Beyond English: Toward Inclusive and Scalable Multilingual Machine Translation with LLMs

Anonymous ACL submission

Abstract

Large language models have significantly advanced Multilingual Machine Translation (MMT), yet the broad language coverage, consistent translation quality, and English-centric bias remain open challenges. To address these challenges, we introduce **LMT**, a suite of **L**arge-scale **M**ultilingual **T**ranslation models centered on both Chinese and English, covering 60 languages and 234 translation directions. In developing LMT, we identify the issue of **directional degeneration**, where symmetric multi-way fine-tuning data lead to an oversample of reverse directions, reducing target-side diversity and degraded performance. To mitigate this, we propose Strategic Downsampling, a simple yet effective method to resolve this degeneration. To further enhance cross-lingual transfer, we introduce **Parallel Multilingual Prompting (PMP)**, a technique that augments instructions with parallel sentences from typologically related auxiliary languages. Through rigorous data curation and proposed training strategies, LMT achieves competitive results, outperforming strong baseline models across a wide range of high-, medium-, and low-resource languages, with notable gains in Chinese-centric directions. We release LMT in four sizes (0.6B/1.7B/4B/8B) to support future research and provide a strong baseline for building more inclusive and advanced large-scale MMT.

1 Introduction

Multilingual machine translation (MMT) seeks to support translation among multiple languages with a unified model, has long been pursued for its parameter efficiency and the promise for cross-lingual transfer (Arivazhagan et al., 2019; Aharoni et al., 2019; Dabre et al., 2021; Costa-jussà et al., 2022). The emergence of large language models (LLMs) has significantly accelerated these goals due to their strong generative and cross-lingual generalization capabilities (Brown et al., 2020; Dubey et al., 2024; Yang et al., 2025). This momentum has spurred an active line of work on adapting LLMs for MMT (Yang et al., 2023; Alves et al., 2024; Xu et al., 2025; Cui et al., 2025; Luo et al., 2025a; Zheng et al., 2025b), as summarized in Table 1.

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.

Despite rapid progress, current LLM-based MMT models still struggle to jointly achieve broad language coverage and consistently strong translation quality — two critical attributes of an ideal MMT system. These persistent limitations often stem from a combination of factors, including sub-optimal adaptation strategies, the scarcity and imbalance of high-quality training data, and dependence on comparatively weaker or legacy base LLMs (Zhu et al., 2024b; Xu et al., 2024; Richburg and Carpuat, 2024). Beyond these issues lies a more fundamental, often overlooked problem: the pervasive English-centric bias inherited from pre-training corpora dominated by English. Consequently, research attention and system design remain excessively English-centric, leaving many languages underserved. Chinese, in particular, despite being among the most widely spoken languages and reflecting substantial real-world translation demand, illustrates this gap in today's LLM-based MMT landscape.

Addressing these gaps is non-trivial and involves

interdependent challenges in both data and training. On the data side, the long-tail distribution of internet corpora leads to foundational LLMs having inherent modeling deficiencies for many low-resource languages, as illustrated in Figure 1. This is compounded by the scarcity of high-quality parallel corpora, particularly for non-English-centric pairs such as those involving Chinese, which limits the effectiveness of supervised adaptation. On the training side, how to design strategies to effectively leverage data to foster robust translation capabilities remains an open question.

To tackle these challenges, we adopt the established Continued Pre-training (CPT) → Supervised Fine-tuning (SFT) pipeline and present LMT, a **L**arge-scale **M**ultilingual machine **T**ranslation model centered on both Chinese and English. LMT covers 60 languages and 234 (English ↔ 59 languages and Chinese ↔ 58 languages) translation directions. On the data front, we developed a comprehensive multi-stage data pipeline comprising large-scale collection, pseudo-parallel synthesis from open-source models, and multi-dimensional filtering, which lays a solid foundation for subsequent adaptation. During SFT, we uncover a critical pitfall-**directional degeneration**, where training with symmetric multi-way data can lead to marked performance degradation in certain directions. Our analysis traces this effect to low target-side diversity, and we propose a simple yet effective strategic downsampling method to resolve it. To further enhance cross-lingual transfer, we propose **Parallel Multilingual Prompting (PMP)**, which augments instructions with a parallel sentence in an auxiliary language chosen for typological proximity to the source or target language.

Combining rigorous data curation with targeted adaptation strategies, LMT delivers consistent gains over strong MMT baselines across diverse language pairs. Notably, it excels on Chinese-centric directions across high-, mid-, and low-resource tiers, often surpassing larger models. In summary, our contributions are threefold:

- We identify and analyze directional degeneration in large-scale multilingual SFT with multi-way data, and present a simple, practical mitigation to alleviate it.
- We introduce Parallel Multilingual Prompting (PMP), an intuitive and effective method that strengthens cross-lingual transfer by augmenting instructions with auxiliary translations in typologically related languages.
- We release **LMT**, a suite of Chinese–English-centric multilingual translation models in four sizes (0.6B/1.7B/4B/8B) covering 60 languages and 234 directions, providing strong baselines for future research.

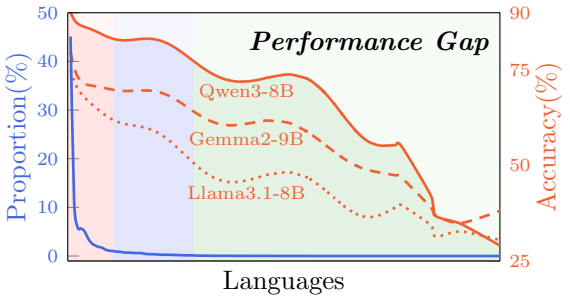


Figure 1: Performance of the base LLMs (orange lines) on the Belebele benchmark across 100+ languages. Languages on the x-axis are sorted by their internet corpus proportion (blue line) and grouped into high-, medium-, and low-resource tiers (shaded regions).

2 Related Work

The Paradigm Shift: From NMT to LLM-based MT The field of machine translation has evolved significantly over the past decade. The introduction of the Transformer architecture (Vaswani et al., 2017) marked the beginning of the Neural Machine Translation (NMT) era, leading to the development of powerful multilingual NMT systems trained from scratch, such as M2M-100 (Fan et al., 2021) and NLLB (Costa-jussà et al., 2022). More recently, the landscape has been reshaped by the rise of pre-trained Large Language Models (LLMs)(Brown et al., 2020; Dubey et al., 2024; Yang et al., 2025). Unlike NMT models built solely for translation, LLMs serve as general-purpose foundation models that can be adapted for any downstream tasks. This has catalyzed a paradigm shift from training bespoke NMT models to adapting existing LLMs (Yang et al., 2023; Alves et al., 2024; Xu et al., 2025; Cui et al., 2025), which not only improves translation quality but also enables advanced capabilities like context-awareness (Wang et al., 2024a, 2025b) and reasoning (Wang et al., 2025a; Chen et al., 2025).

LLM Adaptation for MMT The predominant paradigm for adapting LLMs to MMT involves a multi-stage pipeline, typically consisting of Continued Pre-training (CPT) and Supervised Fine-tuning

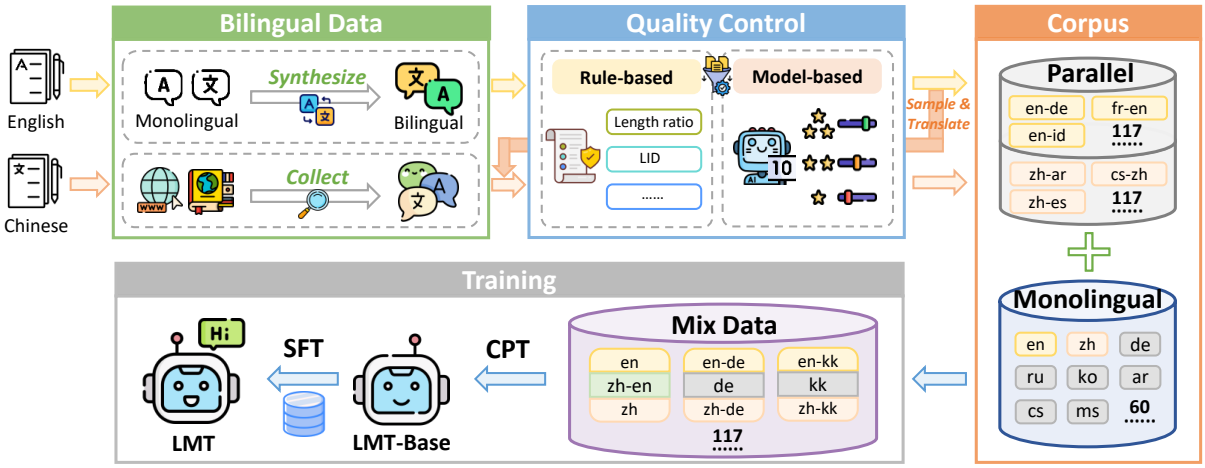


Figure 2: An overview of our methodology for LMT. The pipeline consists of two main stages: a hybrid data curation process (top) to build the training corpus, and a two-stage adaptation (bottom) involving CPT and SFT.

(SFT). Research in CPT has explored various data strategies: some work highlights the importance of monolingual data for enhancing language modeling (Xu et al., 2024), while others demonstrate the effectiveness of mixing bilingual and monolingual data to integrate translation-specific knowledge (Alves et al., 2024; Cui et al., 2025). The SFT stage further aligns the model with high-quality, human-curated data, often leveraging instruction tuning to handle diverse translation requests (Wang et al., 2024b; Luo et al., 2025b). However, several critical gaps remain. First, the potential pitfalls of specific SFT data structures, such as the symmetric, multi-way data commonly used for broad coverage, and the failure modes they might induce, have been largely under-explored. Second, while instruction tuning is prevalent, advanced prompting strategies that actively foster cross-lingual transfer are still relatively unexplored. Finally, much of this work remains predominantly English-centric, leaving a significant need for powerful, open models centered around other major global languages like Chinese. Notably, some recent models (Cui et al., 2025; Zheng et al., 2025a; Cheng et al., 2025) have also made strides in supporting Chinese-centric translation directions. However, their support for this language pair remains limited compared to the broader multilingual capabilities. Our work aims to address these specific gaps.

3 Methodology

3.1 The LMT Adaptation Framework

LMT is designed to cover 60 languages across diverse linguistic families, scripts, and resource conditions. The model is Chinese- and English-centric, supporting 234 translation directions in total (English ↔ 59; Chinese ↔ 58). To promote inclusivity, we support major regional and minority languages of China—such as Uyghur, Tibetan, Mongolian (traditional script), and Cantonese (Yue). A complete list is available in the Appendix Table 6.

Motivated by our preliminary analysis (as shown in Figure 1), which demonstrates Qwen3 (Yang et al., 2025) offers more comprehensive and consistent multilingual performance than several leading open-source LLMs (e.g., Llama 3.1 (Dubey et al., 2024), Gemma 2 (Rivière et al., 2024)), we selected Qwen3 as the backbone across sizes.

Our adaptation pipeline follows a two-stage framework—CPT followed by SFT, as illustrated in Figure 2. CPT operates on a large-scale mixture of monolingual and parallel corpora to integrate wide-coverage translation knowledge. SFT then fine-tunes the model on a carefully curated dataset of high-quality, human-translated parallel sentences formatted as translation instructions to further improve faithfulness and adequacy.

3.2 Data Curation

The foundation of an effective MMT model relies on a high-quality, large-scale, and diverse corpus. Acknowledging that constructing such a corpus is a major challenge, we designed a systematic, multi-stage strategy to curate our training data.

CPT Monolingual Data To ensure broad linguistic coverage and diversity, our monolingual CPT corpus aggregates several high-quality sources. For English and Chinese, we utilized the well-curated

SlimPajama (Soboleva et al., 2023) and Skywork (Wei et al., 2023) corpora, respectively. To cover the remaining languages, we incorporated culturally diverse texts from CulturaX (Nguyen et al., 2024), OpenDataLab (Yu et al., 2025), and Wikimedia, among others (complete sources in Appendix Table 5).

CPT Bilingual Data Building a high-quality, and balanced bilingual corpus is particularly challenging for Chinese↔X. We adopt a hybrid pipeline with three sources (see Figure 2). The foundation of our corpus is data curated from diverse OPUS [1] sub-corpora. To significantly scale this up, we employed pseudo-parallel synthesis using open-source models [2] in two ways: (1) direct synthesis, creating synthetic En/Zh → X pairs by translating monolingual corpora, and (2) pivoted synthesis via English, leveraging typically higher-quality En↔X and En↔Zh models to obtain Zh↔X data. Finally, all data from these three sources was unified and subjected to a rigorous quality control pipeline, using OpusFilter (Aulamo et al., 2020) for heuristic cleaning and 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. Appendix Figures 8 and 9 visualize the final statistics and score distributions.

SFT Data Recent studies emphasize that data quality and diversity are crucial for successful SFT (Zhou et al., 2023; Zhu et al., 2024a). To this end, our SFT dataset compiles a translation set from carefully curated human-translated data. We first establish a broad language coverage using the multi-way parallel datasets Flores-200 dev (Costa-jussà et al., 2022) and NTREX (Federmann et al., 2022). We then incorporate the SMol (Caswell et al., 2025), which provides complementary data for under-represented and low-resource language pairs often absent in other corpora. At last, we further include the WMT14–23 and IWSLT17–24 test sets to enhance style and domain diversity. The final SFT set contains approximately 596K pairs, strategically allocating 3K–20K high-quality examples per direction across 117 directions.

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

[2] Models were selected from strong baselines evaluated in this study, using the best-performing for each language pair.

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 3: 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.

3.3 Parallel Multilingual Prompting

An advantage of MMT models is cross-lingual transfer, where knowledge learned from high-resource languages can benefit low-resource ones. However, this mechanism typically emerges as an implicit and subtle byproduct of joint training on large, mixed-language corpora. Inspired by the powerful in-context learning (ICL) capabilities of LLMs, we propose **Parallel Multilingual Prompting (PMP)** to make this mechanism explicit: we condition both the source and translation on an auxiliary parallel sentence in a carefully chosen language—specifically English for Chinese-centric directions, and typologically similar high-resource languages for English-centric directions. This approach provides semantic and lexical anchors that guide the model toward higher-fidelity translations.

Formulation. Let S be a source sentence in L_S and T the target in L_T. A standard translation prompt (STP) optimizes

$$\max_{\theta} \log P_\theta(T \mid S;\, \tau_{L_S \to L_T}),$$

where $\tau_{L_S \to L_T}$ denotes the translation direction. PMP augments the context with an auxiliary sentence A in L_A:

$$\max_{\theta} \log P_\theta(T \mid S, A;\, \tau_{L_S \to L_A \to L_T}),$$

yielding a many-to-one mapping $S + [A] \to T$ (Figure 3 compares these two prompting strategies).

Selecting the auxiliary language L_A. We select L_A with two complementary policies: (1) Typological similarity for En↔X. For English-centric directions involving medium/low-resource X, we select L_A from the high-resource language that is typologically similar to X. (2) English pivot for Zh↔X. For Chinese-centric directions, we consistently set L_A = En to leverage strong English priors as a universal semantic anchor. This strategy draws from the classic pivot translation paradigm, where English serves as an intermediary between non-English pairs. However, unlike traditional two-step pipelines (e.g., X→En then En→Zh) which lose the original source information in the second step, PMP provides the English translation as an in-context hint alongside the original source, allowing the model to use the English as a semantic guide while preserving the original source's nuances. These two policies are applied only when a suitably similar high-resource candidate is available; otherwise, no auxiliary language is used, and the prompt defaults to use the standard translation format. The full list of participating languages and their auxiliary partners is detailed in the Table 7.

Training and inference. During SFT, PMP and STP are mixed with probability p, thereby balancing direct translation proficiency and cross-lingual transfer. The data required for PMP can be readily constructed from multi-way parallel data (e.g., Flores-200), allowing us to repurpose the same underlying data to create these contextually rich training examples. At inference time, our default setting does not require an auxiliary sentence; the model translates with STP-style prompts, which already yield gains. Optionally, PMP can be enabled by providing an auxiliary translation A, obtained either from a high-quality external MT system or via a self-generation step with our model; both variants can further improve robustness and adequacy.

4 The Pitfall of Directional Degeneration

4.1 An Asymmetric Performance Degeneration

We begin by building a baseline on Qwen3-4B-Base using standard SFT. Following common practice, each parallel pair is used in both directions (e.g., En→X and X→En). While this standard approach yielded expected improvements in the En/Zh→X directions, it led to a counterintuitive performance degradation in the reverse X→En/Zh directions. The outputs exhibit fluent but unfaithful content, often ignoring the source (see Table 3). We term this phenomenon **Directional Degeneration** and hypothesize a *Shallow Mapping Trap* induced by the multi-way data structure. In our 60-language setup, each English or Chinese sentence can be repeated up to 59 times as the target of different sources. This many-to-one redundancy incentivizes a shortcut: mapping diverse foreign inputs to a small, high-frequency set of target patterns, undermining source faithfulness and adequacy.

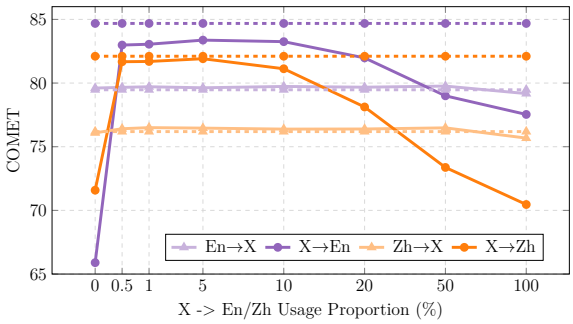


Figure 4: The impact of the Strategic Downsampling proportion (p). Dashed lines represent the use of separate, non-symmetric data for the X→En/Zh directions.

4.2 Mitigation via Strategic Downsampling

To probe the hypothesis, we first run an oracle experiment: we replace the reverse X→En/Zh portion of the SFT data with a separately sampled, non-overlapping subset from our bilingual CPT data, while keeping En/Zh→X unchanged. As shown by the dashed curves in Figure 4, X→En/Zh performance returns to a stable level relative to the fully symmetric baseline (i.e., p=100), suggesting that the collapse stems from symmetric reuse rather than the direction itself.

Although useful, the oracle requires an external pool, making it impractical in many settings. Therefore, we propose a simpler and more practical method: **Strategic Downsampling**. Let p be the retention probability for X→En/Zh direction samples. During SFT, we retain all En/Zh→X instances (100%), while including each X→En/Zh instance with probability p. Varying p, as plotted as solid lines in Figure 4, reveals that even a very small reverse retention (e.g., p= 0.5%) already averts degeneration, and performance peaks around p≈5% before degrading again as redundancy increases. This pattern suggests that only a weak alignment signal is needed for X→En/Zh, given the strong target-side priors in English/Chinese. Based on this observation, we adopt p= 5% for final models.

5 Results and Analyses

5.1 Setup

Training Details We continue pre-training using 90B tokens balanced at a 1:1:1 ratio among monolingual, Chinese-centric bilingual, and English-centric bilingual data. For CPT, 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 3), which we found performs better 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%). We train four sizes—0.6B, 1.7B, 4B, and 8B, and all models were trained on 16 NVIDIA H200 GPUs using the ms-swift framework (Zhao et al., 2024). The 4B and 8B models consume over 2,500 and 4,000 GPU-hours, respectively. Detailed hyperparameters are provided in Appendix 4.

Evaluation Data and Metrics We evaluate on FLORES-200 *devtest* (Costa-jussà et al., 2022). To address the lack of a benchmark for Chinese Mongolian (traditional script), we constructed a new test set by employing professional native speakers for annotation, following a rigorous annotation protocol [3]. We adopt COMET-22 (Rei et al., 2022a) as our primary evaluation metric, and report SacreBLEU (Post, 2018) in the Appendix for reference.

Models for Comparison We compare against two categories. (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-7B/13B (Alves et al., 2024), GemmaX2-28-2B/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-3.3B/54B (Costa-jussà et al., 2022). For fairness, evaluation is conducted on the intersection of language pairs supported by each baseline and LMT, covering Chinese-centric directions when applicable. For brevity, we present LMT-4B and LMT-8B in the main text; complete results for 0.6B/1.7B/4B/8B appear in the Appendix.

[3]We will release this dataset to fill a gap in MT benchmarks and support research on this underrepresented language.

5.2 Overall Performance

Table 2 summarizes the overall evaluation results. Overall, our proposed LMT models (4B and 8B) demonstrate highly competitive performance across most language pairs and translation directions, placing them among the top-tier of open-source MMT models. Notably, LMT-8B performs on par with, or only marginally behind, the current SOTA Seed-X-PPO-7B, despite tackling a substantially broader setting: it supports more than twice as many languages (60 vs. 27) while using less than half the CPT data (90B vs. 200B tokens). Furthermore, LMT-8B consistently outperforms other strong baselines such as TowerInstruct, Aya, LLaMAX3, and NLLB, often by a significant margin. The remarkable efficiency is also validated by our smaller LMT-4B model, which frequently surpasses larger models from other families.

A central objective of this work is to push beyond English-centric MMT. Our bi-centric design—explicitly centering both Chinese and English—proves effective across a wide spectrum of resource tiers. For example, in the X→Zh direction, LMT-8B outperforms the 7× larger NLLB-54B by 7.6 COMET points (87.67 vs. 80.06), underscoring the benefits of our data and modeling choices.

Overall, these results validate the effectiveness of LMT in large-scale MMT. With coverage of 60 languages and 234 directions, LMT maintains robust quality at scale, substantially uplifting the performance of LR (low-resource) languages and narrowing the gap with HR (high-resource) pairs. This suggests that non-English-centric MMT is both feasible and competitive with substantially larger or more narrowly scoped systems.

5.3 Ablation Study

To quantify the contributions of each component in our adaptation pipeline, we conduct an ablation study on our 4B model. Figure 5 presents the results of incrementally adding Strategic Downsampling (SD), Continued Pre-training (CPT), and Parallel Multilingual Prompting (PMP) to a naïve Base + SFT baseline. Each bar in the figure represents the COMET score for a specific translation direction, averaged across all 60 languages.

The Base + SFT model first reaffirms the directional degeneration identified in Section 4. The introduction of SD provides the most critical fix for

# Lang (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 (7/3/0)	TowerInstruct-7B	88.35	88.31	85.28	86.04	89.42	88.21	86.51	85.11	-	-	-	-
	TowerInstruct-13B	88.78	88.51	86.15	86.83	89.92	88.53	87.44	86.66	-	-	-	-
	LMT-4B (Ours)	88.79	88.48	86.63	87.86	89.77	88.75	87.68	87.67	-	-	-	-
	LMT-8B (Ours)	**89.05**	**88.64**	**86.87**	**87.92**	**90.20**	**88.82**	**88.06**	**87.87**	-	-	-	-
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-4B (Ours)	89.10	88.38	86.85	87.57	89.26	88.77	86.52	87.14	88.49	88.58	85.06	86.61
	LMT-8B (Ours)	**89.41**	**88.57**	**87.14**	**87.67**	**89.61**	**88.94**	**86.87**	**87.27**	**88.98**	**88.82**	**85.51**	**86.71**
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-4B (Ours)	89.10	88.38	86.85	87.57	90.75	89.21	88.43	87.64	90.63	88.97	88.02	87.23
	LMT-8B (Ours)	89.41	88.57	87.14	87.67	91.03	89.33	88.76	87.72	90.81	89.14	88.40	87.34
28 (13/8/7)	GemmaX2-28-2B	88.64	88.18	85.92	86.89	87.97	88.76	84.60	86.52	85.23	87.85	81.45	85.06
	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-4B (Ours)	89.10	88.38	86.85	87.57	88.54	88.93	85.60	87.30	86.35	87.90	82.85	85.84
	LMT-8B (Ours)	**89.41**	88.57	**87.14**	**87.67**	**88.86**	89.12	**85.91**	87.50	**86.68**	88.32	**83.36**	86.14
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-4B (Ours)	89.10	88.38	86.85	87.57	88.75	88.80	85.94	87.24	87.01	86.71	83.96	85.10
	LMT-8B (Ours)	89.41	**88.57**	**87.14**	**87.67**	89.06	**88.98**	86.26	**87.42**	**87.34**	**87.17**	**84.31**	**85.54**
40 (13/16/11)	X-ALMA-13B	**89.41**	88.51	-	-	**90.37**	**89.29**	-	-	87.15	87.98	-	-
	LMT-4B (Ours)	89.10	88.38	-	-	89.95	89.06	-	-	86.94	88.12	-	-
	LMT-8B (Ours)	**89.41**	**88.57**	-	-	90.24	89.21	-	-	**87.27**	**88.39**	-	-
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-4B (Ours)	89.10	88.38	86.85	87.57	89.86	89.02	87.04	87.32	86.26	87.31	82.68	85.19
	LMT-8B (Ours)	**89.41**	**88.57**	**87.14**	**87.67**	**90.13**	**89.18**	**87.42**	**87.45**	**86.58**	**87.72**	**83.10**	**85.59**
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-4B (Ours)	89.10	88.38	86.85	87.57	89.86	89.02	87.04	87.32	86.43	87.26	82.95	85.25
	LMT-8B (Ours)	**89.41**	**88.57**	**87.14**	**87.67**	**90.13**	**89.18**	**87.42**	**87.45**	**86.75**	**87.65**	**83.37**	**85.61**
59 (13/18/28)	NLLB-3.3B	87.63	87.36	84.65	76.48	88.32	88.02	84.09	77.78	84.23	85.45	80.60	77.36
	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-4B (Ours)	89.10	88.38	86.85	87.57	89.86	89.02	87.04	87.32	86.92	86.77	83.81	85.06
	LMT-8B (Ours)	**89.41**	**88.57**	**87.14**	**87.67**	**90.13**	**89.18**	**87.42**	**87.45**	**87.23**	**87.20**	**84.20**	**85.48**

Table 2: COMET-22 scores of our LMT models compared with a range of general-purpose multilingual LLMs and dedicated MMT models. Languages are grouped into high-, medium-, and low-resource categories. Evaluation is conducted only on the intersection of language pairs supported by each baseline and our models. The first column (# Lang) 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.

this issue, yielding a massive +11.45 point gain in the X → Zh direction and +5.83 in the X → En direction, validating its role as an essential corrective measure. Next, adding the CPT stage provides a substantial, foundational uplift across all directions, with gains ranging from +3.80 to +8.23 points, with detailed analysis in Appendix A.2. This highlights the crucial role of CPT in strengthening the model's core multilingual capabilities. Finally, the inclusion of PMP adds a consistent final enhancement across all directions, confirming that it's effective for boosting cross-lingual transfer even on top of an already strong CPT-enhanced model.

In summary, this component-wise analysis demonstrates that our model's final performance is attributable to the synergistic effect of its core components: a robust CPT foundation, a critical SFT mitigation for degeneration, and an advanced prompting strategy for transfer. We conclude that their combination is vital for successfully adapting foundation LLMs for high-coverage, robust MMT.

5.4 Analysis of the PMP

Analysis of Inference-Time Strategies To analyze the flexibility of PMP at inference time, we compare four distinct strategies. The evaluation is conducted on the subset of languages that participated in PMP training, with the average results shown in Figure 6 (left). The strategies are: (1) Two-Step Pivot (PT), a traditional pipeline approach; (2) Direct Translation (DT), a standard one-to-one baseline; (3) PMP-O (Oracle), using ground-truth auxiliary translations; and (4) PMP-S (Self-Generated), using model-generated auxil-

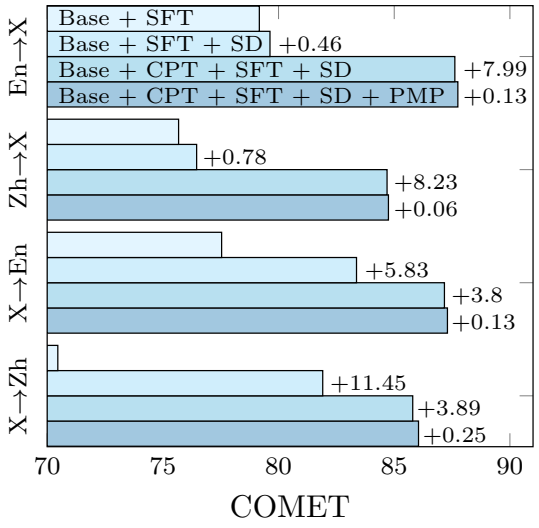


Figure 5: Ablation study on the impact of each component: Strategic Downsampling (SD), Continued Pre-training (CPT), and Parallel Multilingual Prompting (PMP). The annotated values quantify the performance gain of each component over the preceding one.

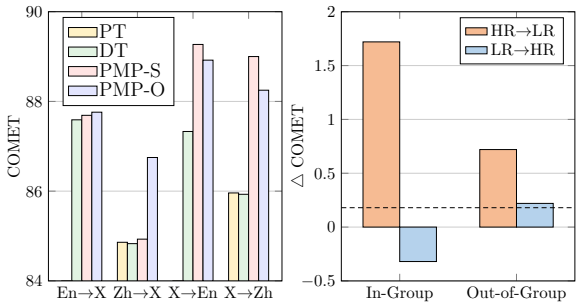


Figure 6: Analysis of the Parallel Multilingual Prompting (PMP). **Left:** Comparison of different inference-time strategies. Both oracle (PMP-O) and self-generated (PMP-S) variants significantly outperform the Direct Translation (DT) and traditional Two-Step Pivot (PT) baselines. **Right:** Zero-shot transfer gains from PMP training. The bars show the performance difference between our final model and an identical model trained without PMP. The dashed line indicates the average baseline gain for languages not involved in PMP.

iary translations. The results show that both PMP variants significantly outperform the DT and PT baselines, confirming the effectiveness of the PMP method. While PMP-O sets the performance upper bound in the En/Zh → X directions as expected, we observe a counter-intuitive result in the X → En/Zh directions, where PMP-S performs best. We hypothesize this is due to the data sparsity for this specific conditional task during training: being relatively undertrained, the model responds more effectively to a self-generated hint that matches its own generative biases than to a perfect but out-of-distribution human-written oracle.

Analysis of Zero-Shot Transfer Gains To quantify the cross-lingual transfer capabilities unlocked by PMP, we plot the performance difference between our final model and an identical model trained without PMP in Figure 6 (right). The comparison is conducted across three zero-shot tiers: (1) In-Group: Translation between the PMP-enhanced languages and their corresponding HR auxiliary languages. (2) Out-of-Group: Translation between the PMP-enhanced languages and an HR language that was not used as the auxiliary (Japanese). This tier tests for the generalization of the PMP enhancement. (3) Control Directions: This tier establishes the baseline zero-shot gain, which is the average performance gain across all mutual translation pairs within HR languages that were not involved in PMP (i.e., Nl, Pt, Ja).

The results reveal that PMP provides substantial gains for HR → LR translation, boosting the In-Group performance by a remarkable +1.8 COMET points. Conversely, a slight performance degradation is observed in the corresponding backward direction. This suggests that the cross-lingual transfer explicitly learned to enhance the HR → LR pathway does not symmetrically benefit the reverse zero-shot direction.

6 Conclusion

In this work, we contribute to overcoming the prevailing English-centric bias in MMT, aiming for both broad language coverage and consistently high translation quality. To this end, we present LMT, a large-scale MMT model centered on both Chinese and English, demonstrating its efficient performance across a wide range of languages, including low-resource and Chinese-centric directions. We also identify directional degeneration, a critical pitfall in SFT with multi-way data, and resolve it by introducing a strategic downsampling approach. Additionally, we propose Parallel Multilingual Prompting (PMP), a simple yet effective technique that leverages typological similarity to enhance cross-lingual transfer. Our results show that LMT achieves competitive performance, often outperforming larger systems, with particularly large gains in Chinese-centric translation. By releasing the LMT suite of models, we provide a strong baseline for building inclusive, scalable, and high-performance MMT systems.

Limitations

While our LMT models demonstrate strong and robust performance, we acknowledge several limitations that primarily point to exciting avenues for future research. First, our evaluation, while comprehensive, is primarily conducted on academic benchmarks like FLORES-200 with COMET-22. Future work could extend this evaluation to a wider range of real-world scenarios and a broader set of proprietary models to further assess generalization and capture more nuanced aspects of translation quality, such as cultural appropriateness. Second, our bi-centric design, which already moves "Beyond English," could be extended in future work to tri- or multi-centric configurations to better serve other major linguistic or geopolitical regions and further the goal of a truly inclusive MMT ecosystem.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL-HLT (1)*, pages 3874–3884. Association for Computational Linguistics.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

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 underrepresented 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.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2021. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5):99:1–99:38.

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.

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. 2025a. Beyond decoder-only: Large language models can be good encoders for machine translation. *Preprint*, arXiv:2503.06594.

Yingfeng Luo, Tong Zheng, Yongyu Mu, Bei Li, Qinghong Zhang, Yongqi Gao, Ziqiang Xu, Peinan Feng, Xiaoqian Liu, Tong Xiao, and JingBo Zhu. 2025b. Beyond decoder-only: Large language models can be good encoders for machine translation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 9399–9431. Association for Computational Linguistics.

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.

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.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation? *CoRR*, abs/2405.20512.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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. 2024a. 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, Fandong Meng, Jie Zhou, and Min Zhang. 2024b. Taste: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6144–6158. 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.

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. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

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.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

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. 2024a. 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.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with

large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2765–2781. 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 Appendix

A.1 Analysis of CPT Bilingual Data Quality Distributions

Figures 8 and 9 present the COMETKiwi score (Rei et al., 2022b) distributions for the English-centric (En → X) and Chinese-centric (Zh → 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 indicates that, on average, the raw parallel data available for these languages is of lower quality. This skew is particularly pronounced for the Chinese-centric low-resource pairs (Figure 9), highlighting the acute challenge of sourcing high-quality, non-English-centric data.

We hypothesize that this phenomenon stems from two potential factors. First, it may reflect the ground truth of data availability; high-quality parallel data for many low-resource languages, especially those paired with Chinese, is genuinely scarcer across web and public corpora. Second, this trend might also expose a potential bias in the quality estimation (QE) model itself. Models like CometKiwi may exhibit performance disparities across different language pairs, potentially due to being trained on less data for certain languages. Their scoring reliability might decrease for pairs that do not involve English or for languages that are typologically distant from their training data. This could lead to systematically lower scores for some language pairs, irrespective of their true quality.

These observations underscore the immense challenge of building truly large-scale, high-quality multilingual datasets. Furthermore, they highlight a critical and ongoing need for the development of more robust and equitable quality estimation models that can perform reliably across a truly diverse range of languages, especially in non-English-centric contexts.

A.2 Analysis of CPT's Impact

This section details the language-specific impact of our Continued Pre-training (CPT) stage, expanding on the aggregated ablation study results in Section 5.3. Figure 7 shows a side-by-side comparison of the COMET scores for two models—Base+SFT+SD (without CPT) and Base+CPT+SFT+SD (with CPT)—across all four main translation directions for each individual language.

Across almost all languages and directions, the inclusion of CPT leads to a positive performance gain, demonstrating the fundamental importance of this adaptation step. However, the magnitude of this gain is not uniform and reveals two interesting trends. First, we observe that the performance uplift from CPT is generally more pronounced for medium- and low-resource languages. This is expected, as the CPT stage is most effective at building a foundational linguistic understanding for languages where the base model's pre-existing knowledge is weakest. Second, the benefits appear particularly significant in Chinese-centric directions (both Zh → X and X → Zh). This indicates that our CPT data curation strategy, which included a large volume of high-quality Chinese-centric parallel data, was successful in strengthening the model's non-English capabilities and making tangible progress towards our "Beyond English" goal.

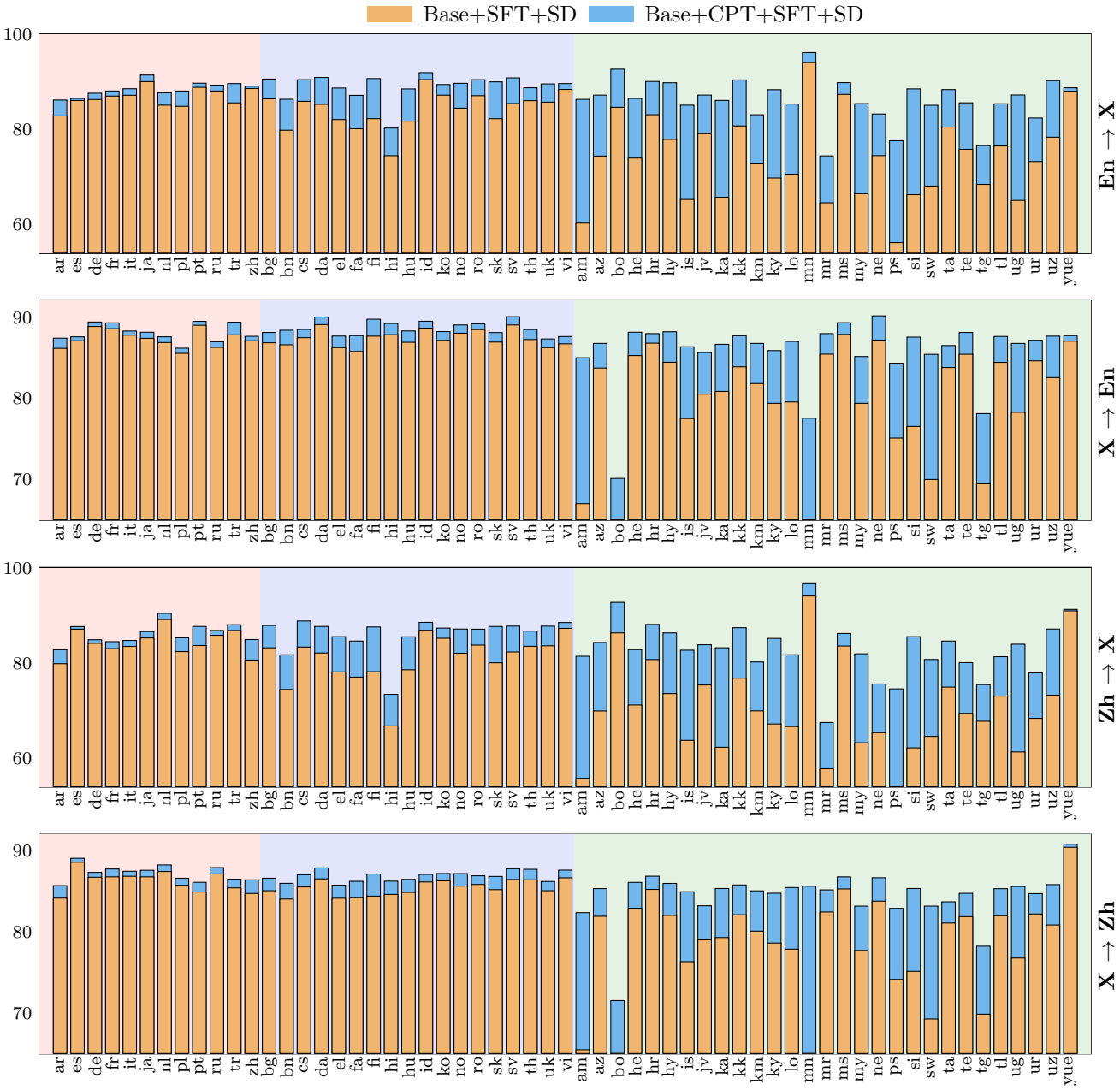


Figure 7: A fine-grained, per-language analysis of the impact of Continued Pre-training (CPT). Languages are grouped and shaded by resource level: high (red), medium (blue), and low (green). Each bar represents the COMET score for a specific language pair. The orange portion of the bar shows the performance of the model without CPT (‘Base+SFT+SD‘), while the total height of the bar represents the final performance after including the CPT stage (‘Base+CPT+SFT+SD‘). The blue portion visually represents the performance gain (ΔCOMET) contributed by CPT.

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 3: Translation error cases: hallucinations and fabricated outputs

Hyperparameter	CPT Stage	Stage2
Learning Rate	2e-5	2e-5
Adam β	(0.9, 0.999)	(0.9, 0.999)
LR Scheduler	cosine	cosine
Number of Epochs	1	1
Global Batch Size	1536	1024
Max Length	2048	1024
Train Steps	40,000	500
Warmup Ratio	0.05	0.01
Weight Decay	0.01	0.01

Table 4: Hyperparameter configuration during three-stage training.

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, no, 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
ug, bo, mn_cn	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 5: 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

Table 6: Detailed information of 60 languages.

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
no	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
mn_cn	Chinese Mongolian	Mongolian	Mongolic	Low

ISO Code	Language	Script	Family	Resource
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

ISO Code	Language	Script	Family	Auxiliary Language	ISO Code
bg	Bulgarian	Cyrillic	Indo-European	Russian	ru
da	Danish	Latin	Indo-European	German	de
fa	Persian	Arabic	Indo-European	Arabic	ar
no	Norwegian	Latin	Indo-European	German	de
ro	Romanian	Latin	Indo-European	Italian	it
sk	Slovak	Latin	Indo-European	Czech	cs
sv	Swedish	Latin	Indo-European	German	de
uk	Ukrainian	Cyrillic	Indo-European	Russian	ru
vi	Vietnamese	Latin	Austroasiatic	French	fr
az	Azerbaijani	Latin	Turkic	Turkish	tr
hr	Croatian	Latin	Indo-European	Polish	pl
is	Icelandic	Latin	Indo-European	German	de
kk	Kazakh	Cyrillic	Turkic	Russian	ru
ky	Kirghiz	Cyrillic	Turkic	Russian	ru
ps	Pashto	Arabic	Indo-European	Arabic	ar
tg	Tajik	Cyrillic	Indo-European	Russian	ru
tl	Tagalog	Latin	Austronesian	Spanish	es
ur	Urdu	Arabic	Indo-European	Persian	fa
uz	Uzbek	Latin	Turkic	Turkish	tr

Table 7: The set of languages that utilize the Parallel Multilingual Prompting (PMP) method and their corresponding auxiliary languages. The auxiliary language for each target was strategically selected from a pool of 11 high-resource languages (excluding English and Chinese), guided by two primary principles: shared **language family** and, where possible, shared **script**. This table includes all languages for which a suitable auxiliary partner was identified based on these criteria. For the remaining languages where no such partner was found, PMP was not applied, and they default to the standard translation format during training.

Direction	COMET				BLEU			
	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B
13 High-resource Languages								
en→ar	82.97	85.82	87.25	87.66	17.64	21.73	25.83	27.45
en→es	84.77	86.24	87.06	87.22	24.82	27.22	29.49	29.95
en→de	83.58	87.04	88.01	88.53	28.92	35.09	38.47	39.87
en→fr	85.70	87.68	88.66	88.86	40.79	46.15	48.87	50.41
en→it	86.17	88.13	88.87	89.16	25.70	28.78	31.03	32.13
en→ja	90.09	91.32	91.69	92.01	30.69	34.79	37.40	38.62
en→nl	84.75	87.20	88.43	88.72	22.32	25.49	26.86	27.99
en→pl	84.48	88.08	89.72	90.11	15.58	19.45	21.60	22.17
en→pt	88.03	89.53	90.02	90.28	42.78	46.76	49.34	50.96
en→ru	86.49	88.80	89.80	90.14	24.92	29.30	32.33	32.79
en→tr	86.72	89.01	90.18	90.76	20.36	23.63	26.98	28.44
en→zh	87.92	88.83	89.49	89.47	44.12	47.16	49.26	49.52
18 Medium-resource Languages								
en→bg	87.90	90.27	91.28	91.36	31.84	36.54	39.53	40.70
en→bn	83.20	85.70	87.12	87.21	13.47	14.98	17.65	18.48
en→cs	87.19	90.09	91.29	91.79	25.48	28.77	30.87	32.82
en→da	88.27	90.46	91.47	91.55	38.27	42.44	46.01	46.75
en→el	84.59	88.05	89.63	89.99	19.39	23.29	25.90	27.03
en→fa	82.58	86.29	88.04	88.46	19.56	22.82	26.05	27.25
en→fi	85.85	90.11	92.42	92.84	15.25	19.23	23.96	25.28
en→hi	76.81	79.65	81.22	81.69	24.11	28.47	31.88	33.25
en→hu	83.81	88.28	89.95	90.29	17.04	21.66	24.10	25.61
en→id	90.62	91.59	92.09	92.23	43.23	46.59	48.54	49.05
en→ko	87.55	88.97	89.77	90.20	26.61	29.62	30.72	32.37
en→no	87.89	89.59	90.45	90.58	28.57	31.34	32.72	33.51
en→ro	86.68	89.76	91.08	91.44	32.33	36.33	39.96	40.89
en→sk	85.90	89.45	91.09	91.40	25.60	29.84	32.43	33.61
en→sv	87.40	90.18	91.29	91.45	36.37	40.77	44.20	45.26
en→th	85.83	87.69	88.96	89.29	11.56	12.92	14.56	15.18
en→uk	85.08	89.01	90.39	90.60	20.86	25.50	30.08	31.32
en→vi	88.25	89.25	89.86	90.05	39.10	41.66	43.69	44.13
29 Low-resource Languages								
en→am	79.76	84.63	87.58	87.82	5.38	8.49	12.03	12.51
en→az	84.31	86.66	88.18	88.72	11.80	13.14	14.53	15.27
en→bo	87.37	90.98	92.22	92.57	1.71	3.33	3.67	3.07
en→he	81.79	86.40	88.49	88.98	17.16	22.95	27.31	29.94
en→hr	86.02	89.55	91.16	91.52	22.48	25.81	30.60	31.09
en→hy	86.42	89.23	90.68	90.89	17.23	20.62	23.44	23.52
en→is	81.43	85.26	86.60	87.14	18.77	22.70	25.48	26.63
en→jv	85.89	87.10	87.84	87.90	23.92	26.94	29.86	30.18
en→ka	81.89	85.78	87.96	88.79	10.89	12.82	15.41	15.84
en→kk	88.46	90.21	91.02	91.25	18.53	21.48	24.07	25.21
en→km	79.32	82.44	83.56	83.82	6.52	7.49	7.86	7.85
en→ky	85.84	87.89	89.26	89.48	11.35	13.09	15.18	15.85
en→lo	80.52	83.30	85.70	85.89	11.57	13.17	14.62	14.80
en→mn_cn	95.25	96.07	96.39	96.31	15.12	19.94	24.63	26.66
en→mr	70.80	74.18	75.91	76.18	11.20	14.22	16.45	17.48
en→ms	88.98	89.84	90.11	90.11	39.25	41.44	43.06	43.48
en→my	77.47	82.95	87.26	88.28	2.72	3.23	4.33	4.51
en→ne	81.45	83.11	84.35	84.70	15.79	18.50	20.79	22.13
en→ps	75.67	77.51	78.65	78.69	9.36	11.26	12.48	12.92
en→si	82.90	87.40	90.24	90.47	10.33	13.66	17.29	17.96
en→sw	82.88	85.07	85.97	86.22	30.52	34.90	38.10	38.84
en→ta	85.42	88.19	89.65	89.96	10.79	13.42	15.19	15.87
en→te	83.10	85.03	86.56	87.03	15.33	18.10	20.85	22.24
en→tg	75.64	76.71	77.54	77.77	15.19	18.98	21.63	22.40
en→tl	83.16	85.07	85.81	86.11	32.19	35.39	38.08	39.12
en→ug	84.05	86.71	88.36	88.52	12.13	15.01	17.84	19.73
en→ur	78.67	81.66	83.51	83.57	17.40	19.70	23.27	23.11
en→uz	88.52	90.01	90.91	91.06	15.18	18.04	21.85	22.23
en→yue	85.88	87.89	88.70	88.99	5.02	8.30	8.01	9.31

Table 8: COMET-22 and SacreBLEU scores of compared models on the FLORES-200 devtest set (En → X).

Direction	COMET				BLEU			
	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B
13 High-resource Languages								
ar→en	84.53	87.09	87.88	88.24	33.56	39.41	42.74	44.28
es→en	86.38	87.28	87.67	87.86	29.17	31.53	33.45	34.83
de→en	88.14	89.15	89.54	89.67	39.99	43.97	45.65	46.24
fr→en	88.43	89.21	89.55	89.58	41.93	44.89	46.71	47.54
it→en	87.09	88.21	88.47	88.57	31.12	34.98	36.56	37.15
ja→en	86.85	87.94	88.44	88.74	25.51	28.71	30.45	32.32
nl→en	86.27	87.37	87.82	87.99	29.92	32.54	34.41	34.98
pl→en	84.47	86.02	86.59	86.74	27.50	30.76	32.69	33.42
pt→en	88.54	89.46	89.81	90.01	45.49	48.69	50.74	51.32
ru→en	85.50	86.72	87.18	87.34	31.91	35.88	38.13	39.33
tr→en	87.12	89.12	89.83	90.00	32.26	37.46	40.32	41.03
zh→en	86.73	87.48	87.83	88.06	28.62	31.85	32.93	34.41
18 Medium-resource Languages								
bg→en	86.72	87.91	88.32	88.46	36.89	40.89	42.79	43.60
bn→en	85.21	87.66	88.86	89.20	25.12	31.79	35.49	37.28
cs→en	86.78	88.28	88.78	89.00	35.38	39.65	41.97	43.20
da→en	88.86	89.98	90.31	90.49	44.61	48.13	49.77	50.23
el→en	85.09	87.33	88.07	88.18	30.94	35.92	38.10	39.49
fa→en	85.32	87.60	88.35	88.62	30.48	35.53	38.58	39.42
fi→en	86.46	89.15	90.17	90.26	27.84	33.22	36.61	37.00
hi→en	86.97	89.11	89.76	90.10	31.72	38.97	41.84	42.82
hu→en	86.30	88.07	88.64	88.83	30.87	35.53	37.45	38.42
id→en	88.61	89.54	89.87	89.99	41.19	45.10	47.70	47.78
ko→en	86.62	88.05	88.75	88.82	25.94	30.51	32.70	33.77
no→en	87.57	88.87	89.37	89.40	41.13	44.04	46.20	46.50
ro→en	88.00	89.27	89.50	89.72	39.91	44.09	45.12	46.18
sk→en	86.32	87.92	88.57	88.71	34.59	39.26	41.63	42.70
sv→en	88.52	89.79	90.28	90.36	44.68	47.94	50.07	50.36
th→en	86.78	88.36	88.93	89.01	28.22	33.16	35.24	36.30
uk→en	85.55	87.17	87.71	87.81	35.01	39.22	41.92	42.81
vi→en	86.70	87.56	88.16	88.24	35.14	37.78	39.48	39.94
29 Low-resource Languages								
am→en	77.78	83.26	86.32	87.28	15.76	26.09	32.75	35.65
az→en	84.26	86.19	87.26	87.46	19.75	24.30	27.03	28.12
bo→en	64.15	68.66	71.85	73.46	6.20	9.83	15.80	18.10
he→en	84.98	87.47	88.58	88.82	35.83	41.76	45.37	46.42
hr→en	85.83	87.81	88.32	88.47	33.44	38.13	39.79	40.56
hy→en	85.31	87.53	88.69	88.94	30.29	34.75	39.57	40.87
is→en	82.28	85.55	86.79	87.27	28.44	34.14	37.34	38.62
jv→en	81.57	84.60	86.11	86.50	32.90	39.00	42.86	43.93
ka→en	83.54	86.20	87.20	87.48	22.45	26.55	30.55	31.13
kk→en	84.96	87.31	88.22	88.47	27.53	33.13	35.96	36.79
km→en	83.37	86.05	87.38	87.88	23.47	29.89	34.15	35.88
ky→en	83.17	85.48	86.33	86.53	20.17	23.86	26.32	27.35
lo→en	83.32	86.17	87.78	88.31	25.63	33.65	38.19	39.42
mn_cn→en	69.36	75.56	78.93	80.52	10.00	28.38	42.65	48.75
mr→en	84.96	87.45	88.61	88.80	27.81	33.00	37.62	38.77
ms→en	88.02	89.03	89.61	89.81	41.94	45.18	48.28	48.87
my→en	80.74	84.26	86.10	86.78	17.59	23.94	26.89	29.12
ne→en	87.81	89.75	90.74	90.96	30.23	37.53	41.02	42.56
ps→en	80.52	83.54	85.19	85.51	24.03	29.64	33.08	34.26
si→en	81.72	85.02	88.53	89.04	19.28	26.38	34.46	35.68
sw→en	81.64	84.56	86.13	86.57	34.29	40.53	44.50	46.40
ta→en	82.22	85.74	87.19	87.70	22.40	29.12	32.96	34.03
te→en	84.09	87.42	88.69	89.13	26.63	35.78	39.06	40.60
tg→en	73.45	77.39	78.87	79.63	26.06	32.08	35.30	38.07
tl→en	84.73	87.13	88.12	88.50	38.80	44.51	48.78	50.46
ug→en	83.68	86.54	87.16	87.64	21.31	26.32	28.70	29.97
ur→en	83.64	86.56	87.74	88.14	26.55	32.75	36.33	38.16
uz→en	85.05	87.34	88.21	88.58	28.58	33.97	36.71	38.02
yue→en	86.53	87.63	87.94	88.08	28.84	32.44	33.66	35.22

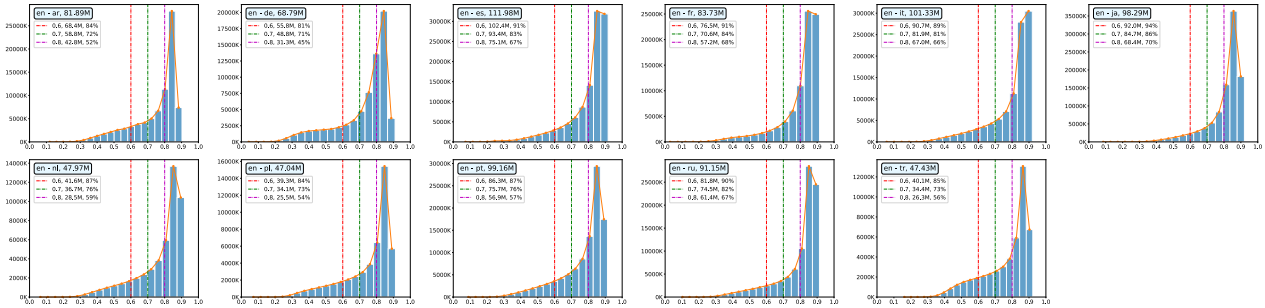
Table 9: COMET-22 and SacreBLEU scores of compared models on the FLORES-200 devtest set (X → En).

Direction	COMET				BLEU			
	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B
13 High-resource Languages								
zh→ar	79.71	82.24	84.04	84.48	8.71	11.17	13.72	14.58
zh→en	81.09	84.42	85.85	86.32	10.47	13.11	15.53	16.31
zh→es	86.73	87.48	87.83	88.06	28.62	31.85	32.93	34.41
zh→de	82.67	84.58	85.38	85.69	16.32	18.94	20.45	20.95
zh→fr	80.41	84.05	85.21	85.65	14.97	18.60	21.78	22.56
zh→it	82.06	84.24	85.28	85.49	21.16	24.93	28.04	29.03
zh→ja	83.75	86.07	87.35	87.46	15.70	18.55	21.03	22.02
zh→nl	88.87	89.74	90.65	90.86	24.08	27.34	29.80	30.90
zh→pl	81.67	84.92	86.01	86.36	13.11	16.14	18.23	19.06
zh→pt	83.08	86.92	88.64	89.04	9.86	13.19	15.37	15.49
zh→ru	84.89	86.51	87.25	87.39	21.17	23.95	26.49	27.40
zh→tr	84.76	87.46	88.71	88.90	14.21	17.35	19.26	20.09
18 Medium-resource Languages								
zh→bg	84.63	87.43	88.80	89.26	17.14	20.32	22.57	23.77
zh→bn	78.06	80.86	82.55	82.92	6.72	7.95	9.75	10.39
zh→cs	84.72	88.06	89.95	90.15	13.30	16.73	19.33	20.07
zh→da	84.95	87.15	88.38	88.74	18.89	21.69	24.16	25.16
zh→el	81.00	85.05	86.80	87.41	10.69	13.17	15.55	16.68
zh→fa	79.75	83.34	85.40	85.75	11.50	14.19	16.19	17.05
zh→fi	81.87	86.67	89.35	89.79	8.83	11.99	14.82	15.04
zh→hi	69.07	72.53	74.73	75.09	12.42	15.78	18.42	19.36
zh→hu	79.62	84.49	87.04	87.26	9.78	12.78	16.50	16.67
zh→id	87.07	88.33	88.87	89.06	23.24	25.83	27.43	28.77
zh→ko	85.14	86.94	87.68	88.06	18.88	21.79	23.41	24.32
zh→no	84.44	86.80	87.77	88.20	13.90	16.38	17.78	18.62
zh→ro	83.24	86.45	87.88	88.26	17.83	20.68	23.63	24.43
zh→sk	83.23	87.22	88.91	89.57	13.10	15.82	18.83	19.63
zh→sv	84.32	86.94	88.41	88.75	17.00	20.38	22.96	24.22
zh→th	83.82	85.60	86.88	87.21	8.20	8.85	10.30	10.80
zh→uk	82.86	86.96	88.62	89.08	10.96	14.79	17.09	18.30
zh→vi	86.97	88.16	88.77	89.00	27.63	30.13	32.04	33.11
29 Low-resource Languages								
zh→am	74.21	79.25	83.38	83.95	2.61	4.16	5.92	6.65
zh→az	80.11	83.46	85.39	85.88	7.95	9.73	10.63	11.23
zh→bo	88.72	91.10	92.72	92.68	1.61	2.11	2.28	2.10
zh→he	78.05	82.65	85.06	85.51	7.97	10.99	14.44	15.69
zh→hr	83.28	87.64	89.51	89.97	12.13	15.60	18.44	19.62
zh→hy	82.24	85.60	87.44	87.98	9.16	11.30	13.41	13.84
zh→is	78.16	82.15	84.28	84.55	10.50	13.07	15.15	16.23
zh→jv	81.64	83.52	84.37	84.87	11.06	13.75	16.17	17.01
zh→ka	77.88	82.77	85.39	86.03	6.61	8.89	10.67	11.47
zh→kk	84.72	86.76	88.19	88.41	10.15	12.33	14.76	15.46
zh→km	75.69	79.08	81.23	81.70	5.01	5.87	6.92	6.98
zh→ky	82.14	85.01	86.61	87.01	7.62	9.08	11.04	11.13
zh→lo	76.58	79.53	82.21	83.01	7.26	8.66	9.24	9.82
zh→mn_cn	96.15	96.60	96.97	96.96	25.34	33.57	41.17	42.45
zh→mr	63.03	66.84	69.02	69.62	6.15	7.95	9.77	9.70
zh→ms	85.16	86.17	86.53	86.83	20.51	22.26	23.46	24.37
zh→my	71.97	78.79	84.08	84.41	1.60	2.26	2.59	2.83
zh→ne	73.18	75.34	77.16	77.75	7.17	8.56	10.13	10.94
zh→ps	70.92	74.26	76.02	76.26	5.10	6.70	8.16	8.25
zh→si	78.49	83.54	87.06	87.33	5.71	7.72	10.64	11.01
zh→sw	78.23	80.78	82.13	82.37	13.87	16.73	19.81	20.39
zh→ta	80.41	83.74	85.92	86.14	5.95	7.11	8.68	8.94
zh→te	75.90	79.55	81.72	82.07	7.20	8.81	10.72	11.31
zh→tg	73.88	74.64	76.24	76.36	8.47	10.85	13.81	14.07
zh→tl	78.69	81.24	81.93	82.41	16.05	20.07	21.65	22.83
zh→ug	80.12	83.41	84.79	85.13	9.20	11.31	13.48	14.50
zh→ur	73.01	76.87	78.90	79.62	9.16	11.56	14.15	14.82
zh→uz	84.90	86.67	88.02	88.16	8.46	10.51	13.14	13.37
zh→yue	90.19	90.93	91.34	91.53	5.52	6.73	6.72	6.34

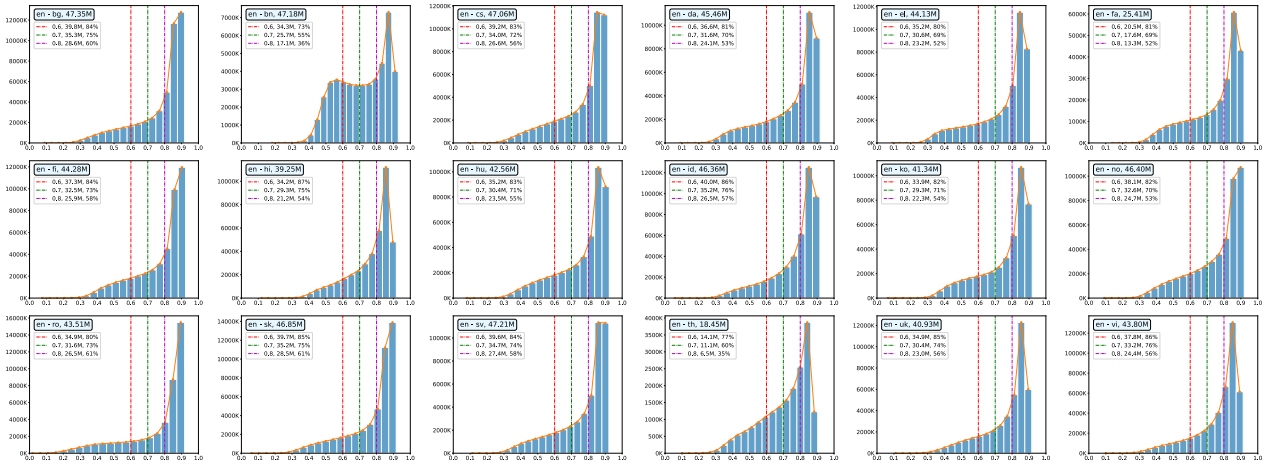
Table 10: COMET-22 and SacreBLEU scores of compared models on the FLORES-200 devtest set (Zh → X).

Direction	COMET				BLEU			
	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B	LMT-0.6B	LMT-1.7B	LMT-4B	LMT-8B
13 High-resource Languages								
ar→zh	82.33	84.95	86.04	86.26	31.79	36.86	39.59	40.39
en→zh	83.99	85.83	86.94	87.14	32.56	37.25	39.91	40.19
es→zh	87.92	88.83	89.49	89.47	44.12	47.16	49.26	49.52
de→zh	86.04	87.31	87.69	87.88	34.09	37.19	38.67	39.14
fr→zh	85.80	87.43	88.01	88.02	36.81	40.31	41.40	41.94
it→zh	86.55	87.44	87.85	87.88	37.54	40.66	41.94	41.80
ja→zh	86.18	87.48	87.84	87.85	34.67	38.30	39.37	39.87
nl→zh	86.80	88.17	88.48	88.71	31.92	35.59	37.27	38.22
pl→zh	84.78	86.34	86.93	87.00	32.49	35.67	37.62	37.73
pt→zh	83.95	85.88	86.49	86.61	31.89	35.63	37.21	37.55
ru→zh	86.46	87.93	88.30	88.38	37.45	40.65	42.05	42.44
tr→zh	84.87	86.26	86.80	86.84	35.11	38.43	39.82	40.49
18 Medium-resource Languages								
bg→zh	84.40	86.39	86.91	87.11	35.67	39.01	40.88	41.34
bn→zh	81.62	84.87	86.37	86.90	27.02	32.68	35.63	37.29
cs→zh	85.06	86.65	87.42	87.25	34.77	38.70	40.94	40.50
da→zh	85.99	87.68	88.32	88.36	37.27	40.90	42.88	42.10
el→zh	82.17	85.21	86.16	86.35	30.55	35.50	38.00	38.69
fa→zh	83.02	86.00	86.70	86.91	30.68	35.93	37.67	38.84
fi→zh	83.35	86.51	87.65	87.78	30.90	36.00	38.72	38.56
hi→zh	82.65	85.64	86.62	87.15	30.19	35.36	37.94	39.18
hu→zh	83.67	86.05	86.84	87.03	32.40	36.83	38.72	39.36
id→zh	85.98	87.13	87.58	87.71	36.96	40.49	41.50	42.22
ko→zh	85.44	87.14	87.67	87.87	32.16	35.77	38.02	38.05
no→zh	85.26	87.13	87.69	87.76	34.91	38.94	40.76	40.69
ro→zh	84.94	86.78	87.45	87.43	36.41	39.94	41.95	41.88
sk→zh	84.61	86.28	87.21	87.13	34.06	38.07	39.90	40.12
sv→zh	85.93	87.68	88.32	88.40	36.53	40.16	42.61	42.38
th→zh	85.72	87.44	88.10	88.22	32.22	36.11	38.47	38.98
uk→zh	83.86	86.03	86.71	86.83	34.54	38.81	40.58	41.03
vi→zh	86.53	87.51	87.96	87.95	35.73	38.24	39.73	40.17
29 Low-resource Languages								
am→zh	73.26	79.39	83.21	84.08	16.53	24.61	31.23	33.02
az→zh	82.36	84.59	85.80	85.90	27.51	31.25	33.64	34.15
bo→zh	64.72	69.65	72.44	74.47	7.84	13.73	19.40	22.06
he→zh	82.45	85.30	86.61	86.71	32.04	37.37	40.38	40.85
hr→zh	84.01	86.24	87.28	87.28	33.55	38.05	40.18	39.81
hy→zh	81.99	84.64	86.45	86.84	30.52	35.03	38.58	39.37
is→zh	80.51	84.08	85.35	85.81	28.44	33.66	36.75	37.73
jv→zh	79.12	82.67	83.85	84.29	29.27	34.26	36.92	37.69
ka→zh	80.88	84.65	86.03	86.36	25.37	32.70	35.97	35.97
kk→zh	82.87	85.19	86.54	86.75	30.63	34.70	37.79	38.23
km→zh	81.26	84.08	85.78	85.94	26.41	30.93	34.43	35.58
ky→zh	81.23	84.41	85.27	85.51	25.83	30.35	32.32	32.99
lo→zh	80.42	83.98	85.99	86.46	25.25	31.53	35.86	36.84
mn_cn→zh	77.98	84.13	87.36	88.33	24.01	50.27	67.48	72.57
mr→zh	80.23	84.43	85.51	86.11	27.00	33.42	36.18	36.93
ms→zh	84.87	86.54	87.18	87.40	36.14	39.72	41.15	41.77
my→zh	76.46	81.64	84.18	84.68	14.78	24.02	29.95	30.89
ne→zh	83.16	85.95	87.22	87.67	29.26	34.29	37.53	38.17
ps→zh	77.75	81.29	83.54	84.01	23.07	29.32	32.98	33.70
si→zh	78.07	82.15	85.98	86.82	19.46	26.80	34.21	35.81
sw→zh	78.09	82.24	84.15	84.31	27.18	33.95	36.99	37.77
ta→zh	78.56	82.95	84.49	85.03	22.72	29.91	33.18	34.49
te→zh	78.81	83.20	85.21	85.73	24.11	32.50	35.67	36.76
tg→zh	73.04	76.91	78.84	79.27	27.95	33.30	37.04	37.74
tl→zh	81.68	84.41	85.65	85.66	32.58	38.30	40.66	40.76
ug→zh	81.64	84.92	86.16	86.33	28.28	33.78	35.70	36.73
ur→zh	80.30	83.74	85.46	86.14	25.39	32.34	35.91	36.89
uz→zh	82.42	85.29	86.40	86.67	30.56	35.88	37.88	38.50
yue→zh	90.67	91.16	91.05	91.07	43.02	45.10	44.93	44.62

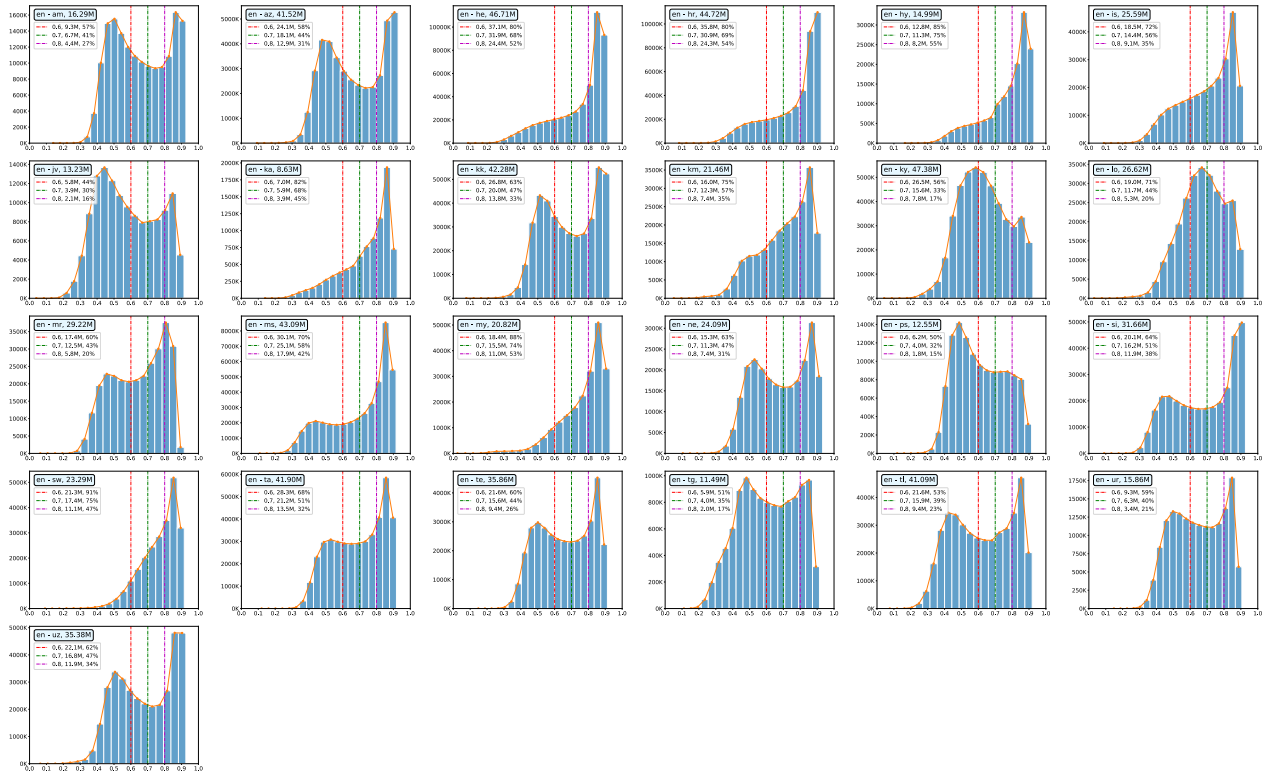
Table 11: COMET-22 and SacreBLEU scores of compared models on the FLORES-200 devtest set (X → Zh).



(a) High Resource Languages

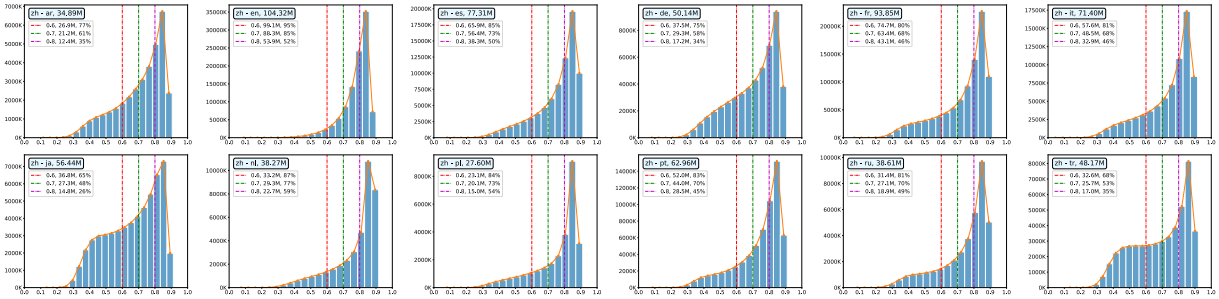


(b) Medium Resource Languages

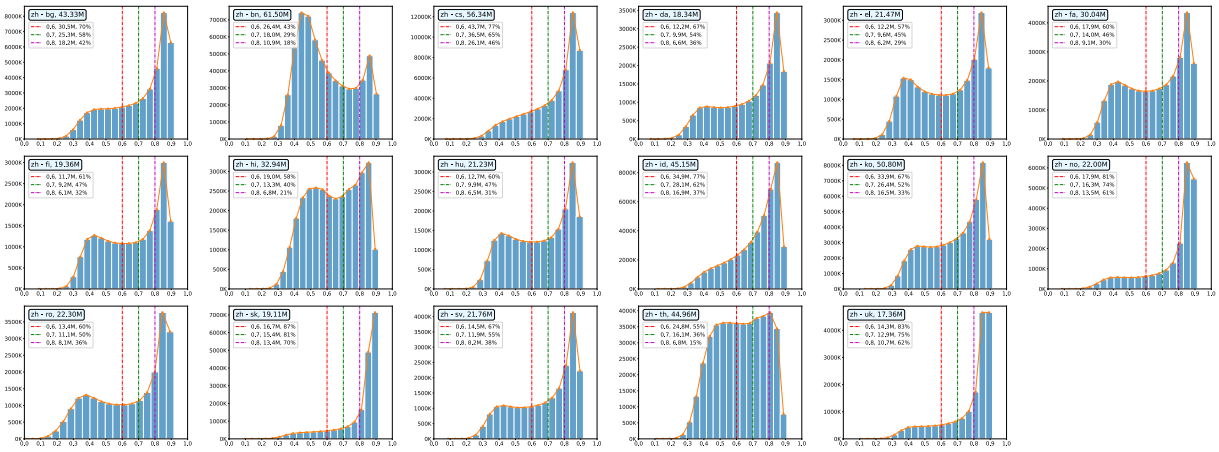


(c) Low Resource Languages

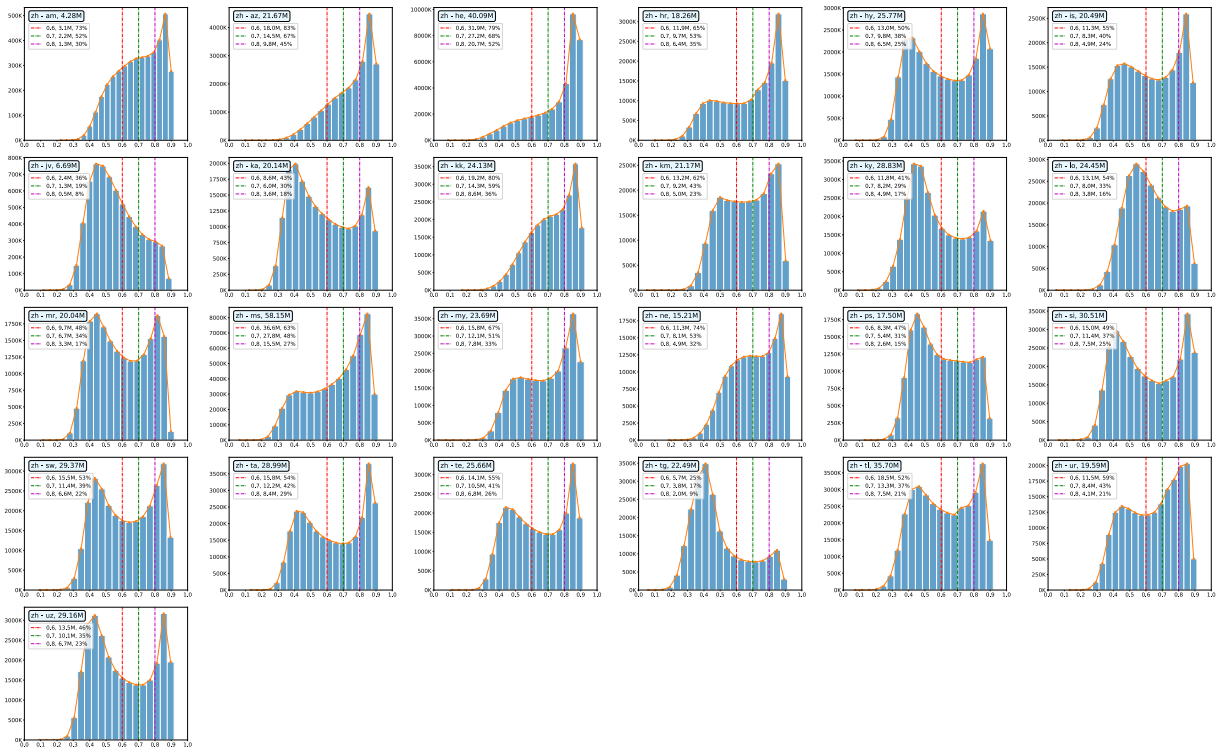
Figure 8: COMETKiwi score distributions for bilingual sentence pairs (En-X) are shown as histograms. Vertical lines indicate quality thresholds at 0.6 (red), 0.7 (green), and 0.8 (magenta), with the legend specifying the number and proportion of sentence pairs exceeding each threshold. Some language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



(a) High Resource Languages



(b) Medium Resource Languages



(c) Low Resource Languages

Figure 9: 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.