Asymmetric Conflict and Synergy in Post-training for LLM-based Multilingual Machine Translation

Anonymous ACL submission

Abstract

The emergence of Large Language Models (LLMs) has advanced the multilingual machine translation (MMT), yet the *Curse of Multilinguality* (CoM) remains a major challenge. Existing work in LLM-based MMT typically mitigates this issue via scaling up training and computation budget, which raises a critical question: *Is scaling up the training and computation budget truly necessary for high-quality MMT, or can a deeper understanding of CoM provide a more efficient solution?* To explore this problem, we analyze the linguistic conflicts and synergy, the underlying mechanism of CoM during post-training phase. We identify an *asymmetric phenomenon in linguistic conflicts and synergy*: the dominance of conflicts and synergy varies in different translation directions, leading to sub-optimal adaptation in existing post-training methods. We further find that a significant bottleneck in MMT appears to lie in post-training rather than multilingual pre-training, suggesting the need for more effective adaptation strategies. Building on these new insights, we propose a direction-aware training approach, combined with group-wise model merging, to address asymmetry in linguistic conflicts and synergy explicitly. Leveraging this strategy, our method fine-tunes X-ALMA-13B-Pretrain—trained only with multilingual pre-training—achieving comparable performance to XALMA-13B (only SFT) while using only 20B pretraining tokens and 17B parameters—5.5× fewer pretraining-tokens and 1.7x fewer model size—with just 0.85 COMET drop on Flores-200 testsets of 50 languages.

1 Introduction

Large language models (LLMs) have shown remarkable general capabilities (Brown et al., 2020; Wei et al., 2022; Dubey et al., 2024) and have advanced multilingual machine translation (Xu et al., 2024a; Yang et al., 2023; Alves et al., 2024). For example, Aya-101 (Aryabumi et al., 2024) expands support to 101 languages and achieves strong performance in multilingual machine translation, while LLaMAX (Lu et al., 2024b) further pushes performance beyond 100 languages. The common practice behind these successes is the large-scale pretraining, which typically involves monolingual pretraining [1], parallel pretraining, or both—followed by a small-scale, high-quality post-training phase. However, as LLMs scale to more languages, they suffer from the issue of *Curse of Multilinguality (CoM)* (Conneau, 2019), which degrades the translation performance.

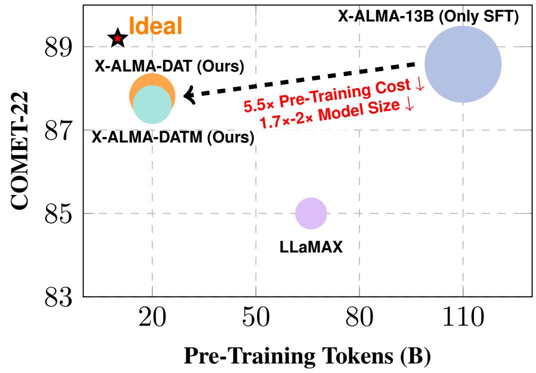


Figure 1: The relationship between pre-training cost, model capacity and translation performance. We evaluate performance on the Flores-200 test sets across 50 languages. **The size of circle denotes model capacity.**

Understanding and mitigating CoM is not new in the MMT literature. In traditional MMT, existing research has identified critical factors such as resource imbalances, limited model capacity, linguistic similarity, and complex interactions between language pairs, particularly for low-resource languages (Arivazhagan et al., 2019; Aharoni et al., 2019; Shaham et al., 2023; Meng and Monz, 2024), and proposed solutions including language-specific modules (Fan et al., 2021; Zhao et al., 2024; Xu

[1] We also refer to this as multilingual pretraining, where data from all languages are mixed during the pretraining process.

et al., 2023), vocabulary optimization (Han et al., 2024), data sampling techniques (Wang et al., 2020; Wang and Neubig, 2019; Lin et al., 2019), and continual learning approach (Liu et al., 2023). Based on these studies, recent LLM-based MMT research focuses on designing increasingly complex training pipelines and modular architectures. For instance, Xu et al. (2024b) proposed a five-stage training pipeline incorporating language-specific modules. However, existing analyses primarily focus on the encoder-decoder paradigm, while current LLM-based approaches heavily rely on scaling up model capacity and computational resources, making them prohibitively expensive. This raises a critical question: *Is scaling up the training and computation budget truly necessary for high-quality MMT, or can a deeper understanding of CoM in LLM-based MMT provides a more efficient solution?*

In this work, we systematically investigate linguistic conflicts and synergy during post-training phase. We conduct extensive experiments with different settings: across 5 to 50 languages, three pretrained LLMs - ALMA-7B-Pretrain, ALMA-13B-Pretrain and X-ALMA-13B-Pretrain, three distinct post-training strategies - multilingual training, group multilingual training, and separate training. We observe a consistent pattern: asymmetry in linguistic conflicts and synergy (Figure 2, Appendix B.1and B.2). For example, in multilingual training, XX→En translation directions experience significant linguistic conflicts, leading to performance degradation, whereas En→XX translations benefit from linguistic synergy, where XX denotes 49 different languages other than English. We further show this asymmetric phenomenon cannot be easily mitigated through existing training approaches, such as group multilingual training (Table 1). **This finding illustrates the need to develop a direction-aware training strategy for optimal post-training.**

Another key finding of our work is that a simple multilingual pre-training stage can be sufficient to equip foundation models with ideal multilingual capabilities, whereas the bottleneck lies in the post-training stage (dotted lines in Figure 2 (g-i)). Motivated by these findings, we propose a novel *Direction-Aware Training* (DAT) approach and build an efficient MMT starting from a relatively efficient base model, the X-ALMA-13B-Pretrain—utilizing only simple multilingual pre-training on 20 billion tokens. Our approach fully leverages the interactive characteristics of different language directions to reduce conflicts while maximizing synergy. We also present a scalable version of the approach, named DATM, which utilizes model merging to further enhance efficiency with only negligible performance degradation.

Through comprehensive evaluations on Flores-200 and WMT23 Benchmark, we demonstrate the effectiveness of our approach. Notably, as shown in Figure 1, compared to X-ALMA (Only SFT) (Xu et al., 2024b), our model X-ALMA-13B-DAT maintains comparable performance while having two advantages: 1) utilizing a simple and efficient training recipe - starting from base models with fewer pre-training tokens and employing a post-training stage. 2) parameter-efficient - we consume 1.7x fewer parameters compared to X-ALMA (Only SFT). These results demonstrate that simple pre-training combined with dedicated post-training can also achieve good multilingual performance.

2 Experimental Settings

In this section, we introduce the basic experimental settings used in Section 3 and Section 4.

2.1 Datasets

We use the high-quality parallel dataset curated by Xu et al. (2024b), covering fifty languages across low-, medium-, and high-resource categories. Following (Xu et al., 2024b), these languages are grouped into eight linguistic groups based on linguistic similarity and a balanced number of languages. Details are provided in Section A in Appendix. The dataset primarily consists of samples from the Flores-200 development set and NTREX (Barrault et al., 2019). For languages in both Flores-200 and WMT'15-22, corresponding test sets are incorporated, yielding an average of 4K examples per language. For evaluation, we use Flores-200 and WMT23 benchmarks to assess performance.

2.2 Models

We select three representative fully open multilingual LLMs for our study: ALMA-Pretrain (Xu et al., 2024a) (7B–13B parameters) and X-ALMA-Pretrain (Xu et al., 2024b) (13B parameters). The ALMA-Pretrain models were pre-trained on 12B or 20B tokens across six languages, while X-ALMA-Pretrain underwent continued pre-training on 20B tokens from 50 languages, both based on LLaMA-2. We exclude other state-of-the-art multilingual models for two key reasons: (1) their pre-trained checkpoints are unavailable, as in the case of Aya-series (Aryabumi et al., 2024) and BigTrans (Yang et al., 2023); or (2) they exhibit suboptimal multilingual performance in certain languages as shown in Xu et al. (2024b); Cui et al. (2025), such as LLaMA-3 (Dubey et al., 2024).

2.3 Training

Fine-tuning Strategies We employ three distinct training strategies for fine-tuning the models: Multilingual Training, Separate Training, and Group Multilingual Training.

- **Multilingual Training** (Tang et al., 2020): This is typically achieved by mixing data from all languages and using it to fine-tune the model. The resulting model is a single model that possesses shared representations across all languages.
- **Group Multilingual Training** (Xu et al., 2024b): We group the languages and then apply multilingual training within each group, resulting in multiple models, each for its respective languages.
- **Separate Training**: Separate tuning involves training a distinct model for each language without considering linguistic synergies or conflicts.

Training Configurations In this work, all models are trained with a learning rate of 2e-3 using an inverse square root scheduler, a weight decay of 0.01, and a warmup ratio of 0.01. The total batch size is set to 128. Fine-tuning is conducted for 1 epoch, with both `max_new_tokens` and `max_source_length` set to 512. Additionally, FP16 precision training is enabled to optimize performance and efficiency. All models are trained on 4 NVIDIA H100 with LoRA (Hu et al., 2022) as Xu et al. (2024a) has shown a negligible performance gap between LoRA tuning and full fine-tuning.

2.4 Evaluation

We set the number of beams to 5 and both `max_new_tokens` and `max_source_tokens` to 512. We evaluate performance mainly using COMET-22 (Rei et al., 2022) and SacreBLEU (Post, 2018).

3 The Phenomenon: *Asymmetry in Linguistic Conflicts and Synergy*

In this section, we investigate the phenomenon of *Asymmetry in Linguistic Conflicts and Synergy* in LLM-based MMT. We begin by illustrating the phenomenon (Section 3.1) and analyzing its distribution across two essential factors: language resources and groups (Section 3.2). Finally, we show how this phenomenon poses challenges to existing post-training strategies (Section 3.3).

3.1 Asymmetry in Linguistic Conflicts and Synergy

We investigate linguistic conflicts and synergy during the post-training phase. To explore this, we utilize three foundation models, as mentioned in Section 2.2, to perform multilingual training with training datasets that include a range of languages, from 5 to 50, and evaluate the average performance on corresponding languages.

To quantify linguistic conflicts and synergy, we compare multilingual training with separate training, where each language pair is trained independently, eliminating cross-lingual interactions.

- **Linguistic Conflicts**: If multilingual training underperforms compared to separate training (i.e., COMET drop), conflicts dominate over synergy.
- **Linguistic Synergy**: If multilingual training outperforms separate training, synergy dominates.
- **Intensity**: the magnitude of the performance gap measures the strength of conflicts/synergy.

Figure 2 displays the results. We can have the following observations:

- **Key Findings 1: Asymmetry in Linguistic Conflicts and Synergy.** As shown in Figures 2 (a), (d), and (g), the average performance decreases with an increase in the number of languages, a phenomenon known as the CoM (Conneau, 2019; Xu et al., 2024b). However, by decomposing the average performance across all language directions into XX→En and En→XX, we uncover an intriguing asymmetry in the distribution of linguistic conflicts and synergies, as illustrated in Figures 2 (b), (c), (e), (f), (h), and (i). Specifically, in the XX→En direction, linguistic conflicts are more dominant, as shown by multilingual training consistently underperforming separate training. Conversely, in the En→XX direction, linguistic synergy is significant, with multilingual tuning consistently outperforming separate training. Furthermore, comparing different models reveals that increasing model capacity (e.g., from 7B to 13B) or incorporating more languages in the pre-training corpus can mitigate conflicts. However, a significant gap remains between separate and multilingual tuning, indicating that simply increasing model capacity and the number of languages in the pre-training corpus cannot fully resolve the issue. We observe similar

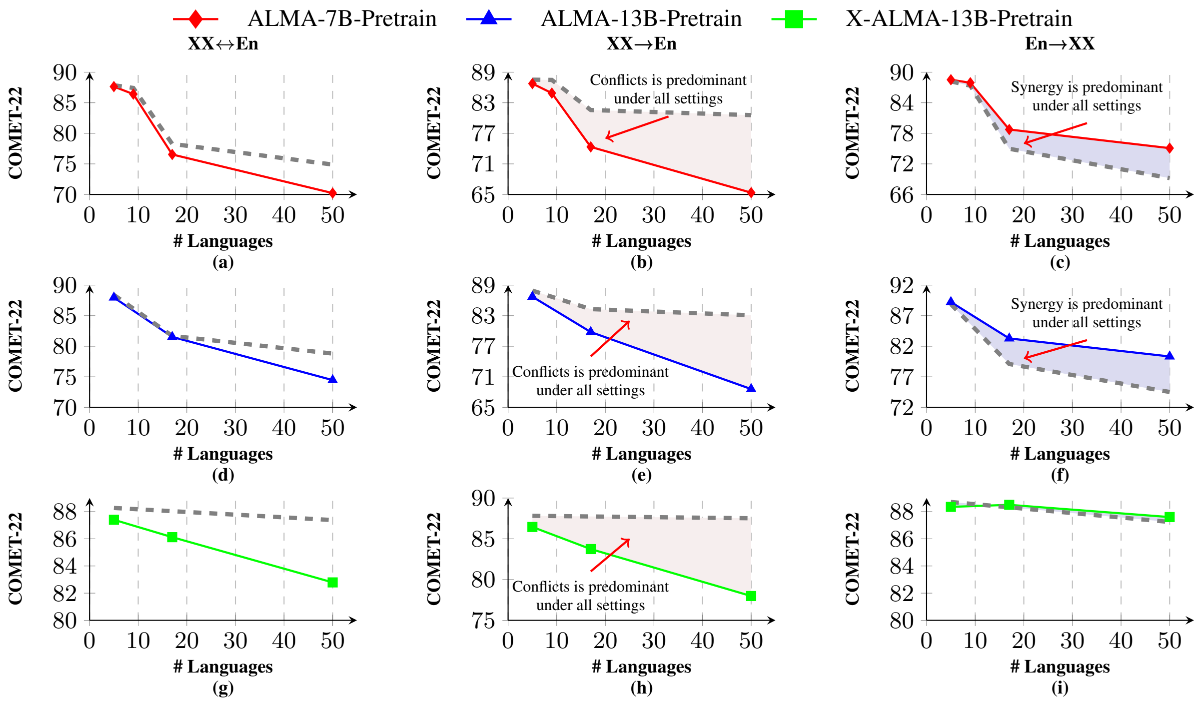


Figure 2: Performance of different models trained on varying numbers of languages. The dotted line represents the performance of separately trained models, serving as a reference point where no language conflicts or synergies occur. Two key findings emerge: (1) Asymmetry in Linguistic Conflicts and Synergy (Figure a–i), highlighting the uneven impact of multilingual training across language pairs; and (2) The Bottleneck of Multilinguality in Post-Training (Figure g–i): While multilingual pre-training provides a solid foundation for handling multiple languages, the multilingual training phase can lead to the CoM.

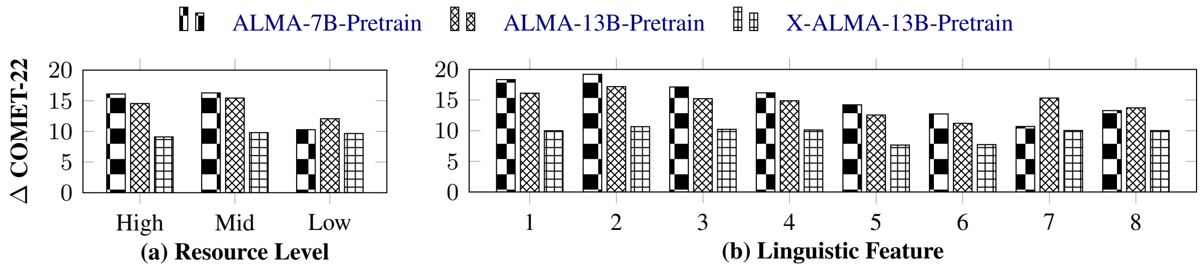


Figure 3: Δ COMET-22 between separate training and multilingual training in XX → En translation, grouped by resource level and linguistic features. The magnitude of Δ COMET-22 denotes the intensity of linguistic conflicts.

findings in terms of SacreBLEU (Appendix B.1) and across different settings (Appendix B.2).

A potential concern regarding this phenomenon is that it may stem from the limited LoRA rank, leading to linguistic conflicts and synergy issues. However, our results (see Appendix B.3) demonstrate that LoRA rank is not the root cause of this phenomenon. Instead, this issue may arise from an inherent limitation in the model's ability to encode source language representations effectively, potentially due to the absence of an encoder component. We leave this for future work.

- **Key Findings 2: Multilingual Pretraining Stage can sufficiently facilitate X-ALMA-13B-Pretrain with ideal multilingual capabilities, whereas the bottleneck may lie in the post-training stage.** Observing the dotted line in Figures 2 (g), (h), and (i), we find that separate training on the X-ALMA-13B-Pretrain model achieves ideal multilingual performance, maintaining average performance as the number of languages increases. However, multilingual training in the post-training stage cannot fully activate this multilingual ability, resulting in the CoM. For instance, Figure 2 (h) shows a significant performance gap between multilingual training and ideal performance, which widens as the number of languages increases. Interestingly, previous work (Xu et al., 2024b) designed complex training regimens with up to five stages, including

Model	Training Type	Group 1		Group 2		Group 3		Group 4	
		XX-En	En-XX	XX-En	En-XX	XX-En	En-XX	XX-En	En-XX
ALMA-13B Pretrain	Group Training	86.0/30.8	**87.7/32.0**	86.3/31.7	**87.7/33.2**	85.2/31.5	**88.0/26.8**	80.3/24.1	**81.3/26.0**
	Separate Training	**88.5/43.4**	86.6/29.9	**88.4/41.4**	87.1/31.7	**86.9/39.3**	86.9/24.7	**81.1/31.0**	74.7/23.7
	Multilingual	72.3/14.5	87.2/31.6	71.2/13.5	87.4/32.6	71.7/13.4	87.5/**26.8**	66.2/10.0	80.6/25.7
		Group 5		Group 6		Group 7		Group 8	
		XX-En	En-XX	XX-En	En-XX	XX-En	En-XX	XX-En	En-XX
	Group Training	82.7/25.8	**80.0**/16.4	82.7/21.9	**81.1**/18.5	77.4/**17.7**	68.7/9.7	73.6/12.6	69.9/6.9
	Separate Training	**83.3/29.4**	75.6/14.2	**83.5/24.8**	75.8/16.8	**77.6**/17.6	57.4/5.1	**76.5/18.5**	55.5/4.0
	Multilingual	70.7/13.1	79.4/ **16.7**	72.3/11.1	80.7/ **18.7**	62.3/4.0	**70.6/11.0**	62.8/5.9	**70.9/8.1**
		Group 1		Group 2		Group 3		Group 4	
		XX-En	En-XX	XX-En	En-XX	XX-En	En-XX	XX-En	En-XX
X-ALMA-13B Pretrain	Group Training	86.3/31.1	**88.8/34.7**	86.6/31.7	**88.6**/33.2	85.9/33.2	89.8/31.3	83.5/26.6	**86.3/29.3**
	Separate Training	**88.9/44.4**	88.5/34.1	**88.6/41.6**	88.3/34.9	**87.5/41.0**	89.7/30.5	**85.9/35.8**	85.3/27.8
	Multilingual	79.0/17.4	88.6/34.3	78.0/16.3	88.5/**35.4**	77.3/15.6	**89.9/31.4**	75.8/13.8	86.2/28.8
		Group 5		Group 6		Group 7		Group 8	
		XX-En	En-XX	XX-En	En-XX	XX-En	En-XX	XX-En	En-XX
	Group Training	85.6/28.2	**89.9/24.0**	86.2/24.6	**89.6/25.8**	86.4/27.0	**81.0/18.8**	83.0/20.6	**87.5/17.6**
	Separate Training	**87.4/35.3**	89.4/23.5	**87.7/30.1**	88.9/22.6	**87.8/33.1**	80.2/17.2	**86.5/31.1**	86.6/15.6
	Multilingual	79.7/17.8	89.7/23.6	80.0/14.7	89.0/22.8	77.8/10.6	80.9/18.3	76.5/11.4	87.2/17.1

Table 1: Performance of ALMA-13B-Pretrain and X-ALMA-13B-Pretrain on 50 languages from the Flores-200 test sets under three training approaches: Group multilingual training, Separate training, and Multilingual training. Results are categorized by language groups. Detailed scores for each group are provided in the Appendix.

three pre-training and two post-training stages with language-specific group training, to address this issue. In contrast, our findings suggest there may be a more efficient way to tackle the CoM. For example, we could start with a base model that only undergoes multilingual pretraining and then apply a dedicated post-training approach to achieve high-quality translation.

3.2 Asymmetry in Conflicts and Synergies Across Languages Groups and Resources

We further address a key question: *Does Asymmetry in Conflicts and Synergies occur across all language pairs, or is it concentrated in specific pairs?* To answer this, we analyze its distribution across different language groups and resource levels.

Figure 3 displays the results. We can have the following observations:

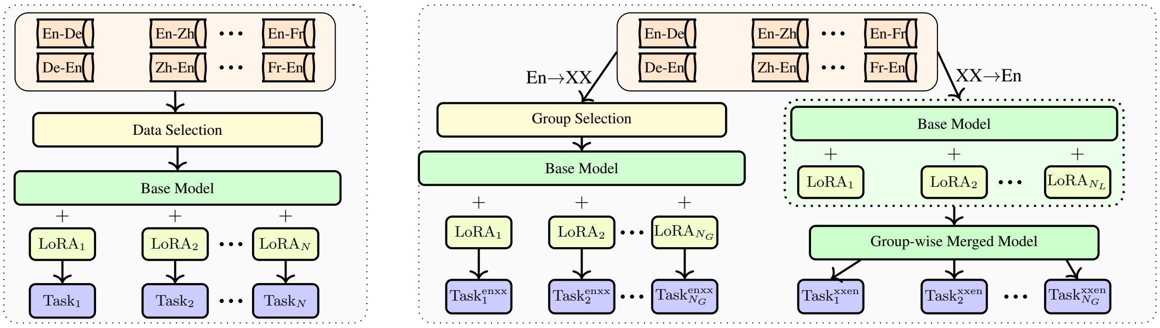
- Asymmetry in linguistic conflicts is consistently observed across languages with varying resource levels and language groups, but its intensity is not uniformly distributed.
- While increasing model capacity or pre-training data can help narrow the performance gap, consistent with findings in previous work (Arivazhagan et al., 2019; Aharoni et al., 2019; Shaham et al., 2023; Meng and Monz, 2024), a substantial gap of nearly 10 COMET-22 points still persists.

3.3 Challenges by Asymmetry in Linguistic Conflicts and Synergy

The asymmetry in linguistic conflicts and synergies may pose challenges for LLM-based MMT, leading to suboptimal performance for existing post-training approaches. Intuitively, translation directions where linguistic conflicts dominate may benefit from post-training strategies that minimize such conflicts. Conversely, translation directions where linguistic synergies prevail may require strategies that effectively enhance high-quality synergy. To see this, we fine-tune foundation models using three key approaches: multilingual training, group multilingual training, and separate training on 50 languages and compare their performance.

Table 1 displays the experimental results on the Flores-200 test set. We observe the following:

- **Key Findings 3: The effectiveness of the existing training strategy exhibits an asymmetrical pattern.**: In XX→En translations, separate training consistently achieves the best performance, followed by group multilingual training, while full multilingual training performs the worst. This result is expected, as linguistic conflict is prominent in these translation directions. By contrast, in En→XX translations, multilingual training or group multilingual training con-



(a) Seperate/Multilingual/Group Training

(b) Direction-aware Training with Group-wise Model Merging

Figure 4: (a) **Separate Training** ($N = N_L$): Each translation task is trained independently using different datasets for different language pairs, with distinct LoRA model weights fine-tuned separately; **Multilingual Training** ($N = 1$): All language pairs are combined to fine-tune a single model with shared LoRA weights; **Group Multilingual Training** ($N = N_G$): Language pairs are grouped as specified in Table 3-4, with an adapter trained for each group. (b) **Group-wise model merging**: For XX→En translation, separate training is applied to each language pair. For En→XX translation, group training is applied, where different tasks share LoRA weights within language groups.

sistently outperforms separate training. This indicates that while linguistic conflicts dominate in the XX→En direction, the En→XX direction benefits from cross-linguistic knowledge transfer, leading to an enhanced translation quality. When model capacity is sufficiently large, the general pattern observed is: group multilingual training > multilingual training > separate training. This highlights two things: 1) linguistic similarity benefits positive cross-linguistic transfer. 2) the widely adopted group multilingual training approach remains insufficient to address the challenges posed by the asymmetry.

These findings underscore the critical impact of asymmetry in linguistic conflicts and synergy phenomenon on the effectiveness of existing training strategies, highlighting the need for novel training approaches to consider such an asymmetry to achieve optimal performance in both directions.

4 Direction-Aware Training and Merging for Efficient LLM-based MMT

In this section, we show how to construct an efficient MMT system by leveraging the insights from Section 3, starting from a base model with simple multilingual pre-training.

4.1 Motivations and Main Ideas

As demonstrated in Section 3, linguistic conflicts and synergy exhibit asymmetry during the post-training stage, posing significant challenges to multilingual translation. A widely adopted technique to mitigate conflicts and enhance synergy is language-specific group multilingual training (Fan et al., 2021; Zhao et al., 2024; Xu et al., 2023, 2024b). However, it still achieves sub-optimal performance.

The state-of-the-art XALMA system (Xu et al., 2024b) achieves high-quality translations by employing eight large language-specific adapters within a MoE framework combined with group multilingual training. However, this approach incurs high computational and storage costs, as each adapter contains up to 15% of the base model's parameters, making large-scale deployment challenging. Additionally, XALMA requires a massive amount of tokens during pre-training, further increasing resource consumption. This raises an important question: *Can we achieve comparable high translation quality in a more efficient manner?*

Intuitively, we could develop a more efficient training approach for high-quality MMT by considering the asymmetry in linguistic conflicts and synergy. To this end, we propose a direction-aware training framework combined with model merging, which fully leverages the inherent asymmetry to enhance both performance and efficiency. Our approach primarily consists of two key components: 1) Direction-aware training strategies for efficiently and effectively mitigating linguistic conflicts and encouraging linguistic synergy and 2) Group-wise model merging for running efficiency.

4.2 Direction-Aware Training Strategies

As shown in Figure 4 (b), we propose a simple yet effective direction-aware training strategy that addresses linguistic conflicts and linguistic synergy separately for different translation directions:

- For XX→En translation directions: We employ separate training to build expert models for each language direction.
- For En→XX translation directions: We adopt

Model	# Tokens (Pre-training)	# Params		FLORES200		WMT23	
		Base/Adapter	Total	XX→En	En→XX	XX→En	En→XX
Existing State-of-the-Art MMT System							
NLLB-3.3B	-	3B/-	3B	80.7	87.4	71.2	81.5
Aya-23-8B	-	8B/-	8B	80.9	74.4	81.5	84.2
Aya-23-35B	-	35B/-	35B	84.9	76.0	82.3	84.1
Aya-101	-	13B/-	13B	86.3	84.1	79.7	80.8
LLaMAX3-Alpaca-8B	66B	8B/-	8B	85.9	84.1	81.0	79.8
X-ALMA-13B (Only SFT, MoE)	110B	13B/16B	29B	**88.2**	**88.9**	**83.2**	**85.6**
Our System							
X-ALMA-13B-DAT (MoE)	20B	13B/4B	17B	87.6	87.8	82.8	84.8
X-ALMA-13B-DATM (MoE)	20B	13B/1B	14B	87.4	87.8	82.1	84.8

Table 2: Performance on Flores-200 and WMT23 benchmarks. The results of baselines are directly sourced from Xu et al. (2024b) as we utilized same generation configuration. Full results are provided in Appendix.

group multilingual training, training one model per language group following Xu et al. (2024b). All training employs LoRA (Hu et al., 2022) with a rank of 16 for parameter efficiency. Using the proposed strategies, we construct a LoRA weight pool of size $N_G + N_L$, where N_G is the number of groups and N_L is the number of languages.

4.3 Group-wise Model Merging

Although the direction-aware training approach achieves promising performance, the number of LoRA weights increases linearly with the number of supported languages, posing challenges for deployment and inference, especially at large language scales. Model merging (Yadav et al., 2024; Zhang et al., 2023) provides a feasible solution to reduce the number of LoRA weights and improve efficiency. However, directly using model merge for efficient MMT is non-trial. In our preliminary experiments, we have two key observations:

- Merging LoRA weights into one for each direction leads to performance degradation. Notably, Dang et al. (2024) find that model merging can improve performance, contrasting our findings. However, this discrepancy may arise because their comparison is against a weaker baseline, such as multilingual training, whereas we compare against the most vigorous baseline—separate training.
- **The degradation effect of model merging exhibits an asymmetric nature.** The performance degradation per parameter in the En→XX direction is **6.86× greater** than in the XX→En direction. A potential explanation is that linguistic synergy plays a crucial role in En→XX directions, while model merging introduces low-quality linguistic synergy, leading to a performance drop.

Therefore, a more dedicated design is needed to preserve performance as much as possible.

Motivated by these observations, we only apply model merging to XX→En directions in a group-wise manner. Specifically, we apply model merging to languages within each group, resulting in N_G LoRA weights. We adopt the TIES (Yadav et al., 2024) for model merging. We also compare this approach with other methods such as DARE-TIES (Yu et al., 2024) and find no significant performance difference. With this approach, we can reduce the number of LoRA weights from $\mathcal{O}(N_L)$ to $\mathcal{O}(N_G)$, improving scalability while lead minimal performance degradation.

4.4 Main Results

We evaluated our models using the Flores-200 test set for 50 languages and the WMT23 test sets for five languages (de, ru, uk, ja, zh). We provide more details in Appendix A. We select existing state-of-the-art open multilingual MT system as baselines:

- **Aya-101** (Üstün et al., 2024): A 13B multilingual LLM supporting 101 languages.
- **LLaMAX** (Lu et al., 2024b): An 8B LLM-based MMT system supporting 102 languages.
- **Aya-23-8B/35B** (Aryabumi et al., 2024): An 8B/35B multilingual LLMs that support 23 languages.
- **XALMA** (Xu et al., 2024b): A 29B multilingual MoE-based MMT system supporting 50 languages, using language-specific adapters and group multilingual training. Notably, since we focus only on the supervised fine-tuning stage, we select the version without preference learning, namely XALMA-13B (Only SFT) to ensure fair

comparison.

Table 2 shows the results. We can have the following observations:

- Both X-ALMA-13B-DAT and X-ALMA-13B-DATM can achieve high translation performance. Compared to previous multilingual LLMs, such as Aya-101, Aya-23-8B, and LLaMAX, our approach consistently outperforms them across both benchmarks and translation directions. Moreover, compared to X-ALMA, our X-ALMA-13B-DAT achieves comparable performance in XX→En directions; however, in En→XX, a significant performance gap remains, up to 0.95 COMET-22 on average.
- Our approach provides an efficient way to build effective MMT. Our model is built upon X-ALMA-13B-Pretrain with only 20 billion tokens of simple multilingual pre-training. Moreover, it utilizes multiple small LoRA weight compositions and achieves relatively high translation performance across all directions, which is consistent with previous work (Zheng et al., 2024)

5 Related Work

5.1 Curse of Multilinguality

Existing research has explored both understanding and addressing this issue in MMT, identifying critical factors such as resource imbalances, limited model capacity, and complex interactions between language pairs, particularly for low-resource languages (Arivazhagan et al., 2019; Aharoni et al., 2019; Shaham et al., 2023). Interestingly, studies have shown that while linguistic similarity enhances positive transfer, dissimilar languages can also act as regularizers, improving training stability (Meng and Monz, 2024). To address these challenges, proposed solutions in recent research include language-specific modules (e.g., adapters, sparse experts) to dynamically allocate capacity and reduce interference (Fan et al., 2021; Zhao et al., 2024; Xu et al., 2023), vocabulary optimization to better support new languages through improved token representations (Han et al., 2024), data sampling techniques to enhance representation for underrepresented languages (Wang et al., 2020; Wang and Neubig, 2019; Lin et al., 2019) and continual learning techniques (Liu et al., 2023). Notably, techniques, such as language-specific modules, have been integrated into LLM-based MMT systems, resulting in substantial improvements in multilingual performance (Xu et al., 2024b). In this work, we systematically investigate how post-training in LLM-based MMT contributes to the CoM, providing a fine-grained analysis of its impact on linguistic conflicts and synergies.

5.2 LLMs for Multilingual MT

Many efforts have been made to adapt LLMs for effective machine translation. A key approach is prompting, which enhances translation performance without additional training (He et al., 2024; Lu et al., 2024a). Beyond this, growing research focuses on fine-tuning open and smaller LLMs to achieve high translation quality while ensuring efficiency (Xu et al., 2024a; Yang et al., 2023; Alves et al., 2024; Aryabumi et al., 2024).

Yang et al. (2023) propose a training pipeline that integrates monolingual pre-training to improve language modeling and parallel instruction fine-tuning for enhanced translation performance. Similarly, Xu et al. (2024a) emphasize the quality over quantity of parallel data, introducing a training recipe: (1) large-scale monolingual pre-training, followed by (2) small-scale, high-quality parallel fine-tuning. Further revisiting the role of parallel data, Guo et al. (2024) highlights its importance in the pre-training stage. Additionally, Xu et al. (2024c) underscore the necessity of alignment in post-training, proposing the CPO algorithm. More recently, with the need to scale models across more languages, Xu et al. (2024b) introduces language-specific modules combined with group training to mitigate language conflicts. In this work, we focus on the post-training stage, which has been underexplored in previous studies, and propose a direction-aware training approach with model merging to achieve efficient and effective MMT.

6 Conclusions

In this work, we systematically investigate linguistic conflicts and synergy during post-training in LLM-based MMT and identify a phenomenon we term asymmetry in linguistic conflicts and synergy. We provide an in-depth analysis of its distribution and challenges for LLM-based MMT. Based on these insights, we propose a direction-aware training approach combined with model merging to build an effective MMT system from X-ALMA-13B-Pretrain with only multilingual pre-training. Our approach highlights the importance of post-training in LLM-based MMT and offers insights into building MMT resource-efficiently.

Limitations

One limitation of this work is that our approach does not surpass state-of-the-art methods like X-ALMA in performance, particularly in En→XX directions, despite requiring less training cost and fewer model parameters. Second, while this work identifies a novel phenomenon and designs an efficient approach leveraging it, it does not provide a deeper or more rigorous analysis of why asymmetry in linguistic conflicts and synergy exists. We leave the analysis of the underlying mechanism of asymmetry in linguistic conflicts and synergy for future work.

Additionally, although this work conducts extensive experiments on fifty languages and three pre-trained models, further scaling is necessary to validate our findings on a broader scale, such as extending to over 100 languages. This would help push the boundaries of multilingual machine translation research, which we also leave for future work.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). ACL.

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

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, et al. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv 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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. 2024. Adapters for altering llm vocabularies: What languages benefit the most? *arXiv preprint arXiv:2410.09644*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019.

Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Junpeng Liu, Kaiyu Huang, Hao Yu, Jiuyi Li, Jinsong Su, and Degen Huang. 2023. Continual learning for multilingual neural machine translation via dual importance-based model division. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12011–12027, Singapore. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024a. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

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

Yan Meng and Christof Monz. 2024. Disentangling the roles of target-side transfer and regularization in multilingual machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1828–1840, St. Julian's, Malta. Association for Computational Linguistics.

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

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

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

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024c. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Haoran Xu, Weiting Tan, Shuyue Li, Yunmo Chen, Benjamin Van Durme, Philipp Koehn, and Kenton Murray. 2023. Condensing multilingual knowledge with lightweight language-specific modules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1575–1587, Singapore. Association for Computational Linguistics.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Jinghan Zhang, Junteng Liu, Junxian He, et al. 2023. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610.

Xinyu Zhao, Xuxi Chen, Yu Cheng, and Tianlong Chen. 2024. Sparse moe with language guided routing for multilingual machine translation. In *The Twelfth International Conference on Learning Representations*.

Tong Zheng, Bei Li, Huiwen Bao, Jiale Wang, Weiqiao Shan, Tong Xiao, and JingBo Zhu. 2024. PartialFormer: Modeling part instead of whole for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7280–7294, Bangkok, Thailand. Association for Computational Linguistics.

A Detailed Experimental Setups

In this section, we will discuss the detailed setup of our experiment, including the datasets.

A.1 Details of Dataset in Section 2.1

Following (Xu et al., 2024b), we present a classification of languages based on linguistic families, scripts, and resource availability in Tables 3- 4. Fifty languages are grouped into eight distinct categories, primarily guided by linguistic similarity while considering a balanced distribution of languages across groups. Each group encompasses a mix of low-, medium-, and high-resource languages to ensure comprehensive multilingual coverage. Additionally, English is included in each group to facilitate English-centric translation and mitigate catastrophic forgetting. This structured grouping provides a well-rounded dataset for multilingual research, enabling robust language modeling and cross-lingual transfer learning.

We train the translation model on X-ALMA-Parallel-Data, a parallel dataset in (Xu et al., 2024b). The distribution of the parallel datasets for each language is illustrated in Figure 5.

The evaluation dataset primarily consists of samples from the Flores-200 development set and NTREX (Barrault et al., 2019). In our experiment, we follow the setting in (Xu et al., 2024b), where the translation sentences are sampled to contain 1012 sentences in each language pair. We also use WMT23 benchmarks to assess performance for evaluation. The distribution of WMT23 for each language is illustrated in Figure 6.

For languages in both Flores-200 and WMT'15-22, corresponding test sets are incorporated, yielding an average of 4K examples per language.

B Additional Experiments

B.1 Asymmetry in Linguistic Conflicts and Synergy in terms of SacreBLEU

As shown in Figure 7, we observe a clear asymmetry in linguistic conflicts and synergy based on the SacreBLEU metric. This aligns with our main findings in the paper, where we used the COMET metric, further reinforcing the consistency of the observed phenomenon across different evaluation measures.

B.2 More Experiments on Asymmetry in Linguistic Conflicts and Synergy

We further design another setting to validate the asymmetry in linguistic conflicts and synergy.

Experimental Setup We select anchor sets of varying sizes and perform post-training using training sets that include different numbers of languages but cover those anchor sets. We then observe the performance changes of these anchor sets. If the performance declines as more languages are included in the training set, this would indicate the presence of linguistic conflicts.

Results Table 5 displays the results. We can clearly observe that in the XX-En directions, the average performance of each anchor set consistently decreases as the number of languages increases. However, this phenomenon is not observed in the En-XX directions, where performance remains relatively stable. The findings are consistent with Section 3.

B.3 Impact of Lora Rank

We observed an asymmetry in linguistic conflicts and synergies. A natural question arises: could this be due to using a low LoRA rank, which might limit learning capacity and, consequently, degrade performance? To address this concern, we selected the ALMA-13B-Pretrain model and trained it on 16 languages using different LoRA ranks, specifically 16 and 32. We then compared the performance of models with these LoRA ranks on the FLores-200 test sets. As shown in Table 6, increasing the LoRA rank did not yield performance improvements. Therefore, we conclude that the observed asymmetry is not attributed to using a low LoRA rank.

C Full Results

Language	ISO-639-1	Script	Family	Subgroup	Resource
English	`en`	Latin	Indo-European	Germanic	High
Group 1: Germanic Languages					
Afrikaans	`af`	Latin	Indo-European	Germanic	Mid
Danish	`da`	Latin	Indo-European	Germanic	Mid
Dutch	`nl`	Latin	Indo-European	Germanic	High
German	`de`	Latin	Indo-European	Germanic	High
Icelandic	`is`	Latin	Indo-European	Germanic	Low
Norwegian	`no`	Latin	Indo-European	Germanic	Low
Swedish	`sv`	Latin	Indo-European	Germanic	High
Group 2: Romance Languages					
Catalan	`ca`	Latin	Indo-European	Italic	High
Galician	`gl`	Latin	Indo-European	Italic	Mid
Italian	`it`	Latin	Indo-European	Italic	High
Portuguese	`pt`	Latin	Indo-European	Italic	High
Romanian	`ro`	Latin	Indo-European	Italic	Mid
Spanish	`es`	Latin	Indo-European	Italic	High
Group 3: Eastern and Southern Slavic Languages					
Bulgarian	`bg`	Cyrillic	Indo-European	Balto-Slavic	Mid
Macedonian	`mk`	Cyrillic	Indo-European	Balto-Slavic	Low
Russian	`ru`	Cyrillic	Indo-European	Balto-Slavic	High
Serbian	`sr`	Cyrillic	Indo-European	Balto-Slavic	High
Ukrainian	`uk`	Cyrillic	Indo-European	Balto-Slavic	Mid
Group 4: Southeast Asian Languages					
French	`fr`	Latin	Indo-European	Italic	High
Indonesian	`id`	Latin	Austronesian	Malayo-Polynesian	Mid
Malagasy	`mg`	Latin	Austronesian	Malayo-Polynesian	Low
Malay	`ms`	Latin	Austronesian	Malayo-Polynesian	Mid
Thai	`th`	Thai	Tai-Kadai	Kam-Tai	Mid
Vietnamese	`vi`	Latin	Austronesian	Vietic	High

Table 3: Detailed information of all languages

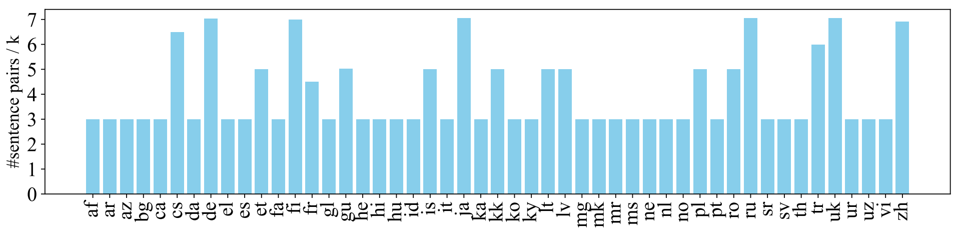


Figure 5: Number of sentences per language pair in X-ALMA-Parallel-Data (Xu et al., 2024b)

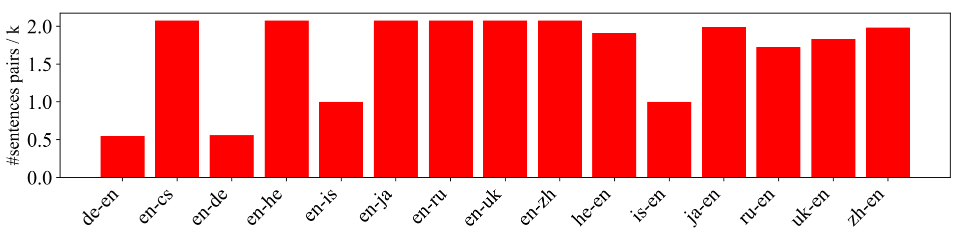


Figure 6: Number of Sentences per language pair in WMT'23

Language	ISO-639-1	Script	Family	Subgroup	Resource
Group 5: Central and Eastern European Languages					
Czech	cs	Latin	Indo-European	Balto-Slavic	Mid
Greek	el	Greek	Indo-European	Graeco-Phrygian	Mid
Hungarian	hu	Latin	Uralic	Finnic	High
Latvian	lv	Latin	Indo-European	Balto-Slavic	Mid
Lithuanian	lt	Latin	Indo-European	Balto-Slavic	Mid
Polish	pl	Latin	Indo-European	Balto-Slavic	High
Group 6: Eurasian Language Mix					
Chinese	zh	Han	Sino-Tibetan	Sinitic	High
Estonian	et	Latin	Uralic	Finnic	Mid
Finnish	fi	Latin	Uralic	Finnic	High
Georgian	ka	Georgian	Kartvelian	Georgian-Zan	Mid
Japanese	ja	Japanese	Japonic	Japanesic	High
Korean	ko	Hangul	Koreanic	Korean	High
Group 7: Indo-Aryan Languages					
Gujarati	gu	Gujarati	Indo-European	Indo-Aryan	Low
Hindi	hi	Devanagari	Indo-European	Indo-Aryan	High
Marathi	mr	Devanagari	Indo-European	Indo-Aryan	Low
Nepali	ne	Devanagari	Indo-European	Indo-Aryan	Low
Urdu	ur	Arabic	Indo-European	Indo-Aryan	Mid
Group 8: Turkic and Semitic Languages					
Arabic	ar	Arabic	Afro-Asiatic	Semitic	High
Azerbaijani	az	Arabic/Latin	Turkic	Common Turkic	Low
Hebrew	he	Hebrew	Afro-Asiatic	Semitic	Mid
Kazakh	kk	Cyrillic	Turkic	Common Turkic	Mid
Kyrgyz	ky	Cyrillic	Turkic	Common Turkic	Low
Persian	fa	Arabic	Indo-European	Iranian	High
Turkish	tr	Latin	Turkic	Common Turkic	High
Uzbek	uz	Latin	Turkic	Common Turkic	Low

Table 4: Detailed information of all languages (cont.)

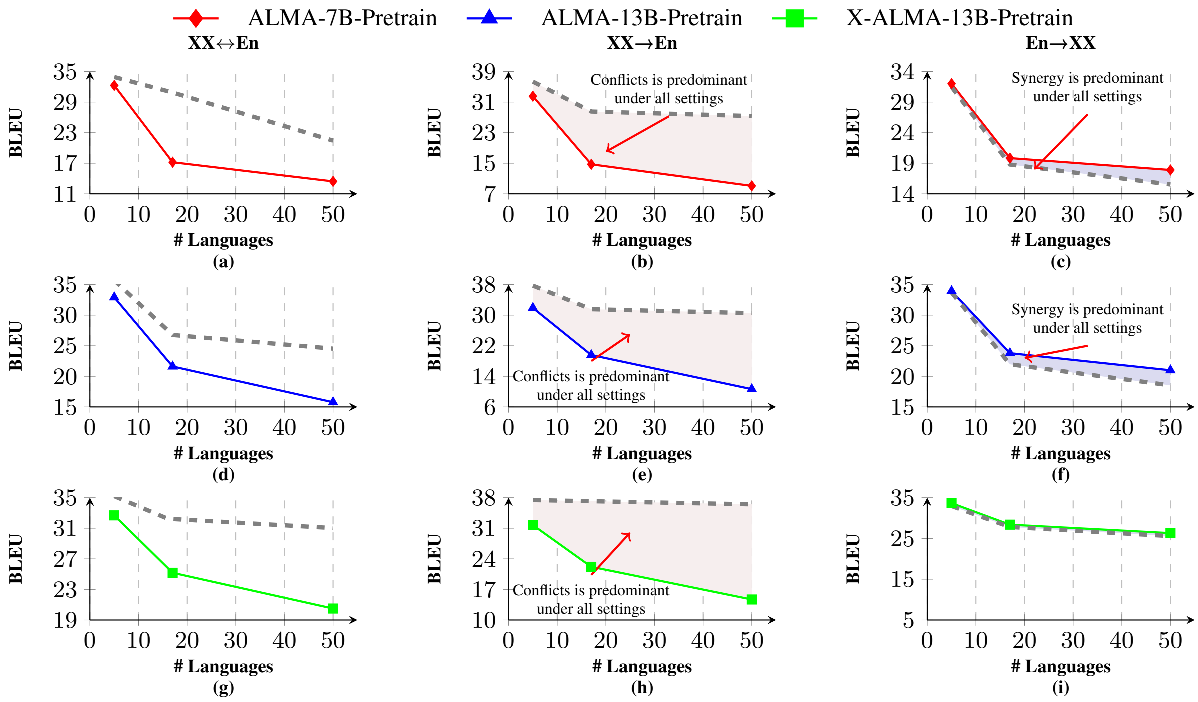


Figure 7: Performance of different models trained on varying numbers of languages (evaluated by SacreBLEU). The dotted line represents the performance of separately trained models, serving as a reference point where no language conflicts or synergies occur. Two key findings emerge: (1) Asymmetry in Linguistic Conflicts and Synergy (Figure a–i), highlighting the uneven impact of multilingual training across language pairs; and (2) The Bottleneck of Multilinguality in Post-Training (Figure g–i), showing that while monolingual pre-training provides an ideal foundation for handling multiple languages, the post-training stage imposes constraints that lead to the curse of multilinguality.

Languages (Trained)	Languages (Test)	XX→En	En→XX	All
ALMA-7B-Pretrain				
5	{De, Zh, Ru, CS}	86.8/32.5	88.5/32.0	87.7/32.3
17	{De, Zh, Ru, CS}	79.8/19.0	88.2/30.9	84.0/25.0
50	{De, Zh, Ru, CS}	76.4/17.7	88.4/31.3	82.4/24.5
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	79.6/18.2	87.4/26.5	83.5/22.4
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	76.5/17.3	87.5/27.2	82.0/22.3
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	74.3/14.7	78.8/19.8	76.6/17.3
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	70.4/13.3	78.7/20.4	74.6/16.7
ALMA-13B-Pretrain				
5	{De, Zh, Ru, CS}	86.7/31.9	89.2/33.9	88.0/32.9
17	{De, Zh, Ru, CS}	83.2/23.3	89.1/34.5	86.2/28.9
50	{De, Zh, Ru, CS}	78.0/18.5	89.0/34.3	83.5/26.4
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	83.5/23.1	89.2/30.6	86.4/26.9
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	78.8/18.9	89.1/30.4	84.0/24.7
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	79.8/19.6	83.3/23.8	81.6/21.7
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	73.4/14.9	83.2/23.6	78.3/19.3
X-ALMA-13B-Pretrain				
5	{De, Zh, Ru, CS}	86.5/31.7	88.4/33.7	87.5/32.7
17	{De, Zh, Ru, CS}	83.6/23.1	89.0/33.9	86.3/28.5
50	{De, Zh, Ru, CS}	80.9/19.5	88.9/33.4	84.9/26.5
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	84.2/23.2	89.9/31.8	87.1/27.5
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro}	81.6/19.8	89.8/31.2	85.7/25.5
17	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	83.7/22.2	88.5/28.4	86.1/25.3
50	{De, Zh, Ru, CS, Ja, Fi, Uk, Ro, Is, Kk, Fr, Lv, Gu, He, Hi, Hu}	80.6/18.2	88.4/28.1	84.5/23.2

Table 5: **The average performance of language in anchor set significantly decreases as the number of trained languages increase in XX-En directions while the average performance in En-XX directions maintain stable. This indicates the linguistic conflicts is predominant in XX-En directions, which is consistent with the findings in Section 3.** Performance was evaluated on Flores200 test sets.

LoRA Rank	En-XX		XX-En		Avg.	
	COMET-22	SacreBLEU	COMET-22	SacreBLEU	COMET-22	SacreBLEU
16	83.3	23.8	79.8	19.6	81.5	21.6
32	83.6	24.1	79.2	19.6	81.4	21.8

Table 6: Performance of ALMA-13B-Pretrain on FLores-200 Test Sets for Different LoRA Ranks.

Group 1 (Af, Da, Nl, De, Is, No, Sv)																
	XX→En								En→XX							
Strategy	Af	Da	Nl	De	Is	No	Sv	Avg.	Af	Da	Nl	De	Is	No	Sv	Avg.
ALMA-13B-Pretrain																
Group Multilingual Training	84.85	86.55	84.82	87.71	85.19	85.63	86.92	85.95	83.77	89.21	87.44	88.15	86.93	88.51	89.72	87.68
Multilingual Training (50)	66.14	70.25	69.87	79.43	80.38	69.23	71.04	72.33	82.65	88.68	87.01	87.98	86.80	88.04	89.22	87.20
Separate Training	87.56	89.60	87.38	89.35	87.04	88.44	89.79	88.45	80.94	88.05	86.81	87.72	86.89	87.35	88.58	86.62
X-ALMA-13B-Pretrain																
Group Multilingual Training	86.35	87.07	84.97	87.45	84.77	86.15	87.32	86.30	86.61	91.11	88.18	87.99	86.75	89.94	91.13	88.82
Multilingual Training (50)	76.65	78.71	76.05	82.53	82.65	77.29	78.91	78.97	86.51	90.93	88.27	87.77	86.30	89.81	90.82	88.63
Separate Training	89.32	90.23	87.46	89.28	86.70	88.93	90.13	88.86	86.36	90.53	87.96	87.66	86.52	89.65	90.70	88.48

Table 7: Result of various training strategies on Group 1 languages. The performance is evaluated by COMET-22.

Strategy	XX→En Ca	XX→En Gl	XX→En It	XX→En Pt	XX→En Ro	XX→En Es	XX→En Avg.	En→XX Ca	En→XX Gl	En→XX It	En→XX Pt	En→XX Ro	En→XX Es	En→XX Avg.
Group 2 (Ca, Gl, It, Pt, Ro, Es)														
ALMA-13B-Pretrain														
Group Multilingual Training	86.29	85.71	85.97	86.74	87.51	85.43	86.28	87.44	86.43	88.10	88.82	88.95	86.39	87.69
Multilingual Training (50)	69.41	67.79	70.34	69.30	82.55	67.91	71.22	87.14	85.52	87.94	88.69	88.67	86.21	87.36
Separate Training	88.59	87.88	88.13	89.33	89.06	87.42	88.40	87.10	84.59	87.83	88.53	88.61	86.09	87.12
X-ALMA-13B-Pretrain														
Group Multilingual Training	86.59	86.23	86.20	87.06	87.88	85.41	86.56	88.08	87.82	88.55	89.44	90.71	86.78	88.56
Multilingual Training (50)	76.67	76.44	76.87	77.03	84.77	75.93	77.95	88.03	87.72	88.32	89.36	90.70	86.68	88.47
Separate Training	88.81	88.46	88.18	89.32	89.36	87.50	88.61	87.77	87.59	88.29	89.21	90.42	86.51	88.30

Table 8: Result of various training strategies on Group 2 languages. The performance is evaluated by COMET-22.

Strategy	XX→En Bg	XX→En Mk	XX→En Ru	XX→En Sr	XX→En Uk	XX→En Avg.	En→XX Bg	En→XX Mk	En→XX Ru	En→XX Sr	En→XX Uk	En→XX Avg.
Group 3 (Bg, Mk, Ru, Sr, Uk)												
ALMA-13B-Pretrain												
Group Multilingual Training	85.48	84.30	85.34	85.44	85.49	85.21	89.16	85.24	89.69	86.80	89.16	88.01
Multilingual Training (50)	69.35	66.40	76.99	68.08	77.54	71.67	88.47	83.57	89.44	86.71	89.16	87.47
Separate Training	87.45	85.78	86.86	87.12	87.26	86.89	88.21	81.19	89.57	86.14	89.21	86.86
X-ALMA-13B-Pretrain												
Group Multilingual Training	86.50	85.77	85.60	86.00	85.84	85.94	90.95	89.55	89.64	88.74	90.18	89.81
Multilingual Training (50)	76.11	75.92	79.33	75.76	79.44	77.31	90.90	89.45	89.64	89.67	89.87	89.91
Separate Training	87.98	87.76	86.78	87.81	87.36	87.54	90.51	89.44	89.34	89.13	90.08	89.70

Table 9: Result of various training strategies on Group 3 languages. The performance is evaluated by COMET-22.

Strategy	XX→En Fr	XX→En Id	XX→En Mg	XX→En Ms	XX→En Th	XX→En Vi	XX→En Avg.	En→XX Fr	En→XX Id	En→XX Mg	En→XX Ms	En→XX Th	En→XX Vi	En→XX Avg.
Group 4 (Fr, Id, Mg, Ms, Th, Vi)														
ALMA-13B-Pretrain														
Group Multilingual Training	86.98	85.93	63.47	84.55	77.34	83.41	80.28	87.60	89.92	63.79	86.85	72.74	87.01	81.32
Multilingual Training (50)	79.99	68.69	53.49	66.90	61.02	67.33	66.24	87.53	89.60	63.29	86.48	70.57	86.07	80.59
Separate Training	89.22	88.76	56.68	87.45	78.61	86.59	81.22	87.46	89.36	39.87	85.64	59.28	86.49	74.68
X-ALMA-13B-Pretrain														
Group Multilingual Training	86.94	86.35	73.33	84.82	85.07	84.24	83.46	87.96	90.75	76.37	87.89	86.51	88.49	86.33
Multilingual Training (50)	83.49	77.59	65.08	75.95	76.97	75.77	75.81	88.07	90.73	75.70	87.80	86.25	88.40	86.16
Separate Training	89.25	89.11	74.75	87.71	87.57	87.21	85.93	87.72	90.75	71.99	87.46	85.84	88.08	85.31

Table 10: Result of various training strategies on Group 4 languages. The performance is evaluated by COMET-22.

Strategy	XX→En Cs	XX→En El	XX→En Hu	XX→En Lv	XX→En Lt	XX→En Pl	XX→En Avg.	En→XX Cs	En→XX El	En→XX Hu	En→XX Lv	En→XX Lt	En→XX Pl	En→XX Avg.
Group 5 (Cs, El, Hu, Lv, Lt, Pl)														
ALMA-13B-Pretrain														
Group Multilingual Training	87.92	81.93	85.57	79.42	76.79	84.70	82.72	91.28	76.51	86.69	69.81	67.41	88.21	79.98
Multilingual Training (50)	79.94	62.88	68.91	68.85	63.48	80.26	70.72	91.11	75.92	86.24	68.90	66.40	87.63	79.37
Separate Training	88.77	82.72	87.66	78.75	75.81	85.94	83.28	90.95	68.11	86.23	61.62	58.89	87.96	75.63
X-ALMA-13B-Pretrain														
Group Multilingual Training	87.30	85.77	85.81	85.63	84.58	84.64	85.62	91.31	89.03	89.57	89.80	90.07	89.69	89.91
Multilingual Training (50)	82.44	76.80	76.78	81.80	78.77	81.73	79.72	91.31	88.88	89.36	89.49	89.67	89.70	89.73
Separate Training	88.58	87.51	87.99	87.58	86.49	86.14	87.38	90.83	88.87	89.03	89.64	89.79	89.61	89.63

Table 11: Result of various training strategies on Group 5 languages. The performance is evaluated by COMET-22.

Group 6 (Et, Fi, Ja, Ka, Ko, Zh)														
	XX→En							En→XX						
Strategy	Et	Fi	Ja	Ka	Ko	Zh	Avg.	Et	Fi	Ja	Ka	Ko	Zh	Avg.
ALMA-13B-Pretrain														
Group Multilingual Training	80.84	87.60	86.18	73.71	85.36	85.22	83.15	73.01	89.85	89.92	61.63	84.44	87.75	81.10
Multilingual Training (50)	71.87	80.54	77.25	58.64	69.78	75.86	72.32	70.26	89.19	89.64	62.08	85.68	87.55	80.73
Separate Training	80.97	89.15	87.38	70.05	86.85	86.71	83.52	61.16	89.58	89.64	41.78	85.28	87.29	75.79
X-ALMA-13B-Pretrain														
Group Multilingual Training	87.23	88.25	86.16	84.34	85.97	85.01	86.16	90.79	92.29	90.44	87.02	88.14	87.46	89.36
Multilingual Training (50)	83.70	84.38	80.55	74.73	77.59	79.11	80.01	90.55	92.00	90.35	86.23	87.97	86.92	89.00
Separate Training	88.75	89.59	87.64	86.37	87.49	86.62	87.74	90.45	92.13	90.40	86.27	87.36	86.99	88.93

Table 12: Result of various training strategies on Group 6 languages. The performance is evaluated by COMET-22.

Group 7 (Gu, Hi, Mr, Ne, Ur)												
	XX→En						En→XX					
Strategy	Gu	Hi	Mr	Ne	Ur	Avg.	Gu	Hi	Mr	Ne	Ur	Avg.
ALMA-13B-Pretrain												
Group Multilingual Training	70.05	83.00	76.42	81.54	76.14	77.43	73.06	69.72	59.17	72.70	68.72	68.67
Multilingual Training (50)	59.78	65.44	60.27	65.85	59.95	62.26	76.01	70.68	61.03	74.13	71.06	70.58
Separate Training	66.96	84.27	76.60	82.62	77.47	77.59	59.24	62.51	45.03	59.68	60.75	57.44
X-ALMA-13B-Pretrain												
Group Multilingual Training	86.43	87.19	85.57	87.82	84.84	86.37	86.68	79.15	74.19	82.71	82.09	80.96
Multilingual Training (50)	79.75	77.66	76.45	79.23	75.84	77.79	86.60	79.37	73.76	82.29	82.21	80.85
Separate Training	87.07	88.83	87.44	89.29	86.49	87.82	85.91	78.57	73.17	81.71	81.84	80.24

Table 13: Result of various training strategies on Group 7 languages. The performance is evaluated by COMET-22.

Group 8 (Ar, Az, He, Kk, Ky, Fa, Tr, Uz)																		
	XX→En									En→XX								
Strategy	Ar	Az	He	Kk	Ky	Fa	Tr	Uz	Avg.	Ar	Az	He	Kk	Ky	Fa	Tr	Uz	Avg.
ALMA-13B-Pretrain																		
Group Multilingual Training	74.17	72.83	70.54	73.58	70.06	75.22	82.01	70.34	73.59	77.14	63.41	71.63	68.57	62.74	71.52	77.61	66.83	69.93
Multilingual Training (50)	60.33	61.77	57.48	63.74	60.45	62.40	76.40	59.61	62.77	75.78	65.72	71.12	71.08	63.12	72.18	78.45	69.64	70.89
Separate Training	80.97	75.94	77.71	72.45	67.25	81.80	85.36	70.53	76.50	73.00	43.13	62.30	46.70	35.71	64.67	74.37	44.37	55.53
X-ALMA-13B-Pretrain																		
Group Multilingual Training	81.70	82.67	83.28	83.53	81.19	83.14	87.15	81.15	82.98	86.40	86.86	87.74	89.41	87.42	86.69	88.50	86.80	87.48
Multilingual Training (50)	74.48	75.55	75.45	78.74	74.45	75.96	83.65	73.35	76.45	86.11	86.67	87.44	89.33	87.22	86.60	88.35	86.20	87.24
Separate Training	86.57	85.65	87.77	87.00	84.49	87.23	88.83	84.21	86.47	85.72	85.81	87.88	88.84	86.45	86.44	87.94	83.71	86.60

Table 14: Result of various training strategies on Group 8 languages. The performance is evaluated by COMET-22.

Group 1 (Af, Da, Nl, De, Is, No, Sv)																
	XX→En								En→XX							
Strategy	Af	Da	Nl	De	Is	No	Sv	Avg.	Af	Da	Nl	De	Is	No	Sv	Avg.
ALMA-13B-Pretrain																
Group Multilingual Training	35.39	31.82	23.66	34.88	28.67	29.25	31.76	30.78	33.13	38.35	24.66	36.48	25.37	27.93	37.83	31.96
Multilingual Training (50)	10.68	12.10	10.23	21.00	24.37	11.10	12.09	14.51	33.00	36.93	24.34	36.82	25.07	27.65	37.18	31.57
Separate Training	52.88	46.74	32.21	44.21	37.39	42.92	47.21	43.37	27.63	35.25	23.39	36.22	25.41	25.40	35.72	29.86
X-ALMA-13B-Pretrain																
Group Multilingual Training	37.40	32.39	23.63	33.66	28.30	29.68	32.74	31.11	40.68	42.78	26.38	36.09	24.23	30.85	41.80	34.69
Multilingual Training (50)	16.45	16.06	11.58	22.41	24.56	14.76	16.14	17.42	40.21	41.95	26.52	35.32	23.82	30.57	41.79	34.31
Separate Training	57.35	48.92	32.66	43.83	35.92	43.87	48.54	44.44	39.08	42.41	26.08	35.04	24.05	30.50	41.25	34.06

Table 15: Result of various training strategies on Group 1 languages. The performance is evaluated by SacreBLEU.

	Group 2 (Ca, Gl, It, Pt, Ro, Es)													
	XX→En							En→XX						
Strategy	Ca	Gl	It	Pt	Ro	Es	Avg.	Ca	Gl	It	Pt	Ro	Es	Avg.
ALMA-13B-Pretrain														
Group Multilingual Training	33.77	31.56	27.82	34.82	35.49	26.81	31.71	39.10	31.01	27.53	41.43	32.21	27.81	33.18
Multilingual Training (50)	11.95	10.61	11.07	11.14	27.33	9.09	13.53	38.50	29.64	27.24	40.89	32.24	27.02	32.59
Separate Training	46.06	41.02	35.75	47.77	43.82	33.89	41.38	37.44	27.43	26.37	40.84	31.85	26.38	31.72
X-ALMA-13B-Pretrain														
Group Multilingual Training	33.79	31.43	27.45	34.99	35.41	26.29	31.56	41.11	35.11	29.31	43.43	37.63	28.15	35.79
Multilingual Training (50)	14.74	13.75	13.19	14.84	28.91	12.40	16.30	40.59	34.85	29.10	43.23	36.83	28.02	35.44
Separate Training	46.01	42.02	35.54	47.73	44.62	33.79	41.62	39.45	33.71	28.55	43.09	36.86	27.55	34.87

Table 16: Result of various training strategies on Group 2 languages. The performance is evaluated by SacreBLEU.

	Group 3 (Bg, Mk, Ru, Sr, Uk)											
	XX→En						En→XX					
Strategy	Bg	Mk	Ru	Sr	Uk	Avg.	Bg	Mk	Ru	Sr	Uk	Avg.
ALMA-13B-Pretrain												
Group Multilingual Training	30.72	31.47	30.09	33.53	31.89	31.54	30.05	22.85	29.03	26.63	25.21	26.75
Multilingual Training (50)	9.96	9.16	18.08	10.21	19.70	13.42	30.21	21.87	29.62	26.79	25.28	26.75
Separate Training	39.56	38.58	36.13	42.05	40.32	39.33	28.31	17.45	29.18	23.40	25.11	24.69
X-ALMA-13B-Pretrain												
Group Multilingual Training	33.56	34.13	30.92	34.43	32.81	33.17	36.66	32.41	28.98	31.32	27.09	31.29
Multilingual Training (50)	12.96	13.49	18.33	14.14	18.85	15.55	36.36	32.30	29.38	32.00	26.90	31.39
Separate Training	40.40	43.17	36.69	44.00	40.71	40.99	35.61	31.28	28.61	30.38	26.45	30.47

Table 17: Result of various training strategies on Group 3 languages. The performance is evaluated by SacreBLEU.

	Group 4 (Fr, Id, Mg, Ms, Th, Vi)													
	XX→En							En→XX						
Strategy	Fr	Id	Mg	Ms	Th	Vi	Avg.	Fr	Id	Mg	Ms	Th	Vi	Avg.
ALMA-13B-Pretrain														
Group Multilingual Training	35.54	30.22	10.13	29.61	13.75	25.54	24.13	43.95	37.60	4.01	28.67	6.73	34.86	25.97
Multilingual Training (50)	24.53	9.34	3.85	9.18	4.22	8.58	9.95	44.11	36.40	4.38	28.37	7.15	33.54	25.66
Separate Training	44.78	41.77	7.63	40.57	16.36	34.66	30.96	43.97	35.24	54.00	25.13	4.11	32.97	23.66
X-ALMA-13B-Pretrain														
Group Multilingual Training	35.04	31.29	16.10	29.44	21.36	26.20	26.57	46.13	39.97	9.99	30.83	9.93	38.68	29.25
Multilingual Training (50)	27.13	13.59	6.83	13.09	10.01	12.26	13.82	45.36	40.01	9.23	29.28	11.36	37.54	28.80
Separate Training	44.69	42.99	19.64	40.79	29.92	36.98	35.84	45.30	39.49	6.48	28.01	9.78	37.43	27.75

Table 18: Result of various training strategies on Group 4 languages. The performance is evaluated by SacreBLEU.

	Group 5 (Cs, El, Hu, Lv, Lt, Pl)													
	XX→En							En→XX						
Strategy	Cs	El	Hu	Lv	Lt	Pl	Avg.	Cs	El	Hu	Lv	Lt	Pl	Avg.
ALMA-13B-Pretrain														
Group Multilingual Training	36.91	23.58	26.31	21.62	19.37	27.05	25.81	31.04	13.04	17.21	10.46	8.63	18.20	16.43
Multilingual Training (50)	22.08	6.40	8.19	11.58	8.37	21.72	13.06	31.02	13.59	17.70	10.33	9.68	18.14	16.74
Separate Training	41.62	27.38	33.58	22.89	19.97	30.96	29.40	30.62	9.04	15.56	6.55	5.59	17.93	14.22
X-ALMA-13B-Pretrain														
Group Multilingual Training	34.26	28.15	25.25	28.89	26.61	26.19	28.22	30.91	23.62	21.91	24.56	22.52	20.28	23.97
Multilingual Training (50)	23.16	12.64	11.68	21.52	16.20	21.59	17.80	30.23	23.40	21.72	23.96	22.27	20.02	23.60
Separate Training	40.95	35.56	34.62	35.99	33.79	31.12	35.34	29.92	22.18	21.17	25.27	22.27	20.05	23.48

Table 19: Result of various training strategies on Group 5 languages. The performance is evaluated by SacreBLEU.

Group 6 (Et, Fi, Ja, Ka, Ko, Zh)														
	XX→En							En→XX						
Strategy	Et	Fi	Ja	Ka	Ko	Zh	Avg.	Et	Fi	Ja	Ka	Ko	Zh	Avg.
ALMA-13B-Pretrain														
Group Multilingual Training	22.49	27.83	22.64	12.35	21.50	23.92	21.79	10.60	18.59	29.63	4.40	6.79	40.49	18.42
Multilingual Training (50)	13.25	18.25	11.69	3.56	7.05	12.82	11.10	10.10	18.29	28.71	5.34	9.87	39.74	18.67
Separate Training	24.10	33.33	26.70	9.22	26.77	28.95	24.84	5.44	17.45	28.49	1.52	9.06	38.79	16.79
X-ALMA-13B-Pretrain														
Group Multilingual Training	29.46	27.76	22.32	20.23	22.24	22.97	24.16	21.92	22.78	31.87	12.59	10.08	39.99	23.21
Multilingual Training (50)	22.32	20.01	12.44	9.06	10.38	14.09	14.72	21.48	21.38	30.89	12.32	11.82	38.65	22.76
Separate Training	35.84	34.19	26.94	27.00	28.37	28.31	30.11	21.52	21.62	31.76	11.54	10.65	38.22	22.55

Table 20: Result of various training strategies on Group 6 languages. The performance is evaluated by SacreBLEU.

Group 7 (Gu, Hi, Mr, Ne, Ur)												
	XX→En						En→XX					
Strategy	Gu	Hi	Mr	Ne	Ur	Avg.	Gu	Hi	Mr	Ne	Ur	Avg.
ALMA-13B-Pretrain												
Group Multilingual Training	10.95	24.45	16.88	19.50	16.89	17.73	8.78	15.68	5.63	8.67	9.90	9.73
Multilingual Training (50)	3.10	5.34	3.71	4.46	3.59	4.04	10.50	17.29	6.25	9.53	11.33	10.98
Separate Training	8.09	26.64	15.40	20.37	17.67	17.63	4.89	9.66	2.09	3.89	5.11	5.13
X-ALMA-13B-Pretrain												
Group Multilingual Training	26.47	29.68	25.56	27.99	25.45	27.03	18.18	26.33	13.57	16.75	18.94	18.75
Multilingual Training (50)	12.55	10.32	10.44	10.05	9.80	10.63	17.78	25.96	13.17	15.90	18.88	18.34
Separate Training	30.82	36.37	32.69	33.91	31.51	33.06	16.55	23.83	12.58	15.01	17.78	17.15

Table 21: Result of various training strategies on Group 7 languages. The performance is evaluated by SacreBLEU.

Group 8 (Ar, Az, He, Kk, Ky, Fa, Tr, Uz)																		
	XX→En									En→XX								
Strategy	Ar	Az	He	Kk	Ky	Fa	Tr	Uz	Avg.	Ar	Az	He	Kk	Ky	Fa	Tr	Uz	Avg.
ALMA-13B-Pretrain																		
Group Multilingual Training	14.67	9.20	14.01	10.83	7.40	14.42	21.09	9.47	12.64	10.04	3.34	10.70	4.42	2.41	11.27	9.99	3.16	6.92
Multilingual Training (50)	5.09	3.87	4.01	5.63	3.46	5.18	15.65	4.19	5.88	11.04	4.76	10.80	5.87	2.87	12.39	11.96	5.05	8.09
Separate Training	28.08	12.64	25.00	11.36	6.59	24.96	28.39	10.72	18.47	8.03	84.00	6.60	1.13	40.00	6.09	7.95	53.00	3.95
X-ALMA-13B-Pretrain																		
Group Multilingual Training	21.40	16.27	24.74	19.94	14.91	20.98	28.69	17.76	20.59	20.13	11.89	25.11	17.26	11.53	21.95	21.56	10.98	17.55
Multilingual Training (50)	11.13	7.89	11.92	12.76	8.08	10.83	20.84	7.77	11.40	19.43	12.11	24.68	17.22	10.65	21.66	20.39	10.52	17.08
Separate Training	38.34	22.16	41.85	30.35	20.97	33.66	35.90	25.28	31.06	18.84	10.10	24.62	15.62	8.89	20.92	19.41	6.56	15.62

Table 22: Result of various training strategies on Group 8 languages. The performance is evaluated by SacreBLEU.

Model	Zh		De		Ru		Ja		Uk		Avg.	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-3.3B	11.4	67.8	20.1	66.6	24.4	76.7	6.8	65.8	33.1	79.0	19.2	71.2
Aya-23-8B	22.6	78.8	32.3	82.1	30.9	81.6	19.8	80.2	39.2	85.0	29.0	81.5
Aya-23-35B	23.5	79.7	32.7	82.3	31.7	82.2	21.3	81.6	39.1	85.7	29.7	82.3
Aya-101	13.8	73.7	34.9	81.6	28.4	81.4	13.9	77.3	34.9	84.5	25.2	79.7
LLaMAX3-Alpaca-8B	22.3	79.3	25.6	79.4	29.4	81.3	17.6	80.1	37.8	84.9	26.5	81.0
X-ALMA-13B (Only SFT, MoE)	23.8	80.3	42.5	85.3	32.8	82.4	20.4	81.6	42.5	86.4	32.4	83.2
X-ALMA-13B-DAT (MoE)	21.1	79.7	38.4	84.6	32.3	82.4	19.8	81.2	39.5	85.9	30.2	82.8
X-ALMA-13B-DATM (MoE)	22.2	79.2	37.5	84.2	29.0	82.0	19.0	80.4	34.0	84.7	28.3	82.1

Table 23: WMT23 XX→En translation results (BLEU and COMET). The results of baselines are directly sourced from Xu et al. (2024b). We keep the same generation configuration as Xu et al. (2024b).

Model	Zh		De		Ru		Ja		Uk		Avg.	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-3.3B	34.8	79.6	33.5	79.7	29.1	83.8	13.8	81.6	25.5	82.8	27.3	81.5
Aya-23-8B	44.5	85.3	29.3	80.4	24.3	84.3	19.3	86.5	24.3	84.3	28.3	84.2
Aya-23-35B	42.8	84.6	30.7	80.7	27.5	84.7	20.6	86.4	24.9	84.0	29.3	84.1
Aya-101	25.4	78.6	25.1	75.1	22.1	83.1	14.1	84.6	19.7	82.7	21.3	80.8
LLaMAX3-Alpaca-8B	34.0	81.5	20.9	73.3	23.5	81.6	11.9	81.8	19.8	80.6	22.0	79.8
X-ALMA-13B (Only SFT, MoE)	47.5	86.1	40.9	84.1	31.5	85.9	22.3	86.8	27.4	85.3	33.9	85.6
X-ALMA-13B-DAT (MoE)	40.3	85.0	35.3	83.3	27.6	85.0	19.0	86.0	23.5	84.6	29.1	84.8
X-ALMA-13B-DATM (MoE)	40.3	85.0	35.3	83.3	27.6	85.0	19.0	86.0	23.5	84.6	29.1	84.8

Table 24: WMT23 En→XX translation results (BLEU and COMET). The results of baselines are directly sourced from Xu et al. (2024b). We keep the same generation configuration as Xu et al. (2024b).

Direction	NLLB-3.3B	LLaMAX3-Alpaca-8B	Aya-101	Aya-23-8B	Aya-23-35B	X-ALMA-13B (only SFT)	X-ALMA-13B-DAT (Ours)	X-ALMA-13B-DATM (Ours)
Group 1 (Af, Da, Nl, De, Is, No, Sv)								
en→af	87.4 / 38.9	86.0 / 38.5	78.8 / 22.5	79.6 / 17.6	81.2 / 26.7	87.5 / 44.2	86.6 / 40.7	
en→da	90.0 / 44.5	88.6 / 38.2	87.6 / 34.2	76.4 / 19.3	82.9 / 29.0	91.8 / 48.6	91.1 / 42.8	
en→de	88.1 / 40.0	85.4 / 31.4	84.3 / 29.3	88.1 / 36.8	88.1 / 37.0	88.7 / 41.2	88.0 / 36.1	
en→is	84.6 / 24.5	81.2 / 18.3	84.3 / 20.9	38.4 / 1.6	51.0 / 5.9	87.2 / 28.0	86.8 / 24.2	
en→nl	87.5 / 27.5	86.3 / 23.3	85.8 / 22.1	87.9 / 26.0	87.7 / 26.6	88.8 / 29.3	88.2 / 26.4	
en→no	88.9 / 33.0	87.8 / 28.0	87.5 / 26.9	77.3 / 15.7	82.4 / 22.1	90.6 / 35.0	89.9 / 30.9	
en→sv	90.7 / 44.3	89.1 / 38.7	86.9 / 31.3	78.3 / 20.8	83.7 / 28.8	91.7 / 47.0	91.1 / 41.8	
af→en	80.3 / 40.6	89.0 / 53.1	86.1 / 43.2	85.3 / 46.9	88.3 / 54.3	89.9 / 58.8	89.3 / 57.4	89.1 / 55.1
da→en	83.0 / 34.4	89.6 / 45.3	89.2 / 42.4	87.7 / 42.6	89.7 / 47.3	90.2 / 49.6	90.2 / 48.9	90.3 / 48.7
de→en	81.3 / 28.6	88.8 / 40.5	88.5 / 39.7	89.3 / 43.9	89.5 / 45.1	89.6 / 45.7	89.3 / 43.8	89.2 / 43.8
is→en	64.2 / 16.2	85.6 / 32.5	82.3 / 27.2	68.0 / 13.0	78.5 / 24.5	87.1 / 37.7	86.7 / 35.9	86.7 / 35.7
nl→en	81.9 / 25.3	87.1 / 30.1	86.9 / 30.1	87.5 / 31.9	87.8 / 33.9	87.6 / 34.2	87.5 / 32.7	87.6 / 32.5
no→en	80.7 / 32.1	88.5 / 41.8	88.1 / 39.5	86.5 / 38.5	88.5 / 43.2	89.1 / 45.7	88.9 / 43.9	88.9 / 44.5
sv→en	82.3 / 35.0	89.5 / 45.6	89.4 / 44.3	87.9 / 42.6	89.5 / 46.9	90.2 / 50.0	90.1 / 48.5	90.1 / 48.7
Group 2 (Ca, Es, Gl, It, Pt, Ro)								
en→ca	87.8 / 43.1	86.5 / 36.3	87.1 / 37.8	81.7 / 25.1	83.9 / 33.1	89.0 / 45.7	88.1 / 41.1	
en→es	86.5 / 28.6	85.0 / 24.1	85.3 / 24.2	86.4 / 27.8	86.2 / 27.7	87.2 / 29.5	86.8 / 28.2	
en→gl	87.3 / 35.7	86.4 / 31.2	86.7 / 32.7	82.7 / 17.2	84.2 / 25.3	88.4 / 39.0	87.8 / 35.1	
en→it	88.5 / 31.3	86.9 / 26.5	87.0 / 25.6	88.4 / 30.2	88.2 / 30.5	89.1 / 32.5	88.6 / 29.3	
en→pt	89.6 / 49.6	88.1 / 41.5	85.3 / 32.5	89.9 / 48.4	89.7 / 48.6	90.2 / 49.9	89.4 / 43.4	
en→ro	90.2 / 37.6	88.1 / 32.7	89.4 / 34.9	90.6 / 37.9	90.7 / 38.4	91.5 / 42.2	90.7 / 37.6	
ca→en	83.7 / 37.9	88.3 / 42.9	87.6 / 41.1	85.8 / 39.5	88.4 / 46.3	89.2 / 48.6	88.8 / 46.0	88.8 / 46.4
es→en	85.3 / 27.1	86.7 / 29.0	86.8 / 28.8	87.4 / 31.3	87.7 / 33.1	87.7 / 34.9	87.5 / 33.8	87.4 / 32.8
gl→en	84.0 / 34.7	88.0 / 38.6	86.9 / 35.5	87.0 / 37.3	88.5 / 41.7	89.0 / 44.9	88.5 / 42.0	88.5 / 42.1
it→en	84.4 / 28.8	87.5 / 31.3	87.4 / 31.2	88.1 / 34.1	88.3 / 36.0	88.3 / 36.9	88.2 / 35.5	88.2 / 35.4
pt→en	86.7 / 42.3	89.1 / 46.3	88.7 / 43.8	89.7 / 49.7	89.9 / 51.5	89.7 / 51.0	89.3 / 47.7	89.4 / 48.4
ro→en	83.0 / 31.4	88.9 / 40.4	88.4 / 37.8	89.5 / 43.5	89.7 / 46.0	89.7 / 46.8	89.4 / 44.6	89.4 / 43.9
Group 3 (Bg, Mk, Ru, Sr, Uk)								
en→bg	90.9 / 40.5	89.0 / 32.2	90.0 / 34.3	73.3 / 6.7	75.7 / 17.0	91.7 / 42.1	90.0 / 36.7	
en→mk	88.8 / 34.4	87.4 / 29.3	88.7 / 30.7	57.1 / 2.9	65.4 / 9.6	90.4 / 37.3	89.6 / 32.4	
en→ru	89.2 / 32.2	87.7 / 26.4	88.3 / 27.2	89.6 / 29.9	89.6 / 31.2	90.1 / 32.3	89.6 / 29.0	
en→sr	89.0 / 33.8	76.2 / 5.8	82.9 / 23.3	61.7 / 0.9	67.4 / 1.1	90.2 / 36.4	88.7 / 31.3	
en→uk	89.1 / 30.3	87.9 / 25.5	88.7 / 25.1	90.2 / 29.4	90.0 / 30.3	90.8 / 31.8	90.2 / 27.1	
bg→en	86.0 / 37.6	87.5 / 38.2	85.4 / 32.9	84.4 / 32.6	86.7 / 38.2	88.4 / 43.4	88.0 / 40.4	88.0 / 40.3
mk→en	84.3 / 37.1	87.2 / 39.8	84.3 / 33.7	78.4 / 25.0	84.6 / 36.2	88.2 / 45.6	87.8 / 43.2	87.6 / 42.7
ru→en	84.2 / 30.7	86.4 / 33.1	86.1 / 32.7	86.7 / 36.1	87.1 / 38.6	87.0 / 38.7	86.8 / 36.7	86.8 / 36.6
sr→en	83.4 / 35.8	87.3 / 40.6	85.0 / 35.0	79.9 / 27.9	85.3 / 37.8	88.4 / 46.2	87.8 / 44.0	87.9 / 43.9
uk→en	83.7 / 33.7	86.8 / 37.0	86.2 / 35.5	87.2 / 40.1	87.7 / 42.0	87.7 / 42.8	87.4 / 40.7	87.3 / 39.8
Group 4 (Fr, Id, Mg, Ms, Th, Vi)								
en→fr	88.3 / 51.1	86.4 / 41.2	85.3 / 38.3	88.3 / 48.9	88.0 / 49.0	88.7 / 51.8	88.0 / 46.1	
en→id	91.2 / 46.4	89.0 / 35.6	90.0 / 38.7	91.2 / 42.9	91.1 / 43.5	91.8 / 48.0	90.8 / 40.0	
en→mg	81.6 / 17.7	56.8 / 2.4	81.1 / 16.1	31.0 / 0.3	41.4 / 0.8	81.8 / 16.8	76.4 / 10.0	
en→ms	89.1 / 41.6	87.4 / 32.5	86.3 / 30.7	87.3 / 22.2	87.2 / 26.7	89.7 / 42.0	87.8 / 30.8	
en→th	84.3 / 5.3	84.8 / 3.7	86.5 / 9.8	61.0 / 0.7	63.2 / 6.1	87.4 / 11.6	86.5 / 10.0	
en→vi	88.0 / 41.8	86.0 / 34.9	85.6 / 31.9	89.0 / 40.3	89.2 / 40.4	89.4 / 43.9	88.5 / 38.7	
fr→en	86.6 / 38.1	88.7 / 41.6	88.6 / 41.2	89.4 / 45.3	89.5 / 47.0	89.6 / 47.8	89.3 / 44.7	89.3 / 45.1
id→en	84.5 / 34.3	89.0 / 40.8	88.4 / 38.8	89.5 / 44.1	89.8 / 45.7	89.6 / 47.3	89.1 / 43.0	89.2 / 43.0
mg→en	63.3 / 13.5	76.0 / 19.6	79.8 / 27.7	47.0 / 1.5	54.1 / 5.3	81.9 / 30.1	74.8 / 19.6	70.6 / 13.8
ms→en	82.1 / 31.4	88.6 / 41.3	87.8 / 39.0	87.3 / 40.0	88.7 / 43.9	89.1 / 46.9	87.7 / 40.8	87.1 / 39.7
th→en	85.9 / 26.8	87.7 / 28.2	85.8 / 26.9	78.1 / 15.2	83.6 / 23.5	88.0 / 32.3	87.6 / 29.9	87.3 / 28.2
vi→en	84.1 / 31.6	87.2 / 33.7	86.6 / 33.6	87.6 / 37.2	87.8 / 38.9	87.9 / 39.8	87.2 / 37.0	87.2 / 36.4

Table 25: Full results for Group 1-4 languages on Flores-200 benchmark. The performance of baselines is directly sourced from Xu et al. (2024b) and we keep the generation configuration of our approach the same as those.

Direction	NLLB-3.3B	LLaMAX3-Alpaca-8B	Aya-101	Aya-23-8B	Aya-23-35B	X-ALMA-13B (only SFT)	X-ALMA-13B-DAT (Ours)	X-ALMA-13B-DATM (Ours)
				Group 5 (Cs, El, Hu, Lt, Lv, Pl)				
en→cs	91.0 / 32.2	88.1 / 24.6	90.0 / 26.7	91.1 / 30.5	91.4 / 32.2	91.5 / 33.8	91.3 / 30.9	
en→el	89.0 / 27.4	86.2 / 20.4	86.6 / 21.4	89.5 / 26.1	89.6 / 27.0	89.8 / 27.9	89.0 / 23.6	
en→hu	89.3 / 26.4	86.6 / 18.2	88.4 / 21.4	51.7 / 3.6	77.0 / 10.8	90.4 / 27.0	89.6 / 21.9	
en→lt	89.3 / 25.2	86.1 / 17.0	89.2 / 22.5	65.4 / 5.4	82.5 / 14.0	91.3 / 28.4	90.0 / 22.5	
en→lv	87.4 / 25.0	85.8 / 21.1	88.6 / 25.0	36.5 / 1.5	62.7 / 7.9	90.7 / 29.3	89.8 / 24.6	
en→pl	88.9 / 21.6	86.7 / 17.2	87.6 / 18.3	89.2 / 20.7	89.8 / 22.4	90.1 / 23.3	89.7 / 20.3	
cs→en	80.1 / 29.4	88.1 / 37.5	87.6 / 35.6	88.5 / 40.7	88.5 / 42.3	89.0 / 43.3	88.6 / 41.0	88.6 / 40.7
el→en	86.1 / 33.0	87.5 / 34.2	86.5 / 32.1	87.8 / 36.1	88.3 / 39.0	87.9 / 38.0	87.5 / 35.6	87.5 / 35.6
hu→en	70.1 / 14.0	87.8 / 32.5	86.4 / 29.9	81.1 / 23.0	86.5 / 32.2	88.7 / 37.3	88.0 / 34.6	88.2 / 35.0
lt→en	67.1 / 12.6	86.0 / 31.0	85.8 / 30.2	80.6 / 24.6	85.4 / 32.9	87.1 / 36.9	86.5 / 33.8	86.3 / 32.8
lv→en	68.1 / 10.4	87.0 / 32.7	86.3 / 32.0	73.4 / 14.1	83.3 / 29.1	87.9 / 38.2	87.6 / 36.0	87.6 / 36.3
pl→en	77.8 / 20.3	85.6 / 28.3	85.6 / 28.0	86.1 / 30.5	86.7 / 33.4	86.5 / 32.8	86.1 / 31.1	86.2 / 31.3
				Group 6 (Et, Fi, Ja, Ka, Ko, Zh)				
en→et	90.5 / 25.0	87.7 / 18.1	90.7 / 21.9	40.5 / 1.5	57.8 / 6.1	91.6 / 26.4	90.8 / 21.9	
en→fi	91.7 / 24.1	89.3 / 17.5	90.3 / 18.9	51.9 / 2.4	70.0 / 8.1	92.7 / 25.3	92.3 / 22.8	
en→ja	87.9 / 22.6	89.0 / 27.5	89.0 / 27.3	90.8 / 30.7	91.0 / 30.9	91.2 / 34.6	90.4 / 31.9	
en→ka	84.6 / 14.8	78.6 / 9.6	85.3 / 11.3	43.3 / 0.4	47.6 / 2.0	87.6 / 14.0	87.0 / 12.6	
en→ko	88.4 / 12.5	85.6 / 8.8	87.4 / 10.2	89.0 / 13.1	89.4 / 12.8	89.3 / 15.0	88.1 / 10.1	
en→zh	82.0 / 32.4	85.6 / 36.3	82.4 / 27.3	87.3 / 40.2	87.5 / 37.3	88.2 / 43.6	87.5 / 40.0	
et→en	62.5 / 7.2	88.3 / 33.6	87.7 / 32.5	74.9 / 15.4	84.2 / 28.9	89.2 / 38.2	88.8 / 35.8	88.4 / 34.7
fi→en	67.7 / 10.2	89.3 / 31.6	88.6 / 29.7	81.3 / 20.4	87.3 / 29.8	90.0 / 36.0	89.6 / 34.2	89.7 / 34.0
ja→en	79.5 / 17.2	87.5 / 24.6	86.5 / 23.5	87.9 / 28.1	88.4 / 30.4	88.1 / 28.9	87.6 / 26.9	87.5 / 26.9
ka→en	84.8 / 25.6	50.7 / 1.2	84.5 / 25.6	60.1 / 3.6	79.4 / 19.4	86.8 / 28.4	86.4 / 27.0	86.3 / 26.9
ko→en	84.9 / 26.2	87.5 / 26.3	87.0 / 26.5	88.0 / 29.4	88.7 / 32.2	88.1 / 30.6	87.5 / 28.4	87.5 / 28.1
zh→en	77.1 / 16.8	86.6 / 25.9	84.5 / 23.1	87.1 / 29.4	87.6 / 32.2	87.1 / 30.4	87.7 / 28.3	86.6 / 28.2
				Group 7 (Gu, Hi, Mr, Ne, Ur)				
en→gu	87.2 / 24.3	82.7 / 13.7	83.9 / 15.6	65.7 / 0.4	62.2 / 1.5	88.2 / 25.0	86.7 / 18.2	
en→hi	80.9 / 34.4	76.6 / 23.5	75.5 / 21.4	79.3 / 25.0	79.1 / 26.0	81.4 / 34.3	79.2 / 26.3	
en→mr	74.3 / 17.1	69.5 / 10.1	69.5 / 10.3	66.7 / 0.9	61.1 / 1.3	75.9 / 18.0	74.2 / 13.6	
en→ne	76.5 / 16.4	78.4 / 10.7	77.5 / 10.5	69.2 / 1.5	68.3 / 1.4	84.0 / 21.5	82.7 / 16.8	
en→ur	81.3 / 22.9	75.6 / 13.4	74.6 / 13.9	63.6 / 0.3	39.1 / 2.4	83.5 / 23.8	82.1 / 18.9	
gu→en	90.2 / 42.3	66.0 / 9.9	82.3 / 28.0	53.6 / 3.4	63.1 / 8.8	90.1 / 40.4	87.0 / 30.8	85.9 / 29.5
hi→en	88.9 / 38.7	88.9 / 35.4	87.5 / 34.6	89.1 / 37.6	89.6 / 40.1	89.8 / 43.0	88.8 / 36.4	88.8 / 36.0
mr→en	87.0 / 34.0	87.3 / 30.6	85.2 / 30.1	68.9 / 7.5	79.9 / 18.4	88.5 / 37.7	87.4 / 32.7	87.4 / 32.3
ne→en	89.7 / 38.0	89.3 / 32.9	84.9 / 31.2	77.0 / 10.0	84.1 / 23.3	90.6 / 41.2	89.3 / 33.9	89.4 / 33.9
ur→en	86.0 / 31.6	86.5 / 30.5	83.9 / 28.1	70.2 / 9.3	80.2 / 21.1	87.7 / 36.4	86.5 / 31.5	86.6 / 31.9
				Group 8 (Ar, Az, Fa, He, Kk, Ky, Tr, Uz)				
en→ar	86.3 / 27.5	82.2 / 14.1	84.1 / 17.2	87.3 / 26.5	87.1 / 27.4	87.8 / 29.1	86.4 / 20.1	
en→az	86.9 / 14.0	80.0 / 7.3	85.6 / 11.5	75.5 / 2.0	67.2 / 3.0	88.2 / 14.0	86.9 / 11.9	
en→fa	86.5 / 22.6	84.5 / 17.7	86.4 / 19.1	87.7 / 23.2	87.6 / 23.8	88.5 / 28.4	86.7 / 22.0	
en→he	87.8 / 30.4	86.2 / 23.3	85.4 / 20.5	88.3 / 27.0	88.2 / 28.9	89.6 / 32.7	87.7 / 25.1	
en→kk	90.0 / 20.6	86.0 / 12.7	89.0 / 17.2	71.0 / 1.2	45.0 / 0.7	90.7 / 22.2	89.4 / 17.3	
en→ky	88.1 / 13.2	82.9 / 7.9	86.6 / 10.4	62.9 / 1.2	49.6 / 0.9	88.5 / 13.2	87.4 / 11.5	
en→tr	89.7 / 29.0	84.2 / 13.8	88.3 / 21.1	88.9 / 23.7	88.7 / 23.6	90.3 / 27.7	88.5 / 21.6	
en→uz	89.8 / 18.6	74.5 / 6.8	88.6 / 12.0	46.5 / 0.5	37.1 / 0.3	90.0 / 16.8	86.8 / 11.0	
ar→en	86.1 / 38.2	86.8 / 35.1	85.8 / 35.0	87.9 / 41.5	87.6 / 43.4	87.5 / 41.2	86.6 / 38.3	86.5 / 36.9
az→en	77.5 / 15.1	70.0 / 7.9	85.2 / 21.5	75.6 / 10.6	82.6 / 17.9	86.7 / 25.8	85.7 / 22.2	86.1 / 23.1
fa→en	83.5 / 29.8	87.6 / 33.1	87.3 / 32.8	87.9 / 36.8	88.5 / 39.6	88.1 / 37.6	87.2 / 33.7	87.4 / 34.2
he→en	86.0 / 39.1	87.2 / 39.5	86.4 / 37.9	88.4 / 43.2	88.9 / 46.6	88.3 / 44.5	87.8 / 41.9	87.6 / 40.9
kk→en	85.0 / 30.2	86.7 / 29.0	86.1 / 29.2	59.9 / 3.5	74.0 / 14.2	87.8 / 33.5	87.0 / 30.4	87.0 / 29.4
ky→en	81.6 / 20.1	84.5 / 20.4	83.0 / 20.4	64.3 / 4.2	74.3 / 11.3	85.4 / 23.5	84.5 / 21.0	84.6 / 21.5
tr→en	75.3 / 16.8	88.6 / 33.4	88.1 / 33.2	88.2 / 35.8	89.6 / 39.3	89.6 / 39.9	88.8 / 35.9	88.5 / 33.0
uz→en	60.7 / 5.3	86.1 / 27.9	84.9 / 28.1	61.3 / 3.9	75.9 / 15.3	86.9 / 32.2	84.2 / 25.3	83.5 / 24.1

Table 26: Full results for Group 5-8 languages on Flores-200 benchmark. The performance of baselines is directly sourced from Xu et al. (2024b) and we keep the generation configuration of our approach the same as those.