Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Building accurate translation-tailored large language models with language-aware instruction tuning*

Changtong ZAN[1], Liang DING[†‡2], Li SHEN[3,4], Yibing ZHAN[3], Xinghao YANG[1], Weifeng LIU[†‡1]

[1]*College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China*

[2]*School of Computer Science, University of Sydney, New South Wales 2006, Australia*

[3]*JD Explore Academy, JD.com Inc., Beijing 100101, China*

[4]*School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China*

[†]E-mail: liangding.liam@gmail.com; liuwf@upc.edu.cn

Received May 30, 2024; Revision accepted Nov. 27, 2024; Crosschecked July 10, 2025

Abstract: Large language models (LLMs) exhibit remarkable capabilities in various natural language processing tasks, such as machine translation. However, the large number of LLM parameters incurs significant costs during inference. Previous studies have attempted to train translation-tailored LLMs with moderately sized models by fine-tuning them on the translation data. Nevertheless, when performing translations in zero-shot directions that are absent from the fine-tuning data, the problem of ignoring instructions and thus producing translations in the wrong language (i.e., the off-target translation issue) remains unresolved. In this work, we design a two-stage fine-tuning algorithm to improve the instruction-following ability of translation-tailored LLMs, particularly for maintaining accurate translation directions. We first fine-tune LLMs on the translation data to elicit basic translation capabilities. At the second stage, we construct instruction-conflicting samples by randomly replacing the instructions with the incorrect ones. Then, we introduce an extra unlikelihood loss to reduce the probability assigned to those samples. Experiments on two benchmarks using the LLaMA 2 and LLaMA 3 models, spanning 16 zero-shot directions, demonstrate that, compared to the competitive baseline—translation-finetuned LLaMA, our method could effectively reduce the off-target translation ratio (up to −62.4 percentage points), thus improving translation quality (up to +9.7 bilingual evaluation understudy). Analysis shows that our method can preserve the model's performance on other tasks, such as supervised translation and general tasks. Code is released at https://github.com/alphadl/LanguageAware_Tuning.

Key words: Zero-shot machine translation; Off-target issue; Large language model; Language-aware instruction tuning; Instruction-conflicting sample

https://doi.org/10.1631/FITEE.2400458 **CLC number:** TP391

1 Introduction

Large language models (LLMs) (Min et al., 2023; Li B et al., 2024) have demonstrated excellent performance on a wide range of natural language processing (NLP) tasks, including reasoning (Wei et al., 2023), summarization (Wang YM et al., 2023; Huang et al., 2024), translation (Hendy et al., 2023), understanding (Zhang HB et al., 2022; Zhong et al., 2023), and evaluation (Lu et al., 2024). LLMs exemplified by GPT-3 (Brown et al., 2020), OPT (Zhang SS et al., 2022), LLaMA (Touvron et al., 2023a), and LLaMA 2 (Touvron et al., 2023b), leverage large-scale monolingual data through pre-training with the causal language modeling task, and exhibit strong

‡ Corresponding authors

* Project supported by the National Natural Science Foundation of China (No. 62372468), the Shandong Natural Science Foundation (No. ZR2023MF008), the Major Basic Research Projects in Shandong Province (No. ZR2023ZD32), and the Qingdao Natural Science Foundation (No. 23-2-1-161-zyyd-jch)

ORCID: Changtong ZAN, https://orcid.org/0000-0002-5467-0937; Liang DING, https://orcid.org/0000-0001-8976-2084; Weifeng LIU, https://orcid.org/0000-0002-5388-9080

zero-shot capabilities with a few demonstration examples. Instruction tuning (Mishra et al., 2022; Wei et al., 2022) further elicits the capacity of LLMs to address general tasks directly with proper guidance, such as task definition. However, due to the significant cost to call state-of-the-art proprietary LLMs, such as GPT-4 (OpenAI, 2024), it is attractive to explore strategies for effectively fitting suitably sized LLMs into specific tasks, such as machine translation (Fu et al., 2023; Xu HR et al., 2024b).

In zero-shot translation (ZST) (Gu et al., 2019; Chen et al., 2023; Zan et al., 2023), the objective is to translate sentences from a source language to a target language, where there is either a lack of direct mapping between the source and target languages in the training data, or the target or source languages are absent during training. Addressing the ZST problem is both vital and challenging, especially for low-resource languages. Recent research demonstrates that building translation-tailored LLMs by fine-tuning translation data can achieve superior translation performance (Liu YJ et al., 2023; Xu HR et al., 2024b; Zeng et al., 2024). However, as illustrated in Fig. 1, our preliminary study shows that when tackling zero-shot directions, translation-tailored LLMs often encounter the off-target translation problem, where the generated translations are in the wrong languages. For example, in De→Fr, the off-target ratio reaches up to 95.0%. We attribute this problem to the fact that pre-training LLMs in the fashion of predicting the next token may lead to overlooking the key information contained in instructions.

Previous studies (Peng et al., 2023; Xu HR et al., 2024b) indicate that introducing more informative prompts during inference, such as preemptively translating prompts into the target language or incorporating few-shot demonstration samples, can be beneficial. Sennrich et al. (2024) modified the decoding process by introducing language-contrastive samples to constrain the decoding process, thus alleviating the off-target problem. Different from the above approaches that focus on the inference stage, our motivation is to fundamentally improve the instruction-following ability (especially the awareness of translation direction) of LLMs.

In this paper, we introduce a simple and effective two-stage fine-tuning algorithm to enhance the effect of instructions in translation-tailored LLMs.

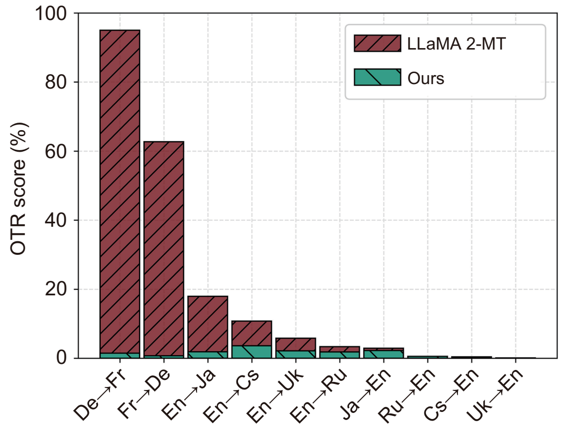


Fig. 1 OTR scores in ZST of WMT benchmark. We present a comparison between LLaMA 2-MT, the LLaMA 2 model fine-tuned on machine translation data, and our method. De: German; Fr: French; En: English; Ja: Japanese; Cs: Czech; Uk: Ukrainian; Ru: Russian

This is accomplished by introducing unlikelihood loss (UL) on instruction-conflicting samples, in which the translation sentence pairs deviate from the specified tasks associated with the given instructions. Initially, we fine-tune the LLMs using a multilingual translation dataset, unlocking the inherent translation capabilities of LLMs. Subsequently, we build upon the pre-tuned model by incorporating translation data along with instruction-conflicting samples. We create instruction-conflicting samples by randomly replacing the task instruction with an incorrect one. These data are used to fine-tune the model, leveraging the unlikelihood training paradigm. Our approach can be viewed as emphasizing the effect of instructions, thereby guiding the model to produce translation in the correct language.

We apply our method in the experiments to fine-tune the currently competitive open-source LLaMA 2 and LLaMA 3 models. Compared with fine-tuning on translation data, the results reveal substantial reductions in the off-target translation ratio (OTR), with 62.3% and 30.5% for the IWSLT benchmark and 29.9% and 18.6% for the WMT benchmark for LLaMA 2 and LLaMA 3 models, respectively. This leads to notable enhancements in the translation quality, as evidenced by the increases of average +9.7/+6.2 and +6.2/+3.9 bilingual evaluation understudy (BLEU) score on the IWSLT and WMT datasets, respectively. Moreover, our method maintains the capability to perform other tasks, such as supervised directions and general tasks. The main contributions are as follows:

1. We reveal the heavy off-target problem of translation-tailored LLMs in ZST settings, and we attribute this problem to the weak instruction (translation direction) following ability.

2. To fundamentally improve the translation direction-following ability, we introduce a two-stage fine-tuning algorithm for LLMs that leverages instruction-conflicting samples.

3. Extensive experiments illustrate the effectiveness of our approach in mitigating the off-target translation problem and producing better translations. Analysis shows that our method will not affect the ability of other tasks, e.g., the general task performance on AlpacaEval and the supervised translation performance.

2 Preliminary

1. Instruction tuning

Instruction tuning aims to refine LLMs by fine-tuning a diverse collection of data characterized by explicit instructions. This refinement process significantly enhances zero-shot performance on previously unseen tasks (Wei et al., 2022). Each instance in the instruction tuning dataset comprises three fundamental components: (1) instruction, a textual representation that describes NLP tasks in natural language; (2) input (optional), supplementary contextual information that provides additional context for the given task; (3) output, the expected response that LLMs should generate. During the tuning process, the model is trained using a teacher-forcing approach (Cho et al., 2014). It models the distribution of output tokens conditioned on the instruction and optionally the input. This training methodology empowers the model to understand and follow instructions effectively. Subsequently, the instruction-tuned model is capable of directly performing unseen tasks by following the appropriate task instructions in a zero-shot manner. In this study, our primary focus is translation-tailored LLMs, where we fine-tune LLMs on multilingual translation data.

2. Unlikelihood training

Welleck et al. (2019) explored a novel approach that encourages the model to assign lower probabilities to improbable generations, in contrast to the traditional likelihood training which focuses on the overall probability distribution of correct sequences. The general training framework comprises two types of updates:

(1) Likelihood update for maximizing the probability of the ground-truth sequence with likelihood loss:

$$\mathcal{L}^{\mathrm{MLE}} = -\sum_{t_n} \ln P(t_n \mid t_0, t_1, \cdots, t_{n-1}),$$

where t_n is the ground-truth token at position n, and MLE is the maximum likelihood estimation.

(2) Unlikelihood update for minimizing the probability of negative samples (tokens that should not occur) with UL:

$$\mathcal{L}^{\mathrm{UL}} = -\sum_{t_n} \sum_{k\in\mathcal{C}_n} \ln\left(1 - P(k \mid t_0, t_1, \cdots, t_{n-1})\right),$$

where $\mathcal{C}_n$ is the set of negative candidate tokens at position n.

We extend this approach to the domain of ZST based on translation-tailored LLMs. We introduce instruction-conflicting samples for unlikelihood updates, emphasizing the impact of translation instructions (especially the translation direction and language) and addressing off-target problems.

3 Methodology

3.1 Pre-tuning on multilingual translation samples

To unlock the translation capabilities of LLMs, we employ a pre-tuning stage using multilingual translation examples. Given a collection of instruction samples $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_N\}$ covering N language pairs, where $\mathcal{D}_i$ denotes a translation parallel corpus of the i^{th} language pair ($i = 1, 2, ..., N$). For each sample $\mathcal{S}_j^i = \left\{\mathbf{ins}_j^i, \boldsymbol{x}_j^i, \boldsymbol{y}_j^i\right\}$ in $\mathcal{D}_i$, the model is trained with the common approach, i.e., MLE. As depicted in Fig. 2, the instruction **ins** for the multilingual translation sample is "Translate the following sentences from Romanian to English," and the corresponding input source sentence $\boldsymbol{x}$ is "Hai să fim cinstiţi: mă autointitulez titan." We train the model to model the mapping from both **ins** and $\boldsymbol{x}$ to the expected translation output $\boldsymbol{y}$; i.e., "I mean, let's be honest, I call myself a titan." Formally, the training loss for each sample $\mathcal{S}_j^i$ is as follows:

$$\mathcal{L}_{\mathcal{S}_j^i}^{\mathrm{MLE}}(\theta) = -\sum_{t_n} \ln P(t_n|\mathbf{ins}_j^i, \boldsymbol{x}_j^i, t_0, t_1, \cdots, t_{n-1}; \theta),$$

where t_n is the n^{th} token of the translation sequence $\boldsymbol{y}$; for instance, the first token of $\boldsymbol{y}$ in the above example is "I." θ represents the trainable model parameters. Minimizing the loss of each token in the translation output unlocks the ability of LLMs to perform translation tasks by following the provided instructions. Then, the final objective of our first fine-tuning stage, pre-tuning on the multilingual translation dataset, can be formulated as follows:

$$\mathcal{L}_{\mathcal{D}}^{\text{MLE}}(\theta) = -\sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{\mathcal{S}_j^i \in \mathcal{D}_i} \mathcal{L}_{\mathcal{S}_j^i}^{\text{MLE}}(\theta).$$

It is worth noting that the instruction **ins** is not limited to translation task definition. Samples from diverse NLP tasks can be formulated into this unified format with proper instruction.

3.2 Unlikelihood training with instruction-conflicting samples

At the second stage, our objective is to improve the LLMs' instruction-following ability and enhance the ZST capability. We employ a dual optimization approach which involves training the model with likelihood loss on multilingual translation samples and UL on samples with conflicting instructions.

1. Instruction-conflicting samples

Although LLMs can achieve impressive performance on tasks within the multilingual translation tuning dataset, they may encounter issues such as ignoring instructions and generating translations in a wrong language during ZST, commonly referred to the off-target problem. To address the off-target issue with unlikelihood training, we define the negative candidate samples by substituting the instruction with a different one while keeping the input and output unchanged. We call this type of samples (instruction-conflicting samples) as the translation pairs deviating from the task associated with the given instructions. As shown in Fig. 2, given a multilingual translation sample $\mathcal{S}_j^i = \left\{\mathbf{ins}_j^i, \boldsymbol{x}_j^i, \boldsymbol{y}_j^i\right\}$ from the instruction tuning dataset $\mathcal{D}$, we randomly select a sample with different instruction $\widetilde{\mathbf{ins}}_j^i$; e.g., "Translate the following sentences from English to Italian." Then, we replace the original correct $\mathbf{ins}_j^i$ to obtain the instruction-conflicting sample $\widetilde{\mathcal{S}}_j^i = \left\{\widetilde{\mathbf{ins}}_j^i, \boldsymbol{x}_j^i, \boldsymbol{y}_j^i\right\}$; e.g., "[Instruction]: Translate the following sentences from English to Italian; [Input]: Hai să fim cinstiţi: mă autointitulez titan; [Output]: I mean, let's be honest, I call myself a titan." in the example. The instruction-conflicting samples contain incorrect outputs, and do not align with the task defined by the instruction.

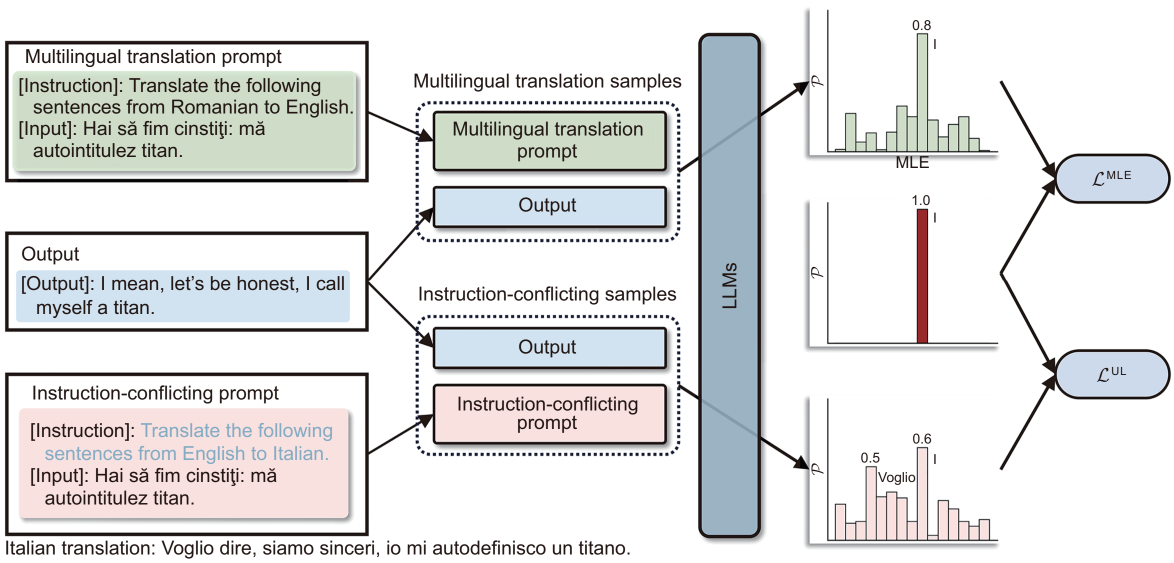


Fig. 2 Overview of our fine-tuning framework for ZST. At the first stage, we pre-tune LLMs on multilingual translation samples, focusing on unlocking the translation ability of LLMs. Subsequently, we introduce instruction-conflicting samples by randomly substituting the instruction component with a different one. We then train the model with $\mathcal{L}^{\text{MLE}}$ on translation data and incorporate a UL $\mathcal{L}^{\text{UL}}$ on the instruction-conflicting samples to assign lower probabilities to wrong language tokens

2. Unlikelihood training with instruction-conflicting samples

Unlikelihood training aims to reduce the probability assigned by the model to negative candidate tokens. Based on the previously defined instruction-conflicting samples, we can extend the unlikelihood training to translation-tailored LLMs, and enhance their ability to follow instructions for translation tasks.

In each update, we optimize the UL for instruction-conflicting samples $\widetilde{\mathcal{S}}_j^i$:

$$\begin{aligned}&\mathcal{L}^{\mathrm{UL}}_{\widetilde{\mathcal{S}}_j^i}(\theta)\\ &=-\sum_{t_n}\ln\left(1-P(t_n|\widetilde{\mathbf{ins}}_j^i,\boldsymbol{x}_j^i,t_0,t_1,\cdots,t_{n-1};\theta)\right),\end{aligned}$$

where t_n represents the n^{th} token of the output. Due to the relatively small difference between **ins** and $\widetilde{\mathbf{ins}}$, such as only two different words in the example, the model pre-tuned on multilingual translation samples usually tends to assign high probability on the output. As shown in Fig. 2, the first token "I" has a high probability, and the model is more confident to generate "I" rather than the correct word "Voglio." The objective of UL on the whole dataset is as follows:

$$\mathcal{L}^{\mathrm{UL}}_{\mathcal{D}}(\theta)=-\sum_{\mathcal{D}_i\in\mathcal{D}}\sum_{\widetilde{\mathcal{S}}_j^i\in\mathcal{D}_i}\mathcal{L}^{\mathrm{UL}}_{\widetilde{\mathcal{S}}_j^i}(\theta).$$

To prevent the potential overfitting on the unlikelihood objective while maintaining the supervised translation ability, we incorporate multilingual translation samples to simultaneously train the model with likelihood loss. The overall objective function in unlikelihood training involves a mixture of likelihood loss and UL, defined as

$$\mathcal{L}_{\mathcal{D}}(\theta)=\mathcal{L}^{\mathrm{MLE}}_{\mathcal{D}}(\theta)+\alpha\mathcal{L}^{\mathrm{UL}}_{\mathcal{D}}(\theta),$$

where α is the mixing hyper-parameter.

4 Experiments and results

In this section, we conduct a series of experiments spanning 16 ZST directions to assess the effectiveness of our algorithm.

4.1 Experimental setup

1. Datasets

We consider the following two widely used datasets:

(1) WMT. Following Jiao et al. (2023) and Liu YJ et al. (2023), we use the development sets from WMT2017 to WMT2020 for instruction tuning. This includes four language directions: En↔Zh (Zh: Chinese) and En↔De, encompassing a total of 51 000 translation sentence pairs. Then, we assess the translation performance on WMT22 test sets, including En↔Cs, En↔Ja, En↔Ru, En↔Uk, and Fr↔De. All these translation language pairs do not exist in the training sets, thus allowing for the evaluation of ZST performance.

(2) IWSLT. Following Qu and Watanabe (2022), we use the IWSLT-17 dataset to evaluate the performance of the models. We consider the four languages (En, Ro, It, and Nl) from MMCR4NLP (Dabre and Kurohashi, 2019). For training, we randomly select 12000 sentence pairs from the training sets, spanning six directions: En↔Nl, En↔It, and En↔Ro (Nl: Dutch; It: Italian; Ro: Romanian). The evaluation is conducted on the test set of IWSLT-17, including Ro↔Nl, Ro↔It, and Nl↔It. All these translation language pairs do not exist in the training sets.

2. Base models

We employ both the 7B size LLaMA 2 (Touvron et al., 2023a) and the 8B size LLaMA 3 (https://llama.meta.com/llama3/) as the base models. LLaMA 2 is a robust language model that has undergone training on 2 trillion tokens, with less than 2% non-English data. In contrast, LLaMA 3 represents a significant advancement, having been pre-trained on more than 6× pre-training data used for LLaMA 2, specifically 14 trillion tokens. This consists of more multilingual data (over 5% non-English data). Moreover, LLaMA 3 has a larger vocabulary than LLaMA 2. These setups ensure that LLaMA 3 models have better capability to process multilingual tasks.

3. Baselines

We consider the following baselines:

(1) LLaMA. We employ the pre-trained LLMs directly for inference. The prompt is the same as those used by the below machine translation (MT).

(2) MT. Following previous works (Jiao et al., 2023; Xu HR et al., 2024b; Zeng et al., 2024), we fine-tune the LLMs on multilingual translation samples, establishing this configuration as our main baseline. We format translation sentence pairs into a unified translation template.

(3) Post-instruction (Post-ins). Following

Liu YJ et al. (2023), we switch the positions of instruction and input, where the model pays more attention to the instruction. Other settings remain consistent with MT.

(4) Prompt in target language (PTL). Instead of using the English prompt during inference, we translate the prompt into the target language during inference, which could potentially provide more guidance information for generating target language words. This setup uses the MT models for inference.

(5) K-shot. In-context learning (Brown et al., 2020) has proven to be an effective way to improve the performance of LLMs. We report the few-shot performance for comprehensive comparison, including 1-shot and 5-shot. MT models are used for inference.

(6) $\mathcal{C}_{\text{src+lang}}$. Following Sennrich et al. (2024), we employ the decoding method by contrasting the translation sentence with language-contrastive input and $\lambda_{\text{lang}} = 0.5$ (λ_{lang} is the hyper-parameter involved in this comparison method). The inference relies on MT models and uses a greedy decoding strategy.

4. Model training

We conduct experiments on the Hugging Face Transformers (Wolf et al., 2020) toolkit. During the pre-tuning phase, we set the learning rate to 2e−5, warmup ratio to 0.03, and batch size to 128. For the IWSLT benchmark, we perform training over 3 epochs, while for the WMT dataset, training is conducted for 1 epoch. During the second stage of training, we set the mixing parameter α to 0.05, learning rate to $2e^{-6}$, batch size to 8, and the number of training steps to 100. We use the final model for evaluation.

5. Evaluation

We adopt SacreBLEU (Post, 2018) to evaluate the translation accuracy, where translations are generated with a beam size of 4. Besides, we compute the ratio of wrong language translation in the generated outputs, i.e., OTR with a publicly available language detector (https://fasttext.cc/docs/en/language-identification.html) (Joulin et al., 2016a, 2016b). We use vLLM (Kwon et al., 2023) to accelerate during inference. Unless otherwise stated, we report results in the ZST direction.

4.2 Main results

We compare the ZST performance of our method and other baseline methods on the WMT and IWSLT benchmarks, as depicted in Tables 1 and 2. Obviously, our method generates more translations in the expected languages with the lowest average OTR scores across 16 directions for both base models. What is especially notable is the performance of our method with the LLaMA 2 base model, which demonstrates a significant average OTR gain of −62.4 percentage points (PPs) on the IWSLT benchmark compared with MT. We attribute this improvement to the better translation direction instruction following the ability of our method. The more the model tends to omit the instruction of language, the greater the benefit from our method. Furthermore, our method outperforms baseline approaches that mitigate off-target problems during inference, such as PTL, K-shot, and $\mathcal{C}_{\text{lang}}$. Our method achieves improvements of up to +14.2/+14.3 average BLEU scores and −21.0/−63.3 PPs average OTR scores on the WMT/IWSLT benchmarks compared with these methods, respectively. Regarding the baseline adjustments during the tuning stage, our method achieves improvements over Post-ins, up to +1.8/+6.9 average BLEU and −9.4/−55.6 PPs average OTR on the WMT/IWSLT benchmarks, respectively. On IWSLT, we also observe that our method shows a slightly lower average BLEU score than that of Post-ins (15.9 vs. 16.1) when LLaMA 3 is used as the base model. Additionally, our method surpasses other robust baseline models in these evaluations.

5 Analysis

To provide a deeper insight into the proposed algorithm, we investigate the following issues: (1) Can increasing the number of training steps lead to greater benefits? (2) What is the impact of the mixing hyper-parameter? (3) How does the size of the translation data in pre-tuning affect the outcome? (4) What role does the model size play? (5) Can continuing training with instruction-conflicting samples maintain supervised translation performance? (6) What is the source of the observed improvements? (7) If we introduce additional general task tuning data, can the model still retain its general task ability?

Table 1 ZST performance on the WMT benchmark

Base model	Method	BLEU score↑										Avg
		Cs←En	Cs→En	Ja←En	Ja→En	Ru←En	Ru→En	Uk←En	Uk→En	Fr←De	Fr→De	
LLaMA 2	LLaMA	0.2	1.3	0.1	0.4	0.3	1.3	0.2	1.9	0.8	0.6	0.7
	MT	18.2	**39.0**	9.6	16.1	21.2	36.4	9.6	**34.3**	4.3	3.2	19.2
	Post-ins	18.8	38.0	12.4	15.8	**22.1**	**36.5**	14.6	34.1	**30.7**	5.3	22.8
	PTL	17.6	**39.0**	10.6	16.1	20.0	36.4	17.8	**34.3**	24.7	**24.6**	24.1
	1-shot	18.8	37.2	11.4	15.5	20.9	34.9	17.7	32.9	3.9	3.2	19.6
	5-shot	18.3	37.0	12.2	15.1	20.9	34.2	**18.4**	31.8	3.7	3.2	19.5
	$\mathcal{C}_{\text{src+lang}}$	3.6	35.5	2.6	13.1	3.2	33.9	2.1	31.8	1.4	0.7	12.8
	Ours	**18.8**	38.9	**12.8**	**16.3**	20.9	35.8	18.0	32.6	29.8	19.4	**24.3**
Base model	Method	OTR score (%)↓										Avg
		Cs←En	Cs→En	Ja←En	Ja→En	Ru←En	Ru→En	Uk←En	Uk→En	Fr←De	Fr→De	
LLaMA 2	LLaMA	90.0	24.2	98.7	16.7	86.6	20.0	90.8	14.9	84.0	85.6	61.2
	MT	10.7	**0.3**	27.9	2.1	6.7	**0.3**	60.4	0.1	90.8	99.3	29.9
	Post-ins	6.3	1.9	9.7	2.7	2.5	0.4	25.1	**0.0**	**4.0**	87.4	14.0
	PTL	18.5	**0.3**	16.0	2.1	13.1	**0.3**	6.2	0.1	19.7	**11.4**	8.8
	1-shot	10.1	0.4	12.6	4.0	6.5	0.5	7.6	0.1	97.7	99.4	23.9
	5-shot	13.8	0.6	11.8	4.9	8.3	0.4	7.1	0.1	98.9	99.6	24.6
	$\mathcal{C}_{\text{src+lang}}$	**3.1**	**0.3**	19.2	**1.4**	**2.4**	0.5	12.5	**0.0**	87.5	97.9	22.5
	Ours	6.6	**0.3**	**2.3**	1.9	2.6	**0.3**	**3.2**	0.1	6.7	31.7	**5.6**
Base model	Method	BLEU score↑										Avg
		Cs←En	Cs→En	Ja←En	Ja→En	Ru←En	Ru→En	Uk←En	Uk→En	Fr←De	Fr→De	
LLaMA 3	LLaMA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MT	20.6	**39.7**	11.5	17.3	**25.2**	**35.9**	19.7	**34.2**	15.3	4.8	22.4
	Post-ins	21.7	**39.7**	12.0	16.4	24.9	**35.9**	18.6	34.1	**33.4**	11.8	24.9
	PTL	20.6	**39.7**	5.8	17.3	24.6	**35.9**	19.6	**34.2**	15.4	10.1	22.3
	1-shot	21.5	38.4	11.7	15.4	24.6	34.0	19.2	32.7	19.1	5.6	22.2
	5-shot	22.0	38.0	13.1	14.9	24.1	33.7	**19.8**	32.4	24.6	8.4	23.1
	$\mathcal{C}_{\text{src+lang}}$	3.0	35.1	2.2	14.5	2.6	31.6	3.0	29.4	1.4	1.7	12.5
	Ours	**23.5**	39.6	**15.6**	**17.7**	24.9	34.8	19.6	33.3	31.0	**27.1**	**26.7**
Base model	Method	OTR score (%)↓										Avg
		Cs←En	Cs→En	Ja←En	Ja→En	Ru←En	Ru→En	Uk←En	Uk→En	Fr←De	Fr→De	
LLaMA 3	LLaMA	100.0	0.1	100.0	0.2	100.0	0.0	100.0	0.0	99.2	99.3	59.9
	MT	10.8	0.4	17.9	2.9	3.3	0.5	5.8	**0.0**	62.7	95.0	19.9
	Post-ins	8.8	0.3	20.6	3.4	2.8	0.5	8.8	**0.0**	1.3	62.7	10.9
	PTL	12.8	0.4	58.7	2.9	4.0	0.5	5.8	**0.0**	60.2	79.5	22.5
	1-shot	10.8	0.4	10.6	3.1	3.5	0.5	4.8	0.1	48.6	92.6	17.5
	5-shot	11.1	0.8	10.2	3.3	4.2	0.5	6.3	0.2	27.6	79.9	14.4
	$\mathcal{C}_{\text{src+lang}}$	**2.2**	**0.1**	8.1	**1.8**	**0.5**	**0.3**	**0.8**	**0.0**	24.4	71.3	11.0
	Ours	3.6	0.3	**1.9**	2.2	1.8	0.5	2.2	0.1	**0.7**	**1.5**	**1.5**

The best results are in bold, except for the OTR scores of LLaMA. ↑: the higher the better; ↓: the lower the better; Avg: average score obtained for all directions

5.1 Effect of the number of unlikelihood training steps

To provide insight into the impact of the number of unlikelihood training steps at our second fine-tuning stage, Fig. 3a presents the ZST performance of our LLaMA 2-based method on the IWSLT benchmark. As observed, the method produces fewer wrong language translations and higher quality translations with more unlikelihood training steps. From the figure, it can be seen that the method achieves its best performance, denoted by near-zero OTR score, after about 40 updates, and this performance is consistently maintained even with further training extending up to 100 steps. This shows that our method can improve the ZST performance without a significant number of updates.

5.2 Effect of α

As mentioned in Section 3.2, our algorithm incorporates a mixing hyper-parameter α to balance the MLE loss and UL. This is an ablation to evaluate the effect of different α values. Fig. 3b shows the ablation results of our LLaMA 2-based methods on the IWSLT benchmark. As expected, a higher α places greater emphasis on the UL, resulting in fewer wrong language translations. Our method fine-tuned with α exceeding 0.02 is unlikely to produce translations in the wrong languages. However, when α increases, there is a slight decrease in the translation quality as reflected by the BLEU score. This decline may be attributed to the potential overfitting on the UL. Future research efforts could aim to alleviate the effects of this potential overfitting issue. In

Table 2 ZST performance on the IWSLT benchmark

Base model	Method	BLEU score↑						Avg
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 2	LLaMA	0.9	0.5	0.5	1.0	0.3	0.7	0.7
	MT	7.9	8.3	2.3	4.8	2.8	4.2	5.1
	Post-ins	9.0	11.2	5.6	7.7	8.1	5.1	7.8
	PTL	10.2	9.4	6.8	7.8	5.3	6.6	7.7
	1-shot	11.5	10.8	3.0	8.5	3.0	6.9	7.3
	5-shot	8.5	9.4	1.7	6.8	1.4	4.5	5.4
	$\mathcal{C}_{\text{src+lang}}$	2.3	2.0	0.7	1.7	0.7	1.8	1.5
	Ours	**17.5**	**16.2**	**15.4**	**12.8**	**12.0**	**14.5**	**14.7**
Base model	**Method**	**OTR score (%)↓**						**Avg**
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 2	LLaMA	86.2	85.5	91.3	84.3	94.7	88.7	88.5
	MT	49.8	39.8	85.8	65.7	80.1	68.8	65.0
	Post-ins	56.8	32.4	71.3	64.7	46.2	77.7	58.2
	PTL	34.4	34.2	49.7	50.2	60.0	52.3	46.8
	1-shot	32.0	26.7	82.2	47.1	81.8	50.1	53.3
	5-shot	46.2	34.4	95.1	57.5	92.3	69.8	65.9
	$\mathcal{C}_{\text{src+lang}}$	31.8	46.0	85.4	65.4	82.8	56.0	61.2
	Ours	**2.7**	**1.5**	**3.8**	**1.5**	**3.8**	**2.5**	**2.6**
Base model	**Method**	**BLEU score↑**						**Avg**
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 3	LLaMA	0.0	0.0	0.0	0.0	0.1	0.0	0.0
	MT	12.0	15.1	3.5	15.9	3.1	8.2	9.6
	Post-ins	17.7	**17.9**	13.5	**19.9**	**12.1**	**15.6**	**16.1**
	PTL	14.5	17.3	8.7	18.4	8.9	10.4	13.0
	1-shot	15.4	17.3	4.0	19.3	8.3	12.6	12.8
	5-shot	15.0	17.8	1.3	19.3	9.6	14.3	12.9
	$\mathcal{C}_{\text{src+lang}}$	1.3	1.5	1.2	2.3	1.1	2.5	1.7
	Ours	**18.2**	**17.9**	**15.6**	18.4	10.5	14.7	15.9
Base model	**Method**	**OTR score (%)↓**						**Avg**
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 3	LLaMA	99.0	99.2	98.9	98.1	99.9	99.7	99.1
	MT	27.4	4.5	80.7	5.7	53.3	30.4	33.7
	Post-ins	5.1	3.6	13.2	4.0	11.7	5.4	7.2
	PTL	19.4	7.4	45.3	5.9	32.1	25.6	22.6
	1-shot	16.3	4.9	79.4	5.7	34.1	19.9	26.7
	5-shot	16.8	3.5	97.5	3.5	32.5	13.9	28.0
	$\mathcal{C}_{\text{src+lang}}$	21.7	2.4	49.9	4.7	34.2	25.6	23.1
	Ours	**2.5**	**1.5**	**4.8**	**1.7**	**6.4**	**2.1**	**3.2**

The best results are in bold. ↑: the higher the better; ↓: the lower the better; Avg: average score obtained for all directions

summary, our experimental results indicate that our method exhibits robustness to varying values of α.

5.3 Results with different sizes of LLMs

To examine the influence of model size, we fine-tune the 13B size LLaMA 2 on the IWSLT benchmark, employing the same experimental setup as the 7B size model. We report the results of MT and our method in Table 3. The 13B model consistently outperforms the 7B model in terms of both reducing the wrong language translation ratio (−7.1 PPs average OTR score) and improving translation quality (+2.7 average BLEU score). This finding is consistent with prior research (Kaplan et al., 2020), which suggests that increasing the number of training parameters yields benefits. However, the off-target problem still exists in the 13B size LLaMA 2-MT method. Additionally, our method achieves a

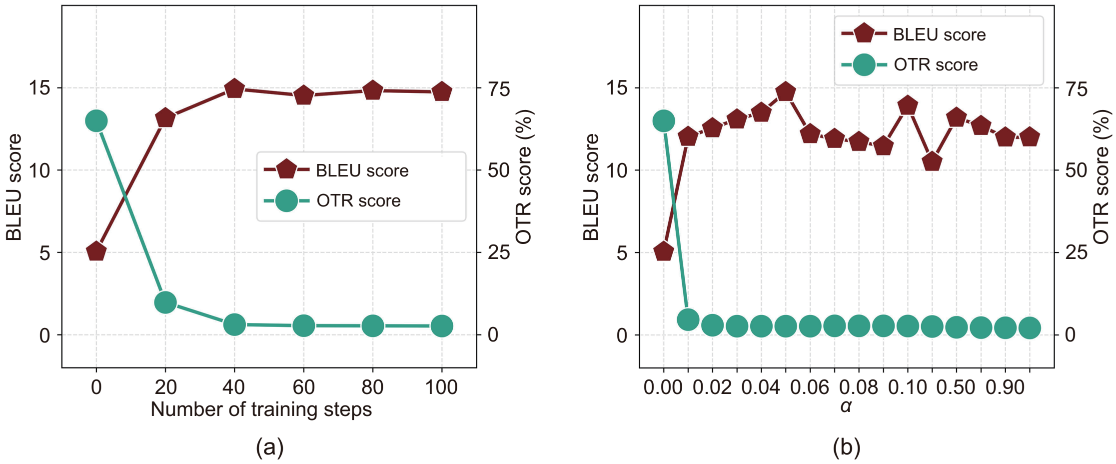


Fig. 3 Ablation studies: (a) number of unlikelihood training steps; (b) the mixing hyper-parameter α

Table 3 Impact of the model size on BLEU and OTR scores on the IWSLT benchmark

Method	Size	BLEU score↑						Avg
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 2-MT	7B	7.9	8.3	2.3	4.8	2.8	4.2	5.1
	13B	11.2	9.2	7.0	6.7	4.8	7.6	7.8
Ours	7B	**17.5**	**16.2**	**15.4**	**12.8**	**12.0**	**14.5**	**14.7**
	13B	**15.5**	**18.8**	**13.8**	**20.4**	**12.9**	**17.4**	**16.5**
Method	Size	OTR score (%)↓						Avg
		It→Nl	Nl→It	It→Ro	Ro→It	Nl→Ro	Ro→Nl	
LLaMA 2-MT	7B	49.8	39.8	85.8	65.7	80.1	68.8	65.0
	13B	38.9	55.1	57.3	69.0	67.1	59.7	57.9
Ours	7B	**2.7**	**1.5**	**3.8**	**1.5**	**3.8**	**2.5**	**2.6**
	13B	**3.4**	**1.1**	**4.6**	**1.2**	**4.3**	**2.7**	**2.9**

The better results are in bold. Avg: average score obtained for all directions; ↑: the higher the better; ↓: the lower the better

significantly lower OTR (2.9% vs. 57.9% average OTR score), leading to a higher quality translation (16.5 vs. 7.7 average BLEU score). The results demonstrate that our algorithm remains effective with larger LLMs.

5.4 Results with different amounts of translation data

Fig. 4 illustrates the ZST performance of LLaMA 2-based method, as measured by BLEU and OTR scores, with multilingual translation datasets of varying size, denoted as n samples. Four settings are considered, with dataset sizes of 12 000, 24 000, 48 000, and 96 000. Although the focus is on ZST performance, increasing the translation dataset size also yields improvements. For LLaMA 2-MT method, tuning with 96 000 samples achieves the best performance (14.2 BLEU score and 24.7% OTR score). Our method attains its peak performance with 96 000 training samples (18.5 BLEU score and 2.7% OTR score). Furthermore, our method demonstrates robustness across different translation dataset sizes and consistently achieves OTR scores close to zero across all four settings, resulting in significantly high BLEU scores. This confirms that our method remains effective with various sizes of fine-tuning dataset.

5.5 Performance on supervised translation

Our algorithm primarily enhances ZST performance through unlikelihood training on instruction-conflicting samples. It raises a question: Does the supervised translation ability persist even after unlikelihood training? As shown in Table 4, we report the performance of MT and ours on the IWSLT and WMT benchmarks. Remarkably, our method successfully retains the supervised translation ability after unlikelihood training with instruction-conflicting samples. Specifically, our final model achieves comparable results compared with LLaMA-MT for LLaMA 2-based (28.0 vs 28.5 in BLEU score) and LLaMA 3-based (28.9 vs. 29.3 BLEU score) models.

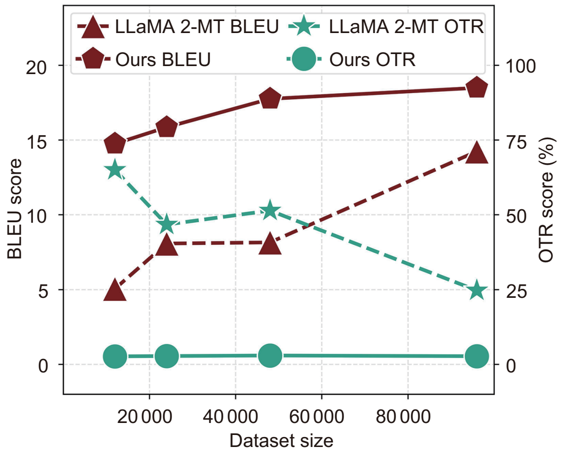


Fig. 4 Impact of the fine-tuning translation dataset size. We report the BLEU and OTR scores on the IWSLT benchmark

Table 4 Supervised translation performance

Base	Method	IWSLT	WMT	Avg
LLaMA 2	MT	**29.4**	**27.6**	**28.5**
	Ours	28.8	27.1	28.0
LLaMA 3	MT	**29.5**	**29.1**	**29.3**
	Ours	29.4	28.3	28.9

The better results are in bold. Avg: average BLEU score. Our method successfully achieves the goal of improving ZST performance without compromising the effectiveness of supervised translation

5.6 Comparison with MT models

To provide more insights into the improvements brought by our method, we introduce model-based translation evaluation metrics to compare MT with our method. The model-based metrics evaluate the semantic similarity of model generations and human translations, including reference-based XComet (https://huggingface.co/Unbabel/XCOMET-XL) and reference-free CometKiwi (https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl). Both of these metrics achieve better correlations with human evaluation than SacreBLEU. As they may assign high scores to language-mismatch translations, such as directly copying the source sentence or incorrectly translating the source sentence into another language, we split the test set into two parts: those where MT models translate the source text into the correct language and those where it translates the source text into an incorrect language. We report the scores for the former, and provide a case study for the latter.

As shown in Table 5, we observe that our method's generated translations achieve comparable semantic similarity to those of MT models when evaluated using model-based metrics. Moreover, our method achieves higher CometKiwi scores (+0.5) in both LLaMA 2- and LLaMA 3-based experiments.

Table 5 Evaluation with XComet and CometKiwi on samples translated into correct language by MT

Base	Method	XComet		Avg
		IWSLT	WMT	
LLaMA 2	MT	**88.6**	88.2	**88.4**
	Ours	88.4	**88.3**	**88.4**
LLaMA 3	MT	85.6	89.1	87.4
	Ours	**85.7**	**89.3**	**87.5**
Base	Method	CometKiwi		Avg
		IWSLT	WMT	
LLaMA 2	MT	**63.1**	66.9	65.0
	Ours	**63.1**	**67.8**	**65.5**
LLaMA 3	MT	64.0	68.8	66.4
	Ours	**64.3**	**69.4**	**66.9**

The better results are in bold. Avg: average score

In Table 6, we present a case study analyzing incorrect language translations by LLaMA 3-based models on the WMT En→Ja translation task. Compared to the MT method, our method generates translations that are more consistent with human translations, and are presented in the correct target language.

These observations demonstrate that our method enhances the model's ability to follow language-specific translation instructions, while preserving its existing translation capabilities.

5.7 Effect on the general task performance

Inspired by Jiao et al. (2023), we examine LLMs fine-tuned on a mixed-task dataset, encompassing both the general task and the translation task. Our objective is to investigate whether our approach can enhance ZST capabilities without compromising the general task performance. We construct the instruction tuning dataset by combining Alpaca (https://github.com/tatsu-lab/stanford_alpaca) with the IWSLT translation dataset, and denote models tuned on this dataset as Alpaca-(*)-MT. The Alpaca dataset consists of 52 000 samples generated by self-instruct (Wang YZ et al., 2023) across various task types. For the second fine-tuning stage of our method, only the translation data part is used. For fine-tuning, we use the same hyper-parameters as the main IWSLT experiments. We use the BLEU and OTR scores on the IWSLT test set for translation evaluation. Following AlpacaEval (https://github.com/tatsu-lab/alpaca_eval), we assess the general task performance with an LLM-based metric. Specifically, we use OpenAI ChatGPT (GPT-3.5-turbo) to perform the judgment automatically while taking LLaMA 2-7B tuned on Alpaca data as the reference model to compute the winning rate on the AlpacaEval dataset. The higher winning rate is the better. The responses are generated with a temperature of 0.3 and a repetition penalty of 1.2.

As shown in Fig. 5a, both Alpaca-7B-MT and Alpaca-13B-MT achieve better ZST performance than models solely fine-tuning with translation samples (10.9/12.8 vs 5.0/7.7 BLEU score and 26.3/24.7 vs. 65.0/57.9 OTR score). This aligns with prior findings (Chung et al., 2024), which suggests that more training tasks bring gains for tasks not present in the training data. By tuning with UL, our method achieves better ZST performance than Alpaca-7B-MT/Alpaca-13B-MT (14.8/16.2 vs. 10.9/12.8 BLEU score and 3.1/2.8 vs. 26.3/24.7

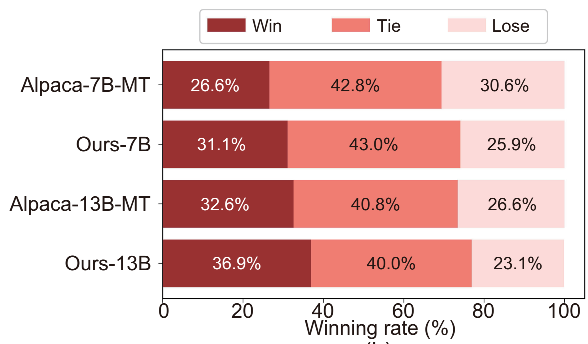
Table 6 Case study of LLaMA 3-based methods on the WMT En→Ja translation task

	Sentence		Sentence
Input	Critics say he remains inseparable from apartheid-era crimes and could have been held accountable for them had he lived longer.	Input	Sources have told The Age and The Sydney Morning Herald that the Gabba is the only major stadium in Australian cricket where ...
Human	家によると、同氏はアパルトヘイト代の犯罪とは切っても切れないにあり、もっとく生きていればその任をわれていたかもしれない。	Human	情筋は、世界模の映像配信に必要な膨大な数の中やデバイスを充分に稼させる会の主源の力供が不十分なのは、オストラリアの大模スタジアムでもガバだけだと……
MT	批判家表示，他始终与种族隔离时期的罪行不可分割，倘若他活得更长一点，就可能被追究责任。 (Translation in Chinese)	MT	Sources have told The Age and The Sydney Morning Herald that the Gabba is the only major stadium · · · (Copy the Input)
Ours	批家によると、彼はアパルトヘイト代の犯罪かられられず、もし生存していたならば任を担する必要があった。	Ours	《The Age》と《Sydney Morning Herald》(SMH)の取材によると、ガバはオストラリアの主要なクリケットスタジ……

We present translation examples for MT generating incorrect language translations, which are highlighted in yellow. "Input" and "Human" indicate the source English sentences and the ground-truth target Japanese sentences, respectively. References to color refer to the online version of this table

Method	BLEU↑	OTR (%)↓
LLaMA2-7B-MT	5.0	65.0
LLaMA2-13B-MT	7.7	57.9
Alpaca-7B-MT	10.9	26.3
Ours-7B	**14.8**	**3.1**
Alpaca-13B-MT	12.8	24.7
Ours-13B	**16.2**	**2.8**

(a)



(b)

Fig. 5 Performance after combining with general tasks: (a) ZST performance; (b) comparative winning rate. We combine the Alpaca and IWSLT benchmarks for fine-tuning. We report the ZST performance on the IWSLT test set. The better results are in bold. We also present the winning rate on the AlpacaEval dataset. The higher winning rate the better

OTR score). For general tasks, as depicted in Fig. 5b, our method achieves a comparable general task performance with Alpaca-7B-MT/Alpaca-13B-MT. These results confirm the effectiveness of our algorithm in addressing off-target translation issues without compromising general task performance.

6 Related works

1. Translation-tailored LLMs

Due to the huge cost to call state-of-the-art LLMs, such as GPT-4 (OpenAI, 2024), there is a need to investigate how to effectively fit a smaller LLM into specific tasks, e.g., machine translation. Note that although there are some powerful sequence-to-sequence-style large-scale pre-trained machine translation models (Liu YH et al., 2020; Zan et al., 2022), this paper mainly focuses on the decoder-only LLMs due to their flexible interaction modes and rich world knowledge. In the field of LLM-based translation, various approaches have been proposed to optimize translation performance (Jiao et al., 2023; Liu YJ et al., 2023; Feng et al., 2024; Li JH et al., 2024; Stap et al., 2024; Xu HR et al., 2024a, 2024b; Zeng et al., 2024; Zhang HB et al., 2024). ParroT (Jiao et al., 2023) proposes fine-tuning the model on machine translation data with a hint that incorporates extra requirements to regulate the translation process. TIM (Zeng et al., 2024) introduces translation samples in comparisons to compute additional preference loss for regularization, exhibiting superior translation ability in both supervised and zero-shot directions. ALMA (Xu HR et al., 2024b) proposes a two-stage approach that first fine-tuned on monolingual data of downstream languages followed by fine-tuning on high-quality translation data, which achieves significant improvement of translation quality. ALMA-R (Xu HR et al., 2024a) aligns the ALMA models with the preferences of translation evaluators using contrastive preference optimization, achieving better performance than human translation on the test set. Liu YJ et al. (2023) presented the position of instruction matters, and that just moving the location of the instruction closer to the output can alleviate the instruction forgetting issue. Zhang HB et al. (2024) showed the overlooking of source sentences in LLMs, and proposed both unsupervised and supervised approaches to improve this issue.

In contrast, we focus on the off-target problem of ZST, where the model fails to follow translation instructions, generating sequences not in the target language. Additionally, we show how instruction-conflicting samples can enhance the influence of instruction, thus mitigating the off-target problem.

2. Unlikelihood training

Unlikelihood training (Welleck et al., 2019) aims to force the model to assign a lower probability to unlikely tokens. This method has been further explored in dialog tasks by Li M et al. (2020), who demonstrated its effectiveness in generating more consistent and coherent human-like dialog. Nogueira dos Santos et al. (2020) used the UL for ranking, and proposed a generative information retrieval approach. Hosseini et al. (2021) proposed the combination of an unlikelihood objective with a reference-based setup for input sentences to model negation with pre-trained BERT (Devlin et al., 2019). Hu et al. (2023) took the semantically-similar or ambiguous tokens as negative information and acquired them via inherent uncertainty for the ASQP task. Zan et al. (2023) proposed a method to alleviate the off-target problem in translation models trained on multilingual translation corpora, which consist of millions of sentence pairs. They analyzed the impact of exposure bias on the off-target translation issue and used incorrect language ID sentences as negative samples.

In this work, we focus on translation-tailored LLMs that reduce the reliance of translation models on expensive, large-scale parallel corpora. We take the instances where translation pairs conflict with the given instructions as negative samples for ZST. Furthermore, we consider a new case that enhances the ability of LLMs to better follow translation instructions and generate translations in the correct language.

7 Conclusions and future work

7.1 Conclusions

We propose a simple two-stage fine-tuning strategy to enhance the instruction-following ability of LLM for translation. The core procedure consists of two main steps: (1) creating instruction-conflicting samples by replacing the translation directions with incorrect ones; (2) training these samples using an additional UL. Experimental results on IWSLT and

WMT, spanning 16 ZST directions, demonstrate the effectiveness of the proposed method, which reduces the OTR and produces translations of higher quality. Furthermore, our approach exerts a negligible influence on other aspects of LLMs, such as supervised translation performance and general task performance.

7.2 Future work

The proposed method contains a mixing hyperparameter α to balance MLE loss and UL in unlikelihood training on instruction-conflicting samples, and the high α may overfit the model on UL. In future work, we may focus on how to balance them adaptively.

This work only focuses on the off-target problem in the ZST of LLMs, which could be seen as a specific type of input-conflicting hallucination. In future work, we will continue to explore the application of the unlikelihood training on general tasks, such as programming, math, and dialog, and more types of hallucinations (Zhang Y et al., 2023), such as fact-conflicting hallucinations and context-conflicting hallucinations. Furthermore, we will apply our method to enhance the instruction-following abilities of different interesting LLM-based scenarios, e.g., safety (Miao et al., 2024; Zhang YQ et al., 2024; Zhong et al., 2024), debiasing (Xu ZY et al., 2024), multimodal analysis (Wang WB et al., 2024), healthcare (Ren et al., 2024), and difficult code generation (Wang S et al., 2024).

Contributors

Changtong ZAN and Liang DING designed the research. Changtong ZAN processed the data and drafted the paper. Li SHEN, Yibing ZHAN, and Xinghao YANG helped organize the paper. Changtong ZAN, Liang DING, and Weifeng LIU revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. https://doi.org/10.48550/arXiv.2005.14165

Chen L, Ma SM, Zhang DD, et al., 2023. On the off-target problem of zero-shot multilingual neural machine translation. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Findings of the Association for Computational Linguistics. Association for Computational Linguistics, Toronto, Canada, p.9542-9558. https://doi.org/10.18653/v1/2023.findings-acl.608

Cho K, van Merrienboer B, Gulcehre C, et al., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proc Conf on Empirical Methods in Natural Language Processing, p.1724-1734. https://doi.org/10.3115/v1/D14-1179

Chung HW, Hou L, Longpre S, et al., 2024. Scaling instruction-finetuned language models. *J Mach Learn Res*, 25(70):1-53.

Dabre R, Kurohashi S, 2019. MMCR4NLP: multilingual multiway corpora repository for natural language processing. https://arxiv.org/abs/1710.01025

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Feng ZP, Chen RZ, Zhang Y, et al., 2024. Ladder: a model-agnostic framework boosting LLM-based machine translation to the next level. https://doi.org/10.48550/arXiv.2406.15741

Fu Y, Peng H, Ou LT, et al., 2023. Specializing smaller language models towards multi-step reasoning. Proc 40th Int Conf on Machine Learning, p.10421-10430.

Gu JT, Wang Y, Cho K, et al., 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. Proc 57th Annual Meeting of the Association for Computational Linguistics, p.1258-1268. https://doi.org/10.18653/v1/P19-1121

Hendy A, Abdelrehim M, Sharaf A, et al., 2023. How good are GPT models at machine translation? A comprehensive evaluation. https://doi.org/10.48550/arXiv.2302.09210

Hosseini A, Reddy S, Bahdanau D, et al., 2021. Understanding by understanding not: modeling negation in language models. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.1301-1312. https://doi.org/10.18653/v1/2021.naacl-main.102

Hu MT, Bai YH, Wu YK, et al., 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In: Rogers A, Boyd-Graber J, Okazaki N (Eds.), Findings of the Association for Computational Linguistics. Association for Computational Linguistics, Toronto, Canada, p.13481-13494. https://doi.org/10.18653/v1/2023.findings-acl.851

Huang YX, Gu HL, Yu ZT, et al., 2024. Enhancing low-resource cross-lingual summarization from noisy data with fine-grained reinforcement learning. *Front Inform Technol Electron Eng*, 25(1):121-134. https://doi.org/10.1631/FITEE.2300296

Jiao WX, Huang JT, Wang WX, et al., 2023. ParroT: translating during chat using large language models tuned with human translation and feedback. https://doi.org/10.48550/arXiv.2304.02426

Joulin A, Grave E, Bojanowski P, et al., 2016a. Bag of tricks for efficient text classification. https://doi.org/10.48550/arXiv.1607.01759

Joulin A, Grave E, Bojanowski P, et al., 2016b. FastText.zip: compressing text classification models. https://doi.org/10.48550/arXiv.1612.03651

Kaplan J, McCandlish S, Henighan T, et al., 2020. Scaling laws for neural language models. https://doi.org/10.48550/arXiv.2001.08361

Kwon W, Li ZH, Zhuang SY, et al., 2023. Efficient memory management for large language model serving with PagedAttention. Proc 29th Symp on Operating Systems Principles, p.611-626. https://doi.org/10.1145/3600006.3613165

Li B, Yang P, Sun YK, et al., 2024. Advances and challenges in artificial intelligence text generation. *Front Inform Technol Electron Eng*, 25(1):64-83. https://doi.org/10.1631/FITEE.2300410

Li JH, Zhou H, Huang SJ, et al., 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Trans Assoc Comput Linguist*, 12:576-592.

Li M, Roller S, Kulikov I, et al., 2020. Don't say that! Making inconsistent dialogue unlikely with unlikelihood training. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.4715-4728. https://doi.org/10.18653/v1/2020.acl-main.428

Liu YH, Gu JT, Goyal N, et al., 2020. Multilingual denoising pre-training for neural machine translation. *Trans Assoc Comput Linguist*, 8:726-742. https://doi.org/10.1162/tacl_a_00343

Liu YJ, Zeng XF, Meng FD, et al., 2023. Instruction position matters in sequence generation with large language models. https://doi.org/10.48550/arXiv.2308.12097

Lu QY, Qiu BP, Ding L, et al., 2024. Error analysis prompting enables human-like translation evaluation in large language models. https://doi.org/10.48550/arXiv.2303.13809

Miao YC, Zhang S, Ding L, et al., 2024. InfoRM: mitigating reward hacking in RLHF via information-theoretic reward modeling. https://doi.org/10.48550/arXiv.2402.09345

Min BN, Ross H, Sulem E, et al., 2023. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*, 56(2):1-40. https://doi.org/10.1145/3605943

Mishra S, Khashabi D, Baral C, et al., 2022. Cross-task generalization via natural language crowdsourcing instructions. Proc 60th Annual Meeting of the Association for Computational Linguistics, p.3470-3487. https://doi.org/10.18653/v1/2022.acl-long.244

Nogueira dos Santos C, Ma XF, Nallapati R, et al., 2020. Beyond [CLS] through ranking by generation. Proc Conf on Empirical Methods in Natural Language Processing, p.1722-1727. https://doi.org/10.18653/v1/2020.emnlp-main.134

OpenAI, 2024. GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774

Peng KQ, Ding L, Zhong QH, et al., 2023. Towards making the most of ChatGPT for machine translation. In: Bouamor H, Pino J, Bali K (Eds.), Findings of the Association for Computational Linguistics. Association for Computational Linguistics, Singapore, p.5622-5633. https://doi.org/10.18653/v1/2023.findings-emnlp.373

Post M, 2018. A call for clarity in reporting BLEU scores. Proc 3rd Conf on Machine Translation: Research Papers, p.186-191. https://doi.org/10.18653/v1/W18-6319

Qu Z, Watanabe T, 2022. Adapting to non-centered languages for zero-shot multilingual translation. Proc 29th Int Conf on Computational Linguistics, p.5251-5265. https://aclanthology.org/2022.coling-1.467

Ren ZY, Zhan YB, Yu BS, et al., 2024. Healthcare Copilot: eliciting the power of general LLMs for medical consultation. https://doi.org/10.48550/arXiv.2402.13408

Sennrich R, Vamvas J, Mohammadshahi A, 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. https://doi.org/10.48550/arXiv.2309.07098

Stap D, Hasler E, Byrne B, et al., 2024. The fine-tuning paradox: boosting translation quality without sacrificing LLM abilities. Proc 62nd Annual Meeting of the Association for Computational Linguistics, p.6189-6206. https://doi.org/10.18653/v1/2024.acl-long.336

Touvron H, Lavril T, Izacard G, et al., 2023a. LLaMA: open and efficient foundation language models. https://doi.org/10.48550/arXiv.2302.13971

Touvron H, Martin L, Stone K, et al., 2023b. LLaMA 2: open foundation and fine-tuned chat models. https://doi.org/10.48550/arXiv.2307.09288

Wang S, Ding L, Shen L, et al., 2024. OOP: object-oriented programming evaluation benchmark for large language models. https://doi.org/10.48550/arXiv.2401.06628

Wang WB, Ding L, Shen L, et al., 2024. WisdoM: improving multimodal sentiment analysis by fusing contextual world knowledge. https://doi.org/10.48550/arXiv.2401.06659

Wang YM, Zhang ZS, Wang R, 2023. Element-aware summarization with large language models: expert-aligned evaluation and chain-of-thought method. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.8640-8665. https://doi.org/10.18653/v1/2023.acl-long.482

Wang YZ, Kordi Y, Mishra S, et al., 2023. Self-instruct: aligning language models with self-generated instructions. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.13484-13508. https://doi.org/10.18653/v1/2023.acl-long.754

Wei J, Bosma M, Zhao VY, et al., 2022. Finetuned language models are zero-shot learners. https://doi.org/10.48550/arXiv.2109.01652

Wei J, Wang XZ, Schuurmans D, et al., 2023. Chain-of-thought prompting elicits reasoning in large language models. https://doi.org/10.48550/arXiv.2201.11903

Welleck S, Kulikov I, Roller S, et al., 2019. Neural text generation with unlikelihood training. https://doi.org/10.48550/arXiv.1908.04319

Wolf T, Debut L, Sanh V, et al., 2020. Transformers: state-of-the-art natural language processing. Proc Conf on Empirical Methods in Natural Language Processing:

System Demonstrations, p.38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu HR, Sharaf A, Chen YM, et al., 2024a. Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation. https://doi.org/10.48550/arXiv.2401.08417

Xu HR, Kim YJ, Sharaf A, et al., 2024b. A paradigm shift in machine translation: boosting translation performance of large language models. https://doi.org/10.48550/arXiv.2309.11674

Xu ZY, Peng KQ, Ding L, et al., 2024. Take care of your prompt bias! Investigating and mitigating prompt bias in factual knowledge extraction. Proc Joint Int Conf on Computational Linguistics, Language Resources and Evaluation, p.15552-15665. https://aclanthology.org/2024.lrec-main.1352

Zan CT, Peng KQ, Ding L, et al., 2022. Vega-MT: the JD explore academy machine translation system for WMT22. Proc 7th Conf on Machine Translation, p.411-422.

Zan CT, Ding L, Shen L, et al., 2023. Unlikelihood tuning on negative samples amazingly improves zero-shot translation. https://doi.org/10.48550/arXiv.2309.16599

Zeng JL, Meng FD, Yin YJ, et al., 2024. TIM: teaching large language models to translate with comparison. https://doi.org/10.48550/arXiv.2307.04408

Zhang HB, Chen Q, Zhang WW, 2022. Improving entity linking with two adaptive features. *Front Inform Technol Electron Eng*, 23(11):1620-1630. https://doi.org/10.1631/FITEE.2100492

Zhang HB, Chen KH, Bai XF, et al., 2024. Paying more attention to source context: mitigating unfaithful translations from large language model. In: Ku LW, Martins A, Srikumar V (Eds.), Findings of the Association for Computational Linguistics. Association for Computational Linguistics, Bangkok, Thailand, p.13816-13836. https://doi.org/10.18653/v1/2024.findings-acl.821

Zhang SS, Roller S, Goyal N, et al., 2022. OPT: open pretrained Transformer language models. https://doi.org/10.48550/arXiv.2205.01068

Zhang Y, Li YF, Cui LY, et al., 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. https://doi.org/10.48550/arXiv.2309.01219

Zhang YQ, Ding L, Zhang LF, et al., 2024. Intention analysis makes LLMs a good jailbreak defender. https://doi.org/10.48550/arXiv.2401.06561

Zhong QH, Ding L, Liu JH, et al., 2023. Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. https://doi.org/10.48550/arXiv.2302.10198

Zhong QH, Ding L, Liu JH, et al., 2024. Rose doesn't do that: boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. https://doi.org/10.48550/arXiv.2402.11889