Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models

Anonymous ACL submission

Abstract

The training paradigm for machine translation has gradually shifted, from learning neural machine translation (NMT) models with extensive parallel corpora to instruction finetuning on multilingual large language models (LLMs) with high-quality translation pairs. In this paper, we focus on boosting many-to-many multilingual translation of LLMs with an emphasis on zero-shot translation directions.[1] We demonstrate that prompt strategies adopted during finetuning are crucial to zero-shot translation and introduce a cross-lingual consistency regularization, XConST, to bridge the representation gap among different languages and improve zero-shot translation performance. XConST is not a new method, but a version of CrossConST (Gao et al., 2023a) adapted for translation instruction finetuning with LLMs. Experimental results on ALMA (Xu et al., 2023), Tower (Team, 2024), and LLaMA-2 (Touvron et al., 2023) show that our approach consistently improves translation performance.

1 Introduction

Large language models (LLMs) such as GPT models (OpenAI, 2023a,b), PaLM models (Chowdhery et al., 2022; Anil et al., 2023), and others have demonstrated remarkable capabilities in multilingual translation (Jiao et al., 2023; Vilar et al., 2023). For example, GPT-4 exhibits superior performance over the state-of-the-art (SOTA) multilingual neural machine translation (NMT) models, such as NLLB models (Team et al., 2022), for high-resource languages (Hendy et al., 2023; Zhu et al., 2023). To further boost LLMs' translation capability, the techniques explored by existing approaches include in-context learning (Lin et al., 2022; Zhu et al., 2023), continual pretraining with parallel data (Yang et al., 2023; Team, 2024), and multilingual finetuning with translation instructions (Zhang et al., 2023b; Zeng et al., 2023; Li et al., 2023; Xu et al., 2023).

Specifically, Xu et al. (2023) introduce an LLM-based training strategy for multilingual translation, which consists of two stages: continual pretraining on monolingual datasets followed by instruction finetuning on a small set of high-quality translation pairs. By leveraging LLaMA-2 as the foundational model, they learn the ALMA models that outperform NLLB-54.5B and GPT-3.5 across ten translation directions on average. Despite their impressive supervised performance in the English-centric directions, the zero-shot translation evaluation among non-English languages is neglected.

In this paper, our primary goal is to provide a simple yet effective strategy for learning many-to-many multilingual translation with LLMs. Inspired by Wu et al. (2021) and Gao et al. (2023a), we investigate the impact of prompt strategies on multilingual translation and show that the zero-shot translation performance varies significantly across different strategies. We then propose a cross-lingual consistency regularization, XConST, for boosting zero-shot translation, where we introduce the explicit constraints to the semantic-equivalent sentence pairs by leveraging Kullback-Leibler (KL) regularization. The contributions of this paper can be summarized as follows:

- We show that prompt strategies are crucial for zero-shot translation, and there isn't a universal strategy that excels in every scenario.
- We propose XConST, a simple but effective training strategy that could consistently boost the zero-shot translation performance of LLMs across different prompt strategies.
- Experimental results on ALMA, Tower, and LLaMA-2 models show that XConST could achieve significant performance improvement on many-to-many multilingual translation.

[1] In this work, supervised or zero-shot translation refers to translating between language pairs seen or unseen during multilingual finetuning with translation instructions.

2 Background

2.1 Language Tag for Multilingual NMT

The multilingual NMT model has a promising capability of zero-shot translation, where it could directly translate between language pairs unseen during training. Language tag is often adopted in multilingual NMT for indicating the translation directions. Wu et al. (2021) investigate four widely used language tag strategies, T-ENC, T-DEC, S-ENC-T-ENC, and S-ENC-T-DEC, demonstrated in Table 1. They show that language tag is not only the translation direction indicator but also crucial for the zero-shot translation performance. They conclude that ignoring the source language tag and placing the target language tag on the encoder side, which is T-ENC, could enhance the semantics representation consistency among languages and boost the zero-shot translation performance.

Prompt Strategy	Template
T-ENC	[tgt]: <SRC>\n<TGT>
T-DEC	<SRC>\n[tgt]:<TGT>
S-ENC-T-ENC	[src] [tgt]: <SRC>\n<TGT>
S-ENC-T-DEC	[src]: <SRC>\n[tgt]:<TGT>
GPT-MT	Translate this from [src] into [tgt]:\n [src]: <SRC>\n[tgt]:<TGT>

Table 1: Different prompt strategies adopted in this paper, where [src] and [tgt] denote the source and target languages, and <SRC> and <TGT> denote the source and target sentences. Note that our examples are slightly different from those in Wu et al. (2021) since our strategies are designed for decoder-only LLM while theirs are for NMT model with encoder-decoder architecture.

2.2 Cross-lingual Consistency Regularization for Multilingual NMT

Gao et al. (2023a) introduce a cross-lingual consistency regularization, CrossConST, to bridge the representation gap among different languages and improve zero-shot translation in multilingual NMT. For each sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective of CrossConST is defined as:

$$\mathcal{L}_{CrossConST}(\theta) = \mathcal{L}_{ce}^{mt}(\theta) + \alpha \mathcal{L}_{kl}^{mt}(\theta), \quad (1)$$

where

$$\mathcal{L}_{ce}^{mt}(\theta) = \ell(f(\mathbf{x}, \mathbf{y}; \theta), \ddot{\mathbf{y}}), \quad (2)$$

ℓ denotes the cross-entropy loss, θ is a set of model parameters, $f(\mathbf{x}, \mathbf{y}; \theta)$ is a sequence of probability predictions, $\ddot{\mathbf{y}}$ is a sequence of one-hot labels,

$$\mathcal{L}_{kl}^{mt}(\theta) = \mathrm{KL}(f(\mathbf{x}, \mathbf{y}; \theta) \| f(\mathbf{y}, \mathbf{y}; \theta)), \quad (3)$$

$\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions, and α is a scalar hyper-parameter that balances $\mathcal{L}_{ce}^{mt}(\theta)$ and $\mathcal{L}_{kl}^{mt}(\theta)$.

3 Datasets and Baseline Settings

In this section, we describe the datasets used in Section 4 as well as the model configurations.

Datasets We consider the English-centric scenario and conduct our experiments on five languages: Czech (cs), German (de), English (en), Russian (ru), and Chinese (zh). We collect test data from WMT17 to WMT20 (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020) and dev data from FLORES-200 (Team et al., 2022) as the training dataset. The detailed statistics of all datasets are summarized in Table 10. For supervised direction evaluation, we collect the test data from WMT22 (Kocmi et al., 2022) resulting in 8 translation directions. For zero-shot direction evaluation, we collect the devtest data from FLORES-200 resulting in 12 translation directions. We evaluate the model performance by case-sensitive SacreBLEU[2] (Post, 2018) and COMET[3] (Rei et al., 2020).

Settings We adopt ALMA-7B-Pretrain (and 13B) from Xu et al. (2023)'s release[4] as our pretrained models which are full-weight continual pretrained on LLaMA-2-7B (and 13B) with monolingual datasets in six languages[5]. We explore both full-weight and low-rank adaptation (LoRA) (Hu et al., 2022) finetuning in our experiments. When LoRA is utilized, we apply LoRA to the down-projection layer in each feed-forward layer and set the rank to 16. Note that only 0.1% parameters (7.7M for the 7B model and 12M for the 13B model) are updated when using LoRA. We finetune the pretrained models with a batch size of 256 and set the maximum sequence length to 512. We use the AdamW optimizer with learning rates 2e-5 and 1e-4 for full-weight and LoRA finetuning respectively. We apply beam search decoding with beam size 5. We train all models with llama-recipes[6] toolkit for one epoch on 8 NVIDIA Tesla A100 GPUs.

4 Methodology

In this section, we investigate the impact of prompt strategies on many-to-many multilingual trans-

[2]We apply zh tokenizer for Chinese and 13a for others.
[3]https://huggingface.co/Unbabel/wmt22-comet-da
[4]https://huggingface.co/haoranxu
[5]Czech, German, English, Icelandic, Russian, and Chinese.
[6]https://github.com/facebookresearch/llama-recipes

Prompt Strategy	LoRA	Supervised	Zero-Shot	Pivot	Supervised	Zero-Shot	Pivot
		Performance on ALMA-7B-Pretrain			*Performance on ALMA-13B-Pretrain*		
T-ENC	×	31.6 / 84.8	21.2 / 84.5	20.6 / 85.4	33.0 / 85.4	**23.4 / 86.3**	21.5 / 85.8
T-DEC	×	31.7 / 84.9	**22.4 / 86.5**	20.3 / 85.2	32.9 / 85.5	19.7 / 81.3	21.6 / 85.9
S-ENC-T-ENC	×	31.6 / 84.8	2.4 / 58.7	20.9 / 85.4	33.1 / 85.5	10.6 / 68.0	21.5 / 85.9
S-ENC-T-DEC	×	31.7 / 85.0	**22.4 / 86.0**	20.7 / 85.4	32.8 / 85.4	**20.3 / 84.0**	21.4 / 85.8
GPT-MT	×	31.7 / 85.0	22.1 / 85.5	20.5 / 85.3	33.0 / 85.4	15.2 / 73.9	21.9 / 86.0
T-ENC	√	32.1 / 84.6	21.9 / 85.4	22.6 / 86.1	34.4 / 85.6	**22.4 / 85.1**	24.7 / 86.8
T-DEC	√	32.5 / 85.2	**22.4 / 86.1**	22.5 / 86.0	34.3 / 85.6	**22.6 / 83.9**	24.7 / 86.8
S-ENC-T-ENC	√	32.2 / 84.8	19.4 / 83.9	22.7 / 86.1	34.5 / 85.6	10.7 / 66.6	24.8 / 86.8
S-ENC-T-DEC	√	32.5 / 85.1	21.7 / 85.2	22.8 / 86.2	34.5 / 85.7	5.1 / 60.5	24.9 / 86.9
GPT-MT	√	32.5 / 85.1	**22.0 / 85.5**	22.7 / 86.1	34.4 / 85.6	14.0 / 70.4	24.8 / 86.8

Table 2: Performance (SacreBLEU / COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA vanilla finetuning on ALMA-7B-Pretrain and ALMA-13B-Pretrain. "LoRA" denotes whether LoRA is utilized during the finetuning procedure. "Pivot" denotes the zero-shot translation performance where we first translate the source language into English before generating the target language. The two highest scores of each setting in the zero-shot directions are marked in bold. The detailed results are summarized in Tables 12 and 13.

lation with LLMs (Section 4.1) and propose a cross-lingual consistency regularization method for boosting zero-shot translation (Section 4.2). We introduce the details of each part below.

4.1 Multilingual Finetuning with Translation Instructions

Following the similar lines of Li et al. (2023) and Xu et al. (2023), we finetune the pretrained LLMs with high-quality parallel datasets as the translation instructions. Consider a decoder-only Transformer (Vaswani et al., 2017) model parameterized by θ. Let $\mathbf{x}$ and $\mathbf{y}$ denote the source and target sentences respectively. A fixed prompt template, denoted as $\mathcal{I}$, is applied to guide the model for translating sentences. The training objective is defined as:

$$\mathcal{L}_{ce}^{llm}(\theta) = \ell(f(\mathbf{x}, \mathbf{y}, \mathcal{I}; \theta), \ddot{\mathbf{y}}), \quad (4)$$

where $f(\mathbf{x}, \mathbf{y}, \mathcal{I}; \theta)$ is a sequence of probability predictions, i.e.,

$$f_j(\mathbf{x}, \mathbf{y}, \mathcal{I}; \theta) = P(y|\mathbf{x}, \mathbf{y}_{<j}, \mathcal{I}; \theta). \quad (5)$$

Note that (4) is a standard causal language modeling (CLM) loss, which predicts the next token in a sequence of tokens, and the model can only attend to tokens on the left. We only compute loss for the target sentence during instruction finetuning.

4.1.1 Does Prompt Strategy Matter for Multilingual Translation?

We adopt five prompt strategies listed in Table 1, where four strategies (T-ENC, T-DEC, S-ENC-T-ENC, and S-ENC-T-DEC) are widely used for learning multilingual NMT, and GPT-MT (Hendy et al., 2023) is a popular translation prompt adopted in the LLM field. We report averaged BLEU and COMET scores of all prompt strategies in Table 2. By checking model performance under different combinations of prompt and finetuning strategies, we have the following observations:

- All prompt strategies achieve comparable performance in the supervised directions. LoRA finetuning slightly outperforms full-weight finetuning, especially regarding the BLEU scores.
- The zero-shot translation performance varies significantly across different prompt strategies, and there is no single strategy that performs the best for all cases. Although 13B models achieve superior supervised performance, they generally underperform 7B models in the zero-shot directions, which can not be resolved by utilizing LoRA.
- When prompt strategies are selected properly, the zero-shot translation could achieve comparable or better performance than the pivot-based approach, which is quite reasonable: pretrained LLMs already have translation capability, and translation instructions are only used to elicit it.

4.1.2 Visualization Analysis

We here conduct experiments on the multi-way parallel devtest data from FLORES-200, where 1012 sentences have translations in different languages. We construct the prompts with `xx` $\rightarrow$ `ru` translation instruction and calculate the sentence representations by extracting the last Transformer layer outputs of the last token in the input sequence. We utilize the models that are LoRA finetuned on ALMA-13B-Pretrain in Table 2 and apply dimension reduction on the 5120-dimensional sentence

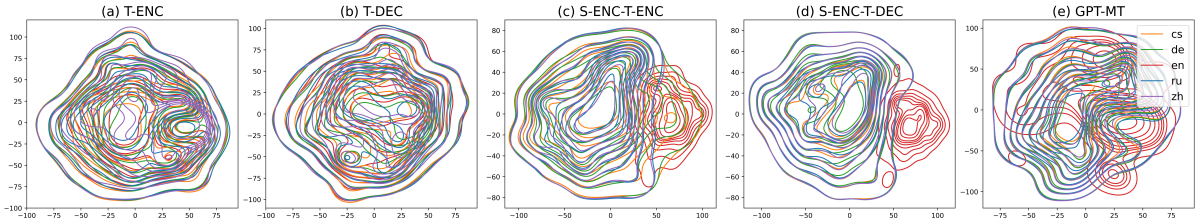

Figure 1: Bivariate kernel density estimation plots of sentence representations after using T-SNE dimensionality reduction on the multi-way parallel devtest data from FLORES-200. The orange line denotes Czech, the green line denotes German, the red line denotes English, the blue line denotes Russian, and the purple line denotes Chinese.

representations with T-SNE (Hinton and Roweis, 2002) and then depict the bivariate kernel density estimation based on the 2-dimensional representations in Figure 1. We can see that our models cannot align these five languages well in the representation space. It is worth mentioning that T-ENC and T-DEC align the languages much better than the other three strategies, which is also reflected in the zero-shot translation performance in Table 2.

4.1.3 Can Off-Target Issue be Mitigated by Instruction Generalization?

We here apply fastText[7] (Joulin et al., 2016, 2017) toolkit for language identification and calculate the off-target ratio (Table 3) in the zero-shot directions for each model demonstrated in Table 2. The off-target ratio is the average percentage of sentences translated into the wrong languages. We can see that the bad translation performance in the zero-shot directions is highly related to the off-target issue which could also be regarded as the failure of LLMs to follow the translation instructions. We then investigate two strategies for improving the instruction-following capability of LLMs.

Prompt Strategy	LoRA	Off-Target 7B (%)	Off-Target 13B (%)
T-ENC	×	5.94	2.33
T-DEC	×	0.11	16.58
S-ENC-T-ENC	×	92.32	60.80
S-ENC-T-DEC	×	1.38	12.78
GPT-MT	×	2.68	38.07
T-ENC	√	3.38	8.23
T-DEC	√	1.08	10.91
S-ENC-T-ENC	√	11.08	66.11
S-ENC-T-DEC	√	3.94	87.22
GPT-MT	√	3.10	50.96

Table 3: The off-target ratio in the zero-shot translation directions for each model demonstrated in Table 2.

- We diversify the translation instructions, where the prompt strategy is randomly selected from Table 1 during instruction finetuning.
- We leverage the additional general instructions, Alpaca dataset (Taori et al., 2023), when finetuning on LLMs with translation instructions, where Alpaca is a dataset of 52K instructions generated by OpenAI's text-davinci-003 engine.

[7] https://fasttext.cc/docs/en/language-identification.html

Method	Supervised	Zero-shot
7B w/ full-weight finetune	31.7 / 85.0	22.1 / 85.5
+ Diversification	31.6 / 85.0	22.2 / 86.1
+ Alpaca dataset	31.0 / 84.7	21.7 / 85.2
7B w/ LoRA finetune	32.5 / 85.1	22.0 / 85.5
+ Diversification	32.7 / 85.2	22.3 / 86.2
+ Alpaca dataset	31.8 / 84.8	17.4 / 78.9
13B w/ full-weight finetune	33.0 / 85.4	15.2 / 73.9
+ Diversification	32.9 / 85.4	12.0 / 68.9
+ Alpaca dataset	31.8 / 84.9	10.0 / 69.4
13B w/ LoRA finetune	34.4 / 85.6	14.0 / 70.4
+ Diversification	34.3 / 85.6	13.1 / 69.7
+ Alpaca dataset	34.2 / 85.5	18.4 / 77.8

Table 4: Performance (SacreBLEU / COMET) on WMT22 and FLORES-200 benchmarks with different instruction finetuning strategies.

We generate translations with GPT-MT prompt strategy and summarize the experimental results in Table 4. We can see that instruction diversification could slightly improve the zero-shot translation performance for 7B models but does not work for 13B models. The zero-shot translation can not be consistently improved by introducing general instructions such as Alpaca, and it has a negative impact on the supervised translation performance, which might be due to the decreased ratio of the translation instructions in the finetuning dataset.

4.2 Cross-lingual Consistency Regularization for Translation Instruction Finetuning

Inspired by CrossConST for multilingual encoder-decoder models (Gao et al., 2023a), a natural question arises: *Can we boost decoder-only LLMs' zero-shot translation by leveraging cross-lingual consistency regularization?* The main idea is to close the

Prompt Strategy	LoRA	Supervised	Zero-Shot	Supervised	Zero-Shot
		Performance on ALMA-7B-Pretrain		*Performance on ALMA-13B-Pretrain*	
T-ENC	×	31.7 / 84.9	22.3 (+1.1) / 86.3 (+1.8)	33.2 / 85.5	23.7 (+0.3) / 86.8 (+0.5)
T-DEC	×	31.6 / 84.9	22.3 (-0.1) / 86.5 (+0.0)	32.8 / 85.4	23.7 (+4.0) / 87.0 (+5.7)
S-ENC-T-ENC	×	31.4 / 84.8	21.6 (+19.2) / 85.8 (+27.1)	33.1 / 85.5	23.5 (+12.9) / 86.7 (+18.7)
S-ENC-T-DEC	×	31.6 / 85.0	22.3 (-0.1) / 86.1 (+0.1)	32.6 / 85.4	24.0 (+3.7) / 87.1 (+3.1)
GPT-MT	×	31.6 / 85.0	22.4 (+0.3) / 86.4 (+0.9)	32.8 / 85.4	23.7 (+8.5) / 87.0 (+13.1)
T-ENC	√	32.1 / 84.7	22.0 (+0.1) / 86.1 (+0.7)	34.3 / 85.7	23.7 (+1.3) / 86.5 (+1.4)
T-DEC	√	32.5 / 85.2	22.1 (-0.3) / 86.3 (+0.2)	34.3 / 85.8	24.3 (+1.7) / 87.0 (+3.1)
S-ENC-T-ENC	√	31.8 / 84.8	21.5 (+2.1) / 85.8 (+1.9)	34.3 / 85.7	24.0 (+13.3) / 86.5 (+19.9)
S-ENC-T-DEC	√	32.5 / 85.1	22.4 (+0.7) / 86.3 (+1.1)	34.1 / 85.7	24.0 (+18.9) / 86.7 (+26.2)
GPT-MT	√	32.5 / 85.1	22.4 (+0.4) / 86.3 (+0.8)	34.2 / 85.7	23.9 (+9.9) / 86.7 (+16.3)

Table 5: Performance (SacreBLEU / COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA cross-lingual consistency finetuning on ALMA-7B-Pretrain and ALMA-13B-Pretrain. The zero-shot performance changes are pointed out in the parentheses. The detailed results are summarized in Tables 14 and 15.

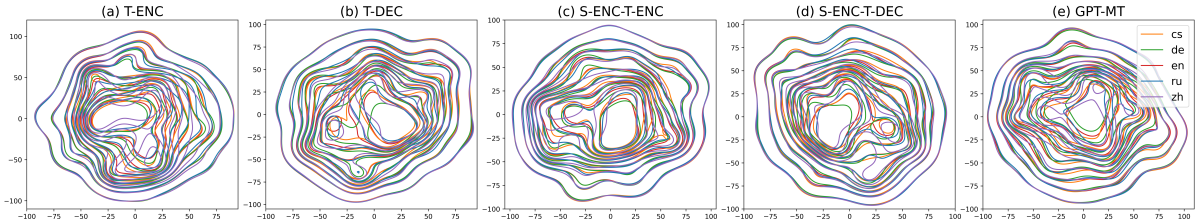

Figure 2: Bivariate kernel density estimation plots of sentence representations after using T-SNE dimensionality reduction on the multi-way parallel devtest data from FLORES-200.

representation gap among semantic-equivalent sentences in different languages and force the output distribution of the model to be consistent among different semantic-equivalent representations. During finetuning, for each sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective of XConST is defined as:

$$\mathcal{L}_{XConST}(\theta) = \mathcal{L}_{ce}^{llm}(\theta) + \alpha \mathcal{L}_{kl}^{llm}(\theta), \quad (6)$$

where

$$\mathcal{L}_{kl}^{llm}(\theta) = \text{KL}(f(\mathbf{x}, \mathbf{y}, \mathcal{I}; \theta) \| f(\mathbf{y}, \mathbf{y}, \mathcal{I}; \theta)). \quad (7)$$

The gradients are backpropagated through both sides of the KL regularization in XConST. Figure 3 illustrates XConST regularization for finetuning multilingual LLM with translation instructions.

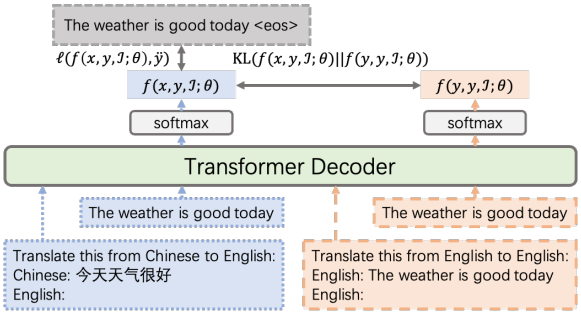

Figure 3: Illustration of XConST for translation instruction finetuning, where the original Chinese-English sentence pair ("今天天气很好", "The weather is good today") and the copied English-English sentence pair ("The weather is good today", "The weather is good today") are fed into the multilingual LLM to generate two output distributions $f(\mathbf{x}, \mathbf{y}, \mathcal{I}; \theta)$ and $f(\mathbf{y}, \mathbf{y}, \mathcal{I}; \theta)$.

4.2.1 Experimental Results

Following the experimental setup in Section 4.1.1, we utilize the cross-lingual consistency regularization during translation instruction finetuning and choose α in (6) from $\{0.05, 0.1, 0.25\}$. We report averaged BLEU and COMET scores of all prompt strategies in Table 5. By checking model performance under different combinations of prompt and finetuning strategies in Tables 2 and 5, we can see that XConST could achieve strong performance in both supervised and zero-shot translation directions across different prompt strategies. For example, XConST improves the zero-shot translation performance by 3.46 and 10.8 COMET scores on average for 7B and 13B models respectively.

Following the experimental setup in Section 4.1.2, we visualize the sentence representations of the multi-way parallel devtest data from FLORES-200 in Figure 2. By comparing the representation spaces illustrated in Figures 1 and 2, we can see that the languages cannot be aligned well in the representation space by leveraging vanilla finetuning procedure, while XConST draws the representations across different languages much closer.

Model	WMT22	FLORES-200
GPT-4 Turbo	36.7 / 87.2	27.4 / 88.3
GPT-3.5 Turbo	34.8 / 86.2	24.6 / 86.9
NLLB-54.5B	31.5 / 81.9	24.5 / 86.2
ALMA-13B-LoRA	34.4 / 85.7	14.8 / 69.8
Ours	34.2 / 85.7	23.9 / 86.7

Table 6: The translation performance of the SOTA models on WMT22 and FLORES-200 benchmarks.

Comparison with SOTA Models We summarize the translation results of several SOTA models on WMT22 and FLORES-200 benchmarks as follows:

- **GPT-3.5 Turbo / GPT-4 Turbo**: The translation performance by leveraging the OpenAI API[8] with greedy decoding for inference. We use the prompt strategy demonstrated in Table 16.
- **NLLB-54.5B**: The largest and best multilingual NMT model with encoder-decoder architecture released by the No Language Left Behind (NLLB) project (Team et al., 2022).
- **ALMA-13B-LoRA**: The best translation model released by Xu et al. (2023), which involves monolingual data finetuning on LLaMA-2 and high-quality parallel data LoRA finetuning.

The experimental results are reported in Table 6. Our model, which is cross-lingual consistency finetuned on ALMA-13B-Pretrain with LoRA and GPT-MT prompt strategy, substantially outperforms ALMA-13B-LoRA in the zero-shot translation directions. Note that the training data of ALMA-13B-LoRA and ours are almost the same except that they include devtest data from FLORES-200 for instruction finetuning. It is also worth mentioning that our model achieves strong performance on the FLORES-200 benchmark compared with NLLB-54.5B which utilizes extensive parallel sentence pairs in those translation directions.

4.2.2 Effect of α

We here investigate the impact of the scalar hyperparameter α in XConST. α is a penalty parameter that controls the regularization strength in our optimization problem. We vary α in $\{0.0, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0\}$ and conduct the experiments with LoRA cross-lingual consistency finetuning on ALMA-13B-Pretrain. Note that cross-lingual consistency finetuning is simplified to vanilla finetuning when $\alpha = 0.0$. The averaged

[8]https://api.openai.com/v1/chat/completions

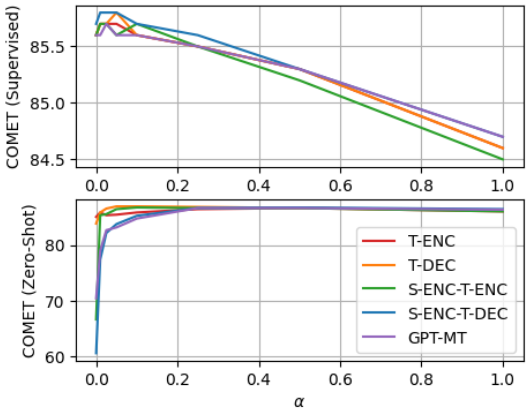

Figure 4: The averaged COMET scores with different α on WMT22 and FLORES-200 benchmarks.

COMET scores are illustrated in Figure 4 due to their better alignment with human evaluations (Freitag et al., 2022). We can see that a small α (e.g. 0.1) is enough to consistently boost the translation performance in the zero-shot directions across different prompt strategies. Meanwhile, an overwhelming regularization ($\alpha = 1.0$) is not plausible for the supervised translation performance.

4.2.3 Does Cross-lingual Consistency Still Work Beyond English-centric Scenario?

Method	WMT22	FLORES-200	
		cs ↔ de	others
T-ENC	34.6 / 85.6	25.9 / 89.2	23.9 / 86.1
+ XConST	34.5 / 85.7	25.8 / 89.2	24.2 / 86.6
T-DEC	34.5 / 85.7	25.9 / 89.3	16.9 / 73.4
+ XConST	34.4 / 85.7	25.6 / 89.2	23.6 / 86.4
S-ENC-T-ENC	34.4 / 85.5	25.8 / 89.1	24.0 / 86.2
+ XConST	34.4 / 85.7	25.9 / 89.2	24.3 / 86.7
S-ENC-T-DEC	34.5 / 85.6	25.6 / 89.1	17.3 / 76.0
+ XConST	34.3 / 85.7	25.5 / 89.1	23.8 / 86.4
GPT-MT	34.3 / 85.7	25.8 / 89.1	17.8 / 74.6
+ XConST	34.3 / 85.7	25.6 / 89.2	23.9 / 86.4

Table 7: Performance on the WMT22 and FLORES-200 benchmarks with LoRA cross-lingual consistency finetuning on ALMA-13B-Pretrain.

We here extend our experiments beyond the English-centric scenario. Specifically, we gather the English-centric datasets used in Section 4.2.1 and supplement it with an additional 22990 cs ↔ de sentence pairs, which are collected from WMT10 to WMT13 (Callison-Burch et al., 2010, 2011, 2012; Bojar et al., 2013) test datasets. FLORES-200 benchmark then includes two supervised directions and ten zero-shot directions. We report averaged BLEU and COMET scores of the baseline and our approach in Table 7. By check-

ing model performance under different combinations of datasets and training strategies in Tables 2, 5 and 7, we can see that adding beyond the English-centric dataset usually could improve the overall zero-shot translation performance except for T-DEC in our experiments. The cross-lingual consistency regularization is complementary to the data-based method and could further improve the zero-shot translation performance.

5 Experiments on More Languages

To ensure the generalization of our findings, we extend our analysis to a broader range of languages and investigate many-to-many machine translation performance with the Tower and LLaMA-2 models. We utilize LoRA finetuning in our experiments and apply the same training and inference configurations discussed in Section 3. We evaluate the model performance on the FLORES-200 devtest datasets.

5.1 Experiments with Tower Model

Dataset Description and Model Configuration We extract the English-centric translation pairs from TowerBlocks[9] dataset, which includes data for multiple translation-related tasks, such as automatic post edition, error span prediction, named-entity recognition, machine translation, and its different variants. The detailed statistics of the translation pairs are summarized in Table 11. We conduct our experiments on ten languages[10] and adopt TowerBase-7B[11] as our pretrained model, which is full-weight continual pretrained on LLaMA-2-7B with monolingual and parallel datasets in those ten languages. We fix α in (6) to be 0.25.

Main Results We report averaged BLEU and COMET scores of all prompt and finetuning strategies in Table 8, where we also include the experimental results of TowerInstruct[12], a language model that results from finetuning TowerBase-7B on the TowerBlock dataset. We can see that XConST consistently improves the zero-shot translation performance across different prompt strategies and performs comparably or even better than TowerInstruct. It is worth mentioning that TowerInstruct utilizes translation pairs beyond English-centric direction and sophisticated instruction diversification mechanisms, while ours adopts a fixed translation prompt template on the English-centric pairs with cross-lingual consistency regularization.

Method	Data	English-centric	Others
TowerInstruct	①	37.0 / 88.6	24.6 / 86.1
T-ENC	②	37.4 / 88.5	23.8 / 85.2
+ XConST	②	37.6 / 88.6	25.0 / 86.0
T-DEC	②	37.6 / 88.6	24.5 / 85.8
+ XConST	②	37.7 / 88.7	25.2 / 86.2
S-ENC-T-ENC	②	37.5 / 88.6	21.6 / 82.8
+ XConST	②	37.7 / 88.6	25.1 / 86.1
S-ENC-T-DEC	②	37.8 / 88.6	24.0 / 85.2
+ XConST	②	37.8 / 88.7	25.3 / 86.2
GPT-MT	②	38.0 / 88.7	24.4 / 85.7
+ XConST	②	38.0 / 88.7	25.2 / 86.2

Table 8: Performance (SacreBLEU / COMET) on FLORES-200 benchmark with LoRA vanilla and cross-lingual consistency finetuning on TowerBase-7B. ① denotes the TowerBlocks dataset. ② denotes the English-centric translation pairs extracted from TowerBlock. The results with all translation pairs in TowerBlock are summarized in Table 17.

5.2 Experiments with LLaMA-2 Model

Dataset Description and Model Configuration We select 28 languages[13] from three different language families (Indo-European-Germanic, Indo-European-Romance, and Indo-European-Slavic), where the LLaMA-2 models achieve impressive few-shot performance in the English-centric translation directions (Zhu et al., 2023). We collect 10K English-centric parallel sentences from the OPUS collection (Tiedemann, 2012) for each language by adopting the data cleaning process as follows: 1) We remove duplicate sentence pairs and also discard sentence pairs wherein the English sentences exceed 2000 characters. 2) Language identification filtering is applied by utilizing the fastText toolkit. 3) Semantic similarity filtering is performed based on the MuSR model (Gao et al., 2023b). In summary, we collect 540K cleaned English-centric sentence pairs covering 28 languages including English. Note that the same sentence pairs are shared by `xx` $\leftarrow$ `en` and `xx` $\rightarrow$ `en` directions. Due to limited

[9] https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.1

[10] Chinese (`zh`), Dutch (`nl`), English (`en`), French (`fr`), German (`de`), Italian (`it`), Korean (`ko`), Portuguese (`pt`), Russian (`ru`), and Spanish (`es`)

[11] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

[12] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1

[13] Afrikaans (`af`), Asturian (`ast`), Belarusian (`be`), Bosnian (`bs`), Bulgarian (`bg`), Catalan (`ca`), Croatian (`hr`), Czech (`cs`), Danish (`da`), Dutch (`nl`), English (`en`), French (`fr`), Galician (`gl`), German (`de`), Icelandic (`is`), Luxembourgish (`lb`), Macedonian (`mk`), Norwegian (`nb`), Occitan (`oc`), Polish (`pl`), Portuguese (`pt`), Romanian (`ro`), Russian (`ru`), Serbian (`sr`), Slovak (`sk`), Slovenian (`sl`), Spanish (`es`), and Ukrainian (`uk`).

Prompt Strategy	xx ↔ en	xx ↔ es	xx ↔ ru	xx ↔ en	xx ↔ es	xx ↔ ru
	Performance on LLaMA-2-7B			*Performance on LLaMA-2-13B*		
T-ENC	29.6 / 82.2	15.1 / 76.8	14.9 / 78.4	32.0 / 83.6	14.6 / 76.1	13.6 / 75.7
+ XConST	29.6 / 82.4	16.2 / 78.5	15.9 / 80.5	31.9 / 83.8	17.8 / 80.1	18.0 / 82.2
T-DEC	29.7 / 82.2	16.4 / 78.7	16.6 / 80.8	31.8 / 83.5	17.8 / 80.3	18.4 / 82.6
+ XConST	29.6 / 82.4	16.7 / 79.2	16.5 / 81.3	31.8 / 83.6	18.3 / 80.8	18.5 / 83.0
S-ENC-T-ENC	29.7 / 82.3	4.2 / 65.5	3.3 / 58.4	32.1 / 83.8	11.3 / 73.9	10.9 / 71.1
+ XConST	29.6 / 82.5	15.9 / 78.5	15.4 / 79.5	32.1 / 83.9	17.8 / 80.5	18.1 / 82.1
S-ENC-T-DEC	29.9 / 82.3	15.5 / 78.0	14.2 / 75.6	32.0 / 83.7	13.5 / 76.6	14.8 / 74.4
+ XConST	29.8 / 82.4	17.0 / 79.4	16.8 / 81.5	31.9 / 83.8	18.3 / 80.9	18.9 / 83.2
GPT-MT	29.9 / 82.3	15.7 / 78.3	13.4 / 74.1	32.1 / 83.8	17.5 / 80.2	16.1 / 77.1
+ XConST	29.8 / 82.4	16.8 / 79.3	16.8 / 81.6	32.0 / 83.8	18.3 / 80.9	18.8 / 83.1

Table 9: Performance (SacreBLEU / COMET) on FLORES-200 benchmark with LoRA vanilla and cross-lingual consistency finetuning on LLaMA-2-7B and LLaMA-2-13B.

computational resources, we only generate Spanish-centric and Russian-centric translations for zero-shot evaluation. We adopt LLaMA-2-7B (and 13B) as our pretrained models and fix α in (6) to be 0.1.

Main Results We report averaged BLEU and COMET scores of all prompt and finetuning strategies in Table 9. All prompt strategies achieve comparable performance in the supervised directions, while the zero-shot translation performance varies significantly. Specifically, S-ENC-T-ENC usually performs the worst in our experiments. XConST consistently boosts the zero-shot translation performance across different prompt strategies. For example, XConST improves the xx ↔ ru translation performance by 7.40 and 6.56 COMET scores on average for 7B and 13B models respectively.

6 Related Work

Recent works on LLMs demonstrate their strong capabilities in multilingual translation (Jiao et al., 2023; Hendy et al., 2023; Zhu et al., 2023). To further improve the translation performance, one direction is to find good in-context learning recipes for machine translation with LLMs (Lin et al., 2022; Agrawal et al., 2023; Zhang et al., 2023a; Zhu et al., 2023). For example, Zhang et al. (2023a) study how to prompt LLMs for machine translation and find that prompt demonstrations and templates have a substantial impact on translation qualities.

Another direction is to utilize parallel datasets for continual pretraining (Yang et al., 2023; Team, 2024) and translation instruction finetuning (Zhang et al., 2023b; Li et al., 2023; Xu et al., 2023). Specifically, Li et al. (2023) focus on eliciting LLMs' translation capability via multilingual finetuning with translation instructions. Although the zero-shot translation performance is investigated, they adopt XGLM-7B (Lin et al., 2022) as the foundation model which has relatively weak multilingual capability (Zhu et al., 2023), and the translation performance is far from satisfactory. Xu et al. (2023) propose a two-stage training strategy for multilingual translation with LLMs and outperforms the SOTA translation engines, such as NLLB-54.5B and GPT-3.5. They however mainly focus on the supervised English-centric directions, and the translation performance among non-English languages is unclear. Our work follows this line and has an emphasis on the zero-shot translation directions. We investigate the impact of the prompt strategies on the zero-shot translation performance and introduce a simple yet effective cross-lingual regularization constraint, which effectively reduces discrepancies in representations across languages.

7 Conclusion

In this paper, we investigate the impact of prompt strategies on many-to-many multilingual machine translation with LLMs and propose a simple but effective cross-lingual consistency regularization method for multilingual finetuning with translation instructions. Through extensive experiments and visualization analysis, we find that: 1) Different prompt strategies achieve comparable performance in the supervised translation directions, while the zero-shot translation performance varies significantly using different prompt strategies. 2) The zero-shot translation performance could be consistently improved by leveraging cross-lingual consistency regularization during translation instruction finetuning across different prompt strategies. For future work, we will explore the effectiveness of the cross-lingual consistency regularization approach on cross-lingual generalization of LLMs across a wide range of tasks and languages.

Limitations

In this paper, we mainly focus on evaluating our approach on the English-centric scenario and beyond. Future research could consider more multilingual machine translation benchmarks with different numbers of languages and training samples and conduct experiments on more challenging training scenarios such as chain configurations where we have multiple bridge languages and different zero-shot distances.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In Proceedings of the Fifth Conference on Machine Translation, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In Proceedings of the Second Conference on Machine Translation, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012.

Findings of the 2012 workshop on statistical machine translation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023a. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In Findings of the Association for Computational Linguistics: ACL 2023, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023b. Learning multilingual sentence representations with cross-lingual consistency regularization. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 243–262, Singapore. Association for Computational Linguistics.

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

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In Advances in Neural Information Processing Systems, volume 15. MIT Press.

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

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions.

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

OpenAI. 2023a. Chatgpt [large language model]. https://chat.openai.com/.

OpenAI. 2023b. Gpt-4 technical report.

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

Ricardo Rei, 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.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Unbabel Research Team. 2024. Announcing tower: An open multilingual llm for translation-related tasks.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3001–3007, Online. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

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

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

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

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Statistics of All Training Datasets

en ↔	#sentences	en ↔	#sentences
cs	22128	ru	27976
de	26398	zh	28788

Table 10: Statistics of the training datasets. Each entry shows the total number of parallel sentence pairs for both directions. Note that en → and en ← directions have an equal number of sentence pairs.

	#sentences		#sentences
en←de	23071	en→de	27175
en←es	1000	en→es	6708
en←fr	3896	en→fr	7883
en←it	1000	en→it	5070
en←ko	1000	en→ko	2782
en←nl	1000	en→nl	3908
en←pt	1000	en→pt	10905
en←ru	15953	en→ru	22038
en←zh	8518	en→zh	10521
de←fr	2889	de→fr	1622

Table 11: Statistics of the translation pairs in TowerBlocks.

B Details of Evaluation Results on ALMA Models

C Prompt Strategy for GPTs

Prompt
Translate this from [src] into [tgt]. Do not provide any explanations or text apart from the translation. \n[src]: <SRC>\n[tgt]:

Table 16: Prompt strategy adopted when using GPTs, where [src] and [tgt] denote the source and target languages, and <SRC> denotes the source sentence.

D More Evaluation Results on Tower Models

Method	Data	English-centric	Others
TowerInstruct	①	37.0 / 88.6	24.6 / 86.1
T-ENC	②	37.4 / 88.6	24.2 / 85.5
+ XConST	②	37.4 / 88.6	25.1 / 86.0
T-DEC	②	37.6 / 88.6	24.7 / 86.1
+ XConST	②	37.5 / 88.6	25.3 / 86.3
S-ENC-T-ENC	②	37.5 / 88.6	24.5 / 85.9
+ XConST	②	37.5 / 88.7	25.3 / 86.3
S-ENC-T-DEC	②	37.7 / 88.7	24.6 / 86.0
+ XConST	②	37.7 / 88.7	25.6 / 86.5
GPT-MT	②	37.9 / 88.7	24.8 / 86.2
+ XConST	②	37.8 / 88.8	25.6 / 86.5

Table 17: Performance (SacreBLEU / COMET) on FLORES-200 benchmark with LoRA finetuning on TowerBase-7B. ① denotes the TowerBlocks dataset. ② denotes all translation pairs extracted from TowerBlock.

	T-ENC (full-weight)					T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.5/88.7	30.2/85.3	26.3/86.7	37.0/84.9	-	29.3/88.7	29.3/85.1	26.8/86.5	35.5/84.3
cs→	42.2/85.7	-	23.1/85.3	21.1/86.4	29.0/82.8	44.1/86.0	-	24.0/86.0	22.6/88.7	30.2/83.6
de→	29.9/83.9	22.0/90.5	-	21.5/87.0	31.8/84.8	30.1/83.7	22.7/90.7	-	22.0/87.9	31.0/84.3
ru→	37.9/84.7	19.4/89.5	21.0/83.1	-	30.2/83.3	40.3/84.7	20.0/89.7	21.5/83.4	-	29.8/83.2
zh→	21.9/78.7	11.7/84.2	13.9/80.6	10.3/76.4	-	21.1/78.0	10.9/83.0	16.1/82.6	12.2/81.8	-
	T-DEC (full-weight)					T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	28.3/88.7	30.7/85.4	26.3/86.9	37.5/85.2	-	29.9/89.3	29.8/85.5	26.9/86.8	36.3/85.0
cs→	41.7/85.4	-	24.1/86.3	22.0/88.8	30.7/84.4	44.2/86.1	-	24.1/86.3	22.3/88.7	30.0/83.7
de→	29.8/84.0	22.0/90.7	-	21.5/88.1	31.3/84.8	30.3/84.2	22.9/90.6	-	21.9/87.7	31.4/84.6
ru→	37.7/84.7	20.2/89.7	21.7/83.3	-	30.3/83.8	40.2/85.0	20.3/89.7	21.9/83.5	-	29.9/83.2
zh→	21.9/78.9	14.2/88.0	16.1/83.0	14.8/87.0	-	22.7/79.8	13.7/87.4	16.3/83.0	13.7/84.6	-
	S-ENC-T-ENC (full-weight)					S-ENC-T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.4/88.7	30.1/85.3	26.2/86.9	36.4/84.7	-	29.7/89.0	29.6/85.1	26.7/86.8	35.8/84.4
cs→	41.8/85.5	-	6.8/67.1	1.5/48.4	0.7/60.6	44.0/86.0	-	24.2/86.1	20.9/87.3	16.2/75.7
de→	30.2/83.9	3.9/62.3	-	2.0/49.4	1.8/62.0	30.1/84.0	22.3/90.0	-	21.6/87.8	21.9/79.2
ru→	38.4/84.8	1.5/58.1	4.6/63.6	-	0.2/59.5	39.9/84.9	20.1/89.1	21.5/83.0	-	23.8/79.5
zh→	22.2/78.7	1.2/58.9	4.6/66.2	0.6/48.3	-	21.6/78.2	11.9/83.4	15.6/82.3	13.0/83.7	-
	S-ENC-T-DEC (full-weight)					S-ENC-T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.6/88.9	30.6/85.4	26.0/86.9	37.3/85.1	-	29.8/89.2	29.4/85.1	26.8/86.8	36.0/84.8
cs→	42.0/85.7	-	23.9/86.2	22.3/88.9	31.2/84.5	44.2/86.0	-	24.3/86.3	22.4/88.3	30.2/83.7
de→	30.0/84.1	22.3/90.8	-	22.1/88.1	32.0/84.7	30.4/84.1	22.8/90.5	-	21.6/87.4	31.3/84.5
ru→	38.0/84.8	20.1/89.7	22.1/83.3	-	30.4/83.5	40.2/85.0	18.3/87.1	22.4/83.5	-	25.2/79.7
zh→	22.5/79.3	13.6/87.8	16.2/82.9	12.7/81.1	-	23.1/79.7	13.5/87.1	15.7/81.9	12.4/82.4	-
	GPT-MT (full-weight)					GPT-MT (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.8/88.7	30.7/85.4	26.5/86.8	36.9/85.0	-	29.8/89.1	29.9/85.3	26.8/86.9	36.2/84.7
cs→	41.6/85.7	-	23.9/86.1	21.6/87.9	30.7/84.3	43.9/86.0	-	24.6/86.2	22.5/88.8	30.2/83.9
de→	30.0/84.1	22.0/90.7	-	21.8/88.0	31.9/84.9	30.4/84.2	22.0/89.1	-	21.6/87.8	31.3/84.7
ru→	37.8/84.7	20.2/89.7	21.7/83.1	-	30.2/83.6	40.4/85.0	19.7/88.5	22.5/83.5	-	29.5/83.3
zh→	22.4/79.4	13.9/87.8	16.1/82.8	10.6/76.7	-	23.0/79.7	10.7/81.3	15.8/82.8	13.9/85.8	-

Table 12: Performance (SacreBLEU/COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA vanilla finetuning on ALMA-7B-Pretrain.

	T-ENC (full-weight)					T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.6/89.4	31.1/85.5	27.6/87.3	39.2/85.7	-	32.1/89.7	31.3/85.3	28.7/87.5	39.3/85.8
cs→	43.4/86.0	-	24.8/86.1	22.7/88.2	32.3/84.5	47.0/86.6	-	24.5/86.2	22.8/88.5	28.3/83.1
de→	30.9/84.3	23.3/91.2	-	23.2/88.0	33.7/85.4	31.1/84.5	24.2/91.2	-	23.2/87.6	32.1/84.3
ru→	38.8/85.0	21.3/90.1	22.4/83.7	-	31.4/84.0	41.4/85.4	21.4/90.3	23.2/83.9	-	30.6/83.5
zh→	23.8/80.1	15.0/88.3	16.0/82.3	14.6/83.3	-	24.6/80.2	13.5/86.2	16.5/82.6	7.9/73.3	-
	T-DEC (full-weight)					T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.6/89.4	31.6/85.5	27.2/87.4	38.4/85.7	-	32.0/89.8	31.3/85.5	29.1/87.5	39.5/85.7
cs→	43.4/86.1	-	25.2/86.7	11.6/66.2	12.7/69.7	46.7/86.7	-	25.7/86.5	20.6/83.6	32.9/84.9
de→	31.5/84.6	23.0/91.3	-	8.9/61.1	33.6/85.4	31.2/84.6	24.1/91.3	-	21.3/83.9	33.9/85.4
ru→	38.2/85.2	21.2/90.2	23.1/83.9	-	28.8/81.6	40.9/85.4	21.7/90.3	24.0/83.7	-	32.5/84.3
zh→	23.6/80.1	15.0/88.7	17.1/83.5	15.8/87.4	-	24.1/80.0	15.9/89.0	7.0/67.3	12.1/76.4	-
	S-ENC-T-ENC (full-weight)					S-ENC-T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	30.1/89.5	31.2/85.7	27.5/87.4	39.5/85.7	-	32.1/89.7	31.1/85.4	28.8/87.5	39.5/85.9
cs→	43.2/85.9	-	8.6/68.2	2.0/49.0	5.7/65.0	47.0/86.5	-	6.9/66.8	4.4/53.8	24.9/79.6
de→	30.9/84.4	14.0/78.9	-	7.5/60.9	23.4/78.7	31.4/84.6	15.4/79.6	-	4.2/53.7	32.5/84.6
ru→	38.5/85.0	16.0/83.1	17.5/78.1	-	27.0/80.6	41.4/85.4	6.4/67.0	2.6/61.3	-	26.4/80.0
zh→	23.6/80.3	2.4/61.8	2.7/62.9	0.6/48.3	-	24.6/80.1	1.7/60.2	2.3/62.6	0.9/49.3	-
	S-ENC-T-DEC (full-weight)					S-ENC-T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.4/89.2	31.2/85.6	27.3/87.3	38.7/85.6	-	32.5/90.0	31.6/85.6	29.2/87.6	39.7/85.8
cs→	43.2/86.1	-	24.1/85.2	23.3/89.2	4.9/63.8	46.4/86.6	-	6.0/66.0	0.9/47.6	1.2/61.5
de→	31.1/84.4	23.1/91.3	-	23.0/88.5	33.5/85.4	31.3/84.6	4.6/64.1	-	1.0/48.0	31.6/82.6
ru→	38.3/85.1	21.1/90.1	22.6/83.4	-	24.7/77.5	40.8/85.4	3.5/61.6	2.1/60.5	-	7.1/64.5
zh→	23.6/80.1	15.1/88.0	12.9/77.7	15.8/87.6	-	24.6/80.2	1.3/59.5	1.6/61.8	0.7/48.5	-
	GPT-MT (full-weight)					GPT-MT (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.6/89.2	31.5/85.4	27.7/87.3	39.3/85.5	-	32.3/89.7	31.4/85.7	28.9/87.5	39.6/85.6
cs→	42.7/86.0	-	22.7/84.3	1.1/47.4	1.8/61.7	46.3/86.6	-	18.4/78.2	2.0/49.2	23.1/77.3
de→	30.9/84.6	23.4/91.2	-	3.2/50.5	33.9/85.6	31.6/84.7	19.3/83.3	-	1.9/48.7	32.3/84.0
ru→	38.6/85.0	21.2/90.1	22.6/83.3	-	22.7/76.7	40.7/85.3	14.9/79.4	20.0/79.6	-	20.9/74.8
zh→	23.5/80.2	15.4/88.3	13.1/78.2	1.4/50.0	-	24.4/80.0	9.4/75.6	5.0/65.7	0.7/48.6	-

Table 13: Performance (SacreBLEU/COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA vanilla finetuning on ALMA-13B-Pretrain.

	T-ENC (full-weight)					T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.8/88.7	30.6/85.4	26.2/86.9	37.3/85.2	-	29.5/88.8	29.3/85.2	26.6/86.6	35.6/84.5
cs→	41.9/85.5	-	23.7/86.0	22.1/88.6	30.7/84.0	44.3/86.1	-	24.4/86.3	22.5/88.8	29.5/83.5
de→	29.5/84.1	21.9/90.5	-	22.2/88.0	32.0/84.8	29.9/83.7	22.1/90.8	-	21.9/88.0	30.4/84.2
ru→	38.3/84.6	19.7/89.7	21.5/83.2	-	30.5/83.5	40.3/84.7	19.9/89.6	22.4/83.5	-	28.3/83.0
zh→	22.1/79.1	13.6/87.8	16.0/82.8	14.3/86.5	-	21.5/77.8	13.1/87.0	15.4/82.5	14.2/86.1	-
	T-DEC (full-weight)					T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	28.1/88.9	30.5/85.5	26.0/86.8	37.2/85.1	-	29.6/89.3	30.0/85.7	27.0/86.9	36.2/85.0
cs→	41.5/85.4	-	24.1/86.1	22.0/88.8	30.4/84.5	43.9/86.1	-	24.2/86.2	22.8/89.0	29.4/83.7
de→	29.7/84.1	22.0/90.9	-	21.7/88.1	31.4/84.7	30.4/84.2	22.3/90.8	-	21.2/87.9	30.6/84.5
ru→	38.0/84.7	19.9/89.8	21.8/83.4	-	29.9/83.6	40.1/85.0	19.9/89.7	22.3/83.7	-	28.5/82.7
zh→	21.9/79.0	13.7/87.8	16.3/83.0	14.8/86.8	-	22.5/79.6	13.9/88.0	15.6/82.7	14.4/86.7	-
	S-ENC-T-ENC (full-weight)					S-ENC-T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.4/88.7	30.1/85.4	26.1/86.9	36.6/84.9	-	29.0/89.0	29.4/85.4	26.8/86.9	35.4/84.6
cs→	41.8/85.5	-	23.8/86.0	21.5/88.1	29.8/83.8	43.5/85.9	-	23.7/85.8	22.0/88.8	28.1/83.1
de→	29.5/83.8	20.6/90.5	-	21.3/87.6	30.8/84.4	29.5/83.9	22.2/90.7	-	21.1/87.8	29.8/84.0
ru→	38.1/84.7	18.9/89.2	21.9/83.1	-	28.1/82.3	40.1/84.8	20.0/89.6	22.3/83.2	-	27.9/82.5
zh→	21.9/78.7	12.7/87.4	15.8/82.5	14.1/85.2	-	21.0/77.8	12.5/86.3	15.4/82.2	13.5/86.0	-
	S-ENC-T-DEC (full-weight)					S-ENC-T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.5/88.8	30.6/85.5	26.3/87.0	37.1/85.1	-	29.6/89.2	29.4/85.2	26.9/86.9	36.1/84.8
cs→	41.3/85.6	-	24.2/86.2	22.1/88.9	30.5/84.3	44.1/86.1	-	24.5/86.4	22.7/88.9	29.8/83.6
de→	30.0/84.0	22.1/90.9	-	20.1/83.8	31.4/84.4	30.3/84.1	22.8/90.8	-	21.7/87.8	31.2/84.4
ru→	38.1/84.8	20.1/89.7	22.0/83.5	-	30.1/83.5	40.2/85.0	20.5/89.8	22.7/83.5	-	29.6/83.3
zh→	22.2/79.2	14.0/87.8	16.6/82.9	14.8/86.7	-	23.0/79.8	13.4/88.0	15.9/83.0	14.1/86.6	-
	GPT-MT (full-weight)					GPT-MT (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	27.6/88.8	30.6/85.4	26.4/86.9	37.6/85.3	-	29.4/89.1	29.9/85.4	27.0/87.0	36.2/84.7
cs→	41.3/85.6	-	24.2/86.2	22.0/88.9	30.6/84.4	43.9/86.0	-	24.6/86.3	22.7/88.8	30.1/83.8
de→	29.6/84.2	22.3/90.9	-	21.4/88.1	31.9/84.7	30.2/84.2	22.5/90.5	-	21.9/87.9	31.0/84.5
ru→	37.9/84.6	19.7/89.6	21.8/83.3	-	30.0/83.5	40.1/84.9	20.1/89.2	22.5/83.5	-	29.9/83.5
zh→	21.4/78.9	13.7/87.8	16.4/82.9	14.5/86.2	-	22.8/79.6	13.5/87.8	16.2/83.1	14.4/86.7	-

Table 14: Performance (SacreBLEU/COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA cross-lingual consistency finetuning on ALMA-7B-Pretrain.

	T-ENC (full-weight)					T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.6/89.5	31.1/85.6	27.2/87.4	39.6/85.8	-	31.8/89.9	31.3/85.7	28.9/87.7	39.5/85.8
cs→	44.0/86.1	-	25.2/86.5	23.3/89.1	32.6/84.7	46.7/86.6	-	26.3/86.5	24.0/89.3	31.8/84.6
de→	31.1/84.5	22.7/91.3	-	23.2/88.4	33.7/85.4	31.2/84.5	23.1/91.2	-	23.8/88.5	33.1/85.4
ru→	39.0/85.1	20.8/90.1	23.5/83.9	-	31.8/84.0	40.9/85.3	21.2/89.9	24.0/84.0	-	31.1/83.9
zh→	23.8/80.2	15.1/88.5	16.8/83.2	15.5/86.4	-	24.3/80.2	14.1/87.6	17.1/82.7	14.6/84.4	-
	T-DEC (full-weight)					T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.1/89.3	30.9/85.6	27.2/87.4	39.0/85.9	-	31.7/90.0	31.5/85.8	29.0/87.5	39.5/85.9
cs→	43.0/85.9	-	24.4/86.5	23.0/89.1	32.3/84.9	46.3/86.6	-	26.3/86.8	23.5/89.0	32.7/85.0
de→	31.1/84.5	22.7/91.3	-	23.4/88.5	33.3/85.4	31.4/84.6	24.2/91.4	-	23.5/88.4	33.9/85.5
ru→	38.5/85.0	20.7/90.2	24.1/84.1	-	32.1/84.3	40.8/85.4	21.9/90.2	24.3/84.3	-	31.9/83.9
zh→	23.4/80.0	15.2/88.6	16.8/83.5	16.2/87.6	-	24.2/80.2	15.5/88.9	17.7/83.6	15.9/87.1	-
	S-ENC-T-ENC (full-weight)					S-ENC-T-ENC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.5/89.4	31.2/85.7	27.3/87.4	39.3/86.0	-	31.8/90.0	31.0/85.6	28.6/87.7	39.5/85.8
cs→	43.7/86.0	-	25.2/86.2	22.9/89.0	32.0/84.9	46.5/86.5	-	26.1/86.7	23.8/89.3	32.2/84.8
de→	31.4/84.5	22.9/91.2	-	23.2/88.3	33.0/85.2	31.4/84.6	24.0/91.4	-	23.9/88.5	34.2/85.5
ru→	38.9/85.1	20.5/90.0	23.3/83.4	-	32.0/84.2	41.1/85.3	21.8/90.0	24.0/84.0	-	32.3/84.1
zh→	23.8/80.2	15.3/88.5	16.3/82.9	15.8/87.1	-	24.5/80.0	13.5/85.9	16.8/82.9	15.3/85.4	-
	S-ENC-T-DEC (full-weight)					S-ENC-T-DEC (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	28.9/89.3	31.2/85.7	27.3/87.4	39.1/85.9	-	31.3/90.0	31.7/86.0	29.1/87.8	39.4/85.9
cs→	42.6/85.9	-	26.1/86.7	23.2/89.3	32.5/85.0	45.6/86.6	-	26.3/86.8	23.3/89.2	31.9/84.8
de→	30.5/84.4	22.2/91.2	-	23.3/88.6	33.5/85.4	31.2/84.6	23.8/91.4	-	23.5/88.6	33.5/85.4
ru→	37.9/84.9	20.9/90.3	24.5/84.2	-	31.9/84.4	40.7/85.2	21.6/90.0	24.4/84.5	-	31.5/84.1
zh→	23.6/80.1	15.3/88.7	17.6/83.9	16.5/87.8	-	23.9/80.0	15.4/88.6	16.4/81.3	15.8/86.1	-
	GPT-MT (full-weight)					GPT-MT (LoRA)				
	en	cs	de	ru	zh	en	cs	de	ru	zh
en→	-	29.2/89.2	31.4/85.7	27.6/87.4	39.2/85.9	-	31.8/89.8	31.5/85.7	28.9/87.5	39.6/85.6
cs→	42.6/85.8	-	25.4/86.6	23.2/89.1	32.1/84.9	45.8/86.6	-	26.1/86.5	23.4/88.8	31.9/84.8
de→	30.8/84.5	22.1/91.1	-	22.9/88.4	33.6/85.6	31.4/84.7	23.8/91.4	-	23.4/88.2	33.2/85.4
ru→	38.4/85.0	20.3/90.1	24.2/84.0	-	31.5/84.3	40.5/85.3	21.6/89.8	24.1/84.4	-	31.5/84.0
zh→	23.6/80.0	15.4/88.7	17.5/83.6	16.3/87.6	-	23.9/80.1	15.0/88.2	16.9/82.6	15.6/86.2	-

Table 15: Performance (SacreBLEU/COMET) on WMT22 and FLORES-200 benchmarks with full-weight and LoRA cross-lingual consistency finetuning on ALMA-13B-Pretrain.