
uk→en	85.55	87.35	87.71	87.64	35.01	39.42	40.15	40.47
vi→en	86.70	87.83	88.21	88.15	35.14	37.47	38.37	38.51
29 Low-resource Languages								
am→en	77.78	83.41	86.36	87.20	15.76	25.52	31.42	33.47
az→en	84.26	86.46	87.42	87.45	19.75	23.76	26.12	26.73
bo→en	64.15	70.09	72.90	74.04	6.20	11.59	15.67	17.72
he→en	84.98	87.72	88.58	88.65	35.83	41.79	44.13	44.58
hr→en	85.83	88.01	88.41	88.37	33.44	38.05	39.41	39.15
hy→en	85.31	87.82	88.75	88.98	30.29	36.15	38.69	40.41
is→en	82.28	85.90	87.04	87.41	28.44	33.98	36.39	37.38
jv→en	81.57	84.96	86.30	86.53	32.90	38.37	41.37	41.94
ka→en	83.54	86.35	87.29	87.62	22.45	26.80	29.33	30.21
kk→en	84.96	87.67	88.29	88.57	27.53	32.83	34.90	35.83
km→en	83.37	86.26	87.56	88.02	23.47	30.17	34.01	35.15
ky→en	83.17	85.84	86.66	86.65	20.17	23.50	25.54	26.25
lo→en	83.32	86.02	87.82	88.31	25.63	31.88	36.87	38.06
mvf→en	69.36	74.64	76.58	77.86	10.00	14.22	18.04	20.45
mr→en	84.96	87.71	88.73	88.73	27.81	33.60	36.46	37.52
ms→en	88.02	89.28	89.59	89.65	41.94	45.06	45.99	46.28
my→en	80.74	84.80	86.62	86.88	17.59	23.72	26.20	27.78
ne→en	87.81	89.95	90.66	90.90	30.23	37.25	39.12	40.79
ps→en	80.52	83.94	85.41	85.76	24.03	29.30	32.95	33.63
si→en	81.72	85.36	88.44	89.00	19.28	26.12	32.61	34.32
sw→en	81.64	84.74	86.26	86.37	34.29	39.21	42.08	43.14
ta→en	82.22	86.08	87.38	87.70	22.40	28.89	31.98	33.11
te→en	84.09	87.62	88.78	89.10	26.63	35.44	37.64	39.50
tg→en	73.45	77.79	79.12	79.49	26.06	32.20	34.85	36.21
tl→en	84.73	87.21	88.15	88.39	38.80	44.41	47.59	48.05
ug→en	83.68	86.76	87.74	87.84	21.31	25.96	28.41	28.83
ur→en	83.64	87.08	87.99	88.23	26.55	33.38	35.50	36.67
uz→en	85.05	87.69	88.42	88.57	28.58	34.00	35.75	36.32
yue→en	86.53	87.81	88.13	88.06	28.84	32.09	32.60	33.58

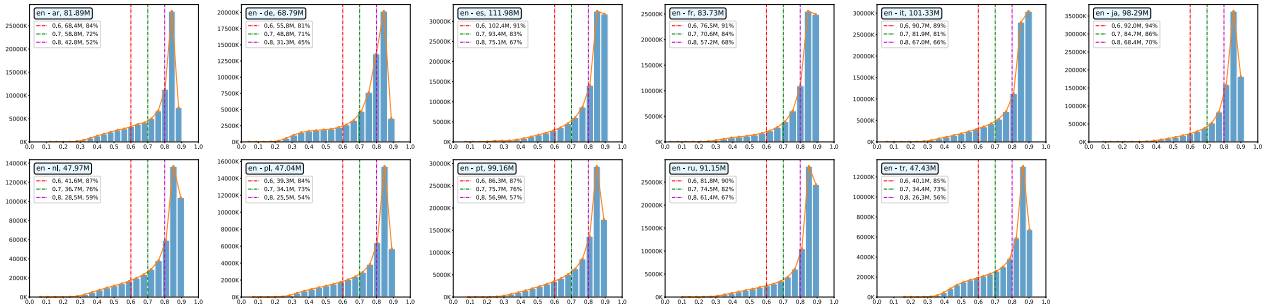
Table 14: COMET-22 and SacreBLEU scores of LMT on the FLORES-200 devtest set (X → En).

Direction	COMET				BLEU			
	0.6B	1.7B	4B	8B	0.6B	1.7B	4B	8B
13 High-resource Languages								
zh→ar	79.71	83.02	84.27	84.31	8.71	10.67	13.14	13.64
zh→en	86.73	87.79	87.96	87.87	28.62	30.95	32.59	32.88
zh→es	82.67	85.05	85.67	86.03	16.32	19.01	20.04	21.03
zh→de	80.41	84.75	85.58	86.10	14.97	18.62	21.15	22.57
zh→fr	82.06	84.86	85.64	85.80	21.16	25.07	27.75	29.36
zh→it	83.75	86.70	87.61	87.71	15.70	18.60	20.35	21.48
zh→ja	88.87	90.71	91.20	91.21	24.08	27.48	29.90	30.81
zh→nl	81.67	85.47	86.44	86.66	13.11	15.77	17.65	18.91
zh→pl	83.08	87.95	89.16	89.41	9.86	12.36	14.86	15.16
zh→pt	84.89	86.97	87.46	87.75	21.17	24.49	26.04	27.65
zh→ru	84.76	88.19	88.94	89.24	14.21	17.14	18.67	20.11
zh→tr	81.09	85.33	86.51	86.80	10.47	12.93	14.87	16.15
18 Medium-resource Languages								
zh→bg	84.63	88.06	89.24	89.39	17.14	19.66	21.96	23.17
zh→bn	78.06	82.14	83.52	83.57	6.72	7.98	9.16	10.22
zh→cs	84.72	89.18	90.32	90.65	13.30	17.02	18.97	19.98
zh→da	84.95	87.85	88.70	88.97	18.89	21.87	24.13	25.02
zh→el	81.00	85.66	87.02	87.53	10.69	12.65	15.54	16.67
zh→fa	79.75	84.04	85.86	85.99	11.50	13.87	16.00	16.39
zh→fi	81.87	87.79	89.79	90.17	8.83	10.96	14.32	15.12
zh→hi	69.07	73.82	75.37	75.50	12.42	15.62	17.77	18.27
zh→hu	79.62	85.75	87.26	87.52	9.78	12.58	15.56	15.94
zh→id	87.07	88.80	89.27	89.35	23.24	25.21	27.07	27.71
zh→ko	85.14	87.68	88.49	88.56	18.88	21.48	23.20	24.20
zh→nb	84.44	87.44	88.22	88.26	13.90	16.60	17.92	18.74
zh→ro	83.24	87.03	88.26	88.51	17.83	21.19	23.41	24.54
zh→sk	83.23	88.15	89.65	89.65	13.10	15.49	18.31	19.44
zh→sv	84.32	87.64	88.70	88.89	17.00	20.12	22.68	23.97
zh→th	83.82	86.58	87.42	87.56	8.20	9.17	10.46	11.87
zh→uk	82.86	87.74	89.13	89.42	10.96	14.72	16.84	17.96
zh→vi	86.97	88.57	89.06	89.12	27.63	29.38	31.07	32.32
29 Low-resource Languages								
zh→am	74.21	81.93	84.53	84.54	2.61	4.24	5.60	6.33
zh→az	80.11	85.27	86.28	86.43	7.95	9.61	10.57	10.75
zh→bo	88.72	92.10	93.26	93.03	1.61	2.39	2.12	2.17
zh→he	78.05	81.34	84.69	84.41	7.97	10.43	14.13	14.53
zh→hr	83.28	88.23	89.83	90.14	12.13	14.94	18.07	18.99
zh→hy	82.24	86.51	88.01	88.19	9.16	11.54	13.25	13.18
zh→is	78.16	82.82	84.73	84.71	10.50	12.47	15.06	15.34
zh→jv	81.64	84.92	85.38	85.35	11.06	14.80	16.08	16.47
zh→ka	77.88	84.30	86.20	86.62	6.61	8.30	10.02	11.27
zh→kk	84.72	87.86	88.69	88.82	10.15	12.27	13.89	14.60
zh→km	75.69	80.56	82.11	82.22	5.01	5.41	6.41	6.52
zh→ky	82.14	86.74	87.36	87.67	7.62	9.53	10.47	11.11
zh→lo	76.58	81.78	83.74	83.86	7.26	9.42	10.01	10.06
zh→mvf	96.15	95.96	96.20	96.25	25.34	25.23	31.97	35.50
zh→mr	63.03	68.39	70.03	70.45	6.15	7.81	9.02	9.65
zh→ms	85.16	86.77	87.09	87.21	20.51	22.09	23.22	24.06
zh→my	71.97	83.07	85.30	85.61	1.60	2.40	2.87	2.88
zh→ne	73.18	77.19	78.04	78.19	7.17	8.69	9.88	10.19
zh→ps	70.92	76.46	77.23	77.25	5.10	6.65	8.09	7.94
zh→si	78.49	85.07	87.50	87.84	5.71	7.68	9.68	10.54
zh→sw	78.23	81.45	82.27	82.63	13.87	16.80	18.80	19.85
zh→ta	80.41	85.38	86.49	86.78	5.95	7.35	8.67	8.84
zh→te	75.90	81.20	82.67	82.92	7.20	9.11	10.35	11.45
zh→tg	73.88	75.87	76.64	76.33	8.47	11.36	13.42	14.02
zh→tl	78.69	81.84	82.61	82.57	16.05	19.96	21.22	22.65
zh→ug	80.12	84.64	85.59	85.49	9.20	11.29	12.73	13.86
zh→ur	73.01	78.45	79.89	80.26	9.16	11.82	13.22	14.51
zh→uz	84.90	87.85	88.65	88.62	8.46	10.58	12.11	12.59
zh→yue	90.19	91.55	91.74	91.84	5.52	8.73	4.76	7.64

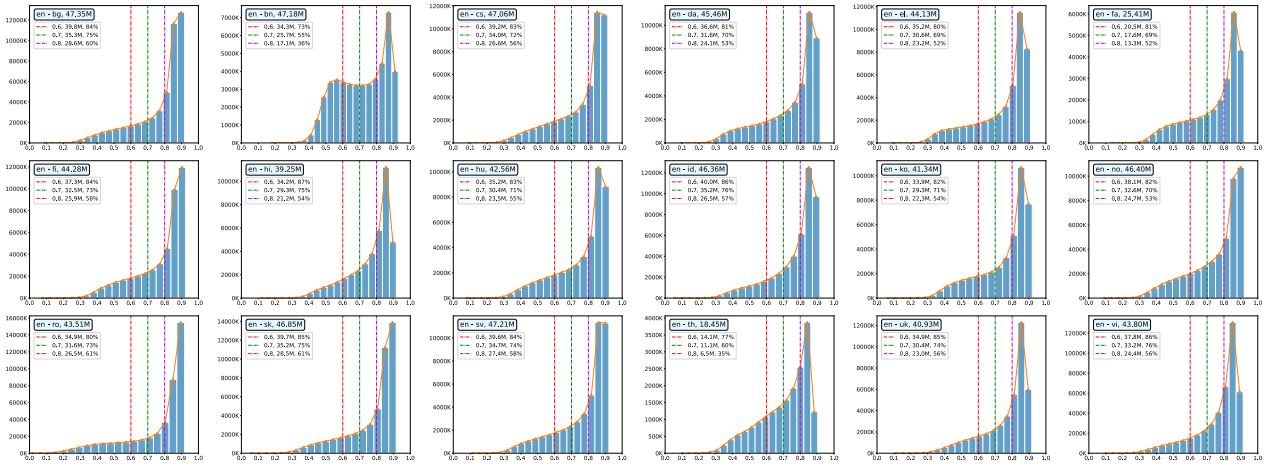
Table 15: COMET-22 and SacreBLEU scores of LMT on the FLORES-200 devtest set (Zh → X).

Direction	COMET				BLEU			
	0.6B	1.7B	4B	8B	0.6B	1.7B	4B	8B
13 High-resource Languages								
ar→zh	82.33	85.82	86.77	86.89	31.79	38.23	40.56	41.53
en→zh	87.92	89.49	89.82	89.84	44.12	48.03	49.21	49.34
es→zh	86.04	87.73	88.33	88.32	34.09	37.72	38.84	39.49
de→zh	85.80	87.94	88.43	88.46	36.81	41.25	42.37	42.21
fr→zh	86.55	87.94	88.49	88.36	37.54	41.60	42.80	42.83
it→zh	86.18	87.93	88.49	88.41	34.67	38.93	40.30	40.79
ja→zh	86.80	88.73	89.36	89.16	31.92	36.96	38.52	38.68
nl→zh	84.78	86.89	87.52	87.53	32.49	36.40	38.20	38.10
pl→zh	83.95	86.61	87.19	87.28	31.89	36.34	38.01	38.38
pt→zh	86.46	88.35	88.86	88.83	37.45	41.65	42.81	43.32
ru→zh	84.87	86.83	87.42	87.40	35.11	39.06	40.93	40.64
tr→zh	83.99	86.68	87.58	87.56	32.56	38.48	40.65	40.64
18 Medium-resource Languages								
bg→zh	84.40	86.94	87.59	87.57	35.67	40.08	41.89	41.68
bn→zh	81.62	85.77	87.13	87.45	27.02	34.16	36.67	37.84
cs→zh	85.06	87.33	88.01	87.82	34.77	39.89	41.41	41.42
da→zh	85.99	88.27	88.98	88.69	37.27	42.14	43.40	42.99
el→zh	82.17	85.81	86.94	86.91	30.55	36.86	39.04	39.39
fa→zh	83.02	86.58	87.53	87.47	30.68	37.07	38.64	39.38
fi→zh	83.35	87.26	88.12	88.25	30.90	37.21	39.83	39.49
hi→zh	82.65	86.27	87.32	87.67	30.19	36.55	39.01	39.84
hu→zh	83.67	86.82	87.61	87.48	32.40	38.49	40.41	39.98
id→zh	85.98	87.72	88.25	88.21	36.96	41.91	42.57	42.99
ko→zh	85.44	87.71	88.28	88.43	32.16	36.68	38.59	39.34
nb→zh	85.26	87.48	88.20	88.26	34.91	39.68	41.22	41.07
ro→zh	84.94	87.45	88.06	88.02	36.41	41.45	42.90	42.70
sk→zh	84.61	87.27	87.77	87.75	34.06	39.29	40.85	41.18
sv→zh	85.93	88.18	88.90	88.85	36.53	41.38	42.79	42.89
th→zh	85.72	88.20	88.78	88.70	32.22	37.83	39.51	39.52
uk→zh	83.86	86.61	87.32	87.19	34.54	39.76	41.31	41.11
vi→zh	86.53	88.24	88.61	88.46	35.73	40.02	40.87	41.02
29 Low-resource Languages								
am→zh	73.26	80.52	83.92	84.74	16.53	26.33	31.58	33.71
az→zh	82.36	85.47	86.37	86.44	27.51	31.84	34.02	34.52
bo→zh	64.72	72.11	74.55	76.06	7.84	17.05	21.15	23.54
he→zh	82.45	86.12	87.27	87.39	32.04	38.68	41.32	41.76
hr→zh	84.01	87.16	87.95	87.91	33.55	38.71	40.51	40.49
hy→zh	81.99	85.80	87.24	87.32	30.52	36.90	39.66	40.35
is→zh	80.51	85.02	86.20	86.30	28.44	34.82	37.69	38.34
jv→zh	79.12	83.44	85.00	85.06	29.27	35.64	37.69	38.54
ka→zh	80.88	85.28	86.86	86.99	25.37	33.32	36.54	37.23
kk→zh	82.87	85.96	87.03	87.13	30.63	36.15	39.02	38.87
km→zh	81.26	84.94	86.58	86.73	26.41	32.17	35.44	36.42
ky→zh	81.23	85.01	86.24	86.14	25.83	31.01	33.91	34.09
lo→zh	80.42	84.79	86.53	87.11	25.25	32.50	36.13	38.08
mvf→zh	77.98	81.50	83.95	85.12	24.01	46.70	54.35	58.64
mr→zh	80.23	85.17	86.60	86.63	27.00	35.08	37.42	37.95
ms→zh	84.87	87.04	87.73	87.86	36.14	40.54	41.99	42.03
my→zh	76.46	83.03	85.10	85.40	14.78	26.54	30.40	31.71
ne→zh	83.16	86.41	87.78	88.13	29.26	35.60	37.99	39.22
ps→zh	77.75	82.73	84.59	84.85	23.07	30.92	34.43	35.06
si→zh	78.07	83.39	86.90	87.38	19.46	27.73	35.46	36.24
sw→zh	78.09	83.09	84.73	85.03	27.18	35.13	37.31	38.63
ta→zh	78.56	83.57	85.34	85.67	22.72	31.58	34.49	35.40
te→zh	78.81	84.35	86.10	86.57	24.11	34.04	37.07	37.82
tg→zh	73.04	77.70	79.72	79.93	27.95	34.58	37.93	38.51
tl→zh	81.68	85.14	86.36	86.39	32.58	38.84	41.48	41.47
ug→zh	81.64	85.73	86.90	87.10	28.28	34.81	37.32	37.18
ur→zh	80.30	84.70	86.38	86.49	25.39	33.71	37.21	37.50
uz→zh	82.42	85.95	87.17	87.16	30.56	37.25	39.21	39.38
yue→zh	90.67	91.42	91.49	91.39	43.02	46.05	45.82	45.25

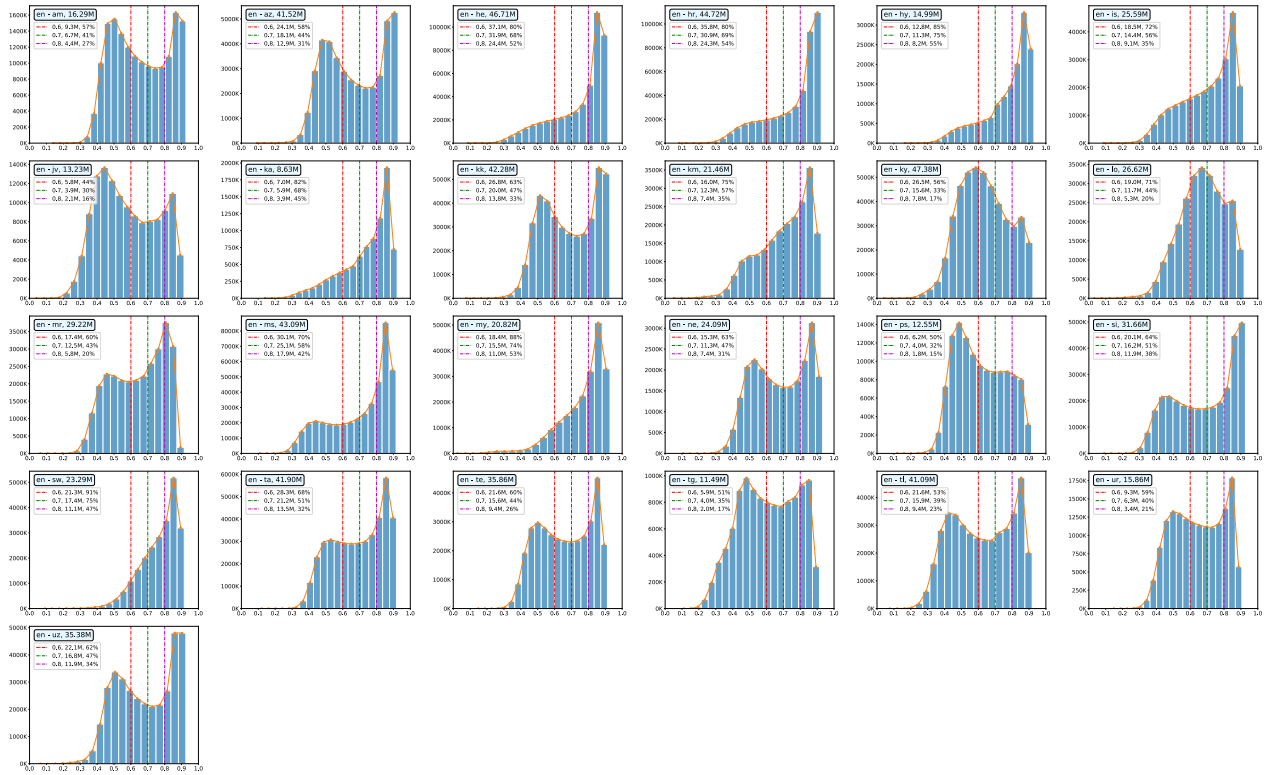
Table 16: COMET-22 and SacreBLEU scores of LMT on the FLORES-200 devtest set (X → Zh).



(a) High Resource Languages

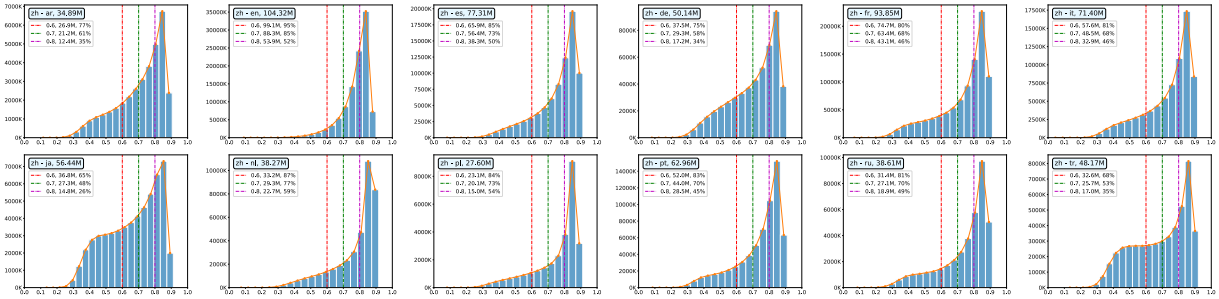


(b) Medium Resource Languages

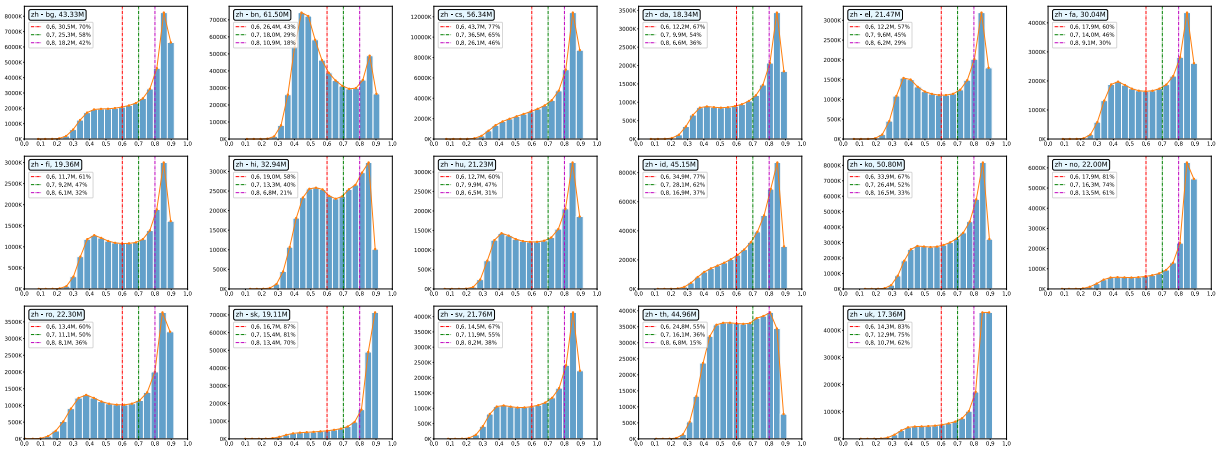


(c) Low Resource Languages

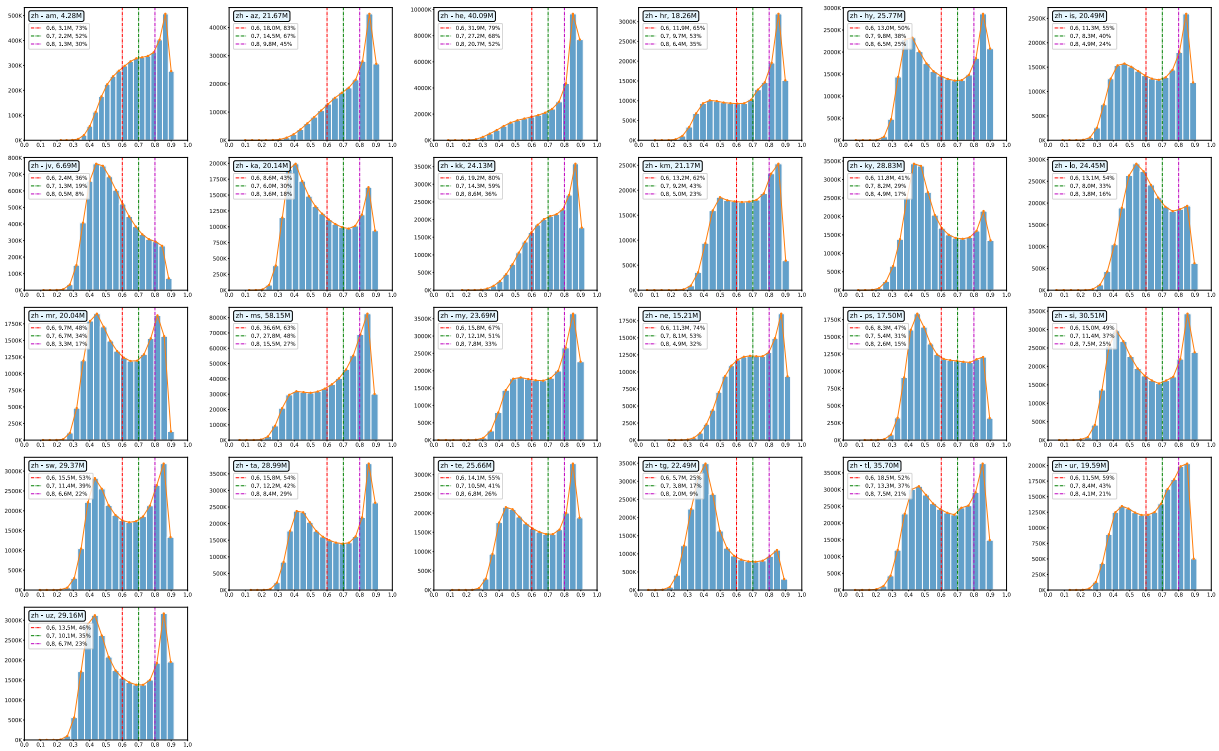
Figure 7: COMETKiwi score distributions for bilingual sentence pairs (En-X) are shown as histograms. Vertical lines indicate quality thresholds at 0.6 (red), 0.7 (green), and 0.8 (magenta), with the legend specifying the number and proportion of sentence pairs exceeding each threshold. Some language pairs are excluded due to COMETKiwi's limited language support.



(a) High Resource Languages



(b) Medium Resource Languages



(c) Low Resource Languages

Figure 8: COMETKiwi score distributions for bilingual sentence pairs (Zh-X) are shown as histograms. Vertical lines indicate quality thresholds at 0.6 (red), 0.7 (green), and 0.8 (magenta), with the legend specifying the number and proportion of sentence pairs exceeding each threshold. Some language pairs are excluded due to COMETKiwi's limited language support.