Improving Translation Faithfulness of Large Language Models via Augmenting Instructions

Anonymous ACL submission

Abstract

Large Language Models (LLMs) present strong general capabilities, and a current compelling challenge is stimulating their specialized capabilities, such as machine translation through low-cost instruction tuning. The standard data for instruction-following is organized as a concatenated series of instructions, inputs, and outputs. Due to the inherent pattern of the attention mechanism in LLMs, these models tend to concentrate more on nearby tokens. Consequently, there is a high risk of forgetting instructions during the decoding process, particularly when dealing with long contexts. To alleviate the instruction forgetting issue on translation, we propose SWIE (**S**egment-**W**eighted **I**nstruction **E**mbedding) and an instruction-following dataset OVERUNDER. SWIE improves the model instruction understanding by adding an instruction representation on the following input and response representations. OVERUNDER improves model faithfulness by comparing over-translation and under-translation samples with the correct translation. We apply our methods to two mainstream open-source LLMs, *i.e.*, BLOOM and LLaMA. Experimental results demonstrate that using SWIE and OVERUNDER in models improves translation performance and faithfulness over the strong baselines. Furthermore, SWIE improves the model performance on various long-context scenarios, including in-context translation, translation on language direction in the instruction-tuning corpus, and translation on zero-shot language pairs. The effectiveness of SWIE is demonstrated on the IFEval instruction-following test set, indicating its potential for broader task applicability.

1 Introduction

In recent years, super closed-source large language models (LLMs) like GPT-4 and ChatGPT have demonstrated remarkable performance on translation tasks without fine-tuning (Jiao et al., 2023b; Hendy et al., 2023; Raunak et al., 2023; He et al., 2023). Considering the hardware constraints in research, many current works employ a small amount of instruction data for fine-tuning to elicit the capabilities of medium-sized models. Typically, instruction-following data is organized by a sequence of the task instruction (for a translation task, it can be "Please translate the sentence from English to Chinese"), the task input, and the output. Some existing studies (Jiao et al., 2023a; Zhang et al., 2023; Zeng et al., 2023) have adopted various instruction data construction and training methods in the translation domain, achieving appealing results with relatively low computational cost.

The localization of the attention mechanism in LLMs has been a widely observed phenomenon. For example, Liu et al. (2023) demonstrates that the model performance significantly degrades when the model must access information in the middle of a long context. In the instruction-following data setting, we hypothesize that this feature leads to a high risk of attention inadequacy and forgetting issues for the instruction placed at the beginning of the text, especially when generating an output with a long context. To verify the above hypothesis, we experimented with translation language direction detection with different sentence lengths and observed that the translation direction accuracy decreases with the input text getting longer, proving the instruction forgetting phenomenon evident in translation tasks. For translation tasks, ignoring instructions can lead to the low quality of translation output, especially the unfaithfulness problem (compassing over-translation and under-translation, *i.e.,* the model translation results contains the content that is not contained in the source or omits the content in the source). Therefore, our work aims to improve the translation faithfulness of instruction-tuning by addressing the above issues.

This paper introduces a novel method for improving instruction tuning named SWIE (**S**egment-**W**eighted **I**nstruction **E**mbedding), which utilizes

trainable adapters to encode instruction and introduces segment weight to enable a natural integration of instruction representations and global representations. To further improve the model translation faithfulness, we present OVERUNDER, an instruction dataset that utilizes our proposed framework to collect contrastive negative samples that specifically target over-translation and under-translation issues.

We evaluate our methods on two machine translation benchmarks and two mainstream backbone models, *i.e.*, BLOOM (Workshop et al., 2022) and LLaMA (Touvron et al., 2023)[1]. SWIE shows wide effectiveness in various machine translation scenarios for LLMs. For the instruction-following scenario, the combination of SWIE and OVERUNDER leads to significant improvements (*e.g.,* up to 2.83 BLEU in LLaMA-7b on the Flores test set). For in-context translation, SWIE notably improves the translation performance by around 7 to 10 BLEU scores. Additionally, we observed that SWIE exhibits further improvements in long-context and zero-shot settings. Furthermore, our human and statistic faithfulness evaluation results indicate that SWIE and OVERUNDER improve the translation faithfulness effectively. Evaluation on instruction-following benchmark IFEval shows SWIE boosts the general instruction-following ability of the models (*e.g.,* 9.54% relative improvement can be seen in prompt-level evaluation). In summary, our contributions are as follows:

- We propose Segment-weighted Instruction Embedding (SWIE) that augments the instruction information in global positions and introduces a translation faithfulness contrastive instruction-tuning dataset OVERUNDER covering the over-translation and the under-translation unfaithfulness negative samples.
- Our experiments show that both SWIE and OVERUNDER consistently improve the translation quality on lexical and semantic metrics. The strength of SWIE can be seen in different scenarios of translation, including direct instruction following and in-context learning.
- According to our further analysis experiments, SWIE shows more significant improvements in long-context and zero-shot scenarios. We also quantified and visualized that SWIE leads to a higher internal instruction attention score. Additionally, the human and statistic evaluation on faithfulness presents that both SWIE and OVERUNDER lead to a more faithful translation, and our evaluation of the IFEval test set shows that SWIE improves the general instruction-following ability.

[1]We acquired the LLaMA weights through the official application form and adhered to the stipulations of the license.

2 Backgound

2.1 Instruction Tuning Formalization

Instruction tuning is one of the alignment methods to make language models meet human preferences. In a typical instruction tuning data item, the initial part of the text is task instruction s, followed by an optional task input x, and the model is expected to generate the task target output y finally (Ye et al., 2022). The standard instruction tuning is trained with maximum likelihood estimation (MLE), and the training objection can be calculated by Equation 1.

$$L_{MLE} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}; x; s) \quad (1)$$

2.2 The Attention Pattern in Translation Instruction-following Data

Most open-source LLMs use the causal decoder architecture because of the wide observation of scaling law on the causal decoder (Raffel et al., 2020; Zhao et al., 2023). In particular, we found that instruction tuning in transformers shows a highly consistent pattern, with concentrated attention on the special tokens of each span, which provides a predictable sign to predict the position importance. We sample an instruction-tuning example on a machine translation task and forward the example on Parrot-hint (Jiao et al., 2023a) based on BLOOMZ-3b[2], and observe in the heapmat in Figure 1 that the two strongest (excluding the beginning token) and global attention concentrated tokens are the end of instruction and the input separately. A similar observation is also claimed by Xiao et al. (2023), which notes that special tokens containing not much meaning can gather attention. The heatmap further reveals that the attention mechanism tends to pay more attention to the nearby text except for the special tokens, and the attention

[2]We perform max pooling on the last 10 layers and multi-heads of the model to generate a heatmap that shows the most significant feature at an abstract level.

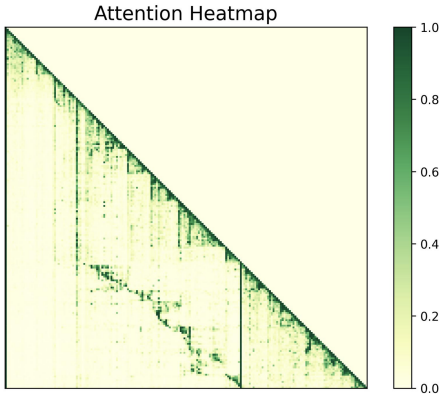

Figure 1: The heatmap is the attention score of a sample translation instruction data. The most prominent and globally focused tokens correspond to the ending tokens of the instruction and input span separately.

score on output spans to the end of instruction gradually decays with the distance getting longer, which can be evidence of instruction forgetting problems.

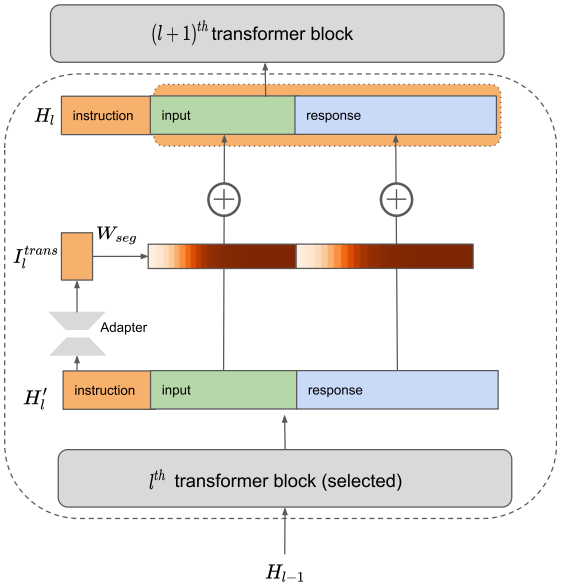

Figure 2: The model structure of SWIE. In the selected transformer layers, an adapter will transform the hidden states of the instruction span, and we get an instruction representation I_l^{trans}. Then, a segment Sigmoid weight W_{seg} is used to control the fusion ratio of instruction representation on different positions.

3 SWIE: Segment-weighted Instruction Embedding

To solve the instruction forgetting problem, we propose segment-weighted instruction embedding. The schematic diagram of the method is shown in Figure 2. An instruction-tuning data can be divided into several segments: instruction, input, output, etc. We extract the instruction of each sample from the hidden states of certain layers[3]. Then, we add a trainable lightweight adapter for the instruction representation to enable the model to learn a new pattern to fusion instruction information with the input and output spans. After instruction states passing the adapter, we fuse the max pooled[4] transformed feature representation with a well-designed segment-aware weight. The two main components of SWIE are described as follows.

3.1 Instruction Adapter

The adapter follows the structure in (Houlsby et al., 2019). The instruction representation can be obtained in the output of each decoder layer, and we use an instruction adapter to re-parameterize instruction.

Let H_l be the hidden output of l^{th} layer, and the H_l^{ins} represents the max pool result of the instruction part in H_l. We use a down-sampling linear layer $\mathbf{L_{down}}$, a Tanh activation layer σ, and an up-sampling linear layer $\mathbf{L_{up}}$ as the adapter, following Equation 2.

$$f(H_l^{ins}) = \mathbf{L_{up}}(\sigma(\mathbf{L_{down}}(H_l^{ins}))) \quad (2)$$

3.2 Segment-Weight and Global Fusion

We use a segment weight containing a Sigmoid function for each span. During our preliminary experiments (Section D.2), we discovered that fusing features directly with a constant weight for all positions would corrupt the model's representation during training.

Combining the observation in Section 2.2, the ending special tokens of each span have concentrated attention scores. We assign the two tokens and the few nearby tokens for weights nearly zero based on the above observation to avoid the attention pattern on the two ending tokens destroyed.

A sentence is tokenized into a list of token indexes and then fed into the model, and for the segment index list, we define the segments by instruction, input, and output, and the segments are separated by our pre-defined special token in the prompts. Assuming the tokenized span list set is $S = \{s_{ins}, s_{input}, s_{output}, \cdots\}$ (ins is an abbreviation for "instruction"), and we assign a correspondent index set $D = \{0, 1, 2, \cdots\}$, where

[3]The position and number of layer selections will be analyzed in Section D.1

[4]Our preliminary experiments indicate that the pooling method is not a sensitive setting.

$D \rightarrow S$. For $s \in S$, let the length of the span token list be N_s, the span index be D_s, the constant bias for Sigmoid function be b, segment weight be $W^s \in \mathbf{R}^{1 \times N_s}$, and the value of w_i^s ($i \in [0, N_s - 1]$) can be calculated as Equation 4. W_{seg} is the concatenation of $W^{s_{ins}}$, $W^{s_{input}}$ and $W^{s_{output}}$.

$$\text{Sigmoid}(x) = \frac{1}{1 - e^{-x}} \tag{3}$$

$$w_i^s = \begin{cases} 0 & s = s_{ins} \\ \text{Sigmoid}(i - b) & s \neq s_{ins} \end{cases} \tag{4}$$

$$H_l := H_l + W_{seg} \cdot f(H_l^{ins}) \tag{5}$$

4 OVERUNDER: A Natural Hallucination Dataset

In the machine translation task, the two most frequent phenomena of model unfaithfulness for fluent output are over-translation and under-translation. Over-translation refers to the situation in which the translated sentence contains words irrelevant to the source sentence, and under-translation refers to the situation in which the translation sentence lacks part of the information from the source sentence. Thus, we prompt gpt-3.5-turbo[5] to mimic the two typical error types, and the prompts are appended in Table.13.

To qualify the extent of under-translation or over-translation errors of generated sentences, we use awesome-align (Dou and Neubig, 2021) to evaluate the word-level cross-lingual alignment rate, and the statistic result is shown in Table.1. For reference corpus, the coverage of both source and target is around 90%, indicating that the statistical metric is roughly accurate. The source sentences of the under-translation dataset mainly cover semantics in the source sentences, but the semantics of the source sentences are significantly reduced in the target sentences. Conversely, the over-translation dataset behaves in the opposite manner. We conduct a detailed analysis of OVERUNDER in the Section A of the appendix.

data	source coverage	target coverage
reference	0.8845	0.8699
under data	0.5800	0.7180
over data	0.6958	0.5771

Table 1: Data statistics of generated over-translation and under-translation data.

[5]https://platform.openai.com/docs/models/gpt-3-5

5 Emprical Experiments

Following (Jiao et al., 2023a), we chose BLOOM and LLaMA (with parameter sizes from 3B to around 7B) as the backbone models. There are 4 translation directions included, De ⇒ En, En ⇒ De, En ⇒ Zh, and Zh ⇒ En. The implement details can be seen in Section B, including training hyper-parameters settings.

5.1 Training Setting

Alpaca The Alpaca dataset (Taori et al., 2023) is a high-quality multi-task instruction-following dataset that contains 52K items. LLMs fine-tuned by the Alpaca dataset are set as baselines with basic instruction following ability.

Parrot-hint We set Parrot-hint (Jiao et al., 2023a) as our strong baseline. The Parrot-hint dataset includes 3 sub-datasets: the Alpaca Dataset, the WMT17-20 dataset, and the MQM instruction dataset. Parrot-hint contains 200K data in total.

OVERUNDER In the training process, we utilize Alpaca Dataset, the WMT17-20 dev sets in Parrot-hint sub-datasets, to ensure the basic ability of the fine-tuned models. The mixup dataset contains instruction-following data without a hint and with a hint, and data with a hint both have an auxiliary task based on translation. Therefore, we use a curriculum learning strategy to fine-tune the data in two stages. Note that the source of positive samples in OVERUNDER is also WMT17-20 dev sets.

5.2 Evaluation

This section introduces the test sets and the evaluation metrics we use.

WMT22 Test Sets WMT22 test sets come from the news translation track of WMT22 competition[6]. The test sets include 1984, 2037, 2037, and 1875 samples for De ⇒ En, En ⇒ De, En ⇒ Zh, and Zh ⇒ En, respectively.

Flores-200 Dev-test Flores-200 is a multi-language translation benchmark. We use the dev-test split as our test set, and there are 1012 samples for each translation direction.

Automatic Evaluation For lexical evaluation, we use BLEU (Papineni et al., 2002); for semantic evaluation, we use COMET with reference. Both

[6]https://github.com/wmt-conference/wmt22-news-systems

Model	De ⇒ En		En ⇒ De		En ⇒ Zh		Zh ⇒ En		Average	
	bleu	comet	bleu	comet	bleu	comet	bleu	comet	bleu	comet
WMT22 Winners										
	33.70	85.46	38.40	88.09	54.30	81.12	33.50	87.84	39.97	85.62
BLOOMZ-3b WMT22										
Alpaca	14.68	68.49	5.55	49.10	20.20	81.46	11.65	75.38	13.02	68.61
Parrot-hint	22.05	75.59	17.80	67.64	33.95	83.70	21.33	78.19	23.78	76.28
w/ SWIE	22.80	75.33	17.55	66.68	34.19	**84.13**	21.58	**78.50**	24.03	76.16
w/ OVERUNDER	22.97	75.37	18.59	69.12	35.05	82.90	21.69	77.83	24.58	76.31
w/ OVERUNDER w/ SWIE	**23.52**	**76.15**	**18.90**	**69.60**	**35.37**	83.66	**21.99**	78.11	**24.95**	**76.88**
BLOOMZ-7b1-mt WMT22										
Alpaca	18.64	73.37	9.97	61.65	25.52	82.31	15.07	77.79	17.30	73.78
Parrot-hint	23.80	77.77	20.58	73.63	35.49	84.61	22.58	78.93	25.61	78.74
w/ SWIE	**25.28**	77.91	19.86	72.93	36.76	**84.76**	22.96	**79.28**	26.22	78.72
w/ OVERUNDER	25.16	78.37	**21.61**	**74.84**	36.76	84.34	23.17	79.21	26.68	79.19
w/ OVERUNDER w/ SWIE	25.25	**78.52**	21.57	74.70	**37.13**	84.52	**23.36**	79.18	**26.83**	**79.23**
LLaMA-7b WMT22										
Alpaca	28.92	82.77	21.72	79.70	17.72	71.96	15.95	74.95	21.07	77.34
Parrot-hint	28.90	82.84	25.96	**82.78**	28.12	79.84	20.61	75.61	25.90	80.27
w/ SWIE	28.72	83.04	26.14	82.22	28.20	78.96	20.22	75.47	25.82	79.92
w/ OVERUNDER	29.27	83.37	**27.20**	82.55	30.26	**80.59**	21.20	**76.58**	26.98	**80.77**
w/ OVERUNDER w/ SWIE	**30.38**	**83.43**	27.10	82.09	**30.69**	80.20	**21.47**	76.50	**27.41**	80.56
LLaMA-7b Flores										
Parrot-hint	40.83	88.50	31.14	85.73	26.96	80.08	22.48	83.62	30.35	84.48
w/ SWIE	**40.88**	88.51	30.89	85.47	27.05	79.27	**22.76**	83.55	30.40	84.20
w/ OVERUNDER	39.57	88.51	32.19	**85.80**	28.73	**81.57**	21.24	83.57	30.43	84.86
w/ OVERUNDER w/ SWIE	40.21	**88.60**	**32.39**	85.78	**29.79**	81.51	21.29	**83.65**	**30.92**	**84.89**

Table 2: Translation performance of LLMs on WMT22 and Flores test sets. The **bolded** scores refer to the best performance under the same or comparable settings.

of them are widely used metrics in machine translation, and we use ScareBLEU[7] and Unbabel/wmt22-comet-da in the evaluation implementation.

5.3 Main Results

The main results are shown in Table.2. For models fine-tuned by Alpaca, the translation performance indicates the basic language ability of the model with an instruction-following format. Overall, we had the following main observations.

Firstly, according to the comparison between OVERUNDER and Parrot-hint, we found that OVERUNDER notably led to performance enhancement. Secondly, according to the comparison between SWIE and Parrot-hint, our method shows a significant 0.5 BLEU scores average improvement on BLOOMZ-7b1-mt, and steady improvements in BLOOMZ-3b can also be seen. Thirdly, by combining the OVERUNDER and SWIE, a further improvement also can be seen in all of the backbones. By combining the dataset and model, we can see further improvements in all of the backbones. When compared with the baseline, there are noticeable enhancements in the overall translation. BLOOMZ-3b has a 1.16 BLEU improvement, while BLOOMZ-7B and LLaMA-7b have 1.22 BLEU and 1.51 BLEU improvements, respectively. On the Flores test set, the combination of OVERUNDER and SWIE also shows the best overall performance in the ablation experiments.

To verify the robustness of SWIE, we have attached the results of the ablation and sensitivity experiments for the layer selection and weight functions in Appendix Section D. Moreover, the experiments in parameter-efficient LoRA setting (Section F) and the significance test (Section E) in various sentence length settings are also provided in the Appendix.

6 Analysis

6.1 In-Context Translation

To evaluate the effectiveness of SWIE in the long-instruction scenario and extend the evaluation to a widely used scenario for LLMs, we conduct experiments on in-context translation with translation

[7] https://github.com/mjpost/sacrebleu

demonstrations. We follow the settings in the main experiments, use WMT22 De⇒En as the test set, and use BLOOMZ-3b as the backbone model. The translation demonstrations are sampled randomly from the Flores test set and concluded in the instruction part. When using 20 translation demonstrations, the token lengths range from 1820 to 2600. Based on the model inference results, we observed that adding in-context demonstrations destroys the translation performance without SWIE because the model forgets the translation instruction and generates irrelevant content. According to the results in Table. 3, the model translation performance decreases with the demonstration numbers getting higher. Meanwhile, SWIE notably leads to BLEU improvements from 7 to 10, which indicates the potential of SWIE for in-context learning scenarios.

N_{demo}	2	5	10	20
OVERUNDER	13.51	7.73	4.58	0.49
OVERUNDER W/ SWIE	**20.28**	**17.82**	**15.16**	**8.63**

Table 3: BLEU scores in in-context translation. N_{demo} means the number of demonstrations.

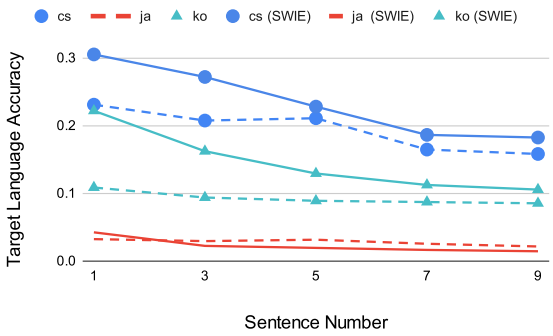

Figure 3: Comparison accuracy for zero-shot translation directions between models with and without SWIE, and cs, ja, and ko representing Czech, Japanese, and Korean, respectively.

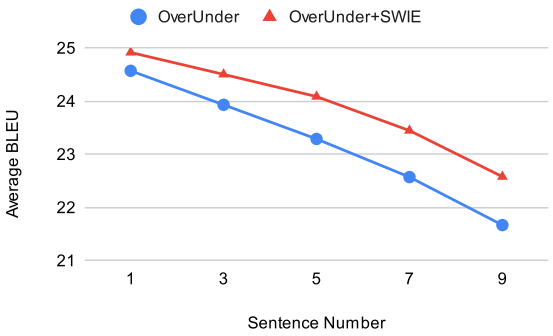

Figure 4: Average BLEU on WMT22 test sets (En ⇔ De, En ⇔ Zh) with different concatenation sentence numbers.

6.2 Long Context Zero-shot Translation

To prove the instruction-forgetting phenomenon, we designed a zero-shot translation direction experiment to determine the relationship between translation quality and the distance between translation outputs and instruction.

We use the Flores test set, which includes about 200 languages, and we select three low-resources contained in the Flores test set for testing. The original test set has 992 sentences for each translation direction. We expanded the test set to multi-sentence (3/5/7/9) using a sliding window and splicing nearby sentences, with the final expanded test sets approximating the original number. The input with 9 sentences has around 1k tokens.

As Figure 3 shows, with the extended test sets sentences getting longer, the worse performance can be seen in the translation accuracy. This phenomenon indicates that the instruction information will be weakened by a longer context. Compared with the Parrot-hint, our method shows much higher accuracy in Czech and Korean for all sentence number settings (*e.g.* 7% and 12% for Czech and Korean in one sentence testing, respectively). Both of the two settings can rarely recognize Japanese.

6.3 Long Context Translation

Similar to the data processing method in Section 6.2, we extend the WMT22 test set to multi-sentence(3/5/7/9) by contacting the nearby sentences with a sliding window, and the input with 9 sentences has around 1k tokens. We compare the original BLOOMZ-3b fine-tuned on OVERUNDER and BLOOMZ-3b with SWIE fine-tuned on OVERUNDER. And the translation results can be seen as Figure 4. We can observe that with the source sentence getting longer, SWIE shows consistently higher translation performance advance.

6.4 Faithfulness Evaluation

In human evaluation, we follow the evaluation setting of (Weng et al., 2020). We simplify the error division and narrow the range to faithfulness problems (including over-translation and under-translation mistakes). Then, we define the faith-

fulness error degree and the corresponding error scores "No Error", "Minor", and "Major" as 0, 1, and 2, respectively. We sample 100 test results from the Zh⇒En test set and engaged three native language speakers, who are undergraduate students, compensating them at a rate of $10 per 1000 sentences to assess the error degree of the output data. Final labels were based on majority voting. The statistical result shows that both SWIE and OVERUNDER have a lower faithfulness error score, and their combination decreases the overall error score to nearly half of the baseline.

We also provide a statistical faithfulness evaluation based on the word-alignment toolkit, which is appended in Section G.

setting	minor↓	major↓	error score↓
Parrot-hint	0.06	0.03	0.12
w/ SWIE	0.09	0.01	0.11
w/ OVERUNDER	0.06	0.02	0.10
w/ SWIE w/ OVERUNDER	**0.05**	**0.01**	**0.07**

Table 4: The human evaluation of translation faithfulness error rate on SWIE and OVERUNDER.

6.5 Visualize Inadequate Attention on Instruction

Our standard instruction-following data item is sequentially organized as instruction, input, and output. The attention score in transformers can show the positions the model addresses more. We sample 20 random translation examples from test sets and report mean results. According to the observation in Figure 1, the attention scores on the ending tokens of instruction and input can represent a global feature of attention on the corresponding spans. Therefore, we simplify the visualization by calculating attention distributions on the special tokens at the end of each span.

Assuming a is the attention score matrix, s is a span belonging to $S = \{s_{ins}, s_{input}, s_{output}, \cdots\}$, the e_s is the special token index of the end of the span s, and the T is the length of instruction data token list. We use C_s to represent the accumulated attention score in a position as shown in Equation 6.

$$C_s = \sum_{i=e_s+1}^{T} a[i][e_s] \tag{6}$$

As depicted in Figure 5, it is evident that the middle layers of the model manifest a considerably higher attention accumulation score on the input spans, whereas the bottom and top layers exhibit more balanced attention on instruction. The phenomenon indicates that the model concentrates on the translation task more on the middle layers. We compute the ratio of the attention score at the ending position of the instruction and the attention score at the ending position of the input. As illustrated in Figure 6, our method leads to a much higher attention ratio on instruction in most layers, implying that the SWIE effectively improves the model performance via enhancing instruction attention.

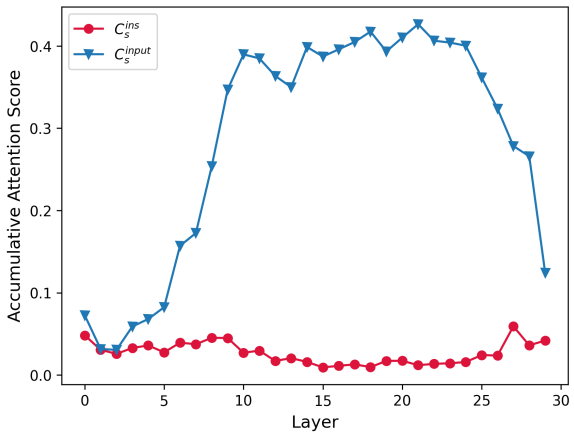

Figure 5: Accumulative attention scores of instruction and input spans on each layer. This figure is based on BLOOMZ-3b, fine-tuned by the Parrot-hint dataset in the origin model structure.

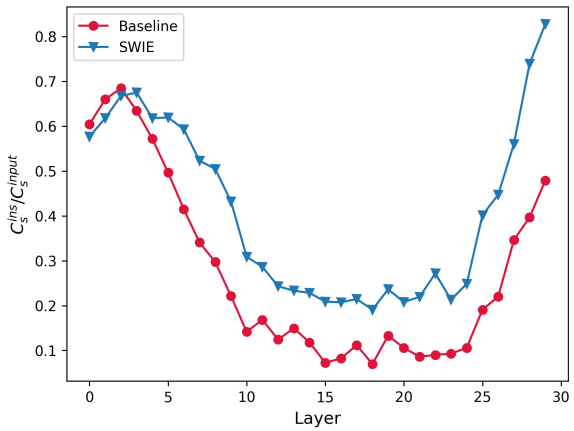

Figure 6: The comparison between models with and without SWIE on attention ratio between the attention accumulation score on instruction C_s^{ins} and the attention accumulation score on input C_s^{input}. This experiment is based on BLOOMZ-3b.

6.6 Instruction-Following Ability Evaluation

We conduct experiments on an instruction-following test set IFEval (Zhou et al., 2023). IFE-

setting	prompt-level	instruction-level
Alpaca	0.1164	0.2110
Alpaca-SWIE	**0.1275**	**0.2230**

Table 5: Instruction following ability evaluation on IFEval test set.

val contains 541 instruction-following test cases and focuses on verifiable instructions, including response word count, keyword frequency, etc. The results in Table 5 show that SWIE strengthens the instruction-following ability, demonstrating the generalization ability of SWIE.

7 Related work

7.1 Instruction Tuning and Variant Methods

Instruction fine-tuning has shown surprising generalization ability on different tasks (Wei et al., 2021; Honovich et al., 2022; Wang et al., 2022). In the context of instruction tuning LLMs for machine translation, Jiao et al. (2023a); Zhang et al. (2023) have proposed multi-task instruction data construction frameworks for instruction tuning open-source LLMs on machine translation. Zeng et al. (2023) proposed a contrastive learning loss to train the model to learn contrastive sample pairs.

On general tasks, existing works are proposed to add instruction or context learning objections to improve instruction fine-tuning generalization ability and performance. Choi et al. (2022) proposed a distilling-based context injection method to preserve the long context information in the fixed model when the model is used in static long prompts situations. Ye et al. (2022) models the instruction in the condition given input and target for tasks with fixed labels. Snell et al. (2022) distills context like task explanation or step-by-step reasoning from the teacher model. Ge et al. (2023) compress the long context into an adding memory slot module for in-context learning.

The above methods focus on diverse or complex instruction modeling but do not stress the risk of instruction forgetting under the premise of position independence and without requiring fixed instructions.

7.2 Translation Faithfulness in Language Models

Faithfulness (also called hallucination) in neural machine translation has been discussed for a long time (Lee et al., 2018; Müller et al., 2020). It is widely observed that the sources of unfaithfulness can be the lack of knowledge or inadequate attention to the source (Ferrando et al., 2022; Raunak et al., 2021). On machine translation hallucination detection benchmarks, existing datasets are constructed by humans or perturbing the translation model (Raunak et al., 2021). Human-making datasets like HalOmi (Dale et al., 2023) are costly and hard to scale up. Datasets generated by the model perturbing method are low quality because the sentences generated are far from the natural text style and the distribution of modern LLMs. Thus, our proposed unfaithful-translation-mimicking dataset construction method can fill the gap with high-quality and fluent negative samples.

8 Conclusion

We proposed SWIE and OVERUNDER, a novel additional model structure for strengthening the attention of the model to instruction, and an effective data construction method for machine translation faithfulness. The experiments on various backbone models and test sets show the effectiveness of SWIE and OVERUNDER in translation quality and faithfulness. The zero-shot long-context translation direction experiment indicates that the origin model structure shows weaker instruction following ability with the input text getting longer, and SWIE alleviates instruction forgetting in different input length settings. Furthermore, the long-context translation experiment shows the SWIE outperforms the corresponding baseline more obviously in a longer input setting. Through the internal attention scores of the models, we visualize the attention distribution on the original model and the attention shift induced by SWIE, thereby corroborating our assumption regarding the necessity for increased attention on instruction. The experiments on the IFEval instruction-following dataset indicate that SWIE also improves the models on general instruction-following tasks. Overall, SWIE effectively mitigates the instruction forgetting issue and enhances both translation quality and faithfulness. Its wide effectiveness in various scenarios and settings indicates the considerable potential of SWIE.

The following aspects can be explored in the future based on our work: (1) investigating explainable and trainable methodologies for constructing segment-weight and (2) extending the data construction method to other tasks.

9 Limitations

Our work focuses on improving the translation faithfulness of LLMs, but there are the following limitations. Firstly, the diversity and scale of the datasets and models in the training process are limited due to the computational resource requirements. Consequently, it remains uncertain whether scaling up the instruction fine-tuning process would unlock greater potential or uncover additional phenomena. Secondly, the SWIE induces approximately 20% inference latency, indicating the potential to boost the efficiency of the method in future work.

References

Eunbi Choi, Yongrae Jo, Joel Jang, and Minjoon Seo. 2022. Prompt injection: Parameterization of fixed inputs. *arXiv preprint arXiv:2206.11349*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *arXiv preprint arXiv:2305.11746*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Javier Ferrando, Gerard I Gállego, Belen Alastruey, Carlos Escolano, and Marta R Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769.

Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

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. *arXiv preprint arXiv:2302.09210*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023b. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. *NIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context. *arXiv preprint arXiv:2209.15189*.

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.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Seonghyeon Ye, Doyoung Kim, Joel Jang, Joongbo Shin, and Minjoon Seo. 2022. Guess the instruction! flipped learning makes language models stronger zero-shot learners. *The Eleventh International Conference on Learning Representations*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

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

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Details of OVERUNDER

Instruction-tuning datasets can be organized flexibly, and the standard format contains instruction, input, and output. After we constructed the over-translation and under-translation contrastive samples based on the WMT17-20 dev set (the data source is the same as the setting in Parrot (Jiao et al., 2023a) with the proposed automatic pipeline, we organized the final instruction data as Figure 7. The total number of samples in the dataset is 54,420.

B Implement details

We use the transformers and DeepSpeed[8] framework for model training and inference. The training hyper-parameters follow the setting of (Jiao et al., 2023a), and we report the results of the best checkpoints within 1.5 epochs. We uniformly set the dim of the instruction adapter to 32 and select the 5th, 6th, and 7th layers to add SWIE. The 3B size models are trained on 8 V100 GPUs, and the 7B size models are trained on 4 A100 (40G) GPUs. We trained all models in DeepSpeed stage 1 with freezing embedding layers to reduce the memory requirement and prevent the models from over-fitting.

C Training Cost Analysis

We use the same device (V100-32G) to train BLOOMZ-3b. The adapter parameters are only 0.02% of the full model, and the train samples per second of SWIE is a 25% decrease compared with the baseline.

setting	parameter size	train samples per second
SWIE	2,360,793,702	30.629
SFT	2,360,294,400	38.384

Table 6: Training cost comparison of SWIE and standard supervised fine-tuning on BLOOMZ-3b.

D Ablation Study

D.1 The Impact of the selected layers

The layer selection is a possible variable on the final model effect and the inference latency. In the

[8]https://github.com/microsoft/DeepSpeed

primary analysis, we select the 5th, 6th, and 7th layers, that is, the bottom three layers of the model. We conducted the sensitivity experiments for layer selection on BLOOMZ-3b, which contains 30 layers in total. We fix the layer number to 3 according to the trade-off for training and inference cost and the model performance, and the results in Table 7 indicate that the selection of top, middle, or bottom three layers is not sensitive for the final overall result. However, adding adapters for all layers shows an obvious decrease, which could be caused by the higher difficulty for a model to learn new features for every layer compared with certain three layers.

layer selection	BLEU(mean)	COMET (mean)
5-7	24.92	76.76
14-16	24.95	76.88
26-28	24.97	76.77
all layer	24.48	76.30

Table 7: Layer sensitivity ablation study.

D.2 The Comparison of Constant Weight and Segment-Weight

As shown in Table.8, we compare the performance on Parrot-hint for weight setting and keep the other settings the same as the main experiments on four language directions (En⇔De and En⇔Zh) of the WMT22 test set. The constant weight setting keeps the same upper bound as the Sigmoid weight. A significant decrease of 0.4 scores on the mean of BLEU and COMET indicates our hypothesis on the necessity of Sigmoid weight.

layer selection	BLEU(mean)	COMET(mean)
Sigmoid	24.03	76.16
Const	23.62	75.79

Table 8: Segmoid weight and constant weight comparison.

E Significance Test

We conduct a significance test to ensure our experiments are significant with random settings. We choose 5 random seeds as initials, including 1, 6, 19, 42, and 3307. The following experiments were conducted using the same settings in Figure 4 and are based on BLOOMZ-3b. As shown in Table 9, the p-values on all sentence length settings are below 0.05, indicating the effectiveness of SWIE is statistically significant.

$N_{sentence}$	1	3	5	7	9
p-value	4.5e-2	1.0e-2	2.3e-2	4.3e-2	4.8e-2

Table 9: This table presents the statistical analysis on the BLEU scores of the experiment in Figure 4, where $N_{sentence}$ means the concatenation number of sentences in the test set.

F SWIE with LoRA

To expand SWIE in light-weight adapter settings, we also provide the experiments of SWIE with LoRA. The hyperparameters related to LoRA follow the setting in Jiao et al. (2023a), and the other settings follow the main experiments. The experimental results in Table 10 show that the performance of SWIE combining LoRA in all length settings is consistently higher than using only LoRA. Meanwhile, the SWIE only increases 1/8 trainable parameters in LoRA settings, maintaining high training efficiency.

$N_{sentence}$	1	3	5	7	9
OVERUNDER	18.61	17.73	16.51	14.85	13.33
OVERUNDER w/ SWIE	18.72	18.04	16.78	15.20	13.53

Table 10: The average BLEU scores for models with and without the incorporation of SWIE under the LoRA setting, where $N_{sentence}$ means the concatenation number of sentences in the test set.

G Statistical Faithfulness Evaluation

There is no widely used standard toolkit on the qualification of word-level machine translation faithfulness yet, so we both conduct statistic-based and human evaluations. The same statistic-based method as Section 4, we use word alignment tools to match the source sentences and the inference sentences word by word, then calculate the recall of source words matching rate and hypothesis words matching rate, and then the ratio can reflect the absence and the redundancy extent. The final scores are derived by averaging the source and target coverage rate on our WMT22 test sets. The result shows in Table.11 that both SWIE and OVERUNDER can improve the faithfulness of results, showing the effectiveness of our proposed method.

setting	score
Parrot-hint	87.94
w/ SWIE	88.28
w/ OVERUNDER	**88.84**
w/ SWIE w/ OVERUNDER	88.80

Table 11: The ablation study of faithfulness score on SWIE and OVERUNDER.

Figure 7: An instance of translation instruction and an instance of OVERUNDER. A faithful translation is given first, then two bad translations with over-translation or under-translation problems are followed.

H Case Study

We select 2 typical hard translation samples which will easily lead to hallucination. In the cases in Table 12, the bold text indicates the better translation part in the generated text. As demonstrated, the model incorporating SWIE exhibits superior performance in terms of semantic faithfulness.

id	type	content
1	reference	In which, “harboring generosity like the ocean” emphasizes the spirit of inclusiveness, reflecting the mindset of “inviting all talents”, which is manifested as absorbing the premium concepts of university operation and talent training of world famous universities; “seeking the truth while upholding morality” lays emphasis on the spirit of the pursuit for truth and kindheartedness, meaning that universities shall accept the cultivation of virtue and the tempering of personality while pursuing for truth; “being brave enough to keep exploring” reflects the spirit of reform and innovation, including innovation in knowledge and science and technology, as well as reform, development and innovation; “building our country” stresses the spirit of patriotism and dedication, meaning that universities shall always share the glory and disgrace with the destiny of the country.
	w/ OVERUNDER	Among them, "Haining River" emphasizes the spirit of openness and inclusiveness, reflects the spirit of "using talent from all over the world", and manifests the high-quality concept of absorbing world-class universities' excellent ideas for education and training; "Qianzhen Zhu De" emphasizes the spirit of seeking truth to the best spirit, namely, the university should pursue the pursuit of truth while receiving the cultivation of virtue and moral cultivation, and reform and innovation; "Xuanqiang" reflects the spirit of reform and innovation, including the pursuit of knowledge and scientific innovation, as well as the reform and development and革故鼎new.
	w/ OVERUNDER w/ SWIE	Among them, "Hainan River" emphasizes the spirit of openness and inclusiveness, reflecting the "heart of using all the talents in the world", and showing the absorption of the high-quality concept of the education and training of world famous universities; "Xuanzhen and De" emphasizes the spirit of seeking truth to the best, which means that while universities pursue truth, they should also receive the cultivation of virtue and the cultivation of personality; "opening up before the people" reflects the spirit of reform and innovation, which includes the pursuit of knowledge and innovation in science and technology, as well as the reform and development, **reform and innovation; "building my country" emphasizes the spirit of patriotism and dedication, which means that universities should always share the fate of the nation.**
2	target	[Nora Books] Authentic, 3 volumes, Roald Dahl Complete Collection, Charlie and the Chocolate Factory + Matilda + Boy: Tales of Childhood, 6-12 years old children’ s book authentic children’ s literature
	w/ OVERUNDER	Nora Book has 6-12 year old children's books for reading children's literature.
	w/ OVERUNDER w/ SWIE	There are three books in the original edition of "Nora Book" and the whole collection of **"Charlie and the Chocolate Factory", "Matilda" and the childhood stories of "Good Boy"** for children aged 6 to 12.

Table 12: Case study for comparison on the faithfulness improvement of SWIE. The **bold** parts mean the more faithful spans of generated translation.

type	prompt
under-translation	You are an unprofessional [source language] to [target language] translator who is not fully faithful to the original text in the translation process there is a problem of omission, *i.e.*, the translation leaves out parts of the original text. Please translate the following [source language] sentence: [source sentence] If the following is a high-quality human [target language] translation: [target sentence] Please give a direct low-quality [target language] translation with omission problems, noting that you are not simply rewriting the previous translation, but need to emulate a translator that may have omissions, *i.e.*, omitting parts of the original text.
over-translation	You are an [source language] to [target language] translator, but your translation is unprofessional. In the translation process, you have not been completely faithful to the original text, resulting in a translation that is not in the original text. This is a translation illusion problem; you need to provide a translation with the illusion problem. Please translate the following [source language] sentence: [source sentence] If the following is a high-quality human [target language] translation: [target sentence] Please give a straightforward and low-quality [target language] translation with an additive or a translation illusion problem. Please note that you need to simulate a translator with possible translation enhancement problems and translate what is not in the original text, rather than simply rewriting the previous translation.

Table 13: The prompts for producing the OVERUNDER dataset.