Boosting Translation Capabilities of Large Language Models with Code-Switching Pretraining

Anonymous ACL submission

Abstract

Recently, there has been significant attention on adapting the translation capabilities of Large Language Models. Represented by ALMA (Xu et al., 2023), a two-stage training recipe has been developed: first, utilizing a large amount of monolingual data for pretraining to enhance proficiency in non-English languages, followed by fine-tuning with a small amount of high-quality bilingual data. However, in the pretraining process, explicit cross-lingual alignment information is not provided, and excessive use of bilingual data can lead to catastrophic forgetting issues, both of which hinder the further advancement of the model's translation abilities. In this article, we address this issue by introducing a new pretraining process based on Code-Switching pretraining data. In this stage of pretraining, we can provide rich cross-lingual alignment information while ensuring that the training data is semantically coherent documents, which helps alleviate catastrophic forgetting. Moreover, the training process relies solely on monolingual data and a pair of traditional machine translation models, making it highly versatile. Experimental results show that our method has improved the translation quality, achieving state-of-the-art results in similar works.

1 Introduction

The rapid development of large language models (LLMs) (Brown et al., 2020b; Chowdhery et al., 2023; Touvron et al., 2023), represented by the GPT series (Brown et al., 2020b), has brought exciting progress to the field of Natural Language Processing (NLP). The powerful language understanding, abstract summarization, and conversational generation capabilities of large models are revolutionizing numerous NLP tasks (Shao et al., 2023; Singhal et al., 2023; Zhang et al., 2024; Min et al., 2023), and the field of machine translation is no exception.

Extensive work has verified that large models can achieve zero-shot and few-shot translation through their powerful in-context learning (Hendy et al., 2023; Zhang et al., 2023a) capabilities, without the need for specific adaptations for translation tasks. However, since large language models are often trained on English as the primary language, the insufficient data in other languages has resulted in most LLMs' translation capabilities falling short compared to commercial traditional models or top commercial LLMs (Jiao et al., 2023b).

ALMA (Xu et al., 2023) has already proven that we can enhance the translation capabilities of LLM through continual training. They first enhance the proficiency of LLM in these non-English languages by adding monolingual data in those languages for Continual Pretraining (CPT), and then stimulate the translation capabilities of LLM by utilizing small amount high-quality bilingual data for Supervised Finetune (SFT). However, in the pretraining phase they proposed, there was no explicit modeling of cross-lingual alignment, which may hinder further quality improvement. In contrast, Guo et al. (2024) attempted to adding sentence-level parallel data during the pretraining phase using an Interlinear text format, but this method has two drawbacks: a) Extensive sentence-level bilingual data has been demonstrated to induce catastrophic forgetting (Xu et al., 2023) and erase previously acquired knowledge. b) The pattern of sentence-level parallel data diverges from that of standard pretraining data, necessitating high-quality, semantically coherent document data.

In studies of traditional machine translation (MT) models, the code-switching strategy (Lin et al., 2020; Yang et al., 2020) (i.e., replacing words or phrases in the current sentence with expressions from another language) has been shown to be effective in aligning multilingual representation spaces. Drawing inspiration from this, we attempted to modify the pretraining corpus of LLMs using a sentence-level code-switching strategy and obtained semantically coherent document data composed of sentences from two languages. Then LLM can learn cross-lingual contextual dependencies and alignment information embedded in the such data through standard pretraining.

To achieve our goals, the most ideal training data is document-level parallel corpora, but such data only exists between high-resource languages and in limited quantities. Nevertheless, leveraging the strong fundamental capabilities of LLMs along with specific markers enables us to utilize traditional MT models for generating document-level back translation data as an alternative.

More specifically, we use monolingual data in English and the target language, along with a pair of traditional MT models, to generate two types of code-switching pretraining data: from English to the target language and from the target language to English. Subsequently, a novel pretraining phase, denote as Code-Switching Continual PreTraining (standard pretraining on the code-switching data), is integrated into the two-stage training recipe suggested by ALMA. In the end, experiments show that our improved training recipe significantly enhances LLM's cross-lingual alignment capability. The translation quality from the target language to English and from English to the target language has both improved. At the same time, we found that in the new pretraining stage, the contribution of code-switching pretraining data in the same direction is greater than in the opposite direction, and we preliminarily analyze that such data may help improve the model's automatic post-editing capabilities.

Our core contributions are as follows:

- The Code-Switching Continual PreTraining stage we proposed can enhance the cross-lingual alignment capability of LLM, addressing the shortcomings of previous work.
- We introduced traditional MT models into the optimization process of LLM's translation capabilities in the form of back translation.
- The final optimized model achieved State-of-the-Art performance in some translation directions compared to similar works.

2 Related Work

2.1 Large Language Models

Large language models generally refer to transformer-based (Vaswani et al., 2017) neural models with billions of parameters. Both open-source models like Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023) and GLM (Zeng et al., 2022) and closed-source models like GPT-3.5/4 (Brown et al., 2020a), Claude (Anthropic) demonstrate enhanced language comprehension and generation capabilities. Mainstream LLMs follow a Decoder-only architecture, expanding their parameter size by layering Transformer decoder units. During training, LLMs initially undergo pretraining on a diverse range of document-level monolingual data (such as internet data, books, code, etc.) to establish a foundational model. Subsequently, they undergo training using algorithms like Supervised Finetune and RLHF (Ouyang et al., 2022) to align with human preferences and ultimately achieve a robust multi-turn Instruct/Chat model for diverse tasks.

When adapting LLM for downstream tasks, there are two common strategies: the Incontext-Learning (ICL) strategy based on prompt (Zhu et al., 2023) technology and various evolving techniques represented by COT (Wei et al., 2022). Another strategy involves fine-tuning (Ding et al., 2023) the model using downstream data, which often achieves higher performance. Technologies like Low-rank Adaptation (LoRA) (Hu et al., 2021), which solve the training cost issue, significantly enhance the applicability of this strategy.

2.2 Machine Translation Task

Traditional Methods The traditional machine translation models, represented by transformers (Vaswani et al., 2017), utilize an Encoder-decoder architecture to autoregressively decode the target language. Among various optimization methods, data augmentation (Burlot and Yvon, 2018) techniques like Back Translation (BT) (Edunov et al., 2018; Hoang et al., 2018; Pham et al., 2021) has been proven to be more effective. BT comprises different variations such as sampling BT, Noise BT, Tag BT (Caswell et al., 2019), and so on. In the training phase, BT incorporates a variety of monolingual data in the target language to boost the language model's capabilities, aiding in producing more natural and accurate outputs (Edunov et al., 2020). Additionally, Forward Translation (FT), which translates the source text into the target language, is frequently paired with BT data.

LLM-based Methods As we mentioned before, when adapting the translation capability of LLMs, there are two types of strategies. The first type

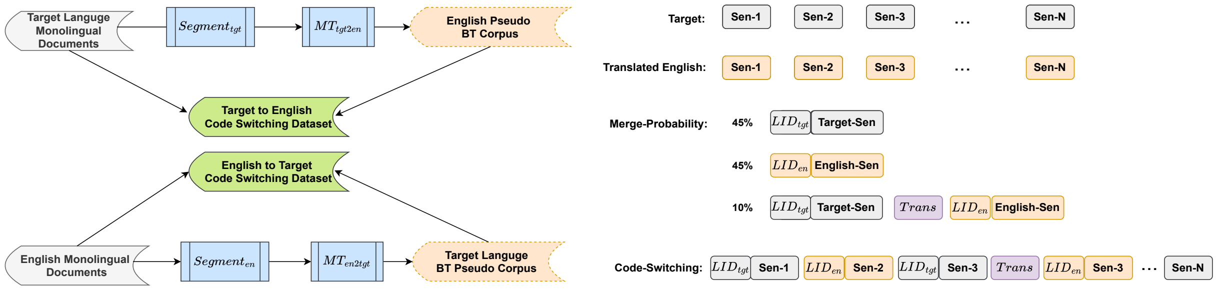

Figure 1: Construction process of Code-Switching pretraining data. The left side displays the key flow nodes involved in data construction, while the right side elaborates on the specifics of constructing Code-Switching data using original monolingual and BT pseudo-corpus. LID and $Trans$ are special tokens.

focuses on harnessing LLMs' Incontext-learning feature and employing prompt techniques to enhance the model's translation ability. Many studies (Hendy et al., 2023; Zhang et al., 2023a; Wang et al., 2023; Gulcehre et al., 2017) have conducted detailed explorations in this direction. Another type involves fine-tuning the model with specific data from translation tasks to achieve better translation quality. Different studies may attempt to fine-tune the model at different training stages. For example, fine-tuning the model with monolingual data (Tan et al., 2023; Yang et al., 2023; Wei et al., 2023) in the target language or domain during the pretraining phase. Alternatively, using translation-related instruction (Li et al., 2024; Zhang et al., 2023b) data during the SFT phase. Xu et al. (2024) aim to enhance translation quality by fine-tuning the model using comparison data with varying quality through reinforcement learning.

3 Methodology

In this chapter, we will describe the details of the code-switching strategy we proposed, as well as the training recipe we optimized for adapting the translation capabilities of LLMs.

3.1 Code-Switching Pretraining Data

In the traditional MT and multilingual language model (MLM) field, the code-switching strategy has been proven to provide cross-lingual alignment information (Lin et al., 2020; Yang et al., 2020). In order to adapt to pretraining tasks for LLMs, we use a sentence-level code-switching strategy and obtain semantically coherent document data composed of sentences from two languages.

In Figure 1, we illustrate the specific approach. We refer to the target language as tgt and English as en. Utilizing a pair of pre-trained traditional MT models, we translate monolingual English and target language corpora to generate BT pseudo-corpus denote as D_{bt}. When constructing Code-Switching pretraining data (D_{cs}), we randomly select original and translated sentences with equal probability, and with a 10% probability, we allow them to appear simultaneously. To effectively differentiate Code-switching data and prevent language confusion during inference, we incorporate some special tokens. The design of special tokens and an example of Code-Switching data are provided in the Appendix A.

3.2 A New Training Recipe

We proposed a new training recipe, in which we added a "Code-Switching Continual Pretraining" stage to ALMA's two-stage training recipe, aiming to more efficiently inject cross-lingual alignment information. Figure 2 illustrates our training recipe and the differences between our work and typical similar works.

Stage-1: Continual Pretraining with Monolingual Data LLMs like LLaMA are pre-trained on English-dominated corpora. They may encounter issues with insufficient comprehension and generation abilities in the target language. By incorporating a large amount of monolingual data in the target language for continual pretraining, we can alleviate this issue. At this stage, we can train with the full set of parameters or utilize LoRA technology to enhance training efficiency. Training data often comes from widely available internet sources, such as Common Crawl (Foundation, 2023), as well as some cleaned versions like OSCAR (Ortiz Su'arez et al., 2019; Kreutzer et al., 2022). It is worth noting that the outcome of this stage is to obtain a foundational LLM with multilingual capabilities, where we can conduct the training process ourselves or obtain pre-trained models from the open-source community.

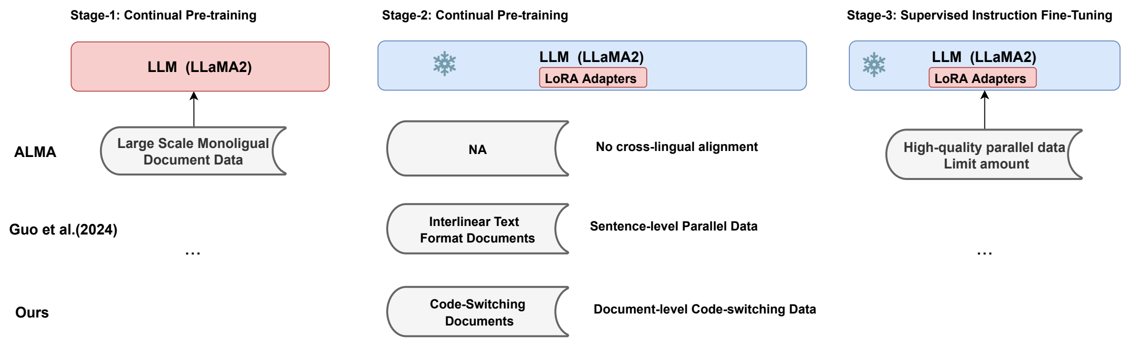

Figure 2: Training process for our and similar works. Overall, we use a three-stage training recipe. And by using Code-Switching strategy, we provide rich cross-lingual alignment information to solve the problems faced in previous works.

Stage-2: Code-Switching Continual Pretraining (CS-CPT) Since our main objective is to enhance the translation capabilities of LLM, cross-lingual alignment information holds significant importance. During the initial training phase, the absence of explicit cross-lingual alignment information necessitates the LLM to learn implicitly, which may not be the most efficient method. We mitigate this issue by performing Continual Pre-training on Code-switching data (presented in 3.1). And CS-CPT offering three key advantages:

- Code-Switching data explicitly provides cross-lingual contextual dependencies, which can compel the model to learn semantic-level alignment relationships.
- Code-Switching data is essentially semantically coherent document data, which maintains consistency with the standard pretraining data format and can alleviate catastrophic forgetting.
- It only requires an additional pair of traditional MT models, making resource consumption and complexity controllable.

We use LoRA technology to carry out the training in this stage, but set the $embed\_tokens$ and lm_head modules to be updatable so that the model can learn token-level alignment information. It is worth mentioning that, as the data pattern is consistent with the first stage, we can even merge them together for training. We also validated this point in the subsequent experimental section.

Stage-3: High-Quality Data Fine-tuning In previous research on adapting LLMs to downstream tasks, it has been confirmed that the quality of data during the SFT phase is more important than the quantity (Zhou et al., 2023; Maillard et al., 2023; Gunasekar et al., 2023) of data. Following the settings of previous works ALMA and Guo et al. (2024), we use a small amount of high-quality bilingual data to fine-tune the model in order to enhance its translation capabilities. To ensure data quality, we collect human-written datasets from WMT development and test sets. We also employ LoRA for training.

4 Experiments

We mainly tested our algorithm on translation tasks in four directions in two language pairs: English-Chinese and English-German. Our experiment design closely follows ALMA to ensure a fair comparison.

4.1 Datasets and Evaluation Metrics

The monolingual dataset we used is sourced from OSCAR. Since the base model we chose (Chinese-LLaMA-2 (Cui et al., 2023)) has already undergone the first stage of pretraining in Chinese, we selected only 0.5B of Chinese and English data from the OSCAR dataset for the second stage of training. For the English-German translation task, we opted to pretrain with 1.5B of German and English monolingual data (the average number in the ALMA's experiments) and similarly used 0.5B for the second stage of training.

For our parallel training data, we collect human-written test datasets from WMT'17 to WMT'20 for EN⇔ZH and EN⇔DE resulting in a total of 37.6K training examples across all languages.

Furthermore, we include the test sets from the WMT22 competition, which are thoughtfully curated to encompass recent content from various domains like news, social media, e-commerce, and conversations.

Models	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
	SoTA models							
NLLB-54B(Team et al., 2022b)	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., 2023a)	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
Guo et al. (2024)	31.14	84.70	30.50	85.62	22.20	79.88	41.10	**86.37**
Parrot-13B(Jiao et al., 2023a)	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., 2023b)	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
Guo et al. (2024)	32.24	85.17	32.53	**86.14**	23.10	80.53	42.30	**86.65**
	Traditional Back Translation Model							
NLLB-distilled-600M-Finetune	26.80	78.53	30.01	85.07	19.72	74.89	33.24	80.76
Ours	Our Recipe with Backbone Model: LLaMA2(Touvron et al., 2023)							
7B Stage1,3	30.05	84.07	30.21	85.55	23.96	79.62	35.31	84.74
7B Stage1,2,3	**31.64**	**85.01**	**31.20**	**85.71**	**26.87**	**80.44**	**41.81**	86.12
13B Stage1,3	31.20	84.43	31.30	85.77	24.31	80.01	37.34	85.27
13B Stage1,2,3	**32.74**	**85.48**	32.49	**86.20**	**27.16**	**81.06**	**42.84**	**86.63**

Table 1: **The main results.** Bold numbers represent the best scores among prior similar studies. After integrating CS-CPT, the translation quality of the model has been significantly improved. Our 7B and 13B models have achieved top performance in most evaluation metrics compare to similar studies. Even the BLEU score for the Zh⇒En direction is on par with that of GPT-4.

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. We rely more on COMET than BLEU due to its better alignment with human evaluations (Freitag et al., 2022).

4.2 Training Setup

Our experiments were carried out using HuggingFace Transformers[1] with open-source LLaMA (Touvron et al., 2023) family as our foundation model. Most of our verification experiments were conducted on 7B model, but we will also report the results of the 13B model to explore the impact of model size.

Specifically, we chose to use Chinese-LLaMA2 (Cui et al., 2023) as the base model for our training because it handles Chinese more efficiently (using expanded vocabulary) and has already completed the first stage of training in Chinese. Building on this foundation, we can proceed with the second and third stages of training for Chinese tasks. For German tasks, we will execute the training of the first and second stages together.

In the training of the first and second stages, we use the LoRA approach to adapt the key, query, value, and output layers of the self-attention mechanism, and the LoRA hyperparameters are set to $R = 32$ and $a = 64$. At the same time, the modules $embed\_tokens$ and lm_head are also set as updatable parameters. We fine-tune the foundation model for one epoch using a batch size of 256, a warm-up ratio of 0.01, and sequences with a maximum of 1024 tokens in total.

During the third stage of training, we follow the ALMA's approach by updating only 0.1% of the parameters using LoRA. We train the model for 2 epochs and select the best model based on the lowest validation loss. For both stages, we adopt deepspeed (Rasley et al., 2020) to accelerate our training.

We employ the NLLB-600M-distil [2] as our traditional MT model for BT pseudo-corpus. Additionally, we leveraged training data from WMT21 to improve the translation quality for the target lan-

[1] https://huggingface.co/docs/transformers/en/index

[2] https://github.com/facebookresearch/fairseq/tree/nllb

guages German and Chinese, thereby ensuring the fundamental quality of the BT pseudo-corpus.

4.3 Baselines

We compare our method against two baseline categories. Firstly, we examine previous studies that share our objective of utilizing LLMs for translation. Secondly, we assess against the latest state-of-the-art translation models.

Prior Similar Work We compare our model with BigTranslate (Yang et al., 2023), which extends LLaMA-1-13B to over 100 translation directions; TIM (Zeng et al., 2023), which uses correct and incorrect examples to help LLM to learn translation; SWIE (Chen et al., 2023), which improves LLM in translation via instruction augmentation; ParroT(Jiao et al., 2023a), through three types of instructions improves the translation performance of LLM after SFT; and BayLing (Zhang et al., 2023b), which uses interactive translation instructions; and ALMA (Xu et al., 2023), a two-stage finetuning method that initially fine-tunes on monolingual data and subsequently on a small set of high-quality parallel data; and Guo et al. (2024), expand on ALMA's approach by introducing an additional stage for fine-tuning with parallel sentences with Interlinear text format.

SoTA Models We focus on the NLLB-54B model, the top-tier translation model in the NLLB family (Team et al., 2022a), as well as the zero-shot capabilities of GPT3.5-text-davinci-003 (**GPT-3.5-D**) and GPT-3.5-turbo-0301 (**GPT-3.5-T**), along with GPT-4[3].

5 Results

Main Results Table 1 summarizes the main results of our experiments. In summary, our final optimized model has shown consistent improvement in translation quality, surpassing ALMA in both BLEU and COMET metrics. The improvement in the Chinese translation task is greater than that in the German task, and the BLEU metric for ZH⇒EN task even on par with GPT-4. Compared to similar works, the 13B model has achieved a leading position in most metrics.

Effectiveness of Code-Switching Continual Pretraining The training in the second stage indeed improved the model's translation ability. Taking Chinese tasks as an example, the COMET scores for ZH⇒EN and EN⇒ZH improved by 0.82 and 1.38, while BLEU scores improved by 2.91 and 6.5, respectively. For the German task, the overall trend is consistent with Chinese, but the improvement is slightly smaller relative to Chinese. This may be because the alignment information in the foundation models for German and English is richer compared to Chinese (with a higher character overlap rate).

Compared with Prior Similar Studies Compared to the strong baseline ALMA, our 7B model achieved an average BLEU improvement of 2.88 and a COMET improvement of 0.73. Our results exceed those of Guo et al. (2024) in the tasks for ZH⇒EN and DE⇒EN, but are on par with theirs in the EN⇒ZH and EN⇒DE directions. But it is important to note that we did not use parallel corpora in our training process. Moreover, unlike the OSCAR data that we employed, they utilized the WMT bilingual training data, which is more closer to the domain of the current test set.

6 Analysis

In this chapter, we will analyze the key points of the model. Some analysis experiments will be conducted on Chinese tasks because Chinese and English have relatively greater linguistic distances.

6.1 Cross-lingual alignment analysis

To verify whether our model in the second stage has achieved the goal of cross-lingual alignment, we referenced relevant works (Lin et al., 2020) and conducted quantitative analyses in two dimensions. Firstly, we calculated the similarity of word embeddings for words with the same meanings in different languages. We selected the top 1000 most frequent words according to the MUSE [4] dictionary. We averaged the sub-word sequences of words to obtain word embeddings and calculated the cosine similarity between the two languages. Additionally, we analyzed representations at the sentence level for sentences with the same meanings. We used the Flores (NLLB Team, 2022) test set to calculate sentence-level embeddings using the same method and computed the corresponding similarities. The results of stage-1 and stage-2 pretraning models are summarized in Figure 3. From the figure, it is evident that, both word and sentence-level similarities have significantly improved after our CS-CPT, regardless of whether it is language pairs with rela-

[3] GPT-3.5-D, GPT-3.5-T and GPT-4 results are sourced from Xu et al., 2023

[4] https://github.com/facebookresearch/MUSE

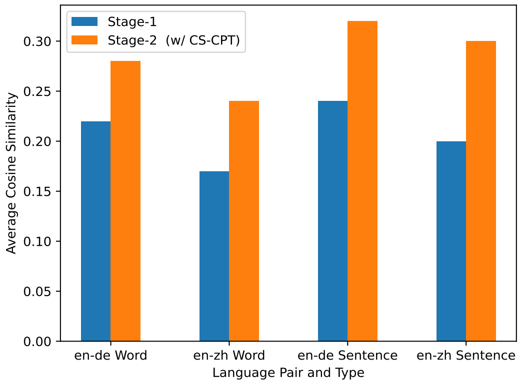

Figure 3: The average cosine similarity results of models from various stages are sourced from the 7B version. We observe an increase in similarity after the second training stage, affirming the effectiveness of our training approach.

tively close distances like EN-DE or distant pairs like Chinese-English. This once again proves that CS-CPT can indeed serve the intended purpose, aligning the model's cross-lingual representations to some extent.

6.2 Using of traditional MT models

When creating Code-Switching data, we introduced a of traditional sentence-level MT model to ensure the method's versatility and overcome challenges in obtaining document-level parallel corpora or document-level MT models. The results in Table 1 indicate that they did not achieve higher translation quality in terms of BLEU and COMET scores compared to the first-stage model. This finding dismisses the idea of LLM gaining knowledge via distillation from pseudo-corpus affirms that the model acquired alignment information beneficial for translation from the Code-Switching data after training in the second stage.

6.3 FT is more effective than BT?

Back translation is more effective than forward translation during the optimization of traditional machine translation models. For instance, when optimizing the ZH⇒EN model, the pseudo-corpus in the EN⇒ZH direction is typically more effective. This is because back translation introduces a large amount of monolingual data for the target language side, enhancing the generation capability of the target language (Edunov et al., 2018). With LLMs having already learned a significant amount of monolingual data during the pre-training phase, the target language's generation ability is already

Models	ZH⇒EN		EN⇒ZH	
	BLEU	CMT	BLEU	CMT
7B Stage-1,3	23.96	79.62	35.31	84.74
7B Stage-1,2,3	26.87	80.44	41.81	86.12
Only D_{cs}^{zh2en}	26.30	80.10	38.01	85.32
Only D_{cs}^{en2zh}	24.70	79.80	39.17	85.60

Table 2: Comparative experimental results of code-switching data between BT and FT. "Only D_{cs}^{zh2en}" means using only ZH⇒EN Code-Switching data for training stage-2.

strong. Does this conclusion still hold when adapting LLM to translation tasks?

To explore this, we conducted comparative experiments on Chinese tasks. Specifically, in our CS-CPT stage, we only used code-switching data in one direction, then obtained the final translation model through the third stage of SFT. The results are summarized in Table 2. We were surprised to find that the improvement brought by forward translation is significantly better than that of back translation. Taking ZH⇒EN task as an example, using only ZH⇒EN direction code-switching data resulted in an improvement ratio of over 70% compared to using a mixture of data from both directions, while the improvement ratio for the quality of EN⇒ZH task was only around 25%. The overall trend for EN⇒ZH task is similar, just not as pronounced as with ZH⇒EN task.

We speculate that apart from bringing benefits in cross-lingual alignment, the forward translation data has also boosted the Automatic Post-Editing (APE) capability of LLM. During the CS-CPT stage, we used special tokens to mark code-switching pseudo data, guiding the model to differentiate between real and pseudo data. In the final SFT stage, the humans-written parallel data inspired the model to output sentences that lean towards real data during translation. By comparing these two types of data, LLM has improved its ability to rewrite machine-translation results into more natural and fluent translations.

To validate our speculation, we conducted a simple test on the APE capabilities of the models from the first and second stages. Specifically, we used the traditional MT model to translate the test set and obtained machine-translation results, then generated APE results using the 3-shot learning. Evaluation results are summarized in Table 3. The APE ability of the second-stage model is stronger than that of the first-stage model, with an aver-

Models	ZH⇒EN		EN⇒ZH	
	BLEU	CMT	BLEU	CMT
NLLB-distil	19.72	74.89	33.24	80.76
Stage1 + APE	20.20	75.24	33.56	82.11
Stage1,2 + APE	20.41	75.78	33.61	82.37

Table 3: Results of APE ability tests for pre-trained models at different stages. The results are all from the 7B version of the model, and the testing method is 3-shot learning. "NLLB-distil" is our traditional MT model used for translating BT pseudo-corpus.

Models	ZH⇒EN		EN⇒ZH	
	BLEU	CMT	BLEU	CMT
7B Stage1,2,3	26.87	80.44	41.81	86.12
CPT-InterLinear	24.12	79.79	37.87	85.54
+ 5-Epoch	23.87	79.36	37.88	85.46
SFT + BT	23.45	78.86	34.57	83.15
+ Full Data	23.01	78.49	34.55	82.71

Table 4: Results of ablation experiments. "CPT + InterLinear" represents replacing D_{cs} with data in InterLinear text format. "SFT + BT" means using BT translation data to replace human-writing data for the stage-3 training with equal data volum. "Full Data" denote using all the BT data.

age COMET improvement of over 0.4 for the final translation results. Further in-depth exploration will be left for future research.

6.4 Ablation for BT Pseudo-Corpus

If we follow the previous work and directly use BT pseudo-corpus in the CPT or SFT stage, how would it compare to the current Code-Switching strategy? To verify this question, we conducted a series of ablation experiments. Firstly, following Guo et al. (2024), we replaced the Code-Switching data with InnerLinear formatted data for the second-stage pretraining and also extended the training time to explore the issue of catastrophic forgetting. Next, we bypassed the second stage and utilized BT pseudo-corpus in the SFT phase, experimenting with varying amounts of data. The results are summarized in Table 4.

From the experimental results, we can draw the following conclusions:

- It is not wise to introduce BT pseudo-corpus in the SFT stage. The improvement in translation quality is not as good as that of human-written data, which aligns with previous findings.
- Using data in InnerLinear Text Format in the second stage can bring limit improvement, and there is a certain gap compared to the Code-Switching strategy in terms of BLEU and COMET metrics. Moreover, as the training time increases, the model indeed exhibits the issue of catastrophic forgetting, with a significant decline in translation quality in the ZH⇒EN direction.

7 Conclusion

In this paper, we focus on the research of adapting the translation capabilities of large models. We attempt to inject cross-lingual alignment information into LLM during the pretraining phase through a Code-Switching strategy, thereby expanding the classic two-stage training recipe. Experiments show that our Code-Switching data constructed based on the back translation strategy achieves desirable results, enhancing the end-to-end translation quality of LLMs. Additionally, we also find in our new training recipe, forward translation data seems to be more efficient, and the model's APE capability may also benefit from the new training stage. Our Code-Switching strategy and the introduction of traditional MT models in the form of back translation into the optimization work of LLM-based translation models may inspire future research to some extent.

8 Limitations

The code-switching data format is consistent with the standard pre-training data format. In theory, we can further increase the amount of monolingual data for additional optimization. This aspect of work needs to be further explored in the future.

Current experiments and analyses are based on translation tasks centered around English. Extending our strategies to non-English translation tasks is also worth further research and optimization.

A more in-depth analysis of the principles behind the effectiveness of Code-Switching data and the internal changes in the model will lead to more meaningful discoveries.

References

Anthropic. claude.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions. *Preprint*, arXiv:2308.12674.

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*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. *Preprint*, arXiv:1908.05204.

Common Crawl Foundation. 2023. Statistics of common crawl monthly archives by commoncrawl. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

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.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models tuned with human translation and feedback. *Preprint*, arXiv:2304.02426.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pretraining multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

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)*, pages 2740–2756.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

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 Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

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.

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.

Hieu Pham, Xinyi Wang, Yiming Yang, and Graham Neubig. 2021. Meta back-translation. In *International Conference on Learning Representations*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

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.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Weiting Tan, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2023. Narrowing the gap between zero-and few-shot machine translation by matching styles. *arXiv preprint arXiv:2311.02310*.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022a. No language left behind: Scaling human-centered machine translation (2022). *URL https://arxiv. org/abs/2207.04672*.

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. 2022b. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

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. *Preprint*, arXiv:2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

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

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636.

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. *arXiv preprint arXiv:2210.02414*.

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

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang

<Chinese> (千站云繁殖池是蜘蛛池的升级版,轮链繁殖池(每一个池可以放10万链接) 2、查一下网站是不是优化过度了,如果网站优化,过度蜘蛛是会对网站进行屏蔽的。**<Chinese>** (目录繁殖池(可日租周租月租),联系官方qq: 3、网站设计的结构不合理,如果网站的页面独立、互相没有关联也会引致收录不理想的。**<English>** (Address breeding pool (Baidu, 360, Sogou, are independent pool) 4), remember to often check the content on your website is not there some sensitive words, involving some sensitive issues search engine is not included. **<Chinese>** <(参数繁殖池(我们引的都是有效蜘蛛) 5、要时常分析网站内容。 **<Translation> <English>** (Parameter breeding pool (we are citing effective spiders)) 5. Analyze website content from time to time.> **<Chinese>** 如果网站上的内容都是采集来的,或者是在其他网站上摘抄的类似内容,蜘蛛往往是不收录的。**<English>** Remember to check the stability of the server or space, and check whether there is improper code in your website program.

Figure 4: An example of Code-Switching data from Chinese to English direction.

Feng. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *Preprint*, arXiv:2306.10968.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

A Code-Switching Data Details

The special tokens used in constructing Code-Switching data include LID and TRANS. Among them, LID consists of the language name enclosed in angled brackets, with "<Chinese>", "<English>", and "<German>" representing Chinese (LID_{zh}), English (LID_{en}), and German (LID_{de}) respectively. TRANS is "<Translation>".

An example of Code-Switching data from Chinese to English direction is shown in Figure 4.