CC-TUNING: A *C*ross-Lingual *C*onnection Mechanism for Improving Joint Multilingual Supervised Fine-Tuning

Yangfan Ye[1], **Xiaocheng Feng**[1,2*], **Zekun Yuan**[1], **Xiachong Feng**[3], **Libo Qin**[4], **Lei Huang**[1] **Weitao Ma**[1], **Yichong Huang**[1], **Zhirui Zhang, Yunfei Lu**[5], **Xiaohui Yan**[5], **Duyu Tang**[5], **Dandan Tu**[5], **Bing Qin**[1,2]

[1]Harbin Institute of Technology [2]Peng Cheng Laboratory [3]The University of Hong Kong
[4]Central South University [5]Huawei Technologies Co., Ltd

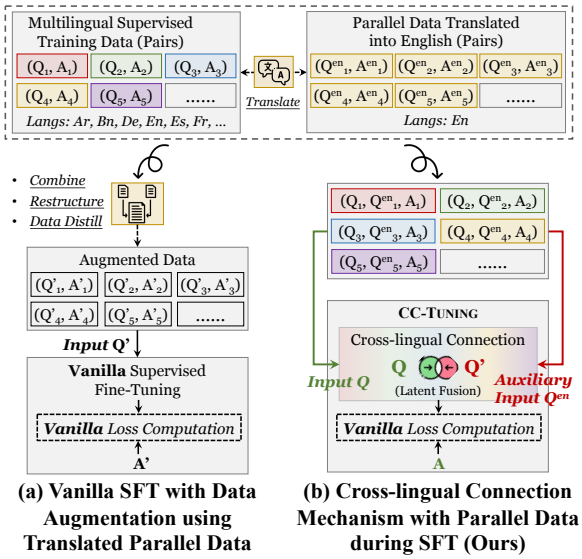
{yfye,xcfeng,qinb}@ir.hit.edu.cn

Abstract

Current large language models (LLMs) often exhibit imbalanced multilingual capabilities due to their English-centric training corpora. To address this, existing fine-tuning approaches operating at the *data-level* (e.g., through data augmentation or distillation) typically introduce implicit cross-lingual alignment, overlooking the potential for more profound, *latent-level*[1] cross-lingual interactions. In this work, we propose CC-TUNING, a novel multilingual fine-tuning paradigm that explicitly establishes a cross-lingual connection mechanism at the latent level. During training, CC-TUNING fuses the feed forward activations from both English and non-English inputs, enabling the model to benefit from both linguistic resources. This process is facilitated with a trainable *Decision Maker* that identifies beneficial activations. Furthermore, during inference, a *Transform Matrix* is utilized to simulate the cross-lingual connection under monolingual setting through representation transformation. Our experiments on six benchmarks covering 22 languages show that CC-TUNING outperforms vanilla SFT and offers a strong latent-level alternative to data-level augmentation methods. Further analysis also highlights the practicality of CC-TUNING and the potential of latent-level cross-lingual interactions in advancing the multilingual performance of LLMs. (Code link: CC-Tuning)

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated exceptional capabilities in handling diverse tasks (Dong et al., 2023; Wei et al., 2022a,b; Shanahan, 2022; Zhao et al., 2023; Liu et al., 2023; Huang et al., 2025) while exhibiting promising generalizability across diverse languages (Ye et al., 2023; Qin et al., 2024; Huo et al., 2025). However, significant performance disparities persist across languages due to the overwhelming dominance of English in training corpora, making balanced multilingual proficiency an ongoing research challenge (Touvron et al., 2023; Zhang et al., 2023; Ye et al., 2024a).



(a) Vanilla SFT with Data Augmentation using Translated Parallel Data **(b) Cross-lingual Connection Mechanism with Parallel Data during SFT (Ours)**

Figure 1: Comparison between vanilla supervised fine-tuning with data augmentation at **data level** (implicit) and our method at **latent activation level** (explicit).

One of the prevailing approaches towards these challenges focuses on joint multilingual supervised fine-tuning (SFT) (Ouyang et al., 2022), which refers to fine-tuning the model with supervised data spanning multiple languages. While effective in principle, these methods encounter the "curse of multilinguality" – a paradoxical phenomenon where expanding language coverage during joint training leads to performance degradation across both high- and low-resource languages (Conneau et al., 2020; Wang et al., 2020).

To address this, current studies primarily focus on data-level interventions through parallel corpus utilization. Common strategies include: multilingual data augmentation with English-aligned parallel examples (Aharoni et al., 2019; Shaham et al.,

*Corresponding Author

[1]*latent-level*: referring to direct manipulation of the model's internal representations (e.g., FFN activations)

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19036–19051
July 27 - August 1, 2025

2024), explicit translation task formulation (Johnson et al., 2017; Tang et al., 2020), and response distillation from resource-rich languages (Zhang et al., 2024). While these methods demonstrate partial success, their reliance on implicitly introducing data-level text alignment overlooks the potential for deeper, latent-level cross-lingual interactions.

We propose CC-TUNING, a novel multilingual fine-tuning paradigm that introduces *explicit cross-lingual connections* at the latent activation level by fusing feed-forward activations from English and non-English languages (Figure 1). This approach is grounded in recent empirical findings highlighting the significant potential of feed-forward activations in improving model's multilingual performance (Ye et al., 2024b). During training, our method leverages parallel bilingual inputs and incorporates a trainable *Decision Maker* to identify linguistically beneficial signals from auxiliary English activations, integrating them into the forward propagation of non-English inputs. Additionally, during inference, an "easy-to-learn" *Transform Matrix* is utilized to simulate the cross-lingual connection without the parallel bilingual inputs, ensuring the practicality of our approach. This latent-level interaction mechanism fundamentally differs from conventional data-level approaches, as it establishes direct interlingual activation connections rather than relying on statistical correlations in training data.

To validate our approach, we conduct extensive experiments across six benchmarks encompassing both natural language understanding and generation tasks, spanning 22 languages using two representative LLMs. Our results highlight the superiority of CC-TUNING over vanilla SFT in multilingual joint learning scenarios. Besides, compared to data-level augmentation or distillation methods that leverage parallel data, CC-TUNING offers a highly effective alternative for facilitating cross-lingual interaction. Additionally, our further ablation studies and analysis also provide strong evidence of the practicality and robustness of CC-TUNING.

2 Related Work

Multilingual Large Language Models. Recently, larger models such as Bloom (Scao et al., 2022), Mala-500 (Lin et al., 2024) and Aya Model (Üstün et al., 2024) have pushed multilingual performance further by leveraging the benefits of greater scale. Generally, multilingual pretraining and fine-tuning are now the two mainstream methods for improving multilingual capabilities. Models such as Sabia (Pires et al., 2023), ChineseLLaMA (Cui et al., 2023), ChineseMixtral (HIT-SCIR, 2024), PolyLM (Wei et al., 2023) and PaLM2 (Anil et al., 2023) have been developed through (continuous) pretraining with large multilingual corpora or language-specific data. Other models like BLOOMz (Muennighoff et al., 2022), m-LLaMA (Zhu et al., 2023), Camoscio (Santilli and Rodolà, 2023), Phoenix (Chen et al., 2023) and Bode (Garcia et al., 2024) have opted for leveraging multilingual or language-specific data directly during SFT stage to foster cross-lingual alignment.

Multilingual Supervised Fine-Tuning. Multilingual SFT is an effective way to enhance the multilingual performance of LLMs. Current research often focuses on data augmentation or distillation techniques to enrich training data and improve model generalization across multiple languages. For instance, Pan et al. (2024) highlighted the importance of diverse, high-quality data for machine translation fine-tuning, while Li et al. (2023) addressed "translationese" by using Google Translate and ChatGPT for multilingual response generation. In terms of instruction tuning, Shaham et al. (2024) showed that adding multilingual examples to English-centric fine-tuning significantly boosts multilingual instruction-following, while Chen et al. (2024) demonstrated the superiority of multilingual tuning over language-specific training. Translation-based fine-tuning has been shown to enhance semantic alignment, as argued by Ranaldi et al. (2024). Similarly, Zhu et al. (2023) combined translation data, cross-lingual tasks, and scaling laws to optimize multilingual performance. Additionally, Zhang et al. (2024) proposed a self-distillation approach leveraging LLMs' internal capabilities in resource-rich languages to enhance multilingual performance.

The above methods primarily focus on enriching training data with parallel data to foster implicit cross-lingual alignment. In contrast, our CC-TUNING emphasizes improving the training paradigm by explicitly incorporating cross-lingual latent interactions into the training process.

3 Method

In this section, we first revisit the vanilla multilingual supervised fine-tuning paradigm, then present the training implementation of CC-TUNING and its specialized configurations during inference stage.

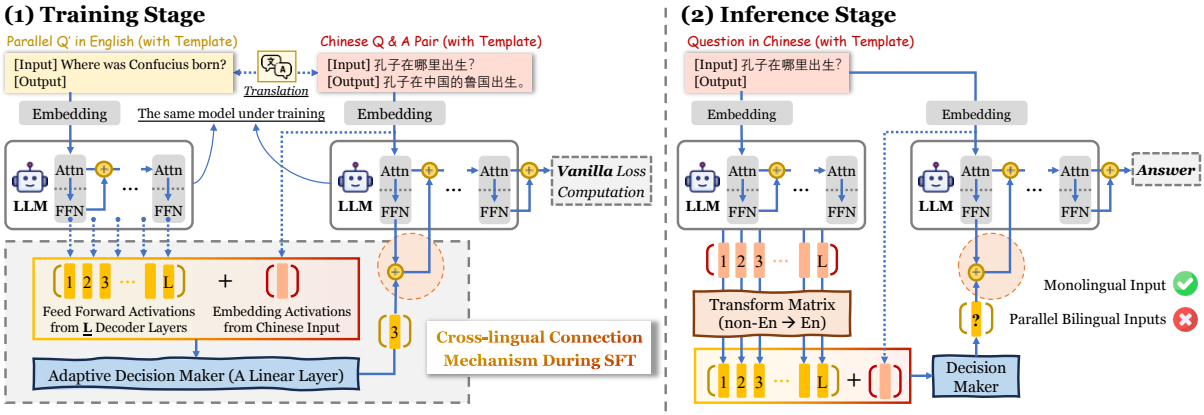


Figure 2: Overview of the cross-lingual connection mechanism in CC-TUNING. In the training stage, CC-TUNING leverages an auxiliary English input alongside the non-English input, while retaining the vanilla loss computation without introducing additional training objectives. In the inference stage, a transform matrix is used to simulate cross-lingual connection in monolingual input scenarios, eliminating the dependence on bilingual parallel input.

3.1 Multilingual Supervised Fine-Tuning

Multilingual supervised fine-tuning enables pre-trained models to better perform downstream tasks across diverse languages through training on annotated multilingual instruction dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where N represents the size of the dataset, x_i denotes the input question or instruction, and y_i is the corresponding expected output or response. The training process is required to minimize the following objective of negative log-likelihood of the predicted output with respect to the ground-truth response. θ denotes the parameters of the model.

$$\mathcal{L}_{SFT}(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log P(y_i|x_i, \theta) \quad (1)$$

Data Augmentation with Parallel Data. For the multilingual instruction dataset D, we define its corresponding English parallel data as D_{en}. Several previous studies have explored enriching the original training data by merging these two datasets, incorporating additional translation task form data constructed from parallel pairs, or utilizing techniques such as distillation. We collectively refer to these augmented datasets as $D_{aug} = \{(x_i^{aug}, y_i^{aug})\}_{j=1}^{M}$. These approaches, in essence, do not alter the SFT process; rather, they introduce additional supervised data, as illustrated below:

$$\mathcal{L}_{SFT_{aug}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log P(y_i|x_i, \theta) + \frac{1}{M}\sum_{j=1}^{M} -\log P(y_i^{aug}|x_i^{aug}, \theta) \quad (2)$$

3.2 CC-TUNING

We will introduce cross-lingual connection mechanism in CC-TUNING in detail, focusing on its implementation during training and inference stages.

3.2.1 Training with Cross-lingual Connection

Motivated by the findings in Ye et al. (2024b), which empirically demonstrate that feed-forward activations from English hold the potential to significantly enhance a model's performance in non-English languages. The cross-lingual connection mechanism in CC-TUNING aims to incorporate the above latent interactions into the multilingual fine-tuning process, enabling the model to benefit from both English and non-English resources as the parameters are updated.

We denote $D = \{(x_i, y_i)\}_{i=1}^{N}$ as a multilingual supervised instruction dataset, where x_i represents the input question for the i-th data point and y_i denotes the corresponding ground-truth response. Besides, CC-TUNING requires auxiliary parallel data, $D^{en} = \{(x_i, x_i^{en}, y_i)\}_{i=1}^{N}$, where x_i^{en} is the English translation of x_i. Generally, the cross-lingual connection mechanism consists of two key operations: (1) adaptive decision maker and (2) latent feed forward connection. Notably, these operations are executed just before the *Response Start Token (RST)*, which marks the beginning of the model's response in the training template. This ensures that our operations can smoothly introduce the intervention into the response generation process. Assuming the training template is structured as "*[Input]* {question} *[output]* {answer}", these operations are executed at the position that is right before the *[output]* token.

Adaptive Decision Maker. Given an auxiliary input x_i^{en}, we first pass it through the model to extract its feed-forward activations $F_i^{en} \in \mathbb{R}^{L\times d} = \{f_{i,l}^{en}\}_{l=1}^{L}$ from L decoder layers, where d is the dimensionality of the hidden states. Notably, prior research has shown that not all feed-forward activations contribute equally and some may degrade performance (Ye et al., 2024b). To mitigate this issue, we introduce a trainable linear layer $W_{DM} \in \mathbb{R}^{d\times L}$, referred to as the *Decision Maker*, which adaptively selects the most beneficial layer. By combining F_i^{en} with the embedding activations $e_i \in \mathbb{R}^d$ of x_i, we integrate features from both English and non-English inputs. The resulting combined features are then fed into the *Decision Maker* along with *Gumbel-Softmax* (Jang et al., 2016) to achieve the identification as follows:

$$H_i = \frac{1}{L}\sum_{l=1}^{L}(f_{i,l}^{en} + e_i)\cdot W_{DM} \tag{3}$$

$$f_{i,s}^{en} = \sum_{l=1}^{L}\text{Gumbel-Softmax}(H_i)_l \cdot f_{i,l}^{en} \tag{4}$$

where $f_{i,s}^{en} \in \mathbb{R}^d$ represents the selected activation from the s-th layer among the L decoder layers.

Latent Feed Forward Connection. The second step aims to transfer the beneficial activation $f_{i,s}^{en}$ identified in the previous step into the forward propagation process of non-English input. When the input x_i is fed into the model, let the output of all L decoders be denoted as $O_i = \{o_{i,l}\}_{l=1}^{L}$, where each $o_{i,l}$ should have been obtained by combining the feed-forward activations $f_{i,l}$ and self-attention activations $a_{i,l}$ through a residual connection. However, the incorporation of $f_{i,s}^{en}$ refines this process by connecting itself with the feed-forward activation $f_{i,1}$ from the first decoder layer. Formally, this modification can be expressed as:

$$\tilde{f}_{i,1} = f_{i,1} + f_{i,s}^{en} \tag{5}$$

The forward propagation of the input x_i then continues with this modification. Consequently, the original decoder outputs $\{o_{i,l}\}_{l=j}^{L}$ will be altered to $\{\tilde{o}_{i,l}\}_{l=j}^{L}$ due to the update of $f_{i,1} \rightarrow \tilde{f}_{i,1}$, leading to new final prediction outcomes $\tilde{o}_{i,L}$.

And within CC-TUNING, the training objective remains the same as the vanilla loss objective in Equation 1. During the tuning process, the model itself, along with the *Decision Maker*, learns to leverage the benefits of both English and non-English resources, improving its multilingual capabilities.

3.2.2 Inference with Transform Matrix

Unlike the training stage, our inference process is conducted without the need for parallel inputs. Instead, we leverages a training-free *Transform Matrix* to simulate the cross-lingual connection.

The role of the *Transform Matrix* W_T here is to achieve the transformation of $F_i = \{f_{i,l}\}_{l=1}^{L} \rightarrow F_i^{en} = \{f_{i,l}^{en}\}_{l=1}^{L}$ in the absence of parallel English input x_i^{en}. Specifically, after training, we firstly sample 1,000 parallel pairs (x_i, x_i^{en}) from the datasets D and D^{en}, and collect their feed-forward activations, F_i and F_i^{en}, respectively. These activations are then stacked and denoted as $A = \{f_{i,l} \mid i = 1, ..., N; l = 1, ..., L\}$ and $B = \{f_{i,l}^{en} \mid i = 1, ..., N; l = 1, ..., L\}$. Therefore, A can be mapped into B as follows through W_T:

$$A \cdot W_T = B \tag{6}$$

To minimize the difference A and B, our objective is defined as follows (Least-Squares optimization):

$$\boldsymbol{W}_T^* = \underset{\boldsymbol{W}_T}{\text{argmin}} \sum_{i=1}^{N}\sum_{l=1}^{L}\left\| f_{i,l}\boldsymbol{W}_T - f_{i,l}^{en}\right\|^2 \tag{7}$$

This problem seeks the optimal $\boldsymbol{W}_T^*$ that minimizes the distance between the source and target representations. Hence, the closed-form solution to this optimization problem is:

$$\boldsymbol{W}_T^* = \left(\sum_{i=1}^{N}\sum_{i=l}^{L}(f_{i,l})^T f_{i,l}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{i=l}^{L}(f_{i,l})^T f_{i,l}^{en}\right) \tag{8}$$

Once the optimal $\boldsymbol{W}_T$ has been learned, it can be applied to the non-English representation to map it to the corresponding English representation. This resulting mapped representation $F_i \cdot W_T$, then substitutes $F_i^{en} = \{f_{i,l}^{en}\}_{l=1}^{L}$ in equations 3, 4, 5, thereby simulating the cross-lingual connection. This alignment effectively eliminates the dependence for bilingual parallel data and enables the simulation of cross-lingual connection in a monolingual scenario.

4 Experiments

4.1 Setup

Models. We selected two representative LLMs: (1) *LLaMA-3.1-8B* (Dubey et al., 2024) and (2) *Qwen2.5-7B* (Yang et al., 2024).

Training Corpus. We totally select 20,236 multilingual instruction pairs from *aya dataset* (Singh

Method	Multilingual Understanding						Multilingual Generation					
	XNLI		XStoryCloze		MMMLU		MKQA		XQuAD		XLSum	
	LLaMA.	Qwen.	LLaMA.	Qwen.	LLaMA.	Qwen.	LLaMA.	Qwen.	LLaMA.	Qwen.	LLaMA.	Qwen.
Baselines												
ML-SFT	31.88	48.23	65.23	70.06	40.20	**50.05**	14.64	14.73	60.42	63.61	12.27	12.40
+EN	35.02	50.76	65.13	71.63	39.62	48.80	13.28	13.05	57.40	62.34	12.04	12.20
+MT	35.90	47.05	69.90	70.50	40.68	47.49	13.56	13.54	58.40	64.03	12.89	12.48
+SDRRL	29.74	52.36	55.82	80.67	28.06	47.28	–	–	–	–	–	–
Ours												
CC-TUNING	**38.42** (+6.54)	51.00 (+2.77)	70.60 (+5.37)	71.43 (+1.37)	**40.74** (+0.54)	49.65 (-0.40)	**15.94** (+1.30)	**14.84** (+0.11)	**61.85** (+1.21)	63.72 (+0.11)	12.88 (+0.61)	12.50 (+0.10)
+EN	32.72 (-2.30)	49.48 (-1.28)	60.94 (-4.19)	64.69 (-6.94)	38.73 (-0.89)	47.35 (-1.45)	14.61 (+1.33)	13.56 (+0.51)	60.89 (+3.40)	62.69 (+0.35)	12.78 (+0.74)	12.63 (+0.43)
+MT	36.44 (+0.54)	48.13 (+1.08)	**73.54** (+3.64)	71.39 (+0.89)	38.87 (-1.81)	49.39 (+1.90)	15.59 (+2.03)	13.77 (+0.23)	61.55 (+3.10)	**64.26** (+0.23)	**13.05** (+0.16)	**12.87** (+0.39)
+SDRRL	29.84 (+0.10)	**53.06** (+0.70)	69.19 (+13.37)	**80.93** (+0.26)	37.77 (+9.71)	47.87 (+0.59)	– –	– –	– –	– –	– –	– –

Table 1: Main results that are the averages of the performance across all languages involved for each dataset. Blue cell indicates better performance than the vanilla ML-SFT under the same training data setting, while Gray cell indicates the opposite. **Bold** numbers indicate the best performance. LLaMA. and Qwen. respectively represent *LLaMA-3.1-8B* and *Qwen2.5-7B*.

et al., 2024) as our training corpus and the multilingual training corpus covers more than 60 languages, ensuring extensive multilingual coverage. Our training processes are conducted on *8 * A800-SXM4-80GB* with the following settings: *batch size=16*, *epochs=3*, *learning rate=1.0e-5*, *warmup ratio=0.1*, and *bf16=true*. The implementation is based on *LLaMA-Factory* (Zheng et al., 2024).

Baselines. More details are in Appendix A.1.

- **ML-SFT** represents vanilla supervised instruction tuning (Ouyang et al., 2022) with original multilingual instruction dataset (data size=N).
- **ML-SFT+EN** incorporates the full parallel English version of the dataset for training, followed by vanilla supervised fine-tuning (data size=$2N$).
- **ML-SFT+MT** constructs additional translation task form data by pairing the original multilingual instruction dataset with its parallel English version and then applies supervised instruction tuning (data size=$2N$).
- **ML-SFT+SDRRL** (Zhang et al., 2024) is a self-distillation-based method that integrates English instruction tuning data and its multilingual code-switched extensions. Additionally, it incorporates partially translated data and completion data for fine-tuning (LLaMA-3.1-8B: data size$\approx 1.2N$, Qwen2.5-7B: data size$\approx 1.6N$).

And **CC-TUNING (+EN, +MT, +SDRRL)** refers to our method applying the cross-lingual connection mechanism and its combination with different above mentioned training data settings.

Evaluation Datasets. We conduct experiments on 6 benchmarks, which can be categorized into:

- **Multilingual Understanding:** (1) *XNLI* (Conneau et al., 2018), a multilingual natural language inference (NLI) dataset, (2) *XStoryCloze* (Lin et al., 2022), a multilingual commonsense reasoning dataset for evaluating story understanding and (3) *MMMLU*, the multilingual version of *MMLU* (Hendrycks et al., 2020), designed to evaluate models' general knowledge.
- **Multilingual Generation:** (1) *MKQA* (Longpre et al., 2021), an open-domain multilingual question answering evaluation dataset, (2) *XQuAD* (Artetxe et al., 2020), a question answering dataset and (3) *XLSum* (Hasan et al., 2021), a multilingual abstractive summarization benchmark comprising professionally annotated article-summary pairs.

For each of the above datasets, we conduct experiments on 10 language subsets, covering a total of 22 languages. For *XNLI*, *XStoryCloze*, *MMMLU*, *MKQA* and *XQuAD* datasets, *Accuracy* metric is used for evaluation. And for *XLSum* dataset, *ROUGE-L* scores are reported. We use greedy decoding with a max of 40 new tokens for each model. Detailed information on the datasets and evaluations can be found in Appendix A.2.

4.2 Main Results

The average results across the different languages involved in each dataset are presented in Table 1. The detailed results for different languages can be found in Table 7, 8. Note that the results of applying **+SDRRL** to NLG tasks are not reported, as it may lead to deviations from the prompt language in model responses, as shown in Appendix A.3.

(1) CC-TUNING outperforms vanilla SFT in joint multilingual learning scenarios. The results in Table 1 demonstrate that under the same multilingual training data settings of original data, +MT and +SDRRL, CC-TUNING significantly outperforms vanilla SFT in both multilingual understanding and multilingual generation tasks. However, under the +EN setting, where more than half of the training data is in English, the cross-lingual connection becomes an EN2EN connection. This shift undermines the core goal of CC-TUNING—to promote cross-lingual latent interaction—leading to a notable decline in performance, which also emphasizes CC-TUNING's alignment with its motivation and use case in joint multilingual learning scenarios.

(2) CC-TUNING with original training data outperforms data augmentation and distillation methods on *LLaMA-3.1-8B*. As observed on *LLaMA-3.1-8B*, CC-TUNING, even when trained solely with the original dataset (data size = N), outperforms the data augmentation and distillation approaches of ML-SFT+EN (data size = $2N$), +MT (data size = $2N$), and +SDRRL (data size $\approx 1.2N$), which utilize larger training set. This suggests that, compared to implicitly introducing cross-lingual alignment information at the data level, the explicit latent-level cross-lingual connection mechanism in CC-TUNING provides a compelling alternative for facilitating cross-lingual interaction.

4.3 Ablation Studies

We perform ablation studies to assess the following aspects: (1) the effectiveness of the *Transform Matrix*, (2) the necessity of the *Decision Maker*, and (3) the advantages of feed-forward activations in facilitating cross-lingual interactions.

(1) The *Transform Matrix* aligns well with the effect of using parallel bilingual inputs. We verify whether the *Transform Matrix* W_T can effectively achieve the alignment by evaluating the mean squared error (MSE) between $f_{i,l} \cdot W_T$ and $f^{en}_{i,l}$ as well as the performance difference $|\Delta|$ between using parallel bilingual inputs during inference and applying the *Transform Matrix*. The results in Table 2 show that the MSE value reaches the order of magnitude as low as 10^{-2}, indicating that the *Transform Matrix* effectively transforms $f_{i,l}$ into $f^{en}_{i,l}$. Additionally, the small performance difference $|\Delta|$ further suggests that the *Transform Matrix* serves as an effective substitute for parallel bilingual inputs, achieving great alignment.

Method ($\|M\| = 1000$)	XNLI	XStoryCloze	MMMLU	MKQA	XQuAD	XLSum
	Model: LLaMA-3.1-8B					
MSE value	*MSE* = $\frac{1}{N \times L}\sum_{i=1}^{N}\sum_{l=1}^{L}(\frac{1}{d}\|f_{i,l} \cdot W_T - f^{en}_{i,l}\|_2^2)$					
CC-TUNING	0.012427	0.013256	0.013196	0.021572	0.015428	0.014314
+EN	0.014215	0.012734	0.012917	0.018137	0.015919	0.014046
+MT	0.020413	0.021251	0.023016	0.027770	0.021769	0.025639
+SDRRL	0.017896	0.019859	0.017098	–	–	–
AVG.MSE	0.016238	0.016775	0.016557	0.022493	0.017705	0.018000
$\|\Delta\|$ value	$\|\Delta\|$ = | *Result(Parallel Bilingual Input) - Result(Transform Matrix)* |					
CC-TUNING	0.16	0.60	0.03	0.01	0.21	0.28
+EN	0.08	0.31	0.25	0.01	0.08	0.16
+MT	0.01	0.23	0.36	0.12	0.07	0.10
+SDRRL	0.06	0.56	0.11	–	–	–
AVG.$\|\Delta\|$	0.08	0.43	0.19	0.05	0.12	0.18

Table 2: The results of mean squared error between feed-forward representations in English and the transformed representations after applying the *Transform Matrix*, as well as the performance difference $|\Delta|$ between using parallel bilingual inputs and applying *Transform Matrix*.

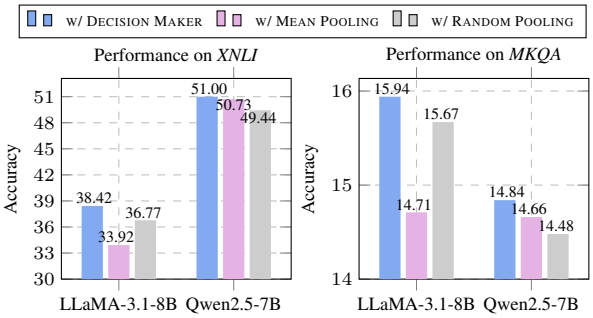


Figure 3: Performance comparisons of using *Decision Maker*, *Mean Pooling* and *Random Pooling* strategy on *XNLI* and *MKQA* datasets.

(2) The *Decision Maker* plays a crucial role. To verify the necessity of the *Decision Maker*, we replaced it with two alternative strategies—*Mean Pooling* and *Random Pooling*—during both training and inference, and compared their performance in Figure 3. In *Mean Pooling*, the feed-forward activations from all layers are averaged, while in *Random Pooling*, a single activation is randomly selected from the set of feed-forward activations across all layers. The results demonstrate that the performance with the *Decision Maker* significantly outperforms the other two strategies, confirming that

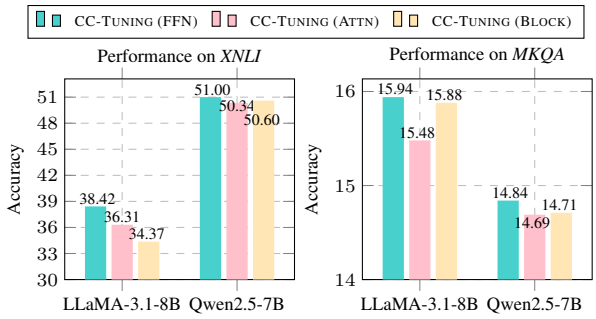


Figure 4: Performance comparisons of utilizing feed forward activations, self-attention activations and whole decoder block activations for cross-lingual connection on *XNLI* and *MKQA* datasets.

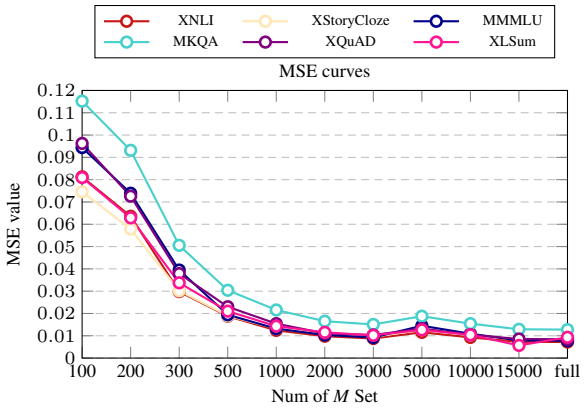


Figure 5: The curves of mean squared error between feed-forward representations in English and the transformed representations after applying the *Transform Matrix*, as the amount of parallel data used to acquire the *Transform Matrix* increases.

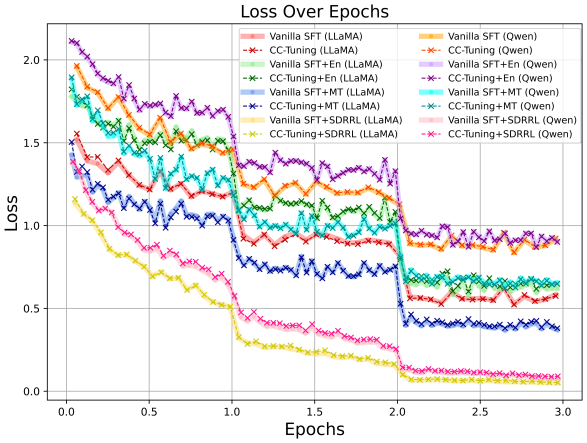


Figure 6: The training loss curves of vanilla supervised fine-tuning and CC-TUNING under different training settings (models and training data).

the *Decision Maker* effectively serves its role in beneficial activation identification and contributes to the overall training paradigm of CC-TUNING.

(3) Feed-forward activations contribute the most in cross-lingual connection. In addition to investigating cross-lingual connections at the feed-forward activation level, we also explored the potential contributions of self-attention activations and whole decoder block activations. Our results, as shown in Figure 4, indicate that feed-forward activations have the most pronounced impact on cross-lingual connections within the CC-Tuning paradigm. This finding highlights the crucial role of feed-forward activations in facilitating cross-lingual latent interactions, which well match the findings presented in Dai et al. (2022), where FFN stores factual knowledge, as well as the motivation of cross-lingual feed forward transplantation operation in Ye et al. (2024b).

5 Further Analysis

5.1 Practicality Analysis

(1) Is the *Transform Matrix* difficult to learn? Figure 5 presents the variation in MSE values between $f_{i,l} \cdot W_T$ and $f_{i,l}^{en}$ as the amount of parallel data, $|M|$, used to acquire the *Transform Matrix* increases. We observe that when $|M| = 1000$, the MSE value starts to converge between 0.01 and 0.02, and subsequently exhibits a stable trend. This indicates that only a thousand of parallel data are sufficient to effectively align $f_{i,l}$ with $f_{i,l}^{en}$ through the *Transform Matrix*, suggesting that the *Transform Matrix* is relatively easy to learn.

(2) Does incorporating cross-lingual connection substantially interfere with model training and model inference? During **training**, as shown in Figure 6, the loss curves of vanilla SFT and CC-TUNING are closely aligned, suggesting that the incorporation of cross-lingual connection on top of vanilla SFT introduces only negligible interference to the overall training process. This is primarily because no additional training objectives are introduced. In terms of training overhead, our statistics show that the training time for CC-TUNING is approximately ***1.12∼1.16 times*** that of vanilla SFT (Table 4). Moreover, the additional linear layer *Decision Maker* accounts for only ***0.0016%*** and ***0.0013%*** of the total parameter count in *LLaMA-3.1-8B* and *Qwen2.5-7B*, respectively—proportions so small that they are practically negligible. During **inference**, the time cost for inference with the *Transform Matrix* is also approximately ***1.1 times*** that of vanilla inference (Table 5).

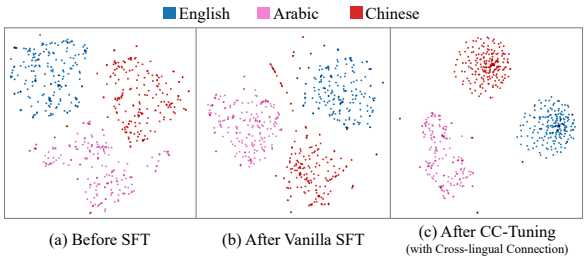


Figure 7: t-SNE visualizations of output representations by *LLaMA-3.1-8B* before fine-tuning, after vanilla supervised fine-tuning and after CC-TUNING.

5.2 Multilingual Representation Analysis

To analyze the impact of CC-TUNING on multilingual representations, we employ t-SNE (Van der Maaten and Hinton, 2008) to visualize the representations of 200 sentences sampled from *XNLI* in parallel across English, Arabic, and Chinese.

As depicted in Figure 7 (c), after applying CC-TUNING, the multilingual representations show a significantly more compact clustering. This indicates that CC-TUNING has already facilitated a certain level of cross-lingual interaction through the cross-lingual connection mechanism, allowing the multilingual representations after CC-TUNING require less extensive sharing with representations from other languages in high-dimensional space. And the boundaries between different language representations become more distinct, suggesting that CC-TUNING alleviates the mutual dependency between representations of different languages, enabling the model to exhibit clearer and more distinct multilingual modeling capabilities.

5.3 Beneficial Layer Distribution Analysis

In this section, we present the distribution of the layer with the highest probability of being selected by the *Decision Maker* across NLU and NLG tasks, as shown in Figure 8. This analysis explores layer-wise effectiveness within the cross-lingual connection. The distribution results indicate that LLMs tend to predominantly utilize the middle layers for both NLU and NLG tasks (*LLaMA-3.1-8B*: 19; *Qwen2.5-7B*: 17), which suggests that the middle layers may capture more valuable and generalized knowledge, potentially acting as a bridge between representations in different languages. Additionally, we observe that the beneficial layers identified in NLG tasks are more diverse, likely due to the inherent complexity of generation tasks. In contrast, NLU tasks—primarily focused on selecting from predefined options (e.g., A, B, C, or D)—are less

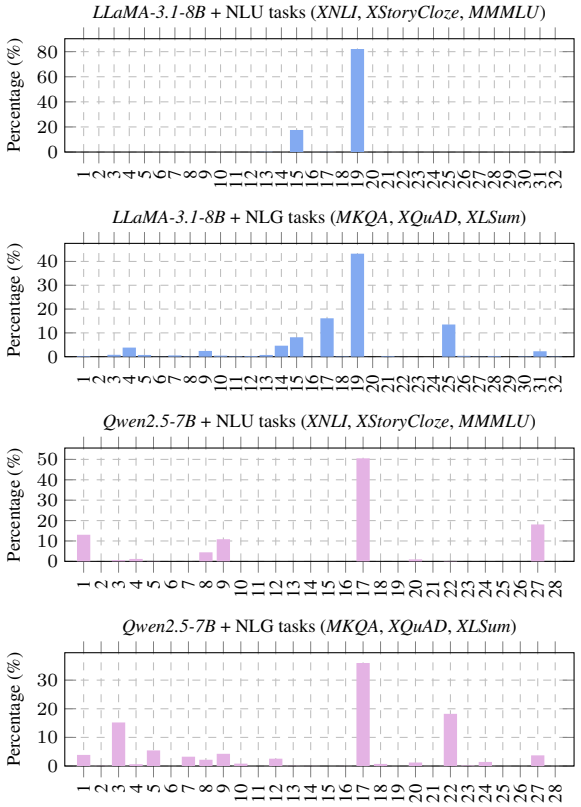


Figure 8: The distribution of the layer with the highest probability of being selected after the *Decision Maker* over NLU and NLG tasks.

complex, and thus, the layer distribution tend to be more concentrated.

5.4 Performance on Cross-lingual Task

We further conduct additional experiments evaluating the performance of CC-TUNING in cross-lingual scenarios on *XQuAD* under following settings: (1) "*X*-to-English": the question is given in language *X*, and the model is explicitly prompted to respond in English. (2) "English-to-*X*": the question is given in English, and the model is explicitly prompted to respond in language *X*.

The results in Table 3 show that CC-TUNING outperforms vanilla ML-SFT in both settings, highlighting its effectiveness in cross-lingual scenarios. The advantage is more pronounced in the *X*-to-English" setting, where the model is given non-English questions. This aligns with the motivation behind CC-TUNING: the models can benefit more when processing non-English inputs by leveraging the latent activations from English. Moreover, the performance gains on English questions under "English-to-*X*" setting are relatively smaller, which is also consistent with the observations under +EN setting in Table 1.

XQuAD (Ask in *X*, Answer in English)	en	ar	de	el	hi	ru	th	tr	vi	zh	AVG
ML-SFT	72.61	15.29	30.84	21.60	12.18	15.21	18.82	25.38	33.87	13.95	25.97
CC-TUNING	74.62	18.82	35.71	24.54	12.27	17.14	25.29	27.23	35.21	17.90	28.87
XQuAD (Ask in English, Answer in *X*)	en	ar	de	el	hi	ru	th	tr	vi	zh	AVG
ML-SFT	72.61	20.34	40.50	18.91	21.09	20.08	17.48	29.24	31.93	27.48	29.97
CC-TUNING	75.29	18.99	40.34	20.67	23.36	20.34	19.41	29.33	33.45	28.49	30.97

Table 3: Results on the cross-lingual QA task with *LLaMA-3.1-8B*. The symbol *X* refers to either the input prompt language or the required response language, as specified by the corresponding configuration.

5.5 Language Confusion Analysis

Language confusion refers to the cases where the model fails to consistently response in the user's desired language, or the appropriate language given the context. Here we employ the *lid.176.bin* model from *fastText*, which can identify 176 languages, to evaluate the alignment between model responses and input languages.

The results in Table 6 show that language confusion phenomenon frequently occured in baseline SDRRL. Since SDRRL is designed to facilitate knowledge distillation from resource-rich to low-resource languages, the training data under this setup often contains inconsistencies between input and output languages. While this issue is partially mitigated through code-switching and the integration of external parallel corpora, we observed that it still frequently causes deviations from the prompt language in model responses, making SDRRL less suitable for generation tasks. In contrast, CC-TUNING, along with other baselines, do not exhibit significant language confusions.

6 Conclusion

In this paper, we propose CC-TUNING, a novel multilingual fine-tuning paradigm that establishes a cross-lingual connection mechanism at latent level to address the imbalanced multilingual capabilities of current LLMs. During training, CC-TUNING fuses the feed forward activations from both English and non-English inputs, enabling the model to benefit from both languages. During inference, we simulate the cross-lingual connection using only monolingual input through representation transformation techniques. Extensive experiments across six benchmarks covering 22 languages demonstrate that CC-TUNING outperforms vanilla supervised fine-tuning and serves as a strong latent-level alternative to data-level augmentation approaches. Our results also highlight the importance of rethinking multilingual training paradigms beyond superficial data manipulation, suggesting that deeper architectural interventions may unlock greater potential in LLMs' multilingual capabilities.

Limitations

This work exhibits several limitations worth noting. Firstly, though several ablation experiments are conducted to validate the benefits of our training paradigm, we believe there is much more to explore and investigate in latent cross-lingual interactions. Such interactions should not only be limited to the form discussed in our work. Secondly, our experiments were conducted on LLaMA-3.1-8B and Qwen2.5-7B. While these models represent important milestones in open-source LLM development, the evaluation across more LLMs would improve the generalizability of our findings across the broader LLM ecosystem. Thirdly, due to the computational constraints, we did not conduct comparisons between LLMs of different model sizes (particularly larger models), resulting in a lack of insights into the impact of model capacity on performance.

Acknowledgements

Xiaocheng Feng is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62276078, U22B2059), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018). We also thank Huawei Technologies Co., Ltd for supporting part of the computing resources and funding.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation.

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024. Introducing bode: A fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*.

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*.

HIT-SCIR. 2024. Chinese-mixtral-8x7b: An open-source mixture-of-experts llm. https://github.com/HIT-SCIR/Chinese-Mixtral-8x7B.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, et al. 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.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

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, Veselin Stoyanov, and Xian Li. 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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15395–15406, Bangkok, Thailand. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

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.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an italian instruction-tuned llama. *Preprint*, arXiv:2307.16456.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Murray Shanahan. 2022. Talking about large language models. *ArXiv preprint*, abs/2212.03551.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

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

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

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

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024a. GlobeSumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10803–10821, Miami, Florida, USA. Association for Computational Linguistics.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Libo Qin, Yichong Huang, Lei Huang, Weitao Ma, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, et al. 2024b. Xtransplant: A probe into the upper bound performance of multilingual capability and culture adaptability in llms via mutual cross-lingual feed-forward transplantation. *arXiv preprint arXiv:2412.12686*.

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. *arXiv preprint arXiv:2402.12204*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 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)*, 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*.

A Experiment Details

A.1 Baselines Settings

This section introduces the details of different training data settings.

- **+EN** combines the original multilingual dataset D with its translated parallel English dataset D^{en}, resulting in a total training dataset size of $N + N = 2N$.
- **+MT** constructs additional translation task form data by pairing the original multilingual dataset D with its translated parallel English dataset D^{en} as follows:

  ```
  {
  "instruction": "Translate the following sentence from English to Spanish.\n The category corresponds to politics.",
  "output": "La categoría corresponde a política. "
  }
  ```

  N pairs of parallel data from D and D^{en} can be constructed into N additional samples of translation task form data, resulting in a total training dataset size of $N + N = 2N$.
- **+SDRRL** (Zhang et al., 2024) is a self-distillation-based method that integrates English instruction tuning data and its multilingual code-switched extensions. Additionally, it incorporates partially translated data and completion data for fine-tuning (LLaMA-3.1-8B: data size≈$1.2N$, Qwen2.5-7B: data size≈$1.6N$).

A.2 Datasets and Evaluations

A.2.1 Datasets

The language subsets used in the 6 evaluation datasets involved in our experiments and the data size used for each language subset are as follows:

Involved Languages (10 languages each dataset)

XNLI: en, ar, el, hi, ru, sw, th, tr, ur, zh

XStoryCloze: en, ar, es, eu, hi, id, ru, sw, te, zh

MMMLU: en, ar, bn, es, hi, id, ko, pt, sw, yo

XQuAD: en, ar, de, el, hi, ru, th, tr, vi, zh

MKQA: en, ar, de, ja, ko, pt, ru, tr, vi, zh

XLSum: en, ar, fr, hi, id, ru, sw, tr, ur, vi

A total of 22 unique languages are involved

Questions (ask in Chinese)

```
1.  "...\n问题：  黑豹队的防守丢了多少分？\n\n您的答案：
2.  "...\n问题：  贾里德在职业生涯中有多少次擒杀？\n\n您的答案："
3.  "...\n问题：  卢克·坎克利贡献了多少次擒抱？\n\n您的答案："
4.  "...\n问题：  约什·诺曼拦截了多少球？\n\n您的答案："
5.  "...\n问题：  本赛季谁为球队贡献的擒杀最多？\n\n您的答案："
6.  "...\n问题：  2015年黑豹队的防守有多少次拦截记录？\n\n您的答案："
7.  "...\n问题：  谁带领黑豹队擒杀？\n\n您的答案："
8.  "...\n问题：  有多少名黑豹队防守球员入选了职业碗？\n\n您的答案："
9.  "...\n问题：  托马斯·戴维斯有多少次迫使掉球？\n\n您的答案："
10. "...\n问题：  本赛季哪个球员拦截次数最多？\n\n您的答案："
11. ...
```

Answers

```
1.  "The answer is 308 points."
2.  "Jared Allen has 136 career sacks."
3.  "在他们身后，黑豹队的三名首发线卫中有两人入选了职业碗：托马斯·戴维斯和卢克·坎克"
4.  "四次"
5.  "Jared Allen"
6.  "The answer is 24."
7.  "Jared Allen"
8.  "11人"
9.  "Thomas Davis forced four fumbles."
10. "The answer is: Josh Norman"
11. ...
```

Figure 9: Examples of the deviations from the prompt language in model responses when applying +SDRRL.

Sample Size

XNLI: $1000 \times 10 = 10000$ (parallel)
XStoryCloze: $1511 \times 10 = 15110$ (parallel)
MMMLU: $1000 \times 10 = 10000$ (parallel)
MKQA: $1000 \times 10 = 10000$ (parallel)
XQuAD: $1190 \times 10 = 11900$ (parallel)
XLSum: $100 \times 10 = 1000$ (non-parallel)

A.2.2 Evaluations

XNLI, *XStoryCloze*, and *MMMLU* all belong to the multiple-choice category. For these datasets, a model's response is considered correct only if it contains the correct option and excludes all other options. For the short QA generative dataset *MKQA* and *XQuAD*, a model's answer is deemed correct if the gold answer appears in the model's response.

A.3 Model Responses with +SDRRL

The results of applying **+SDRRL** to NLG tasks are not reported in the main body, as it may lead to deviations from the prompt language in model responses. Since **+SDRRL** aims to achieve distillation from resource-rich languages to low-resource languages, many of the training data's input and output languages under this setup are inconsistent. Although this issue is partially mitigated through code-switching and the incorporation of external parallel corpora, we still observed that it easily leads to deviations from the prompt language in model responses, making it unsuitable for NLG tasks. As in the examples shown in Figure 9, only 3 of the 10 questions given are correctly answered in Chinese, while the rest are all answered in English.

Training Time Cost *(h:m:s)*	*LLaMA-3.1-8B*	*Qwen2.5-7B*
ML-SFT	01:36:43	01:33:10
CC-TUNING	01:51:58	01:45:15
Time Cost Ratio	1.16	1.13
ML-SFT+EN	03:08:25	03:03:02
CC-TUNING+EN	03:34:28	03:25:30
Time Cost Ratio	1.14	1.12
ML-SFT+MT	03:08:24	03:04:13
CC-TUNING+MT	03:34:19	03:25:59
Time Cost Ratio	1.14	1.12
ML-SFT+SDRRL	01:52:17	02:23:00
CC-TUNING+SDRRL	02:08:52	02:41:20
Time Cost Ratio	1.15	1.13

Table 4: Comparisons of training time cost.

Inference Time Cost *(s)*	*LLaMA-3.1-8B*	*Qwen2.5-7B*
vanilla inference	2012.26	1898.89
inference w/ *Transform Matrix*	2209.90	2064.50
Time Cost Ratio	1.10	1.09

Table 5: Comparisons of inference time cost on the Arabic subset of *XNLI* dataset.

Consistency	MKQA	XQuAD	XLSum
LLaMA-3.1-8B			
Base Model	0.880	0.988	0.879
ML-SFT	0.879	0.901	0.983
ML-SFT+EN	0.911	0.888	0.988
ML-SFT+MT	0.827	0.888	0.986
ML-SFT+SDRRL	0.360	0.447	0.352
CC-TUNING	0.972	0.967	0.993
Qwen2.5-7B			
Base Model	0.963	0.915	0.955
ML-SFT	0.996	0.999	0.999
ML-SFT+EN	0.994	0.998	0.998
ML-SFT+MT	0.995	0.998	0.981
ML-SFT+SDRRL	0.541	0.642	0.582
CC-TUNING	0.996	0.997	0.997

Table 6: Input and output language consistency results.

Models	Dataset: XNLI										
	en	ar	el	hi	ru	sw	th	tr	ur	zh	Avg
Vanilla Model (*LLaMA-3.1-8B*)	36.20	15.20	23.90	31.80	29.70	28.20	29.70	28.10	24.50	8.90	25.62
ML-SFT (*LLaMA-3.1-8B*)	12.90	35.50	35.80	31.10	34.50	31.20	31.60	37.70	33.10	35.40	31.88
+EN	46.60	38.70	24.40	32.20	32.70	31.90	32.00	37.60	38.90	35.20	35.02
+MT	47.20	35.10	22.10	35.80	40.90	31.00	32.90	39.20	36.90	37.90	35.90
+SDRRL	29.80	29.70	29.70	29.80	29.70	29.80	29.70	29.70	29.80	29.70	29.74
CC-TUNING (*LLaMA-3.1-8B*)	51.10	40.50	38.90	33.20	42.50	30.70	37.10	39.00	35.10	36.10	38.42
+EN	39.50	32.00	33.40	29.20	34.50	30.00	31.20	31.70	34.50	31.20	32.72
+MT	48.70	36.20	37.30	30.40	39.70	31.10	32.50	38.00	33.00	37.50	36.44
+SDRRL	29.70	29.60	29.70	31.00	29.50	29.00	29.70	29.70	30.70	29.80	29.84
Vanilla Model (*Qwen2.5-7B*)	90.20	43.60	40.40	41.80	58.70	11.60	46.90	42.50	33.60	49.00	45.83
ML-SFT (*Qwen2.5-7B*)	60.60	52.60	44.80	45.90	56.60	29.70	51.20	48.10	36.30	56.50	48.23
+EN	81.70	54.10	39.70	43.30	59.60	30.50	50.70	49.30	39.70	59.00	50.76
+MT	61.60	49.60	36.70	45.80	56.80	29.70	51.60	48.10	42.00	48.60	47.05
+SDRRL	81.60	56.60	34.00	51.00	60.20	33.10	55.20	54.10	38.90	58.90	52.36
CC-TUNING (*Qwen2.5-7B*)	78.90	51.80	41.70	44.00	60.10	30.70	52.30	50.60	40.50	59.40	51.00
+EN	72.00	54.30	43.10	47.00	56.50	28.60	51.70	48.40	35.90	57.30	49.48
+MT	64.40	51.40	36.70	43.70	57.70	29.70	52.50	49.40	37.90	57.90	48.13
+SDRRL	83.80	56.30	33.40	46.00	60.00	32.20	58.00	58.30	42.30	60.30	53.06

Models	Dataset: XStoryCloze										
	en	ar	es	eu	hi	id	ru	sw	te	zh	Avg
Vanilla Model (*LLaMA-3.1-8B*)	49.70	41.69	14.89	28.72	48.44	60.03	36.47	49.70	6.02	19.72	35.54
ML-SFT (*LLaMA-3.1-8B*)	88.62	65.32	21.91	64.86	70.81	83.39	43.61	62.54	63.40	87.82	65.23
+EN	77.04	40.44	65.45	59.36	77.10	79.62	49.44	60.49	55.46	86.90	65.13
+MT	91.73	77.70	85.04	60.82	75.78	80.48	62.14	57.84	20.91	86.57	69.90
+SDRRL	72.93	65.32	45.80	20.32	68.63	64.13	66.05	45.14	49.64	60.29	55.82
CC-TUNING (*LLaMA-3.1-8B*)	89.34	73.73	64.26	51.09	79.48	79.81	70.22	58.24	55.33	84.45	70.60
+EN	69.36	66.71	75.38	25.74	62.48	75.78	49.77	50.36	49.24	84.58	60.94
+MT	87.43	73.99	87.62	57.91	82.06	82.86	76.44	59.43	38.65	89.01	73.54
+SDRRL	86.96	71.67	70.42	34.61	80.61	77.17	77.63	50.56	66.18	76.04	69.19
Vanilla Model (*Qwen2.5-7B*)	85.97	85.84	91.40	18.07	78.76	69.89	91.33	17.94	55.92	75.84	67.09
ML-SFT (*Qwen2.5-7B*)	92.12	78.89	93.51	52.95	79.48	79.48	71.21	37.06	28.92	86.96	70.06
+EN	78.23	54.27	91.00	56.25	81.80	87.36	71.34	44.74	61.02	90.27	71.63
+MT	82.06	56.12	92.19	57.64	82.20	88.15	73.92	29.52	57.78	85.44	70.50
+SDRRL	93.85	88.42	94.51	62.61	82.00	89.15	93.05	52.88	62.01	88.22	80.67
CC-TUNING (*Qwen2.5-7B*)	91.59	81.60	91.00	54.86	77.96	80.68	78.82	35.94	37.59	84.25	71.43
+EN	39.38	37.46	90.40	55.26	79.55	86.70	85.24	45.00	42.55	85.31	64.69
+MT	65.32	66.64	91.66	55.06	81.67	86.43	75.12	52.75	53.47	85.77	71.39
+SDRRL	93.45	90.87	92.26	57.58	82.26	88.68	94.51	57.25	59.03	93.45	80.93

Models	Dataset: MMMLU										
	en	ar	bn	es	hi	id	ko	pt	sw	yo	Avg
Vanilla Model (*LLaMA-3.1-8B*)	45.40	28.20	17.70	11.30	25.10	26.40	25.00	11.70	16.20	0.60	20.76
ML-SFT (*LLaMA-3.1-8B*)	57.40	41.60	31.70	51.20	37.60	44.70	39.40	51.70	20.70	26.00	40.20
+EN	56.80	35.90	32.00	50.90	36.00	40.80	40.50	48.10	29.80	25.40	39.62
+MT	59.20	37.80	33.10	51.50	37.60	42.80	42.50	48.10	28.30	25.90	40.68
+SDRRL	53.70	32.20	23.00	27.80	35.30	30.50	27.20	36.70	12.80	1.40	28.06
CC-TUNING (*LLaMA-3.1-8B*)	57.50	41.30	33.40	51.70	37.60	43.30	41.70	46.80	27.00	27.10	40.74
+EN	56.50	38.10	30.80	49.90	37.00	40.70	39.00	47.30	26.80	21.20	38.73
+MT	55.70	36.80	30.70	49.60	36.10	39.90	38.70	48.40	26.70	26.10	38.87
+SDRRL	53.10	38.60	33.40	46.40	36.50	41.10	35.70	47.20	31.70	14.00	37.77
Vanilla Model (*Qwen2.5-7B*)	68.20	53.50	43.70	64.00	47.40	60.90	46.00	62.40	17.90	1.30	46.53
ML-SFT (*Qwen2.5-7B*)	69.80	53.30	42.00	65.60	41.10	59.50	55.70	62.60	28.60	22.30	50.05
+EN	65.90	53.20	37.90	64.60	40.30	56.80	55.70	62.10	28.20	23.30	48.80
+MT	60.50	51.40	40.00	63.90	39.50	58.30	49.00	63.00	28.80	20.50	47.49
+SDRRL	66.00	46.00	38.30	60.70	41.50	56.90	52.10	58.60	29.90	22.80	47.28
CC-TUNING (*Qwen2.5-7B*)	69.10	52.80	40.50	65.10	41.20	59.10	54.90	62.30	30.90	20.60	49.65
+EN	67.10	54.50	38.10	64.20	41.90	55.50	53.90	61.30	24.10	12.90	47.35
+MT	66.60	53.10	40.30	64.70	41.70	60.50	53.60	63.10	29.90	20.40	49.39
+SDRRL	66.30	50.60	39.30	60.30	41.60	56.30	52.30	57.50	28.10	26.40	47.87

Table 7: The detailed performance results of different language subsets on NLU tasks (XNLI, XStoryCloze, MMMLU) across all involved models and baselines.

Models	Dataset: MKQA										
	en	ar	de	ja	ko	pt	ru	tr	vi	zh	Avg
Vanilla Model (*LLaMA-3.1-8B*)	22.50	5.20	3.50	3.50	3.00	4.80	4.90	15.70	6.80	5.70	7.56
ML-SFT (*LLaMA-3.1-8B*)	33.60	4.90	23.30	9.10	6.00	20.70	10.00	15.60	14.30	8.90	14.64
+EN	26.00	6.30	18.80	9.70	5.70	19.40	10.00	15.60	13.00	8.30	13.28
+MT	29.20	5.80	19.50	9.10	5.70	17.10	10.70	15.50	14.40	8.60	13.56
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*LLaMA-3.1-8B*)	32.00	6.00	24.10	10.90	6.30	22.40	10.50	17.80	18.20	11.20	15.94
+EN	27.70	6.40	20.20	11.10	7.30	20.50	9.50	17.20	15.50	10.70	14.61
+MT	32.10	6.90	21.90	10.70	7.30	21.10	10.10	17.70	17.30	10.80	15.59
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
Vanilla Model (*Qwen2.5-7B*)	1.00	6.60	8.50	10.40	8.30	7.20	7.50	10.10	15.70	15.20	9.05
ML-SFT (*Qwen2.5-7B*)	30.30	6.60	19.10	11.20	8.90	19.70	9.40	12.10	15.80	14.20	14.73
+EN	27.80	6.90	14.80	10.80	7.40	17.50	8.10	10.10	14.70	12.40	13.05
+MT	27.10	6.90	16.10	9.60	7.90	19.40	8.60	11.00	14.70	14.10	13.54
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*Qwen2.5-7B*)	30.3	7.2	18.60	11.6	8.5	20.5	8.3	12.9	15.80	14.7	14.84
+EN	27.60	5.90	15.10	10.80	7.60	19.20	9.60	12.90	13.60	13.30	13.56
+MT	29.30	6.50	16.10	11.20	7.90	18.30	8.50	12.50	13.30	14.10	13.77
+SDRRL	–	–	–	–	–	–	–	–	–	–	–

Models	Dataset: XQuAD										
	en	ar	bn	es	hi	id	ko	pt	sw	yo	Avg
Vanilla Model (*LLaMA-3.1-8B*)	72.18	52.86	58.07	47.73	61.26	43.87	46.97	51.93	53.53	68.32	55.67
ML-SFT (*LLaMA-3.1-8B*)	72.61	56.13	64.62	52.18	60.00	47.73	58.49	53.70	64.87	73.87	60.42
+EN	63.28	53.45	62.02	51.09	57.73	46.39	58.40	50.84	61.18	69.66	57.40
+MT	71.76	53.78	60.84	50.84	58.99	49.33	54.03	47.73	65.21	71.51	58.40
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*LLaMA-3.1-8B*)	75.29	55.29	64.96	51.34	62.27	52.10	60.42	54.20	67.82	74.79	61.85
+EN	69.08	58.32	64.03	52.77	60.59	51.51	60.25	52.69	66.64	73.03	60.89
+MT	77.73	55.29	63.45	53.61	61.34	52.18	56.72	53.11	68.24	73.78	61.55
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
Vanilla Model (*Qwen2.5-7B*)	53.19	71.26	71.01	50.17	49.92	56.39	64.62	57.98	77.31	89.24	64.11
ML-SFT (*Qwen2.5-7B*)	79.92	66.97	70.08	40.00	46.39	53.95	64.96	56.05	72.77	85.04	63.61
+EN	74.29	64.54	69.41	36.13	47.39	54.37	64.03	59.41	73.19	80.59	62.34
+MT	79.33	65.13	69.33	41.01	50.67	52.61	67.56	58.24	72.69	83.70	64.03
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*Qwen2.5-7B*)	79.24	64.12	71.34	39.75	47.06	53.61	65.71	57.73	74.03	84.62	63.72
+EN	72.18	64.45	68.40	41.01	47.31	54.37	66.30	58.99	72.94	80.92	62.69
+MT	77.98	67.31	71.26	39.75	49.41	52.86	69.33	57.82	72.27	84.62	64.26
+SDRRL	–	–	–	–	–	–	–	–	–	–	–

Models	Dataset: XLSum										
	en	ar	fr	hi	id	ru	sw	tr	ur	vi	Avg
Vanilla Model (*LLaMA-3.1-8B*)	6.60	3.88	11.91	1.02	5.63	7.62	3.53	5.74	1.54	9.59	5.71
ML-SFT (*LLaMA-3.1-8B*)	24.36	9.67	18.66	1.94	13.72	14.47	8.05	11.07	6.64	14.14	12.27
+EN	22.46	10.62	19.66	2.97	13.72	14.02	6.76	7.14	5.77	17.27	12.04
+MT	25.74	11.06	19.50	3.97	14.78	14.74	7.56	9.58	7.16	14.78	12.89
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*LLaMA-3.1-8B*)	25.00	10.87	19.46	3.02	13.46	15.55	8.63	10.01	7.20	15.63	12.88
+EN	23.76	10.26	21.45	3.67	14.30	14.45	8.94	9.94	6.15	14.92	12.78
+MT	27.57	11.38	21.08	3.23	13.34	15.71	9.14	10.88	4.41	13.73	13.05
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
Vanilla Model (*Qwen2.5-7B*)	10.45	3.59	10.86	0.00	5.43	6.89	2.73	3.54	3.09	4.21	5.08
ML-SFT (*Qwen2.5-7B*)	24.13	12.20	22.10	0.33	14.89	16.10	5.95	8.04	5.47	14.74	12.40
+EN	23.75	11.70	20.14	0.33	14.97	15.61	6.90	8.51	5.78	14.36	12.20
+MT	26.72	12.32	21.47	0.67	14.00	15.29	5.66	8.73	5.12	14.78	12.48
+SDRRL	–	–	–	–	–	–	–	–	–	–	–
CC-TUNING (*Qwen2.5-7B*)	23.22	10.75	22.21	0.62	14.47	17.61	6.47	8.37	5.43	15.84	12.50
+EN	25.06	12.79	19.58	0.33	14.71	15.63	7.12	10.62	5.39	15.01	12.63
+MT	25.84	11.46	22.62	1.00	15.77	16.43	5.69	9.37	5.06	15.46	12.87
+SDRRL	–	–	–	–	–	–	–	–	–	–	–

Table 8: The detailed performance results of different language subsets on NLG tasks (MKQA, XQuAD, XLSum) across all involved models and baselines.