Cross-Connected Mixture-of-Expert LoRA for Multilingual Neural Machine Translation

Anonymous ACL submission

Abstract

Generative Large Language Models (LLMs) and the associated pre-training & fine-tuning paradigms have achieved significant advancements in various NLP tasks. However, Multilingual Neural Machine Translation (MNMT) systems encounter capacity constraints when scaling to numerous languages with fixed model size, resulting in degraded translation quality, particularly for supervised tasks. Furthermore, the scarcity of parallel corpora for non-English language pairs limits expansion to new translation directions. This paper presents CrossLoRA, a novel MNMT framework that combines Low-Rank Adaptation (LoRA) with a Mixture-of-Experts (MoE) architecture featuring cross-connected language-specific experts. Our approach establishes dedicated experts for individual languages while enabling strategic interaction between source and target language experts during the translation process. To achieve any-to-any translation capability, we tailor a two-staged fine-tuning paradigm for CrossLoRA framework with a self-contrastive semantic enhancement, fine-tuning using English as the pivot language, followed by pseudo-corpus generation and subsequent fine-tuning with the generated data. Experimental results on multilingual translation datasets confirm the quality improvement and parameter efficiency of CrossLoRA framework. Our findings provide an effective recipe for fine-tuning LLMs to achieve any-to-any translation capability. Our code is available at: `https://anonymous.4open.science/r/CrossL-3FBF/`.

1 Introduction

Recently the emergence of various generative Large Language Models (LLMs) (OpenAI et al., 2024; Grattafiori et al., 2024; Qwen et al., 2025) has significantly advanced numerous NLP tasks, including Multilingual Neural Machine Translation (MNMT) (Bahdanau et al., 2015). By integrating prompt engineering methods with pre-training and fine-tuning paradigms (Zhang et al., 2023a), as illustrated in Figure 1(a), conventional LLMs can fully leverage their translation capabilities. The superior performance of LLMs in translation is primarily attributed to their billions of trainable parameters (Xu et al., 2024b), while fully fine-tuning these models demands substantial computing resources, limiting practical applications (Zhang et al., 2024). To address this challenge, Parameter-Efficient Fine-Tuning (PEFT) methods (Han et al., 2024), such as Low-Rank Adaptation (LoRA) (Hu et al., 2022) shown in Figure 1(b), enable smaller models (e.g., 7B parameters) to gain significant improvements on MNMT tasks in computationally efficient settings (Zhang et al., 2023b; Chen et al., 2024a).

Despite these methods facilitating a balance between high-quality translation and manageable computational costs, challenges persist in fine-tuning MNMT tasks. The limited availability of parallel corpora for non-English-centric pairs constrains model capabilities, impeding expansion to additional directions through supervised fine-tuning approaches (Guzmán et al., 2019; Ranathunga et al., 2023). Additionally, in multilingual scenarios, the generalization capability of simple LoRA adapters is limited. While introducing Mixture-of-Experts (MoE) framework is an effective solution for enhancing model generalization (Shazeer et al., 2017; Lepikhin et al., 2021), this approach suffers from routing fluctuations when the number of experts is limited (Dai et al., 2022). Even with Mixture-of-LoRAs (MoLoRA) framework (Zadouri et al., 2024; Zhu et al., 2023), which combines MoE and LoRA as illustrated in Figure 1(c), scaling the number of experts or assigning specific experts to different translation directions becomes computationally prohibitive in scenarios with numerous translation directions.

To address the above-mentioned issues, we propose a novel framework for MNMT which

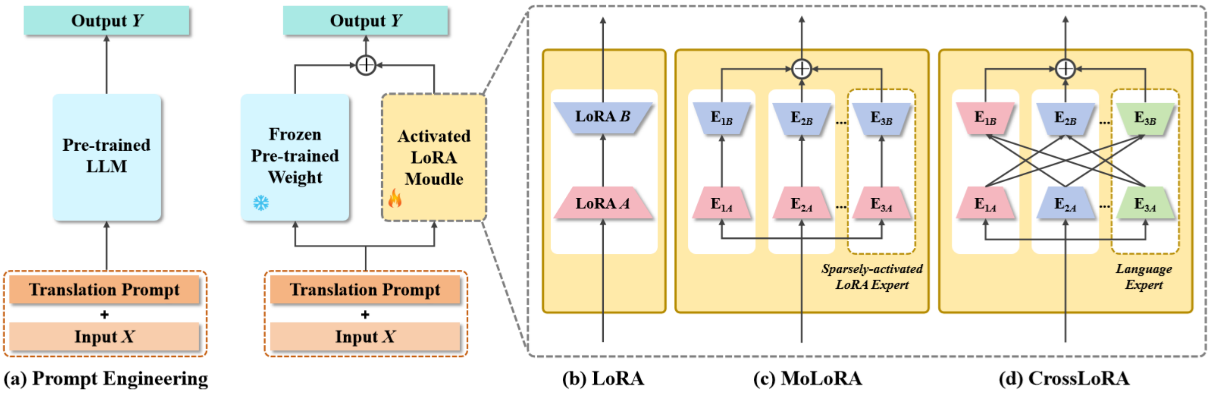

Figure 1: Illustrations of (a) prompt engineering method for NMT, compared with LLM fine-tuning process with (b) LoRA, (c) MoLoRA framework, and (d) our proposed CrossLoRA framework.

cross-connects experts within the MoE structure, named **CrossLoRA**. Specifically, instead of training experts specialized in particular translation directions, each LoRA expert is designated as a language-specific expert. The *LoRA A* and *LoRA B* modules correspond to the source and target sides of the translation process, respectively. Dedicated cross-connected activations between experts facilitate translation between two distinct languages, as shown in Figure 1(d). Combined with the static language router, the number of experts required to support diverse translation directions can be significantly reduced. Additionally, we tailor a two-stage fine-tuning process to enable efficient translation in multilingual language directions, as seen in Figure 2. In the first fine-tuning stage, a pivot language (e.g., English) serves as the "hub" language to establish translations from and to all target languages. Following this initial fine-tuning, pseudo-corpora for any-to-any translation directions are generated using the first-stage fine-tuned LoRA modules. By consolidating these corpora, we facilitate a second fine-tuning stage to achieve comprehensive any-to-any translation capability. We further employ the self-contrastive learning method to enhance the robustness and semantic representation capability in translations. Consequently, the required data for non-English language pairs are significantly reduced, allowing the fine-tuned LLMs to achieve promising outcomes in terms of both language coverage and translation quality. The main contributions of this paper can be summarized as follows:

- We introduce the CrossLoRA framework for fine-tuning LLMs on multilingual translation tasks. By incorporating cross-connected language-specific experts alongside the static language routers, the proposed framework enables the fine-tuned model to achieve broad language coverage and precise translation, even with a limited number of experts.

- Based on the CrossLoRA framework, we design a two-stage fine-tuning process with sequential cross-connected activations, allowing LLMs to perform any-to-any language translation without being constrained by the limitations of multilingual corpora.

- Extensive evaluations across LLMs demonstrate that our approach achieves superior quality improvements with computational efficiency, enabling fine-tuned general-purpose LLMs to outperform specialized NMT models in multilingual translation tasks.

2 Related Works

2.1 Sparse Mixture-of-Experts

Sparse expert models have gained prominence for enhancing model capacity while maintaining computational efficiency (Fedus et al., 2022a). The MoE framework, initially designed to overcome scalability limitations of monolithic models (Shazeer et al., 2017), has become a cornerstone in deep learning for tasks requiring task-specific specialization (Chen et al., 2022). In the Transformer architecture (Shazeer et al., 2018), MoE is widely adopted in Multi-Task Learning (MTL) and has been integrated into LLMs to address diverse NLP tasks (Wang et al., 2023; Fedus et al., 2022b).

The combination of MoE and LoRA has further advanced parameter-efficient fine-tuning. MoLoRA (Zadouri et al., 2024), a pioneering approach for resource-constrained environments, combines MoE with LoRA to improve task adaptability. Subsequent studies have extended this framework by introducing task-adaptive gating mechanisms (Liu et al., 2024), addressing data conflicts in instruction datasets (Chen et al., 2024b), and mitigating knowledge forgetting through localized balancing constraints (Dou et al., 2024).

In multilingual translation, MoE-based methods such as MoE-LGR (Li et al., 2023) leverage linguistic typology to group languages, while smoothed gating networks with token-level feature mixing (Liu et al., 2022) enhance language-specific feature extraction. However, challenges persist in balancing computational overhead and performance, particularly when scaling to diverse language pairs with limited experts (Tourni and Naskar, 2024).

2.2 LLM-Based Multilingual Translation

Generative LLMs are widely used in multilingual translation due to their broad language coverage and robust performance (Yang et al., 2023; Zeng et al., 2024). However, their deployment is constrained by high computational costs and reliance on large-scale parallel corpora (Zhang et al., 2023a). To address these challenges, researchers employ two strategies: Parameter-efficient fine-tuning methods like LoRA (Xu et al., 2024a,b) and adapters (Stickland et al., 2021) reduce the number of trainable parameters while maintaining performance, and data synthesis techniques such as pseudo-corpus generation (Pan et al., 2024) and data augmentation (Liu et al., 2023; Lu et al., 2024) alleviate data scarcity in low-resource settings. Despite these advancements, existing approaches still struggle with arbitrary language pair translation and computational efficiency. In this paper, the proposed CrossLoRA framework aimed at simultaneously addressing both computational efficiency and data scarcity issues in multilingual translation.

3 Methodology

3.1 Preliminaries

In this subsection, we briefly introduce the Low-Rank Adaptation (LoRA) method, as depicted in Figure 1(b), followed by the Mixture-of-LoRAs (MoLoRA) framework based on LoRA method.

When employing the LoRA adapter, the pre-trained model's weight matrix W_0 is kept frozen, while a trainable low-rank decomposition matrix ΔW, which can be further decomposed into the paired *LoRA A* and *LoRA B* modules, is incorporated into the selected linear layer of the model framework. The update can be formulated as follows:

$$y = (\Delta W + W_0)x = (BA + W_0)x \quad (1)$$

Here, $A \in R^{r \times d_i}$ and $B \in R^{d_o \times r}$ represent the coordinated low-rank matrices corresponding to *LoRA A* and *LoRA B* modules respectively, with $r \ll min(d_i, d_o)$ refers to the selected LoRA rank. x denotes the input sequence, and y is the corresponding output. Given that only the low-rank matrices A and B get updated, the LoRA method significantly reduces the number of parameters required for downstream fine-tuning.

Building upon the LoRA framework, the MoLoRA method further integrates the MoE framework. As illustrated in Figure 1(c), the structure of a MoLoRA component comprises a set of n LoRA experts, denoted as $E_1, E_2, \ldots, E_n$, which are tasked with adapting the pre-trained layer during the fine-tuning stage. Each expert E_i can be further comprised into two trainable low-rank weight matrices, E_{iA} and E_{iB}, which relate to the previous *LoRA A* and *LoRA B* modules respectively. In addition, the MoLoRA module includes a token-level expert router denoted as θ^{MoL} for computing routing weight. The routing weight s_i^{MoL} related to expert E_i is computed by the equation below:

$$s_i^{MoL} = \theta^{MoL}(x)_i = softmax(W^{MoL}x)_i, \quad (2)$$

where W^{MoL} represents the weight matrix of the router. The final output y for integrating n experts in the module is calculated as follows:

$$y = W_0 x + \sum_{i=1}^{n} s_i^{MoL} E_{iB} E_{iA} x \quad (3)$$

3.2 Cross-Connected Language Experts

We modify MoLoRA and introduce the CrossLoRA framework in Figure 1(d), which crossly connects the experts in the MoE structure. Each expert in the MoE structure is regarded as an expert of one specific language, and the cross-connected experts act as the translation between two distinct languages. For the languages involved in the translation task, a specific LoRA expert is assigned for each language. To enhance clarity, we can consider a simplified scenario involving a restricted set of three languages: German, English, and Chinese, as is shown in Figure 2, where the language experts are labeled as De, En, and Zh.

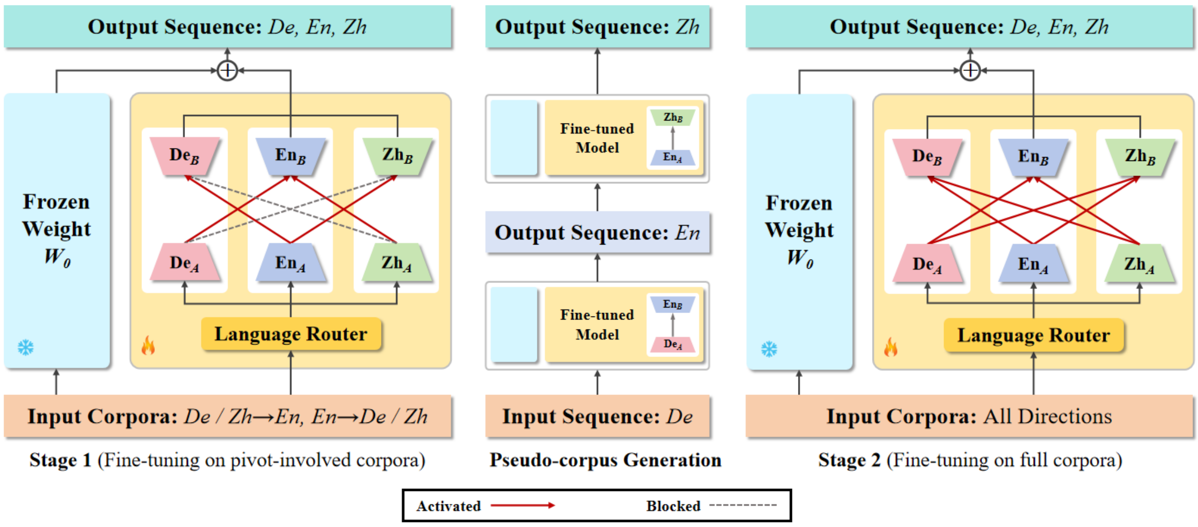

Figure 2: Our CrossLoRA fine-tuning process. The figure depicts an example involving a limited set of three languages (De, En, and Zh) for MNMT task. Within the CrossLoRA module, a red pathway represents a route initially activated in Stage 1 or Stage 2 respectively, while the gray dashed line denotes a blocked route.

Each expert, such as De, is further decomposed into two low-rank weight matrices, denoted as De_A and De_B modules. These modules correspond to scenarios where German is designated as either the source or the target language, respectively, which is activated only when involved in the specific translation process. More specifically, considering a specific translation direction De⇒En, when German is set as the source language and English is set as the target language for translating a sequence pair, only De_A as well as En_B can be activated while the remaining experts stay frozen. To ensure the accurate activation of the corresponding source and target language low-rank weight matrices when translating a sequence pair, we deployed a static language router that outputs the corresponding routes based on pre-set language labels.

For a specific case of translating a sequence x from source language x_{src} into the target language x_{tgt}, the target output y can be calculated by the following formula:

$$y = W_0 x + \sum_{i=1}^{n} \sum_{j=1}^{n} f(x; i, j) E_{jB} E_{iA} x, \quad (4)$$

where n is the number of language experts, and $f(x; i, j)$ is the gating function of the static router:

$$f(x; i, j) = \begin{cases} 1 & \text{if } i = x_{src} \text{ and } j = x_{tgt} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

In this way, only the corresponding low-rank weight matrices in each expert module are properly activated.

3.3 Staged Fine-Tuning on CrossLoRA

To achieve any-to-any translation on the CrossLoRA framework with limited parallel training data, we further tailor a staged fine-tuning process, as illustrated in Figure 2, which can be outlined as follows:

Stage 1. Firstly, CrossLoRA module is trained on English-centric corpora. For the case illustrated in the figure, assume only three language experts are involved: German (De), English (En) and Chinese (Zh). With the increasing diversity of languages, non-English language pairs often exhibit limited or non-existent parallel text resources. To address this issue, English is designated as the pivot language. The primary objective of Stage 1 fine-tuning is to enhance the model's translation capabilities in both En⇒Any and Any⇒En directions, thereby augmenting the model to generate high-quality pseudo-corpora. Thus, the training data used for training in this stage includes translation corpora with English as the source language (En⇒De, En⇒Zh) and translation corpora with English as the target language (De⇒En, Zh⇒En), as shown by the red solid arrows in Figure 2 (Stage 1). For translation directions not involving English (De⇒Zh and Zh⇒De), the corresponding routes among matrices remain inactive, as shown by the gray dashed lines. On the other hand, the weight matrices of all experts in the CrossLoRA module are updated, thereby enhancing the model's ability to understand all English-involved translation directions.

Pseudo-corpus Generation. English is regarded as the pivot language for creating the pseudo corpus required for the subsequent training stage. For instance, to obtain parallel pseudo corpora for the De⇒Zh translation, we employ CrossLoRA fine-tuned after Stage 1 to translate the German sequence into English, followed by translating the English sequence into Chinese. Parallel corpora for the Zh⇒De translation are obtained in a similar manner. All generated corpora undergo language identification to ensure accuracy and avoid off-target translations. Theoretically, we can obtain parallel pseudo corpora in any translation direction among the languages involved in the translation model.

Stage 2. In the final stage, the reinitialized CrossLoRA model is fine-tuned using both the training data from Stage 1 and all the previously generated parallel pseudo-corpora. All routes are now activated, which enables each language expert to be applicable across all source and target languages. This comprehensive approach enhances the model's translation efficacy in all directions, ensuring optimized performance regardless of the specific language pair.

3.4 Self-Contrastive Semantic Enhancement

In the translation task, given a labeled sequence pair (x_j, y_j) in the parallel training corpora $D\left\{(x_j, y_j)\right\}_{j=1}^{M}$, where x_j and y_j represent the source and target sequence, respectively. The training objective for the translation model is to minimize the following Negative Log-Likelihood (NLL) loss function:

$$\mathcal{L} = -\frac{1}{M}\sum_{j=1}^{m}\log\mathcal{P}^{w}\left(y_j \mid x_j;\theta\right) \tag{6}$$

where θ is the set of trainable parameters. To further improve regularization capability, we take R-Drop (Liang et al., 2021) to reduce the inconsistency existing in training and inference. Due to the dropout mechanism of randomly deactivating units within a model, each forward pass effectively utilizes distinct sub-models. Consequently, we input x_j through two separate forward passes of the network to obtain two distributions of model predictions, denoted as $\mathcal{P}_1^{w}\left(y_j \mid x_j\right)$ and $\mathcal{P}_2^{w}\left(y_j \mid x_j\right)$. In each training step, the R-Drop method seeks to regularize the model's predictions by minimizing the bidirectional Kullback-Leibler (KL) divergence between the two output distributions for the same sample, and the corresponding KL-divergence loss is formulated as:

$$\begin{aligned}\mathcal{L}_{kl} = \frac{1}{2M}\sum_{j=1}^{m}&\left(\mathcal{D}_{kl}\left(\mathcal{P}_1^{w}\left(y_j \mid x_j\right)\|\mathcal{P}_2^{w}\left(y_j \mid x_j\right)\right)\right.\\ &\left.+\mathcal{D}_{kl}\left(\mathcal{P}_2^{w}\left(y_j \mid x_j\right)\|\mathcal{P}_1^{w}\left(y_j \mid x_j\right)\right)\right)\end{aligned} \tag{7}$$

With these two forward passes, the original learning objective is reformulated as a bidirectional NLL loss:

$$\begin{aligned}\mathcal{L}_{nll} = -\frac{1}{2M}\sum_{j=1}^{n}&\left(\log\mathcal{P}_1^{\omega}\left(y_j \mid x_j\right)\right.\\ &\left.+\log\mathcal{P}_2^{\omega}\left(y_j \mid x_j\right)\right)\end{aligned} \tag{8}$$

Finally, the CrossLoRA model can be optimized by minimizing a composite loss function that incorporates both the modified NLL loss and the contrastive loss:

$$\mathcal{L}_{\text{Reg}} = \mathcal{L}_{nll} + \alpha\cdot\mathcal{L}_{kl} \tag{9}$$

where α is the coefficient weight to control the proportion of KL-divergence loss.

4 Experiments

4.1 Dataset and Metrics

For our parallel training data, we utilize the training set of the OPUS-100 dataset (Tiedemann, 2012), an English-centric multilingual corpus, along with the development set of Flores-200 dataset (NLLB Team et al., 2022). Following the ALMA model's configuration (Xu et al., 2024a), we select six languages—English (En), German (De), Chinese (Zh), Russian (Ru), Czech (Cs) and Icelandic (Is)—with English serving as the pivot language. To comprehensively evaluate the model's translation performance, we test all 30 directions. Given the lack of non-English-centric test data in OPUS-100, our experiment's test data comprises test sets from OPUS-100 that involve English and Flores-200 for other translation directions. For Stage 1 training data, we randomly sample 20k parallel sentence pairs for each of the 10 language pairs. For Stage 2 fine-tuning pseudo data, using the fine-tuned model, we generate 20k parallel sentence pairs for each non-English-centric directions. See Appendix A.2 for detailed data settings.

We employ a commonly adopted sentence-level translation prompt template (Hendy et al., 2023), which can be formulated as "*Translate the following* $\{src\}$ *sentences into* $\{tgt\}$*:* ", where $\{src\}$

Models	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
ALMA-7B (English-pivot)	24.45	78.12	17.66	80.79	19.92	79.90
M2M100-12B	24.06	74.59	18.98	82.52	20.68	79.88
BigTranslate-13B	22.02	72.95	18.94	81.88	19.98	78.90
NLLB-3.3B	27.85	77.01	20.53	82.88	22.97	80.92
`LLaMA-3-8B-Instruct`	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base	18.41	68.61	13.94	77.41	15.43	74.48
+LoRA	26.11	76.43	16.20	79.80	19.50	78.68
+MoLoRA (Top-k)	27.08	77.02	17.45	80.29	20.66	79.20
+MoLoRA (Static)	28.38	77.30	19.14	81.39	22.22	80.03
+CrossLoRA						
— Stage 1	29.50	78.26	13.95	77.35	19.13	77.65
— Stage 2	**29.69**	**78.74**	**20.60**	**81.94**	**23.63**	**80.88**
`Qwen2.5-7B-Instruct`	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Base	19.09	70.37	12.85	76.55	14.93	74.49
+LoRA	25.84	75.92	16.44	79.76	19.57	78.48
+MoLoRA (Top-k)	26.79	76.88	17.67	80.20	20.71	79.09
+MoLoRA (Static)	28.28	77.60	20.31	81.35	22.97	80.10
+CrossLoRA						
— Stage 1	29.14	78.29	12.93	76.58	18.33	77.15
— Stage 2	**29.52**	**78.37**	**21.04**	**81.83**	**23.87**	**80.68**

Table 1: The overall results in all directions. Except for CrossLoRA, which is evaluated across both Stage 1 and Stage 2, all other LoRA-based methods report only Stage 2 outcomes. **Bold results** highlight the highest scores among fine-tuning approaches for the same backbone model, demonstrating that CrossLoRA outperforms all competitors and achieves competitive performance with state-of-the-art multilingual translation systems.

and $\{tgt\}$ denote the respective source and target languages of the specific translation direction. For evaluation metrics, we utilize the SacreBLEU (Post, 2018) and COMET-22 (Rei et al., 2022) to evaluate translation quality.

4.2 Implementation Details

The CrossLoRA framework is applied to state-of-the-art base LLMs, including Qwen2.5-7B-Instruct (Qwen et al., 2025) and LLaMA-3-8B-Instruct (Grattafiori et al., 2024).

During the fine-tuning phase, our setup features a batch size of 32, training for 3 epochs, and a learning rate of 5e-4. The coefficient weight of KL-divergence loss $\alpha = 0.1$. Given the number of languages in translation, the defined number of experts is fixed at 6. For the LoRA configurations, we set the *lora rank* $r = 16$, *lora alpha* $\alpha_l = 64$, *lora dropout* $p = 0.1$.

4.3 Baselines

To ensure a fair evaluation, we compare CrossLoRA with the following LoRA-based methods using identical staged fine-tuning configurations:

- **LoRA**. Scales the *lora rank* and *lora alpha* parameters within a single LoRA adapter, yielding comparable parameter counts.
- **MoLoRA (Top-k)**: We employ MoLoRA adapter with the same number of experts as CrossLoRA alongwith a top-1 router, activating one expert per translation process.
- **MoLoRA**. **Static**: A MoLoRA adapter equipped with a static router, designating specific experts for each language pair, thereby expanding the total number of experts to 30. This configuration ensures consistent expert activation, eliminates routing fluctuations but also substantially increases training costs.

In addition to the aforementioned LoRA-based methods, we compare our model with prior studies that exhibit robust multilingual translation capabilities, specifically **M2M100-12B** (Fan et al., 2021), **BigTranslate** (Yang et al., 2023) and **NLLB-3.3B** (NLLB Team et al., 2022) from the

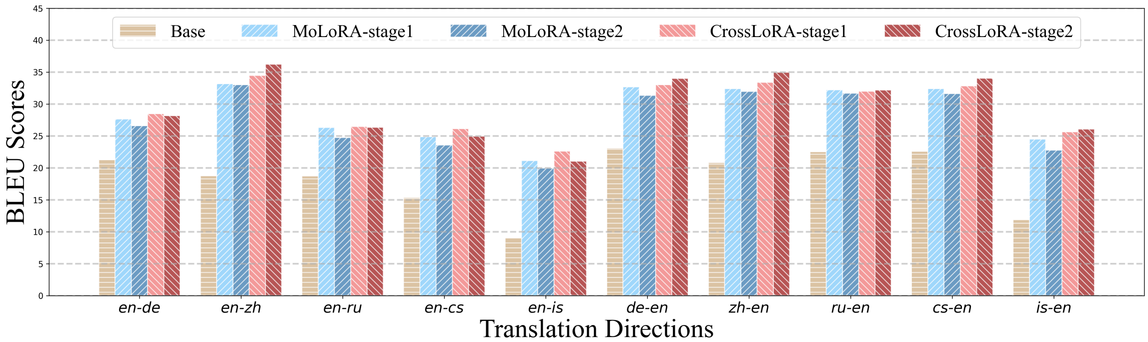

Figure 3: Detailed results of CrossLoRA after Stage 1 & Stage 2 fine-tuning in all translation directions involving English, based on LLaMA-3-8B-Instruct. A comparison is made between MoLoRA (with top-k routing) and CrossLoRA, highlighting that CrossLoRA benefits from multilingual collaborative training in Stage 2, while MoLoRA experiences expert fluctuations when the number of experts is insufficient.

NLLB model family. We also include **ALMA-7B** (Xu et al., 2024a), an English-centric model that employs a staged fine-tuning strategy. Notably, ALMA-7B's performance in non-English-centric directions is evaluated via an English pivot translation pipeline.

4.4 Main Results

We report the overall results across all translation directions in Table 1. In summary, after Stage 2 fine-tuning, the proposed CrossLoRA method outperforms other LoRA-based fine-tuning methods, and the optimal model surpasses previous state-of-the-art translation models.

Compared with backbone LLMs. After Stage 1 fine-tuning, CrossLoRA achieves significant improvements in all directions involving English, while maintaining the translation performance of the backbone model in other directions. Following Stage 2 fine-tuning, CrossLoRA exhibits substantial performance gains across all translation directions relative to the backbone models, particularly for non-English-centric directions.

Compared with LoRA-based fine-tuning methods. CrossLoRA demonstrates a more substantial enhancement compared to all other LoRA-based methods on average, showing marginal improvements in both evaluation metrics. Specifically, MoLoRA with top-k routing exhibits better average performance than pure LoRA fine-tuning, while MoLoRA with static routing achieves comparable performance but at the cost of significantly increased computational overhead. CrossLoRA outperforms both MoLoRA configurations. Additionally, detailed results for English-involved directions during the staged fine-tuning process are shown in Figure 3. After Stage 2 fine-tuning, MoLoRA with top-k routing experiences expert fluctuations when the number of experts is insufficient, leading to a general performance decline in English-involved translation directions. In contrast, CrossLoRA benefits from stronger generalization ability under the same parameters, leveraging multilingual collaborative training to achieve performance improvements in most directions.

Compared with prior studies. Both backbone models fine-tuned with CrossLoRA outperforms previous professional multilingual translation models. Notably, while the ALMA model exhibits strong performance in English-centric translation directions compared to other baselines, its efficacy in non-English-centric directions is markedly constrained by reliance on an English-pivot pipeline-based approach for any-to-any translation. CrossLoRA's distinct advantage lies in its ability to minimize dependency on large-scale non-English parallel corpora, which were traditionally deemed essential for robust multilingual translation. This highlights its parameter-efficient design without compromising translation quality.

5 Ablation Studies

Beyond the main results, we further explore the CrossLoRA framework with diverse configurations to deepen our understanding. All experiments are conducted on LLaMA-3-8B-Instruct.

5.1 Fine-tuning Data Configuration

To evaluate the fine-tuning data configuration, we conduct ablation experiments with two additional fine-tuning configurations. As shown in Table 2, **Pseudo-corpora + Stage1** involves fine-tuning the

Methods	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Pseudo-corpora + Stage 1	28.99	78.50	20.37	81.51	23.24	80.51
All + Stage 1	27.78	77.57	20.44	81.43	22.89	80.14
All + Reinitialized	**29.69**	**78.74**	**20.60**	**81.94**	**23.63**	**80.88**

Table 2: The ablation study on the fine-tuning data configurations for Stage 2, based on LLaMA-3-8B-Instruct. The best scores are marked in **bold**. The newly fine-tuned CrossLoRA model achieves the best overall performance.

Methods	English-centric		non-English-centric		Average		Trainable
	BLEU	COMET	BLEU	COMET	BLEU	COMET	Parameters
1 Shared Source Expert	28.78	77.57	19.56	81.11	22.63	79.93	1.24%
1 Shared Target Expert	28.61	77.60	19.41	80.97	22.48	79.85	
3 Experts	29.32	78.66	20.22	81.89	23.25	80.81	1.06%
6 Experts	**29.69**	**78.74**	**20.60**	**81.94**	**23.63**	**80.88**	2.08%

Table 3: The ablation study on the merged language experts, based on LLaMA-3-8B-Instruct model. The best scores are marked in **bold**.

Stage 1 checkpoint using only generated pseudo-corpora. **All + Stage1** uses the same checkpoint but includes both pseudo-corpora and English-pivot corpora from Stage 1. The main experiment adopts the **All + Reinitialized** setup, which fine-tunes a reinitialized CrossLoRA model using both pseudo-corpora and English-pivot corpora from Stage 1.

The results indicate that the reinitialized CrossLoRA network, when trained with combined corpora, achieves the overall best performance. In contrast, Stage 1 checkpoint-based models exhibit knowledge forgetting, improving new directions while degrading English-centric translations. The reinitialized model avoids this issue by synergistically learning language features across all data, achieving consistent gains across translation directions as the optimal configuration.

5.2 Merged Language Experts

Exploring the application of expert compression techniques within the CrossLoRA framework is crucial for further improving parameter efficiency. Thus, we conduct experiments using two distinct expert compression strategies:

Shared Source & Target Side Language Expert. Building on HydraLoRA's asymmetric MoE design (Tian et al., 2024), we test configurations where a single merged expert handles all source inputs or target outputs. This approach enables shared parameterization between source and target sides to minimize redundancy.

Language Group Experts. Drawing from the integration of language typology in MoE-based translation systems (Li et al., 2023), we merge languages into typologically grouped experts (see Table 4). For example, English, German, and Icelandic share one expert. This reduces the total expert count and trainable parameters by half (from 2.08% to 1.06%), while preserving CrossLoRA's architecture.

The results, presented in Table 3, indicate that despite a significant reduction in the number of trainable parameters required for fine-tuning, the merged language group expert configuration only experiences a slight decrease in overall performance. This suggests that CrossLoRA can be efficiently scaled to support more languages while preserving translation quality, offering promising potential for future multilingual extensions.

6 Conclusion

In this paper, we propose a novel CrossLoRA framework designed for fine-tuning LLMs on downstream multilingual translation tasks. The proposed approach integrates the LoRA technique with the MoE framework, deploying transactional language experts. Building upon this foundation, we tailor a staged training approach that enables the model to acquire the capability for any-to-any translation with a limited training corpus. Experiments conducted across various translation directions have proven the effectiveness and parameter efficiency of CrossLoRA.

For future work, we plan to conduct more in-depth research on the CrossLoRA architecture, which includes expanding the range of supported languages and investigating the impact of pseudo-corpus size & quality on model performance.

Limitations

While this article presents an efficient framework for fine-tuning LLMs on multilingual translation task, several limitations warrant further investigation:

Language Coverage Constraints. Although CrossLoRA mitigates dependence on non-English parallel data, our experiments are constrained to 6 languages (including one low-resource language: Icelandic). While Section 5.2 demonstrates its theoretical scalability via linguistic expert ablations, systematic evaluation is required to validate its capability under expanded conditions. Key questions remain: (1) Can translation quality be balanced across a significantly larger set of languages? (2) How does the framework perform when integrating additional low-resource languages?

Diversified Training Process. This work focuses on supervised fine-tuning of LLMs utilizing parallel corpora. However, recent advances in translation enhancement include continual pre-training with monolingual data (Xu et al., 2024a) and preference learning approaches (Xu et al., 2024b). Further exploration of integrating more methods with CrossLoRA is essential for enhancing its adaptability to diverse training paradigms.

Model Diversity Constraints. The proposed CrossLoRA framework is evaluated on LLaMA-3-8B-Instruct and Qwen2.5-7B-Instruct, which demonstrate strong performance but restrict generalization insights across diverse architectures and scales. Future research should investigate its effectiveness on models with varying capabilities to validate robustness and adaptability beyond current baselines.

Pseudo-Corpus Generation Optimization. While we employs synthetic pseudo-corpora for training, current rule-based filtering strategies struggle to guarantee high-quality data generation. Additionally, integrating quality assessment models introduces computational overhead, limiting scalability. Given that high-quality training corpora directly impact model performance, it is worthwhile to explore efficient pseudo-corpus generation paradigms that balance data quality and resource efficiency.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Kaidi Chen, Ben Chen, Dehong Gao, Huangyu Dai, Wen Jiang, Wei Ning, Shanqing Yu, Libin Yang, and Xiaoyan Cai. 2024a. General2specialized llms translation for e-commerce. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024*.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024b. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *Preprint*, arXiv:2401.16160.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. 2022. Towards understanding the mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Stablemoe: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Loramoe: Alleviate world knowledge forgetting in large language models via moe-style plugin. *Preprint*, arXiv:2312.09979.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*.

William Fedus, Jeff Dean, and Barret Zoph. 2022a. A review of sparse expert models in deep learning. *Preprint*, arXiv:2209.01667.

William Fedus, Barret Zoph, and Noam Shazeer. 2022b. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

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 *The Tenth International Conference on Learning Representations, ICLR 2022*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021*.

Shangjie Li, Xiangpeng Wei, Shaolin Zhu, Jun Xie, Baosong Yang, and Deyi Xiong. 2023. Mmnmt: Modularizing multilingual neural machine translation with flexibly assembled moe and dense blocks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*.

Junpeng Liu, Kaiyu Huang, Jiuyi Li, Huan Liu, Jinsong Su, and Degen Huang. 2022. Adaptive token-level cross-lingual feature mixing for multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*.

Xiner Liu, Jianshu He, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. A scenario-generic neural machine translation data augmentation method. *Electronics*.

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*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shilong Pan, Zhiliang Tian, Liang Ding, Haoqi Zheng, Zhen Huang, Zhihua Wen, and Dongsheng Li. 2024. Pomp: Probability-driven meta-graph prompter for llms in low-resource unsupervised neural machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022*.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter

Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, and 1 others. 2018. Mesh-tensorflow: deep learning for supercomputers. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017*.

Asa Cooper Stickland, Alexandre Bérard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for nmt: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation, WMT 2021*.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*.

Isidora Chara Tourni and Subhajit Naskar. 2024. Direct neural machine translation with task-level mixture of experts models. *Preprint*, arXiv:2310.12236.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. 2023. Language-routing mixture of experts for multilingual and code-switching speech recognition. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023*.

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

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Forty-first International Conference on Machine Learning, ICML 2024*.

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

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. 2024. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2024*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *International Conference on Machine Learning, ICML 2023*.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023*.

Yun Zhu, Nevan Wichers, Chu-Cheng Lin, Xinyi Wang, Tianlong Chen, Lei Shu, Han Lu, Canoee Liu, Liangchen Luo, Jindong Chen, and Lei Meng. 2023. Sira: Sparse mixture of low rank adaptation. *Preprint*, arXiv:2311.09179.

A Appendix

A.1 Training Details

We hereby supplement the model training configuration not mentioned in the main text. For both backbone LLMs, we fine-tune the models using a warm-up ratio of 5e-4, a maximum sequence length of 512 tokens, and a weight decay of 0.02. LoRA adapters are applied to the *gate_proj*, *up_proj*, and *down_proj* modules of the backbone LLMs. Stage 1 fine-tuning requires 3 epochs, while Stage 2 requires 2 epochs. Model training process is conducted on 2 NVIDIA A800 GPUs, with each GPU handling 4 batches and employing a gradient accumulation step of 4, resulting in an effective batch size of 32.

A.2 Data Settings

For the fine-tuning data details:

Stage 1 Fine-Tuning: The pre-divided development subset from OPUS-100 serves as our development set. The training data consists of the randomly sampled OPUS-100 train subset combined with the full Flores-200 development subset.

Pseudo-Corpora Generation: To generate pseudo-corpora after Stage 1 fine-tuning, the monolingual backbone sequences required for generation are randomly sourced from the non-overlapping portions of the OPUS-100 training set and the Stage 1 training set. This ensures that the generated pseudo-corpora introduce new data not seen during the initial training phase.

To enhance the quality of the pseudo-corpus, inspired by Junczys-Dowmunt (2018), we implement rule-based filtering strategies, specifically: (1) Target language detection to exclude sequences misaligned with the intended target language; (2) Sequence-length filtering to remove pseudo-pairs with significant disparities in source and target lengths, which often indicate low-quality translations. These filters systematically exclude noisy or unreliable pseudo-corpus entries, ensuring higher fidelity in downstream training tasks.

Stage 2 Fine-Tuning: The training set in this stage comprises the generated pseudo-corpora, supplemented by the Flores-200 development set. About 10% of the combined data is allocated as the evaluation set, with the remaining 90% used for model training. Detailed data statistics are summarized in Table 5.

A.3 The Effect of R-Drop

To scrutinize the impact of employing R-Drop regularization, we compare the CrossLoRA model based on LLaMA-3-8B-Instruct, fine-tuned with and without R-Drop. Corresponding results are presented in Table 6. The ablation reveals that self-contrastive semantic enhancement improves the generalization capability of the CrossLoRA model, achieving substantial performance gains across all translation directions relative to the baseline, without additional inference costs.

A.4 Necessity of Stage 1 Fine-Tuning

The primary objective of Stage 1 fine-tuning is to enhance the model's performance in English-centric translation directions, thereby generating high-quality parallel pseudo-corpora from available English-centric data for subsequent training. To validate the necessity, we conduct an ablation study: Fine-tuning the model using pseudo-corpora generated by the backbone model and the NLLB-3.3B model. The results are summarized in Table 7.

Experimental findings demonstrate that despite the additional computational overhead introduced by Stage 1, the higher-quality pseudo-corpora it generates significantly improve translation performance after Stage 2 fine-tuning. This improvement is particularly pronounced in non-English-centric translation directions.

A.5 Full Results of Main Experiment

In Table 8 and Table 9, We present the specific performance of the CrossLoRA model based on the

Language	Language Family
(En) English (De) German (Is) Icelandic	Germanic, Indo-European
(Cs) Czech (Ru) Russian	Balto-Slavic, Indo-European
(Zh) Chinese	Sino-Tibetan

Table 4: The languages selected in the main experiment and their corresponding language families.

Training Stage	Directions	Parallel Data		
		train	dev	test
Stage 1	En⇔Any	20997	2000	2000
Stage 2	En⇔Any	20997	2000	2000
	others	18997	2000	1012

Table 5: The statistics for the data we utilize for main experiments.

Configurations	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
w/o R-Drop	29.13	78.41	20.09	81.38	23.10	80.39
w/ R-Drop	**29.69**	**78.74**	**20.60**	**81.94**	**23.63**	**80.88**

Table 6: Results of the ablation study on the effect of R-Drop regularization, based on the LLaMA-3-8B-Instruct backbone model. Higher scores are marked in **bold**. Employing the R-Drop method results in a comprehensive performance improvement.

Pseudo-corpora Source	English-centric		non-English-centric		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
LLaMA-3-8B-Instruct	28.68	77.70	18.93	80.02	22.18	79.25
NLLB-3.3B	**29.72**	78.67	20.29	81.55	23.43	80.59
CrossLoRA Stage 1	29.69	**78.74**	**20.60**	**81.94**	**23.63**	**80.88**

Table 7: Results of the ablation study on the effect of Stage 1 training, based on the LLaMA-3-8B-Instruct backbone model. Higher scores are marked in **bold**.

Models	Zh⇒En			En⇒Zh			De⇒En		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	29.75	50.11	79.51	28.07	44.35	80.58	28.08	45.06	75.24
M2M100-12B	27.66	51.72	78.97	27.76	45.06	79.81	30.90	50.48	78.27
LLaMA-3-8B-Instruct	20.84	39.64	74.93	18.76	33.86	73.47	23.07	37.75	71.10
CrossLoRA Stage 1	33.37	55.59	81.17	34.45	51.15	82.26	33.00	53.12	79.58
CrossLoRA Stage 2	34.95	58.13	82.19	36.21	53.93	83.36	33.99	52.28	80.16
Models	En⇒De			Ru⇒En			En⇒Ru		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	27.54	43.22	78.24	29.03	47.86	76.80	28.41	43.21	82.51
M2M100-12B	27.29	45.93	76.48	26.65	46.34	76.97	23.39	36.81	79.54
LLaMA-3-8B-Instruct	21.26	34.14	70.36	22.54	38.50	71.10	18.74	30.84	73.84
CrossLoRA Stage 1	28.47	48.15	78.31	32.01	54.26	79.06	26.47	45.59	81.28
CrossLoRA Stage 2	28.16	45.80	78.33	32.17	52.33	78.94	26.34	45.21	81.95
Models	Cs⇒En			En⇒Cs			Is⇒En		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	31.10	47.71	76.15	28.11	41.25	81.39	25.47	43.63	72.63
M2M100-12B	26.12	41.56	76.31	21.19	32.69	77.70	16.41	38.40	64.67
LLaMA-3-8B-Instruct	22.58	38.45	70.06	15.38	25.64	71.02	11.89	21.46	55.51
CrossLoRA Stage 1	32.82	54.07	79.88	26.14	45.55	81.11	25.65	48.69	71.62
CrossLoRA Stage 2	34.02	55.64	80.37	24.95	44.71	81.61	26.06	49.42	72.18
Models	En⇒Is			De⇒Zh			Zh⇒De		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	22.98	38.37	67.03	25.11	43.38	79.31	18.17	41.43	80.52
M2M100-12B	13.21	32.84	57.15	27.24	48.11	84.06	16.47	39.34	80.09
LLaMA-3-8B-Instruct	9.05	17.29	54.71	16.81	32.70	76.69	13.26	33.07	77.25
CrossLoRA Stage 1	22.63	44.89	68.32	16.56	32.00	76.45	13.34	34.01	77.80
CrossLoRA Stage 2	21.05	43.02	68.35	37.96	56.18	86.21	22.54	48.29	81.58
Models	De⇒Ru			Ru⇒De			De⇒Cs		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	25.29	46.45	87.12	24.17	49.12	81.89	24.13	47.40	89.48
M2M100-12B	22.07	43.26	86.55	21.30	45.92	80.52	23.35	46.50	89.61
LLaMA-3-8B-Instruct	17.79	36.24	82.52	17.84	39.78	77.34	17.11	37.73	85.10
CrossLoRA Stage 1	17.85	36.19	82.40	17.78	39.93	77.23	17.00	37.62	84.79
CrossLoRA Stage 2	27.23	50.17	87.57	28.44	54.32	82.03	16.69	41.55	86.49

Table 8: Part 1 of the full results for all translation directions of the main experiment.

LLaMA-3-8B-Instruction backbone LLM across all translation directions in the main experiment. The performance metrics include BLEU scores, ROUGE-L, and COMET scores. For comparison, the table also includes the performance of prior studies and the backbone LLM baseline.

From the table, it is evident that after fine-tuning in CrossLoRA Stage 1, the model's scores have significantly improved in translation directions involving English, while maintaining the backbone model's performance in other non-English-involved directions. After further fine-tuning in Stage 2, with the addition of pseudo-corpus to the training data, the model achieves substantial improvements in translation directions not involving English, reaching or even exceeding the performance of specialized translation models.

Models	Cs⇒De			De⇒Is			Is⇒De		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	26.02	51.21	84.59	18.19	43.15	82.37	20.63	44.89	78.80
M2M100-12B	24.00	49.15	83.55	13.72	37.02	79.35	18.99	42.91	78.30
LLaMA-3-8B-Instruct	20.26	42.71	80.47	8.06	26.90	72.05	9.92	24.22	66.90
CrossLoRA Stage 1	20.34	42.90	80.80	8.23	27.17	72.31	10.05	24.10	67.12
CrossLoRA Stage 2	30.69	56.65	84.77	13.11	38.37	72.49	21.62	48.72	79.26
Models	Zh⇒Ru			Ru⇒Zh			Zh⇒Cs		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	17.50	36.90	85.57	24.96	42.84	80.22	15.64	36.42	86.20
M2M100-12B	15.76	34.68	85.39	26.10	46.57	83.60	14.87	35.39	86.59
LLaMA-3-8B-Instruct	12.25	27.87	81.79	23.02	71.13	79.93	11.29	28.05	82.84
CrossLoRA Stage 1	12.11	27.40	81.51	22.87	41.15	79.43	11.35	28.40	83.16
CrossLoRA Stage 2	21.68	44.27	86.76	38.32	57.87	85.92	18.74	42.91	87.43
Models	Cs⇒Zh			Zh⇒Is			Is⇒Zh		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	24.38	42.81	79.50	12.81	34.72	79.63	20.83	38.99	77.27
M2M100-12B	26.96	47.80	84.38	9.89	30.84	77.44	21.14	42.30	80.73
LLaMA-3-8B-Instruct	18.49	34.99	77.71	6.34	21.53	71.12	16.20	33.11	73.70
CrossLoRA Stage 1	18.55	35.30	78.21	6.54	21.85	71.71	16.11	32.90	73.09
CrossLoRA Stage 2	40.89	59.69	86.87	10.69	31.98	66.31	31.61	53.75	83.64
Models	Cs⇒Ru			Ru⇒Cs			Cs⇒Is		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	24.26	45.55	87.82	21.10	43.93	88.36	16.43	40.52	81.39
M2M100-12B	21.65	43.34	87.76	20.18	42.64	88.99	12.49	35.55	76.42
LLaMA-3-8B-Instruct	17.69	36.25	83.09	15.04	35.02	83.93	7.97	26.07	71.36
CrossLoRA Stage 1	17.72	35.96	82.90	15.16	35.20	83.87	7.76	25.87	70.62
CrossLoRA Stage 2	20.42	41.40	87.55	16.50	37.62	86.87	12.48	36.70	76.54
Models	Is⇒Cs			Ru⇒Is			Is⇒Ru		
	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET	BLEU	ROUGE	COMET
NLLB-3.3B	17.32	38.99	84.35	15.23	39.21	80.58	18.43	38.32	82.64
M2M100-12B	16.05	36.89	82.39	11.49	33.54	74.37	15.96	35.35	80.20
LLaMA-3-8B-Instruct	10.77	27.58	77.63	7.31	25.06	70.51	11.60	27.48	76.28
CrossLoRA Stage 1	10.85	26.98	77.46	7.02	25.30	70.29	11.74	26.59	75.80
CrossLoRA Stage 2	12.15	31.51	77.95	12.59	36.07	75.18	14.76	33.32	81.20

Table 9: Part 2 of the full results for all translation directions of the main experiment.