Enhancing LLM Language Adaption through Cross-lingual In-Context Pre-training

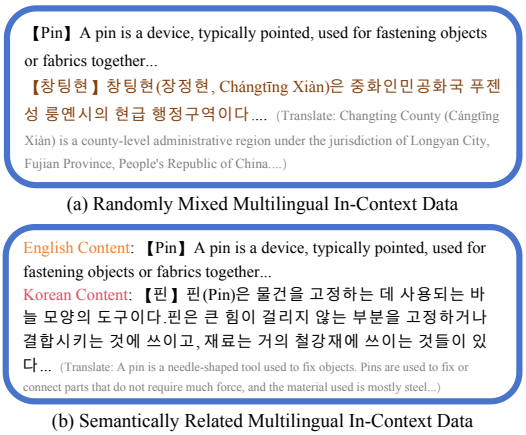
Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit remarkable multilingual capabilities despite English-dominated pre-training, attributed to cross-lingual mechanisms during pre-training. Existing methods for enhancing cross-lingual transfer remain constrained by parallel resources, suffering from limited linguistic and domain coverage. We propose Cross-lingual In-context Pre-training (CrossIC-PT), a simple and scalable approach that enhances cross-lingual transfer by leveraging semantically related bilingual texts via simple next-word prediction. We construct CrossIC-PT samples by interleaving semantic-related bilingual Wikipedia documents into a single context window. To access window size constraints, we implement a systematic segmentation policy to split long bilingual document pairs into chunks while adjusting the sliding window mechanism to preserve contextual coherence. We further extend data availability through a semantic retrieval framework to construct CrossIC-PT samples from web-crawled corpus. Experimental results demonstrate that CrossIC-PT improves multilingual performance on three models (Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-1.5B) across six target languages, yielding performance gains of 3.79%, 3.99%, and 1.95%, respectively, with additional improvements after data augmentation.

1 Introduction

Recent state-of-the-art (SOTA) large language models (LLMs) (Achiam et al., 2023; Anthropic; Reid et al., 2024) have demonstrated remarkable multilingual capabilities. These models are typically pre-trained on massive web-crawled corpora, where English text overwhelmingly dominates in the quantity (Brown et al., 2020; Dubey et al., 2024). However, current LLMs exhibit unexpectedly strong performance on non-English languages that cannot be fully explained by their relative data proportions during pre-training. Researchers have attributed this phenomenon to cross-lingual transfer in LLM training, where linguistic patterns and knowledge acquired from high-resource languages (particularly English) appear to transfer effectively to enhance performance on the other languages (Artetxe et al., 2020; Scao et al., 2022; Wang et al., 2024).



(a) Randomly Mixed Multilingual In-Context Data

(b) Semantically Related Multilingual In-Context Data

Figure 1: Existing works randomly mix multilingual texts (a) in an input window. Our approach groups semantically related texts (b) to enhance cross-lingual transfer.

A series of works have explored methods for interpreting and enhancing cross-lingual transfer during language model pre-training. Blevins and Zettlemoyer (2022) revealed that even in English-dominated pre-training data, millions of non-English tokens can be identified, which are crucial for multilingual capabilities. Some studies have attempted to analyze cross-lingual transfer abilities from perspectives of shared vocabulary and representation similarity(Patil et al., 2022; Lin et al., 2023), though their conclusions primarily apply to specific language groups. The predominant research paradigm has focused on explicitly enhancing cross-lingual transfer through exploiting supervision signals, such as parallel corpora(Zhang et al., 2024b; Ming et al., 2024; Ji et al., 2024;

Gosal et al., 2024; Gilabert et al., 2024), code-switching datasets(Singh et al., 2024; Yoo et al., 2024), or fine-grained signals like cross lingual entity links(Yamada and Ri, 2024). These approaches, however, remain constrained by the limited quantity, domain coverage, and morphological diversity of available bilingual resources (e.g., dictionaries, and parallel sentence pairs).

Our approach builds upon the fundamental principle of LLM pre-training: contextual modeling through next-word prediction (NWP) loss optimization within fixed-length text windows. Since LLMs could effectively learn monolingual semantics through this mechanism, we hypothesize that extending NWP optimization on semantically related cross-lingual content - using source language context to predict target language sequences - could enhance cross-lingual transfer capabilities. As illustrated in Fig.1(b), our method constructs **Cross-lingual In-context** samples by interleaving semantically related bilingual text pairs. Subsequently, we optimize LLMs through standard NWP loss computation on these composite samples. The proposed **Cross**-lingual **I**n-**C**ontext **P**re-**T**raining (**CrossIC-PT**) eliminates the reliance on parallel corpora, and could be applied to different types of text, providing a simple and scalable paradigm for cross-lingual transfer learning.

To validate our method, we implement the proposed **CrossIC-PT** method through continued pre-training (CPT) on existing LLMs (Dubey et al., 2024; Yang et al., 2024). This strategy converges faster than training from scratch, providing a cost-effective solution for multilingual experimentation (Zheng et al., 2024). Leveraging the readily available multilingual Wikipedia data, we construct a cross-lingual in-context corpus by concatenating two bilingual Wikipedia articles on the same entity, as illustrated in Fig.2. To mitigate context window length constraints, we segment article pairs into bilingual sub-pairs, using a dedicated [SPLIT] token as delimiters (Fig.2(b)). We further optimize the sliding window mechanism, ensuring that the next window starts from the token after the last [SPLIT] of the current window, thereby maintaining context coherence and enhancing cross-lingual alignment learning. To further assess the generalizability of our method, we develop a cross-lingual semantic retrieval framework build upon that extends beyond Wikipedia data by incorporating web-crawled text. As shown in Fig.3, this framework retrieves semantically related paragraphs from the English Fineweb_edu (Lozhkov et al., 2024) dataset using title and partial content keywords from the target-language Wikipedia articles as query.

We conducted experiments in six languages based on three LLMs (Llama-3.1-8B, Qwen2.5-7B, Qwen2.5-1.5B) and tested them on seven tasks. The CrossIC-PT model, built on Wikipedia, improved average performance by 3.79%, 3.99%, and 1.95% compared to the base models, respectively. The expansion of the data further boosted performance by 0.73% for Llama-3.1-8B.

Our contributions can be summarized as follows:

- We propose **CrossIC-PT**, a novel method that enhances LLMs' cross-lingual transfer by leveraging semantically related in-context data.
- To address input window length limitations, we design a window-split strategy with a [SPLIT] token and an optimized sliding window mechanism to maintain cross-lingual contextual coherence.
- We also design a cross-lingual semantic retrieval framework to augment training data, which further enhances model performance, proving the robustness and scalability of our approach.

2 Related Work

Many existing works focus on collecting multilingual data to enhance LLMs' cross-lingual capabilities (Yang et al., 2024; Dubey et al., 2024; Ming et al., 2024; Ji et al., 2024). Samples from different languages are randomly packed into fixed window sizes (e.g., 4096) without cross-contamination in self-attention. Even so, these models already demonstrate multilingual ability. Based on this, we hypothesize that concatenating semantically related English and target language data (Fig.1(b)) could enhance cross-lingual transfer by leveraging implicit supervision signals.

Cross-lingual supervision signals have been proven effective in enhancing LLMs' cross-lingual transfer abilities (Singh et al., 2024; Yamada and Ri, 2024). Most methods rely on bilingual corpora as explicit supervision signals (Zhang et al., 2024b; Ming et al., 2024; Ji et al., 2024; Gosal et al., 2024; Gilabert et al., 2024). Some works, like (Zhang et al., 2024b) distills translation pairs from LLMs through back-translation to create supervision signals. Others, such as (Singh et al., 2024; Yamada and Ri, 2024), apply code-switching techniques to replace or augment words with English translations. (Yoo et al., 2024) also explores code-switching at various levels using curriculum

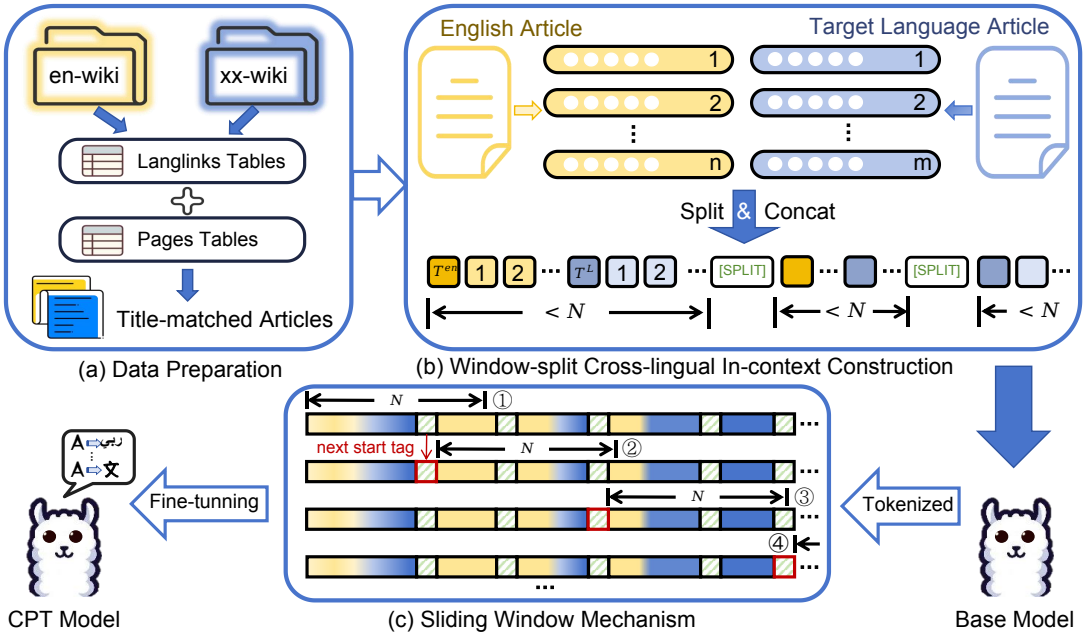


Figure 2: The implementation process of our method, CrossIC-PT, which constructs cross-lingual in-contexts based on Wikipedia data and performs continued pre-training (CPT) on existing multilingual models. Here, N represents the input window length of the model. The T indicates the title of the articles, and L indicates the target language.

learning. However, parallel corpora have restricted types, domains (most bilingual corpora are short sentence level bitexts, and usually extracted from news websites), and quantity. Synthetic parallel documents built by back-translation, however, are limited in text Quality. In contrast, our method constructs semantically related document pairs from the authentic data on the Internet, which is more scalable and less problematic.

3 Method

Multilingual LLM pre-training typically packs documents from different languages randomly into the fixed-size context window. We hypothesize that concatenating semantically related English and target language corpora, predicting the next words based upon not only monolingual and cross-lingual context could enhance cross-lingual transfer ability. We call this concatenated sample **Cross-lingual In-context** data, where English serves as the guiding context for learning the target language. Based on this, we propose **CrossIC-PT**, a pre-training method leveraging cross-lingual in-context data.

As LLMs are pre-trained with a fixed tokens window size (e.g. 4096 tokens), cross-lingual in-context data, which are usually two times longer than the vanilla monolingual documents, may exceed the size limit. Simplifying the packing by length may break the cross-lingual relationship. To address this problem, we carefully design a bilingual-aware window-split strategy to construct cross-lingual in-context data. Additionally, to avoid the traditional sliding window mechanism from splitting the concatenated context, we further optimize the sliding window mechanism to ensure context coherence.

We take advantage of Wikipedia data to implement our method, as shown in Fig.2, consisting of three key steps: (1) **Data preparation**, where we extract and align bilingual article pairs from Wikipedia (Sec. 3.1); (2) **Window-split cross-lingual in-context construction**, where we split multilingual contexts to match the length of the input window (Sec. 3.2); and (3) training with an optimized **sliding window mechanism** to enhance cross-lingual representation learning (Sec. 3.3). In order to test the generalization of our approach, we propose a cross-lingual semantic retrieval framework to augment the training data (Sec. 3.4).

3.1 Data Preparation

To obtain aligned article pairs in English and the target language (denoted L), we utilize three key tables from **Wikimedia** with three steps:

1. **Langlinks Table for Language** L: It contains article ID mappings between language L and other languages with matching titles, along with the corresponding title names T. This table helps identify

English article IDs and title names that match those in language L, mapping as $(ID^L, (ID^{en}, T^{en}))$.

2. **English Pages Table**: The 'pages' table of English provides article IDs and their corresponding title. We use it to remove English articles with blank or invalid titles from the initial mappings in step (1), yielding the final ID pairs (ID^L, ID^{en}).

3. **Articles Tables for English and Language L**: The 'articles' tables for both languages contain the article ID and full information on the web page, which includes the article content. Using the bilingual article ID pairs (ID^L, ID^{en}), we extract the corresponding article pairs with matching titles.

To ensure completeness, we also perform the reverse mapping (ID^{en}, ID^L), and combine the results with the forward mappings to obtain a comprehensive set of bilingual article pairs. This process ensures that we capture all possible title-matched articles between English and the target language.

3.2 Window-split Cross-lingual In-Context Construction

To fit within the context size N, we set a strategy for processing long article pairs by segmenting them into paragraphs and aligning them sequentially. Specifically, for each bilingual article pair (A_{en}, A_L), we extract the title T and split the articles into paragraphs by signal "\n\n":

$$A_{en} = [p_1^{en}, p_2^{en}, ..., p_n^{en}], \quad A_L = [p_1^L, p_2^L, ..., p_m^L].$$

We iteratively select paragraph pairs (p_i^{en}, p_i^L) until adding the k-th pair would exceed the length N, and then concat the paragraphs as follows:

$$(T^{en}, p_1^{en}; p_2^{en}; ...; p_{k-1}^{en}; T^L, p_1^L; p_2^L; ...; p_{k-1}^L),$$

with all English paragraphs preceding the target language L paragraphs, and the delimiter as "\n\n". Each concatenated sequence is terminated with a special `[SPLIT]` token to mark the end of the context window. If the paragraphs of one language are exhausted before the other, we continue concatenating paragraphs from the remaining language until the length limit N is reached or all paragraphs are used. This process converts each bilingual article pair into one or more window-split multilingual in-contexts, each fitting within the length limit N.

3.3 Pre-training Method

3.3.1 Sliding Window Mechanism

In standard pre-training, the sliding window mechanism concatenates all training data and sliding with a fixed window size. However, this can randomly break down our cross-lingual in-contexts, disrupting coherence. To address this, we optimize the sliding window by the introduced tag `"[SPLIT]"`. Specifically, all the windows set the start boundary after the last `"[SPLIT]"` token, as shown in Fig. 2. The tokens remain between the end boundary and the latest `"[SPLIT]"` token will be dropped. In this way we could try best to preserve the cross-lingual coherence within the window.

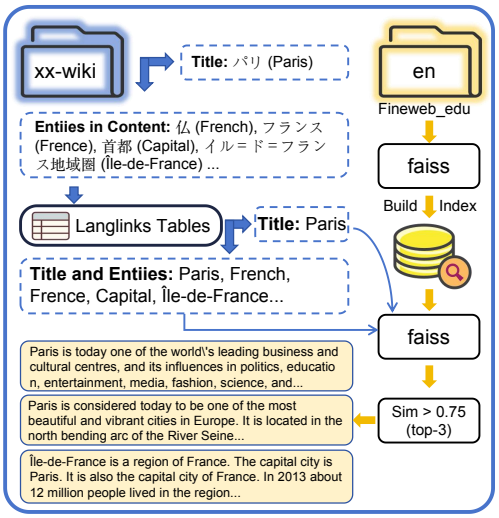


Figure 3: The framework of cross-lingual semantic retrieval based on FAISS similarity search tool.

3.3.2 Training Strategy

As discussed earlier, continual pre-training (CPT) is cost-effective for cross-lingual transfer. So we adopt it in all our experiments. Recent studies, such as (Whitehouse et al., 2024), show that Low-Rank Adaptation (LoRA) is highly competitive with full fine-tuning, especially in low-data and cross-lingual transfer scenarios. In our experiments, we also adopt LoRA during continual pre-training and results reveal that LoRA consistently provides better and more stable performance.

3.4 Data Augmentation via Retrieval

To validate our approach, we use the Wikipedia corpus, which includes data in nearly 200 languages linked by matched titles. While the content across languages is not strictly parallel, it covers the same topics, making it suitable for our needs. To enhance the generalization of our method, we introduce a cross-lingual semantic retrieval framework based on the FAISS similarity search tool (Johnson et al., 2019), as shown in Fig.3. This framework augments the training data by incorporating relevant

English articles from the Fineweb_edu (Lozhkov et al., 2024) dataset, retrieved using title and content keywords (up to 10 per article) extracted from the Wikipedia data.

First, keywords are extracted from the target-language Wikipedia page and mapped to English via the `langlinks` table. Fineweb_edu is then indexed using FAISS for similarity calculations. We employ a two-step retrieval process using FAISS: (1) retrieval based on title keywords, and (2) retrieval based on both title and content keywords. The final similarity score is the average of these two steps, balancing the importance of the titles (which may be ambiguous) and content keywords. Based on empirical observations, we set a similarity threshold of 0.75 and retrieved up to three relevant samples per target-language article to construct window-split cross-lingual in-context data. These samples are combined with the original Wikipedia data to form an augmented dataset.

4 Experiments

4.1 Training Data

Our training data is primarily sourced from Wikipedia (denoted as W), with token counts for English and each target language listed in Table 1. We selected six target languages L: Arabic (ar), Spanish (es), Japanese (ja), Korean (ko), Portuguese (pt), and Thai (th). To further expand the dataset, we retrieved relevant English data from a subset of Fineweb_edu (denoted as F), which has a file size of 17.44GB. The token counts for the augmented data are also provided in Table 1.

data	language	ar	es	ja	ko	pt	th
W	en	1.53B	1.88B	1.32B	1.01B	1.48B	0.42B
	L	0.67B	1.57B	1.28B	0.37B	0.81B	0.18B
W+F	en	0.12B	0.10B	0.06B	0.04B	0.05B	0.10B
	L	0.12B	0.13B	0.08B	0.03B	0.05B	0.06B

Table 1: The token counts for the data from Wikipedia (W) and augmented data from Wikipedia and Fineweb_edu (F).

4.2 Training Settings

We conducted experiments on three base models: Llama-3.1-8B (Dubey et al., 2024), Qwen2.5-7B (Yang et al., 2024), and Qwen2.5-1.5B (Yang et al., 2024). For LoRA, we set the rank to 64, alpha to 128, and dropout to 0.05. The input window length N was set to 4096, with a batch size of 128. All models were trained for one epoch, using a warmup ratio of 0.05, a cosine learning rate scheduler, and the AdamW optimizer. We randomly selected 0.1% of the data as the validation set, with a seed number of 32. For Llama-3.1-8B and Qwen2.5-7B, the models after one epoch of training were used as the final models. For Qwen2.5-1.5B, we validated the model every 100 steps and saved the checkpoint with the lowest validation loss as the final model. The training was performed on 8 A100 GPUs.

4.3 Benchmark

We evaluated our models on several tasks from the latest multilingual and multitask benchmark, P-MMEVAL (Zhang et al., 2024a), which includes: generation (FLORES-200 (Costa-jussà et al., 2022)), understanding (XNLI (Conneau et al., 2018), MHELLASWAG [1]), knowledge (MMMLU[2]), logical reasoning (MLOGIQA), and mathematical reasoning (MGSM (Shi et al., 2023)). To further assess the models' paragraph comprehension abilities, we incorporated a reading comprehension task (MRC). The MRC test data includes TydiQA-GoldP (Clark et al., 2020) for Arabic (ar) and Korean (ko), XQuAD (Artetxe et al., 2020) for Spanish (es), Portuguese (pt), and Thai (th), and 1,200 samples from JaQuAD (So et al., 2022) for Japanese (ja). Details of the evaluation setting can be found in Appendix A.

4.4 Baselines

In addition to the base models (Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-1.5B), we included the following baselines:

- **EMMA-500** (Ji et al., 2024): A model CPT on Llama-2-7B (Touvron et al., 2023) with 136B tokens covering over 500 languages.
- **LEIA** (Yamada and Ri, 2024): A method that randomly adds English translations of entities to target-language Wikipedia data for pre-training, leveraging cross-lingual entity supervision. We reproduced this method using the provided code to construct the data and perform CPT on Llama-3.1-8B, ensuring the target-language token count matched ours. We conducted experiments with three random seeds (32, 111, 222) and reported the mean and variance of the results.
- **Mix-PT**: A method that uses our title-matched article pairs from Fig.2(a) for pre-training.

[1] https://huggingface.co/datasets/alexandrainst/m_hellaswag
[2] https://huggingface.co/datasets/openai/MMMLU

Model		Languages					
		ar	es	ja	ko	pt	th
Llama-2-7B	base	24.77	37.10	37.76	35.05	40.90	23.27
	EMMA-500	30.14	31.18	32.77	32.49	28.06	33.31
Llama-3.1-8B	base	37.96	42.11	43.02	43.82	44.36	38.79
	LEIA	37.04±0.49	44.03±0.24	44.86±0.89	44.11±0.48	44.48±0.58	42.90±0.82
	Mix-PT	38.09	43.46	44.81	44.75	46.45	42.38
	CrossIC-PT	**40.57**	**45.49**	**47.27**	**46.87**	**49.09**	**43.51**
Qwen2.5-7B	base	50.91	54.71	56.95	55.52	56.49	53.81
	Mix-PT	54.48	58.71	57.69	57.39	60.30	56.19
	CrossIC-PT	**55.97**	**59.44**	**59.00**	**59.03**	**61.59**	**57.33**
Qwen2.5-1.5B	base	37.83	43.90	42.26	39.75	44.35	41.40
	Mix-PT	38.14	44.37	41.85	39.48	45.63	40.92
	CrossIC-PT	**40.21**	**45.09**	**43.96**	**41.47**	**48.25**	**42.23**

Table 2: The average results of our CrossIC-PT model, based on three base LLMs (Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-1.5B), are compared with corresponding baselines across six target languages. The cross-lingual in-context datasets used in CrossIC-PT are sourced from Wikipedia.

4.5 Results

4.5.1 Base Results

The average results of the baselines and our method, based on data from Wikipedia in six languages, are shown in Table 2. Detailed results for each task can be found in Appendix B. CrossIC-PT consistently improves the performance of the base LLMs and outperforms other baselines, demonstrating the effectiveness of using semantically related cross-lingual in-context corpora for pre-training.

Compared to the base LLMs, our CrossIC-PT method improves performance by 3.79%, 3.99%, and 1.95% on Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-1.5B, respectively, across six languages. Notably, in Portuguese (pt), CrossIC-PT improves performance by 4.73% on Llama-3.1-8B, surpassing the strongest baseline by 2.64%. The performance gains for Qwen2.5 models are more pronounced as model size increases, which may be attributed to the fact that CPT performance is influenced by the initial capabilities of the model.

Our method consistently improves performance across all languages. The improvement in Thai is less noticeable on Qwen2.5-1.5B, likely due to the smaller dataset size. The LEIA method shows significant gains in some languages (Spanish, Japanese, and Thai), but its performance is unstable and data-dependent. For instance, the standard deviation for Japanese and Thai exceeds 0.8. This suggests that the implicit supervision signals from our cross-lingual in-context data are more robust and adaptable across languages compared to the entity-alignment signals used by LEIA.

Data	Model	Languages					
		ar	es	ja	ko	pt	th
	Llama-3.1-8B	37.96	42.11	43.02	44.14	44.36	38.79
W	Mix-PT	38.09	43.46	44.81	44.75	46.45	42.38
	CrossIC-PT	40.57	45.49	47.27	46.87	49.09	43.51
W+F	Mix-PT	40.19	44.58	44.75	44.48	46.62	42.05
	CrossIC-PT	**41.18**	**46.93**	**48.10**	**47.32**	**49.97**	**43.72**

Table 3: The results of our CrossIC-PT model and Mix-PT baseline with Wikipedia-based data and augmented data constructed on Wikipedia and Fineweb_edu data.

The Mix-PT model is a strong baseline, trained on non-concatenated title-matched article pairs from Wikipedia, and improves performance across all six languages compared to the three base LLMs. However, our method improves the average performance by 2.15% over the Mix-PT model on Llama-3.1-8B. Our method further enhances Mix-PT by concatenating cross-lingual data and designing an optimized sliding window mechanism.

4.5.2 Results of Data Augmentation

To explore the generalization of our method, we propose a cross-lingual semantic retrieval framework (shown in Fig. 3) to augment the training data, with results reported in Table 3. After retrieval, the data volume increased by 0.06B–0.23B. Although this is a relatively small increase, it improved the average performance of our method by 0.73%. This demonstrates that even when English data is not perfectly aligned with target languages, semantically related still aids cross-lingual transfer. The simplicity of the semantic similarity retrieval process allows easy extension to various data sources.

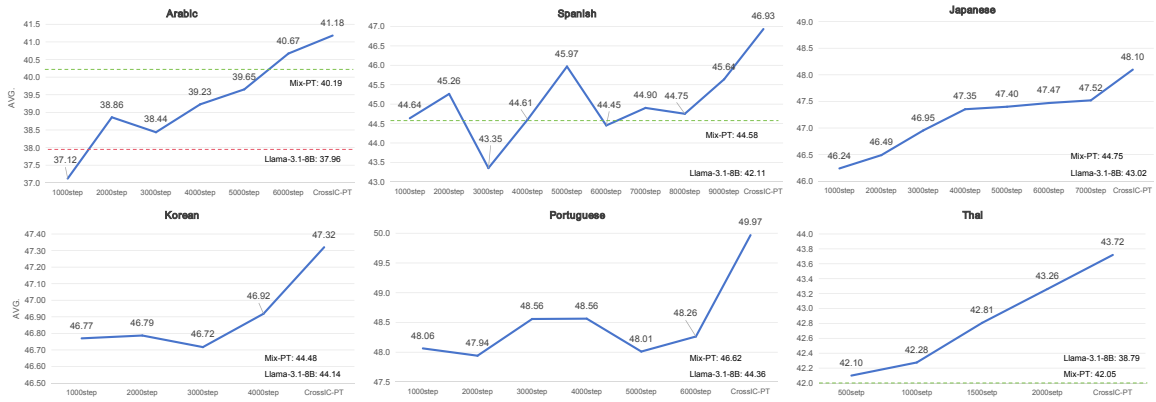


Figure 4: Performance progression of CrossIC-PT across intermediate checkpoints based on Llama-3.1-8B. Our method outperforms the baseline LLM early on, indicating quick acquisition of cross-lingual transfer capabilities, maintaining a slow upward trend as data volume increases.

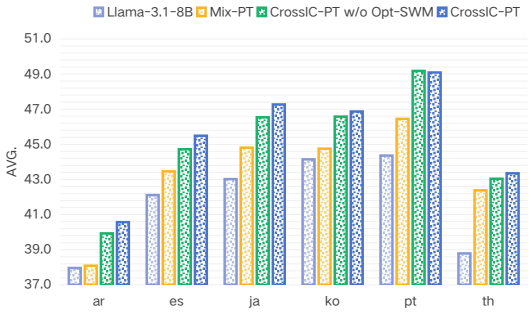


Figure 5: Ablation results of CrossIC-PT without optimized sliding window mechanism (Opt-SWM).

Additionally, we saved several intermediate checkpoints to assess the impact of data volume on performance. As shown in Fig. 4, at earlier checkpoints, our method outperformed the baseline LLM in all six languages and surpassed the strong baseline Mix-PT in four languages. This suggests that CrossIC-PT can quickly acquire useful cross-lingual transfer capabilities from the cross-lingual in-context data. Although performance improvements became slower as data volume increased, a consistent upward trend was still observed.

4.6 Ablation Study on Sliding Window Mechanism

We conduct an ablation study to assess the impact of our optimized sliding window mechanism (Opt-SWM), which introduces the `[SPLIT]` token and ensures each window starts after the last `[SPLIT]` token. Specifically, we compare the performance of CrossIC-PT with and without Opt-SWM, denoted as CrossIC-PT w/o Opt-SWM. The results shown in Fig.5 reveal that even without Opt-SWM, using only window-split cross-lingual in-context data, CrossIC-PT consistently improves performance across all languages. The addition of the optimized sliding window mechanism further enhances performance, highlighting its role in maintaining cross-lingual in-context coherence and improving language transfer. This demonstrates the effectiveness of all the steps in our design.

5 Analysis

We believe concatenating semantically related English and target language text in cross-lingual in-context data helps the model better understand the target language guided by the English context. Thus, we set the order as English first, followed by the target language. To verify if this direction is more beneficial, we analyze the concatenation order and test the model's performance in English to ensure there is no catastrophic forgetting.

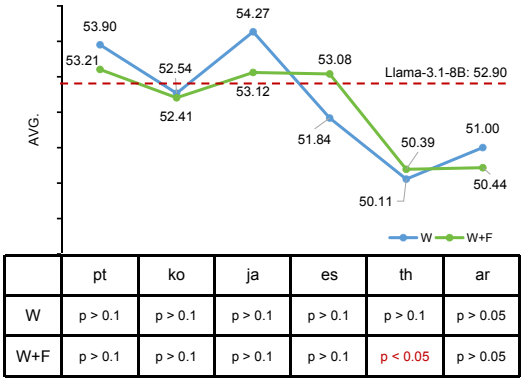
5.1 Analysis of Concatenation Direction

To evaluate the impact of concatenation direction on performance, we compare the original direction (English first, target language second) with the reverse direction (target language first, English second), as well as a 1:1 random mix of both directions. Previously, we only reported results for the en-xx direction in the translation task. In this experiment, we also provide results for the xx-en direction on FLORES-200.

The average results of six languages across tasks are presented in Table 4. The effect of data concatenation order on translation tasks is most pronounced and fits the intuition. The best translation performance occurs when the concatenation direction matches the translation direction. When combining both directions, CrossIC-PT consistently outperforms the Mix-PT method in translation

Model	XLOGIQA	XHELLASWAG	MMMLU	XNLI	MRC	FLORESE-200		MGSM	AVG.
						en-xx	xx-en		
Llama-3.1-8B	33.25	35.33	40.10	56.17	57.49	38.56	29.41	38.00	41.04
Mix-PT	34.75	36.68	43.05	59.17	60.36	39.63	32.72	36.96	42.92
CrossIC-PT	**36.00**	**39.71**	43.15	**62.17**	**63.02**	**41.39**	30.44	**39.68**	**44.44**
CrossIC-PT$_{mix}$	34.75	32.33	**43.55**	58.33	62.20	40.75	33.53	36.96	42.80
CrossIC-PT$_{reverse}$	35.50	33.69	43.00	57.67	62.41	39.51	**34.12**	36.40	42.79

Table 4: The average task results of CrossIC-PT with mix two directions (CrossIC-PT$_{mix}$) and the reverse direction (CrossIC-PT$_{reverse}$) of cross-lingual in-context data.



	pt	ko	ja	es	th	ar
W	p > 0.1	p > 0.1	p > 0.1	p > 0.1	p > 0.1	p > 0.05
W+F	p > 0.1	p > 0.1	p > 0.1	p > 0.1	p < 0.05	p > 0.05

Figure 6: The average results of each target language model in English tasks. The p is the significant score between the CrossIC-PT model and Llama-3.1-8B.

tasks, showing that even non-parallel bilingual data improves translation. Overall, the English-first, target language-second concatenation gives the best results, aligning with our intention of using English as context to guide the target language learning.

5.2 Performance on English Tasks

To prevent catastrophic forgetting, it's important to ensure English performance is maintained. To verify this, we tested the performance of six target language models on English tasks, using the same tasks as before. The results are shown in Fig.6.

The upper part of Fig.6 shows the average performance of each target language model on English tasks, with the x-axis ordered by the performance gap between Llama-3.1-8B's performance on the target language and English. The trend suggests that a larger performance gap corresponds to a greater impact on English performance after training. For example, the English performance of Thai (th) and Arabic (ar) is lower. However, it is primarily due to a significant drop in one task. To further investigate, the lower part of Fig.6 presents the statistical significance ("p") of the performance differences between target language models and the base model, Llama-3.1-8B, across seven tasks. The results show that, except for the Thai model trained with data augmentation (which exhibits a significant drop in English performance), there are no significant differences for other target language models. This suggests that CrossIC-PT improves performance in target languages while effectively preserving English capabilities. We believe this is likely due to the inclusion of at least 50% of English tokens in the cross-lingual in-context corpus, which helps mitigate severe forgetting. This result further validates the robustness and practicality of CrossIC-PT for cross-lingual transfer.

6 Conclusion

Our work explores a special angle by focusing on semantically related multilingual in-context to enhance the cross-lingual transfer capability of LLMs. We hypothesize that concatenating semantically related English and target language corpora as Cross-lingual In-context data is easily accessible and provides an implicitly cross-lingual supervision signal. Building on this hypothesis, we propose CrossIC-PT, a pre-training method based on cross-lingual in-context data. We implement our method using Wikipedia data and employ continual pre-training of existing LLMs on this data. To address the limitations posed by input window length during model training, we design a window-split strategy coupled with an optimized window sliding mechanism. Experimental results demonstrate that CrossIC-PT enhances multilingual performance across three models—Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-1.5B—across six target languages, achieving performance gains of 3.79%, 3.99%, and 1.95%, respectively, compared to base models. Further improvements are observed after data augmentation using a semantic retrieval framework. Our approach is simple to scale for multilingual LLM pre-training and offers an efficient way to expand data volume.

Limitations

To our knowledge, this work has the following limitations:

- Due to resource constraints, our experiments were limited to a context window length of 4096 tokens. Longer windows could better preserve the completeness of articles and enable the concatenation of similar multilingual data from more than two languages, potentially further enhancing cross-lingual transfer.
- Our experiments focused on validating the effectiveness of concatenated cross-lingual in-context data, so we performed continued pretraining on monolingual data rather than mixing multilingual data. While this choice aligns with our research goals, our approach also provides valuable insights for developers working on multilingual LLMs.
- Our data expansion method, based on retrieval, currently demonstrates how to retrieve additional English data from external sources using target-language Wikipedia data. However, this approach can be easily extended to retrieve more diverse data. Wikipedia's broad domain coverage makes it an ideal hub for retrieving both target language and English data from other sources. By controlling the retrieval process with appropriate similarity thresholds, the retrieved bilingual data can be used to construct high-quality cross-lingual in-context data.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, et al. 2023. Gpt-4 technical report.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 4623–4637.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3563–3574. Association for Computational Linguistics.

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, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits,

David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Javier García Gilabert, Carlos Escolano, Aleix Sant Savall, Francesca de Luca Fornaciari, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Investigating the translation capabilities of large language models trained on parallel data only. *CoRR*, abs/2406.09140.

Gurpreet Gosal, Yishi Xu, Gokul Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming Chen, Biswajit Mishra, Natalia Vassilieva, Joel Hestness, Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Onkar Pandit, Satheesh Katipomu, Samta Kamboj, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Soundar Doraiswamy, Mohamed El Karim Chami, and Preslav Nakov. 2024. Bilingual adaptation of monolingual foundation models. *CoRR*, abs/2407.12869.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *CoRR*, abs/2409.17892.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2023. mplm-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *CoRR*, abs/2412.04003.

Vaidehi Patil, Partha Pratim Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. *ArXiv*, abs/2203.01976.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A three-pronged approach to cross-lingual adaptation with multilingual llms. *CoRR*, abs/2406.17377.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. Jaquad: Japanese question answering dataset for machine reading comprehension. *CoRR*, abs/2202.01764.

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 Moly-

bog, 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hetong Wang, Pasquale Minervini, and E. Ponti. 2024. Probing the emergence of cross-lingual alignment during llm training. *ArXiv*, abs/2406.13229.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1202–1228. Association for Computational Linguistics.

Ikuya Yamada and Ryokan Ri. 2024. LEIA: facilitating cross-lingual knowledge transfer in language models with entity-based data augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7029–7039. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum learning for multilingual transfer in llms. *CoRR*, abs/2411.02460.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024a. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *CoRR*, abs/2411.09116.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11189–11204. Association for Computational Linguistics.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7725–7738. Association for Computational Linguistics.

A The Setting for Evaluation

The prompts of each task we used are shown in Table 5. Since our method aims to transfer English capabilities to target languages, the prompts are primarily designed in English, and the demonstrations are also selected from English data. For the mathematical reasoning task (MGSM), we conducted an 8-shot test; for the reading comprehension task (MRC), we adopted a zero-shot setting to evaluate the model's understanding of the target language; for other tasks, we set up a 5-shot test. For multiple-choice tasks (e.g., XNLI, MMLU, XHELLASWAG, XLOGIQA), we directly obtain answers by predicting the next logits. For other tasks, we use greedy search to generate answers and extract the final answer through regular expression matching.

Task	Prompt
XLOGIQA	`Passage: {context}\nQuestion: {question}\nChoices:\nA. {option_a}\nB. {option_b}\nC. {option_c}\nD. {option_d}\nAnswer:`
XHELLASWAG	`{premise}\nOptions: \nA. {option_1}\nB. {option_2}\nC. {option_3}\nD. {option_4}\nQuestion: Which is the correct ending for the sentence from A, B, C, and D? \nAnswer:`
MMMLU	`The following is a multiple-choice question.\n\n{question}\nA. {option_a}\nB. {option_b}\nC. {option_c}\nD. {option_d}\n\nAnswer:`
XNLI	`Take the following as truth: {premise}\nThen the following statement: "{hypothesis}" is\nOptions:\nA. true\nB. inconclusive\nC. false\nAnswer:`
MRC	`Refer to the passage below and answer the following question:\nPassage: {context}\nQuestion: {question}\nAnswer: Based on the passage, the answer to the question is "`
FLORES-200	`Translate from [source] to [target].\n[source]: </X>\n[target]:`
MGSM	`Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of "[The answer is ]". Do not add anything other than the integer answer after "The answer is".\n\n{question}`

Table 5: Task Prompts."[]" represents optional content. For the FLORESE task, the "[source]" indicates the source language, and "[target]" indicates the target language of translation. For MGSM, "[The answer is]" is the translation of "The answer is " according to the test language.

B Results of Tasks

The average results of our method and the baseline across six languages in each task are shown in Table 6.

Model		XLOGIQA	XHELLASWAG	MMMLU	XNLI	MRC	FLORES-200	MGSM	AVG.
Llama-3.1-8B	base	33.96	35.33	38.96	55.84	56.72	34.83	37.40	41.86
	LEIA	34.93±0.64	37.74±0.16	38.48±0.07	60.00±0.69	58.02±0.25	34.35±0.06	37.36±0.11	42.98±0.25
	Mix-PT	35.00	35.45	41.96	57.64	60.88	36.02	36.33	43.32
	CrossIC-PT	**35.83**	**38.98**	**42.17**	**61.11**	**62.99**	**38.32**	**38.87**	**45.47**
	Wikipedia+Fineweb_edu en								
	Mix-PT	33.75	36.44	**41.88**	57.92	63.55	36.17	36.73	43.78
	CrossIC-PT	**36.25**	**41.49**	41.83	**62.36**	**65.49**	**38.13**	**37.87**	**46.20**
Qwen2.5-7B	base	44.79	60.76	48.67	62.78	65.07	35.06	66.00	54.73
	Mix-PT	45.42	61.74	49.83	**76.11**	69.38	34.79	**64.93**	57.46
	CrossIC-PT	**46.46**	**63.31**	**50.00**	**76.11**	**71.69**	**39.05**	64.47	**58.73**
Qwen2.5-1.5B	base	36.46	36.72	41.21	50.14	63.21	24.27	39.07	41.58
	Mix-PT	37.08	35.46	42.21	55.42	65.47	21.68	34.80	41.73
	CrossIC-PT	**39.17**	**38.01**	**42.46**	**58.33**	**65.89**	**25.48**	**35.40**	**43.53**

Table 6: The average results of our method and the baseline across six languages in each task.