Changtong ZAN, Liang DING, Li SHEN, Yibing ZHAN, Xinghao YANG, Weifeng LIU, 2025. Building accurate translation-tailored large language models with language-aware instruction tuning. *Frontiers of Information Technology & Electronic Engineering*, 26(8):1341-1355. https://doi.org/10.1631/FITEE.2400458

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

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

Liang DING
E-mail: liangding.liam@gmail.com
ORCID: https://orcid.org/0000-0001-8976-2084

Weifeng LIU
E-mail: liuwf@upc.edu.cn
ORCID: https://orcid.org/0000-0002-5388-9080

Motivation

1. Proprietary large language models (LLMs) demonstrate strong translation capabilities. However, their practical application is limited by significant operational costs.
2. Fine-tuning smaller LLMs on translation data has emerged as a cost-effective strategy to achieve superior performance. However, these models frequently encounter a critical “off-target translation” problem in zero-shot settings, generating output in an incorrect language.
3. Existing approaches attempt to mitigate this off-target issue by modifying the inference process. Nevertheless, these methods fail to fundamentally improve the models’ core instruction-following ability and awareness of translation direction.

Main idea

1. The paper proposes a two-stage fine-tuning algorithm to enhance the instruction-following ability of translation-tailored LLMs, thereby alleviating the off-target translation problem.
2. The first stage fine-tunes the LLM on a multilingual translation dataset to unlock and establish its foundational translation capabilities.
3. The second stage introduces instruction-conflicting samples and leverages unlikelihood loss to actively penalize the model for generating translations that deviate from the specified task, thereby forcing it to strictly adhere to the correct translation direction.

Method

1. To unlock the translation capabilities of LLMs, a pre-tuning stage is employed, where the model is trained on a collection of multilingual translation samples using the common maximum likelihood estimation (MLE) approach.

2. To address the off-target issue and improve the LLMs' instruction-following ability, instruction-conflicting samples are created by substituting the original instruction with a different one while keeping the input and output unchanged.

Method (Cont'd)

3. To prevent potential overfitting on the unlikelihood objective while maintaining the supervised translation ability, multilingual translation samples are incorporated to simultaneously train the model with likelihood loss.

4. To enhance the ability to follow instructions for translation tasks, unlikelihood training is used, which aims to reduce the probability assigned by the model to the output tokens of the instruction-conflicting samples.

Method (Cont'd)

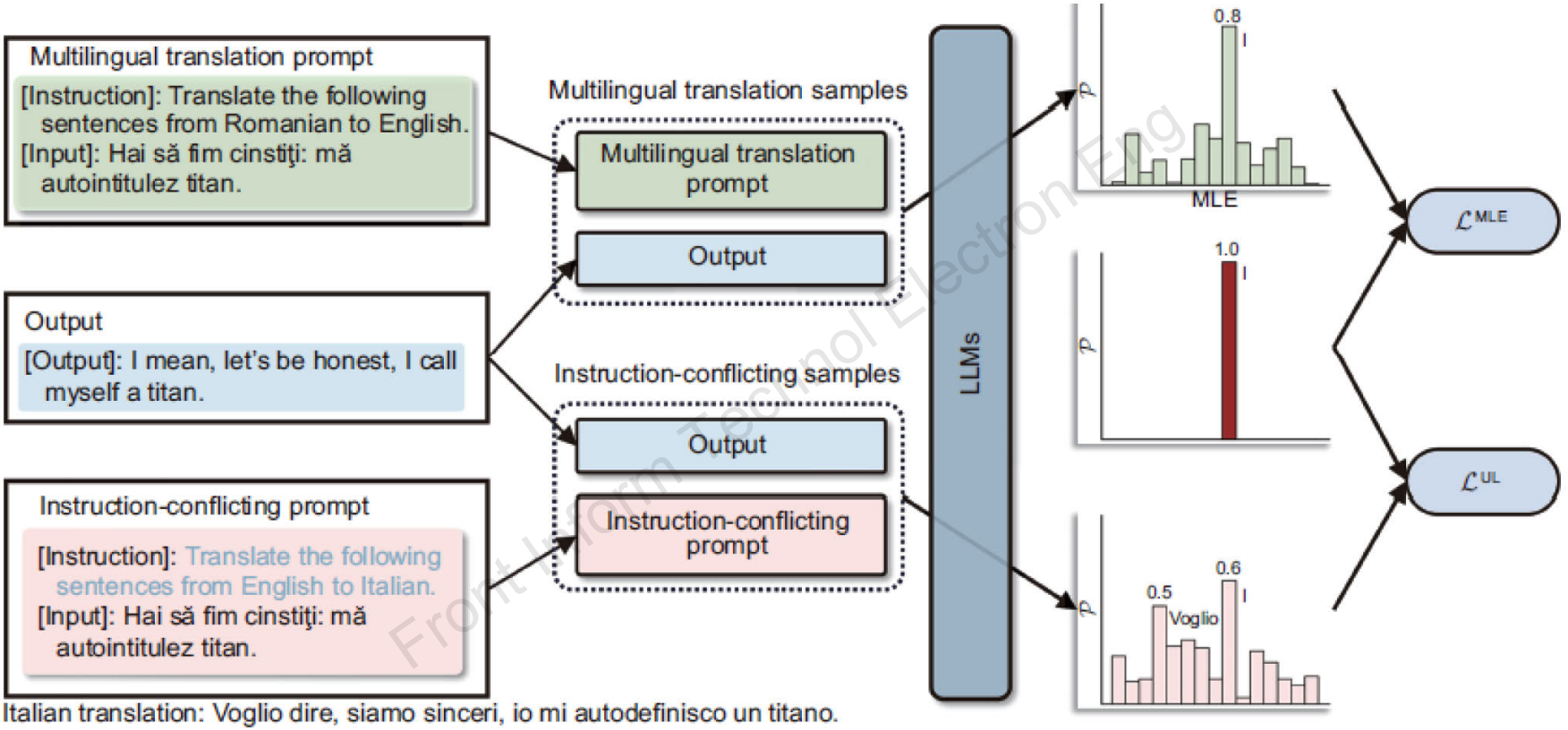


Fig. 2 Overview of our fine-tuning framework for ZST. At the first stage, we pre-tune LLMs on multilingual translation samples, focusing on unlocking the translation ability of LLMs. Subsequently, we introduce instruction-conflicting samples by randomly substituting the instruction component with a different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

Major results

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

Major results (Cont'd)

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

Major results (Cont'd)

Ablation study

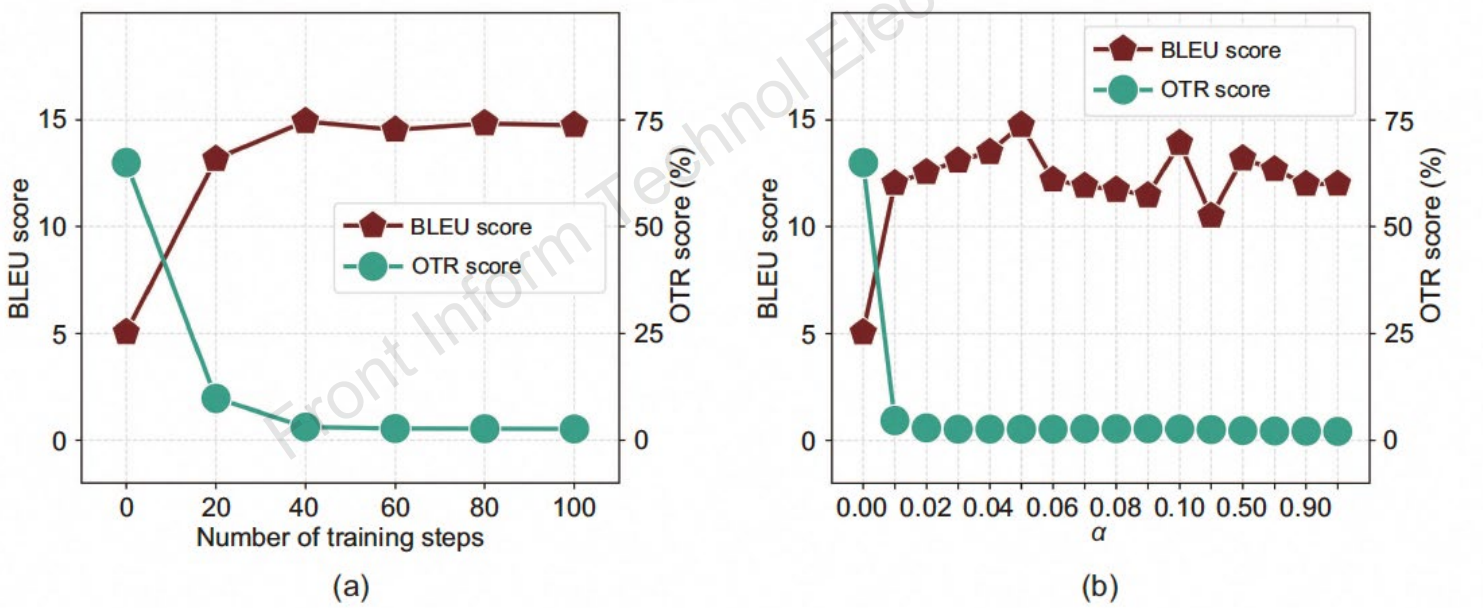


Fig. 3 Ablation studies: (a) number of unlikelihood training steps; (b) the mixing hyper-parameter α

Major results (Cont'd)

Results with different sizes of LLMs

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

Major results (Cont'd)

Results with different amounts of translation data

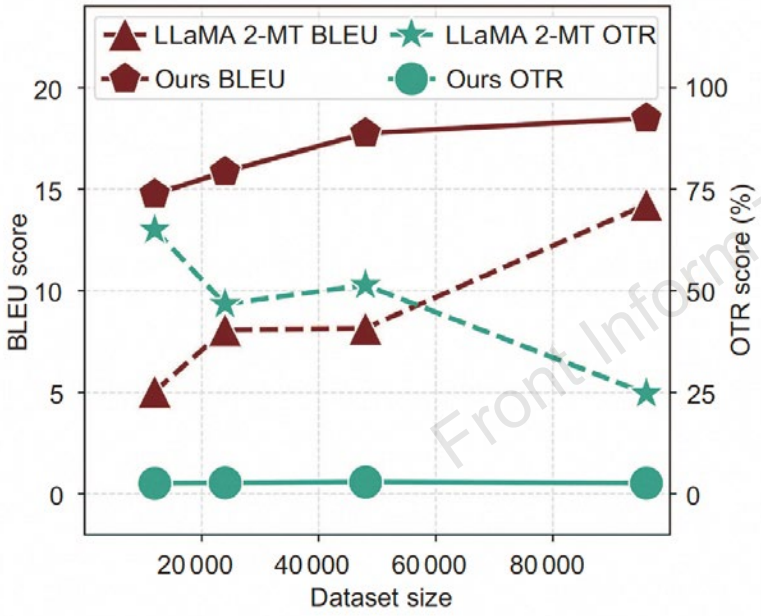


Fig. 4 Impact of the fine-tuning translation dataset size. We report the BLEU and OTR scores on the IWSLT benchmark

Performance on supervised translation

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

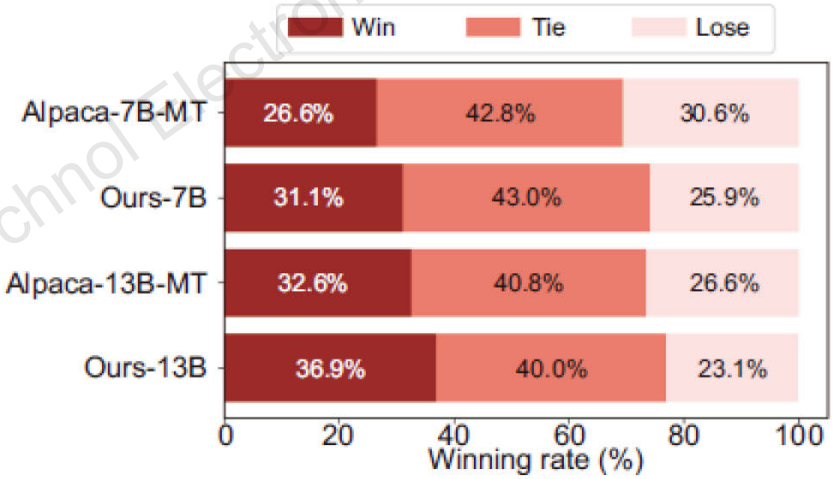
The better results are in bold. Avg: average BLEU score. Our method successfully achieves the goal of improving ZST performance without compromising the effectiveness of supervised translation

Major results (Cont'd)

Performance after combining general tasks

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

Conclusions

1. A two-stage fine-tuning strategy is proposed to enhance LLM's instruction-following for translation by training on instruction-conflicting samples.

2. The method proves to be effective, significantly reducing the off-target translation ratio and improving translation quality on the IWSLT and WMT benchmarks.

3. The proposed approach has a negligible negative impact on the model's standard translation capabilities and general task performance.

Changtong ZAN received his B.S. degree in electronic information engineering from China University of Petroleum (East China) in 2019. He is currently a Ph.D. candidate in the College of Control Science and Engineering at the same institution. His research interests include large language models and machine translation.

Liang DING received his Ph.D. degree in computer science from the University of Sydney in 2022. Formerly the Head of the NLP Research Group at JD Explore Academy, JD.com Inc., he has published over 100 papers in premier venues like ACL, EMNLP, and NeurIPS, where he also serves as Area Chairs. He has won numerous AI competitions, including SuperGLUE/GLUE and WMT. His leadership on a large-scale language model project garnered him the 2022 WAIC SAIL Award and JD.com's highest technical award—Technology Golden Award.

Weifeng LIU is currently a professor with the College of Control Science and Engineering, China University of Petroleum (East China), China. He received the B.S. degree in automation and business administration and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2002 and 2007, respectively. He has authored or co-authored numerous papers in top journals and prestigious conferences, including 11 ESI highly-cited papers and 3 ESI hot papers. He serves as an associate editor for *Neural Processing Letter*, a co-chair for IEEE SMC Technical Committee on Cognitive Computing, and a guest editor of special issues for *Signal Processing*, *IET Computer Vision*, *Neurocomputing*, and *Remote Sensing*. His current research interests include pattern recognition and machine learning.