Language Adaptation of Large Language Models: An Empirical Study on LLaMA2

Shumin Wang[1,*] **Yuexiang Xie**[2] **Bolin Ding**[2] **Jinyang Gao**[2] **Yanyong Zhang**[1,†]
[1]University of Science and Technology of China [2]Alibaba Group
shuminwang@mail.ustc.edu.cn {yuexiang.xyx, bolin.ding, jinyang.gjy}@alibaba-inc.com
yanyongz@ustc.edu.cn

Abstract

There has been a surge of interest regarding language adaptation of Large Language Models (LLMs) to enhance the processing of texts in low-resource languages. While traditional language models have seen extensive research on language transfer, modern LLMs still necessitate further explorations in language adaptation. In this paper, we present a systematic review of the language adaptation process for LLMs, including vocabulary expansion, continued pre-training, and instruction fine-tuning, which focuses on empirical studies conducted on LLaMA2 and discussions on various settings affecting the model's capabilities. This study provides helpful insights covering the entire language adaptation process, and highlights the compatibility and interactions between different steps, offering researchers a practical guidebook to facilitate the effective adaptation of LLMs across different languages.

1 Introduction

The popularity of Large Language Models (LLMs), such as GPT-4 (OpenAI and et al., 2024), LLaMA (Touvron et al., 2023a,b; Abhimanyu Dubey and et al., 2024), Mistral (Jiang et al., 2023), and so on (Chowdhery et al., 2023; Jiang et al., 2024; Li et al., 2023; Yang et al., 2023; Bai et al., 2023), has witnessed skyrocketing increase in recent years. It is worth noting that a vast majority of LLMs are primarily trained on English corpus data, with only limited allocations to texts in other languages. For example, in the training corpus of LLaMA2 (Touvron et al., 2023b), English constitutes a substantial 89.7%, while other prevalent languages, such as Chinese and French, account for less than 0.2%. The dominance of English in training data can introduce *language bias*: LLMs skew the performance towards English texts and show diminished effectiveness when processing non-English languages (Navigli et al., 2023).

Training LLMs from scratch with a multilingual balanced corpus entails several practical challenges. Firstly, many non-English languages suffer from a lack of sufficient high-quality data, often requiring billions or even trillions of tokens across diverse data types, such as web texts, books, articles, and so on. Constructing such training data is crucial for LLMs to achieve satisfactory performance (Shen et al., 2023), necessitating substantial efforts from experts in data collection and preprocessing. Secondly, the computational resources needed to train LLMs from scratch can be unaffordable for many organizations and research groups. Last but not least, the linguistic diversity represented in the corpus can bring unique complexities in training LLMs (Conneau et al., 2019). The wide range of linguistic rules and structures across different languages makes it much more challenging to develop LLMs that are truly balanced and effective.

As a result, researchers (Jiang et al., 2023; Cui et al., 2023) propose to apply language adaptation techniques on the well-trained LLMs to enhance their capacity in processing low-source languages. Previous studies (Gogoulou et al., 2021; Zoph et al., 2016; Lample and Conneau, 2019; Artetxe and Schwenk, 2019; Conneau et al., 2019; Xue et al., 2020) have made remarkable progress in transferring the ability of traditional language models, such as BERT (Devlin et al., 2018), across languages (a.k.a. cross-lingual transfer). However, due to the differences in model architecture and training objectives between LLMs and traditional language models, it remains an open question whether these existing studies can be effectively adopted for LLMs. Specifically, LLMs typically employ causal language modeling as their training task, follow a pretraining-alignment paradigm, and utilize in-context learning for application. These technologies differ from those employed in traditional

[*]Work done as an intern at Alibaba Group.
[†]Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7195–7208
January 19–24, 2025.

language models, which primarily focus on fine-tuning classifiers for specific tasks.

To fulfill this gap, we systematically outline the pipeline of language adaptation for LLMs, including vocabulary expansion, continued pre-training, and instruction fine-tuning. We conduct a comprehensive experimental study to analyze potential training techniques and data construction methods throughout the entire process, investigating the effects of these factors when performing language adaptation on LLMs. Specifically, we conduct language adaptation experiments on LLaMA2 (Touvron et al., 2023b), focusing on the adaptation from English to Chinese, and also supplementing our study with adaptations to other non-English languages, including Italian, Thai, Portuguese, German, French, and Japanese.

We provide helpful insights from the following three perspectives:

- For vocabulary expansion, we experimentally analyze the impact of vocabulary size and initialization methods on model performance, highlighting the importance of these factors in improving overall model performance.
- Regarding continued pre-training, we focus on the constructions of training datasets, investigating the importance of source language texts and the effects of translation texts;
- We explore the relationship between continued pre-training and instruction fine-tuning, highlighting their compatibility. Besides, we offer valuable insights on constructing effective instructions using both source language instructions and translation instructions.

By conducting such a comprehensive empirical study, we aim to provide the research community with a practical guidebook for the language adaptation of LLMs, offering valuable insights and recommendations to assist researchers in effectively transferring LLMs across various languages for real-world applications.

2 Language Adaptation of LLMs

Drawing from previous studies (Artetxe et al., 2019; Lample and Conneau, 2019; Zoph et al., 2016; Zhu et al., 2023; Zhang et al., 2023), the language adaptation pipeline for LLMs typically consists of three key steps: *vocabulary expansion*, *continued pre-training (CPT)*, and *instruction fine-tuning (IFT)*, as illustrated in Figure 1. In this section, we outline the challenges and open questions associated with each step and provide corresponding insights and discussions.

2.1 Vocabulary Expansion

The vocabulary maintains the token units used by the tokenizer within the modern architecture of language models, which is one of the critical factors for the remarkable capability of LLMs to understand and generate human-like text.

In language adaptation of LLMs, the vocabulary of the source model is usually learned from a dataset sampled from the training corpus, therefore it lacks the vocabulary of the target language. For example, the vocabulary of LLaMA2 contains 32,000 tokens, of which only around 300 are Chinese characters. Without containing enough Chinese characters in the vocabulary, LLaMA2 has to encode a single out-of-vocabulary Chinese character with 3-4 Unicodes that have very poor semantic relevance. This phenomenon also exists in other non-Latin languages, such as Thai, Russian, Japanese and others. For other Latin languages such as Italian, German, and French, using a vocabulary trained only in English might also lead to suboptimal model performance (Petrov et al., 2024; Ahia et al., 2023).

As a result, vocabulary expansion techniques are implemented in LLMs' language adaptation to enhance the model's ability in processing the target language, and to improve the encoding efficiency. Specifically, a target-language vocabulary is learned and merged with the original vocabulary to form an expanded one. During the process of vocabulary expansion, two crucial settings are worth noting and should be carefully determined: *the number of expanded tokens* and *the initialization of their token embeddings*.

The Number of Expanded Tokens The effectiveness of LLMs on target languages is closely related to the number of expanded tokens. On the one hand, a small number of expanded tokens might result in limited improvement brought by the vocabulary expansion. On the other hand, a large number of expanded tokens can cause a significant performance drop without continued pre-training on the corpus in target languages, and requires more training resources to achieve the model convergence.

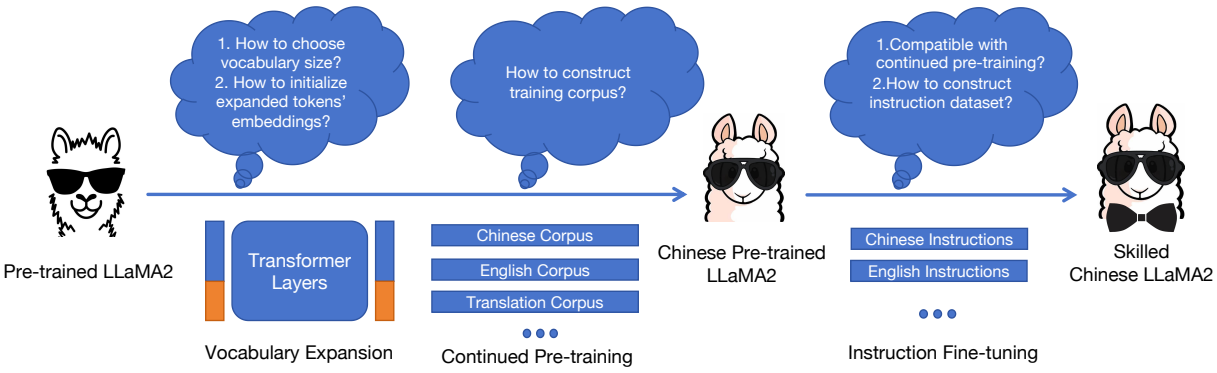

Figure 1: Overall pipeline and questions of language adaptation (using En-Zh of LLaMA2 as an example) of LLMs.

The Initialization of Token Embeddings Effectively initializing the embeddings of the expanded tokens helps to capture the relationship between these expanded tokens and original tokens in vocabulary, providing a suitable starting point for continued pre-training. The naive approach is applying random initialization, which indicates learning the representations of expanded tokens from scratch.

Recently, advanced approaches have been proposed to initialize token embeddings based on the bilingual model or the overlap tokens between languages (Minixhofer et al., 2021; Dobler and de Melo, 2023). To be more specific, a mapping from the expanded tokens to the original tokens is built by recognizing related words between the source and target language. Besides, benefited from applying the Byte-level BPE (Wang et al., 2020) tokenizer, LLMs like GPT-2 (Lagler et al., 2013) and LLaMA2 (Touvron et al., 2023b) are able to encode out-of-vocabulary words by Unicode. When the embeddings of these Unicodes are trained based on the corpus containing target language texts, it is also a feasible solution to initialize the expanded tokens with the aggregation of the corresponding Unicode-encoded tokens.

An empirical study of the aforementioned initialization methods in vocabulary expansion can be found in Section 3.2 and 3.3.

2.2 Continued Pre-training

Continued pre-training in language adaptation involves loading the parameters of trained LLMs and conducting pre-training tasks (such as next-word prediction) using a corpus that includes texts in the target language. While it is evident that continued pre-training can lead to significant improvements in both understanding and generation for the target language (Zeng et al., 2023; Colossal-AI, 2023), the construction of a suitable dataset emerges as a critical problem that needs to be further explored.

For dataset construction in the pre-training of LLMs, existing studies primarily focus on the quality and diversity of data (Li et al., 2023; Lee et al., 2023), both of which are widely recognized as critical factors in model performance. In this paper, we provide explorations and discussions on some additional important yet under-explored aspects that are integral to the language adaptation process.

Translation Data One critical factor is the usage of translation data, which can promote alignment between source and target languages. Previous studies (Lample and Conneau, 2019; Zhang et al., 2023) have pointed out the value of translation data in the cross-lingual transfer of traditional language models. However, the effectiveness of such translation data in language adaptation for modern LLMs, which employ causal language modeling as their training task, remains unclear. This gap calls for further investigations.

Texts in Source Language Inspired by previous studies (Wang et al., 2024; Ye et al., 2023), multilingual pre-trained language models can implicitly align different languages even without the need for multilingual parallel data. It is also recommended to incorporate a mixture of texts in both the source and target languages within the dataset to mitigate catastrophic forgetting (Kirkpatrick et al., 2017). These phenomena suggest that source language texts may significantly impact language adaptation, and the quality of these source texts is an important yet under-explored issue.

We provide experimental results and analysis of these two factors in Section 3.4 and 3.5.

2.3 Instruction Fine-tuning

Instruction fine-tuning (Wei et al., 2021) is a promising approach for aligning pre-trained LLMs with their downstream applications based on various instruction data. In the context of language adaptation on LLMs, we propose to investigate two questions that are general and highly impactful: *the compatibility of instruction fine-tuning and continued pre-training*, and *the construction of instruction datasets*.

Compatibility of Instruction Fine-tuning and Continued Pre-training Previous studies (Zhu et al., 2023; Zhang et al., 2023) have demonstrated that direct conducting instruction fine-tuning on pre-trained LLMs (e.g., LLaMA) with target language instructions can effectively endow the model with target language abilities, it is questionable whether conducting continued pre-training on general-purpose data before instruction fine-tuning would provide additional benefits, especially when training resources are limited.

Meanwhile, Ye et al. (2023) shows that languages that LLaMA has limited exposure to during pre-training exhibit better language adaptation potential during instruction fine-tuning. This inspires us to investigate whether continued pre-training provides this benefit for instruction fine-tuning.

Construction of Instruction Dataset Previous studies (Pires et al., 2019; Wu and Dredze, 2019) demonstrate that Multilingual-BERT, a multilingual variant of BERT (Devlin et al., 2018), can be fine-tuned for a specific task in one language and successfully perform the same task in another language. Meanwhile, a recent study (Ye et al., 2023) shows that generative pre-trained LLMs can enhance their downstream performance in one language by utilizing instruction fine-tuning on another language. Drawing inspiration from these findings, we consider that the model's ability to process the target language may benefit from instructions provided in the source language. Therefore, an empirical study and further discussions are provided in Section 3.7 to show the model performance when varying the ratio of different language texts in the instruction data.

Furthermore, recent studies (Zhang et al., 2023; Zhu et al., 2023) propose to use translation instructions between the source and target languages during instruction fine-tuning to assist the model in aligning instructions from both languages. However, they do not explore whether this approach is beneficial for tasks beyond translation, which remains a valuable area for further research.

The aforementioned experimental results and discussions can be found in Section 3.6 and 3.7.

3 Experiments

3.1 Settings

Model & Language In the experiments, we use LLaMA2-7B (hereinafter referred to as LLaMA2) as the initial model, English as the source language, and Chinese as the main target language. We conduct supplementary experiments using Italian, Thai, Japanese, Portuguese, German, and French. The related details can be found in Appendix A. We use the same hyperparameters as those in the pre-training of LLaMA2, except for reducing the learning rate to half of its original value. More implementation details can be found in Appendix B.

Datasets For continued pre-training, we construct a general-purpose corpus consisting of Chinese and English texts in a ratio of 3:1. Unless otherwise specified, we use CCI corpora (BAAI, 2023) as the source of Chinese corpus, and RedPajama (TogetherComputer, 2023) as that of English corpus. For instruction fine-tuning, we select the English QA dataset provided by FLAN (Wei et al., 2021) and the subset of Chinese in the multi-lingual QA dataset provided by XP3 (Muennighoff et al., 2022) as the instruction dataset. The details of datasets we use for other languages are introduced in Appendix A.

Evaluation Comparisons are conducted mainly on C-Eval (Huang et al., 2023) evaluation set, which contains 13,948 Chinese multiple-choice questions covering 52 different subjects and is a widely used benchmark dataset to comprehensively evaluate the model's ability to process Chinese texts. The models are evaluated in a 5-shot manner with the officially recommended prompts, as shown in Appendix C. For languages other than Chinese, we employed two testing metrics, M3Exam and MGSM, more detailes are provided in Appendix A.

To alleviate the uncertainty caused by the 5-shot in-context learning evaluation mode on the experimental results, we generate different contexts by adjusting the order of example questions in the prompt. Each result is tested three times, and the average value is reported as the final score.

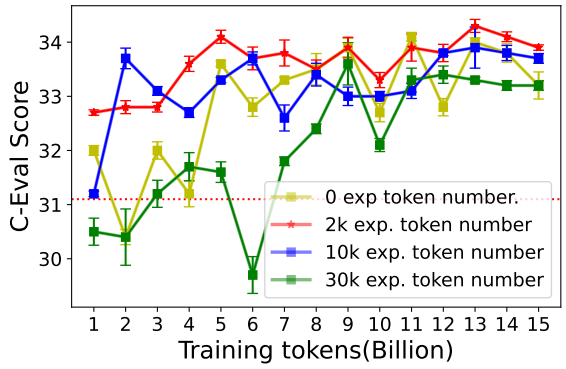

Figure 2: The change of training loss and C-EVAL score during continued pre-training w.r.t. extending different numbers of Chinese words in the vocabulary. Numerical results are in Table 5 in Appendix G.

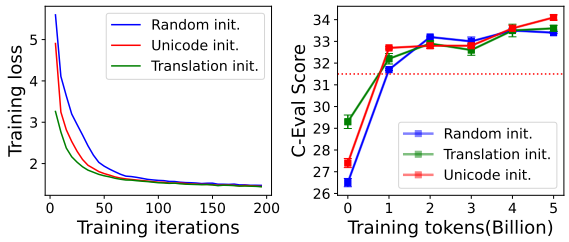

Figure 3: The training loss and C-EVAL score during continued pre-training w.r.t. different initialization of the embeddings of expanded words. Numerical results are in Table 6 in Appendix G.

3.2 The Number of Expanded Tokens

We implement the Byte-level BPE algorithm via SentencePiece package (Kudo and Richardson, 2018) to train a Chinese vocabulary on CCI dataset, and select (following the order generated by BPE) 2k/5k/30k to expand the original vocabulary from LLaMA2. Therefore, the sizes of the expanded vocabulary become 34k/37k/62k, respectively. We initialize these expanded tokens with Unicode initialization (please ref Section 3.3) The corpus for continued pre-training, as introduced in Section 3.1, consists of a mixture of Chinese and English texts amounting to 15 billion tokens.

The evaluation results on the C-Eval benchmark, shown in Figure 2, demonstrate the effectiveness of vocabulary expansion. Specifically, appropriately expanding the vocabulary (denoted as "2k exp. token number" in the figure) before training can achieve better performance improvements compared to training without expanding the vocabulary. It can also be found that massively expanding the vocabulary can significantly harm model performance, especially when the training data is limited (less than 10 billion tokens). Therefore, for low-resource languages, it is crucial to expand the vocabulary carefully. In settings similar to ours, we empirically recommend 2k tokens as the scale for vocabulary expansion. We also recommend conducting a preliminary experiment to determine the scale of vocabulary expansion before executing large-scale CPT.

3.3 Initialization of Token Embeddings

We conduct experiments with three different approaches to initialize the embeddings of the expanded tokens, and then perform continued pertaining with the same settings for a fair comparison. Specifically, the training dataset consists of 5 billion tokens and the size of the expanded tokens is 2k. For more details on initialization approaches: (i) **Random Initialization**: The embeddings of the expanded tokens are drawn from a normal distribution; (ii) **Unicode Initialization**: Benefited from BPE algorithm, LLaMA2 can encode out-of-vocabulary token as several Unicodes accordingly. Thus we can initialize the expanded tokens using the mean of embedding vectors corresponding to Unicode-encoded tokens. (iii) **Translation Initialization**: Inspired by previous studies (Minixhofer et al., 2021), we use the average of the embedding vectors corresponding to the English translations of Chinese words.

We show the training loss curve and the evaluation results on C-Eval benchmark in Figure 3. Based on C-Eval scores, translation initialization can endow the model with certain Chinese proficiency even without training, and it may hold an advantage when training data is extremely scarce (less than 1B tokens). This aligns with the performance of loss curves during model training and is consistent with research in traditional language models (Dobler and de Melo, 2023; Minixhofer et al., 2021). However, unlike traditional language models, as the training data adopted for LLMs increases, the performance differences between different initializations become very small.

For further investigation, for models utilizing different initialization methods, we calculate the average cosine similarity of the embedding layers of the expanded tokens, respectively. The experimental results shown in Table 1 indicate that, during the training process, the text representations across different models tend to be similar. This phenomenon suggests that, although different ini-

#Tokens	Uni.-Rand.	Uni.-Trans.	Rand.-Trans.
0B	5e-5	0.11	2e-5
2B	0.08	0.29	0.02
5B	0.13	0.37	0.10

Table 1: Cosine similarity between expanded tokens' embeddings of different models. Uni. refers to models whose token embeddings are initialized with Unicode, Rand. refers to models whose token embeddings are initialized randomly while Trans. refers to models whose token embeddings are initialized by translation.

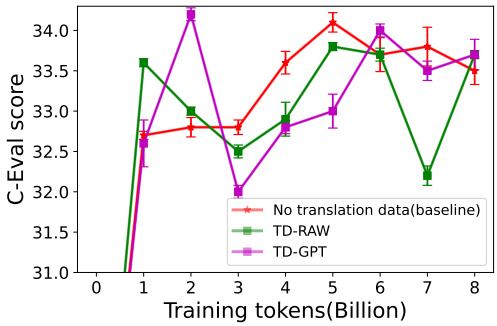

Figure 4: The C-Eval scores of the model w.r.t. adding different translation corpora during the continued pre-training process. Numerical results are in Table 7 in Appendix G.

tialization methods bring differences in textual representations, extensive training on the same dataset guides the models toward learning in similar directions. We believe that after extensive training, the inherent characteristics of the training data itself become the primary factor affecting the model's capabilities and therefore, the impact of different initialization methods on the model would gradually diminish.

3.4 Translation Data in Continued Pre-training

Inspired by previous studies (Lample and Conneau, 2019; Yudong Li et al., 2023), we propose to add some translation data when performing continued pre-training. We conduct experiments with the following open-source **T**ranslation **D**ata: (i) **TD-Raw**: Chinese-English translation data collected from CCMatrix (Schwenk et al., 2019b) and WikiMatrix (Schwenk et al., 2019a). We filter out data with garbled characters and formatting errors and produce 1 billion tokens; (ii) **TD-GPT**: We use *GPT-3.5-turbo* provided by OpenAI to translate a portion of CCI dataset, generating approximately 0.1 billion tokens of translation data.

Due to the fact that CCMatrix and WikiMatrix are primarily composed of short sentences collected from the internet, which differs significantly from the commonly used LLM pre-training corpora (TogetherComputer, 2023), we consider their data quality to be lower. On the other hand, CCI is a large corpus prepared for LLM training, consisting of many high-quality articles, thus we believe that the quality of the data translated from it is higher. To be more specific, the average length of data in TD-Raw is 141.6 tokens while is 3105.4 in TD-GPT. We use TD-Raw to represent translation data collected from the internet which is rich in quantity but poor in quality, and use TD-GPT to represent expert-level translation data which is rich in quality but relatively poor in quantity.

The evaluation results on C-Eval benchmark are shown in Figure 4. From the experimental results, it can be observed that different translation corpora provided a boost to the model's Chinese proficiency in the early stages of training (within 1-3B tokens). Additionally, the ranking of the max gain effect is TD-GPT > TD-Raw, which aligns with the ranking based on data quality rather than quantity. However, as training progresses, the model's performance fluctuates and ultimately converges to a performance similar to that achieved without the addition of translation corpora.

Therefore, for languages with extremely scarce training data (less than 3B tokens), adding high-quality expert-level translation data (even though the scale might be smaller) during continued pre-training is a better choice.

3.5 The Quality of English Texts

The dataset used for continued pre-training is a mixture of texts in both source languages and target languages. While previous studies mainly focus on the data quality of the target language texts, which is proven to be critical for the model ability in processing the target language texts, we conduct an empirical study to show that the data quality of the source language texts is also important.

Specifically, we prepare two datasets for continued pre-training, with the only difference between these two datasets being the English part of the training corpus: one is collected from CCMatrix (Schwenk et al., 2019b), and another is collected from RedPajama (TogetherComputer, 2023). CCMatrix is a multi-lingual translation corpora, containing relatively short bilingual parallel sentences. In contrast, Redpajama is a high-quality English corpus prepared for LLM pre-training with

Models	M3Exam			MGSM			C-Eval	MMLU
	it	th	pt	de	fr	ja	zh	en
High En	**42.5**$_{\pm 0.6}$	**27.3**$_{\pm 0.8}$	**31.8**$_{\pm 0.4}$	**12.7**$_{\pm 1.7}$	**9.2**$_{\pm 0.7}$	**6.4**$_{\pm 1.0}$	**34.1**$_{\pm 0.1}$	41.2$_{\pm 0.1}$
Low En	40.4$_{\pm 0.2}$	22.3$_{\pm 0.9}$	26.9$_{\pm 1.1}$	12.1$_{\pm 0.8}$	5.6$_{\pm 0.3}$	2.7$_{\pm 0.2}$	31.9$_{\pm 0.1}$	37.8$_{\pm 0.1}$
Vanilla	37.7$_{\pm 0.1}$	21.8$_{\pm 0.3}$	96.9$_{\pm 0.9}$	11.2$_{\pm 1.4}$	8.9$_{\pm 0.8}$	4.0$_{\pm 1.7}$	31.1$_{\pm 0.2}$	**42.5**$_{\pm 0.2}$

Table 2: The evaluation score of models trained with different quality of English texts and vanilla LLaMA2. Due to varying differences of the target language from English, it/pt/de/fr models are trained with 2B tokens, while zh/th/ja models are trained with 5B tokens.

Lang.	IFT_{1B}	CPT_{X+1B}	$CPT_X + IFT_{1B}$
zh	31.2$_{\pm 0.2}$	33.0$_{\pm 0.1}$	**34.1**$_{\pm 0.1}$
it	42.0$_{\pm 0.6}$	42.5$_{\pm 0.6}$	**43.3**$_{\pm 1.2}$
th	26.5$_{\pm 0.3}$	26.2$_{\pm 0.5}$	**32.1**$_{\pm 0.3}$
pt	35.4$_{\pm 1.1}$	31.8$_{\pm 0.4}$	**40.9**$_{\pm 0.1}$

Table 3: The model's evaluation scores under different settings. The subscript indicates the number of consumed tokens. X is 5B for Chinese/Thai and 1B for Italian/Portuguese.

carefully collected documents from books, articles, and the internet. Therefore, we believe that Redpajama has better data quality.

We perform continued pre-training on these two datasets with the same settings, and we test the model's English proficiency using the English evaluation metric MMLU. These results shown in Table 2 indicate that using the Redpajama corpus as English data better maintains the model's English performance. Meanwhile, we perform experiments on six languages respectively, and notice that the quality of English text significantly impacts the model's performance in the target language, both for Latin languages (it/pt/de/fr) and non-Latin languages (zh/th). Therefore, ensuring the quality of the source language corpus is crucial when performing language adaptation on LLMs.

3.6 Compatibility of Continued Pre-training and Instruction Fine-tuning

Instruction fine-tuning involves training pre-trained LLMs with supervised data that consists of instructions or demonstrations, which aims to improve the model's ability to understand and generate responses that align with the given instructions for specific tasks. In this section, we provide some experimental results to study the compatibility of continued pre-training and instruction fine-tuning in the language adaptation of LLMs.

We conduct instruction fine-tuning on two kinds of models, one of which has already been continued pre-training on a mixture of target language and English texts with a ratio of 3:1, while the other one has not. Here we construct the instruction datasets based on the XP3x (using the instructions in the target language) and FLAN (consisting of instructions in English), with a ratio of 1:1 to produce 1 billion tokens. An empirical study on the construction of the instruction datasets can be found in Section 3.7.

The experimental results are shown in Table 3. From these results, we observe that the model's performance improves significantly when instruction fine-tuning is following continued pre-training, compared to directly performing instruction fine-tuning or only conducting continued pre-training without instruction fine-tuning. Therefore, for the language adaptation of LLMs, the *pretrain-then-finetune* paradigm should be still considered as one of the suitable approaches, which can be particularly important for low-resource languages where instruction data is limited.

3.7 Instruction Dataset Construction

We investigate how to construct useful instruction datasets for instruction fine-tuning in language adaptation from two aspects: (i) The ratio of the instructions in English (the source language) and the target languages; (ii) Adding some translation instructions for explicitly cross-lingual alignment.

We generate a series of instruction datasets by adjusting the ratio of target language instructions to English instructions. We follow the settings in Section 3.6 to conduct continued pre-training on LLaMA2, and then perform instruction fine-tuning on the resulting models, using these instruction datasets. For comparisons, we also perform instruction fine-tuning on the model without continued pre-training.

The experimental results are shown in Figure 5, from which we can observe that for most languages,

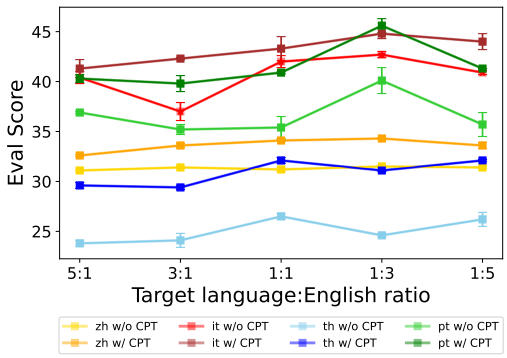

Figure 5: The evaluation score of models' performance after instruction fine-tuning with different English to target language ratio in training data. Numerical results can be found in Table 8 in Appendix G.

Settings	w/o Trans. Data	w/ Trans. Data
w/o CPT	**31.2**$_{\pm 0.2}$	30.9$_{\pm 0.1}$
w/ CPT	34.1$_{\pm 0.1}$	**34.8**$_{\pm 0.1}$

Table 4: The evaluation results (C-Eval score) on the usage of translation instructions during the instruction fine-tuning, with or without prior continued pre-training.

the models can benefit from English instructions and the best ratio between the target language and English is around 1:1 to 1:3. We also find that models have undergone CPT can get more benefits from English instructions(ref Appendix E for further discussion).

Besides, we propose to add translation instructions for explicitly cross-lingual alignment. As a result, we collect 0.1 billion tokens of high-quality translation instructions between English and Chinese (shown in Appendix D) generated by GPT-3.5-turbo, and add these data into the dataset for instruction fine-tuning. The experimental results shown in Table 4 indicate that the usage of high-quality translation data can greatly enhance the model's performance during the instruction fine-tuning process in the target language, only when the instruction fine-tuning following a continued pre-training process.

In summary, in order to effectively apply instruction data for cross-lingual alignment in specific tasks, it is crucial for the models to possess fundamental capabilities in understanding the target languages, which can be acquired through continued pre-training. Note that these observations are consistent with the findings in Section 3.6.

4 Related Works

Cross-Lingual Transfer of language models Researchers (Zoph et al., 2016) focus on transferring the capabilities of traditional language models, such as BERT, in downstream tasks across different languages. For example, Artetxe et al. (2019) trains a multi-language BERT model by only replacing the tokenizer and freezing the transformer parameters except for the embedding layer. Meanwhile, Dobler and de Melo (2023) and Minixhofer et al. (2021) guide the initialization of the target language embedding layer utilizing existing target language tokens in original models or additional bilingual embeddings of source and target languages.

Further, Lample and Conneau (2019) achieves outstanding performance by using translation corpora during continued pre-training. Recently, (Wu et al., 2023) have studied the difficulties of cross-language transfer, and found that compared to the grammatical differences between languages, the vocabulary differences brought by embedding layers have the greatest impact on cross-language transfer. It has also been found that non-Latin languages that cannot share words with English have greater difficulty performing cross-language transfer than Latin languages that can share part of their vocabulary with English(Choenni et al., 2023).

These studies provide some insights for language adaptation of LLMs, but their efficiency on LLMs remains to be studied.

Language Adaptation of LLMs Recently, researchers (Cui et al., 2023; Zhang et al., 2023; Colossal-AI, 2023) have attempted to conduct language adaptation for LLMs, but are limited to exploring a single implementation approach and lack comparative research on different training factors. Some studies have provided empirical research during the instruct fine-tuning process. Ye et al. (2023) reveals the different potentials of different pre-trained models in language adaptation. Zhu et al. (2023) and Shaham et al. (2024) utilize translation data and multilingual instructions to enhance the language adaptation process. Other studies (LLaMA2-ChineseTeam, 2023; Yudong Li et al., 2023) investigated the effect of continued pre-training. However, they do not consider other processes of language adaptation or the impact of the compatibility between different processes on experimental outcomes.

5 Conclusions

The limited ability of LLMs to process low-resource language texts motivates the development of language adaptation, particularly for adapting English-based LLMs to other languages. In this study, we provide a comprehensive review of the language adaptation pipeline for LLMs, focusing on empirical investigations using LLaMA2, and offering valuable insights and practical strategies for various processes, including vocabulary expansion, continued pre-training, and instruction fine-tuning.

Specifically, our observations can be summarized as follows:

- Vocabulary expansion is beneficial, but the number of expanded tokens should be determined carefully. We empirically show that an expansion size of 2,000 words is appropriate for LLaMA2.
- High-quality source language data is crucial for CPT, and high-quality translation data can make improvements in CPT, especially when training data is scarce.
- The combined use of both CPT and IFT results in more effective language adaptation, even with limited CPT data. We recommend employing English instructions alongside translated instruction data during IFT, particularly after prior CPT.

Such an empirical study serves as a guide for researchers seeking to effectively adapt LLMs to target languages, while also inspiring further research aimed at expanding the applicability of LLMs.

Limitations

Due to limitations in (high-quality) training data and computational resources, our experiments primarily concentrated on adapting the language from English to Chinese. Although we have supplemented our key findings with experiments on other languages, where possible, we attempt to conduct more validation across a wider range of languages and utilize more training tokens in future work.

References

Abhinav Jauhri Abhimanyu Dubey and Abhinav Pandey et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *arXiv preprint arXiv:2305.13707*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

BAAI. 2023. Https://data.baai.ac.cn/details/BAAI-CCI.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during llm fine-tuning. *arXiv preprint arXiv:2305.13286*.

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.

Colossal-AI. 2023. Https://hpc-ai.com/blog/one-half-day-of-training-using-a-few-hundred-dollars-yields-similar-results-to-mainstream-large-models-open-source-and-commercial-free-domain-specific-llm-solution.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Konstantin Dobler and Gerard de Melo. 2023. Focus: Effective embedding initialization for specializing pretrained multilingual models on a single language. *arXiv preprint arXiv:2305.14481*.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2021. Cross-lingual transfer of monolingual models. *arXiv preprint arXiv:2109.07348*.

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*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Alycia Lee, Brando Miranda, Sudharsan Sundar, and Sanmi Koyejo. 2023. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. *arXiv preprint arXiv:2306.13840*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

LLaMA2-ChineseTeam. 2023. Https://github.com/LlamaFamily/Llama2-Chinese.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2021. Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. *arXiv preprint arXiv:2112.06598*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*.

OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *Preprint*, arXiv:2401.01854.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

TogetherComputer. 2023. Redpajama: an open dataset for training large language models.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12159–12173, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. 2023. Oolong: Investigating what makes transfer learning hard with controlled studies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3280–3289.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Yuhao Feng Yudong Li, Cheng Hou Zhe Zhao, Wen Zhou Bizhu Wu, Xiaoqin Wang Hao Li, Yaning Zhang Wenhang Shi, Siri Long Shuang Li, Yiren Chen Xianxu Hou, Ningyuan Sun Jing Zhao, and Xiaoshuai Chen Wenjun Tang. 2023. Https://github.com/CVI-SZU/Linly.

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 2023 Conference on International Joint Conference on Artificial Intelligence (AI and Social Good Track)*.

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

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A Settings of experiments with more languages

To further validate the critical conclusions of our experiments, which may vary due to language differences, we conduct additional experiments using four Latin languages (Italian, Portuguese, German, French) and two non-Latin languages, Thai and Japanese (noting that Chinese is a non-Latin language).

Datasets For continued pre-training, we utilize the CC100 (Lin et al., 2022) dataset as the corpus source for the target language and the RedPajama (TogetherComputer, 2023) dataset as the corpus source for English. The CC100 dataset is a large-scale multilingual corpus. However, it is not specifically prepared for pre-training modern LLMs, which may lack in quality and quantity. Therefore, we impose certain limitations on the scale of our experiments. For instruction finetuning, we use the XP3x (Muennighoff et al., 2022) multilingual QA dataset as the source of instruction data for the target language.

Evaluation Due to the lack of comprehensive evaluation metrics for multilingual assessments, we use M3Exam (Zhang et al., 2024) and MGSM (Shi et al., 2022) as our evaluation benchmarks for the model's capabilities in the target languages. M3Exam is a multilingual dataset comprised of local human examination questions. Its format is similar to C-Eval and covers multiple subjects, providing a relatively comprehensive evaluation of LLMs. MGSM is a multilingual math problem dataset derived from the translation of GSM8K, focusing on assessing the model's mathematical abilities across different languages. Since these datasets do not support all languages, we use the M3Exam benchmark for Italian and Thai, and use the MGSM benchmark for German, Japanese, and French.

B Hyper parameters of continued pre-training and instruction finetuning

Hyper param	Value
seq. length	4096
batch size	2048
maxLR	1.5e-4
minLR	1.5e-5
weight decay	0.1
β_1	0.9
β_2	0.95

C Prompt of C-Eval testing

以下是中国关于科目考试的单项选择题，请选出其中的正确答案。
{Question 1}
A. {Choice A}
B. {Choice B}
C. {Choice C}
D. {Choice D}
答案:{Correct Answer}
[k-shot demo, k is 0 in the zero-shot case]
{Test Question}
A. {Choice A}
B. {Choice B}
C. {Choice C}
D. {Choice D}
答案:

D Prompt of transforming bilingual data into translation instructions

Type	Prompt
Zh-En Translation	请将下面的中文句子翻译成英文:\n 中文:{Chinese Document}\n English:{English Document}
En-Zh Translation	Please translate this English sentence into Chinese:\n English:{English Document}\n 中文:{Chinese Document}
Zh example	以下是含义相同的中英文例句:\n 中文:{Chinese Document}\n English:{English Document}
En example	The following are example sentences in English and Chinese with the same meaning:\n English:{English Document}\n 中文:{Chinese Document}
Naive combine	中文:{Chinese Documenr}\n English:{English Document}

E The effect of CPT on English instructions

The experimental results in Table 8 indicate that CPT plays an important role in the utilization of English instructions for target language capability.

We take the ratio of the target language:English = 5:1 as the baseline. For Chinese, the model can get almost no benefit from a higher English ratio

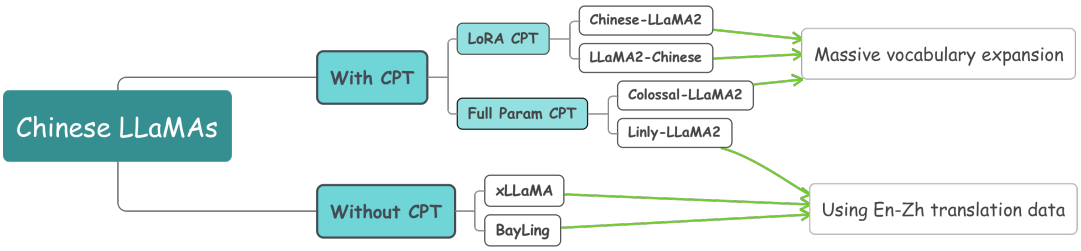

Figure 6: Existing studies on En-Zh cross-lingual transfer for LLaMA or LLaMA2.

without prior CPT. However, with prior CPT conducted, the model's performance gets significant gain with a ratio of 1:1 and 1:3. For Italian and Portuguese, the model gets higher performance gain with CPT (3.5 and 5.3) than without CPT (2.3 and 3.2). For Thai, the performance gain with or without CPT is close, but the best performance occurs in a higher English ratio (Thai:English=1:5) with CPT.

F Existing "Chinese LLaMA" studies

In recent years, several studies have been conducted on the cross-lingual transfer of LLaMA or LLaMA2 into Chinese (as summarized in Figure 6). These studies employ diverse strategies, yielding varying performance outcomes. For example, x-LLaMA (Zhu et al., 2023) and Bayling (Zhang et al., 2023) use bilingual translation corpora in Chinese and English when conducting instruction tuning on LLaMA to enhance LLaMA's ability to follow Chinese instructions. Chinese-LLaMA (Cui et al., 2023), Chinese-LLaMA2 and LLaMA2-Chinese (LLaMA2-ChineseTeam, 2023) expand the vocabulary of LLaMA and LLaMA2 with massive Chinese words (more than 20k tokens) and utilize LoRA (Hu et al., 2021) for continued pre-training and fine-tuning on Chinese corpora to inject Chinese knowledge into the model. Linly-LLaMA2 (Yudong Li et al., 2023) conduct restrained vocabulary expansion (around 8000 tokens) for LLaMA2 and combine unsupervised Chinese corpus, unsupervised English corpus, translation corpus, English instruction data, and Chinese instruction data as its training corpus conducting full-parameter continued pre-training on LLaMA2. Colossal-LLaMA2 (Colossal-AI, 2023) also conducts full-parameter continued pre-training on LLaMA2 with carefully selected Chinese Corpus and achieves impressive results.

These Chinese LLaMA models exhibit significant strategy differences during the training process and do not use a unified data source (some projects utilize large-scale non-public private datasets). Therefore, a fair comparison and study of the impact of various factors on the cross-lingual transfer process with a limited training corpus are hindered.

G Numerical results in the experiments

In this section, we provide the detailed numerical experimental results.

Table 5 shows the Chinese performance of the model during the continued pre-training process after expanding the vocabulary with different scales. The results indicate the effectiveness of vocabulary expansion and show that an expansion scale of 2k words is appropriate for LLaMA2.

Table 6 illustrates how the model performance varies during the CPT process after different embedding initializations. Although there are significant differences in initial model performance, the performance of the different models converges as training progresses.

Table 7 shows the model performance after adding different translation corpora during the CPT process. In the early stages of training, the translation corpora provided significant assistance, but as training progressed, the model's performance became unstable, ultimately approaching the performance of the model without the translation corpora.

Table 8 displays the performance of the model in IFT when experiencing or not experiencing CPT, using different ratios of the target language to English. As discussed in Appendix E, the model is able to benefit more from English instructions after undergoing CPT.

	0 exp.	2k exp.	10k exp.	30k exp.
1B	$32.0_{\pm 0.1}$	$\mathbf{32.7}_{\pm 0.1}$	$31.2_{\pm 0.1}$	$30.5_{\pm 0.2}$
2B	$30.4_{\pm 0.1}$	$32.8_{\pm 0.1}$	$\mathbf{33.7}_{\pm 0.2}$	$30.4_{\pm 0.5}$
3B	$32.0_{\pm 0.2}$	$32.8_{\pm 0.1}$	$\mathbf{33.1}_{\pm 0.1}$	$31.2_{\pm 0.2}$
4B	$31.2_{\pm 0.2}$	$\mathbf{33.6}_{\pm 0.1}$	$32.7_{\pm 0.1}$	$31.7_{\pm 0.3}$
5B	$33.6_{\pm 0.1}$	$\mathbf{34.1}_{\pm 0.1}$	$33.3_{\pm 0.1}$	$31.6_{\pm 0.2}$
6B	$32.8_{\pm 0.2}$	$\mathbf{33.7}_{\pm 0.2}$	$\mathbf{33.7}_{\pm 0.1}$	$29.7_{\pm 0.3}$
7B	$33.3_{\pm 0.1}$	$\mathbf{33.8}_{\pm 0.2}$	$32.6_{\pm 0.2}$	$31.8_{\pm 0.1}$
8B	$\mathbf{33.5}_{\pm 0.3}$	$\mathbf{33.5}_{\pm 0.2}$	$33.4_{\pm 0.2}$	$32.4_{\pm 0.1}$
9B	$33.9_{\pm 0.2}$	$\mathbf{33.9}_{\pm 0.2}$	$33.0_{\pm 0.2}$	$33.6_{\pm 0.4}$
10B	$32.7_{\pm 0.2}$	$\mathbf{33.3}_{\pm 0.1}$	$33.0_{\pm 0.1}$	$32.1_{\pm 0.1}$
11B	$\mathbf{34.1}_{\pm 0.1}$	$33.9_{\pm 0.2}$	$33.1_{\pm 0.1}$	$33.3_{\pm 0.2}$
12B	$32.8_{\pm 0.2}$	$\mathbf{33.8}_{\pm 0.2}$	$\mathbf{33.8}_{\pm 0.1}$	$33.4_{\pm 0.2}$
13B	$34.0_{\pm 0.1}$	$\mathbf{34.3}_{\pm 0.1}$	$33.9_{\pm 0.4}$	$33.3_{\pm 0.1}$
14B	$33.8_{\pm 0.2}$	$\mathbf{34.1}_{\pm 0.1}$	$33.8_{\pm 0.1}$	$33.2_{\pm 0.1}$
15B	$33.2_{\pm 0.2}$	$\mathbf{33.9}_{\pm 0.1}$	$33.7_{\pm 0.1}$	$33.2_{\pm 0.1}$

Table 5: Detailed C-Eval Scores(main/std error) in Figure 2

Training tokens	0B	1B	2B	3B	4B	5B
Random init.	$26.5_{\pm 0.2}$	$31.7_{\pm 0.0}$	$\mathbf{33.2}_{\pm 0.0}$	$\mathbf{33.0}_{\pm 0.2}$	$33.5_{\pm 0.1}$	$33.4_{\pm 0.0}$
Unicode init.	$27.4_{\pm 0.2}$	$\mathbf{32.7}_{\pm 0.1}$	$32.8_{\pm 0.1}$	$32.8_{\pm 0.1}$	$\mathbf{33.6}_{\pm 0.1}$	$\mathbf{34.1}_{\pm 0.1}$
Translation init.	$\mathbf{29.3}_{\pm 0.3}$	$32.2_{\pm 0.2}$	$32.9_{\pm 0.1}$	$32.6_{\pm 0.2}$	$33.5_{\pm 0.3}$	$33.6_{\pm 0.1}$

Table 6: Detailed C-Eval Scores(main/std error) in Figure 3

Training tokens	1B	2B	3B	4B	5B	6B	7B	8B
No trans. data	$32.7_{\pm 0.1}$	$32.8_{\pm 0.1}$	$\mathbf{32.8}_{\pm 0.1}$	$\mathbf{33.6}_{\pm 0.1}$	$\mathbf{34.1}_{\pm 0.1}$	$\mathbf{33.7}_{\pm 0.2}$	$\mathbf{33.8}_{\pm 0.2}$	$33.5_{\pm 0.2}$
TD-Raw	$\mathbf{33.6}_{\pm 0.1}$	$33.0_{\pm 0.1}$	$32.5_{\pm 0.1}$	$32.9_{\pm 0.2}$	$33.8_{\pm 0.1}$	$\mathbf{33.7}_{\pm 0.1}$	$32.2_{\pm 0.1}$	$\mathbf{33.7}_{\pm 0.1}$
TD-GPT	$32.6_{\pm 0.3}$	$\mathbf{34.2}_{\pm 0.1}$	$32.0_{\pm 0.1}$	$32.8_{\pm 0.1}$	$33.0_{\pm 0.2}$	$34.0_{\pm 0.1}$	$33.5_{\pm 0.1}$	$\mathbf{33.7}_{\pm 0.2}$

Table 7: Detailed C-Eval Scores(main/std error) in Figure 4

Language	CPT	5:1	3:1	1:1	1:3	1:5
zh	w/o	$31.1_{\pm 0.1}$	$31.4_{\pm 0.1}$	$31.2_{\pm 0.2}$	$31.5_{\pm 0.1}$	$31.4_{\pm 0.1}$
	w/	$32.6_{\pm 0.1}$	$33.6_{\pm 0.2}$	$34.1_{\pm 0.1}$	$\mathbf{34.3}_{\pm 0.1}$	$33.6_{\pm 0.3}$
it	w/o	$40.4_{\pm 0.6}$	$37.0_{\pm 0.9}$	$42.0_{\pm 0.6}$	$\mathbf{42.7}_{\pm 0.3}$	$40.9_{\pm 0.2}$
	w/	$41.3_{\pm 0.9}$	$42.3_{\pm 0.1}$	$43.3_{\pm 1.2}$	$\mathbf{44.8}_{\pm 0.5}$	$44.0_{\pm 0.8}$
th	w/o	$23.8_{\pm 0.3}$	$24.1_{\pm 0.7}$	$\mathbf{26.5}_{\pm 0.3}$	$24.6_{\pm 0.2}$	$26.2_{\pm 0.7}$
	w/	$29.6_{\pm 0.3}$	$29.4_{\pm 0.2}$	$\mathbf{32.1}_{\pm 0.3}$	$31.1_{\pm 0.1}$	$\mathbf{32.1}_{\pm 0.1}$
pt	w/o	$36.9_{\pm 0.1}$	$35.2_{\pm 0.5}$	$35.4_{\pm 1.1}$	$\mathbf{40.1}_{\pm 1.3}$	$35.7_{\pm 1.2}$
	w/	$40.3_{\pm 0.4}$	$39.8_{\pm 0.8}$	$40.9_{\pm 0.1}$	$\mathbf{45.6}_{\pm 0.7}$	$41.3_{\pm 0.0}$

Table 8: Numerical results in figure 5, the evaluation results w.r.t. using different ratios of the target language to English instructions during the instruction fine-tuning, with or without prior continued pre-training.