A Novel Paradigm Boosting Translation Capabilities of Large Language Models

Anonymous ACL submission

Abstract

This paper presents a study on strategies to enhance the translation capabilities of large language models (LLMs) in the context of machine translation (MT) tasks. The paper proposes a novel paradigm consisting of three stages: Secondary Pre-training using Extensive Monolingual Data, Continual Pre-training with Interlinear Text Format Documents, and Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning. Previous research on LLMs focused on various strategies for supervised fine-tuning (SFT), but their effectiveness has been limited. While traditional machine translation approaches rely on vast amounts of parallel bilingual data, our paradigm highlights the importance of using smaller sets of high-quality bilingual data. We argue that the focus should be on augmenting LLMs' cross-lingual alignment abilities during pre-training rather than solely relying on extensive bilingual data during SFT. Experimental results conducted using the Llama2(Touvron et al., 2023) model, particularly on Chinese-Llama2(Cui et al., 2023) after monolingual augmentation, demonstrate the improved translation capabilities of LLMs. A significant contribution of our approach lies in Stage2: Continual Pre-training with Interlinear Text Format Documents, which requires less than 1B training data, making our method highly efficient. Additionally, in Stage3, we observed that setting instructions consistent with the source language benefits the supervised fine-tuning process. Experimental results demonstrate that our approach surpasses previous work and achieves superior performance compared to models such as NLLB-54B(Team et al., 2022) and GPT3.5-text-davinci-003, despite having a significantly smaller parameter count of only 7B or 13B. This achievement establishes our method as a pioneering strategy in the field of machine translation.

1 Introduction

Translation capabilities of large language models (LLMs)(Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) in machine translation (MT) tasks have been explored extensively in previous research(Jiao et al., 2023; Zeng et al., 2023; Chen et al., 2023; Xu et al., 2023; Yang et al., 2023; Zhang et al., 2023). However, achieving significant improvements in translation quality through supervised fine-tuning (SFT) strategies has proven challenging. Traditionally, machine translation relies on vast amounts of parallel bilingual data, but SFT only requires a small amount of high-quality bilingual data, highlighting a crucial distinction. It is a naive approach to consider using vast quantities of parallel bilingual data during SFT. However, experiments have shown that increasing the data volume yields limited improvements in quality and even leads to performance degradation on certain test sets. **Thus, the question arises: are extensive parallel bilingual data useless in SFT, or are they being misused?**

In this paper, we propose a novel training paradigm, consisting of three stages, to boost the translation capabilities of LLMs. Our contributions include refining the training strategy for downstream tasks and emphasizing the enhancement of LLMs' cross-lingual alignment abilities during pre-training. These contributions address the limitations observed in previous research.

Our training paradigm comprises the following stages:

Stage 1: Secondary Pre-training using Extensive Monolingual Data. Consistent with previous findings(Xu et al., 2023), we validate the effectiveness of monolingual data augmentation. Specifically, we perform SFT on a Chinese-Llama2(Cui et al., 2023) model, which undergoes monolingual data augmentation, thereby demonstrating improved translation capabilities.

Stage 2: Continual Pre-training with Interlinear Text Format Documents. We construct interlinear text format documents from aligned bilingual data and utilize them for continual pre-training of LLMs. Experimental results demonstrate the critical importance of this stage, resulting in a significant improvement in translation quality, particularly for English-Other translations. Stage 2 stands as a pivotal contribution in our paper, requiring less than 1B training data, thereby enhancing training efficiency.

Stage 3: Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning. In SFT, we discover that using instruction aligned with the source language of the translation notably improves performance. Leveraging source-language consistent instructions during SFT yields significant enhancements.

In summary, this paper presents a novel paradigm to boost the translation capabilities of large language models (LLMs)(Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) in machine translation tasks. Our contributions involve refining the training strategy for downstream tasks and emphasizing the improvement of LLMs' cross-lingual alignment abilities during pre-training. We introduce a three-stage training paradigm, highlighting the effectiveness of secondary pre-training, continual pre-training with interlinear text format documents, and leveraging source-language consistent instruction for supervised fine-tuning. These contributions address the limitations observed in previous research and pave the way for improved translation quality.

2 Related Work

2.1 Large Language Models

Foundation Model Foundation Model, a product of pre-training, is a prominent type of Large Language Model. It has gained substantial recognition in recent years for its impressive capabilities in natural language processing tasks. The most prevalent architectural framework for such models is the Transformer, which employs a series of self-attention mechanisms to process input text efficiently.

Among the state-of-the-art Large Language Models, notable examples include GPT-3(Brown et al., 2020) and Llama2(Touvron et al., 2023). These models have been widely lauded for their exceptional proficiency in understanding and generating natural language text. They showcase the remarkable potential of Foundation Models, pushing the boundaries of language processing and setting new benchmarks in various applications.

Instruct/Chat Model Instruct/Chat Model, a variant of Large Language Models, is specifically developed through the process of Supervised Fine-Tuning (SFT). Unlike Foundation Models, which are pre-trained, Instruct/Chat Models undergo additional supervised training to enhance their performance in specific tasks such as instruction following or conversational dialogue.

Supervised Fine-Tuning involves training the model on labeled datasets, where human annotators provide examples of desired input-output behavior. This approach enables Instruct/Chat Models to learn task-specific skills and exhibit improved performance in situations that require language understanding, generation, and interaction.

Noteworthy advancements have been observed in Instruct/Chat Models, with notable examples including models like DialoGPT and ChatGPT. These models have exhibited remarkable outcomes in conversational scenarios, demonstrating their potential in enabling interactive and engaging human-like conversations.

2.2 Machine Translation Task

Machine Translation Task refers to the process of automatically translating text from one language to another using computational methods. It aims to bridge language barriers and enable effective communication across different languages.

Traditional Methods Traditional approaches to machine translation have primarily relied on encoder-decoder(Vaswani et al., 2017) models. These models consist of an encoder that encodes the source language sentence into a fixed-length representation, followed by a decoder that generates the translated sentence in the target language. These methods heavily depend on large bilingual parallel corpora, which align source sentences with their corresponding translations. The prevailing understanding is that the larger the amount of data available and the larger the model, the better the translation quality.

LLM-based Methods In recent years, Language Model (LM)-based approaches have gained attention in the field of machine translation. These approaches leverage pre-trained language models,

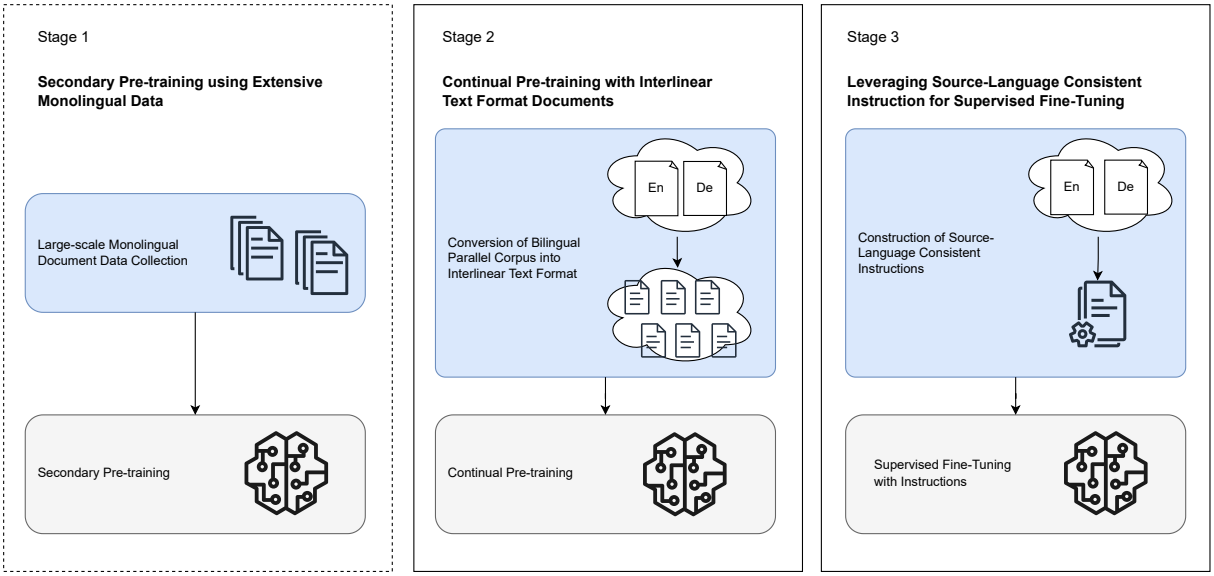

Abbildung 1: The overall of our approach. Stage 1: Secondary Pre-training using Extensive Monolingual Data. Stage 2: Continual Pre-training with Interlinear Text Format Documents. Stage 3: Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning. *It should be noted that Stage 1 is considered non-essential.*

such as the GPT (Generative Pre-trained Transformer)(Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) series, and adapt them for translation tasks.

One line of LLM-based methods focuses on zero-shot or few-shot translation by incorporating in-context learning(Hendy et al., 2023). By conditioning the LLM on a source sentence, the model can generate translations in the target language without explicitly using parallel data. This approach has shown promising results in enabling translation for language pairs with limited or no parallel resources.

Another approach involves using a small amount of high-quality bilingual parallel data to construct translation-guiding instructions. These instructions explicitly define the translation behavior by providing source-language consistent cues during the supervised fine-tuning (SFT) process. By utilizing these specially crafted instructions, the LM can be fine-tuned to perform translation more accurately and robustly.

Overall, LLM-based methods present alternative approaches to machine translation, exploring the potential of leveraging pre-trained models and incorporating limited parallel resources or high-quality instructions to enhance translation quality.

3 A New Training Recipe

We propose an innovative training strategy to enhance the translation capabilities of Large Language Models. Our approach consists of three stages: (1) Secondary Pre-training using Extensive Monolingual Data, (2) Continual Pre-training with Interlinear Text Format Documents, and (3) Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning.

3.1 Stage 1: Secondary Pre-training using Extensive Monolingual Data

In this stage, our aim is to enhance the training of large language models (LLMs) by utilizing diverse monolingual data. Currently, existing large models, such as Llama, are primarily pre-trained on English-centric corpora, resulting in relatively weaker comprehension and generation abilities in non-English languages. To expand the multilingual generation capabilities of LLMs, we suggest an incremental pre-training approach using extensive monolingual data.

It is important to note that this stage primarily focuses on enhancing the intrinsic multilingual capacity of LLMs. While it is inherently related to machine translation tasks, it is not essential. On the one hand, we can select an existing LLM that already demonstrates robust multilingual capabilities as the base model for further training. On the other hand, even LLMs with limited multilingual support can benefit from the subsequent stages outlined in our methodology.

3.2 Stage 2: Continual Pre-training with Interlinear Text Format Documents

Interlinear Text Format Documents Interlinear Text Documents are a specific type of parallel text resource that consists of source sentences and their corresponding translations displayed in a aligned format. Each source sentence is accompanied by its translation, typically presented word-by-word or phrase-by-phrase, to facilitate a clear interlingual correspondence. See Figure 2.

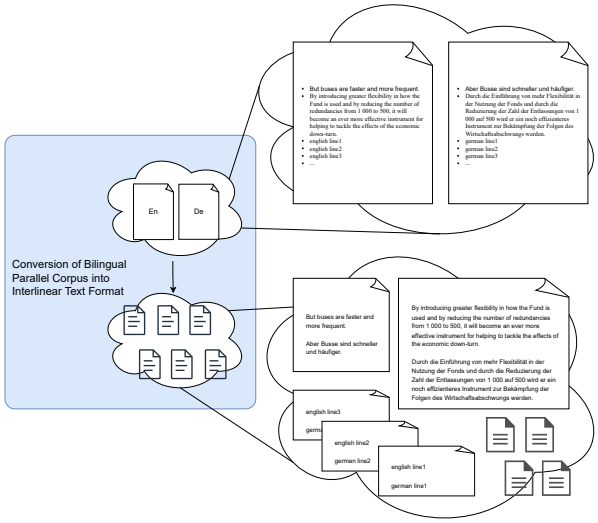

Abbildung 2: Interlinear Text Format Documents

Utilizing Interlinear Text Documents offers several advantages for language understanding and translation tasks. Firstly, these documents provide explicit linguistic alignment at a fine-grained level, enabling the model to capture syntactic and semantic correspondences across languages. This aligns closely with the goals of machine translation, as it facilitates accurate encoding of source language information and improves the quality of generated translations. Additionally, interlinear data contributes to the learning of interlingual representations, allowing the model to better understand the relationship and transferability between languages.

Continual Pre-training To leverage the benefits of Interlinear Text Documents, we propose a Continual Pre-training strategy based on the LoRA(Hu et al., 2021) (Language Optimization with a Reinforced Adversary) framework. LoRA is a robust and effective pre-training approach for language models, introduced in recent research.

By leveraging the inherent alignment information present in Interlinear Text Documents, the model learns to align and generate translations that maintain syntactic and semantic consistency with the source sentences. This continual training process allows the model to progressively improve its ability to capture cross-lingual correspondences, resulting in enhanced translation quality.

3.3 Stage 3: Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning

Source-Language Consistent Instruction In the field of machine translation, SSource-Language Consistent Instructionrefers to the practice of constructing translation instructions that maintain consistency with the source language, aiming to achieve better results. This approach involves generating instructions that are closely related to the source language. By providing more accurate and clear guidance for supervised fine-tuning of models, this technique enhances translation quality.

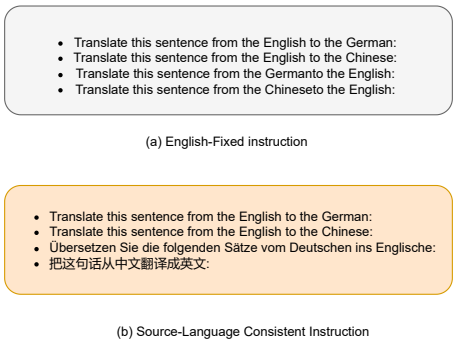

Abbildung 3: Instruction

To illustrate this concept, let's consider translations in English⇔Chinese and English⇔German. Traditional approaches typically employ a standardized English-Fixed instruction such as "*Translate this sentence from the source language to the target language:*". However, in Source-Language Consistent Instruction, the instruction varies based on the language pair. For English-to-Chinese translation, the instruction would be "把这句话从中文翻译成英文："(Translate this sentence from Chinese to English). Similarly, for German-to-English translation, the instruction would be "*Übersetzen Sie die folgenden Sätze vom Deutschen ins Englische:*"(Translate the following sentences from German to English). By utilizing language-specific instructions, there is a semantic consistency established between the instruction and the source language, resulting in clearer and more accurate guidance. As shown in Figure 3.

Supervised Instruction Fine-Tuning Supervised Instruction Fine-Tuning for machine translation tasks incorporates two pivotal aspects. Firstly, akin to the earlier phase of Continual Pre-training, we employ LoRA(Hu et al., 2021) to finely tune specific parameters of Language Learning Models (LLMs), thereby enhancing their efficiency. LoRA(Hu et al., 2021) plays a crucial role in preventing model overfitting and leads to notable performance improvements. With this approach, we judiciously fine-tune a subset of model parameters using low-rank updates, striking a delicate balance between model adaptation and computational efficiency.

Secondly, as emphasized in prior studies(Zhou et al., 2023; Maillard et al., 2023; Xu et al., 2023), LLMs exhibit benefits from a limited yet high-quality dataset. To ensure optimal data quality during the fine-tuning process, we leverage exceptional data sources. In line with previous research, we make use of meticulously curated human-written datasets derived from the WMT test data. These datasets undergo rigorous quality control measures, rendering them an ideal choice for fine-tuning purposes.

4 Experiments

4.1 Datasets and Evaluation Metrics

For Stage 2, we utilized the WMT bilingual training dataset consisting of English⇔German (En⇔De) and English⇔Chinese (En⇔Zh) sentence pairs. The En⇔De dataset comprised approximately 4.5 million pairs, while the En⇔Zh dataset contained around 25 million pairs. Due to the higher number of En⇔Zh pairs compared to En⇔De, we sampled 4.5 million En⇔Zh pairs for our experiments. Overall, the combined dataset contained nearly 1 billion tokens.

For Stage 3, we employed the newstest2017-2020 dataset for both En⇔Zh and En⇔De translation tasks. This dataset included a total of 37.6 thousand sentence pairs for each language direction. To ensure consistency across the source language and target language, we organize these sentence pairs into Source-Language Consistent Instructions.

We additionally incorporate the test sets from the WMT22 competition, which are carefully curated to include more recent content from diverse domains such as news, social media, e-commerce, and conversations. The test sets for the De⇒En, En⇒De, Zh⇒En, and En⇒Zh tasks consist of 1984, 2037, 1875, and 2037 samples, respectively.

For automatic evaluation, we utilize SacreBLEU, which implements BLEU(Papineni et al., 2002), and COMET(Rei et al., 2020) from Unbabel/wmt22-comet-da. SacreBLEU calculates similarity based on n-gram matching, while COMET leverages cross-lingual pretrained models for evaluation.

4.2 Setup

We conducted our experiments using HuggingFace Transformers with open-source LLMs from the LLaMA family. Specifically, we utilized LLaMA2-7b(Touvron et al., 2023) with matched parameters as our foundation model. Additionally, we included LLaMA2-13b(Touvron et al., 2023) to explore the impact of different model sizes.

Due to computational constraints, we did not reproduce the foundational experiments from Stage 1. After Stage 1, we selected Chinese-LLaMA2(Cui et al., 2023) as our new foundation model. Chinese-LLaMA2(Cui et al., 2023) is an extended and optimized version of Llama-2(Touvron et al., 2023), specifically tailored for Chinese language understanding and instruction comprehension. It incorporates a larger Chinese vocabulary and undergoes incremental pretraining on a large-scale Chinese dataset, which further enhances its semantic understanding capabilities.

For Stage 2, Continual Pre-training, and Stage 3, Supervised Fine-Tuning, we referred to the hyper-parameters employed in the Chinese-LLaMA2(Cui et al., 2023) project. During Stage 2, we trained the model for 1 epoch, and for Stage 3, we extended the training to 3 epochs.

Our experiments were conducted on 8 Nvidia A100 GPUs, utilizing DeepSpeed(Rasley et al., 2020) ZeRO 2 for model parallelization.

4.3 Baselines

We evaluate our method against two baseline categories, consistent with previous studies. Firstly, we compare our approach to prior studies that share our goal of leveraging LLMs for translation. Secondly, we benchmark against the current state-of-the-art (SoTA) translation models. It's important to note that this comparison may not be entirely fair due to disparities in training data and model architectures. For example, there is a significant contrast between the 175B GPT-3.5 model and our 7B model. Nevertheless, by utilizing the same test

Models	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
SoTA models								
NLLB-54B(Team et al., 2022)	26.89	78.94	34.50	86.45	16.56	70.70	27.38	78.91
GPT-3.5-D, zero-shot	30.90	84.79	31.80	85.61	25.00	81.60	38.30	85.76
GPT-3.5-T, zero-shot	33.10	85.50	34.40	87.00	26.60	82.90	44.90	87.00
GPT-4	33.87	85.62	35.38	87.44	27.20	82.79	43.98	87.49
Prior Similar Studies								
TIM-7B(Zeng et al., 2023)	27.91	82.80	25.59	82.56	19.33	75.46	19.33	75.46
Parrot-7B(Jiao et al., 2023)	29.80	83.00	26.10	81.60	20.20	75.90	30.30	80.30
SWIE-7B(Chen et al., 2023)	30.48	82.97	27.21	82.36	21.30	76.48	31.24	80.63
ALMA-7B(Xu et al., 2023)	29.56	83.95	30.31	85.59	**23.64**	**79.78**	36.48	85.05
Parrot-13B(Jiao et al., 2023)	31.10	83.60	28.10	82.60	21.70	76.70	31.70	81.00
BigTranslate-13B(Yang et al., 2023)	23.35	80.68	21.48	78.81	14.16	74.26	28.56	81.31
Bayling-13B(Zhang et al., 2023)	27.34	83.02	25.62	82.69	20.12	77.72	37.92	84.62
ALMA-13B(Xu et al., 2023)	31.14	84.56	31.47	85.62	**25.46**	**80.21**	39.84	85.96
(Ours)	Our Recipe with Backbone Model: LLaMA-2(Touvron et al., 2023)							
7B Stage3	30.02	84.09	25.40	82.30	20.59	76.18	30.60	80.40
7B Stage2,3	**31.14**	**84.70**	**30.50**	**85.62**	21.97	78.45	**39.00**	**85.79**
13B Stage3	31.70	84.39	28.80	83.87	21.40	77.68	35.90	84.23
13b Stage2,3	**32.24**	**85.17**	**32.53**	**86.14**	22.57	79.05	**40.40**	**85.98**
(Ours)	Our Recipe with Backbone Model: Chinese-LLaMA2(Cui et al., 2023)							
7B Stage3	25.20	78.32	12.50	69.19	20.90	76.40	35.00	84.32
7B Stage2,3	30.10	83.96	29.90	83.86	22.20	**79.88**	**41.10**	**86.37**
13B Stage3	26.13	78.65	12.79	72.23	21.40	78.28	37.34	85.27
13b Stage2,3	30.21	84.26	30.41	84.72	23.10	**80.53**	**42.30**	**86.65**

Tabelle 1: The overall results.

set, we can gain insights into our model's current performance.

In the category of prior similar work, we compare our model to the following approaches: BigTranslate(Yang et al., 2023), which extends LLaMA-1-13B to cover over 100 translation directions; TIM(Zeng et al., 2023), which leverages correct and incorrect examples to aid LLMs in learning translation; SWIE(Chen et al., 2023), which enhances LLMs in translation through instruction augmentation; BayLing(Zhang et al., 2023), which incorporates interactive translation instructions; and ALMA(Xu et al., 2023), a two-stage fine-tuning method that initially fine-tunes on monolingual data and subsequently on a small set of high-quality parallel data.

In the SoTA models category, we consider the following: the NLLB-54B(Team et al., 2022) model, the largest and best translation model released in the NLLB family; the zero-shot performance of GPT3.5-text-davinci-003 (GPT-3.5-D) and GPT-3.5-turbo-0301 (GPT-3.5-T). Additionally, we present the zero-shot results for GPT-4. To evaluate these baselines, we employ the same test data and evaluation metrics, reporting BLEU(Papineni et al., 2002) and COMET(Rei et al., 2020) scores as provided in their respective papers.

5 Results and Analysis

As shown in Table 1, **overall, our results outperform all previous studies, NLLB-54B(Team et al., 2022), and GPT-3.5-D, except for a slight lag in Zh⇒En. Even our 7B model surpasses the results of other works. Particularly in the En⇒Zh direction, our BLEU score is approximately 2.5 higher than the previous state-of-the-art.** These findings are a testament to the effectiveness of our approach.

5.1 Assessing the Impact of Stage 1

Just as mentioned earlier, we didn't specifically train Llama2 in Stage 1, but instead directly utilized the Chinese-Llama2(Cui et al., 2023) that had undergone similar training. Since Chinese-Llama2(Cui et al., 2023) was only trained in Chinese during Stage 1, our main analysis focuses on its performance in Zh⇒En and En⇒Zh translations.

As shown in Table 1, our findings align with previous research conclusions that incremental training on monolingual data is beneficial. Furthermore, we discovered that this benefit primarily affects the target language in translation tasks. For example, we observed a significant improvement in the performance of the 7B model on the En⇒Zh test

set, where the BLEU score increased from 30.60 to 35.00, a substantial improvement of 4.4 points. However, the improvement in the Zh⇒En direction was limited, indicating that the role of Stage 1 is to enhance generation rather than comprehension.

Additionally, we found that performing incremental training on only one monolingual dataset had disastrous effects on translation tasks in other languages. For example, on the En⇒De test set, the BLEU score plummeted from 25.40 to 12.50. Therefore, for multilingual translation, it is crucial to conduct Stage 1 training on multiple languages.

5.2 Measuring the Effectiveness of Stage 2

As shown in Table 1, Llama2(Touvron et al., 2023) demonstrates improved quality across various test sets after Stage 2 training. An interesting observation, considering Llama2 as a large model primarily focused on English, is that the enhancement in English-Other translations is particularly noteworthy after Stage 2 Training. For instance, the 7B model exhibits remarkable improvements in En⇒De, with the BLEU score increasing from 25.40 to 30.50, and in En⇒Zh, where it rises from 30.60 to 39.00. The magnitude of these improvements is quite significant! Encouragingly, there are also improvements observed in translations from other languages to English.

Models	De⇒En	En⇒De	Zh⇒En	En⇒Zh
Stage 3	30.02	25.40	20.59	30.60
w/o	30.40	25.20	18.39	28.30
Stage2,3	31.14	30.50	21.91	39.00
w/o	31.00	30.23	18.93	38.69

Tabelle 2: Results of the comparative experiments based on Llama2-7B(Touvron et al., 2023) model. [w/o] means using English-Fixed Instruction.

An even more intriguing finding is that, as mentioned before, since Chinese-Llama2(Cui et al., 2023) only underwent Stage 1 training on Chinese, its translation performance substantially deteriorates in the En⇒De direction. However, with the magical touch of Stage 2 training, these capabilities are miraculously restored. The 7B model, on En⇒De, rebounds from 12.50 to 29.90, approaching the results of the original Llama2(Touvron et al., 2023). These outcomes effectively affirm the effectiveness of Stage 2.

5.3 Analyzing the Outcomes of Stage 3

To evaluate the effectiveness of Source-Language Consistent Instruction in Stage 3, we conducted a comparative experiment using English-Fixed Instruction. The results of the experiment are presented in Table 2. It is evident that in the En⇒De and De⇒En directions, the performance of these two types of instructions is quite similar. However, in the Zh⇒En and En⇒Zh directions, the use of Source-Language Consistent Instruction clearly outperforms.

We believe that the similarity between English and German, as they belong to the same language family, contributes to the lack of noticeable differences. However, when dealing with cross-language pairs, employing Source-Language Consistent Instruction further enhances the translation quality.

5.4 Comparing the Difference with ALMA

We have noticed that our work shares some similarities with ALMA(Xu et al., 2023) in terms of the process, involving Continual Pre-training followed by Supervised Fine-Tuning. However, there are notable differences between our approaches.

ALMA(Xu et al., 2023) suggests that the impact of bilingual data is reduced in the era of large models. In contrast, we firmly believe in the continued strength of bilingual data and its application in Continual Pre-training through Interlinear Text Format Documents. While ALMA(Xu et al., 2023) acknowledges the effectiveness of conducting Continual Pre-training on monolingual data, we have also validated this finding in our own work and reached the same conclusion. However, it is important to note that our approach primarily enhances the multilingual generation capability of the large model itself, rather than being specifically tailored to translation tasks. Furthermore, ALMA(Xu et al., 2023) utilizes a significantly larger training dataset, ranging from 13B to 20B, compared to our own.

6 Ablation Study: What if we directly employ a large quantity of translation data for SFT?

Both Continual Pre-training and Supervised Fine-Tuning involve incremental training on the original model. However, if we skip Stage 2 training and directly utilize the translation data from Stage 2 as instruction data for SFT, i.e., conducting SFT directly with a substantial amount of translation

data, will it yield consistent improvement?

Data Size	De⇒En	En⇒De	Zh⇒En	En⇒Zh
37.6K	30.02	25.40	20.59	30.60
0.4M	30.20	25.60	18.49	31.74
4M	30.66	25.12	20.77	32.22

Tabelle 3: Results of the ablation experiments based on Llama2-7B(Touvron et al., 2023) model under different Stage 3 data size.

To address this question, we conducted an ablation experiment. Our Stage 3 training data consisted of 37.6k samples. Randomly selecting and merging some data from the Stage 2 training data with the Stage 3 training data, we created three sets: 37.6K, 0.4M, and 4M. The experimental results are presented in Table 3.

We found that augmenting the training data in Stage 3 slightly improved translation quality for certain test sets. This indicates that a small amount of high-quality data is sufficient for the SFT stage.

Now, our focus is solely on the translation task. However, if we were conducting multi-task SFT, it is unlikely that other tasks would have as extensive data as machine translation. Therefore, using a large amount of translation data during SFT would result in the problem of imbalanced data distribution with other tasks. Hence, the optimal approach would still be to utilize this substantial data during the earlier stage of Continual Pre-training.

7 Conclusions

In this study, we have introduced a novel paradigm for enhancing the translation capabilities of large language models in machine translation tasks. Our three-stage approach, including Secondary Pre-training using Extensive Monolingual Data, Continual Pre-training with Interlinear Text Format Documents, and Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning, addresses the limitations of previous strategies and offers notable improvements in translation quality. We emphasize the significance of pre-training stages in enhancing LLMs' cross-lingual alignment abilities and the effectiveness of using a smaller but high-quality set of bilingual data during supervised fine-tuning. Notably, Stage2, which involves Continual Pre-training with Interlinear Text Format Documents, stands out as a highly efficient method, requiring minimal training data. Furthermore, aligning the instructional setting with the source language during supervised fine-tuning, as observed in Stage3, yields positive effects. The findings from this paper contribute to advancing the field of machine translation and offer valuable insights for optimizing the translation capabilities of large language models. Future research can explore additional language pairs, alternative data augmentation techniques, and different pre-training strategies to further refine our proposed paradigm.

8 Limitations

Despite notable contributions, this study has certain limitations. Firstly, the proposed method exhibits slightly reduced performance in the Zh⇒En translation direction, necessitating further analysis and improvements. Secondly, the presence of illusionary translations within large models was observed but not extensively explored. Future research should delve deeper into this phenomenon. Lastly, while the paper primarily focuses on SFT for machine translation, opportunities exist to explore SFT techniques in diverse contexts such as style translation and colloquial translation. Addressing these limitations would further enhance the effectiveness and applicability of the proposed methods.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

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. *CoRR*, abs/2302.09210.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2740–2756. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine

translation: Boosting translation performance of large language models.

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

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

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.

A Example Appendix

This is an appendix.