PREALIGN: Boosting Cross-Lingual Transfer by Early Establishment of Multilingual Alignment

Jiahuan Li[♣], **Shujian Huang**[†♣], **Aarron Ching**[♠], **Xinyu Dai**[♣] and **Jiajun Chen**[♣]
[♣]National Key Laboratory for Novel Software Technology, Nanjing University, China
[♠]Independent Researcher
lijh@smail.nju.edu.cn, {huangsj,daixinyu, chenjj}@nju.edu.cn

Abstract

Large language models demonstrate reasonable multilingual abilities, despite predominantly English-centric pretraining. However, the spontaneous multilingual alignment in these models is shown to be weak, leading to unsatisfactory cross-lingual transfer and knowledge sharing. Previous works attempt to address this issue by explicitly injecting multilingual alignment information during or after pretraining. Thus for the early stage in pretraining, the alignment is weak for sharing information or knowledge across languages. In this paper, we propose PREALIGN, a framework that establishes multilingual alignment prior to language model pretraining. PREALIGN injects multilingual alignment by initializing the model to generate similar representations of aligned words and preserves this alignment using a code-switching strategy during pretraining. Extensive experiments in a synthetic English to English-Clone setting demonstrate that PREALIGN significantly outperforms standard multilingual joint training in language modeling, zero-shot cross-lingual transfer, and cross-lingual knowledge application. Further experiments in real-world scenarios further validate PREALIGN's effectiveness across various languages and model sizes. [1]

1 Introduction

Large language models (Brown et al., 2020; Touvron et al., 2023a,b) have drastically changed the research paradigm of multilingual language processing. Despite being trained on mainly English texts, they still exhibit reasonable ability for other languages (Touvron et al., 2023a,b; Wang et al., 2024), and have established multilingual alignment to some extent (Devlin et al., 2019; Conneau and Lample, 2019; Lin et al., 2022). However, researchers (Wang et al., 2024; Gao et al., 2024; Zhang et al., 2023; Qi et al., 2023) have found the spontaneous alignment between languages in these model is still relatively weak, leading to weak cross-lingual factual knowledge retrieval (Wang et al., 2024; Gao et al., 2024) and inconsistency behaviors given the same input (Qi et al., 2023; Zhang et al., 2023).

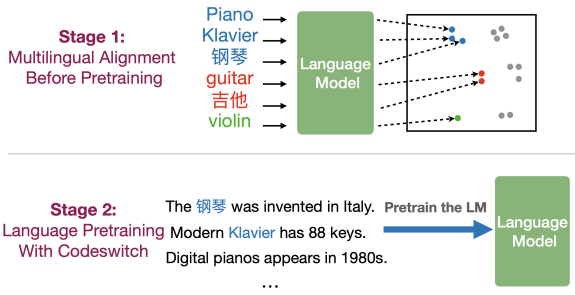

Figure 1: The illustration of PREALIGN. Words in blue, red and green represent translations of *piano, guitar* and *violin*, respectively.

A handful of works (Reimers and Gurevych, 2020; Cao et al., 2020; Wu and Dredze, 2020; Chaudhary et al., 2020; Yang et al., 2021; Tang et al., 2022; Feng et al., 2022; Gao et al., 2024) try to mitigate the problem by explicitly injecting alignment information using existing supervision data. They either construct cross-lingual prediction tasks (Chaudhary et al., 2020; Yang et al., 2021) or train models to produce similar representations of aligned words or sentences (Tang et al., 2022; Wu and Dredze, 2020; Reimers and Gurevych, 2020). However, the improvements are somewhat mixed and the establishment of multilingual alignment requires a long training process either *during* or *after* pretraining (Dufter and Schütze, 2020), which prevents the model from effectively performing cross-lingual transfer at earlier stage in pretraining.

In this paper, we introduce PREALIGN, a framework designed to enhance the alignment of pretrained language models. PREALIGN differs from

† The Corresponding author.

[1]The code of this paper is available at https://github.com/NJUNLP/PreAlign

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10246–10257
November 12–16, 2024

prior methods by integrating the multilingual alignment information *before* extensive language pretraining and maintaining it throughout the pretraining process. This proactive alignment effectively enhances the learning of cross-lingual knowledge in the pretraining corpus, therefore advancing cross-lingual transfer. Therefore, model's proficiency in target languages at early training stage is enhanced, leading to improvement of model's ability to acquire knowledge at that stage.

More specifically, before large-scale language pretraining, PREALIGN first collects multilingual translation pairs between English and languages to be transferred, and inject this information into the model by pre-pretraining it to produce similar representations of aligned pairs. In order to maintain the established multilingual alignment across the pretraining phase, we propose an input-only codeswitching strategy, which only substitutes words in the input text to its aligned words, and optimizes model using language modeling objective. The illustration of PREALIGN is shown in Figure 1.

We firstly conduct experiments on an English to English-Clone setting (K et al., 2020; Dufter and Schütze, 2020; Schäfer et al., 2024), where English-clone is a synthetic language that shares identical grammar and vocabulary distribution with English, but has no vocab overlap. This allows us to study cross-lingual transfer on a more controlled environment. Experiments demonstrate that PREALIGN improves language ability on English-Clone by strengthening the cross-lingual transfer of knowledge and abilities from English. Further analysis shows that the early established multilingual alignment can be kept throughout large-scale language pretraining and generalize to other unaligned words. Experiments on real-world settings (including Chinese, German, Arabic and Russian) validate the effectiveness of PREALIGN across different languages and model scales.

2 Related Work

2.1 Understanding Cross-lingual Ability of Pretrained language models

Many works attempt to analyze the cross-lingual ability of LLMs. Dufter and Schütze (2020); Conneau et al. (2020) try to explain factors that contributes to spontaneous multilingual alignment developed in pretrained language models, including under-parameterization, shared model architectures and pivot words across languages. Other works investigate the working mechanism of multilingual representations. Wendler et al. (2024) find that English-centric models works on a concept space that is close to English when processing other languages. Gaschi et al. (2023); Hämmerl et al. (2024) discuss the relationship between multilingual alignment and cross-lingual transfer. Recently, Gao et al. (2024); Qi et al. (2023) analyze multilingual knowledge alignment in existing LLMs, and find that multilingual training and instruction tuning can only lead to shallow alignment, i.e. LLMs can achieve similar task performances and consistent responses across languages, yet cannot apply knowledge across languages.

Our paper differs from theirs in that we focus on improving models' cross-lingual ability and successfully unlocks the ability of cross-lingual knowledge transferring.

2.2 Enhancing Cross-lingual Ability of Pretrained Language Models

Other studies also seek to enhance the cross-lingual capabilities of pretrained language models. These typically utilize explicit alignment signals, such as parallel sentences and dictionaries. They can be categorized based on when the alignment occurs: during pretraining or post-pretraining.

On the first category, Yang et al. (2020); Chaudhary et al. (2020) perform codeswitching on the monolingual data to make model better capture cross-lingual relation and dependency. Hu et al. (2021) train the model to produce consistent word alignment matrices between source and target language and similar representations for parallel sentences. Chi et al. (2022) explores multilingual replaced token detection and translation replaced token detection task. Tang et al. (2022) further maximize the cosine similarity of aligned word embeddings to explicitly inject multilingual alignment.

On the second category, researchers enhance the multilingual alignment after pretraining. Earlier works either optimizes pretrained models to produce similar representations for parallel sentences (Reimers and Gurevych, 2020; Pan et al., 2021; Feng et al., 2022) or parallel words (Cao et al., 2020; Wu and Dredze, 2020). Recent works on large language models typically train the model to produce consistent responses (She et al., 2024) or performing cross-lingual instruction-following tasks (Zhu et al., 2024b,a).

PREALIGN differs from all above works in that it establishes multilingual alignment before language

pretraining, therefore facilitating the cross-lingual transfer at early pretraining stage.

3 The PreAlign Method

In this section, we present PREALIGN, a simple and effective framework that advances the establishment of multilingual alignment before language pretraining.

3.1 Injecting Multilingual Alignment before Language Pretraining

PREALIGN aims to inject multilingual alignment information before large-scale language model pretraining, which facilitates cross-lingual transfer as soon as possible. This involves two stages: collection of multilingual alignment table and alignment injection via contrastive learning.

Collection of multilingual alignment table The collection of multilingual alignment table can either leverage the off-the-shelf multilingual dictionaries such as MUSE, or rely on machine translation models. In this paper, we take the second method: we first extract from an English monolingual corpus $\mathcal{D}$ the collections of all unique words $\mathcal{W} = \{w\}_i^N$, where N is the number of unique words. For each word w, we translate it to all considered target languages, and denote the translation results as $T(w)$. We collect diverse translations for each word using GPT-4. More details can be found in Appendix A.

Alignment injection via contrastive learning After the multilingual alignment table is collected, PREALIGN initializes models' parameters using a contrastive alignment objective, which optimizes the model to produce similar representations for aligned words. Specifically, given an English word w_i and its available translations across other languages $T(w_i)$, PREALIGN firstly obtains representations of each layer for each $w \in T(w_i)$:

$$h_w^l = \text{MeanPool}(f(w, l)) \tag{1}$$

where $l = 0, 1, \cdots, L, L+1$; $f(w, l)$ for $1 \leq l \leq L$ denotes the l-th Transformer layer representations of the model's encoding of w; $f(w, 0)$ and $f(w, L+1)$ denotes the word embedding and output embedding of w, respectively. Note that since w could be tokenized to multiple subwords, PREALIGN aggregates them into a single representation using mean-pooling operator.

PREALIGN then leverages a contrastive learning objective (Khosla et al., 2021) to establish alignments between words in different languages:

$$\mathcal{L}_{\text{align}}^l = \sum_{\substack{w_j \in \mathcal{W} \\ w_i \in T(w_j)}} \log \frac{\exp(\, d(h_{w_i}^l, h_{w_j}^l)/\tau)}{\sum_{w_k \in \mathcal{B}} \exp(\, d(h_{w_j}^l, h_{w_k}^l)/\tau)} \tag{2}$$

where $\mathcal{B}$ is the set of all words in current mini-batch, τ is the temperature parameter. $cos(\cdot, \cdot)$ is the cosine similarity function. The final learning objective is the sum of contrastive loss of all layers:

$$\mathcal{L}_{\text{align}} = \sum_{l=0}^{L+1} \mathcal{L}_{\text{align}}^l \tag{3}$$

To prevent the initialization from being trapped in a local minima that is not suitable for the subsequent language modeling, we also add an auxiliary language modeling loss beside the contrastive objective in practice [2]:

$$\mathcal{L}_{\text{joint}} = \alpha \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{LM}} \tag{4}$$

Note that, the $\mathcal{L}_{LM}$ objective in the pre-alignment stage only serves to regularize the optimization process, rather than performing large-scale pretraining. In practice, this stage only consumes 5% pretraining data.

3.2 Maintaining Multilingual Alignment via Input-only Codeswitching

PREALIGN injects multilingual alignment information before language pretraining. However, it is possible that this information could be quickly forgotten if not continuously reinforced. Inspired by prior research (Chaudhary et al., 2020; Yang et al., 2021) demonstrating that codeswitching effectively promotes multilingual alignment, we propose using the codeswitching technique to sustain this alignment throughout the pretraining process.

Originally, codeswitching was applied to both the input sequence and the target tokens in raw data, which can exacerbate the issue of multilingual script mixing in the outputs of decoder-only models. To address this, we propose an input-only codeswitching strategy that affects only the input. The distinction between the traditional codeswitching and our input-only codeswitching is illustrated in Figure 2.

[2]The training data for $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{LM}}$ are independently sampled in each mini-batch.

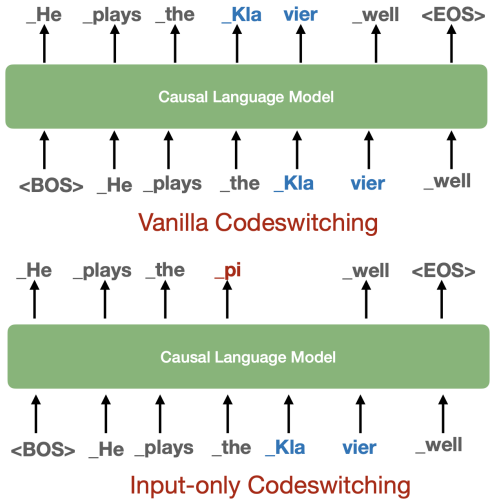

Figure 2: Comparison between vanilla codeswitching and the proposed input-only codeswitching. The original English sentence is *He plays the piano well*, and *Klavier* is the German translation of *piano*.

Formally, given a subword sequence $X_{<i}, x_i^1, \cdots, x_m^i, X_{>i}$, where $X_{<i}$ and $X_{>i}$ are the subword sequences before and after the i-th word, respectively; $x_i^1, \cdots, x_i^m$ is the subword sequence of the i-th words. Suppose the i-th word is substituted by $y_i^1, \cdots, y_i^n$ after codeswitching, then the language modeling objective after the original codeswitching is

$$\begin{aligned}\mathcal{L}_{\text{vanilla\_CS}} = p(X_{<i}) \cdot p(X_{>i}|X_{<i}, y_i^1, \cdots, y_i^n) \\ \cdot p(y_i^1|X_{<i}) \\ \cdot \prod_{j=2}^{n} p(y_i^j|X_{<i}, y_i^1, \cdots, y_i^{j-1})\end{aligned} \tag{5}$$

In Equation 5, the item $p(y_i^1|X_{<i})$ requires the model to generate words in one language given prefixes in another language. To mitigate this, our input-only codeswitching modifies the objective to be

$$\begin{aligned}\mathcal{L}_{\text{input\_only\_CS}} = p(X_{<i}) \cdot p(X_{>i}|X_{<i}, y_i^1, \cdots, y_i^n) \\ \cdot p(x_i^1|X_{<i}).\end{aligned} \tag{6}$$

Equation 6 omits the prediction objective of subwords in the word after codeswitching ($p(y_i^1|X_{<i})$), therefore preventing the generation results contain scripts from other languages. In this paper, we use a codeswitching ratio of 5%.

4 Evaluation of Cross-Lingual Transfer

To investigate the cross-lingual transfer effects that is close to situations in current LLMs, we design the evaluation in an English-dominated setting, where most of the pretraining data is English. For the examined target language, the amount of pretraining data is much less than English. Intuitively, the language ability of the target language will be much weaker than English. However, it is still interesting to know to what extent the language abilities and knowledge could transfer from English to the target language.

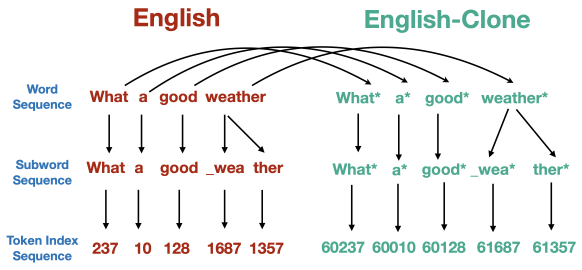

Figure 3: Illustration of the creation of English-Clone.

4.1 Languages

4.1.1 Synthetic Language: En-Clone

We construct a synthetic language called En-Clone, by cloning all English tokens by a one-to-one mapping. En-Clone shares the same linguistic properties with English, such as vocabulary distribution, subword segmentation, grammar and syntax, yet they have no word overlapping. See Figure 3 for an illustration.

This synthetic setting provides many benefits. Firstly, the English to En-Clone setting arguably forms the easiest setting for testing the cross-lingual transferring ability of LLMs, since it does not involve the discrepancy of word ordering and possibly complex one-to-many/many-to-one alignments between real-world languages. Therefore, this setting can serve as a sanity-check for cross lingual transferring methods.

Secondly, since the golden alignment between English and En-Clone is trivial to get, we can easily achieve *perfect* alignment at the initialization stage by setting the input and output embedding of aligned tokens to be identical. In this way, hidden states of all intermediate layers would also be identical. This provides us a chance to analyze the upper-bound performance of our method.

4.1.2 Real-World Languages

We also experiment with real-world languages to examine the effect of PREALIGN in more complex situations. We select Chinese, Russian, German

and Arabic as our target languages, which spans four different language families, serving as good representatives of world languages. Note that in this case, the alignment is established for 4 target languages at the same time.

4.2 Evaluation Metrics

The cross-lingual transfer effects are evaluated in the following 3 aspects:

Target Language Modeling (LM) The first evaluation metric is the language modeling performance (perplexity) of the target language. Given the same amount of target language data, this can reflect the extent of language ability transferred from English to the target language.

Zero-shot Cross-lingual Transfer (ZS-CLT) Another common way to evaluate model's cross-lingual ability is zero-shot cross-lingual transfer, where we finetune models with training data of a given task in the source language, and test model's ability on the same task in target languages. We use the commonly-used XNLI (Conneau et al., 2018) dataset for ZS-CLT evaluation.

Cross-lingual Knowledge Application (CLKA) Large language models acquire extensive world knowledge from their pretraining corpora, which might be described in different languages. It is ideal for LLMs to learn knowledge from texts in one language and apply it across other languages.

In order to evaluate models' ability to perform such cross-lingual knowledge application, we propose a setting where the model is trained with certain English texts describing synthetic knowledge, and test the injected knowledge in the target language. Each synthetic knowledge is a triplet like (subject, relation, object), where relations are extracted from WikiData (Vrandečić and Krötzsch, 2014), and subjects and objects are artificial entities. We assess the model's knowledge retention by comparing the likelihood of different statements, including one correct statement and three distractors generated by randomly substituting named-entities for the original object in the knowledge statement. See Appendix A for examples of synthesized knowledge.

4.3 Experiment Settings in General

Pretraining Dataset We adopt CulturaX (Nguyen et al., 2023) as the pretraining dataset. CulturaX is a multilingual pretraining corpus that has been rigorously cleaned. For English, we randomly select 10 billion tokens from CulturaX as the pretraining data. For each language to be transferred to (Zh, De, Ru and Ar in the real-world setting, and En-Clone in the synthetic setting), we randomly select 100 million tokens, which is 1% of the data in English.

Model Configuration We adopt the GPT-2 style Transformer architecture for our model. As the defaulting setting, our model contains 12 Transformer layers with a hidden dimension of 1024. The number of total non-embedding parameters is about 150 million. We use AdamW (Kingma and Ba, 2017) optimizer with a global batch size of about 1 million tokens. The learning rate is decayed from $3e-4$ to $3e-5$ following a cosine scheduler.

Baselines We compare PREALIGN 's performance with the following methods:

- Joint Training, where we pretrain the model on the mix of 10 billion English tokens and 100 million tokens in the target language.
- Only Target, where we only pretrain the model on 100 million tokens in the target language.
- Full Target, where we pretrain the model on 10 billion tokens in the target language. This can serve as an upper-bound performance for the target language.

5 Experiments on Synthetic Setting

We start our evaluation on the English to En-Clone setting, which allows us to better control the relationship between the source and target language.

5.1 General Results

We present results on LM, ZS-CLT and CLKA in Table 1.

Joint Training achieves spontaneous multilingual transfer to some extent. Table 1 shows that compared to Only-Target, Joint training achieves notable improvements on LM despite there are neither parallel signal or pivot words between English and English-clone. Surprisingly, the model could successfully transfer the ability to perform NLI task from English (79.8) to English-Clone (74.9). However, this transfer does not work well on CLKA, which is consistent with previous findings (Gao et al., 2024) that cross-lingual knowledge transfer is hard to achieve by multilingual pretraining.

	#Tokens		LM (ppl. ↓)		ZS-CLT (acc. ↑)		CLKA (acc. ↑)
	En	En-Clone	En	En-Clone	En	En-Clone	En-Clone
Only Target	-	0.1B	-	47.2	-	-	-
Joint Training	10B	0.1B	16.1	21.6	79.8	74.9	26.5
PREALIGN	10B	0.1B	**15.9**	**16.5**	**80.1**	**79.3**	**90.3**
Full Target	-	10B	-	16.2	-	-	-

Table 1: Performance of PREALIGN and other methods on language modeling, zero-shot cross-lingual transfer (ZS-CLT) and cross-lingual knowledge application (CLKA).

PREALIGN improves over Joint Training on all evaluation tasks. We can also see that PREALIGN significantly outperforms Joint Training on all three evaluation metrics. On the LM evaluation, PREALIGN even achieves performance comparable to Full-Target, using only 1% data. For ZS-CLT, the performance gap between two languages are narrowed. For the CLKA the accuracy is greatly improved (from 27.7 to 64.6). All the results demonstrate the effectiveness of PREALIGN for facilitating cross-lingual transfer. It is worth noticing that the performance of English are also improved, suggesting the learning of English-Clone also helps English as well.

PREALIGN outperforms methods that establish alignment during and post pretraining. We experiment with performing contrastive alignment during and post the pretraining process (Wu and Dredze, 2020), and compared them to PREALIGN in Table 2. For the post-pretraining alignment, we add an additional LM loss to reduce catastrophic forgetting. It can be seen that on-the-fly alignment can degrade model's performance, which we hypothesize that excessively optimizing the model on the limited size of word-level multilingual alignment during pretraining might have a negative impact on language ability. For post-pretraining alignment, we can observe a improvement over Joint Training on LM and ZS-CLT, but nearly no effect on CLKA. However, PreAlign outperforms both on-the-fly alignment and post alignment on all three evaluation protocols.

5.2 In-Depth Investigation for CLKA

As the performance for CLKA varies for different methods, we further examine the learning dynamics of CLKA. We segment the pretraining process into shorter periods, each consisting of 250 training steps, and evaluate the CLKA accuracy for each period. More specifically, for each period, knowledge of different frequence are provided to the model, and assessment occurs immediately after each period using the corresponding model checkpoint. For comparison, we evaluate all four combinations of languages for training and testing the knowledge. Figure 4 shows the results.

	LM	ZS-CLT	CLKA
Joint Training	21.6	74.9	26.5
On-the-fly alignment	22.1	74.3	26.5
Post alignment	19.7	75.5	28.4
PreAlign	**16.5**	**79.3**	**90.3**

Table 2: Comparison of performing contrastive alignment at different stage. On-the-fly alignment: performing alignment during the pretraining. Post alignment: performing alignment when the pretraining is done.

Knowledge learning ability correlates with language ability. We can see from the top-left of Figure 4, where we test English knowledge in English language, models' knowledge completion accuracy after each learning period rapidly grows as the pretraining goes on. This indicates that the models' ability to acquire knowledge correlates with their language modeling ability. The final performance also correlates with the knowledge frequency in the learning period as expected.

Early cross-lingual transfer enhance target language ability, facilitating knowledge learning. In the top-right of Figure 4 where we test English-Clone knowledge in English-clone language, we observe a similar trend as the top-left figure. However, the growing rate of PREALIGN is higher compared to Joint Training especially when frequency of knowledge is low, thanks to better transfer of language ability from English to English-Clone.

PREALIGN unlocks CLKA. From the bottom two figures in Figure 4, we can see the CLKA ability of Joint Training is greatly weaker than PREALIGN, close to the random guessing performance. This renders PREALIGN a promising method for learning truly multilingual knowledge alignment.

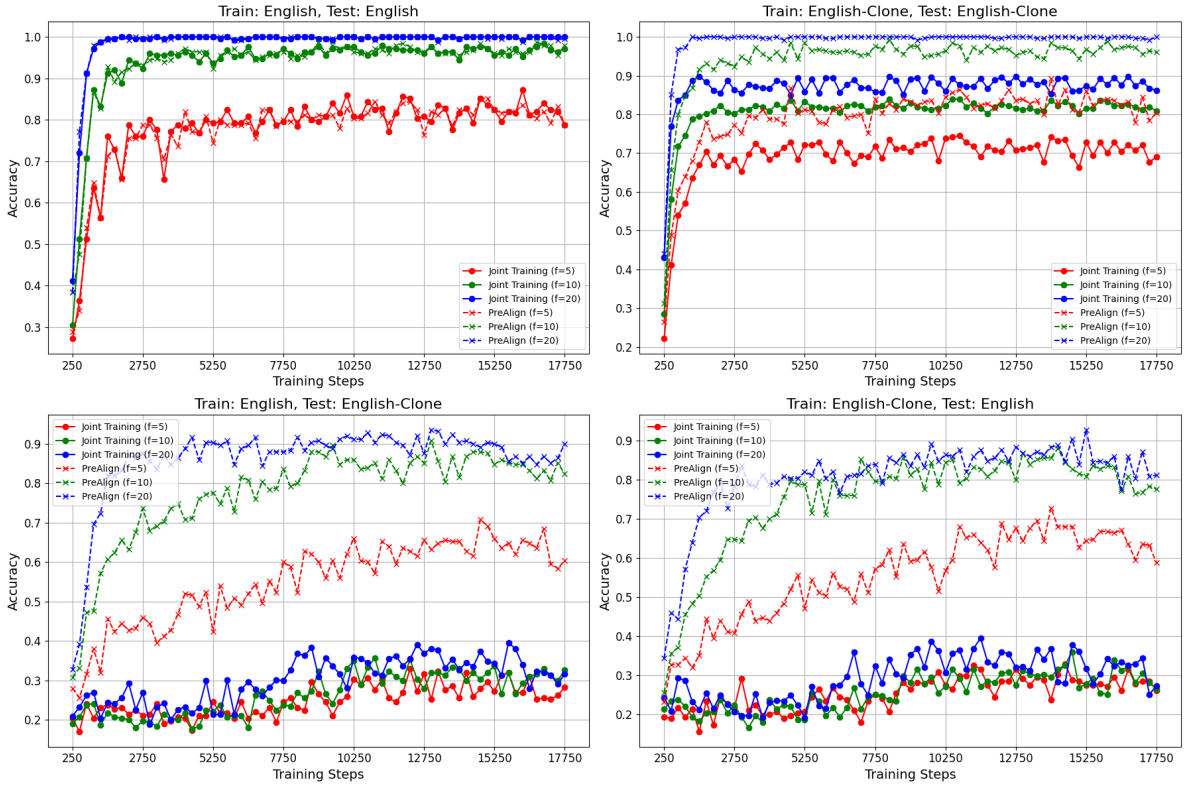

Figure 4: Knowledge application accuracy at each training period of different models. f indicates the frequency of the test knowledge.

5.3 Ablation Study

In this section, we present an ablation study of the proposed methods. The results are in Table 3.

Solely input-only CS helps LM and ZS-CLT, but not CLKA. Comparing Line #1 and Line #2, we can see that adding input-only CS to the pretraining stage can bring improvements to language modeling and downstream cross-lingual transferring performance, which is consistent with findings in previous works (Chaudhary et al., 2020; Yang et al., 2021). However, the improvement on CLKA is much smaller (27.7 $\rightarrow$ 32.6).

Multilingual alignment initialization significantly facilitates CLT, especially CLKA. By establishing multilingual alignment before language model pretraining, all considered metrics that evaluating cross-lingual transfer are significantly improved (Line #1 vs. Line #3 and Line #2 vs. Line #4). Notably, this brings a much better CLKA performance, highlighting the importance of early multilingual alignment for knowledge transferring.

Combining Multi-Align Init with input-only codeswitching achieves the best performance. Finally, by comparing Line #4 vs. Line #2 and Line #3, we can see the proposed two strategies all contributes to the good performance that PRE-ALIGN achieves.

Input-only codeswitching causes less mixed-script problem. We also compare the proposed input-only codeswitching strategy with the vanilla codeswitching strategy in Table 4, in terms of both English language modeling performance and the ratio that generation results contains En-clone tokens. It can be seen that when the training time codeswitching ratio is to 5%, adopting vanilla codeswitching strategy would result in 4.17% sentences contains En-clone tokens, which would significantly decrease the generation quality in real-world settings. However, the input-only codeswitching strategy proposed in this paper effectively decrease the ratio to 0.02%, and achieves better English LM perplexity.

5.4 Maintaining Multilingual Alignment across Pretraining.

In order to understand how the injected multilingual alignment information evolves during pretraining, we compute the similarity of aligned word embedding at different training period (every 250

	Joint Training	Multi-Align Init	Input-only CS	LM (ppl. ↓)	ZS-CLT (acc. ↑)	CLKA (acc. ↑)
#1	✓			21.6	74.9	26.5
#2	✓		✓	19.7	76.1	30.2
#3	✓	✓		17.1	77.8	85.7
#4	✓	✓	✓	**16.5**	**79.3**	**90.3**

Table 3: Ablations of PREALIGN. Multi-Align Init: using multilingual alignment objective to initialize LM. Input-only CS: the proposed data augmentation method by only codeswitching the input words. All reported performance are evaluated in English-Clone.

	LM	Mixed-Script Ratio
Original CS	17.1	4.17%
Input-only CS	16.5	0.02%

Table 4: Comparison of the original codeswitching strategy and the proposed input-only codeswitching strategy. Note the mixed-script ratio in the table refers to the portion of random English samples that contains English-clone scripts during inference.

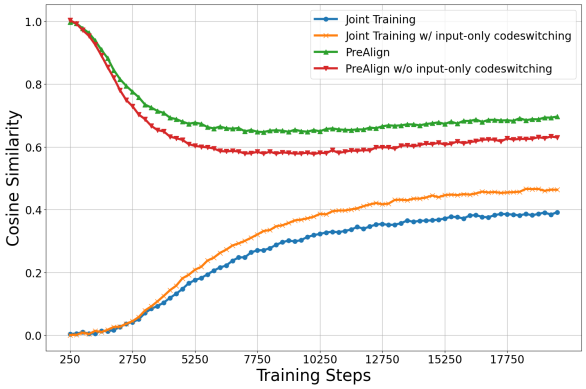

Figure 5: The evolution of word embeddings' cosine similarity between aligned words from different models.

	LM	ZS-CLT	CLKA
Joint Training	21.6	74.9	26.5
PREALIGN			
$\beta = 25\%$	17.0	78.2	80.2
$\beta = 50\%$	16.8	78.6	83.1
$\beta = 75\%$	16.6	78.8	88.4
$\beta = 100\%$	**16.5**	**79.3**	**90.3**

Table 5: Performance of PREALIGN when using different portion of aligned word pairs. For reference, we also list the performance of Joint Training.

Figure 6: Language modeling perplexity on Seen and Unseen words categorized according to multilingual alignment stage.

training steps). Figure 5 illustrates the results.

Firstly, we can see that despite there are no vocabulary overlap between English and English-clone, the embedding similarity of aligned words still grows during Joint-Training, which is consistent with findings in previous works (Dufter and Schütze, 2020). This indicates the ability of spontaneous establishment of multilingual alignment of language models. Secondly, the aligned similarity score of PREALIGN is near perfect as designed, and despite the score decreases at the beginning of pretraining, it maintains to be significantly higher than Joint-Training throughout the pretraining process. Finally, the codeswitching strategy is helpful for both Joint Training and PREALIGN, as it accelerates the increment of Joint Training's aligning similarity score, and helps slow down the decrement of PREALIGN's aligning similarity score.

5.5 Generalization to Unseen Word Pairs

In previous experiments, we assumes that we can collect translations for all words in the pretraining corpus. However, in real-world settings, this might be impractical. Therefore we present an investigation on whether we can only collect alignment table of high-frequency words, and generalize the alignment to words unseen in the alignment table.

Specifically, we sort words in our unique word set according to their frequency, and only train PREALIGN model based on the top β word alignment. Table 5 shows the results. We can see that when using the most frequent 25% words for multilingual alignment, PREALIGN can already achieve significant improvements over Joint Training. This indicates the alignment information can be general-

	LM(ppl. ↓)					ZS-CLT(acc. ↑)					CLKA(acc. ↑)			
	En	Zh	De	Ar	Ru	En	Zh	De	Ar	Ru	Zh	De	Ar	Ru
150M														
Joint Training	25.7	99.7	43.5	46.9	49.8	**80.6**	64.6	63.5	58.3	62.0	26.2	25.1	26.8	26.3
PREALIGN	**25.4**	**91.1**	**39.8**	**40.7**	**44.6**	**80.6**	**69.2**	**67.5**	**60.8**	**65.1**	**53.1**	**57.2**	**51.6**	**55.5**
400M														
Joint Training	20.3	79.8	32.5	34.8	39.6	82.3	65.8	65.3	56.9	63.7	37.8	39.5	36.1	37.7
PREALIGN	**19.9**	**75.2**	**28.3**	**30.7**	**33.6**	**82.4**	**70.0**	**69.3**	**65.6**	**68.2**	**63.8**	**66.5**	**64.7**	**63.6**
1.3B														
Joint Training	**15.8**	62.2	24.0	27.7	31.2	**84.3**	70.8	70.6	63.7	68.6	49.6	44.1	45.5	48.0
PREALIGN	16.1	**58.0**	**23.3**	**25.3**	**29.4**	83.9	**74.0**	**72.9**	**68.2**	**71.4**	**71.1**	**73.9**	**72.7**	**72.5**

Table 6: Performance of Joint Training and PREALIGN across different scale of models on language modeling, zero-shot cross-lingual transfer (ZS-CLT) and cross-lingual knowledge application (CLKA).

ize between words.

To better validate this, we split all words into Seen and Unseen according to their appearance during the multilingual alignment phase. We then compute the test LM perplexity of seen words and unseen words, and present the results in Figure 6. It can be seen that PREALIGN not only can effectively leverage seen words to enhance the language modeling ability, but only can generalize the alignment information to unseen words.

6 Experiments on Real-world Settings

We validate the effectiveness of PREALIGN under real-world settings. Performances of LM, ZS-CLT and CLKA is shown in Table 6.

PREALIGN are also effective under real-world scenarios. It can be seen from Table 6 that PREALIGN can still achieve substantially better performance compared to the original Joint Training method. This improvements is consistent across different model scales, rendering the effectiveness of PREALIGN in real-world scenarios. Interestingly, the transferring effect from English is more preeminent for German and Russian than Chinese and Arabic, indicating typological similarity between language might also play important roles in cross-lingual transferring effectiveness.

Enlarging models is beneficial for CLKA. We can also see that although Joint Training gets near-random performance at the small scale, the performance grows with the scale of model parameters. This indicates that the ability of spontaneous multilingual alignment only appears on larger models, which is consistent with finding in Qi et al. (2023).

7 Conclusion

We present the PREALIGN framework in this paper. It advances the establishment of multilingual alignment prior to language pretraining, and maintain it throughout pretraining using an input-only codeswitching strategy. Through extensive experiments and analysis, both on synthetic and real-world settings, we demonstrate the effectiveness of PREALIGN for facilitating cross-lingual ability and knowledge transfer.

Limitations

The main limitation of this paper is scale of studied models and datasets. Although we proved the effectiveness of PREALIGN up to 1.3B models, it is still very small compared to LLMs nowadays. Whether the findings in the paper holds on larger settings still remains to be explored.

Another limitation is that we only test simple factual knowledge in this paper. In real worlds, knowledge may take more complex forms, and the effectiveness of PREALIGN on these settings need to examined.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120) and Nanjing University-China Mobile Communications Group Co.,Ltd. Joint Institute.

References

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *Preprint*, arXiv:2010.12566.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pre-training and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *ArXiv*, abs/2404.04659.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10922–10943, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. Language imbalance can boost cross-lingual generalisation. *Preprint*, arXiv:2404.07982.

Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *ArXiv*, abs/2401.06838.

Henry Tang, Ameet Deshpande, and Karthik Narasimhan. 2022. Align-mlm: Word embedding alignment is crucial for multilingual pre-training. *Preprint*, arXiv:2211.08547.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F. Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *Preprint*, arXiv:2309.04766.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, ShuangZhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9386–9393.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Ziqing Yang, Wentao Ma, Yiming Cui, Jiani Ye, Wanxiang Che, and Shijin Wang. 2021. Bilingual alignment pre-training for zero-shot cross-lingual transfer. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 100–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024a. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *Preprint*, arXiv:2405.01345.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. *Preprint*, arXiv:2401.07817.

A Experimental Details

Collection of the multilingual table We recognize words using the *word_tokenize* function from NLTK library. The word set consists of all words, including named entities, that appears above 20 times in the pretraining corpus. We translate all words using GPT-4, by asking it to generate 5 most common translations of a given word (without placing it in a context).

Training details During the multilingual alignment stage, we set the τ to be 0.1. During the language pretraining stage, we independently sample sentences for LM loss and word pairs for the alignment loss at each training step. We ran all experiments on 8×A100 GPUs. The multilingual alignment stage takes about 500 steps, and the language pretraining stage takes about 24000 steps. The running time of different sizes of models ranges from 4 hours to 24 hours.

Example of the synthesized knowledge We collect relations from WikiData, and ask GPT-4 to compose templates for each relation. We then fill in the person name to synthesize knowledge about people. For example, if the subject, relation and object are *Oprah Winfrey, godparent* and *Tyler Perry* respectively, then the composed knowledge is *Oprah Winfrey is the godparent of Tyler Perry.*

B Experimental results on the PAWSX dataset

To further validate the effectiveness of PreAlign, we conduct additional experiments on the PAWSX dataset (Yang et al., 2019), and show the result in Table 7. It can be seen that PreAlign can still bring consistent improvements across different model scales.

	En	De	Zh
150M			
Joint Training	**90.6**	71.5	78.9
PREALIGN	90.1	**76.1**	**83.7**
400M			
Joint Training	**92.6**	75.1	83.3
PREALIGN	92.3	**78.9**	**85.6**
1.3B			
Joint Training	94.1	79.7	85.2
PREALIGN	**94.3**	**82.4**	**87.9**

Table 7: ZS-CLT performance of Joint Training and PREALIGN across different scale of models on the PAWSX dataset.