Beyond English-Centric Machine Translation by Multilingual Instruction Tuning Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance on Machine Translation (MT) among various natural languages. However, many LLMs are English-dominant and only support some high-resource languages, they will fail on the non-English-Centric translation task. In this work, we propose a Multilingual Instruction Tuning (MLIT) method to improve the LLMs on non-English-Centric translation. We design a multilingual instruction method which leverage the English sentence as reference to help LLMs understand the source sentence. In order to solve the problem of difficulty in obtaining multilingual parallel corpora of low-resource languages, we train a to-English LLM to generate English reference so that our MLIT method only needs bilingual data. We experiment on LLaMA2 foundation and extensive experiments show that MLIT outperforms the baselines and some large-scale language models. We further demonstrate the importance of English reference in both training and inference processes.

1 Introduction

Large language models (LLMs) have shown remarkable achievement across various NLP tasks (Brown et al., 2020; Ouyang et al., 2022; Zhang et al., 2022). For machine translation, generative LLMs achieve a competitive translation quality, especially on these high-resource language pairs (Hendy et al., 2023; Vilar et al., 2022). The models can be prompted to do so by designing a prompt such as "Translate the following sentence from French to English".

However, most of the existing LLMs are English-dominant. They only support several high-resource natural languages. For example, LLaMA (Touvron et al., 2023) covers 20 languages, BLOOM (Workshop et al., 2022) supports 46 languages, and GLM (Du et al., 2022; Zeng et al., 2022) only supports English and Chinese. So they still fall short for non-English-Centric language translation. Even these very large models such as GPT-3.5 cannot rival the traditional supervised encoder-decoder state-of-the-art (SoTA) models (Hendy et al., 2023; Zhang et al., 2023a; Jiao et al., 2023). Obviously, a large population in the world cannot be benefited. As shown in Figure 1, even ChatGPT (OpenAI, 2022) will make some mistakes on non-English translation directions.

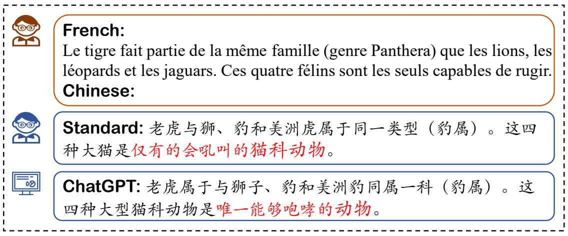


Figure 1: The results of standard output and ChatGPT output on French-to-Chinese translation. The general meaning of the translation is correct. However, ChatGPT makes logical mistakes in the red part. The red part of standard answer is "the only **catamount** that roars", but the ChatGPT translation is "the only **animal** that roars".

To equip LLMs with much more multilingual ability, we propose a Multilingual Instruction Tuning (MLIT) method to fine-tune LLMs. Our method focuses on non-English translation task. We design a multilingual instruction which includes the source language, target language and English to fine-tune LLMs. In this way, these English-dominant models can better understand the translation sentence based on the English reference, and transfer the knowledge from English to other languages.

Specifically, our MLIT method is consisting of three steps. First, we train a to-English LLM to generate English sentence based on the source sentence. In the second step, we design a multilingual instruction (X-En-Y, where X represents the source language and Y represents the target language) based on parallel sentences to train a non-English-Centric translation model. Finally, we leverage the to-English model to generate English reference and then predict target sentence based on the non-English-Centric model. We evaluate our method on both low-resource and high-resource language pairs based on LLaMA foundation. Our MLIT method achieves better results on all test sets.

In summary, this paper makes the following contributions:

- We propose a Multilingual Instruction Tuning (MLIT) method to fine-tune the LLMs on non-English machine translation task. We add the English sentence to instruction as reference in order to transfer knowledge from English to other languages. MLIT method improves the capability of low-resource translation.
- We solve the problem of difficulty in obtaining multilingual parallel sentences of low-resource languages. Our framework only uses 1K bilingual sentences of source and target languages. We train LLMs to generate other languages' instruction to build the multilingual instruction instead of leveraging multilingual parallel data.
- We propose a framework which can be applied on many foundation models. Extensive experiments show that our method has a significant improvement over all test pairs and even outperforms some large-scale models.

2 Background

2.1 Machine Translation for Low-Resource Languages

With the development of large-scale language modeling techniques, LLMs have achieved remarkable improvements in machine translation (Kim et al., 2021; Costa-jussà et al., 2022). They have opened up new possibilities for building more effective translation systems (Brown et al., 2020; Chowdhery et al., 2023; Sanh et al., 2022). However, due to the unbalanced training resources, most of these models focus on high-resource languages. Low-resource machine translation have attracted a lot of attention (Haddow et al., 2022; Ramesh et al., 2022). While most of these focus on translations on English-Centric languages (between English and other languages). Fan et al. (2021) emphasizes the importance on improving translation among non-English languages.

2.2 Cross-Lingual Method for LLMs on Machine Translation

Large language models (LLMs) can be prompted to perform very high-quality machine translation. It is assumed that the model is pretrained on enough training data in both source and target languages. However, most LLMs is trained primarily on English data. When it comes to low-resource languages, the model struggles to output high quality translations (Koehn and Knowles, 2017). Lu et al. (2023) proposed a novel framework, Chain-of-Dictionary (CoD), which augments LLMs with prior knowledge with the chains of multilingual dictionaries for a subset of input words. Ghazvininejad et al. (2023) proposed a method for incorporating dictionary knowledge into prompting-based MT (DIPMT). Their prompt is designed as follows:

> Translate the following sentence to English:
> <*source-sentence*>
> In this context, the word <*word X in source-language*> means <*word X in target-language*>; the word <*word Y in source-language*> means <*word Y in target-language*>.
> The full translation to English is:

Jiao et al. (2023) proposed a pivot prompting method for distant languages, which asks LLMs to translate the source sentence into a high-resource pivot language before into the target language, improving the translation performance noticeably:

> Please provide the <*pivot-language*> translation first and then the <*target-language*> translation for this sentence:
> <*source-sentence*>

Nearly all the existing LLMs have a strong capability on English and get weaker on other languages. Most of the methods concentrate on English-Centric machine translation and prompting method, ignore the non-English-Centric translation. In this paper, we will improve the LLMs' ability on non-English-Centric translation through our multilingual instruction tuning method with the help of a small amount of bilingual data.

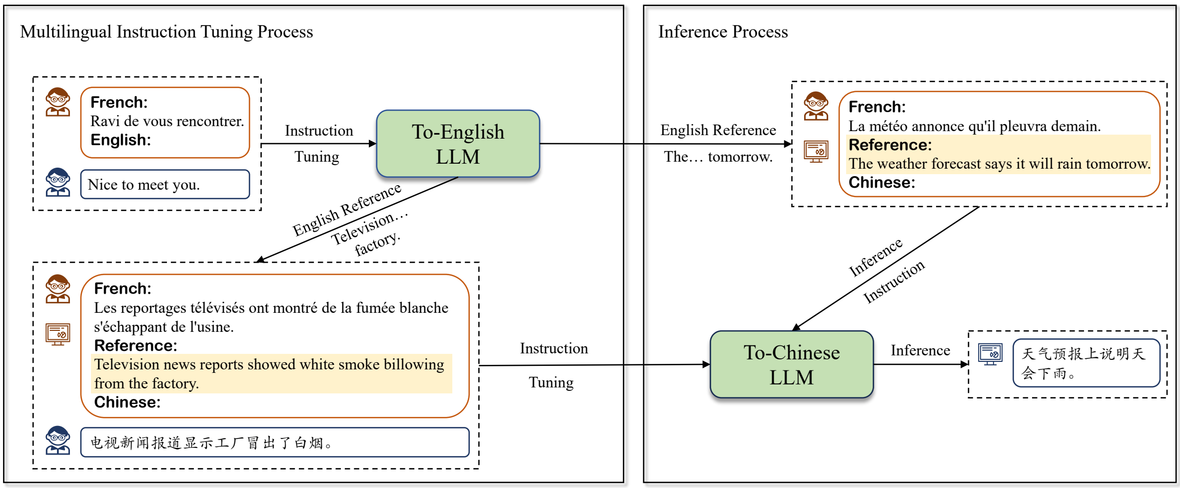


Figure 2: The main framework of our proposed method. Multilingual Instruction Tuning (MLIT) process contains two parts. First, we train a to-English LLM based on the bilingual instruction. Then we generate English reference and combine them with the bilingual sentence as the multilingual instruction. The inference process leverage to-English LLM generate the English reference and transfer it with the source sentence to Multilingual Instruction Tuned model to generate the corresponding translation.

3 Methodology

In this section, we introduce the details of our Multilingual Instruction Tuning (MLIT) method. We first introduce the format of instruction. Then we show the two components of MLIT: to-English translation model in Section 3.2 generates English reference for training and inference processes. MLIT method in Section 3.3 trains the LLMs with multilingual instruction. Finally, we introduce the way to predict target sentence in Section 3.4. The framework of our method is shown in Figure 2.

3.1 Instruction Design

Due to the strong capabilities of existing large language models on English, we still choose the English instruction for training. We have experimented with various forms of instruction, and the results show that the simplest form of prompt has the best effect. The complex instruction, such as *"Translate the following sentence from French to Chinese."*, may affect translation abilities of LLMs. The format of our instruction is as follows:

> Human:
> <source-language>: <*source-sentence*>
> Reference: <*English-sentence*>
> <target-language>:
> Assistant:
> <*target-sentence*>

We leverage the parallel sentences of <*source-language*> and <*target-language*> to generate the instruction for non-English-Centric translation. As for the English reference, we train a model to generate based on the <*source-sentence*>. As shown in Figure 2, the orange part denotes the instruction of Human, and the blue part denotes the instruction of Assistant.

3.2 To-English Translation Model

To-English translation model aims to generate the English instruction as reference in our multilingual instruction. Let L_s and L_e represent source language and English, S_s represents the source sentence. We leverage bilingual parallel sentence with the format in Section 3.1 to train this model, just as shown in Figure 2. The formulation can be expressed as follows:

$$S_e = \arg\max p_\theta(e_1, e_2, ...|L_s, L_e, S_s) \quad (1)$$

where S_e denotes the English sentence, e_i denotes the ith generated English token, p denotes the probability of the generation model and θ denotes the parameter. We evaluate the impact of the quality of generated English sentences on subsequent training and inference.

3.3 Multilingual Instruction Tuning

After achieving the to-English model, we further propose the Multilingual Instruction Tuning

(MLIT) method to train the non-English translation model.

Specifically, we want to use the strong capability of large language models' ability in English to help the LLMs understand sentences in other languages, so as to achieve a better performance on the non-English translation task. To do this, based on the original bilingual parallel instruction, we add the English reference to build the multilingual instruction. However, we only use the bilingual sentence S_s and S_t of the source and target language, L_s and L_t. We leverage the to-English translation model in Section 3.2 to generate the corresponding English sentence S_e of the source sentence. With this approach, we get multilingual instruction and then use them for the training step, just as shown in the left part of Figure 2. Formally, the MLIT method is determined as:

$$S_t = \arg\max p_\theta(t_1, t_2, ...|L_s, L_t, S_s, S_e) \quad (2)$$

where t_i denotes the ith generated token of target sentence.

3.4 Inference

After the Multilingual Instruction Tuning Process, we finally leverage the two LLMs in Section 3.2 and 3.3 to predict the target sentence. Specifically, we first generate the English reference based on the source sentence using the to-English translation model. Then we combine the source sentence and English reference to non-English-Centric translation and infer the target sentence. The inference process is similar to the form of Eq. 2. However, compared with the training process, the quality of English reference has a greater impact on the inference process. We will prove this in Section 4.5.

4 Experiments

4.1 Settings

Datasets. To assess the effectiveness of our proposed model on machine translation, we conduct evaluations usings the devtest subset of the FLORES-200 dataset (Costa-jussà et al., 2022). For each language, it contains 1012 parallel sentences encompassing various fields and topics. We choose 8 language pairs for to-Chinese translation and 5 language pairs for to-French translation, which contains both high-resource and low-resource languages, to evaluate our method.

Implementation Settings. We select a representative and common open source large language model as our foundation models for our study: Atom[1]. Specifically, we experiment on the Atom-7B scale model, which is based on the LLaMA2 (Touvron et al., 2023). We leverage the dev subset of the FLORES-200 dataset for training. Specifically, we leverage the source-English parallel data to train To-English translation model. Then we combine the source-target parallel data and the generated English sentence by To-English translation model based on source sentence to train the MLIT-trained LLM. All the two training processes are full fine-tuned and conducted on 4 A100 GPUs with 40GB of RAM for 6 epochs. And the inference processes are conducted on 1 A100 GPUs with 40GB of RAM costing 20 minutes (1012 pieces of data).

Baselines. For our foundation models, we leverage the bilingual instruction to train our baseline. Besides, we choose four mostly used instruction methods for machine translation to evaluate: 1) Chain of Thought. 2) Mixed Instruction. 3) Chained Multilingual Instruction. 4) Pivot Prompting (we use a two-step pivot-based mehtod, first train a source-English model, and then train an English-target model). The other format of instruction is appended in Appendix A. Meanwhile, we compare our method with BigTranslate[2] (Yang et al., 2023), which is a multilingual translation model that enhances the LLaMA with multilingual translation capability on more than 100 languages. Besides, BayLing[3] (Zhang et al., 2023b) has a good multilingual capability, we choose its 13B version to compare. Meanwhile, we evaluate the performance of ChatGPT (OpenAI, 2022) (we use gpt-3.5-turbo API). For all the open-source LLMs, we execute their publicly accessible prompt or the same prompt as our method to acquire the baseline findings. As for ChatGPT, we evaluate it with 11 kinds of prompts and choose the best score, the prompts are appended in Table 5.

4.2 Main Results

Table 1 presents the results in chrF++ and COMET system[4] (Rei et al., 2022) on FLORES-200 dataset for translating from 8 source languages to Chinese. Our method is based on Atom-7B foundation. We compare our method with four instruction tuned

[1] https://github.com/FlagAlpha/Llama2-Chinese
[2] https://github.com/ZNLP/BigTranslate
[3] https://github.com/ictnlp/BayLing
[4] model: Unbabel/wmt22-comet-da

model	fr	de	es	id	ro	ru	ja	th	avg
chrF++									
BigTranslate-13B(Yang et al., 2023)	17.6	17.1	17.5	12.3	17.3	15.7	13.6	2.8	14.2
BayLing-13B(Zhang et al., 2023b)	20.5	19.9	19.5	17.6	21.0	17.4	6.6	3.1	15.7
ChatGPT(OpenAI, 2022)	**24.4**	**24.4**	22.5	24.0	**23.9**	**22.7**	**20.8**	**18.3**	**22.6**
Atom-7B+Bilingual Instruction Tuning	21.8	21.8	20.6	21.2	21.2	21.0	18.6	12.3	19.8
Atom-7B+Chain of Thought	17.9	19.2	19.0	19.4	20.5	21.3	17.2	11.8	18.3
Atom-7B+Mixed Instruction	15.0	14.9	15.7	17.2	15.5	15.6	13.9	9.8	14.7
Atom-7B+Chained Multilingual Instruction	15.8	17.4	16.9	19.0	18.2	19.8	14.3	10.0	16.4
Atom-7B+Pivot Prompting	22.4	22.4	22.0	22.7	22.2	22.1	17.3	12.1	20.4
Atom-7B+MLIT	24.1	22.0	**23.8**	**26.1**	23.0	23.6	19.0	12.8	21.8
COMET									
BigTranslate-13B(Yang et al., 2023)	0.76	0.76	0.76	0.71	0.65	0.72	0.52	0.48	0.67
BayLing-13B(Zhang et al., 2023b)	0.75	0.76	0.76	0.72	0.68	0.75	0.56	0.49	0.68
ChatGPT(OpenAI, 2022)	**0.82**	0.82	**0.81**	**0.78**	**0.82**	**0.83**	**0.76**	**0.74**	**0.80**
Atom-7B+Bilingual Instruction Tuning	0.77	0.79	0.78	0.73	0.69	0.74	0.57	0.48	0.69
Atom-7B+Chain of Thought	0.69	0.71	0.70	0.66	0.67	0.73	0.57	0.45	0.65
Atom-7B+Mixed Instruction	0.66	0.66	0.67	0.60	0.63	0.70	0.53	0.45	0.61
Atom-7B+Chained Multilingual Instruction	0.64	0.67	0.70	0.64	0.66	0.70	0.54	0.46	0.63
Atom-7B+Pivot Prompting	0.74	0.80	0.78	0.75	0.73	0.74	0.60	0.50	0.71
Atom-7B+MLIT	0.81	**0.83**	**0.81**	0.75	0.77	0.77	0.60	0.53	0.73

Table 1: Main results of MLIT method in chrF++ and COMET system for MT on the FLORES-200 dataset. We experiment on the **to-Chinese** translation task based on Atom-7B foundation. The "underline" signifies the better score between all the baselines. The "**bold**" indicates the best score among all the test set of each language pairs.

baselines, pivot prompting model and some large scale language models on both high-resource and low-resource languages. Compared with all the baselines, the results show that our MLIT method achieves better results among all the language pairs, and the improvement is more significant on high-resource languages.

As for the instruction tuned baselines, the Bilingual Instruction Tuning (BIT) method and Pivot Prompting method achieve better results. We think, compared with other baselines, they have simpler forms which is more suitable for small-scale models. Besides, compared with the COT, CMI and Pivot Prompting baselines, our MLIT method do not directly leverage English generated by the original model. We train a to-English model for generation which reduces the noise caused by the quality of the generated English, which can be proved in Section 4.5. In this way, our method achieves better results than the baselines.

As depicted in Table 1, compared with the large scale language models, our method achieves better results (achieving improvements of 6.1% and 0.05 on the BayLing-13B model). Meanwhile, the performance of our model is close to ChatGPT and even exceeds its performance on two language pairs. Besides, the results show that the large scale models have similar performance among all the languages on non-English translation task which demonstrates the robustness of large-scale models.

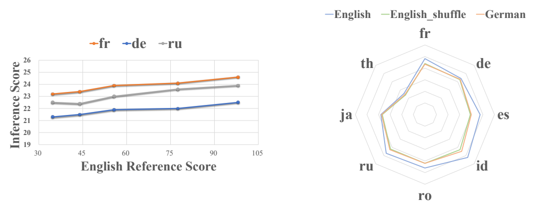
4.3 Translation to High-Resource Language

The results in Section 4.2 show the significant improvement on to low-resource translation. In this section, we demonstrate the robustness of our approach on to high-resource translation compared with the baselines and some state-of-the-art translation models. We report the results on to-French translation in Table 2. The results show that MLIT method achieves better scores on both foundations (with 2.4% and 0.03% improvements of chrF++ and COMET on average accuracy). The results prove that MLIT efficiently improves the translation ability on both low-resource and high-resource languages.

Compared with the high-resource translation, Table 2 shows that our propoesd method does not have such a big advantage over large scale models. However, it still achieves the best average score. Under the high-resources condition, ChatGPT shows more obvious advantages and achieves the best performance on all language pairs. Meanwhile, ChatGPT has a relatively stable performance on all experimental data, and the score gap is small

model	de-fr		es-fr		id-fr		ru-fr		th-fr		avg	
	chrF++	COMET	chrF++	COMET	chrF++	COMET	chrF++	COMET	chrF++	COMET	chrF++	COMET
BigTranslate-13B(Yang et al., 2023)	44.5	0.78	47.5	0.77	38.0	0.73	38.8	0.75	13.4	0.54	36.4	0.71
BayLing-13B(Zhang et al., 2023b)	52.1	0.78	49.4	0.78	42.7	0.71	49.4	0.73	26.8	0.57	44.1	0.71
ChatGPT(OpenAI, 2022)	**61.4**	**0.86**	**56.1**	**0.88**	**57.7**	**0.88**	**57.3**	**0.84**	**47.7**	**0.79**	**56.0**	**0.85**
Atom-7B+Bilingual Instruction Tuning	48.9	0.80	46.7	0.79	45.8	0.75	46.1	0.74	24.7	0.58	42.4	0.73
Atom-7B+Chain of Thought	48.9	0.74	46.0	0.72	44.1	0.73	45.4	0.68	22.9	0.57	41.5	0.69
Atom-7B+Mixed Instruction	44.2	0.69	40.9	0.69	41.0	0.60	40.2	0.64	21.2	0.55	37.5	0.63
Atom-7B+Chained Multilingual Instruction	45.6	0.73	43.1	0.71	43.9	0.66	44.7	0.68	22.0	0.53	39.9	0.66
Atom-7B+Pivot Prompting	51.0	0.82	46.7	0.80	48.2	0.77	46.3	0.76	23.2	0.57	43.1	0.74
Atom-7B+MLIT	51.5	0.83	46.7	0.82	51.1	0.80	49.8	0.76	27.0	0.63	45.5	0.77

Table 2: Results of MLIT method in chrF++ and COMET for MT on the FLORES-200 dataset. We experiment on the **to-French** translation task based on Atom-7B foundation.



(a) Standard reference (b) Other reference

Figure 3: The relationship between the quality English reference in training process and the inference score. We evaluate the different quality of standard English reference and other kind of reference using the chrF++ score.

between each language pair. These experiments prove that the languages that the foundation model supports plays an important role on translation.

4.4 The Impact of MLIT on Training

4.4.1 The impact of reference quality on training

To explore how instruction tuning affect the model, we generate different quality of English reference for MLIT. We first experiment on three language pairs (fr-zh, de-zh, ru-zh), which contains both high-resource and low-resource language pairs. As shown Figure 3(a), with the increase of the English reference quality, the scores of the prediction change very little in all the experimented language pairs.

Besides, we continuously experiment on three different settings: (1) The original English reference of MLIT. (2) We shuffle the order of the original English reference. (3) We leverage German as reference. As shown in Figure 3(b), these two new settings decrease model performance a little, especially the German reference. These results indicate that The MLIT does not teach the model new knowledge (when the given reference is wrong in setting (2), it can performer well), but transfer the knowledge through the reference (the performance of the model will decrease on references of a weaker language in setting (3)).

4.4.2 MLIT improves the model's basic ability

To evaluate what improvements MLIT has brought during the training phase, we generate the instruction with the blank reference for our instruction tuned model (the format of the blank reference is appended in Appendix B). We compare the results with the bilingual instruction tuned model. Our model has no additional information for inference with the blank reference. As shown in Figure 4, with the same inference setting, our model achieves a better average score of all the languages. For the high-resource language pairs, our MLIT method can effectively enhance the basic capabilities of the model. However, our approach has limitations in this regard for low-resource languages. We think this may cause by the foundation model is weak on the low-resource, so it is hard to improve it. We will explore this issue in subsequent work.

4.5 How Does English Reference Affect Inference

To evaluate the impact of the English reference in inference, we generate difference quality of English reference for inference. We experiment on French to Chinese translation. The results is shown in Figure 5. As we can see, the translation accuracy is directly proportional to the quality of the English reference. Although there is a drop in accuracy in the middle part of the figure, they fluctuate on references of similar quality. The results also proves the truth, that compared with the pivot prompting method, our method maintains the source sentence and adds English sentence as reference to reduce the noise of the inaccurate English.

Besides, we evaluate the parallel English reference of the input French sentence. Table 3 shows the **upper limit** of the improvement brought by

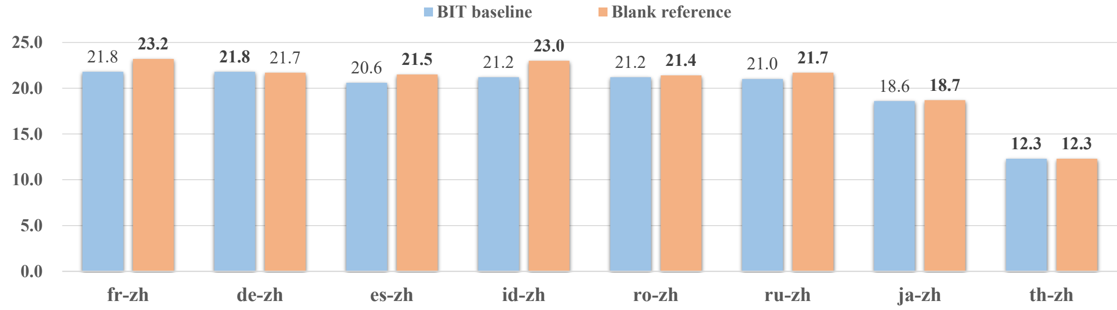


Figure 4: The accuracy comparison between the bilingual instruction tuned baseline and the MLIT model with blank reference for inference.

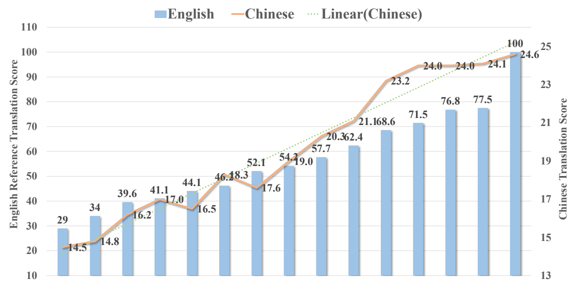


Figure 5: The results of the impact of reference on inference. The primary axis represents the chrF++ score of the English reference, and the secondary axis represents the chrF++ score of the Chinese translation. We plotted the trend line of the secondary axis relative to the primary axis.

Figure 6: The results of the case study. We choose French to Chinese translation task. It contains the input instruction and the outputs of the standard translation, baselines and our proposed method.

English reference, and our model is gradually approaching this upper limit. Meanwhile, we evaluate the MLIT trained model with blank reference. We regard this as the **lower limit** of the model. Table 3 shows that the lower limit of our model is better than the BIT baseline, which prove that we improve the translation ability through MLIT. Compared with the lower limit, the bad English reference will bring noise and affect the translation. This section shows the importance of English reference and proves the effectiveness of our method.

4.6 Case Study

To further understand the improvement of our proposed method, we provide a case study that contains the standard answer ang the outputs generated by the baselines and our method. As depicted in Figure 6, the standard translation contains two pieces of information, one is an introduction to animal classification and the other is saying that *"who is the only catamount that roars"*. For the BigTranslate model, some of the information was not translated, and secondly, it missed the second part information. BayLing, ChatGPT and our BIT baseline make the same mistake, which expands the scope (catamount to animal). In this case, only Google Translate and our method give the right translation. This indicates that our proposed MLIT can help the model to better understand sentences and their logical information on the non-English translation task. And this capability is essential to the translation task, because understand the sentence is the first step of translation. This observation further validates the effectiveness of MLIT.

4.7 MLIT Works Well on Large Scale Models

In this section, we apply the MLIT inference process to ChatGPT. We want to explore whether our method can narrow the gap between ChatGPT and

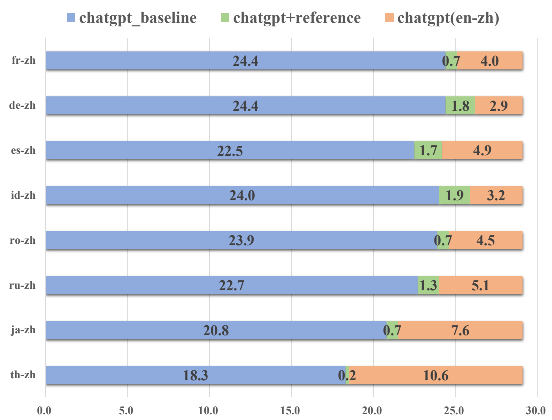


Figure 7: The results of our method on ChatGPT. The blue part represents the baseline of ChatGPT. The green part indicates the improvements of adding the English reference compared with the baseline. The orange part represents the gap between adding reference model and English to Chinese translation score.

our proposed model in low-resource translation. We generate English reference using ChatGPT to build the mulilingual prompt for inference. As shown of the blue and green part in Figure 7, our method achieves better results compared with the baseline. These results demonstrate the effectiveness of our method on large scale language models.

However, the improvement is limited. We conducted the English to Chinese translation to explore the limitation. As shown in Figure 7, what limits the performance of ChatGPT on Chinese-Centric translation is its lack of Chinese capabilities. So, the English to Chinese translation ability is a major problem of LLMs on low-resource tasks.

5 Related Work

5.1 Instruction Tuning

In recent years, LLMs have undergone rapid development. One of the major issue with LLMs is the mismatch between the training object and the users' object (Brown et al., 2020; Fedus et al., 2022; Rae et al., 2021; Thoppilan et al., 2022) . Instruction tuning method is proposed to address this mismatch, which is an efficient technique to make the LLMs perform complex and diverse tasks in the unified form. Generally, todays' LLMs, such as ChatGPT (OpenAI, 2022), use instruction tuning via supervised learning in the second training step (Sanh et al., 2022; Wei et al., 2022; Mishra et al., 2021). The instructions serve to constrain the model's outputs and provides a channel for humans to intervene with the model's behaviors (Zhang et al., 2023c). The LLMs can rapidly adapt to a specific domain with the help of Instruction tuning.

5.2 Multilingual Generalization

Training a universal translation system between multiple languages has shown enormous improvement for translating low-resource languages (Gu et al., 2020; Arivazhagan et al., 2019). Most studies focus on the unbalanced problem of each language in multilingual translation. Some works explore how to design the shared and language-dependent model parameters (Wang et al., 2018; Lin et al., 2021; Xie et al., 2021; Wang and Zhang, 2022). Other studies work on how to train the multilingual translation model when the training data are quite unbalanced across languages (Zhou et al., 2021; Huang et al., 2022). Recently, with the emergence of Large Language Models (LLMs), nontraining-based cross-lingual learning has gained more attention (Brown et al., 2020; Ahuja et al., 2023; Winata et al., 2022; Zeng et al., 2023; Huang et al., 2023).

Compared to their work, we propose the multilingual instruction tuning (MLIT) method to improve the LLMs on non-English translation, which only need cross-lingual parallel data.

6 Conclusion

In this work, we proposed multilingual instruction tuning (MLIT) method for non-English machine translation. Specifically, MLIT method consists of a to-English translation model and a multilingual instruction translation model. We leverage the to-English model to generate English instruction as reference to guide the non-English translation. The experiments show that our method outperforms the baselines on all the language pairs. Besides, our proposed model achieves a better performance than ChatGPT on some language pairs. The extensive experiment shows the contributions of MLIT on both training and inference processes.

7 Limitations

In this work, we focus on the non-English-Centric translation. The results prove that the low resource language capability of the foundation model is still a main reason that limits the further improvement of the model which is proved in Section 4.7. Therefore, improving the foundation model on other language remains an urgent issue that needs to be addressed in the future.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2020. Improved zero-shot neural machine translation via ignoring spurious correlations. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 1258–1268. Association for Computational Linguistics (ACL).

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

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*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022. Unifying the convergences in multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6822–6835.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. 2021. Scalable and efficient moe training for multitask multilingual models. *arXiv e-prints*, pages arXiv–2109.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. *arXiv preprint arXiv:2105.09259*.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *arXiv preprint arXiv:2104.08773*, pages 839–849.

OpenAI. 2022. Openai: Introducing chatgpt. In *https://openai.com/blog/chatgpt*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Qian Wang and Jiajun Zhang. 2022. Parameter differentiation based multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11440–11448.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

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*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigt, and Jie Yang. 2023. Greenplm: cross-lingual transfer of monolingual pre-trained language models at almost no cost. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6290–6298.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

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

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023c. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. 2021. Distributionally robust multilingual machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5664–5674.

A Details of Our Baselines' Instruction

We train our baselines with four kinds of prompts: 1) Bilingual Instruction Tuning (BIT) method: BIT method is the traditional method which leverage bilingual parallel sentences for training. 2) Chain of Thought (COT) method: COT method explicitly makes the model perform the translation process of first to English reference and then to non-English language with the bilingual parallel sentences. 3) Mixed Instruction (MI) method: MI method leverage both source-English and source-target parallel data to train a baseline which is the same data with our MLIT method. 4) Chained Multilingual Instruction (CMI) method: this method leverage multilingual parallel data in one instruciton which using more rigorous data than our methods.

B The Format of Blank Reference Instruction

The blank reference only contains the source sentence. The English reference of this instruction is blank, which leverage the same information as the BIT method for inference.

<source-language>: <*source-sentence*>
Reference: \n
<target-language>:

C Translation Quality of Different Reference

We generate different quality and kinds of English reference to evaluate the influence during inference. The two reference with scores (46.2 and 68.6) represents different quality reference. Bilingual baseline represents represents the BIT-trained method. Blank English reference is the same with Appendix B. Parallel English reference represents the parallel English reference of the source sentence.

model	score
Bad English reference (46.2)	18.3
Bilingual baseline	21.8
Bad English reference (68.6)	23.2
Blank English reference	23.2
Our MLIT method	24.1
Parallel English reference	24.6

Table 3: Results of different quality of English reference on inference. We evaluate two bad references with its chrF++ score. We leverage the bilingual instruction trained Atom as the baseline. We use source-language-only instruction and the parallel English instruction as the upper and lower limits of our MLIT model.

D ChatGPT Prompts

We evaluate the performance of ChatGPT using the following prompts. We report the best score of these prompts in Section 4.

Baseline	Instruction Format
BIT	<*SRC*>: <*SRC-sentence*> <*TGT*>:
COT	Please translate the following <*SRC*> sentence first into *English*, then into <*TGT*> : <*SRC-sentence*>
MI	<*SRC*>: <*SRC-sentence*> *English*: <*SRC*>: <*SRC-sentence*> <*TGT*>:
CMI	Consider the following <*SRC*> sentence and its *English* translation. Please translate it into <*TGT*>. <*SRC*>: <*SRC-sentence*> *English*: <*English-sentence*> <*TGT*>: Please translate the following <*SRC*> sentence first into *English*, then into <*TGT*> : <*SRC-sentence*>

Table 4: The instruction used for baselines. <*SRC*> and <*TGT*> denote source and target languages, respectively. <*SRC-sentence*> represents the source language to be translated.

ID	Prompt Format
1	Translate the following sentence from <*SRC*> to <*TGT*>: <*SRC-sentence*>
2	Translate the following <*SRC*> sentences into <*TGT*>: <*SRC-sentence*>
3	Provide the <*TGT*> equivalent for the following <*SRC*> sentences: <*SRC-sentence*>
4	Please provide the <*TGT*> translation for this sentence: <*SRC-sentence*>
5	What is the <*TGT*> version of this <*SRC*> sentence? <*SRC-sentence*>
6	What do the following sentence mean in <*TGT*>? <*SRC-sentence*>
7	What is the translation of this <*SRC*> sentence in <*TGT*>? <*SRC-sentence*>
8	How do this <*SRC*> sentence translate to <*TGT*>? <*SRC-sentence*>
9	I want you to act as a machine translation expert for <*SRC*> to <*TGT*>. <*SRC-sentence*>
10	You are a helpful assistant that translates <*SRC*> to <*TGT*>: <*SRC-sentence*>
11	<*SRC*>: <*SRC-sentence*> \n <*TGT*>:

Table 5: The prompts used for ChatGPT translation. <*SRC*> and <*TGT*> denote source and target languages, respectively. <*SRC-sentence*> represents the source language to be translated.