SiLP: Enhancing Non-Dominant Language Capabilities with a Selective Bidirectional Language Projection Framework

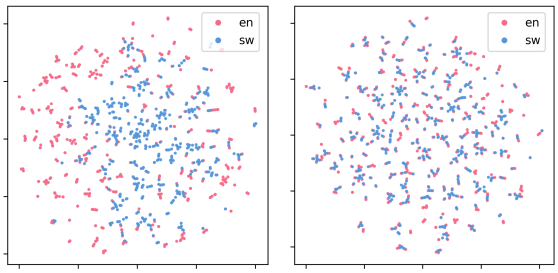
Junpeng Liu[1]*, Jiuyi Li[2]*, Kaiyu Huang[3], Bo Jin[2], Degen Huang[4]†, Hui Xiong[1]†
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Dalian University of Technology [3]Beijing Jiaotong University
[4]Fuyao University of Science and Technology
junpengliu@hkust-gz.edu.cn xionghui@ust.hk

Abstract

Current large language models (LLMs) often exhibit performance imbalances between dominant languages (e.g., English) and non-dominant ones due to the skewed distribution of pretraining data. A common strategy to address this issue is to enhance cross-lingual alignment, thereby facilitating non-dominant language processing. However, existing methods typically rely on additional training objectives or language-specific parameters, which increase training complexity and cost. In this work, we propose a selective bidirectional language projection framework that enables efficient multilingual alignment and language shift using the intrinsic parameters. Specifically, we first identify the layers most sensitive to language projection between non-dominant and dominant languages through neuron activation analysis. We then perform sequential language projection within the selected layers by mapping non-dominant representations into the dominant language space and reverting them before generation. The bidirectional projection benefits the subsequent instruction tuning in non-dominant languages. Experiments on seven benchmarks demonstrate that our method remarkably enhances the performance of non-dominant languages. Further analyses indicate that our method learns better internal representations and exhibits strong generalization capabilities.

1 Introduction

Large language models (LLMs) have demonstrated impressive multilingual capabilities across diverse understanding and generation tasks (OpenAI, 2023; Google et al., 2023; AI@Meta, 2024). However, imbalanced pretraining data biases models toward dominant languages (e.g., English), resulting in a significant performance gap with non-dominant ones (Zhang et al., 2024; Zhu et al., 2024; Fan et al., 2025). Consequently, achieving balanced multilingual proficiency in LLMs remains a critical and persistent challenge in this field (Zhang et al., 2023; Ye et al., 2025).



(a) LLaMA3-8B (b) LLaMA3-8B+Translation

Figure 1: (a) Representation visualization of the 15th layer in LLaMA3-8B via t-SNE. (b) The internal representations are aligned after fine-tuning with the Swahili-English translation data, indicating that translation effectively supports cross-lingual alignment.

A prevalent strategy to mitigate this performance gap is multilingual alignment (Li et al., 2024a; Yoon et al., 2024; Zhang et al., 2025; Liu and Niehues, 2025), which aims to align the dominant and non-dominant languages in the shared latent semantic space, thereby enhancing the non-dominant capabilities via knowledge transfer. Recent studies (Zhao et al., 2024b; Huo et al., 2025) have further suggested a three-step processing framework: non-dominant language inputs are first mapped into a language-agnostic representation space, followed by semantic understanding or reasoning, and finally projected back into the target language. Within this paradigm, some works focus on aligning intermediate layers via additional explicit supervision, such as contrastive learning (Li et al., 2024a; Zhang et al., 2025) or deep supervised fine-tuning (Huo et al., 2025). However, these methods require balancing multiple training objectives and often

*Equal Contribution
†Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43307–43325

incur significant increases in training complexity and cost. Other works (Zhang et al., 2025; Ye et al., 2025; Wang et al., 2025) explore explicit language shift mechanisms between the dominant and non-dominant languages. While promising, these approaches typically rely on additional parameters to implement the shift. Therefore, how to achieve effective language projection using only intrinsic model parameters remains underexplored.

Recent findings (Zhao et al., 2024b; Tezuka and Inoue, 2025) reveal that LLMs have specialized neurons supporting cross-lingual latent transitions. This observation suggests that language projection may be achieved solely with intrinsic model parameters. Motivated by this insight, we conduct preliminary experiments on LLaMA3-8B (AI@Meta, 2024) by fine-tuning the lower layers on parallel sentences and analyzing the resulting internal representations, as translation can be viewed as a natural form of language shift. As shown in Figure 1, representations become well-aligned in the middle layer, showing that translation can effectively support cross-lingual projection. However, performing full-parameter training risks degrading the model's inherent reasoning and understanding abilities, which may ultimately harm performance in non-dominant languages.

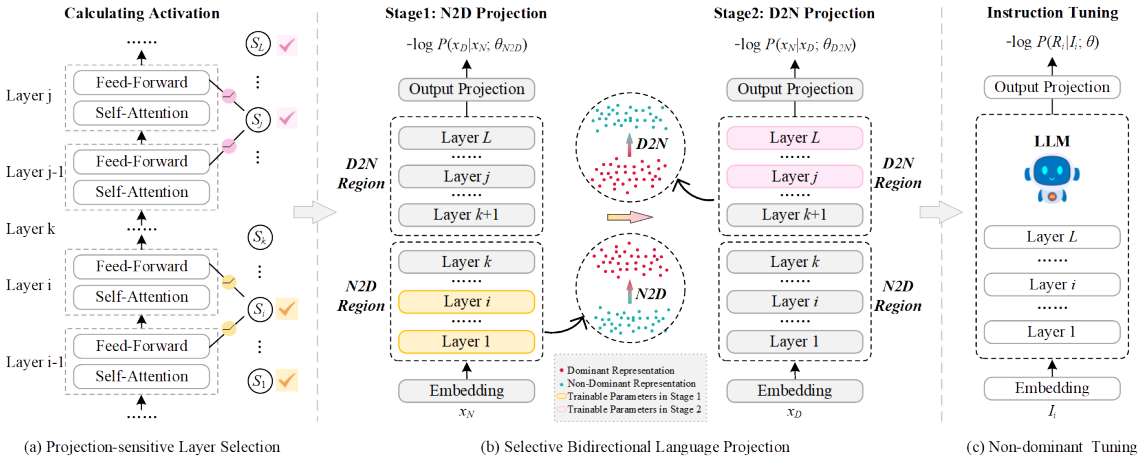
In this study, we propose a **s**elective b**i**directional **l**anguage **p**rojection framework (SiLP) that enables efficient multilingual alignment without introducing external parameters. Following the three-step processing paradigm, our approach leverages intrinsic LLM parameters to facilitate both understanding and generation in non-dominant languages. Specifically, we first identify the layers that are most sensitive to language projection between non-dominant and dominant languages through neuron activation analysis. We then utilize parallel data to perform sequential language projection within these selected layers, enabling efficient language shifts in both the lower and upper parts of the LLM. Finally, we fine-tune the aligned model on instruction data in the target non-dominant language, thereby strengthening its non-dominant capabilities while maintaining high alignment efficiency. Our main contributions are summarized as follows:

- We propose SiLP, a selective bidirectional language projection framework to enhance the non-dominant language capabilities by aligning their representations with the dominant counterparts using the intrinsic model parameters. We also propose an activation analysis method to select the projection-sensitive layers for efficient projection.
- We conduct extensive evaluations across different tasks and languages, and experimental results show that our method remarkably outperforms strong baselines, achieving an average performance gain of up to +2.85 points.
- In-depth analysis confirms that our method can deliver substantial performance improvements with limited parallel data. Furthermore, the method proves effective across different LLM models and scales, demonstrating strong generalization capabilities.

2 Related Work

Multilingual LLMs Large language models pretrained on large-scale multilingual datasets demonstrate remarkable multilingual capabilities. These models have shown proficiency in a wide range of understanding (Bandarkar et al., 2024; Niklaus et al., 2023; Li et al., 2024a) and generation (Li et al., 2024c; Zhang et al., 2024; Zhao et al., 2024a) tasks. However, most LLMs still exhibit limited performance in low-resource languages due to imbalanced distribution of pretraining data, resulting in a significant cross-lingual performance gap (OpenAI, 2023; Zhang et al., 2024). Considerable efforts have been made to improve performance in non-dominant languages by constructing multilingual training samples via translation (Zhu et al., 2023; Ranaldi et al., 2024), but these methods often face challenges such as high annotation costs and low-quality translations (Muennighoff et al., 2023; Tan et al., 2024; Chen et al., 2024). In this work, we propose a bidirectional language projection framework designed to align cross-lingual representations and enhance model performance in non-dominant languages with limited parallel data.

Multilingual Alignment To enhance the performance of LLMs in non-dominant languages, substantial efforts have been devoted to cross-lingual alignment. One line of work aligns non-dominant languages with dominant ones by incorporating multilingual parallel data for joint training or introducing auxiliary training objectives. Typical approaches include integrating translation-based instruction data (Li et al., 2023, 2024b; Zhang et al., 2024), aligning internal representations through contrastive training (Li et al., 2024a; Huo et al.,



(a) Projection-sensitive Layer Selection (b) Selective Bidirectional Language Projection (c) Non-dominant Tuning

Figure 2: Overview of our selective bidirectional language projection framework. First, it identifies the layers most sensitive to language projection via neuron activation analysis. Second, the selected layers are sequentially fine-tuned with parallel data to align the representations of dominant and non-dominant languages. Lastly, the aligned model is fine-tuned with instruction data in the target non-dominant language.

2025; Zhang et al., 2025), aligning questions via translation to English (Zhu et al., 2024), selectively fine-tuning FFN in the lower layers (Fan et al., 2025), and aligning lexicons via random code-switching (Chai et al., 2025; Zhang et al., 2024). Another line of work focuses on leveraging language shift to project the representations between dominant and non-dominant languages. Representative methods involve employing language vectors to better learn unified representations (Xu et al., 2023; Zhang et al., 2025), performing cross-lingual intervention (Wang et al., 2025) or transformation (Ye et al., 2025) in the inference stage. Recent research (Zhao et al., 2024b; Tezuka and Inoue, 2025) reveals the internal mechanisms in LLMs: converting diverse language inputs into a dominant language (e.g., English) in the lower layers, while performing the reverse mapping in the upper layers. These findings further underscore the necessity of multilingual alignment. Building upon these insights, we propose a translation-based alignment framework that first identifies layers pertinent to language projection and then conducts selective fine-tuning with those selected intrinsic parameters to facilitate multilingual alignment.

3 Method

As illustrated in Figure 2, our SiLP framework contains the following three steps: (1) Projection-sensitive layer selection, which identifies the layers most strongly associated with language projection. (2) Selective bidirectional language projection, which sequentially aligns dominant and non-dominant language representations in both directions using the selected layers. (3) Instruction tuning in non-dominant languages, strengthens non-dominant language performance.

3.1 Projection-sensitive Layer Selection

Our goal is to enable language projections between dominant and non-dominant languages in the latent space, thereby enabling LLMs to better process non-dominant language inputs by aligning them with dominant-language representations and generating outputs in the non-dominant language. To achieve this efficiently, we first identify the layers most sensitive to cross-lingual projection. We use translation as a probing task for this purpose, since it inherently requires mapping between languages. Specifically, we measure changes in neuron activation between two consecutive feed-forward network (FFN) sublayers during translation, which allows us to identify the layers most involved in cross-lingual projection. The FFN module is formalized as:

$$\mathbf{h}^l_{out} = \text{act\_fn}(\mathbf{h}^l_{in}\mathbf{W}^l_1) \cdot \mathbf{W}^l_2 \quad (1)$$

where $\text{act\_fn}(\cdot)$ is the activation function and the resulting activation values are denoted as $\mathbf{A}$. For each translation pair (x, y), we conduct a forward pass through the model from the source to the target language and calculate the activation differences between adjacent layers. Formally, the activation values of neuron i in layers $l-1$ and l during the n-th forward pass is represented as $\mathbf{A}^{l-1}_{i,n}$ and $\mathbf{A}^{l}_{i,n}$, respectively. The activation change between

adjacent layers across N forward propagations is defined as:

$$S_l(x,y) = \frac{1}{MN}\left|\sum_{n=1}^{N}\sum_{i=1}^{M}\left(\mathbf{A}_{i,n}^{l} - \mathbf{A}_{i,n}^{l-1}\right)\right| \quad (2)$$

where M is the total number of intermediate neurons in each FFN layer. S_l quantifies the average changes in activation patterns when the model translates from source to target language. The layers with larger changes are considered to play a more critical role in cross-lingual mapping and thus better suited for interlingual representation projection.

3.2 Bidirectional Language Projection

Layer Allocation. After calculating S_l for all layers, we assign the most sensitive layers to each projection direction. Prior studies (Zhao et al., 2024b; Tezuka and Inoue, 2025) empirically show that the lower layers tend to convert diverse input languages into English, whereas the upper layers reverse this process. Following this observation, we separately select the projection-sensitive layers from the lower and upper parts of the model as follows:

$$L_{N2D} = \arg\max_{\text{top}k_1}\{S_1, S_2, ..., S_k\} \quad (3)$$

$$L_{D2N} = \arg\max_{\text{top}k_2}\{S_{k+1}, S_{k+2}, ..., S_{L-1}\} \quad (4)$$

where L_{N2D} denotes the top-k_1 layers responsible for projecting non-dominant languages to the dominant ones (N2D projection), while L_{D2N} represents the top-k_2 layers for the reverse projection (D2N projection). After identifying the projection-sensitive layers, we leverage translation tasks to perform bidirectional language projection.

N2D Projection. To make the model effectively process the non-dominant language in the lower layers, we first conduct N2D projection to align non-dominant language representations with those of the dominant language counterparts, in which the model already exhibits strong proficiency. This alignment allows the model to handle non-dominant language inputs in a manner similar to dominant ones, thereby facilitating knowledge transfer from dominant to non-dominant languages. Specifically, given a parallel sentence pair (x_N, x_D), where x_N is an input in a non-dominant language and x_D is the corresponding translation in the dominant language. The N2D projection trains the model to generate the response x_D given the input x_N, which is formulated as follows:

$$\mathcal{L}_{N2D} = -\log P(x_D|x_N, \theta_{L_{N2D}}) \quad (5)$$

where L_{N2D} is the layers allocated to N2D projection and $\theta_{L_{N2D}}$ denotes the corresponding parameters. By learning this projection in the intermediate representation space, the model develops a shared representation space that bridges the semantic gap between dominant and non-dominant languages, enabling more effective cross-lingual generalization in the subsequent layers.

D2N Projection. To enable the model to generate appropriate responses in the target non-dominant language, we further introduce an inverse projection step that maps dominant-language representations back to the non-dominant language prior to generation. Specifically, we reverse the parallel sentence pair to obtain (x_D, x_N) and perform D2N projection at the upper layers of the model, which is formulated as:

$$\mathcal{L}_{D2N} = -\log P(x_N|x_D, \theta_{L_{D2N}}) \quad (6)$$

where L_{D2N} is the layers at which the D2N projection is applied, $\theta_{L_{D2N}}$ is the corresponding parameters. This projection ensures that knowledge acquired from the dominant language is preserved and effectively leveraged for generation in the non-dominant language.

3.3 Multi-stage Training

To fully leverage the transferability of the dominant language, we propose a multi-stage training strategy, as shown in Figure 2. Since the inherent semantic gaps in vanilla LLMs may result in suboptimal performance, we first perform selective bidirectional language projection sequentially. Specifically, we apply N2D projection to align non-dominant language representations with the dominant language space, followed by D2N projection to enable effective generation in the target non-dominant language.

After establishing bidirectional alignment, we finally fine-tune the model on instruction dataset in the target non-dominant language to enhance the task-specific performance. Formally, given the non-dominant instruction dataset $\mathcal{D} = \{(I_i, R_i)\}_{i=1}^{N}$, the training objective is formulated as:

$$\mathcal{L}_{TFT} = -\frac{1}{N}\sum_{i=1}^{N}\log P(R_i|I_i, \theta) \quad (7)$$

where I_i denotes the input question or instruction, R_i is the corresponding output or response, and θ denotes the model parameters.

Model	Understanding				Generation		
	BELE.	***MMMLU***	***XCOPA***	***XStoryCloze***	***MKQA***	***XQuAD***	***XLSUM***
				LLaMA3-8B			
SFT	57.97	38.92	69.64	74.68	18.24	60.13	13.67
TFT	61.35	40.18	65.72	76.87	22.34	70.83	19.68
QAlign	56.90	38.55	66.60	78.38	20.30	24.49	7.57
SLAM	58.04	38.42	70.64	80.05	18.18	59.76	13.93
Incline	60.07	40.43	**73.16**	81.26	18.64	66.84	16.71
DFT	62.63	40.87	69.16	75.55	22.32	65.85	19.73
Ours	**66.00**	**42.67**	72.48	**81.49**	**22.80**	**71.64**	**19.98**
				Qwen2.5-7B			
SFT	75.94	52.58	72.32	81.28	20.24	70.97	18.09
TFT	77.61	53.55	73.04	88.19	22.30	71.09	19.69
QAlign	63.44	49.43	78.04	79.34	19.90	19.46	6.85
SLAM	67.28	49.77	74.28	79.19	18.20	65.49	18.26
Incline	75.99	55.17	**78.20**	85.90	20.92	71.02	18.94
DFT	77.74	53.78	76.32	88.40	20.36	67.02	19.77
Ours	**78.39**	**55.25**	74.88	**90.39**	**23.38**	**71.46**	**19.99**

Table 1: The averaged results of different languages across seven tasks within two distinct models. Detailed results for each language are presented in Tables 10-16, Appendix B.

4 Experiments

4.1 Experimental Settings

Training setup. We conduct experiments on two LLMs: *LLaMA3-8B* (AI@Meta, 2024) and *Qwen2.5-7B* (Yang et al., 2025). We adopt the multilingual instruction dataset Bactrian-X (Li et al., 2023) as the training corpus, and construct parallel sentences using the inputs from Bactrian-X for projection-sensitive layer detection and language projection. Further details about training setups are presented in Appendix A.1.

Evaluation. We conduct evaluations on seven multilingual benchmarks, covering four natural language understanding (NLU) tasks (*BELEBELE* (Bandarkar et al., 2024), *MMMLU* (Lai et al., 2023), *XCOPA* (Ponti et al., 2020), *XStoryCloze* (Lin et al., 2022)) and three generation (NLG) tasks (*MKQA* (Longpre et al., 2021), *XQuAD* (Artetxe et al., 2020), *XLSUM* (Hasan et al., 2021)). We report *ROUGE-L* scores for *XLSUM* dataset while *Accuracy* for other datasets.

Baselines. We implement the baselines for comparison: **SFT** (Ouyang et al., 2022), **TFT** (Chen et al., 2024), **QAlign** (Zhu et al., 2024), **SLAM** (Fan et al., 2025), **Incline** (Wang et al., 2025) and **DFT** (Huo et al., 2025).

4.2 Main Results

The average results across different languages on multilingual NLU and NLG tasks are presented in Table 1. Our method remarkably outperforms the vanilla SFT and TFT models in both multilingual understanding and generation tasks. The results demonstrate that our method effectively enhances the model's capabilities in the target non-dominant languages via multilingual alignment. Although QAlign and SLAM also utilize translation-based alignment, they do not consistently improve generation tasks, as they were originally designed for reasoning tasks. For instance, SLAM aligns multilingual representations in the upper layers of the model, which can lead to incorrect target-language responses in summarization. Our method also surpasses the training-free cross-lingual alignment method Incline on most testsets, except XCOPA. We conjecture that the improvements of Incline on non-dominant languages may be constrained by its performance on the dominant language, since it intervenes with dominant-like representations during inference. Compared with DFT, which also introduces language conversion in the intermediate layers, our method achieves more stable improvements while consuming less GPU memory during training. Overall, our method achieves the best average performance across all tasks, demonstrating that the language projection not only facilitates the understanding of non-dominant languages but also strengthens their generative capability.

4.3 Ablation Studies

(1) Feed-forward network contributes most in language projection. We study the contributions of different modules in LLMs to language projec-

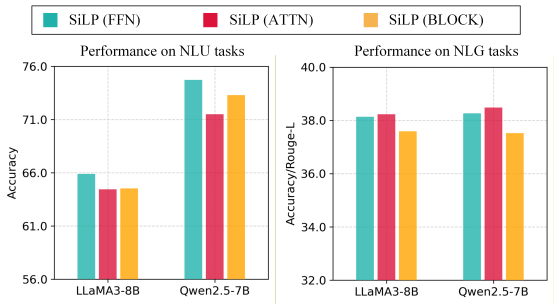


Figure 3: Performance comparisons of using feed-forward, self-attention and whole decoder block for model training on understanding and generation tasks.

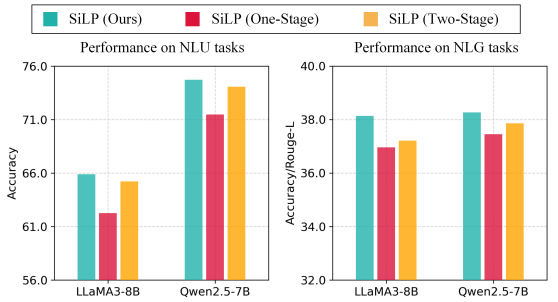
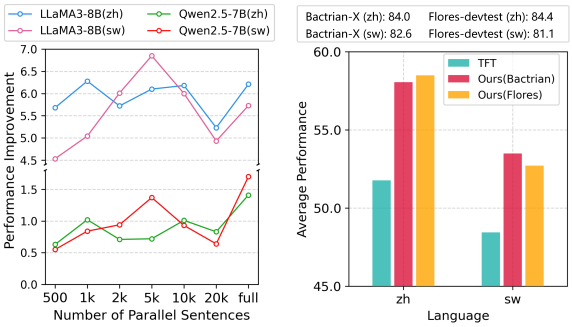


Figure 4: Performance comparisons of different training strategies on understanding and generation tasks.

tion, including self-attention, feed-forward and the whole decoder block. As shown in Figure 3, our method achieves optimal overall performance when applied to the feed-forward network. While the attention modules exhibit slightly stronger capabilities in NLG tasks, they substantially underperform in NLU tasks. These findings highlight the critical role of FFN in cross-lingual alignment. Our results align well with prior work by Tezuka and Inoue (2025), which identifies transfer neurons within FFN modules that map representations between the shared and language-specific latent spaces.

(2) Training strategy significantly impacts performance. To assess the impact of different training strategies, we implement two alternative approaches: one-stage (joint) training and two-stage training. In one-stage training, bidirectional language projection and TFT are jointly optimized within a single phase. In contrast, two-stage training first performs bidirectional language projection simultaneously, followed by TFT. As shown in Figure 4, our three-stage training strategy outperforms the two alternatives, demonstrating that performing sequential language projection before TFT can effectively preserve the transfer capability from the dominant language to non-dominant ones.



(a) Number of Samples (b) Data Source

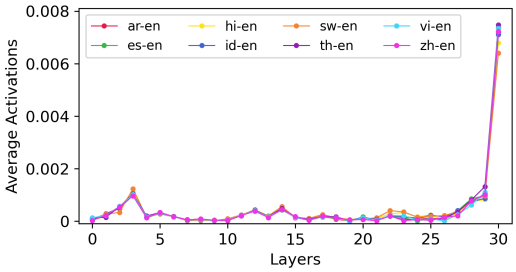
Figure 5: Impact of translation data: (a) Performance improvement concerning the number of translation samples. (b) Performance comparison of different translation data sources in LLaMA3-8B.

5 Analysis

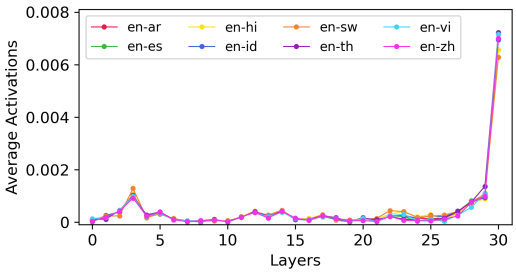
5.1 Impact of Parallel Data

A Small amount of parallel data makes improvements. To investigate how the amount of parallel data used in the language projection stage affects model performance, we conduct experiments with varying amounts of parallel data in the language projection stage and present the results on Chinese and Swahili in Figure 5(a). As shown in Figure 5(a), even a small set of 500 parallel sentences can yield noticeable improvements in both languages, demonstrating the high efficiency of language projection. However, increasing the amount of translation data does not consistently improve performance across languages. For example, while performance on Swahili continues to improve as the data increases from 500 to 5k, the performance on Chinese fluctuates and even declines in some cases. We conjecture that Swahili may require more parallel data than Chinese to establish effective alignment with English, since it is a low-resource language.

High-quality parallel data brings greater improvements. By default, the parallel data are constructed with inputs from Bactrian-X dataset. To examine the impact of parallel data quality on model performance, we additionally sample parallel sentences from the Flores devtest (Costa-Jussà et al., 2022) for comparison. We adopt the reference-free metric COMET-Wiki (wmt22-cometkiwi-da) (Rei et al., 2022) to measure the quality of the parallel data. The COMET scores and performance results under the 1k-sample setting are shown in Figure 5(b). Our method consistently outperforms



(a) N2D Projection

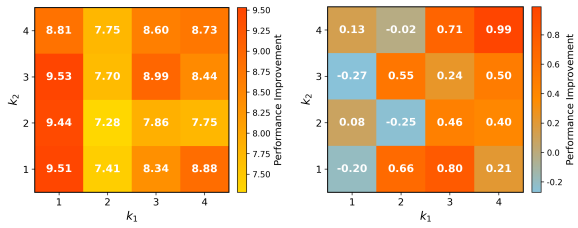


(b) D2N Projection

Figure 6: Layer-wise average activation changes across different language pairs in the LLaMA3-8B.

Model	LLaMA3-8B		Qwen2.5-7B	
	NLU	NLG	NLU	NLG
TFT	61.03	37.62	73.10	37.70
Ours	**65.86**	**38.14**	**74.73**	**38.27**
w/o L_{N2D}	64.72	37.89	73.77	37.77
w/o L_{D2N}	65.03	38.07	73.56	37.95

Table 2: The impact of bidirectional language projection on the average results of all benchmarks.



(a) NLU (b) NLG

Figure 7: Average improvement for Chinese and Swahili on 7 tasks at different (k_1,k_2) settings in LLaMA3-8B.

the vanilla TFT across both parallel datasets, and the improvement is clearly correlated with the quality of the parallel data. This indicates that performance is more dependent on the data quality rather than the data source.

5.2 Impact of Bidirectional Language Projection

Projection-sensitive FFNs locate in the lower and upper layers. We analyze the distribution of layers that exhibit higher sensitivity to language projection. Figure 6 presents the average activation changes between adjacent layers in LLaMA3-8B when translation instructions across different language pairs are provided. The results show that the average activation changes follow a similar trend for both N2D and D2N projections across language pairs. Higher activation changes are generally observed in the lower and upper layers, suggesting that these layers are more sensitive to language projection. Similar patterns can also be found in Qwen2.5-7B, as shown in Figure 12. Accordingly, we select the lower layers with the largest activation changes for the N2D projection and the upper layers with the largest activation changes for the D2N projection.

Bidirectional projection is necessary. Based on the ablation results in Table 2, we can observe that removing either N2D or D2N projection from the model leads to a decline in performance across both NLU and NLG tasks on LLaMA3-8B and Qwen2.5-7B. The results demonstrate that each direction of the language projection contributes to the overall effectiveness. Aligning non-dominant languages with the dominant language alone is insufficient, and mapping knowledge back from the dominant language to non-dominant languages is equally important for maintaining strong cross-lingual generalization.

Impact of different training layers. To evaluate how k_1 and k_2 affect performance, we conduct experiments under different (k_1, k_2) settings and present the results in Figure 7. The performance on NLU tasks can be remarkably improved when only one layer is used in both the bottom and top sides. This indicates that our language projection method can effectively align non-dominant languages with the dominant language, thereby enabling the model to benefit from the representations of the dominant language. In contrast, for NLG tasks, larger values of k_1 and k_2 are needed to ensure that the model responds correctly in the target non-dominant language. Overall, pronounced performance improvements for NLU and NLG tasks are achieved when k_1 and k_2 fall within the range $\{3, 4\}$. Based on these observations, we empirically set both k_1 and k_2 to 4 in the experiment to achieve balanced per-

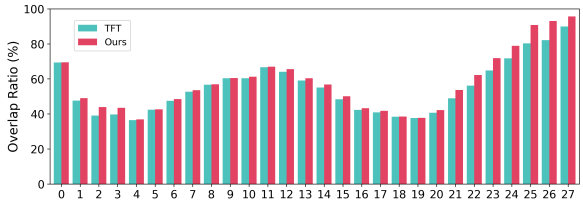
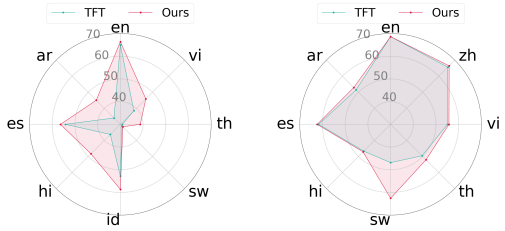


Figure 8: Comparison of the overlap ratio between English and Swahili activated neurons in LLaMA3.2-3B.



(a) LLaMA3-8B (b) Qwen2.5-7B

Figure 9: Average improvement achieved on other languages when finetuning LLaMA3-8B and Qwen2.5-7B with Chinese and Indonesian, respectively.

formance on NLU and NLG tasks. We also compare the performance of different layer selection methods and present the results in Figure 10. Our projection-sensitive layer selection method outperforms the random selection and all-layer selection methods, demonstrating the effectiveness of our layer selection strategy.

5.3 Representation Analysis

To understand the mechanism of bidirectional language projections, we visualize the internal sentence representations across different languages using t-SNE (Van der Maaten and Hinton, 2008) for visualization. Specifically, we encode parallel English and Swahili sentences from the Flores dataset and obtain sentence representations through mean pooling over all tokens. The representation distributions at the 15th and 31st layers are presented in Figures 1 and 11, respectively. At the 15th layer, representations of the two languages align better after applying the N2D projection, indicating that our method effectively brings non-dominant language representations closer to those of the dominant language. At the 31st layer, while representations remain clustered by language, they also exhibit greater proximity to each other. To quantify cross-lingual alignment, we employ the accuracy of similarity search tasks as a quantitative indicator of cross-lingual representation alignment following Liu et al. (2022). As presented in Figure 13, the search accuracy is notably improved with the D2N projection, demonstrating that representations are semantically aligned despite possessing distinct language-specific spaces.

5.4 Neuron Activation Analysis

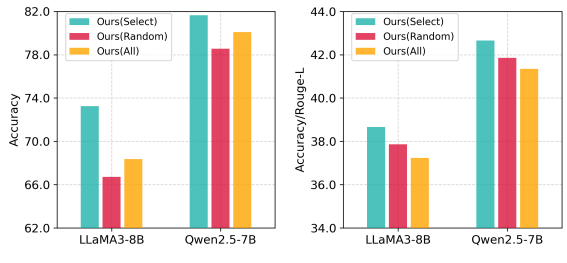
We further analyze the activation patterns in each FFN layer for non-English languages compared to English to facilitate understanding of our method. Following Fan et al. (2025), we compute the overlap ratio of activated neurons when processing parallel sentences in different languages. A higher overlap indicates that the model activates similar neural pathways for non-dominant languages as it does for English, suggesting successful transfer of English capabilities. The results for English-Swahili are shown in Figure 8. SiLP consistently achieves higher neuron activation overlap than TFT, particularly in the middle-to-deep layers. This indicates that our bidirectional projection enables the model to simulate the activation pattern of English when processing non-dominant languages, leading to effective reuse of English-centric knowledge and resulting in performance gains.

5.5 Generalization across Languages

To investigate the cross-lingual transferability of our method, we further evaluate the model's performance on languages that are not involved in instruction tuning. Specifically, we fine-tune LLaMA3-8B and Qwen2.5-7B separately on Chinese and Indonesian datasets, and evaluate their performance on other languages. As illustrated in Figure 9, finetuning the model on one language consistently improves performance across diverse languages, regardless of language family. The improvements on LLaMA3-8B are more pronounced, likely due to its more imbalanced pretraining data distribution, which allows greater benefit from internal representation alignment. These results confirm the efficacy of our method in facilitating cross-lingual transfer.

5.6 Performance across Different Scales

We further assess the generalization of our method across model scales. Experiments are conducted on the LLaMA family (3B and 13B) and the Qwen2.5 family (0.5B, 1.5B, and 14B), with average results summarized in Table 3. The results show that our method stably outperforms baselines across all model scales. Specifically, for the LLaMA fam-



(a) NLU (b) NLG

Figure 10: Comparison of average performance on the Chinese dataset under different layer selection methods.

ily, our method shows average improvements of 1.17 and 0.83 points at the 3B and 13B scales, respectively. For the Qwen2.5 family, our method achieves average improvements of 0.74, 0.47, and 1.24 points at the 0.5B, 1.5B, and 14B scales, respectively. Comparing the results within the Qwen2.5 family, we find that the improvements on larger models are greater than those on smaller models, which is likely due to the inherent multilingual capacities of the base models. These findings confirm that our method generalizes effectively across different model families and scales.

5.7 Comparison of Layer Selection Methods

To evaluate the necessity of the projection-sensitive layer selection strategy, we conduct experiments to compare our method with other layer selection methods, including random selection and all-layer selection. Random selection performs language projection in the randomly selected layers in the lower and upper parts of the model, while all-layer selection utilizes all layers for language projection. The results are presented in Figure 10. The results show that our projection-sensitive layer selection method outperforms the other two methods, demonstrating the effectiveness of our layer selection strategy.

5.8 Experiments with Non-English languages as the Dominant Language

Language projection aims to align the representations of non-dominant languages with those of the dominant language. To evaluate the generalization of our method, we conduct experiments using French as the dominant language instead of English, with a 1k parallel dataset constructed from the Flores devtest. The results in Table 4 show that the average performance on both NLU and NLG tasks is still improved for both LLaMA3-8B and Qwen2.5-7B. However, since the two base models generally perform worse in French than in English, the improvements are relatively smaller compared to those under the English-dominant settings.

Model	*BELE.*	*XCOPA*	*XQuAD*	*XLSUM*
		LLaMA3.2-3B		
SFT	45.07	51.56	61.15	13.43
TFT	52.01	62.16	66.32	19.37
Ours	**53.64**	**63.96**	**67.02**	**19.93**
		LLaMA2-13B		
SFT	34.15	41.84	49.44	13.47
TFT	46.72	56.60	53.45	19.42
Ours	**47.92**	**57.64**	**54.13**	**19.84**
		Qwen2.5-0.5B		
SFT	34.42	53.88	41.86	14.57
TFT	35.55	54.16	48.98	16.50
Ours	**35.98**	**55.56**	**49.34**	**17.28**
		Qwen2.5-1.5B		
SFT	54.54	63.44	61.05	16.35
TFT	61.53	68.08	64.47	17.82
Ours	**62.29**	**68.58**	**64.80**	**18.11**
		Qwen2.5-14B		
SFT	83.90	87.12	72.33	17.09
TFT	84.63	86.12	71.99	19.03
Ours	**85.39**	**87.36**	**72.78**	**21.20**

Table 3: Average performance across languages under models of different scales and families.

Model	LLaMA3-8B		Qwen2.5-7B	
	NLU	NLG	NLU	NLG
TFT	61.03	37.62	73.10	37.70
Ours	**63.58**	**38.35**	**73.62**	**37.85**

Table 4: Average performance on all benchmarks with French as the dominant language for language projection. Detailed results are presented in Table 17.

6 Conclusion

We propose a selective bidirectional language projection framework to enhance the non-dominant capabilities of LLMs. Our method performs sequential language projection between the dominant and non-dominant languages within the projection-sensitive layers, thereby enabling efficient language shifts without introducing additional parameters. We conduct extensive experiments on diverse languages from different families on LLaMA3-8B and Qwen2.5-7B. The results on seven understanding and generation tasks demonstrate that our method significantly enhances the performance of target languages. Further analyses reveal that our method can efficiently bridge the representation gap and generalize to unseen languages.

Limitations

While our method effectively enhances the non-dominant language capabilities of LLMs, our method has the following limitations. First, our experiments currently focus on a single non-dominant language. Future work will expand evaluations to multiple languages and investigate cross-lingual interactions, which may further improve the transfer ability while reducing the cost for deploying separate models per language. Second, due to computational resource limitations, our experiments are constrained to two base models with up to 14B parameters. Validating the scalability and generalizability of our approach across larger and more diverse model architectures remains an important research direction. Lastly, the performance on extremely low-resource languages still falls behind other languages. In the future, we will explore modifying the model architecture or introducing additional training data to further enhance their performance.

Acknowledgments

We sincerely thank all the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057).

References

AI@Meta. 2024. Llama 3 model card.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. 2025. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23550–23558.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yuchun Fan, Yongyu Mu, YiLin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. SLAM: Towards efficient multilingual reasoning via selective language alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9499–9515, Abu Dhabi, UAE. Association for Computational Linguistics.

Gemini Team Google, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, and 1 others. 2025. Enhancing non-english capabilities of english-centric large language models through deep supervision fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24185–24193.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning

from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

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

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024b. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

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*, pages 10097–10113, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31730–31780, Suzhou, China. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

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, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. 2025. CC-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19036–19051, Vienna, Austria. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang Sun, Yujiu Yang, and Furu Wei. 2025. ShifCon: Enhancing non-dominant language capabilities with a shift-based multilingual contrastive framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4818–4841, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2024a. Lens: Rethinking multilingual enhancement for large language models. *arXiv preprint arXiv:2410.04407*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

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

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational*

Linguistics: ACL 2024, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Appendix

A Experiments

A.1 Experimental Settings

Training Setup. We employ two representative LLMs: *LLaMA3-8B* (AI@Meta, 2024) and *Qwen2.5-7B* (Yang et al., 2025) as the base models to validate the effectiveness of our method. We adopt the multilingual instruction dataset Bactrian-X (Li et al., 2023) as the training corpus, which contains 67k instruction-response pairs in each language. Since the instructions in Bactrian-X are constructed through translation, we directly utilize the instruction content from their respective datasets to obtain the translation pairs for projection-sensitive layer detection and language projection. We utilize *English* as the dominant language by default, since its data generally predominates in the pre-training corpora. Our experiments are implemented based on the LLaMAFactory framework (Zheng et al., 2024) using DeepSpeed (Rasley et al., 2020) on 4 NVIDIA A800-SXM4-80GB GPUs. The training duration is empirically set to 2 epochs with the learning rate of 2e-5, cosine learning rate scheduler with warm-up ratio of 0.03, maximum sequence length of 1024, and global batch size of 128. AdamW optimizer (Loshchilov and Hutter, 2019) is used to update the parameters during the training process with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Mixed precision training and ZeRO are applied within the DeepSpeed training framework to accelerate the training process and reduce memory usage (Rasley et al., 2020). We set k to 20 to categorize the N2D and D2N regions following Tezuka and Inoue (2025). Other hyperparameters are set according to Zheng et al. (2024).

Evaluation Tasks. We conduct evaluations on seven multilingual benchmarks, which can be categorized into understanding and generation tasks.

- **Multilingual Understanding**: (1) *BELEBELE* (Bandarkar et al., 2024), a multiple-choice reading comprehension dataset, (2) *MMMLU* (Lai et al., 2023), the multilingual version of MMLU (Hendrycks et al., 2020), designed to evaluate models' general knowledge, (3) *XCOPA* (Ponti et al., 2020), a multilingual dataset for causal commonsense reasoning, and (4) *XStoryCloze* (Lin et al., 2022), a multilingual commonsense reasoning dataset for evaluating story understanding.

Code	Language	Family
ar	Arabic	Afro-Asiatic
es	Spanish	Indo-European
en	English	Indo-European
hi	Hindi	Indo-European
id	Indonesian	Austronesian
sw	Swahili	Niger-Congo
th	Thai	Tai-Kadai
vi	Vietnamese	Austro-Asiatic
zh	Chinese	Sino-Tibetan

Table 5: Details of language information in this work.

- **Multilingual Generation**: (1) *MKQA* (Longpre et al., 2021), an open-domain question answering evaluation dataset, (2) *XQuAD* (Artetxe et al., 2020), a cross-lingual question answering dataset, and (3) *XLSUM* (Hasan et al., 2021), a multilingual abstractive summarization benchmark comprising multiple long news texts into a single sentence.

Baselines. For comparison, we consider the following baseline methods that enhance the multilingual capabilities of LLMs using different multilingual instruction fine-tuning methods:

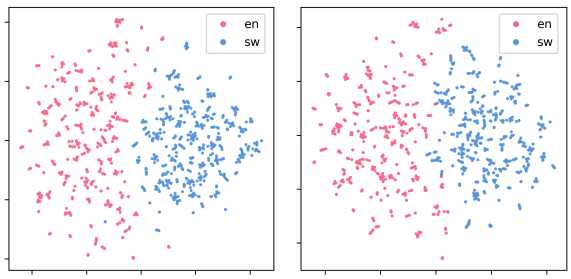
- **SFT** (Ouyang et al., 2022): which performs instruction fine-tuning with the English datasets.
- **TFT** (Chen et al., 2024), which is instruction-tuned using the target datasets translated from the original English datasets.
- **QAlign** (Zhu et al., 2024), which first aligns questions by translating the question in target languages into English before instruction-tuning.
- **SLAM** (Fan et al., 2025), which identifies and fine-tunes the FFN sublayers responsible for handling multilingualism in the lower-level layers. We fine-tune the first four FFN sublayers following their settings.
- **Incline** (Wang et al., 2025), which is an intervention approach that transforms source language representations into target language representation space during inference. We utilize Flores-devtest to calculate the transform matrix and tune the intervention parameter α on each testset within $[-1, 1]$.
- **DFT** (Huo et al., 2025), which incorporates additional supervision in the internal layers of

Dataset	#Lang	Languages	#samples	Metric
BELEBELE	8	Arabic, Spanish, Hindi, Indonesian, Swahili, Thai, Vietnamese, Chinese	900	Accuracy
MMMLU	6	Arabic, Spanish, Hindi, Indonesian, Vietnamese, Chinese	1000	Accuracy
XCOPA	5	Indonesian, Swahili, Thai, Vietnamese, Chinese	500	Accuracy
XStoryCloze	6	Arabic, Spanish, Hindi, Indonesian, Swahili, Chinese	1151	Accuracy
MKQA	5	Arabic, Spanish, Thai, Vietnamese, Chinese	1000	Accuracy
XQuAD	6	Arabic, Spanish, Hindi, Thai, Vietnamese, Chinese	1190	Accuracy
XLSUM	8	Arabic, Spanish, Hindi, Indonesian, Swahili, Thai, Vietnamese, Chinese	100	ROUGE-L

Table 6: Dataset Statistics.

Task	Pattern	Verbalizer
BELEBELE	Given the following passage, query, and answer choices, output the letter corresponding to the correct answer. Passage: `{passage}` Query: `{query}` Choices: `(A){choice_1}` `(B){choice_2}` `(C){choice_3}` `(D){choice_4}`	(A)||(B)||(C)||(D)
MMMLU	Given the following query and answer choices, output the letter corresponding to the correct answer. Query: `{query}` Choices: `(A){choice_1}` `(B){choice_2}` `(C){choice_3}` `(D){choice_4}`	(A)||(B)||(C)||(D)
XCOPA	Given the following premise, query, and answer choices, output the letter corresponding to the most plausible answer. Premise: `{premise}` {% if question == "cause" %} Which of the following two options is the cause of the above premise? {% else %} Which of the following two options is the result of the above premise? Choices: (A) `{choice_1}` (B) `{choice_2}`	(A)||(B)
XStoryCloze	Given the following story, query, and answer choices, output the letter corresponding to the correct answer. Story: `{story}` Query: Which of the following two options is more likely to be the ending of the given story? Choices: (A) `{choice_1}` (B) `{choice_2}`	(A)||(B)
MKQA	`{question}`	`{answer}`
XQUAD	Given the following passage, answer the question. Passage: `{passage}` Question: `{question}`	`{answer}`
XLSUM	Summarize this passage in one sentence. `{passage}`	`{summary}`

Table 7: The example prompt templates used for evaluation.



(a) LLaMA3-8B (b) LLaMA3-8B+Translation

Figure 11: Representation visualization of the 31st layer in LLaMA3-8B and our method fine-tuned with the English-Swahili translation data.

the model to constrain the conversion of the target language into English and to enforce reasoning in English. We follow their implementation for the feature-based method.

A.2 Evaluations

Datasets. We select a subset of representative languages from the entire set of languages in Bactrian-X, covering both high and low-resource languages for training and evaluation. The language subsets and the test samples used in the seven evaluation tasks in our experiments are presented in Tables 5 and 6. The evaluation prompt templates in English for each task are presented in Table 7. The templates are translated into target languages using ChatGPT for evaluation.

Metrics. For *BELEBELE*, *MMMLU*, *XCOPA*, *XStoryCloze*, *MKQA* and *XQuAD* datasets, *Accuracy* is utilized as the evaluation metric. Specifically, for the multiple-choice tasks of *BELEBELE*, *MMMLU*, *XCOPA* and *XStoryCloze*, a response is considered correct only if it contains the correct answer choice(s) and excludes all others. In contrast, for the generative QA tasks of *MKQA* and *XQuAD*, a response is deemed correct if it contains the gold answer. For *XLSUM* dataset, *ROUGE-L* scores are reported. We employ a greedy decoding strategy with a maximum of 40 new tokens for generation.

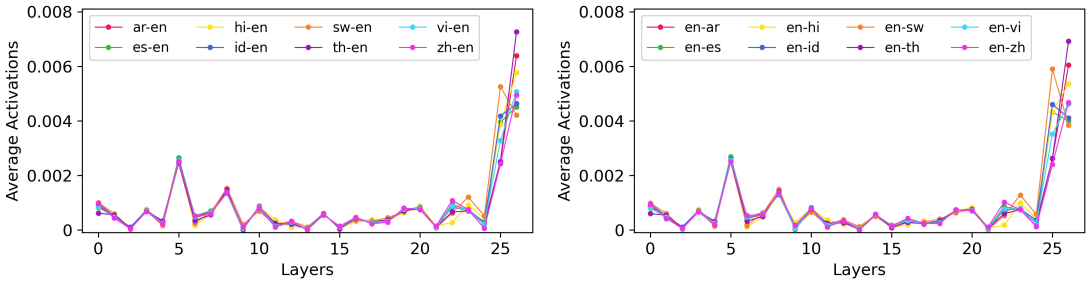
B Detailed Results

B.1 Results on Seven Benchmarks

Detailed performance results of LLaMA3-8B and Qwen2.5-7B on *BELEBELE*, *MMMLU*, *XCOPA*, *XStoryCloze*, *MKQA*, *XQuAD* and *XLSUM* dataset are listed in Tables 10-16, respectively.

B.2 Results on Similarity Search

We employ the accuracy of similarity search tasks as a quantitative indicator of cross-lingual representation alignment. We sample 500 parallel sentences



(a) N2D Projection

(b) D2N Projection

Figure 12: Layer-wise average activation changes across different language pairs in the Qwen2.5-7B.

Model	#Para.(LP)	#Para.(IT)
Ours(ATTN)	4.18%	29.80%
Ours(FFN)	17.55%	89.20%
Ours(BLOCK)	21.73%	100%

Table 8: Comparison of the amount of trainable parameters in each stage. '#Para.(LP)' denotes the number of parameters involved in language projection, while '#Para.(IT)' denotes the number of parameters involved in non-dominant instruction tuning.

Task	SFT	Projection Language			
		ar	hi	sw	zh
BELEBELE	70.00	72.66	69.44	71.78	70.78
MMMLU	48.50	50.50	48.70	49.70	49.40
XCOPA	84.00	84.80	84.00	85.20	82.20
XStoryCloze	78.62	75.58	80.34	79.15	73.39
MKQA	45.90	46.10	44.10	45.50	45.70
XQuAD	78.32	76.97	77.98	78.49	77.90
XLSUM	20.35	20.96	20.20	21.00	21.68

Table 9: Comparison of English performance between models with and without our SiLP on LLaMA3.2-3B.

between English and Swahili from Flores devtest, and obtain their sentence presentations across all layers via mean pooling. The average top-1 accuracy of sentence similarity research is presented in Figure 13. The search accuracy is notably improved with the D2N projection, demonstrating that representations are semantically aligned despite possessing distinct language-specific spaces.

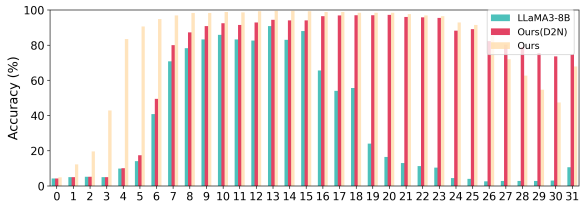


Figure 13: The similarity search top-1 accuracy on Flores devtest on English-Swahili language pair.

B.3 Discussion on Training Cost

We compare the number of parameters required for language projection and non-dominant instruction tuning on LLaMA3-8B, as summarized in Table 8. As shown in Figure 3, Ours(FFN) outperforms Ours(BLOCK) with fewer trainable parameters, yet has a higher training cost than Ours(ATTN). Compared to vanilla SFT and TFT, Ours(FFN) introduces only about 6.7% additional parameters (not external parameters) during fine-tuning.

B.4 Results on Experiments with French as the Dominant Language

To evaluate the generalization of our method, we conduct experiments using French as the dominant language instead of English, with a 1k parallel dataset sampled from the Flores devtest. The detailed results on each benchmark are presented in Table 17. The results show that the average performance on both NLU and NLG tasks is still improved in most cases for both LLaMA3-8B and Qwen2.5-7B.

B.5 Analysis on English Performance

To evaluate whether our alignment method can retain their English capabilities, we perform SFT with English instruction data on (i) vanilla LLaMA3.2-3B and (ii) the aligned LLaMA3.2-3B with our bidirectional language projection. The English performance for each projection language is summarized in Table 9. The model with bidirectional projection achieves comparable or slightly better English performance after instruction-tuning, confirming that the alignment of intrinsic parameters does not induce catastrophic forgetting or compromise the model's core reasoning and linguistic foundations in English.

Model	ar	es	hi	id	sw	th	vi	zh	AVG.
				LLaMA3-8B					
SFT	57.44	73.22	49.89	62.44	38.33	57.56	58.22	66.67	57.97
TFT	62.55	76.11	48.33	62.67	58.22	54.67	65.67	62.56	61.35
QAlign	53.44	69.11	44.67	58.33	44.67	54.22	63.44	67.33	56.90
SLAM	57.33	72.78	49.44	61.67	37.78	57.22	60.89	67.22	58.04
Incline	60.33	73.78	53.56	62.22	38.44	57.33	64.88	70.00	60.07
DFT	69.00	63.89	49.67	55.44	53.44	80.67	61.78	67.11	62.63
Ours	70.33	76.00	55.11	63.67	61.44	60.33	70.56	71.67	66.14
				Qwen2.5-7B					
SFT	82.44	87.00	62.11	82.44	47.00	76.11	83.33	87.11	75.94
TFT	78.33	85.44	60.89	83.67	72.11	73.00	81.67	85.78	77.61
QAlign	76.33	82.33	52.78	33.44	34.67	65.89	80.00	82.11	63.44
SLAM	75.89	82.67	53.89	60.22	32.11	66.00	80.67	86.78	67.28
Incline	83.78	86.77	60.78	82.67	48.00	77.88	81.33	86.67	75.99
DFT	79.33	85.78	60.89	83.00	71.89	72.67	81.89	86.44	77.74
Ours	80.67	86.33	61.78	83.77	73.44	72.44	82.00	86.67	78.39

Table 10: Detailed performance results of different languages on the BELEBELE task across all models.

Model	ar	es	hi	id	vi	zh	AVG.
				LLaMA3-8B			
SFT	32.00	42.50	31.70	43.90	40.40	43.00	38.92
TFT	36.00	43.60	36.00	43.90	42.00	39.60	40.18
QAlign	34.70	41.00	34.60	41.80	39.30	39.90	38.55
SLAM	32.40	42.20	32.00	41.40	40.00	42.50	38.42
Incline	33.90	45.90	35.40	43.20	40.30	43.90	40.43
DFT	45.90	42.40	34.40	41.90	35.80	44.80	40.87
Ours	38.50	47.80	37.00	45.00	43.40	44.30	42.67
				Qwen2.5-7B			
SFT	45.90	62.70	38.60	54.30	52.90	61.10	52.58
TFT	48.20	64.50	41.80	60.30	53.70	52.80	53.55
QAlign	45.30	58.90	39.50	44.20	50.90	57.80	49.43
SLAM	44.40	59.90	38.00	46.30	48.80	61.20	49.77
Incline	50.50	64.50	44.30	56.10	54.30	61.30	55.17
DFT	68.00	55.50	42.00	58.30	48.10	50.80	53.78
Ours	50.80	64.00	43.30	60.40	55.00	58.00	55.25

Table 11: Detailed performance results of different languages on the MMMLU task across all models.

Model	id	sw	th	vi	zh	AVG.
			LLaMA3-8B			
SFT	77.60	51.20	65.80	71.20	82.40	69.64
TFT	75.00	56.60	67.80	67.40	61.80	65.72
QAlign	71.60	58.60	66.20	58.80	77.80	66.60
SLAM	78.00	54.40	66.20	72.00	82.60	70.64
Incline	78.40	55.80	73.80	74.20	83.60	73.16
DFT	74.40	58.00	68.80	63.20	81.40	69.16
Ours	75.80	63.80	70.00	71.80	85.00	73.28
			Qwen2.5-7B			
SFT	85.00	31.60	71.80	81.00	92.20	72.32
TFT	87.20	74.00	34.60	82.20	87.20	73.04
QAlign	82.20	64.20	78.20	75.20	90.40	78.04
SLAM	83.40	45.60	72.60	76.40	93.40	74.28
Incline	85.40	49.60	74.00	89.40	92.60	78.20
DFT	86.00	75.80	62.80	83.40	73.60	76.32
Ours	87.20	76.00	37.60	85.40	88.20	74.88

Table 12: Detailed performance results of different languages on the XCOPA task across all models.

Model	ar	es	hi	id	sw	zh	AVG.
			LLaMA3-8B				
SFT	74.59	89.68	52.90	84.70	54.47	91.73	74.68
TFT	76.83	87.82	71.61	83.39	56.98	84.58	76.87
QAlign	76.64	85.44	75.25	71.60	73.26	88.09	78.38
SLAM	74.89	89.27	80.54	84.77	59.30	91.53	80.05
Incline	76.24	90.07	80.41	84.17	66.18	90.47	81.26
DFT	62.01	90.14	68.89	86.43	65.59	80.24	75.55
Ours	84.65	87.93	80.41	84.58	62.48	88.88	81.49
			Qwen2.5-7B				
SFT	82.79	95.43	76.97	87.36	53.87	91.26	81.28
TFT	86.63	93.58	83.85	90.21	82.13	92.72	88.19
QAlign	90.73	84.85	81.34	47.12	78.76	93.25	79.34
SLAM	81.34	92.06	67.84	87.16	52.61	94.10	79.19
Incline	93.64	96.10	87.76	87.75	53.93	96.23	85.90
DFT	83.72	93.51	86.57	89.21	87.36	90.00	88.40
Ours	92.26	95.57	87.29	90.47	83.06	93.71	90.39

Table 13: Detailed performance results of different languages on the XStoryCloze task across all models.

Model	ar	es	th	vi	zh	AVG.
			LLaMA3-8B			
SFT	10.20	28.50	15.40	23.90	13.20	18.24
TFT	13.10	36.20	17.30	29.40	15.70	22.34
QAlign	13.30	29.00	17.60	28.80	12.80	20.30
SLAM	10.40	27.90	15.50	24.10	13.00	18.18
Incline	11.00	28.60	15.10	24.80	13.70	18.64
DFT	12.90	35.60	17.70	29.60	15.80	22.32
Ours	13.20	35.90	17.90	30.90	16.10	22.80
			Qwen2.5-7B			
SFT	11.40	31.90	17.60	23.40	16.90	20.24
TFT	15.10	31.10	16.50	29.40	19.40	22.30
QAlign	11.20	30.70	17.30	26.30	14.00	19.90
SLAM	11.80	27.60	14.50	20.40	16.70	18.20
Incline	13.60	32.10	17.80	24.50	16.60	20.92
DFT	13.00	30.30	13.50	26.00	19.00	20.36
Ours	15.40	33.70	16.30	29.70	21.80	23.38

Table 14: Detailed performance results of different languages on the MKQA task across all models.

Model	ar	es	hi	th	vi	zh	AVG.
				LLaMA3-8B			
SFT	41.68	69.08	59.41	60.50	69.41	60.67	60.13
TFT	63.87	76.47	65.88	66.72	77.65	74.37	70.83
QAlign	61.60	28.07	12.61	13.03	23.03	8.57	24.49
SLAM	40.34	69.66	59.58	59.92	68.57	60.50	59.76
Incline	56.97	74.11	62.26	62.35	71.76	73.61	66.84
DFT	58.40	75.46	60.67	61.18	71.43	67.98	65.85
Ours	64.45	77.14	67.90	66.81	77.90	75.63	71.64
				Qwen2.5-7B			
SFT	67.06	80.76	47.30	73.45	77.06	80.20	70.97
TFT	66.47	78.82	49.42	69.08	80.67	82.10	71.09
QAlign	10.00	29.26	29.16	13.45	24.20	10.67	19.46
SLAM	62.02	74.87	33.95	72.10	68.57	81.43	65.49
Incline	66.89	80.50	47.48	72.26	77.39	81.60	71.02
DFT	61.60	76.47	46.55	63.61	77.23	76.63	67.02
Ours	66.56	78.32	49.33	70.50	81.09	82.94	71.46

Table 15: Detailed performance results of different languages on the XQUAD task across all models.

Model	ar	es	hi	id	sw	th	vi	zh	AVG.
LLaMA3-8B									
SFT	2.37	16.58	6.46	11.81	18.53	16.75	16.48	20.38	13.67
TFT	15.21	20.97	19.61	19.59	21.98	17.65	18.85	23.76	19.68
QAlign	13.90	4.57	0.30	5.78	6.21	17.58	3.72	8.53	7.57
SLAM	2.88	16.53	6.76	11.93	20.02	16.34	16.61	20.36	13.93
Incline	11.23	17.62	17.01	14.80	16.65	17.08	16.68	22.62	16.71
DFT	15.90	21.09	19.58	19.98	21.87	17.79	18.68	22.92	19.73
Ours	16.08	21.18	19.25	20.02	23.24	17.96	17.90	24.24	19.98
Qwen2.5-7B									
SFT	13.13	19.59	13.84	21.30	19.71	17.16	16.80	23.16	18.09
TFT	18.03	21.35	17.12	20.55	21.11	16.13	18.75	24.56	19.69
QAlign	4.01	5.78	2.34	6.28	6.02	15.96	3.87	10.53	6.85
SLAM	9.56	19.93	15.37	22.22	18.00	18.44	17.06	25.47	18.26
Incline	16.66	19.91	16.67	19.11	18.78	18.13	18.30	23.92	18.94
DFT	18.46	21.10	16.60	19.90	22.80	16.51	18.25	24.55	19.77
Ours	19.00	21.77	17.01	20.26	23.65	16.85	18.25	23.11	19.99

Table 16: Detailed performance results of different languages on the XLSUM task across all models.

Model		ar	es	hi	id	sw	th	vi	zh	AVG.
Dataset: BELEBELE										
LLaMA3-8B	TFT	62.55	76.11	48.33	62.67	58.22	54.67	65.67	62.56	61.35
	Ours	67.22	79.67	52.22	61.33	63.89	60.56	70.78	58.22	64.24
Qwen2.5-7B	TFT	78.33	85.44	60.89	83.67	72.11	73.00	81.67	85.78	77.61
	Ours	79.44	85.78	60.89	83.89	71.89	72.67	81.89	86.56	77.88
Dataset: MMMLU										
LLaMA3-8B	TFT	36.00	43.60	36.00	43.90	–	–	42.00	39.60	40.18
	Ours	36.00	43.20	35.20	44.10	–	–	42.60	42.00	40.52
Qwen2.5-7B	TFT	48.20	64.50	41.80	60.30	–	–	53.70	52.80	53.55
	Ours	49.20	64.20	43.60	60.50	–	–	54.20	52.00	53.95
Dataset: XCOPA										
LLaMA3-8B	TFT	–	–	–	75.00	56.60	67.80	67.40	61.80	65.72
	Ours	–	–	–	76.00	62.40	67.60	70.60	82.20	71.76
Qwen2.5-7B	TFT	–	–	–	87.20	74.00	34.60	82.20	87.20	73.04
	Ours	–	–	–	87.20	74.40	37.80	80.60	87.60	73.52
Dataset: XStoryCloze										
LLaMA3-8B	TFT	76.83	87.82	71.61	83.39	56.98	–	–	84.58	76.87
	Ours	85.97	91.19	64.99	69.89	69.03	–	–	85.84	77.82
Qwen2.5-7B	TFT	86.63	93.58	83.85	90.21	82.13	–	–	92.72	88.19
	Ours	86.57	95.76	86.57	90.60	81.54	–	–	93.71	89.13
Dataset: MKQA										
LLaMA3-8B	TFT	13.10	36.20	–	–	–	17.30	29.40	15.70	22.34
	Ours	12.10	37.90	–	–	–	18.50	32.90	16.30	23.54
Qwen2.5-7B	TFT	15.10	31.10	–	–	–	16.50	29.40	19.40	22.30
	Ours	14.60	32.60	–	–	–	15.40	28.50	20.80	22.38
Dataset: XQuAD										
LLaMA3-8B	TFT	63.87	76.47	65.88	–	–	66.72	77.65	74.37	70.83
	Ours	64.20	77.56	68.07	–	–	67.06	78.40	75.80	71.85
Qwen2.5-7B	TFT	66.47	78.82	49.42	–	–	69.08	80.67	82.10	71.09
	Ours	66.81	78.49	49.41	–	–	70.42	81.26	83.27	71.61
Dataset: XLSUM										
LLaMA3-8B	TFT	15.21	20.97	19.61	19.59	21.98	17.65	18.85	23.76	19.68
	Ours	16.29	20.70	19.28	20.00	20.44	18.50	18.28	23.85	19.67
Qwen2.5-7B	TFT	18.03	21.35	17.12	20.55	21.11	16.13	18.75	24.56	19.69
	Ours	18.15	21.61	16.34	20.35	22.70	16.56	18.22	22.58	19.56

Table 17: Detailed results on NLU and NLG tasks with French as the dominant language for language projection.