AlignX: Advancing Multilingual Large Language Models with Multilingual Representation Alignment

Mengyu Bu[1,3], **Shaolei Zhang**[1,3], **Zhongjun He**[4], **Hua Wu**[4], **Yang Feng**[1,2,3†]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
[4] Baidu Inc.
{bumengyu23z, zhangshaolei20z, fengyang}@ict.ac.cn

Abstract

Multilingual large language models (LLMs) possess impressive multilingual understanding and generation capabilities. However, their performance and cross-lingual alignment often lag for non-dominant languages. A common solution is to fine-tune LLMs on large-scale and more balanced multilingual corpora, but such approaches often lead to imprecise alignment and suboptimal knowledge transfer, struggling with limited improvements across languages. In this paper, we propose AlignX to bridge the multilingual performance gap, which is a two-stage representation-level framework for enhancing multilingual performance of pre-trained LLMs. In the first stage, we align multilingual representations with multilingual semantic alignment and language feature integration. In the second stage, we stimulate the multilingual capability of LLMs via multilingual instruction fine-tuning. Experimental results on several pre-trained LLMs demonstrate that our approach enhances LLMs' multilingual general and cross-lingual generation capability. Further analysis indicates that AlignX brings the multilingual representations closer and improves the cross-lingual alignment.[1]

1 Introduction

Multilingual large language models (LLMs), trained on extensive multilingual corpora, demonstrate impressive capabilities across a wide range of NLP tasks (Brown et al., 2020; Touvron et al., 2023a; Üstün et al., 2024). However, they still exhibit a strong language bias towards high-resource languages, predominantly English, resulting in inferior performance and cross-lingual alignment for other languages (Qi et al., 2023; Chen et al., 2023; Zhu et al., 2024c; Chua et al., 2024).

To alleviate this problem, current mainstream methods implicitly inject cross-lingual alignment information at data level, such as continual pre-training on large-scale multilingual corpora (Cui et al., 2023; Yang et al., 2023; Xu et al., 2024), multilingual general instruction fine-tuning (Li et al., 2023; Zhang et al., 2024c), and cross-lingual instruction fine-tuning on translation pairs (Zhang et al., 2023; Zhu et al., 2023). Such data-level methods adjust multilingual semantic representations by implicitly injecting alignment information through translation pairs or large-scale multilingual corpora. However, the impact of such methods on the semantic space is uncontrollable and inefficient, often resulting in imprecise alignment and suboptimal knowledge transfer.

A classic assumption in multilingual NLP is that more consistent multilingual representations facilitate easier knowledge transfer (Pan et al., 2021; Tang et al., 2022). Wendler et al. (2024) provide evidence by showing that English-centric LLMs internally pivot through English during processing. To investigate this further, we explore how LLMs process multilingual data across layers, and reveal an align-then-diverge pattern (Figure 2). From the lower to intermediate layers, the model aligns multilingual representations to enable knowledge sharing, while from the intermediate to upper layers, it gradually diverges these representations to produce language-specific outputs. Building on this assumption and finding, an intuitive approach to enhancing multilingual capabilities is to align multilingual semantic spaces at the intermediate layer for better knowledge sharing, while preserving language-specific features in higher layers for accurate generation.

To achieve this, we propose AlignX, a two-stage and representation-level framework for enhancing the multilingual performance of pre-trained LLMs. In the first phase, we efficiently align multilingual representations during continual pre-

†Corresponding author: Yang Feng.
[1]The code is available at https://github.com/ictnlp/AlignX.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6460–6489
November 4-9, 2025

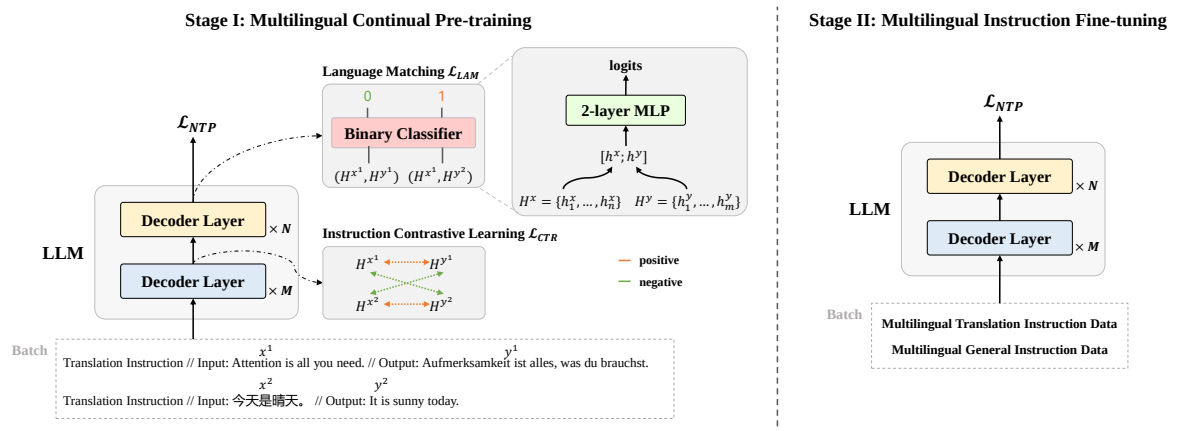


Figure 1: The framework of our AlignX, which consists of two stages. The first stage, multilingual continual pre-training, aligns multilingual representations within LLMs. The second stage, multilingual instruction fine-tuning, stimulates the multilingual capabilities of LLMs and maintains pre-established multilingual alignment.

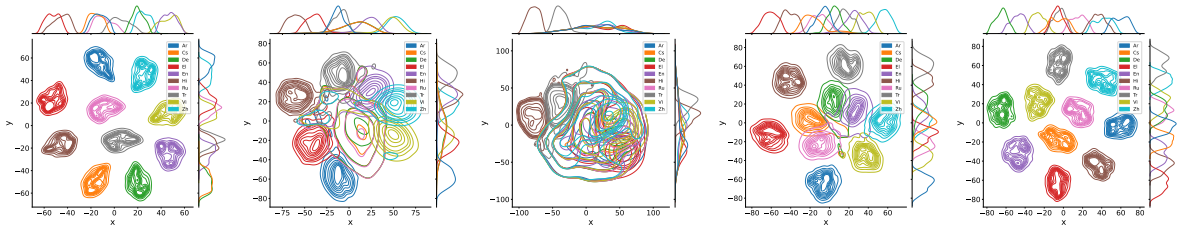
training. Specifically, we perform multilingual semantic alignment at the intermediate layer using the instruction contrastive learning task to promote knowledge sharing, and integrate language-specific features at the output layer via the language matching task to ensure accurate target language generation. In addition, we provide overall constraints through the standard language modeling task. In the second stage, we fine-tune the model with multilingual instruction data, comprising both translation and multilingual general instruction data, to stimulate the general capability while preserving the multilingual alignment information established in the first stage.

We experiment on five pre-trained LLMs and evaluate performance across five widely used multilingual general and generation benchmarks. The results indicate that AlignX effectively enhances multilingual general and cross-lingual generation capabilities. Extending AlignX to 51 languages further improves performance, highlighting the benefits of increasing the number of aligned languages in enhancing knowledge sharing.

2 Related Work

Data-level Cross-lingual Alignment in Multilingual LLMs Many works aim to enhance LLMs' multilingual capabilities through data-level approaches. Since translation pairs inherently contain language alignment information, researchers often use translation corpora to boost language proficiency (Zhang et al., 2023; Alves et al., 2024; Zhu et al., 2024b). Xu et al. (2024) first utilize massive monolingual data to enhance language generation and then leverage a small but high-quality translation dataset to inject alignment information. Zhu et al. (2023) leverage both multilingual translation instruction data and general task instruction data to build semantic alignment across languages. Zhang et al. (2024c) propose a self-distillation method that transfers high-resource language capabilities to enhance multilingual performance using translation and code-switched pairs. Compared to these approaches, our representation-level AlignX aligns cross-lingual representations more efficiently.

Representation-level Cross-lingual Alignment in Multilingual LLMs Representation engineering provides a powerful lens for analyzing internal representations of LLMs (Zhang et al., 2024a; Yu et al., 2024). Following this, many recent works investigate internal cross-lingual capability of LLMs (Zhong et al., 2024; Zhao et al., 2025). Wendler et al. (2024) suggest that English-centric LLMs use English as an internal pivot language. Given the existence of an internal pivot language, it is intuitive that aligning multilingual representations facilitates more efficient knowledge transfer. Li et al. (2024a) bridge the multilingual representation gap through multilingual contrastive learning and cross-lingual instruction tuning. Li et al. (2024b) establish word-level multilingual alignment before pre-training and then pre-train on code-switched text. AlignX differs by performing multilingual semantic alignment at the intermediate layer and language feature integration at the output layer. This integration is crucial for preserving language-specific features and enhancing accurate language generation.



(a) Layer 1 (b) Layer 6 (c) Layer 16 (d) Layer 20 (e) Layer 32

Figure 2: Preliminary visualization of multilingual representations from some representative LLaMA3 layers after dimension reduction. We leverage the FLORES-101 dev set, which is multi-way parallel, and use average-pooled hidden states for dimension reduction. Appendix A presents detailed visualization process and results.

3 Preliminary Analysis

We visualize the multilingual representations across different layers of LLaMA3 after dimension reduction. Specifically, we use the FLORES-101 dev set, perform mean-pooling on hidden states of different layers, and reduce dimension to 2-dim using t-SNE (Van der Maaten and Hinton, 2008). Figure 2 presents the results of some representative layers, revealing a clear pattern: from the lower to intermediate layers, the LLM progressively aligns multilingual representations, facilitating knowledge transfer. After the intermediate layers, multilingual representations gradually diverge into distinct language-specific subspaces, ultimately leading to different language outputs. This reveals a pattern in how LLMs handle multilingual data: the LLM exhibits a natural tendency to process multilingual representations in an align-then-diverge manner, though the alignment remains suboptimal. Following this, we propose AlignX to strengthen this pattern.

4 Method

To enhance this align-then-diverge pattern of LLMs, we propose AlignX, a two-stage and representation-level framework for enhancing the multilingual performance of pre-trained LLMs. AlignX contains two stages: 1) multilingual continual pre-training, which explicitly aligns multilingual representations via multilingual semantic alignment and language feature integration; and 2) multilingual instruction fine-tuning, which combines multilingual general instruction data with multilingual translation instruction data to stimulate general capabilities while preserving the pre-established alignment. Figure 1 illustrates the framework of AlignX.

4.1 Multilingual Continual Pre-training

In the first stage, we conduct multilingual continual pre-training and efficiently align multilingual representations through multilingual semantic alignment and language feature integration, guided by multilingual instruction contrastive learning $\mathcal{L}_{CTR}$ and language matching learning $\mathcal{L}_{LAM}$, respectively. This stage leverages English-centric translation instruction data.

Multilingual Semantic Alignment Our preliminary visualization reveals that LLMs tend to align multilingual representations during processing. To enhance this trend, we explicitly introduce a contrastive learning task at the intermediate layer, encouraging the model to produce similar representations for translation pairs.

Formally, given the dataset $\mathcal{D}$ and a translation instruction data $I^i = \{\mathbf{x}^i, \mathbf{y}^i\}$, we compute its l-th layer hidden states in model $f(\theta)$ as follows:

$$f_l(I^i) = \{f_l(I^i)_1, \ldots, f_l(I^i)_n\} \quad (1)$$

We extract the hidden states of $\mathbf{x}^i$ and $\mathbf{y}^i$ based on their respective token index ranges and get $I^i_{\mathbf{x}^i}$ and $I^i_{\mathbf{y}^i}$, and subsequently apply mean pooling to obtain the sentence representation $h^{\mathbf{x}^i}$ and $h^{\mathbf{y}^i}$:

$$h^{\mathbf{x}^i} = g(f_l(I^i_{\mathbf{x}^i})), h^{\mathbf{y}^i} = g(f_l(I^i_{\mathbf{y}^i})) \quad (2)$$

where $g(\cdot)$ denotes mean pooling operation.

We take $(\mathbf{x}^i, \mathbf{y}^i)$ as the positive example and randomly select $\mathbf{y}^j$ to form the negative example $(\mathbf{x}^i, \mathbf{y}^j)$. Then, the objective of multilingual instruction contrastive learning is:

$$\mathcal{L}_{CTR} = - \mathop{\mathbb{E}}_{\mathbf{x}^i, \mathbf{y}^i \in \mathcal{D}} \log \frac{e^{sim(h^{\mathbf{x}^i}, h^{\mathbf{y}^i})/\tau}}{\sum_{\mathbf{y}^j} e^{sim(h^{\mathbf{x}^i}, h^{\mathbf{y}^j})/\tau}} \quad (3)$$

where $sim(\cdot)$ calculates the similarity of different sentences, and we use cosine similarity in our

work. τ is a temperature hyper-parameter. To simplify the implementation, we sample negative examples within every training batch.

Language Feature Integration Intuitively, after aligning the multilingual semantic space in the intermediate layer, it becomes more difficult for the model to distinguish between output languages. To bridge this gap, we leverage a language matching task to inject language features. Specifically, we train a language matching classifier that predicts whether a given sentence pair is in the same language (matched) or different languages (unmatched). The classifier takes the hidden states from the model's final layer as input and is optimized using a binary classification task.

Formally, given mean-pooled sentence representations $h^{\mathbf{x}}$ and $h^{\mathbf{y}}$, we concatenate and feed them into the matching classifier to get a logit. The ground truth is whether $\mathbf{x}$ and $\mathbf{y}$ belong to the same language, denoted as y. The language matching task is optimized with the cross-entropy loss:

$$\mathcal{L}_{LAM} = - \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\in\mathcal{D}} \log M([h^{\mathbf{x}}; h^{\mathbf{y}}], y) \quad (4)$$

where $M(\theta)$ denotes the language matching classifier, a 2-layer MLP in our work. To simplify the implementation, we sample sentence pairs within every training batch.

Final Loss Finally, we optimize the representation alignment objectives $\mathcal{L}_{CTR}$ and $\mathcal{L}_{LAM}$ with the standard next token prediction loss of LLMs, using multilingual translation instruction data. The next token prediction loss $\mathcal{L}_{NTP}$ is:

$$\mathcal{L}_{NTP} = - \mathop{\mathbb{E}}_{I^i\in\mathcal{D}} \sum_j \log P(I^i_j | I^i_{<j}) \quad (5)$$

The final loss of multilingual continual pre-training stage is:

$$\mathcal{L} = \mathcal{L}_{NTP} + \alpha_1 \mathcal{L}_{CTR} + \alpha_2 \mathcal{L}_{LAM} \quad (6)$$

where α_1 and α_2 are hyper-parameters to balance the losses.

4.2 Multilingual Instruction Fine-tuning

After multilingual representation alignment, we leverage multilingual instruction fine-tuning to effectively stimulate general capabilities while maintaining the alignment established in the first stage. We mix multilingual translation instruction data, which implicitly provides alignment information (Zhu et al., 2023), and multilingual general instruction data, which efficiently stimulates multilingual general capabilities (Chen et al., 2024). To ensure diversity and multilingual balance, we sample English-centric translation pairs uniformly across languages, pairing each with a randomly selected translation instruction. General instruction data are also sampled uniformly across languages. Empirically, we maintain a 1:3 ratio of translation data to general instruction data and a 1:5 ratio of second-stage data to first-stage data. In this stage, we optimize the model with the next token prediction loss in Equation (5).

5 Experiment

5.1 Experiment Setup

Base LLMs We leverage LLaMA-7B (Touvron et al., 2023a), LLaMA2-7B (Touvron et al., 2023b), LLaMA3-8B-Instruct (Dubey et al., 2024), Gemma-2B (Team et al., 2024) and Mistral-7B-v0.3 (Jiang et al., 2023) as base LLMs.

Language Setup In the main experiment, we focus on ten languages: Arabic (Ar), Czech (Cs), German (De), Greek (El), English (En), Hindi (Hi), Russian (Ru), Turkish (Tr), Vietnamese (Vi) and Chinese (Zh), spanning diverse families and resource levels, and refer to this setup as **AlignX**. To validate scalability, we extend to 51 languages using LLaMA3-8B-Instruct, denoting this configuration as **AlignX (51langs)** to distinguish it from the original 10-language setup.

Baselines We compare with these baselines: (1) **CPT-then-SFT** follows the same two-stage multilingual instruction tuning as AlignX but without $\mathcal{L}_{CTR}$ and $\mathcal{L}_{LAM}$; (2) **ParroT-7B** (Jiao et al., 2023) leverages chat data to improve translation ability; (3) **BayLing1-7B** (Zhang et al., 2023) and **BayLing2-8B** (Zhang et al., 2024b) use multi-turn interactive translation instruction data for cross-lingual alignment; (4) **BigTrans-13B** (Yang et al., 2023) applies continual pre-training on a large-scale multilingual corpus; (5) **Tower-7B** (Alves et al., 2024) first trains on multilingual data then fine-tunes on translation instruction data.

Training Dataset We construct the multilingual translation instruction data from OPUS-100 (Zhang et al., 2020) and extract multilingual general instruction data from Bactrian-X (Li et al.,

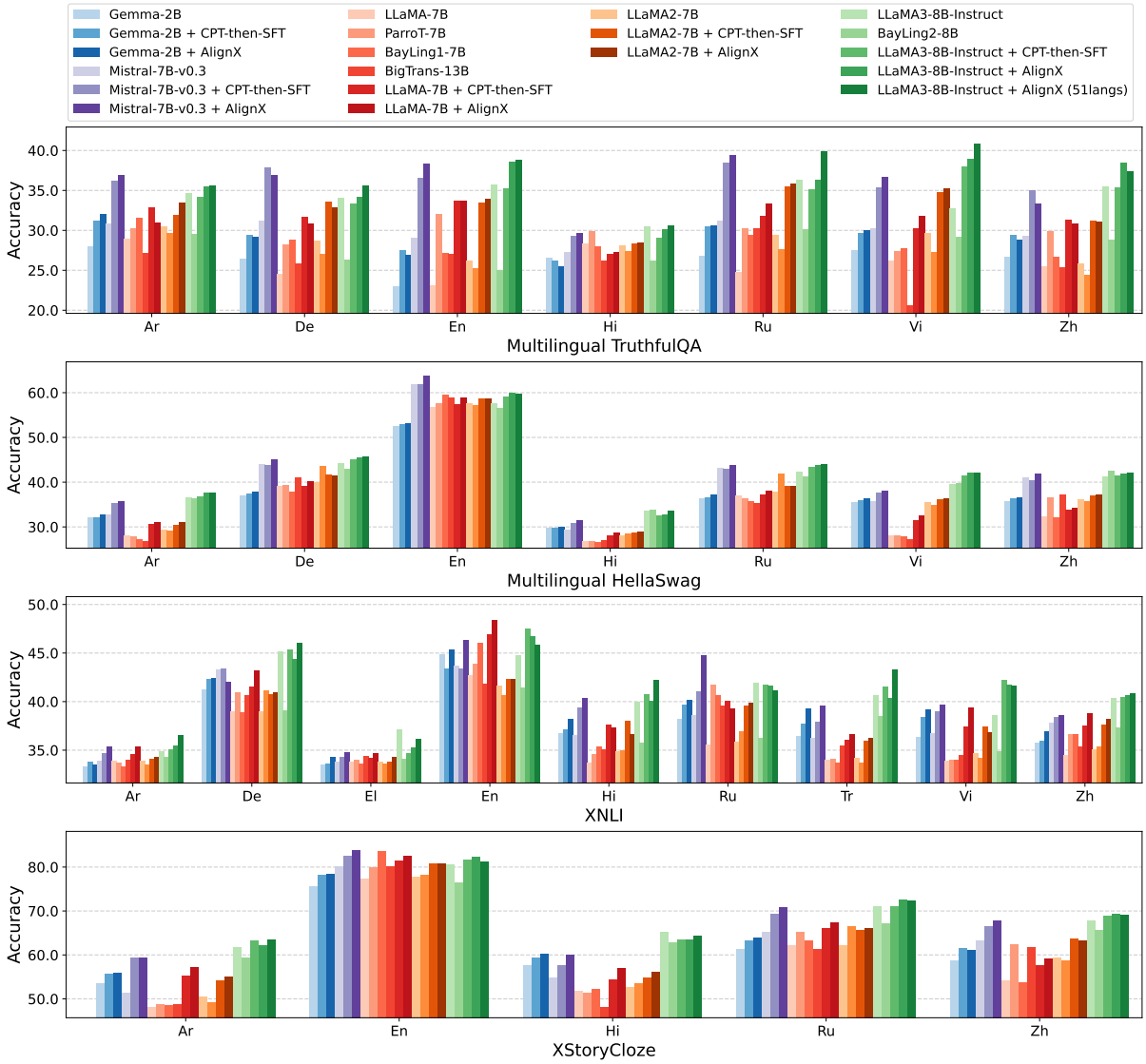


Figure 3: Overall performance on multilingual general benchmarks. All the benchmarks are evaluated using Accuracy metric. For comparison purposes, models with the same base LLM have the same color scheme, with the lightest being the base LLM and the darkest being AlignX.

2023). Appendix B.1 and B.2 present details about data processing and data statistics, respectively.

Evaluation Benchmarks We evaluate multilingual general and cross-lingual generation capabilities on five benchmarks: multilingual HellaSwag (Zellers et al., 2019), multilingual TruthfulQA (Lin et al., 2022), XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2021b), evaluated with Accuracy metric; and FLORES-101 (Costa-jussà et al., 2022), evaluated with BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) metrics. See Appendix B.3 for details.

Configuration In AlignX, the language matching classifier is a 2-layer MLP with an intermediate dimension of 128 and an output dimension of 2. For training, we optimize using AdamW optimizer with a learning rate of 2e-6, training for 2 epochs per stage with a batch size of 128. We empirically set $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, and $\tau = 0.1$. For evaluation, we use the *MMT-LLM* framework (Zhu et al., 2024c) for FLORES-101 and the *lm-evaluation-harness* framework (Gao et al., 2024b) for other general benchmarks. All tasks are evaluated in a 1-shot setup.

5.2 Main Results

AlignX improves multilingual general capability. Figure 3 shows the results on multilingual TruthfulQA, multilingual Hellaswag, XNLI, and XStoryCloze, with detailed results in Appendix C.1. While these benchmarks are out of our train-

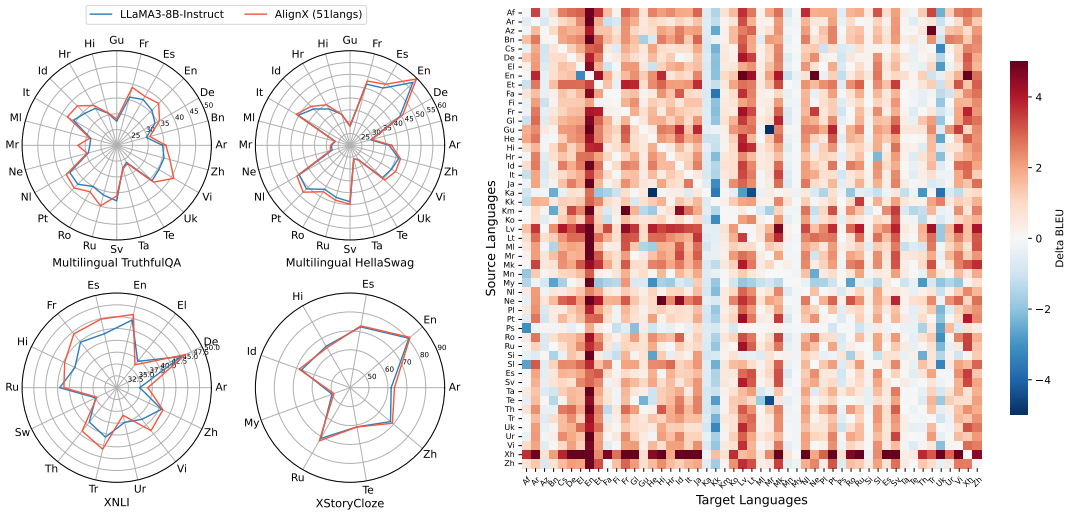
Methods	X-En	X-Ar	X-Cs	X-De	X-El	X-Hi	X-Ru	X-Tr	X-Vi	X-Zh	Avg.
Gemma-2B Based											
Gemma-2B	30.46	3.77	6.39	8.53	4.67	3.41	4.84	4.11	8.22	7.19	8.16
CPT-then-SFT	31.62	**7.36**	7.94	13.40	**8.99**	7.16	**11.05**	7.24	14.01	10.55	11.93
AlignX	**31.68**	6.94	**8.75**	**13.84**	8.83	**7.40**	10.63	**7.94**	**15.87**	**11.02**	**12.29**
Mistral-7B-v0.3 Based											
Mistral-7B-v0.3	30.96	3.89	14.29	**17.88**	2.26	3.49	15.68	5.78	9.71	10.50	11.44
CPT-then-SFT	**34.29**	**8.13**	**15.05**	17.35	**6.84**	6.17	**16.88**	9.94	13.86	12.56	**14.11**
AlignX	33.85	8.00	14.58	16.54	6.77	**6.74**	16.59	**10.34**	**14.18**	**12.78**	14.04
LLaMA-7B Based											
LLaMA-7B	20.81	0.74	5.70	9.05	0.80	0.97	6.60	1.22	1.10	1.60	4.86
ParroT-7B	17.91	0.22	3.28	4.65	0.45	0.26	1.86	1.59	1.67	1.81	3.37
BayLing1-7B	21.64	0.59	5.09	5.67	0.78	0.54	3.83	1.78	1.87	3.62	4.54
BigTrans-13B	21.30	2.10	5.27	5.31	1.38	**4.03**	4.66	3.80	2.54	4.95	5.53
CPT-then-SFT	**27.02**	3.21	**9.38**	**11.90**	3.34	2.66	**10.48**	4.79	6.26	**5.64**	8.47
AlignX	25.46	**3.85**	9.01	11.39	**4.51**	3.36	9.96	**5.60**	**7.84**	5.30	**8.63**
LLaMA2-7B Based											
LLaMA2-7B	25.74	1.37	8.99	12.28	1.27	1.62	8.68	2.41	9.69	5.48	7.76
Tower-7B	30.48	1.51	7.74	12.75	1.33	1.79	10.15	2.26	8.56	9.26	8.58
CPT-then-SFT	31.61	3.46	**13.50**	**17.73**	2.74	2.71	14.39	5.37	13.53	10.52	11.55
AlignX	**31.99**	**5.06**	13.44	17.62	**4.22**	**4.17**	**15.19**	**6.30**	**15.55**	**11.64**	**12.52**
LLaMA3-8B-Instruct Based											
LLaMA3-8B-Instruct	34.52	14.16	19.40	22.99	16.45	16.42	21.18	16.37	24.65	17.04	20.32
BayLing2-8B	35.21	13.83	15.09	18.33	14.26	16.02	18.14	14.09	22.41	15.94	18.33
CPT-then-SFT	36.50	14.82	19.68	23.98	**17.82**	16.09	21.52	15.26	24.65	19.08	20.94
AlignX	37.36	14.98	19.66	23.58	17.73	16.26	21.31	14.72	23.51	18.40	20.75
AlignX (51langs)	**38.70**	**16.12**	**20.71**	**24.58**	17.63	**17.77**	**22.00**	**16.90**	**25.88**	**19.08**	**21.94**

Table 1: The overall performance on the FLORES-101 benchmark. In each model group, we fine-tune the base LLM (first row) to obtain CPT-then-SFT and AlignX. We report averaged BLEU scores for each target language. "X" denotes all other nine languages except the target language. "Avg." denotes average scores on all translation directions. We bold the highest scores.

ing distribution, AlignX achieves improvements on all five base LLMs, demonstrating the generalizability of AlignX. In contrast, data-level methods exhibit varying degrees of forgetting across languages, indicating that cross-lingual alignment achieved solely from continual pre-training on large parallel datasets struggles to generalize beyond translation tasks. These findings demonstrate the importance of multilingual representation alignment for enhancing multilingual capability and generalizability.

AlignX significantly enhances cross-lingual generation capability. Table 1 presents results on the FLORES-101 benchmark, with statistical significance test, COMET scores, and additional details in Appendix C and a case study in Appendix D. The results show that AlignX achieves the highest scores on all base LLMs, with average +4.13, +2.6, +3.77, +4.76, and +0.43 BLEU scores on Gemma-2B, Mistral-7B-v0.3, LLaMA-7B, LLaMA2-7B, and LLaMA3-8B-Instruct, respectively. This demonstrates AlignX's versatility, as it improves across LLMs with varying multilingual capabilities. Notably, BigTrans-13B uses a larger base LLM, 300M translation pairs, and some non-English translation directions (e.g., Hi-Zh). In contrast, AlignX relies solely on less than 1M English-centric translation pairs yet still achieves higher translation scores, indicating that representation alignment effectively enhances LLMs' cross-lingual capability.

Scaling AlignX to 51 languages further improves multilingual general and cross-lingual generation capabilities. To further validate the effectiveness of AlignX across a broader range of languages, we scale AlignX to 51 languages, listed in Appendix B.4. Figure 4 presents the results on multilingual general and translation benchmarks, with detailed results in Appendix C. These results indicate that AlignX performs effectively under the 51-language setup, enhancing both multilingual general and cross-lingual generation capabilities. To illustrate the impact of language scaling, we compare AlignX (51-langs) and AlignX (10-langs) under the 10-language setup, as shown in Figure 3 and Table 1. Surprisingly, increasing the number of languages from 10 to 51 leads to further gains in multilingual general and generation tasks, without exhibiting the "curse of multilinguality" (Zhu et al., 2024a). This highlights the



(a) Multilingual General Benchmarks (b) FLORES-101 Benchmark

Figure 4: Performance comparison of LLaMA3-8B-Instruct (blue) and AlignX (red) under the 51-language setup. (a): Accuracy metric results on multilingual general benchmarks. (b): BLEU differences on FLORES-101 benchmark, with red showing better AlignX performance and blue showing better LLaMA3-8B-Instruct performance.

Methods	TruthfulQA	HellaSwag	XNLI	IWSLT2017			
				X-En	En-X	Non-En	OTR
LLaMA-7B	24.77	41.79	35.70	31.20	16.08	10.90	30.92
w/o $\mathcal{L}_{CTR}$	31.34	44.03	36.07	36.13	28.68	20.01	4.89
w/o $\mathcal{L}_{LAM}$	**32.22**	43.99	**36.27**	**36.81**	29.39	18.59	12.59
AlignX	32.12	**44.11**	36.00	36.25	**29.88**	**21.21**	**4.16**

Table 2: Averaged results of variant models on LLaMA for multilingual general and IWSLT2017 benchmarks. "w/o $\mathcal{L}_{CTR}$" and "w/o $\mathcal{L}_{LAM}$" remove the instruction contrastive learning and language matching in the first stage, respectively. "Non-En" indicates translation between four non-English languages. "OTR" (off-target ratio) represents the proportion of outputs in incorrect target languages, thus the lower the better. We bold the best results.

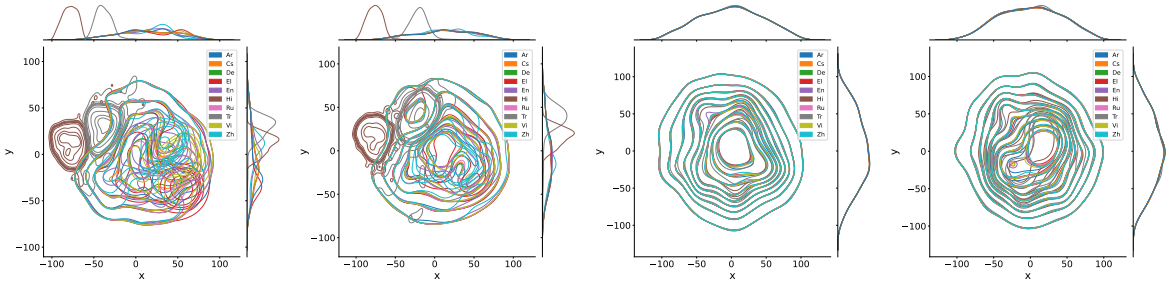
efficacy of multilingual representation alignment across a large number of languages.

6 Analysis

In this section, we provide an in-depth analysis of AlignX, examining the following aspects: the role of two auxiliary objectives, improvements in multilingual representation alignment, enhancements in cross-lingual generation and knowledge transfer, the impact of corpus size, and gains in cross-lingual alignment. Appendix E provides detailed comparisons with representation-level AFP, along with efficiency analysis. Additional evaluation benchmarks are reported in Appendix B.3.

Instruction contrastive learning facilitates knowledge sharing, and language matching promotes more accurate cross-lingual generation. We conduct experiments on several variant models to investigate the effectiveness of training objectives. We experiment with German, English, Italian, Dutch, and Romanian. These five languages belong to the Indo-European language family, which are relatively similar. We extract the training and test sets from the IWSLT2017 dataset, follow the same data construction process as the main experiment, and experiment on LLaMA-7B. Table 2 presents the results, highlighting the effectiveness of multilingual representation alignment. Specifically, instruction contrastive learning facilitates knowledge sharing, primarily benefiting relatively low-resource languages, but increases incorrect linguistic output in cross-lingual generation. To mitigate this off-target issue, language matching serves as a regularizer for output languages, enhancing LLMs' ability to distinguish output languages.

AlignX effectively brings the multilingual representations closer. To intuitively understand the effectiveness of multilingual representation alignment, we visualize multilingual



(a) Base Model (b) xSFT (c) CPT (d) AlignX

Figure 5: Visualization of multilingual representations from the base model, xSFT model, CPT model, and AlignX model after dimension reduction. The CPT and xSFT models train only the first and second stages of multilingual instruction fine-tuning, respectively. The visualization process follows the same steps as in Figure 2.

W/T/L	Grammar	Language Fit
X-El	173 / 174 / 103	192 / 138 / 120
All	1098 / 2901 / 501	1223 / 2611 / 666

Table 3: Automatic evaluation results using GPT-4o on cross-lingual generation. We compare paired outputs of AlignX and LLaMA3-8B-Instruct, where "W/T/L" indicates that AlignX produces better, comparable, or worse translations, respectively. "X" denotes all other nine languages except the target language.

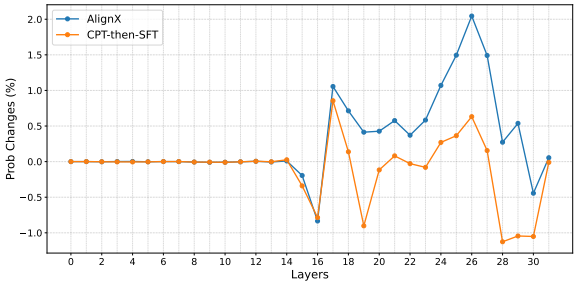


Figure 6: Layer-wise probability changes of unembedding the correct English answer token at non-final layers during non-English translation among German, Russian, and Chinese (e.g., De-Ru). Probabilities are averaged across samples.

representations of LLaMA3, data-level xSFT, and representation-level CPT, AlignX. Figure 5 presents visualization results, demonstrating the implicit multilingual alignment of xSFT and explicit multilingual alignment of AlignX. Comparing Figures 5(a) and 5(b), xSFT preliminarily aligns multilingual representations without explicit auxiliary tasks, although some distant languages still exist. This alignment capability comes from translation pairs, which naturally contain alignment information. The comparison in Figures 5(b) and 5(d) shows that multilingual representation alignment in the first stage is important in bringing multilingual representations closer. Appendix A presents detailed results for AlignX.

AlignX achieves better grammatical correctness and more accurate target language according to GPT-4o evaluation. To further examine how AlignX enhances cross-lingual generation, particularly in low-resource scenarios, we employ GPT-4o as an automatic evaluator focusing on the *grammar* and *language fit* dimensions. Evaluations are conducted under the 10-language setup. For each translation direction, we randomly sample 50 instances and compare the outputs from LLaMA3-8B-Instruct and AlignX. As shown in Table 3, detailed in Appendix E.3, AlignX demonstrates stronger grammatical correctness and more appropriate target language across languages, especially low-resource languages (e.g., Greek).

AlignX facilitates knowledge transfer via cross-lingual concept alignment in intermediate layers. Following Wendler et al. (2024), we analyze how AlignX enhances knowledge transfer and influences cross-lingual representations. Specifically, we measure the likelihood that non-English inputs activate English-centric representations in intermediate layers through German, Russian, and Chinese translation tasks (e.g., German–Chinese). For each direction, we compute the probability of unembedding the correct English token at non-final layers, averaged across samples, and report the change relative to LLaMA2 (e.g., AlignX – LLaMA2). Figure 6 reveals that in the concept space (Wendler et al., 2024), approximately layers 16 to 28, associated with language-agnostic processing, AlignX exhibits higher probabilities of retrieving the English token, whereas CPT-then-SFT fluctuates around zero. This indicates that AlignX

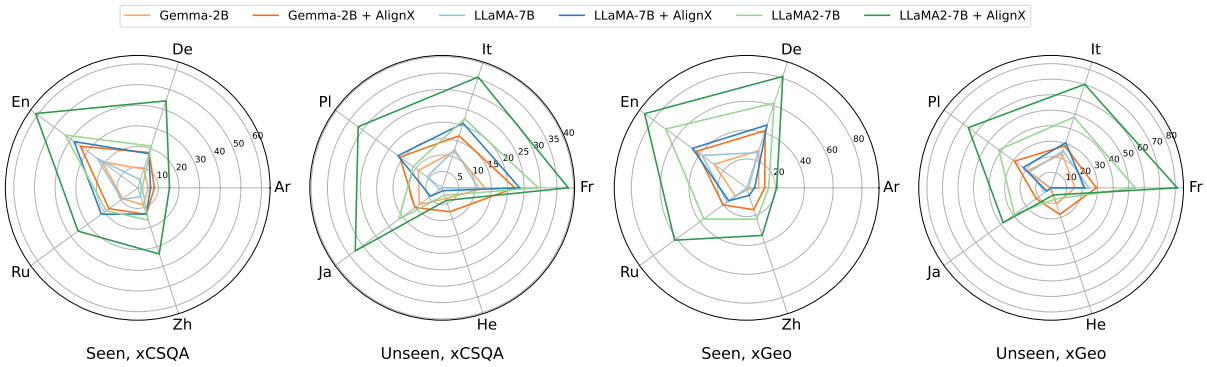


Figure 7: Results of the general cross-lingual knowledge alignment evaluation with CLiKA (Gao et al., 2024a). We present the re-scaled accuracy (RA) scores, where higher values (greater than 0) indicate better performance. "Seen" and "Unseen" denote whether these languages are included during AlignX training. xCSQA and xGeo evaluate *Basic* knowledge and *Factual* knowledge, respectively.

Methods	WMT14		WMT17	
	De-En	En-De	Zh-En	En-Zh
LLaMA-7B Based				
LLaMA-7B	28.95	18.30	15.75	10.58
xSFT	28.83	19.15	16.16	18.69
AlignX	31.45	22.25	18.37	24.86
AlignX†	**32.16**	**24.79**	**19.43**	**29.15**
LLaMA2-7B Based				
LLaMA2-7B	31.25	20.32	21.41	20.97
xSFT	30.81	20.42	18.88	24.51
AlignX	32.04	23.59	19.76	29.13
AlignX†	**33.94**	**26.70**	**22.00**	**32.18**
LLaMA3-8B Based				
LLaMA3-8B	33.66	23.70	25.03	31.60
xSFT	34.77	26.45	**25.06**	35.77
AlignX	**36.09**	27.23	24.84	36.26
AlignX†	35.70	**28.33**	24.32	**36.46**

Table 4: The results for the WMT14EnDe and WMT17EnZh benchmarks. The xSFT model only trains the second stage of multilingual instruction fine-tuning. AlignX† is a variant that increases the training data in the first stage from 50K to 250K examples per language direction. We bold the highest scores.

better aligns cross-lingual representations with the English-centric concept space, leveraging English as an interlingua to improve cross-lingual transfer.

Scaling up the dataset of multilingual representation alignment consistently enhances non-English language generation. We investigate the impact of scaling the first-stage training dataset for multilingual representation alignment on language generation, using German (De), English (En), and Chinese (Zh) with LLaMA-7B. Training and test sets are drawn from WMT14EnDe and WMT17EnZh. In Stage 1, we vary the corpus size for each language pair from 50K to 250K, while Stage 2 training follows the same settings as in the main experiment. As presented in Table 4, AlignX mainly enhances the performance in the non-English translation directions, consistent with Table 2. Moreover, extending corpora size in the first stage further enhances performance, especially in non-English generation.

AlignX achieves better cross-lingual alignment. To validate AlignX's impact on cross-lingual alignment, we use the CLiKA framework (Gao et al., 2024a), evaluating *Basic* and *Factual* knowledge with the xCSQA and xGeo benchmarks. We compute re-scaled accuracy scores, which exclude the interference of random baseline and question difficulty for better cross-lingual comparison. Results in Figure 7 show that AlignX improves cross-lingual alignment across multiple languages, reducing multilingual performance gaps. Remarkably, AlignX also generalizes well to unseen languages, with cross-lingual alignment correlating strongly with similar languages from training, suggesting generalization is influenced by language family similarities.

7 Conclusion

In this paper, we propose AlignX, an efficient two-stage representation-level framework for enhancing multilingual performance of multilingual LLMs. Results on several pre-trained LLMs and multiple widely used benchmarks show that AlignX effectively enhances multilingual general capabilities and cross-lingual generation capabilities. The analysis further demonstrates the impact of AlignX on LLMs, including bringing multilingual representations closer and improving the cross-lingual alignment of LLMs.

Limitations

Given LLMs with uneven multilingual capabilities, AlignX enhances both the multilingual general capability and cross-lingual generation capability, and alleviates the imbalance of multilingual capabilities, which matches our motivation. However, the multilingual performance of the final model is the outcome of the combined influence of the base model and AlignX, meaning that the imbalance remains unavoidable. This suggests that we are still far from achieving fully balanced multilingual capabilities.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments on this paper. This work was supported by the grant from the National Natural Science Foundation of China (No. 62376260).

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

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.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2024. Crosslingual capabilities and knowledge barriers in multilingual large language models. *arXiv preprint arXiv:2406.16135*.

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.

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

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

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*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024a. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. 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 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024b. A framework for few-shot language model evaluation.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

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.

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, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2024b. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. *arXiv preprint arXiv:2407.16222*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021b. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Henry Tang, Ameet Deshpande, and Karthik Narasimhan. 2022. Align-mlm: Word embedding alignment is crucial for multilingual pre-training. *arXiv preprint arXiv:2211.08547*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

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

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10862–10884.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024a. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, Shoutao Guo, Yan Zhou, Xiaodong Liu, and Yang Feng. 2024b. Bayling 2: A multilingual large language model with efficient language alignment. *arXiv preprint arXiv:2411.16300*.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024c. 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. 2025. Lens: Rethinking multilingual enhancement for large language models.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, Deyi Xiong, et al. 2024a. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8411–8423, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024c. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. 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 Detailed Visualization across Layers on LLaMA3

We use the FLORES-101 dev set, with 997 sentences per language. We directly input the sentence into the LLM to obtain the hidden state sequence, perform average pooling to derive the sentence representation, and apply t-SNE (Van der Maaten and Hinton, 2008) for dimension reduction to 2-dim. Detailed visualization results for both the base LLM and the AlignX LLM, based on LLaMA3-8B, are presented in Figure 8 and 9, respectively.

B Detailed Information for Training Dataset

B.1 Data Processing

For multilingual translation instruction data, we randomly select English-centric translation pairs from the OPUS-100 (Zhang et al., 2020) corpus and sample a translation instruction from the translation instruction set for each pair. For instruction diversity, we prompt ChatGPT to generate a set of ten translation instructions, as shown in Table 5. For multilingual general instruction data, we primarily draw from the Bactrian-X (Li et al., 2023) dataset, a multilingual version of the Alpaca (Taori et al., 2023) dataset available in 52 languages[2]. The instructions are translated using Google Translate, and the responses are generated with GPT-3.5-Turbo. We filter out off-target responses using the *langid* toolkit (Lui and Baldwin, 2012). In the first stage, we sample 50k translation pairs per translation direction, resulting in 0.9M translation instruction data in total. In the second stage, we sample 2.5k translation instruction data per translation direction and 10k general instruction data per language, resulting in 145k mixed instruction data in total.

B.2 Statistics on Corpora Size and Languages

Table 6 presents statistics on corpora size for our approach and some typical data-level systems. Table 7 presents information on languages involved in this work.

[2]Since Greek (El) is not included in Bactrian-X, we obtain the Greek Alpaca dataset from https://github.com/NJUNLP/x-LLM.

B.3 Details about Evaluation Benchmarks

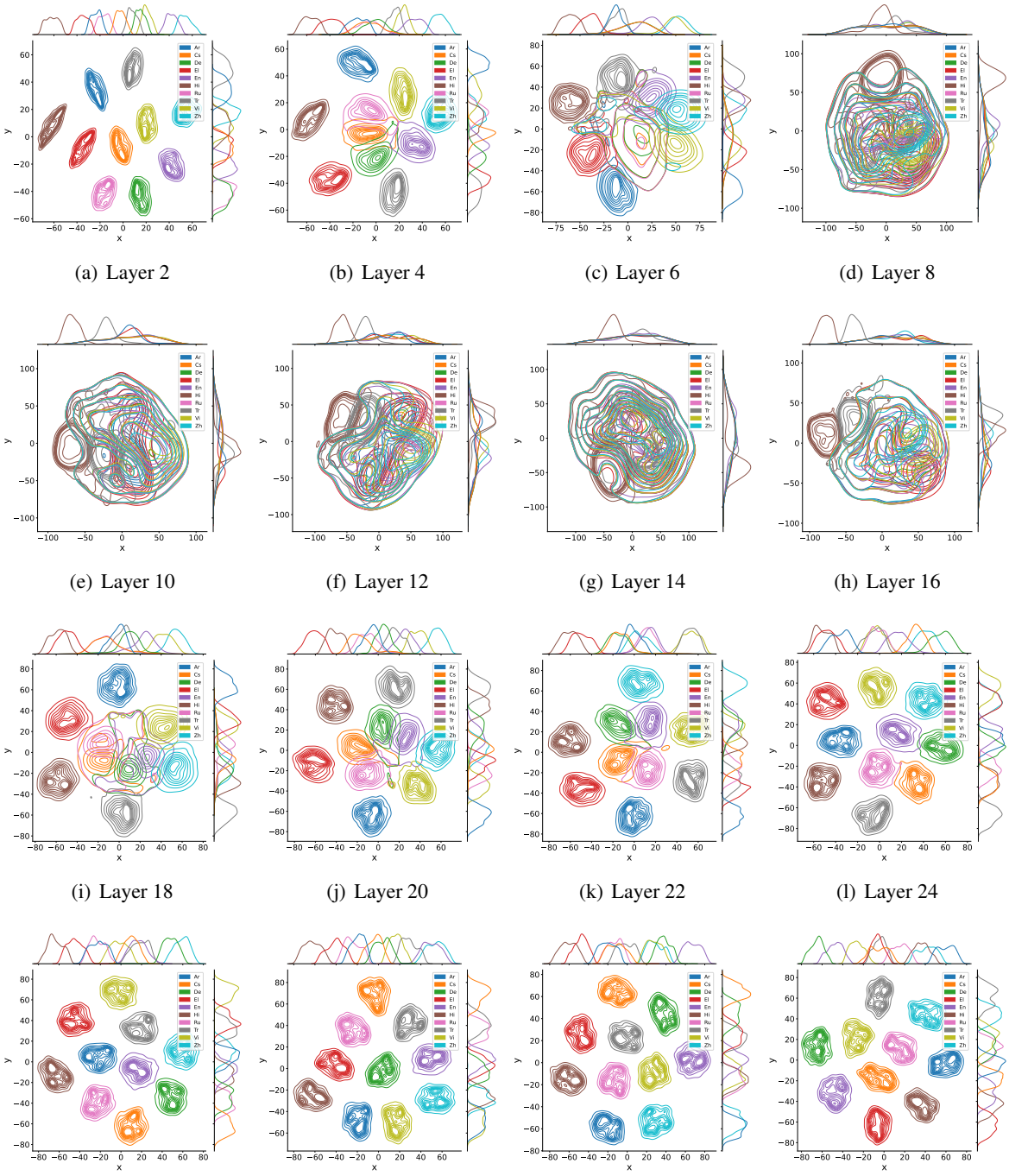
We evaluate multilingual general capabilities and cross-lingual generation capabilities on the following benchmarks:

- **Multilingual TruthfulQA** (Lin et al., 2022): This is the multilingual version of the TruthfulQA benchmark and evaluates knowledge and truthfulness capabilities.
- **Multilingual HellaSwag** (Zellers et al., 2019): This is the multilingual version of the HellaSwag benchmark and evaluates commonsense reasoning and contextual understanding capabilities.
- **Cross-lingual Natural Language Inference (XNLI)** (Conneau et al., 2018): This benchmark evaluates language transfer and cross-lingual sentence classification.
- **XStoryCloze** (Lin et al., 2021b): This is a multilingual commonsense reasoning benchmark for evaluating story understanding, story generation, and script learning.
- **FLORES-101** (Costa-jussà et al., 2022): This is a multilingual machine translation benchmark and evaluates the cross-lingual generation capability.

The multilingual HellaSwag and TruthfulQA benchmarks are obtained from Okapi (Lai et al., 2023), translated by ChatGPT. We evaluate the trained languages. In the analysis section, we further experiment on the following benchmarks:

- **IWSLT2017** (Cettolo et al., 2017), **WMT14**[3] and **WMT17**[4]: These benchmarks are classical multilingual translation benchmarks and evaluate cross-lingual generation capabilities of different language subsets.
- **xCSQA** (Lin et al., 2021a): This is the multilingual version of CSQA and evaluate cross-lingual commonsense reasoning.
- **xGeo** (Gao et al., 2024a): This benchmark evaluates cross-lingual knowledge alignment.

[3]https://huggingface.co/datasets/wmt/wmt14
[4]https://huggingface.co/datasets/wmt/wmt17



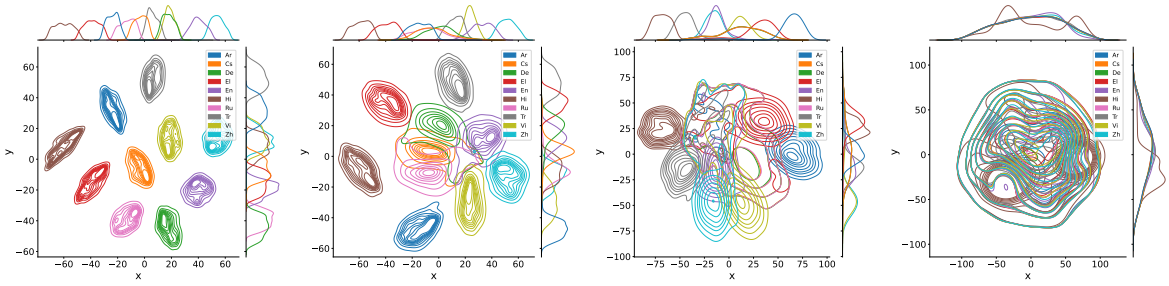
(a) Layer 2 (b) Layer 4 (c) Layer 6 (d) Layer 8

(e) Layer 10 (f) Layer 12 (g) Layer 14 (h) Layer 16

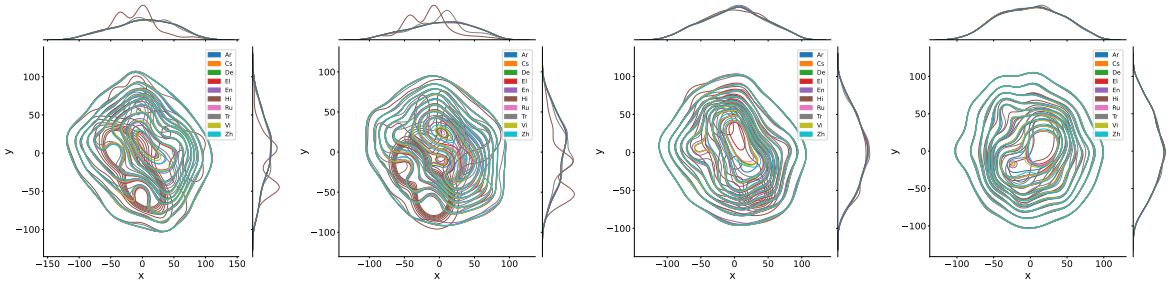
(i) Layer 18 (j) Layer 20 (k) Layer 22 (l) Layer 24

(m) Layer 26 (n) Layer 28 (o) Layer 30 (p) Layer 32

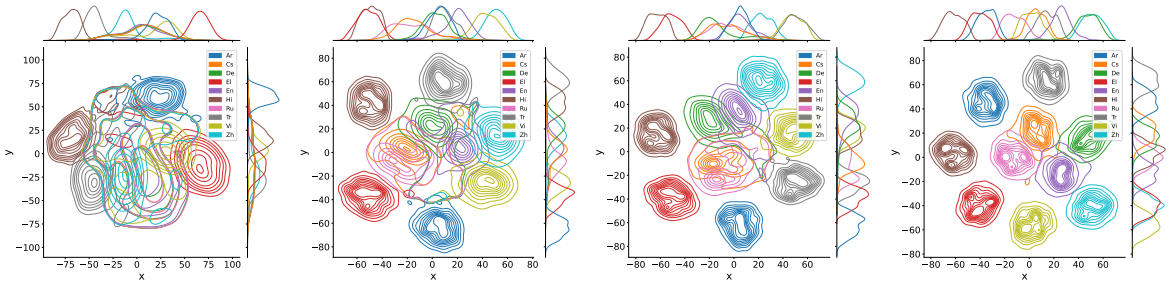
Figure 8: Preliminary visualization of the multilingual representations across layers in the LLaMA3 after dimension reduction. We leverage the FLORES-101 dev set, which is multi-way parallel.



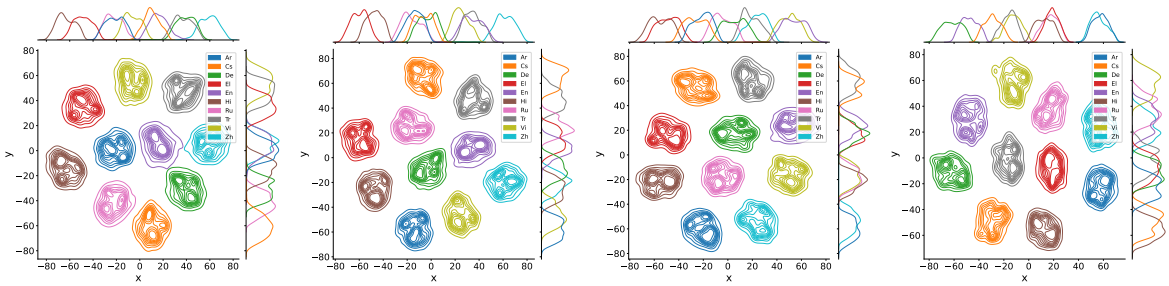
(a) Layer 2 (b) Layer 4 (c) Layer 6 (d) Layer 8



(e) Layer 10 (f) Layer 12 (g) Layer 14 (h) Layer 16



(i) Layer 18 (j) Layer 20 (k) Layer 22 (l) Layer 24



(m) Layer 26 (n) Layer 28 (o) Layer 30 (p) Layer 32

Figure 9: Visualization of the multilingual representations across layers in AlignX, trained on LLaMA3, after dimension reduction. We leverage the FLORES-101 dev set, which is multi-way parallel.

Translation Instruction Set
Translate this text from {src_lang} to {tgt_lang}.
Convert this sentence from {src_lang} to {tgt_lang}.
Change this paragraph from {src_lang} to {tgt_lang}.
Render this message from {src_lang} to {tgt_lang}.
Translate this phrase from {src_lang} to {tgt_lang}.
Turn this text from {src_lang} to {tgt_lang}.
Rewrite this statement from {src_lang} to {tgt_lang}.
Provide a translation from {src_lang} to {tgt_lang} for this text.
Offer a {tgt_lang} translation for this text from {src_lang}.
Give a {tgt_lang} version of this text from {src_lang}.

Table 5: The translation instruction set in this work. "src_lang" denotes the source language, and "tgt_lang" denotes the target language.

Methods	Languages	Data per Language	Total Data
x-LLaMA (Zhu et al., 2023)	6	736.1K sentences	4.4M sentences
SDRRL (Zhang et al., 2024c)	15	150K sentences*	2.3M sentences*
BayLing1 (Zhang et al., 2023)	4	75.5K sentences	302K sentences
BayLing2 (Zhang et al., 2024b)	158	20.3K sentences	3.2M sentences
ParroT (Jiao et al., 2023)	3	67.4K sentences	202.2K sentences
BigTrans (Yang et al., 2023)	102	880.4M tokens + 2.4k sentences	89.8B tokens + 241K sentences
ALMA (Xu et al., 2024)	6	121.7B tokens + 9.8K sentences	703B tokens + 58.7K sentences
AlignX (10langs)	10	104.5K sentences	1.0M sentences
AlignX (51langs)	51	112.9K sentences	5.8M sentences

Table 6: Statistics on corpus size for our approach and some typical systems. We list the number of fine-tuned languages, the total corpus size, and the average corpus size per language. "*" indicates that the detailed corpus size is not given in the paper and we estimate a lower bound.

ISO 639-1	Language	ISO 639-1	Language	ISO 639-1	Language
Af	Akrikaans	Hr	Croatian	Pl	Polish
Ar	Arabic	Id	Indonesian	Ps	Pashto
Az	Azerbaijani	It	Italian	Pt	Portuguese
Bn	Bengali	Ja	Japanese	Ro	Romanian
Cs	Czech	Ka	Georgian	Ru	Russian
De	German	Kk	Kazakh	Si	Sinhala
El	Modern Greek	Km	Khmer	Sl	Slovenian
En	English	Ko	Korean	Sv	Swedish
Es	Spanish	Lt	Lithuanian	Ta	Tamil
Et	Estonian	Lv	Latvian	Te	Telugu
Fa	Persian	Mk	Macedonian	Th	Thai
Fi	Finnish	Ml	Malayalam	Tr	Turkish
Fr	French	Mn	Mongolian	Uk	Ukrainian
Gl	Galician	Mr	Marathi	Ur	Urdu
Gu	Gujarati	My	Burmese	Vi	Vietnamese
He	Hebrew	Ne	Nepali	Xh	Xhosa
Hi	Hindi	Nl	Dutch	Zh	Chinese

Table 7: The languages and corresponding language codes used in this work.

B.4 Languages Included in the 51-Language Setup

To further validate the effectiveness of AlignX across a broader range of languages, we conduct experiments involving 51 languages. We select languages from the intersection of Bactrian-X and OPUS-100, with the addition of Greek. The languages are: Akrikaans (Af), Arabic (Ar), Azerbaijani (Az), Bengali (Bn), Czech (Cs), German (De), Modern Greek (El), English (En), Spanish (Es), Estonian (Et), Persian (Fa), Finnish (Fi), French (Fr), Galician (Gl), Gujarati (Gu), Hebrew (He), Hindi (Hi), Croatian (Hr), Indonesian (Id), Italian (It), Japanese (Ja), Georgian (Ka), Kazakh (Kk), Khmer (Km), Korean (Ko), Lithuanian (Lt), Latvian (Lv), Macedonian (Mk), Malayalam (Ml), Mongolian (Mn), Marathi (Mr), Burmese (My), Nepali (Ne), Dutch (Nl), Polish (Pl), Pashto (Ps), Portuguese (Pt), Romanian (Ro), Russian (Ru), Sinhala (Si), Slovenian (Sl), Swedish (Sv), Tamil (Ta), Telugu (Te), Thai (Th), Turkish (Tr), Ukrainian (Uk), Urdu (Ur), Vietnamese (Vi), Xhosa (Xh), Chinese (Zh).

C More Evaluation Metrics and Detailed Results

C.1 Multilingual General Benchmarks

Table 11, 12 and 13 show results on multilingual TruthfulQA, multilingual HellaSwag, and XNLI benchmark, respectively, and Table 15, 16, 17 and 18 present results on corresponding benchmarks under 51-language setup.

C.2 Multilingual Translation Benchmark

To better evaluate the quality of multilingual translation, we report the results for statistical significance test in Table 19, and COMET (Rei et al., 2020)[5] metric in Table 20, respectively. We present detailed BLEU results for each base LLMs on the multilingual translation benchmark in Table 21, 23, 24, 25 and 26. We present the averaged BLEU scores under the 51-language setup in Table 27.

D Case Study for Multilingual Translation

To intuitively understand the superiority of AlignX translations, we provide some representative, diverse, and multilingual translation cases of LLaMA2-7B and AlignX. We use online Google Translate to translate non-English text into English for clear understanding. Figure 10 presents the cases. While LLaMA2 fails to provide accurate translations, AlignX provides accurate and concise translations. While LLaMA2 fails to provide accurate translations in these cases, including mistranslations, omissions, or even the incorrect target language, AlignX provides accurate and concise translations.

E Additional Analysis

E.1 Comparison with Representation-level AFP

We further compare our approach with AFP (Li et al., 2024a), a SOTA representation-level method. AFP enhances multilingual semantic alignment through an auxiliary contrastive learning objective applied to translation pairs and cross-lingual instruction data. We evaluate AFP on 10 languages, using cross-lingual instructions constructed from approximately 1.15M instances—comparable in scale to AlignX (1.14M instances). Following AFP's optimal hyperparameters, results in Table 8 demonstrate that AlignX outperforms AFP in both multilingual understanding and cross-lingual generation.

We attribute AlignX's superior performance to two key factors. First, AlignX aligns more closely with the LLMs' inherent "align-then-diverge" pattern by combining contrastive learning with language matching, thereby mitigating the performance degradation in generation when aligning semantic representations alone. Second, AlignX integrates contrastive learning and language matching directly within the instruction format. This encourages the model to naturally learn intra-sentence language shifts, leading to more accurate cross-lingual generation compared to AFP's cross-lingual instruction tuning.

E.2 Efficiency Analysis

To evaluate the training efficiency, we compare the training cost between AlignX and CPT-then-SFT, using the same two-stage dataset but without the two auxiliary tasks. All experiments are conducted under identical hardware and hyperparameter settings, including 4×A800-80G GPUs, fixed training steps, batch size, and micro-batch size. As shown in Table 9, AlignX incurs only minimal additional cost, as the auxiliary tasks are integrated

[5]https://huggingface.co/Unbabel/wmt22-comet-da

Translation Direction	Source	Reference	LLaMA2 Outputs	AlignX Outputs
Ru-De	Все животные, которые первоначально прибыли на эти острова, добрались сюда вплавь, прилетели или их прибило течением.	Alle Tiere, die ursprünglich auf den Inseln eintrafen, kamen entweder durch Schwimmen, Fliegen oder Anschwemmen hierher. **(Translation: All the animals that originally arrived on the islands came here either by swimming, flying or washing ashore.)**	"All the animals that first arrived on these islands came by swimming, flying or were washed ashore."	Alle Tiere, die ursprünglich auf diesen Inseln ankamen, kamen hierher zu Fuß, zu Wasser oder wurden von den Strömungen hierher getrieben. **(Translation: All the animals that originally arrived on these islands came here on foot, by water or were carried here by the currents.)**
De-Ar	Dennoch sollten Sie den Rat von Behörden befolgen, alle Schilder beachten und insbesondere Augenmerk auf Sicherheitswarnungen richten.	مع هذا، اطلب النصائح من السلطات، وامتثل لجميع الإشارات، وانتبه جيداً لتحذيرات السلامة. **Translation: However, seek advice) from authorities, obey all signs, and pay close attention to safety warnings.)**	ينبغي أن يتم تقديم الرد للمؤسسات الحكومية ويبلغ كل شاشة ويبلغ من أي ملاحظات الأمن ويبلغ من أي ملاحظات الأم **Translation: The response should be) provided to the government institutions and every screen should be informed and any security notes should be reported and any mother notes should be reported.)**	ومع ذلك يجب أن تتبع توجيهات السلطات وتحتفظ بجميع الأشكال ويجب أن تركز في أولويات الأمان والحذر. **Translation: However, you should follow) the directions of the authorities and keep all forms and should focus on safety and caution priorities.)**
El-En	Από την άλλη πλευρά, οι συνθήκες πάγου και χιονιού είναι φυσιολογικές σε πολλές χώρες, όπου η κυκλοφορία συνεχίζεται ως επί το πλείστον χωρίς διακοπές όλο το χρόνο.	On the other hand, icy and snowy conditions are normal in many countries, and traffic goes on mostly uninterrupted all year round.	From the other side, the patterns of snowflakes and icicles are found in many places, where the cycle continues without interruptions throughout the year.	From the other side, the conditions of snow and ice are natural in many countries, where the circulation continues throughout the year without interruptions.
En-Zh	The Internet combines elements of both mass and interpersonal communication.	互联网结合了大众传播和人际传播的要素。 **(Translation: The Internet combines elements of mass and interpersonal communication.)**	互联网结合了两种通信方式的特点。 **(Translation: The Internet combines the characteristics of both modes of communication.)**	互联网把大众传播和个人交流的元素结合在一起。 **(Translation: The Internet combines elements of mass communication and personal communication.)**

Figure 10: Multilingual translation cases of LLaMA2-7B and AlignX.

	TruthfulQA	HellaSwag	XNLI	XStoryCloze	FLORES-101 (COMET)	FLORES-101 (OTR)
LLaMA-7B Based						
LLaMA-7B	25.91	35.44	35.70	58.75	48.25	48.78
AFP	29.00	36.18	37.30	62.64	62.17	15.37
AlignX	**31.26**	**37.68**	**39.24**	**64.70**	**62.78**	**8.42**
LLaMA2-7B Based						
LLaMA2-7B	28.35	37.79	35.92	60.50	54.34	43.76
AFP	31.79	38.35	36.31	62.66	64.11	17.68
AlignX	**32.99**	**38.99**	**37.76**	**64.29**	**68.64**	**8.07**

Table 8: Comparison between AFP and AlignX. We report averaged Accuracy, COMET, and OTR (off-target ratio) scores under the 10-language setup. We bold the best results.

The Template for Automatic GPT-4o Evaluation

You are a multilingual translation quality evaluator. Given a translation task, including translation direction, source text, and ground truth, compare two translation outputs across multiple evaluation dimensions. "win" stands for output1 is better, "lose" stands for output2 is better, and "tie" stands for two outputs are similar. Output strictly in the following JSON format and **do not include any explanations or reasoning**.

Input:
translation_direction: {translation_direction}
source: {source}
ground_truth: {ground_truth}
output1: {output1}
output2: {output2}

Output Format:
{"grammar": "win/tie/lose", "target_language_fit": "win/tie/lose"}

Figure 11: Template used for automatic GPT-4o evaluation. We randomly stack the translation outputs from LLaMA-8B-Instruct and AlignX, treating them as output1 and output2.

Relative Speed	Training (Stage 1)
CPT+SFT	1.00×
AlignX	0.95×

Table 9: Training efficiency comparison. AlignX introduces negligible additional cost compared to the baseline.

into the same forward pass during Stage 1. Stage 2 and inference incur no further overhead, as no extra parameters or computations are involved beyond CPT-then-SFT.

E.3 Detailed GPT-4o Evaluation Results

We present the GPT-4o evaluation template in Figure 11 and provide detailed results in Table 10, grouped by target language. For high-resource languages (e.g., English and German), GPT-4o predominantly produces neutral judgments. In contrast, for low-resource languages (e.g., Arabic, Greek, and Czech), GPT-4o shows a clear preference for AlignX outputs. These findings highlight the effectiveness of the language-matching task, which proves particularly beneficial in low-resource scenarios.

W/T/L	Grammar	Language Fit
X-En	52 / 394 / 4	80 / 359 / 11
X-Ar	185 / 191 / 74	191 / 165 / 94
X-Cs	127 / 264 / 59	135 / 248 / 67
x-De	103 / 325 / 22	102 / 309 / 39
X-El	173 / 174 / 103	192 / 138 / 120
X-Hi	149 / 203 / 98	144 / 179 / 127
X-Ru	81 / 333 / 36	93 / 320 / 37
X-Tr	108 / 277 / 65	123 / 238 / 89
X-Vi	53 / 380 / 17	80 / 339 / 31
X-Zh	67 / 360 / 23	83 / 316 / 51
All	1098 / 2901 / 501	1223 / 2611 / 666

Table 10: Detailed automatic evaluation results using GPT-4o on cross-lingual generation. We compare paired outputs of AlignX and LLaMA3-8B-Instruct, where "W/T/L" indicates that AlignX produces better, comparable, or worse translations, respectively. "X" denotes all other nine languages except the target language.

Methods	Ar	De	En	Hi	Ru	Vi	Zh	Avg.
Gemma-2B Based								
Gemma-2B	27.94	26.40	23.01	**26.52**	26.78	27.52	26.65	26.40
CPT-then-SFT	31.18	**29.44**	**27.54**	26.26	30.46	29.68	**29.44**	**29.14**
AlignX	**32.08**	29.19	26.93	25.49	**30.58**	**30.06**	28.81	29.02
Mistral-7B-v0.3 Based								
Mistral-7B-v0.3	30.79	31.22	29.01	27.30	31.22	30.19	29.31	29.86
CPT-then-SFT	36.22	**37.82**	36.60	29.24	38.45	35.41	**35.03**	35.54
AlignX	**36.87**	36.93	**38.31**	**29.62**	**39.47**	**36.69**	33.38	**35.89**
LLaMA-7B Based								
LLaMA-7B	28.98	24.49	23.10	28.33	24.75	26.24	25.51	25.91
BayLing1-7B	30.27	28.17	32.07	**29.88**	30.20	27.39	29.95	29.70
ParroT-7B	31.57	28.81	27.17	27.94	29.44	27.77	26.65	28.48
BigTrans-13B	27.17	25.89	27.05	26.26	30.20	20.64	25.38	26.08
CPT-then-SFT	**32.86**	**31.73**	**33.66**	27.04	31.85	30.19	**31.35**	31.24
AlignX	30.92	**30.84**	**33.66**	27.30	**33.38**	**31.85**	30.84	**31.26**
LLaMA2-7B Based								
LLaMA2-7B	30.53	28.68	26.19	28.07	29.44	29.68	25.89	28.35
Tower-7B	29.62	27.03	25.21	27.43	27.66	27.26	24.37	26.94
CPT-then-SFT	31.95	**33.63**	33.41	28.33	35.53	34.78	**31.22**	32.69
AlignX	**33.51**	32.87	**33.90**	**28.46**	**35.79**	**35.29**	31.09	**32.99**
LLaMA3-8B-Instruct Based								
LLaMA3-8B-Instruct	34.67	34.01	35.74	30.53	36.29	32.74	35.53	34.22
BayLing2-8B	29.50	26.27	24.97	26.26	30.08	29.17	28.81	27.87
CPT-then-SFT	34.15	33.38	35.25	29.11	35.15	37.96	35.41	34.34
AlignX	35.45	34.14	38.56	30.14	36.29	38.98	**38.45**	36.00
AlignX (51langs)	**35.58**	**35.66**	**38.80**	**30.66**	**39.85**	**40.89**	37.44	**36.98**

Table 11: Detailed results on multilingual TruthfulQA benchmark. "Avg." denotes average scores on all languages. We bold the highest scores.

Methods	Ar	De	En	Hi	Ru	Vi	Zh	Avg.
Gemma-2B Based								
Gemma-2B	32.12	36.94	52.60	29.74	36.39	35.60	35.82	37.03
CPT-then-SFT	32.14	37.50	53.01	29.85	36.65	35.93	36.37	37.35
AlignX	**32.74**	**37.83**	**53.22**	**30.10**	**37.17**	**36.32**	**36.68**	**37.72**
Mistral-7B-v0.3 Based								
Mistral-7B-v0.3	32.73	43.96	61.80	29.42	43.24	35.73	40.99	41.12
CPT-then-SFT	35.21	43.83	61.82	30.77	43.01	37.58	40.47	41.81
AlignX	**35.83**	**45.16**	**63.71**	**31.44**	**43.85**	**38.05**	**41.91**	**42.85**
LLaMA-7B Based								
LLaMA-7B	28.03	39.04	56.84	26.83	37.00	28.04	32.27	35.44
BayLing1-7B	27.77	39.38	57.61	26.85	36.36	28.11	36.53	36.09
ParroT-7B	27.22	37.93	**59.58**	26.58	35.75	27.83	32.20	35.30
BigTrans-13B	26.74	**41.12**	58.99	26.99	35.34	27.21	**37.23**	36.23
CPT-then-SFT	30.69	39.22	57.49	28.08	37.23	31.54	33.77	36.86
AlignX	**31.10**	40.14	58.97	**28.69**	**38.14**	**32.55**	34.17	**37.68**
LLaMA2-7B Based								
LLaMA2-7B	29.30	39.96	57.60	28.04	37.95	35.43	36.26	37.79
Tower-7B	29.06	**43.70**	57.30	28.46	**41.82**	34.83	35.69	38.69
CPT-then-SFT	30.42	41.65	**58.75**	28.78	39.15	36.08	36.93	38.82
AlignX	**31.05**	41.51	58.67	**28.97**	39.17	**36.33**	**37.25**	**38.99**
LLaMA3-8B-Instruct Based								
LLaMA3-8B-Instruct	36.51	44.20	57.61	33.63	42.41	39.53	41.36	42.18
BayLing2-8B	36.29	42.88	56.58	**33.90**	41.22	39.81	**42.61**	41.90
CPT-then-SFT	36.80	45.00	59.14	32.58	43.47	41.52	41.48	42.86
AlignX	**37.63**	45.61	**59.97**	32.66	43.84	42.10	41.97	43.40
AlignX (51langs)	37.59	**45.73**	59.66	33.52	**43.98**	**42.20**	42.05	**43.53**

Table 12: Detailed results on multilingual HellaSwag benchmark. "Avg." denotes average scores on all languages. We bold the highest scores.

Methods	Ar	De	El	En	Hi	Ru	Tr	Vi	Zh	Avg.
Gemma-2B Based										
Gemma-2B	33.29	41.29	33.49	44.86	36.79	38.19	36.43	36.39	35.82	37.39
CPT-then-SFT	**33.86**	42.33	33.61	43.45	37.19	39.68	37.75	38.39	36.02	38.03
AlignX	33.49	**42.45**	**34.34**	**45.38**	**38.19**	**40.20**	**39.32**	**39.24**	**36.99**	**38.84**
Mistral-7B-v0.3 Based										
Mistral-7B-v0.3	33.90	**43.33**	33.86	43.69	36.55	38.63	36.27	36.71	37.79	37.86
CPT-then-SFT	34.66	43.45	34.34	43.45	39.44	41.04	37.91	38.96	38.39	39.07
AlignX	**35.38**	42.09	**34.78**	**46.35**	**40.40**	**44.74**	**39.64**	**39.72**	**38.63**	**40.19**
LLaMA-7B Based										
LLaMA-7B	33.90	39.04	33.86	42.77	33.69	35.62	34.02	33.90	34.54	35.70
BayLing1-7B	33.69	40.96	33.98	43.86	34.58	**41.73**	34.13	34.02	36.67	37.07
ParroT-7B	33.37	38.92	33.61	46.02	35.34	40.72	33.69	33.98	36.63	36.92
BigTrans-13B	34.02	40.64	34.38	41.89	35.06	39.60	35.46	34.54	35.38	36.77
CPT-then-SFT	34.58	41.57	34.22	46.95	37.63	40.08	36.10	37.47	37.51	38.46
AlignX	**35.38**	**43.25**	**34.66**	**48.39**	**37.31**	39.32	**36.67**	**39.40**	**38.80**	**39.24**
LLaMA2-7B Based										
LLaMA2-7B	33.94	39.00	33.86	41.69	34.94	35.86	34.22	34.70	35.06	35.92
Tower-7B	33.57	41.16	33.61	40.64	35.02	36.99	33.73	34.18	35.38	36.03
CPT-then-SFT	34.14	40.76	33.82	**42.33**	**38.03**	39.56	36.02	**37.43**	37.67	37.75
AlignX	**34.34**	**40.96**	**34.34**	42.29	36.67	**39.88**	**36.31**	36.87	**38.19**	**37.76**
LLaMA3-8B-Instruct Based										
LLaMA3-8B-Instruct	34.90	45.18	**37.11**	44.74	39.96	**41.97**	40.68	38.63	40.40	40.40
BayLing2-8B	34.34	39.12	34.14	41.45	35.74	36.31	38.47	34.90	37.35	36.87
CPT-then-SFT	35.14	45.38	34.74	**47.55**	40.80	41.77	41.53	**42.21**	40.52	41.07
AlignX	35.46	44.38	35.26	46.75	40.12	41.61	40.36	41.73	40.68	40.71
AlignX (51langs)	**36.59**	**46.10**	36.18	45.86	**42.21**	41.16	**43.33**	41.61	**40.84**	**41.54**

Table 13: Detailed results on XNLI benchmark. "Avg." denotes average scores on all languages. We bold the highest scores.

Methods	Ar	En	Hi	Ru	Zh	Avg.
Gemma-2B Based						
Gemma-2B	53.61	75.65	57.71	61.35	58.70	61.40
CPT-then-SFT	55.66	78.23	59.43	63.27	**61.61**	63.64
AlignX	**55.86**	**78.42**	**60.29**	**63.86**	61.15	**63.92**
Mistral-7B-v0.3 Based						
Mistral-7B-v0.3	51.36	80.15	54.86	65.32	63.34	63.00
CPT-then-SFT	**59.43**	82.66	57.71	69.29	66.58	67.13
AlignX	59.36	**83.92**	**60.03**	**70.95**	**67.84**	**68.42**
LLaMA-7B Based						
LLaMA-7B	48.11	77.43	51.89	62.14	54.20	58.75
BayLing1-7B	48.71	80.01	51.42	65.19	**62.54**	61.57
ParroT-7B	48.51	**83.59**	52.35	63.20	53.74	60.28
BigTrans-13B	48.84	80.28	48.18	61.35	61.75	60.08
CPT-then-SFT	55.26	81.47	54.40	66.05	57.58	62.95
AlignX	**57.18**	82.53	**57.11**	**67.44**	59.23	**64.70**
LLaMA2-7B Based						
LLaMA2-7B	50.50	77.83	52.61	62.14	59.43	60.50
Tower-7B	49.24	78.23	53.47	66.58	58.84	61.27
CPT-then-SFT	54.27	80.87	54.86	65.78	**63.73**	63.90
AlignX	**55.06**	**80.94**	**56.06**	**66.05**	63.34	**64.29**
LLaMA3-8B-Instruct Based						
LLaMA3-8B-Instruct	61.75	80.61	**65.25**	71.01	67.90	69.30
BayLing2-8B	59.43	76.51	62.81	67.11	65.65	66.30
CPT-then-SFT	63.20	81.73	63.53	71.08	68.89	69.69
AlignX	62.14	**82.33**	63.53	72.60	**69.36**	69.99
AlignX (51langs)	**63.47**	81.34	64.39	**72.47**	69.23	**70.18**

Table 14: Detailed results on XStoryCloze benchmark. "Avg." denotes average scores on all languages. We bold the highest scores.

TruthfulQA	Ar	Bn	De	En	Es	Fr	Gu	Hi	Hr
LLaMA3-8B-Instruct	34.54	29.71	34.01	35.74	36.76	35.83	27.63	30.53	33.55
AlignX (51langs)	**35.58**	**30.86**	**35.66**	**38.80**	**37.90**	**39.01**	**28.32**	**30.66**	**34.59**
	Id	**It**	**Ml**	**Mr**	**Ne**	**Nl**	**Pt**	**Ro**	**Ru**
LLaMA3-8B-Instruct	34.32	35.89	28.53	28.53	29.33	37.20	37.44	35.04	36.29
AlignX (51langs)	**37.66**	**38.06**	**28.82**	**32.33**	**29.72**	**38.47**	**39.09**	**36.97**	**39.85**
	Sv	**Ta**	**Te**	**Uk**	**Vi**	**Zh**	**Avg.**		
LLaMA3-8B-Instruct	37.60	27.46	26.24	35.84	35.41	35.53	33.29		
AlignX (51langs)	**36.05**	**27.73**	**26.67**	**36.49**	**40.89**	**37.44**	**34.90**		

Table 15: Detailed results on multilingual TruthfulQA benchmark under the 51-language setup. We bold the highest scores.

HellaSwag	Ar	Bn	De	En	Es	Fr	Gu	Hi	Hr
LLaMA3-8B-Instruct	36.54	29.08	44.24	57.61	47.97	46.78	28.06	**33.63**	37.73
AlignX (51langs)	**37.59**	**29.61**	**45.73**	**59.66**	**49.79**	**48.13**	**28.31**	33.52	**39.78**
	Id	**It**	**Ml**	**Mr**	**Ne**	**Nl**	**Pt**	**Ro**	**Ru**
LLaMA3-8B-Instruct	41.73	45.34	26.43	27.86	27.86	44.79	46.05	41.45	42.76
AlignX (51langs)	**43.27**	**46.25**	**26.61**	**28.24**	**28.66**	**45.87**	**48.01**	**42.62**	**43.98**
	Sv	**Ta**	**Te**	**Uk**	**Vi**	**Zh**	**Avg.**		
LLaMA3-8B-Instruct	43.83	**26.02**	26.93	39.36	40.22	41.31	38.48		
AlignX (51langs)	**44.89**	25.97	**27.10**	**40.72**	**42.20**	**42.05**	**39.52**		

Table 16: Detailed results on multilingual HellaSwag benchmark under the 51-language setup. We bold the highest scores.

XNLI	Ar	De	El	En	Es	Fr	Hi	Ru	Sw
LLaMA3-8B-Instruct	34.86	45.18	**37.11**	44.66	41.69	42.29	39.80	**42.05**	**34.94**
AlignX (51langs)	**36.59**	**46.10**	36.18	**45.86**	**44.86**	**44.54**	**42.21**	41.16	34.66
	Th	**Tr**	**Ur**	**Vi**	**Zh**	**Avg.**			
LLaMA3-8B-Instruct	39.28	40.72	**37.51**	38.55	40.52	39.94			
AlignX (51langs)	**41.73**	**43.33**	35.98	**41.61**	**40.84**	**41.12**			

Table 17: Detailed results on XNLI benchmark under the 51-language setup. We bold the highest scores.

XStoryCloze	Ar	En	Es	Hi	Id	My
LLaMA3-8B-Instruct	61.75	80.61	72.60	**65.25**	67.31	**50.10**
AlignX (51langs)	**63.47**	**81.34**	**73.13**	64.39	**68.56**	49.04
	Ru	**Te**	**Zh**	**Avg.**		
LLaMA3-8B-Instruct	71.01	61.02	67.90	66.40		
AlignX (51langs)	**72.47**	**61.09**	**69.23**	**66.97**		

Table 18: Detailed results on XStoryCloze benchmark under the 51-language setup. We bold the highest scores.

System Comparison	Statistical Significance
Gemma-2B v.s. AlignX	84 / 90
Mistral-7B-v0.3 v.s. AlignX	70 / 75
LLaMA-7B v.s. AlignX	81 / 86
LLaMA2-7B v.s. AlignX	90 / 90
LLaMA3-8B-Instruct v.s. AlignX	41 / 55
LLaMA3-8B-Instruct v.s. AlignX (51langs)	79 / 87

Table 19: Statistical significance tests for multilingual translations. We present the results in the *a / b* format, indicating that AlignX outperforms corresponding base LLMs in *b* translation directions, where the improvements in *a* translation directions are significant (p_value < 0.05).

Methods	X-En	X-Ar	X-Cs	X-De	X-El	X-Hi	X-Ru	X-Tr	X-Vi	X-Zh	Avg.
Gemma-2B Based											
Gemma-2B	84.55	58.84	65.20	65.51	61.16	47.02	58.03	59.54	66.23	68.94	63.50
CPT-then-SFT	**85.39**	66.02	71.24	68.44	**75.27**	**71.04**	**73.17**	**74.44**	72.29	53.28	71.06
AlignX	85.29	**67.31**	**72.18**	**73.44**	70.88	53.34	71.07	67.10	**76.59**	**75.71**	**71.29**
Mistral-7B-v0.3 Based											
Mistral-7B-v0.3	84.48	55.65	75.34	55.99	74.30	48.51	75.88	64.37	75.44	39.13	64.91
CPT-then-SFT	**86.12**	67.44	**78.00**	67.01	**78.79**	62.68	**77.28**	72.97	**82.13**	45.18	71.76
AlignX	85.89	**69.75**	77.36	68.06	78.42	**64.36**	74.21	**73.96**	80.72	**47.93**	**72.07**
LLaMA-7B Based											
LLaMA-7B	75.00	40.37	54.13	59.23	41.90	31.76	54.66	39.34	39.68	46.44	48.25
ParroT-7B	73.26	41.27	50.66	52.59	44.57	37.55	42.05	48.77	51.12	48.96	49.08
BayLing1-7B	75.42	45.47	55.07	55.91	44.80	38.46	48.74	46.24	49.15	55.85	51.51
BigTrans-13B	74.06	47.67	55.28	54.75	41.81	43.11	50.54	51.01	46.16	57.95	52.23
CPT-then-SFT	**81.97**	52.74	**69.79**	53.93	**64.40**	**52.98**	68.70	**57.34**	**71.23**	37.58	61.07
AlignX	80.87	**57.97**	69.28	**67.98**	57.80	41.33	**70.23**	56.32	62.20	**63.79**	**62.78**
LLaMA2-7B Based											
LLaMA2-7B	77.49	43.00	61.16	64.16	40.35	34.73	57.70	45.95	61.93	56.93	54.34
Tower-7B	82.20	49.46	62.54	45.12	65.12	45.22	67.40	62.23	61.20	36.98	57.75
CPT-then-SFT	84.70	55.94	**75.96**	55.52	**74.11**	**50.75**	77.10	**71.81**	**77.67**	42.27	66.58
AlignX	**85.08**	**60.82**	75.92	**77.14**	56.04	41.80	**79.61**	57.69	75.23	**77.05**	**68.64**
LLaMA3-8B-Instruct Based											
LLaMA3-8B-Instruct	86.62	75.12	83.98	81.97	78.77	63.49	84.51	78.50	83.26	82.10	79.83
Bayling2-8B	86.41	75.09	78.40	76.46	78.41	**77.82**	76.66	**81.51**	81.55	64.92	77.72
CPT-then-SFT	80.23	71.73	75.22	72.08	76.80	74.31	72.10	76.38	77.32	57.85	73.40
AlignX	86.85	78.78	84.07	81.65	**81.24**	63.67	85.18	78.51	83.31	82.70	80.60
AlignX (51langs)	**87.61**	**79.81**	**84.70**	**82.49**	78.25	63.98	**85.73**	79.20	**84.60**	**83.46**	**80.98**

Table 20: The averaged COMET (Rei et al., 2020) scores on FLORES-101. "X" denotes all other training languages except the target language. "Avg." denotes average scores on all translation directions. We bold the highest scores.

Gemma-2B	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	13.18	20.98	26.93	13.48	16.34	22.70	14.71	29.78	18.80	19.66
Ar	29.79	-	4.28	6.26	3.19	1.43	1.89	1.57	6.01	4.69	6.57
Cs	34.67	5.08	-	10.62	3.99	0.39	7.37	3.23	2.86	7.35	8.40
De	40.67	5.55	6.71	-	5.94	5.42	4.21	4.37	8.65	10.00	10.17
El	29.11	1.16	3.37	5.85	-	0.73	0.91	1.18	4.51	3.11	5.55
Hi	27.36	0.60	2.70	4.00	1.01	-	1.00	2.55	2.97	4.72	5.21
Ru	31.54	4.08	9.18	10.39	6.97	2.69	-	3.18	6.69	8.07	9.20
Tr	25.48	1.19	1.96	2.73	0.91	0.78	0.61	-	2.19	1.58	4.16
Vi	29.69	1.87	2.97	3.20	2.25	2.19	1.72	2.82	-	6.41	5.90
Zh	25.79	1.23	5.40	6.78	4.26	0.69	3.13	3.34	10.31	-	6.77
Avg.	30.46	3.77	6.39	8.53	4.67	3.41	4.84	4.11	8.22	7.19	8.16
CPT-then-SFT	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	14.09	21.22	27.78	15.28	16.77	23.02	16.17	30.21	20.10	20.52
Ar	30.41	-	6.05	11.50	8.67	4.97	4.91	4.11	11.68	6.19	9.83
Cs	35.66	7.92	-	15.86	9.71	5.77	17.92	6.43	9.44	10.50	13.25
De	41.12	9.13	11.57	-	11.09	9.48	16.40	10.29	19.39	14.03	15.83
El	30.81	5.53	4.08	11.71	-	3.80	8.76	3.82	9.59	7.72	9.54
Hi	28.47	4.61	3.05	8.22	5.27	-	3.93	3.67	5.18	4.95	7.48
Ru	33.20	7.92	5.43	14.34	10.56	8.62	-	7.52	17.19	12.52	13.03
Tr	26.69	4.82	3.32	8.29	5.15	2.44	5.05	-	6.96	7.23	7.77
Vi	31.75	5.79	8.01	10.31	7.50	5.70	8.98	6.74	-	11.75	10.73
Zh	26.51	6.40	8.77	12.55	7.70	6.86	10.49	6.44	16.49	-	11.36
Avg.	31.62	7.36	7.94	13.40	8.99	7.16	11.05	7.24	14.01	10.55	11.93
AlignX	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	14.47	21.49	27.83	15.38	16.87	23.13	16.02	29.83	19.81	20.54
Ar	30.27	-	7.20	12.72	8.31	5.70	6.20	5.45	16.38	8.58	11.20
Cs	35.73	7.42	-	15.86	9.59	5.60	18.18	7.78	12.31	11.15	13.74
De	41.47	8.63	9.96	-	10.97	8.33	14.82	10.40	18.89	13.04	15.17
El	31.14	4.58	5.71	12.21	-	4.72	8.35	5.20	15.56	8.25	10.64
Hi	27.77	3.93	3.38	8.49	5.32	-	2.08	5.08	8.90	6.42	7.93
Ru	33.55	7.88	10.54	15.80	10.11	8.97	-	7.31	17.46	12.78	13.82
Tr	26.68	3.60	3.90	8.72	5.13	3.12	4.56	-	7.46	7.23	7.82
Vi	31.65	5.70	7.62	10.34	7.13	5.83	9.56	7.22	-	11.91	10.77
Zh	26.82	6.24	8.96	12.55	7.54	7.42	8.82	7.04	16.03	-	11.27
Avg.	31.68	6.94	8.75	13.84	8.83	7.40	10.63	7.94	15.87	11.02	12.29

Table 21: Detailed BLEU scores of the FLORES-101 benchmark on Gemma-2B.

Mistral-7B-v0.3	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	6.81	27.30	32.95	4.24	6.50	28.21	11.44	17.66	17.92	17.00
Ar	27.37	-	9.83	13.82	1.54	2.06	9.39	2.71	7.27	6.00	8.89
Cs	39.36	4.44	-	24.67	2.94	3.60	23.31	5.81	9.89	12.60	14.07
De	43.69	4.88	22.01	-	2.88	4.17	23.23	7.58	12.35	13.55	14.93
El	26.20	3.08	10.98	14.53	-	2.04	12.24	3.67	7.65	6.55	9.66
Hi	21.61	1.94	5.76	8.91	1.04	-	5.69	3.73	5.04	5.73	6.61
Ru	35.88	4.80	20.29	22.52	3.11	3.97	-	6.53	11.75	12.59	13.49
Tr	26.89	2.68	8.40	12.69	1.51	3.75	12.30	-	7.00	9.11	9.37
Vi	30.29	3.76	11.15	15.48	1.73	2.56	15.23	5.40	-	10.44	10.67
Zh	27.33	2.59	12.85	15.39	1.38	2.76	11.55	5.11	8.81	-	9.75
Avg.	30.96	3.89	14.29	17.88	2.26	3.49	15.68	5.78	9.71	10.50	11.44
CPT-then-SFT	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	16.21	28.75	32.61	11.61	12.61	28.31	18.36	24.52	20.24	21.47
Ar	32.65	-	12.41	14.45	5.37	3.60	13.14	6.51	11.12	8.93	12.02
Cs	40.75	9.28	-	20.96	7.68	6.31	21.98	10.29	14.58	14.07	16.21
De	45.30	9.76	18.86	-	8.64	7.85	21.75	12.04	16.98	15.03	17.36
El	33.75	7.07	15.00	17.68	-	4.52	16.59	8.08	12.68	10.29	13.96
Hi	27.87	4.80	9.68	11.71	4.34	-	10.75	7.66	8.71	8.93	10.49
Ru	37.33	8.84	20.55	21.34	7.83	6.07	-	10.39	14.85	13.68	15.65
Tr	29.47	5.17	6.16	8.63	5.13	5.64	11.59	-	8.66	10.09	10.06
Vi	32.05	6.42	10.92	14.07	5.74	4.36	14.50	7.36	-	11.81	11.91
Zh	29.40	5.65	13.13	14.70	5.21	4.57	13.27	8.80	12.68	-	11.93
Avg.	34.29	8.13	15.05	17.35	6.84	6.17	16.88	9.94	13.86	12.56	14.11
AlignX	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	15.90	28.19	31.06	11.62	13.11	27.37	18.98	22.05	19.47	20.86
Ar	32.52	-	12.67	14.44	5.55	4.26	14.00	7.24	12.28	10.33	12.59
Cs	39.89	8.52	-	21.16	7.55	6.80	21.38	10.30	15.11	13.89	16.07
De	44.36	9.46	19.45	-	8.74	8.37	21.10	11.85	16.30	14.52	17.13
El	33.53	6.96	8.77	14.92	-	5.18	15.25	8.28	13.45	11.15	13.05
Hi	28.03	4.64	7.93	8.55	3.89	-	10.25	7.68	9.62	8.53	9.90
Ru	36.74	8.54	19.64	19.23	7.94	6.55	-	11.02	15.72	13.90	15.48
Tr	29.35	5.33	8.47	11.20	4.93	6.21	11.71	-	9.85	10.74	10.87
Vi	32.13	6.79	12.59	14.27	5.75	4.84	14.68	8.36	-	12.49	12.43
Zh	28.11	5.89	13.51	14.07	4.98	5.31	13.57	9.38	13.23	-	12.01
Avg.	33.85	8.00	14.58	16.54	6.77	6.74	16.59	10.34	14.18	12.78	14.04

Table 22: Detailed BLEU scores of the FLORES-101 benchmark on Mistral-7B-v0.3.

LLaMA-7B	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	2.40	19.71	27.24	2.16	2.71	21.16	3.85	3.08	5.65	9.77
Ar	13.56	-	0.77	3.00	0.10	0.10	1.41	0.17	0.22	0.15	2.16
Cs	34.83	1.17	-	16.94	1.19	1.49	14.86	1.55	1.45	1.52	8.33
De	40.18	1.44	11.38	-	1.45	1.79	13.55	2.31	1.97	2.86	8.55
El	16.40	0.14	1.73	4.85	-	0.20	2.77	0.28	0.45	0.45	3.03
Hi	9.89	0.09	0.59	2.29	0.11	-	0.80	0.25	0.15	0.06	1.58
Ru	32.22	1.10	11.29	14.84	1.27	1.40	-	0.90	0.99	2.71	7.41
Tr	10.19	0.15	0.99	2.77	0.31	0.42	1.47	-	0.93	0.59	1.98
Vi	9.93	0.10	1.43	2.12	0.36	0.31	0.90	0.75	-	0.38	1.81
Zh	20.13	0.09	3.39	7.40	0.23	0.35	2.44	0.96	0.68	-	3.96
Avg.	20.81	0.74	5.70	9.05	0.80	0.97	6.60	1.22	1.10	1.60	4.86
ParroT-7B	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	0.53	18.50	28.85	1.15	1.10	13.69	3.19	2.58	12.62	9.13
Ar	7.82	-	0.57	0.71	0.18	0.18	0.22	0.54	0.60	0.13	1.22
Cs	32.60	0.19	-	2.71	0.45	0.15	0.58	2.02	2.29	0.76	4.64
De	38.36	0.22	2.39	-	0.47	0.16	0.62	2.06	2.18	0.64	5.23
El	9.95	0.17	0.97	1.38	-	0.14	0.35	1.05	1.25	0.22	1.72
Hi	5.67	0.11	0.52	0.54	0.19	-	0.24	0.59	0.71	0.11	0.96
Ru	29.72	0.17	1.95	2.25	0.40	0.13	-	1.53	1.76	0.98	4.32
Tr	9.56	0.20	1.58	1.92	0.44	0.15	0.25	-	2.02	0.45	1.84
Vi	7.78	0.19	1.53	1.76	0.46	0.17	0.32	1.74	-	0.36	1.59
Zh	19.76	0.18	1.55	1.71	0.34	0.18	0.44	1.63	1.68	-	3.05
Avg.	17.91	0.22	3.28	4.65	0.45	0.26	1.86	1.59	1.67	1.81	3.37
BayLing1-7B	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	2.34	19.97	30.54	1.86	1.89	17.23	3.68	3.98	19.80	11.25
Ar	11.23	-	0.61	0.78	0.17	0.17	0.22	0.59	0.64	0.26	1.63
Cs	36.89	0.55	-	5.94	1.16	0.56	6.50	2.21	2.29	2.71	6.53
De	42.72	0.87	10.99	-	1.34	0.81	7.86	2.70	2.68	5.71	8.41
El	14.16	0.22	1.38	1.50	-	0.19	0.42	0.92	1.18	0.38	2.26
Hi	8.71	0.15	0.63	0.80	0.16	-	0.29	0.63	0.64	0.22	1.36
Ru	33.66	0.35	5.92	3.65	0.87	0.29	-	1.66	1.62	2.04	5.56
Tr	11.29	0.26	1.93	2.85	0.57	0.44	1.04	-	2.14	0.90	2.38
Vi	11.25	0.32	1.96	2.57	0.49	0.27	0.56	1.96	-	0.58	2.22
Zh	24.81	0.26	2.44	2.42	0.41	0.23	0.39	1.67	1.64	-	3.81
Avg.	21.64	0.59	5.09	5.67	0.78	0.54	3.83	1.78	1.87	3.62	4.54
BigTrans-13B	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	5.01	25.23	25.50	3.02	9.48	19.04	11.49	6.03	15.56	13.37
Ar	10.09	-	1.90	1.28	0.55	1.95	1.76	0.82	0.99	1.87	2.36
Cs	36.23	2.84	-	4.80	1.55	4.96	6.08	4.20	2.99	7.93	7.95
De	38.06	2.36	3.82	-	1.54	4.57	4.78	4.04	2.84	4.21	7.36
El	5.61	0.15	0.71	0.77	-	0.57	0.70	0.84	0.87	0.59	1.20
Hi	12.81	1.16	0.97	1.47	0.74	-	1.09	1.37	0.97	4.30	2.76
Ru	30.17	2.61	2.66	2.82	1.70	4.38	-	3.10	2.51	3.80	5.97
Tr	20.54	1.42	2.37	3.01	1.04	2.50	2.28	-	2.23	3.57	4.33
Vi	12.89	0.73	1.87	2.30	0.79	1.02	1.03	2.19	-	2.74	2.84
Zh	25.26	2.65	7.86	5.84	1.49	6.81	5.19	6.11	3.44	-	7.18
Avg.	21.30	2.10	5.27	5.31	1.38	4.03	4.66	3.80	2.54	4.95	5.53

Table 23: Detailed BLEU scores of the FLORES-101 benchmark on LLaMA-7B (part 1).

CPT-then-SFT	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	10.30	24.08	28.80	8.22	7.93	23.71	11.49	15.14	12.37	15.78
Ar	23.79	-	6.01	7.70	0.99	0.66	6.44	2.06	3.38	2.80	5.98
Cs	36.37	4.49	-	19.09	4.68	3.38	18.49	5.82	7.65	7.81	11.98
De	41.17	4.49	14.53	-	5.38	3.68	16.39	7.13	8.14	8.48	12.15
El	24.94	0.61	2.48	3.98	-	0.15	4.24	2.03	3.50	2.71	4.96
Hi	17.95	0.49	3.99	5.45	0.16	-	4.09	1.82	2.21	1.54	4.19
Ru	33.53	3.84	16.94	17.33	4.45	3.00	-	5.55	6.92	7.20	10.97
Tr	19.88	1.56	3.55	6.63	2.24	2.56	6.10	-	4.26	3.72	5.61
Vi	23.27	1.64	4.78	8.00	1.97	1.25	6.45	3.58	-	4.12	6.12
Zh	22.32	1.49	8.06	10.12	1.99	1.34	8.40	3.61	5.17	-	6.94
Avg.	27.02	3.21	9.38	11.90	3.34	2.66	10.48	4.79	6.26	5.64	8.47
AlignX	**En**	**Ar**	**Cs**	**De**	**El**	**Hi**	**Ru**	**Tr**	**Vi**	**Zh**	**Avg.**
En	-	11.55	22.81	26.19	9.79	8.86	21.46	13.14	17.60	12.14	15.95
Ar	22.59	-	6.35	8.11	2.88	1.35	6.24	2.89	4.94	2.74	6.45
Cs	34.25	4.53	-	17.22	5.97	3.77	17.34	6.56	8.94	7.20	11.75
De	38.44	5.02	11.58	-	6.35	3.94	13.66	7.52	9.46	7.39	11.48
El	24.57	2.69	4.15	5.48	-	2.25	7.95	3.28	6.33	2.95	6.63
Hi	16.30	1.09	4.21	5.25	1.67	-	4.00	2.77	3.26	1.68	4.47
Ru	31.75	4.03	15.09	15.97	5.71	3.63	-	6.08	8.63	6.71	10.84
Tr	18.18	1.80	3.96	7.05	2.92	2.97	4.92	-	4.67	3.20	5.52
Vi	22.52	2.11	5.23	7.52	2.86	1.82	6.23	3.90	-	3.71	6.21
Zh	20.51	1.79	7.69	9.70	2.41	1.64	7.83	4.27	6.72	-	6.95
Avg.	25.46	3.85	9.01	11.39	4.51	3.36	9.96	5.60	7.84	5.30	8.63

Table 24: Detailed BLEU scores of the FLORES-101 benchmark on LLaMA-7B (part 2).

LLaMA2-7B	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	3.76	23.28	30.07	3.01	4.20	22.54	6.87	24.64	14.65	14.78
Ar	16.20	-	1.93	4.62	0.14	0.05	1.08	0.33	2.62	0.39	3.04
Cs	37.30	1.73	-	20.18	1.65	2.16	16.04	2.49	10.35	7.96	11.10
De	41.71	1.99	17.75	-	1.98	2.32	16.12	4.35	17.50	9.61	12.59
El	16.96	0.17	2.14	4.93	-	0.14	1.58	0.49	2.65	0.38	3.27
Hi	11.31	0.05	1.87	3.20	0.11	-	1.04	0.71	1.43	0.27	2.22
Ru	33.69	1.69	16.27	17.95	1.79	1.86	-	2.27	14.16	7.74	10.82
Tr	17.52	0.48	2.85	6.90	0.62	1.33	3.87	-	4.28	2.38	4.47
Vi	30.75	1.42	8.78	13.12	1.09	1.08	10.81	2.48	-	5.96	8.39
Zh	26.23	1.08	6.05	9.57	1.06	1.45	5.08	1.74	9.57	-	6.87
Avg.	25.74	1.37	8.99	12.28	1.27	1.62	8.68	2.41	9.69	5.48	7.76

Tower-7B	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	3.80	20.85	38.97	2.94	4.86	33.97	5.79	21.00	25.39	17.51
Ar	21.32	-	2.12	4.13	0.54	0.21	0.94	0.73	3.69	1.62	3.92
Cs	39.63	1.72	-	20.98	1.59	1.96	15.29	1.99	6.69	13.60	11.49
De	45.31	2.24	14.98	-	1.93	2.79	23.09	3.44	13.20	17.58	13.84
El	22.87	0.25	2.60	6.38	-	0.35	2.36	1.05	4.72	1.34	4.66
Hi	18.38	0.09	1.35	2.56	0.34	-	0.94	0.73	2.54	0.55	3.05
Ru	38.30	2.22	13.68	16.48	1.94	2.38	-	2.46	12.57	9.64	11.07
Tr	21.47	0.85	2.74	6.72	0.64	1.24	5.36	-	3.40	4.85	5.25
Vi	35.66	1.31	5.06	10.44	1.13	0.88	5.33	2.27	-	8.75	7.87
Zh	31.39	1.10	6.26	8.08	0.92	1.44	4.04	1.90	9.22	-	7.15
Avg.	30.48	1.51	7.74	12.75	1.33	1.79	10.15	2.26	8.56	9.26	8.58

CPT-then-SFT	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	9.07	25.85	32.45	6.52	8.32	26.44	11.77	26.20	18.41	18.34
Ar	26.07	-	10.17	12.74	0.51	0.57	11.25	2.67	10.43	7.48	9.10
Cs	39.60	4.25	-	22.33	3.80	2.54	18.13	6.31	15.26	12.85	13.90
De	44.03	5.87	19.76	-	4.59	4.38	20.01	8.17	18.96	14.65	15.60
El	27.86	0.10	9.49	14.07	-	0.12	9.00	1.55	7.27	3.64	8.12
Hi	23.80	0.22	7.39	10.49	0.08	-	8.26	2.67	6.31	6.09	7.26
Ru	35.72	4.96	19.78	21.96	3.95	3.69	-	6.57	16.62	13.44	14.08
Tr	24.45	1.24	5.62	11.40	1.59	1.29	7.62	-	7.10	6.74	7.45
Vi	34.36	2.61	11.15	18.09	1.80	1.31	15.57	4.31	-	11.34	11.17
Zh	28.60	2.80	12.26	16.08	1.79	2.17	13.20	4.33	13.59	-	10.54
Avg.	31.61	3.46	13.50	17.73	2.74	2.71	14.39	5.37	13.53	10.52	11.55

AlignX	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	10.48	25.76	32.04	7.67	8.65	26.37	12.64	26.56	18.31	18.72
Ar	27.55	-	10.26	12.96	3.00	1.81	11.54	3.32	12.62	7.77	10.09
Cs	39.59	6.14	-	22.39	5.09	4.34	20.06	7.07	16.93	13.66	15.03
De	44.07	6.67	18.69	-	5.42	4.92	19.67	8.71	19.17	14.65	15.77
El	28.80	2.34	11.51	15.16	-	2.97	12.76	2.99	13.37	7.47	10.82
Hi	24.34	1.59	7.47	10.62	1.92	-	8.65	3.83	9.70	7.03	8.35
Ru	35.72	6.36	20.37	21.90	5.38	4.91	-	7.31	17.80	14.03	14.86
Tr	25.25	2.91	4.89	10.58	2.63	3.47	8.47	-	8.64	8.52	8.37
Vi	34.38	4.95	10.15	17.40	3.88	3.16	15.76	5.15	-	13.28	12.01
Zh	28.24	4.07	11.85	15.52	3.02	3.28	13.39	5.64	15.17	-	11.13
Avg.	31.99	5.06	13.44	17.62	4.22	4.17	15.19	6.30	15.55	11.64	12.52

Table 25: Detailed BLEU scores of the FLORES-101 benchmark on LLaMA2-7B.

LLaMA3-8B-Instruct	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	19.84	29.14	34.97	22.59	24.18	29.12	24.19	34.37	22.71	26.79
Ar	34.30	-	17.84	21.34	15.69	13.92	20.16	14.14	23.71	14.93	19.56
Cs	39.19	14.84	-	26.63	17.30	16.78	24.19	16.65	24.76	16.88	21.91
De	42.08	15.43	23.80	-	18.67	18.60	24.54	18.44	26.25	18.41	22.91
El	34.10	13.76	19.31	22.65	-	14.38	21.03	15.08	24.36	16.07	20.08
Hi	33.51	12.02	15.59	19.43	13.87	-	17.10	15.59	21.47	14.84	18.16
Ru	34.68	14.57	21.57	23.66	16.85	15.93	-	15.84	24.68	17.28	20.56
Tr	32.65	12.40	14.76	19.72	14.68	15.59	18.19	-	20.63	15.73	18.26
Vi	33.58	13.26	17.46	20.66	14.73	14.22	19.81	13.74	-	16.54	18.22
Zh	26.57	11.35	15.15	17.88	13.69	14.22	16.52	13.68	21.62	-	16.74
Avg.	34.52	14.16	19.40	22.99	16.45	16.42	21.18	16.37	24.65	17.04	20.32

BayLing2-8B	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	20.71	26.02	33.80	20.33	25.40	27.74	21.87	34.28	25.99	26.24
Ar	34.86	-	13.58	16.54	13.10	13.14	16.34	12.12	21.00	11.17	16.87
Cs	38.11	14.61	-	15.44	14.73	15.56	19.20	12.43	20.29	16.25	18.51
De	43.00	15.51	19.83	-	16.57	18.13	22.67	15.81	25.58	18.62	21.75
El	34.11	12.56	14.11	18.44	-	13.31	16.53	12.59	20.76	13.86	17.36
Hi	33.67	12.36	12.85	17.46	12.25	-	15.31	14.83	20.99	14.05	17.09
Ru	35.71	13.32	15.45	19.52	14.75	14.58	-	13.61	22.36	14.76	18.23
Tr	33.16	11.77	9.06	11.59	11.18	14.82	13.07	-	15.08	13.16	14.77
Vi	34.36	13.04	11.14	16.35	12.24	14.14	17.00	10.61	-	15.64	16.06
Zh	29.88	10.60	13.81	15.80	13.16	15.10	15.38	12.96	21.39	-	16.45
Avg.	35.21	13.83	15.09	18.33	14.26	16.02	18.14	14.09	22.41	15.94	18.33

CPT-then-SFT	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	23.02	30.38	36.01	24.94	27.11	29.83	24.74	35.22	25.35	28.51
Ar	37.32	-	18.60	22.10	18.11	12.86	20.37	13.00	24.03	16.91	20.37
Cs	40.18	15.50	-	26.77	17.68	15.60	24.25	14.69	24.81	19.03	22.06
De	42.84	16.09	23.83	-	19.42	17.60	24.05	17.24	26.34	20.18	23.07
El	35.75	13.35	18.13	22.86	-	12.76	21.31	11.81	22.75	17.84	19.62
Hi	35.91	12.58	16.97	20.65	15.51	-	18.35	14.93	21.31	17.36	19.29
Ru	35.91	15.26	21.78	24.01	20.02	15.51	-	15.94	24.22	18.43	21.23
Tr	34.98	12.60	14.68	21.68	14.83	14.05	18.19	-	20.29	17.88	18.80
Vi	36.49	13.55	16.56	22.35	14.94	14.32	19.82	11.02	-	18.78	18.65
Zh	29.16	11.45	16.18	19.38	14.94	15.00	17.51	14.00	22.88	-	17.83
Avg.	36.50	14.82	19.68	23.98	17.82	16.09	21.52	15.26	24.65	19.08	20.94

AlignX	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	23.21	29.89	35.39	24.93	26.57	29.53	23.51	35.05	24.47	28.06
Ar	37.62	-	18.80	22.27	17.28	13.05	20.05	12.90	23.04	15.80	20.09
Cs	41.63	16.12	-	26.91	18.22	15.67	24.44	13.48	22.81	18.60	21.99
De	44.68	16.26	23.09	-	19.08	18.11	24.30	15.42	24.40	19.56	22.77
El	35.23	14.59	18.94	23.08	-	13.42	20.73	12.42	22.65	16.08	19.68
Hi	36.92	12.63	17.05	20.41	15.18	-	17.99	15.08	20.38	17.09	19.19
Ru	36.88	15.20	22.10	23.92	18.97	15.34	-	15.10	23.89	18.03	21.05
Tr	36.40	11.68	13.83	19.90	15.19	15.51	18.01	-	17.75	17.53	18.42
Vi	36.90	14.12	16.94	21.83	15.90	14.06	19.71	11.34	-	18.42	18.80
Zh	30.02	11.02	16.31	18.51	14.83	14.62	17.07	13.22	21.61	-	17.47
Avg.	37.36	14.98	19.66	23.58	17.73	16.26	21.31	14.72	23.51	18.40	20.75

AlignX (51langs)	En	Ar	Cs	De	El	Hi	Ru	Tr	Vi	Zh	Avg.
En	-	23.64	30.43	36.44	19.37	27.91	30.36	23.14	36.11	25.45	28.09
Ar	39.31	-	18.85	22.65	18.05	14.72	20.18	14.79	24.62	16.18	21.04
Cs	42.14	16.72	-	27.83	19.15	17.73	24.99	16.57	25.86	19.19	23.35
De	45.96	17.50	24.77	-	19.71	19.26	24.77	18.85	27.44	20.25	24.28
El	38.63	15.91	20.76	23.91	-	15.55	21.90	16.09	25.21	17.70	21.74
Hi	37.27	13.73	17.67	20.60	15.72	-	18.25	17.15	22.09	17.20	19.96
Ru	38.36	15.71	22.37	25.10	19.17	16.31	-	16.56	25.16	18.65	21.93
Tr	36.90	13.93	16.31	22.03	15.86	16.75	19.33	-	22.25	17.97	20.15
Vi	38.34	15.21	18.47	22.70	15.93	15.15	20.43	14.32	-	19.10	19.96
Zh	31.42	12.70	16.74	19.95	15.69	16.55	17.78	14.60	24.22	-	18.85
Avg.	38.70	16.12	20.71	24.58	17.63	17.77	22.00	16.90	25.88	19.08	21.94

Table 26: Detailed BLEU scores of the FLORES-101 benchmark on LLaMA3-8B-Instruct.

FLORES-101	X-Af	X-Ar	X-Az	X-Bn	X-Cs	X-De	X-El	X-En	X-Et
LLaMA3-8B-Instruct	16.64	11.71	**7.47**	8.10	16.21	19.48	14.15	30.30	10.65
AlignX (51langs)	**16.67**	**13.75**	7.18	**8.41**	**17.70**	**21.29**	**16.02**	**34.87**	**12.77**
	X-Fa	**X-Fi**	**X-Fr**	**X-Gl**	**X-Gu**	**X-He**	**X-Hi**	**X-Hr**	**X-Id**
LLaMA3-8B-Instruct	14.30	12.80	26.06	18.12	4.10	12.30	14.97	13.39	19.37
AlignX (51langs)	**14.92**	**13.06**	**28.29**	**19.47**	**4.48**	**13.19**	**16.60**	**14.70**	**21.25**
	X-It	**X-Ja**	**X-Ka**	**X-Kk**	**X-Km**	**X-Ko**	**X-Lv**	**X-Lt**	**X-Ml**
LLaMA3-8B-Instruct	19.11	16.27	**2.79**	**8.00**	0.76	12.45	10.64	10.35	3.01
AlignX (51langs)	**20.36**	**18.32**	2.09	6.38	**1.17**	**13.50**	**13.29**	**12.32**	**3.14**
	X-Mr	**X-Mk**	**X-Mn**	**X-My**	**X-Nl**	**X-Ne**	**X-Pl**	**X-Pt**	**X-Ps**
LLaMA3-8B-Instruct	6.84	16.07	**2.68**	**0.95**	17.28	6.94	14.07	23.46	**3.24**
AlignX (51langs)	**7.43**	**18.41**	2.39	0.81	**19.20**	**7.86**	**14.83**	**25.28**	2.91
	X-Ro	**X-Ru**	**X-Si**	**X-Sl**	**X-Es**	**X-Sv**	**X-Ta**	**X-Te**	**X-Th**
LLaMA3-8B-Instruct	19.29	17.91	1.84	12.93	19.02	19.05	3.67	**4.09**	12.16
AlignX (51langs)	**20.18**	**19.38**	**1.88**	**14.47**	**19.81**	**21.29**	**3.80**	3.92	**12.30**
	X-Tr	**X-Uk**	**X-Ur**	**X-Vi**	**X-Xh**	**X-Zh**	**Avg.**		
LLaMA3-8B-Instruct	13.68	16.79	8.65	20.63	0.85	14.45	12.35		
AlignX (51langs)	**15.02**	**15.64**	**9.09**	**22.34**	**2.90**	**16.66**	**13.39**		

Table 27: The averaged BLEU scores on FLORES-101 under the 51-language setup. "X" denotes all other training languages except the target language. "Avg." denotes average scores on all translation directions. We bold the highest scores.