Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment

Anonymous ACL submission

Abstract

Multilingual pre-trained models (mPLMs) have shown impressive performance on cross-lingual transfer tasks. However, the transfer performance is often hindered when a low-resource target language is written in a different script than the high-resource source language, even though the two languages may be related or share parts of their vocabularies. Inspired by recent work that uses transliteration to address this problem, our paper proposes a transliteration-based post-pretraining alignment (PPA) method aiming to improve the cross-lingual alignment between languages using diverse scripts. We select two areal language groups, **Mediterranean-Amharic-Farsi** and **South+East Asian Languages**, wherein the languages are mutually influenced but use different scripts. We apply our method to these language groups and conduct extensive experiments on a spectrum of downstream tasks. The results show that after PPA, models consistently outperform the original model (up to 50% for some tasks) in English-centric transfer. In addition, when we use languages other than English as sources in transfer, our method obtains even larger improvements. We will make our code and models publicly available.

1 Introduction

Recent mPLMs such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have shown remarkable performance on cross-lingual transfer tasks by learning cross-lingual representations from monolingual corpora (Pires et al., 2019; Artetxe et al., 2020a). Despite their impressive performance, these models still exhibit limitations in cross-lingual transfer involving low-resource languages. Deshpande et al. (2022) showed that the downstream performance of mPLMs is correlated with the degree of alignment between word embeddings across languages. Another factor that hinders the knowledge transfer is the script diversity or script barrier of represented languages, which has been observed even in the case of related languages (Anastasopoulos and Neubig, 2019; Muller et al., 2021). The script barrier problem can also be viewed from the perspective of representation alignment. Wen-Yi and Mimno (2023) showed that token representations from different scripts could be almost perfectly linearly separated, indicating that models struggle to learn a common representation space. Therefore, post-training is required to boost zero-shot cross-lingual transfer in tasks like sentence retrieval, text classification, or sequence labeling, all of which benefit from better cross-lingual alignment (Hämmerl et al., 2024).

Many post-training alignment strategies use objectives that rely on bilingual dictionaries or parallel data to align the representations of mPLMs (Cao et al., 2020; Wang et al., 2020; Schuster et al., 2019; Pan et al., 2021). However, dictionaries and parallel corpora are often limited in the data scale or the number of languages they cover (Artetxe et al., 2020b), which might be impractical for building strong supervision signals for many low-resource languages. Another alternative that improves cross-lingual alignment is to use *transliteration* (a process of converting the text of a language from one script to another). Transliteration can improve the lexical overlap, especially for related languages (Moosa et al., 2023). Different from translation, transliterations can be obtained nearly for free using well-performing rule-based transliteration tools (Hermjakob et al., 2018). Therefore, several works have shown improvements in cross-lingual transfer by pre-training or fine-tuning models with data transliterated into a common script (Murikinati et al., 2020; Muller et al., 2021; Purkayastha et al., 2023; Moosa et al., 2023). However, these works require the use of a single common script by the model. This is restrictive for many tasks as the process of transliteration can be lossy and non-invertible.

Recently, Liu et al. (2024a) proposed a sequence-level contrastive learning objective to improve the alignment across different scripts at a large scale (for more than 500 languages), using sentences in both their original scripts and their Latin-script transliterations, validated by English-centric cross-lingual zero-shot transfer evaluations. However, there are three major limitations in their setup. First, the contrastive objective only manipulates the sequence-level representations in the middle layer, which does not directly contribute to better alignment in the token-level space. Second, not every language pair has extensive lexical overlap that can boost cross-lingual transfer: transliteration-based alignment makes more sense for mutually influenced languages. Lastly, English alone as a transfer source language does not fully exploit the alignment benefit, as it does not have the most lexical overlap with other languages.

To this end, we propose a new transliteration-based post-training alignment method that works on both **sequence** and **token** levels. Our method does not rely on parallel data. Instead, similar to Liu et al. (2024a), we use the monolingual data in their original scripts and their Latin transliteration obtained by using `Uroman` (Hermjakob et al., 2018), a rule-based transliteration tool. We investigate the impact of the strategy by focusing on two groups of languages: **Mediterranean-Amharic-Farsi** and **South+East Asian Languages**, described more in detail in Section 4.2. The languages in each group share areal features but differ in scripts. Some languages are closely related as members of the same language family (e.g., Semitic and Sino-Tibetan). Additionally, languages in each group have extensive lexical overlap due to historical contact and geographical proximity (e.g., Chinese and Korean or Turkish and Arabic). As these languages are written in different scripts, transliteration can help to better exploit the shared linguistic properties and thus improve the cross-lingual transfer performance.

We leverage our method to post-train Glot500 (ImaniGooghari et al., 2023) (a continually pre-trained model from XLM-R on more than 500 languages) on the selected language groups and evaluate the zero-shot cross-lingual transfer performance on three types of downstream tasks: sentence retrieval, text classification, and sequence labeling. The evaluation is done with English as the source language, as well as with three other source languages of different scripts in each group. We show that our method consistently improves the downstream task performance across different languages and scripts. Moreover, our method further boosts the transfer performance when better source languages are chosen, as the performance depends on the degree of alignment between the source and target languages – precisely the alignment that our approach boosts.

Our contributions can be summarized as follows: (i) We propose a transliteration-based post-training alignment method that operates on both sequence and token levels, aiming to bridge the script barrier in mPLMs; (ii) We investigate the impact of our method on two areal groups of languages with different scripts and show consistent improvements in zero-shot cross-lingual transfer; (iii) We systematically explore how different source languages influence the zero-shot transfer performance of our obtained transliteration-aligned models.

2 Related Work

Many recent works have proposed pre-training or fine-tuning alignment methods to improve cross-lingual transfer in mPLMs. Cao et al. (2020) proposed a fine-tuning embedding alignment objective between word pairs procured in an unsupervised fashion from parallel data using statistical word alignment models (Dyer et al., 2013). Chaudhary et al. (2020) improved alignment during pre-training by using bilingual dictionaries to replace words in original sentences with translations in other languages. Similarly, Tang et al. (2022) used bilingual dictionaries to explicitly align the embeddings of the same words in different languages during pre-training. Wei et al. (2021) proposed a hierarchical contrastive learning pre-training method, which uses parallel data to align representations at the word and sentence levels. Similarly, Hu et al. (2021) proposed a pre-training method with explicit alignment signals from parallel data that encourages symmetry at both word and sentence levels. Pan et al. (2021) combined contrastive learning with translation language modeling (Conneau and Lample, 2019) as a post-training alignment method that uses parallel data as well. While these methods have shown improvements in cross-lingual transfer, they have the limitation of requiring parallel data or bilingual dictionaries, which may be hard to acquire for many low-resource languages.

Transliteration is a process of converting the text of a language from one script to another (Wellisch et al., 1978). This process does not involve translating meanings but rather represents the original symbols as closely as possible in the target script. Different works have proposed transliteration-based methods to address the script barrier problem in multilingual models. Murikinati et al. (2020) used transliteration to a common script to improve cross-lingual morphological inflection. Khemchandani et al. (2021) exploited language relatedness between Indian languages and leveraged transliteration to a common script to adapt multilingual models to low-resource languages. Muller et al. (2021) analyzed the behavior of multilingual models on unseen languages and found that languages written in different scripts do not benefit from transfer learning. They proposed transliteration to the high-resource source language script as a solution to address the script barrier. Purkayastha et al. (2023) showed that fine-tuning multilingual models on data transliterated into Latin script improves cross-lingual transfer for low-resource languages. Similarly, Moosa et al. (2023) pre-trained models from scratch on data transliterated into a common script for the Indic languages and showed improvements in cross-lingual transfer. Our work is most related to TRANSLICO proposed by Liu et al. (2024a), where a sequence-level contrastive learning objective is used to encourage alignment across different scripts without restricting the models to a common script, using data in their original scripts and their Latin-script transliteration. However, their English-centric evaluation setup limits the ability to fully reveal the impact of transliteration. This paper systematically explores how transliteration-based alignment enhances the performance of using various source languages.

3 Methods

We present a transliteration-based post-training alignment method that can be used to fine-tune existing encoder-only mPLMs for improved alignment across languages using different scripts, boosting cross-lingual transfer performance. Our method consists of three objectives: masked language modeling, sentence-level alignment, and token-level alignment. All objectives are trained on combined original and transliterated data. The transliterated data is obtained by converting the original data into Latin script. The transliteration process uses `Uroman` (Hermjakob et al., 2018), a rule-based system that can convert nearly all character sets into a common Latin script. The overall method is illustrated in Figure 1, and we introduce the three objectives in detail in the following.

3.1 Masked Language Modeling

Given an input sequence in its original script: X_i^{orig} or its transliterated version: X_i^{latn}, we apply the naive MLM objective (Devlin et al., 2019) to predict randomly masked tokens in both sequences:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}\left[-\sum_{m\in\mathcal{M}} \log p_{\text{MLM}}(X_{i,m}|\boldsymbol{h}_{i,m})\right]$$

where $\mathcal{M}$ is the set of masked positions in the input sentence X_i (either X_i^{orig} or X_i^{latn}) and $P_{\text{MLM}}(X_{i,m}|\boldsymbol{h}_{i,m}$ is the probability of predicting token $X_{i,m}$ giving $\boldsymbol{h}_{i,m}$, the final contextualized representation at the position m in the ith sequence. The probability is computed by an MLM head. Fine-tuning with MLM on original data is necessary to preserve the model's knowledge. On the other hand, the mPLM has very limited knowledge about the transliterated data, which makes the MLM objective on transliterated data crucial for learning useful cross-script representations. We refer to the MLM objective for the original data (resp. transliterated data) as $\mathcal{L}_{\text{MLM}}^{orig}$ (resp. $\mathcal{L}_{\text{MLM}}^{latn}$).

3.2 Sentence-Level Alignment

We treat an input sequence in the original script X_i^{orig} and its transliterated version X_i^{latn} as having the same semantics. Therefore, we apply a sequence-level contrastive learning objective, similar to SimCSE (Gao et al., 2021), to encourage the model to learn similar sequence-level representations for the original and transliterated sequences. This setting is analogous to other works that apply contrastive learning on pairs formed by an original sentence and its English translation (Chi et al., 2021; Pan et al., 2021). In our context, Latin acts as a pivot script, encouraging better cross-lingual alignment of representations in different scripts.

Following Liu et al. (2024a), we apply the contrastive learning objective on a given batch of original and transliterated sequences $B = \{(X_i^{orig}, X_i^{latn})\}_{i=1}^N$. Each batch defines positive contrastive pairs (X, X^+) where X is the original sequence and X^+ is its transliterated version or vice versa, i.e., (X_i^{orig}, X_i^{latn}) or (X_i^{latn}, X_i^{orig}). For each positive pair, the negative examples are formed by all other sequences in the batch $B^- = B \setminus \{(X_i^{orig}, X_i^{latn})\}$ (slightly abusing notation).

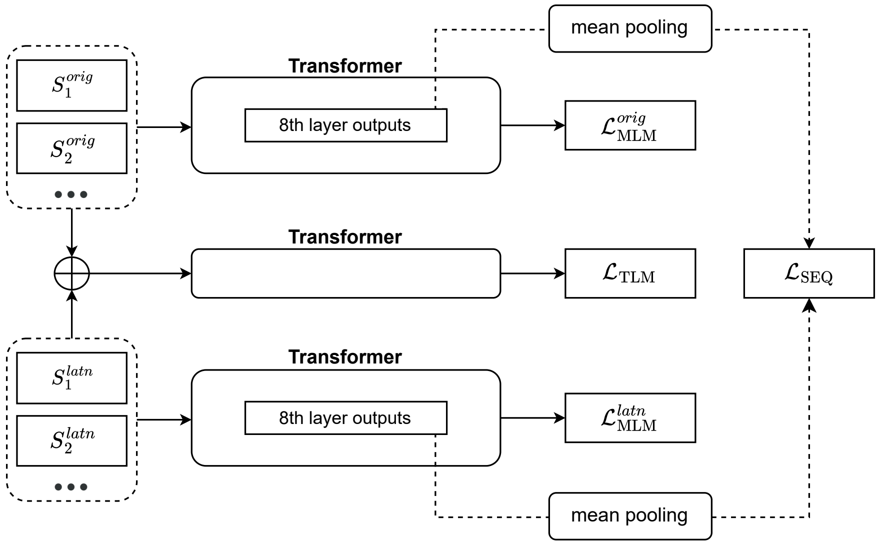

Figure 1: Overview of our transliteration-based post-training alignment method consisting of three objectives: masked language modeling ($\mathcal{L}^{orig}_{\text{MLM}}$ and $\mathcal{L}^{latn}_{\text{MLM}}$), sentence-level alignment ($\mathcal{L}_{\text{SEQ}}$), and token-level alignment ($\mathcal{L}_{\text{TLM}}$).

The contrastive loss is then defined as:

$$\mathcal{L}_{\text{SEQ}} = \mathbb{E}\left[-\log \frac{e^{\text{sim}(f(X),f(X^+))/\tau}}{e^{\text{sim}(f(X),f(X^+))/\tau} + \text{NEG}}\right]$$

where $\text{NEG} = \sum_{(X,X^-)\in B^-} e^{\text{sim}(f(X),f(X^-))/\tau}$, f is defined as mean pooling over the 8th layer output contextualized embeddings (ignoring the special tokens output except for [mask] token), sim is the dot product, and τ is the temperature set to 1.

3.3 Token-Level Alignment

The sentence-level alignment objective helps the model to learn similar sentence representations for the original and transliterated sequences. This is useful for improving performance in sentence-level downstream tasks like sentence retrieval or classification. However, this objective manipulates the output of a middle layer, which does not directly contribute to better alignment in the token-level space. For token-level tasks like NER and POS tagging, alignment at the token level might be more beneficial. Therefore, we propose a token-level alignment objective that further encourages the model to align the representations of the original and transliterated words. We adapt the translation language modeling objective introduced by Conneau and Lample (2019), which is equivalent to applying the MLM objective on a concatenated bilingual sentence pair. Specifically, given a sentence pair (X_i^{orig}, X_i^{latn}), we apply the MLM objective on the concatenated sequence $X_i^{orig} \oplus X_i^{latn}$ or $X_i^{latn} \oplus X_i^{orig}$, where concatenation order is randomly chosen during training. The intuition is that, to predict a token masked in the original sentence, the model can either attend to surrounding tokens in the original script or their transliterations and vice versa. This encourages the model to align the representations in the original script and the Latin script. We refer to this objective as **transliteration language modeling** (TLM) and the loss as $\mathcal{L}_{\text{TLM}}$.

The overall training objective combines the masked language modeling, sentence-level alignment, and token-level alignment objectives:

$$\mathcal{L} = \mathcal{L}^{orig}_{\text{MLM}} + \mathcal{L}^{latn}_{\text{MLM}} + \mathcal{L}_{\text{SEQ}} + \mathcal{L}_{\text{TLM}}$$

4 Experiments

4.1 General Setups

We use the Glot500 model (ImaniGooghari et al., 2023), a state-of-the-art multilingual encoder-only model pre-trained on more than 500 languages, as our source model for all our experiments. We fine-tune Glot500 on two groups of languages using the proposed transliteration-based post-training alignment method. The languages for each group are selected based on **areal features** so that they have some lexical overlap in different degrees and cover various scripts. We then evaluate the two resulting models on several downstream tasks in a zero-shot cross-lingual transfer manner. Apart from the standard transfer setting with English as the source language, we also evaluate the model's transfer capabilities with three other source languages of different scripts for each language group.

Figure 2: Geographical distribution of languages selected in each group. **Mediterranean-Amharic-Farsi** are shown with circles, while **South+East Asian Languages** are shown with squares.

4.2 Languages, Data and Models

The two language groups are named as **Mediterranean-Amharic-Farsi** and **South+East Asian Languages**. We visualize each group's geographical distribution of the selected languages in Figure 2. The languages within each group are spoken in adjacent areas and there is a long history of linguistic influence between them. For example, Arabic has had extensive contact with languages such as Turkish and Persian (Versteegh, 2001). The data for each language is sampled from the Glot500-c training dataset (ImaniGooghari et al., 2023). We sample 10% of the available data for each language, or a minimum of 10k sentences, whichever is larger. The data is then transliterated into Latin script using `Uroman` (Hermjakob et al., 2018). Table 1 shows each group's languages and the number of sampled sentences. In total, Mediterranean-Amharic-Farsi consists of 10 languages, 5 scripts and around 16M sentences, while South+East Asian Languages consists of 10 languages, 7 scripts and around 4M sentences. We fine-tune Glot500 using our alignment method on each group separately. We then select the best checkpoint for each group by validating the checkpoints' performance on the Tatoeba (Artetxe and Schwenk, 2019) sentence retrieval dataset, which contains 1000 English-aligned sentences. We compute the top-10 retrieval accuracy based on the cosine similarity of the averaged 8th-layer contextual embeddings. The best checkpoint for each group is regarded as our final aligned model.

Language	Script Code	Language Code	Num. Sent.
Mediterranean-Amharic-Farsi			
Macro Lang. Arabic	Arab	ara	2.4M
Standard Arabic	Arab	arb	15k
Moroccan Arabic	Arab	ary	10k
Egyptian Arabic	Arab	arz	348k
Macro Lang. Farsi	Arab	fas	1.8M
Amharic	Ethi	amh	286k
Greek	Grek	ell	2.2M
Hebrew	Hebr	heb	1.8M
Turkish	Latn	tur	2.9M
Maltese	Latn	mlt	4M
South+East Asian Languages			
Macro Lang. Chinese	Hani	zho	2.4M
Classical Chinese	Hani	lzh	10k
Yue Chinese	Hani	yue	48k
Wu Chinese	Hani	wuu	22k
Korean	Hang	kor	646k
Lao	Laoo	lao	10k
Lahu	Latn	lhu	10k
Burmese	Mymr	mya	94k
Tibetan	Tibt	bod	27k
Thai	Thai	tha	773k

Table 1: Basic information and number of sampled sentences for each language in the two language groups.

4.3 Downstream Tasks

We evaluate the resulting aligned model for each group on several downstream tasks (described below). For each task, we use **four** different source languages, **English** and **three other source languages** belonging to the same group that use different scripts. The evaluation is done in a zero-shot cross-lingual transfer manner: we fine-tune the models on the train set of a given source language, select the best checkpoint, and compute the macro F1 score (except for SR-B where we compute the average top-10 retrieval accuracy) on the test sets of the remaining languages in each group. Note that no training step is needed for the retrieval task: we directly use the sentences from the source

	SR-B				Taxi1500				SIB200				NER				POS			
	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr
Glot500																				
tur_Latn	63.2	77.6	51.8	32.2	63.0	63.5	**55.6**	**44.1**	81.4	79.8	82.1	79.8	74.1	71.4	71.7	70.4	70.4	49.4	63.8	66.5
mlt_Latn	50.4	55.0	69.2	44.8	54.1	57.3	53.0	**46.1**	81.8	79.3	82.8	82.6	69.2	60.0	68.2	66.9	81.1	59.8	76.9	74.2
ell_Grek	48.6	58.6	src	40.8	**64.0**	**61.4**	src	**44.1**	77.7	74.4	src	81.7	72.7	72.1	src	72.6	**86.1**	58.3	src	67.8
heb_Hebr	21.8	27.8	33.8	src	35.4	**44.9**	39.0	src	77.6	73.9	**81.7**	src	48.9	58.7	52.6	src	**68.3**	70.1	**60.6**	src
amh_Ethi	52.8	64.6	51.2	33.2	7.2	**10.4**	12.7	**15.1**	73.1	74.9	74.1	74.8	43.8	**52.5**	54.0	46.6	**66.5**	**65.1**	**64.3**	73.6
ara_Arab	-	-	-	-	-	-	-	-	-	-	-	-	57.2	src	56.7	61.5	**84.6**	src	**63.2**	**78.0**
arz_Arab	24.8	33.6	52.8	44.8	35.1	43.1	45.2	**42.7**	79.7	src	81.8	80.4	**58.4**	75.1	63.8	65.2	-	-	-	-
ary_Arab	15.2	16.4	29.0	28.4	35.8	**40.3**	**41.6**	39.6	79.9	80.2	**84.0**	82.0	-	-	-	-	-	-	-	-
arb_Arab	14.6	23.0	29.0	32.2	-	-	-	-	79.9	79.9	82.8	81.2	-	-	-	-	-	-	-	-
fas_Arab	89.2	src	72.4	40.2	71.0	src	59.2	**48.7**	-	-	-	-	**49.7**	66.2	58.2	50.6	71.5	67.2	60.9	72.0
Average	42.2	44.5	48.6	37.0	45.7	45.8	43.8	**40.1**	78.9	77.5	81.3	80.4	59.3	65.1	60.8	62.0	72.8	61.6	65.0	72.0
Ours																				
tur_Latn	**81.0**	**91.2**	**77.8**	**49.4**	**64.8**	**65.1**	54.6	38.2	**85.6**	**86.0**	**84.8**	**85.6**	**77.0**	**73.1**	**76.9**	**73.2**	**73.6**	**52.8**	**66.4**	**68.0**
mlt_Latn	**85.6**	**93.4**	**90.4**	**59.4**	**66.6**	**60.9**	**58.4**	42.5	**86.2**	**85.8**	**84.6**	**85.4**	**75.2**	**72.0**	**73.9**	**75.2**	**83.1**	**63.0**	**77.1**	**77.8**
ell_Grek	**68.0**	**85.4**	src	**45.2**	63.6	60.4	src	36.4	**82.3**	**81.4**	src	**82.1**	**73.8**	**73.9**	src	**75.3**	85.9	**58.7**	src	**70.7**
heb_Hebr	**29.0**	**32.0**	**42.8**	src	**45.3**	44.5	**45.8**	src	**79.0**	**79.8**	79.9	src	**51.9**	**61.6**	**57.2**	src	67.7	**71.3**	59.8	src
amh_Ethi	**63.6**	**79.4**	**64.8**	**49.4**	**7.6**	9.6	**17.0**	11.3	**77.2**	**77.9**	**76.6**	**76.9**	**45.0**	51.7	54.0	**50.0**	66.2	63.9	63.7	**74.9**
ara_Arab	-	-	-	-	-	-	-	-	-	-	-	-	**59.9**	src	**60.4**	**65.1**	65.3	src	62.3	77.7
arz_Arab	**56.4**	**79.4**	**82.2**	**69.6**	**41.8**	**44.1**	**49.6**	42.4	**83.1**	src	**83.4**	**82.8**	58.3	**76.7**	**65.6**	**68.3**	-	-	-	-
ary_Arab	**47.6**	**66.2**	**66.2**	**65.4**	**39.0**	37.3	39.0	**40.5**	**83.2**	**82.7**	82.9	**83.3**	-	-	-	-	-	-	-	-
arb_Arab	**44.4**	**55.0**	**56.0**	**49.6**	-	-	-	-	**82.8**	**83.3**	**83.1**	**83.1**	-	-	-	-	-	-	-	-
fas_Arab	**89.6**	src	**87.0**	**57.8**	**71.9**	src	**63.2**	37.5	-	-	-	-	48.6	63.1	**62.3**	**56.9**	**71.6**	**69.1**	**61.5**	71.3
Average	**62.8**	**72.7**	**70.9**	**55.7**	**50.1**	**46.0**	**46.8**	35.5	**82.4**	**82.4**	**82.2**	**82.7**	**61.2**	**67.4**	**64.3**	**66.3**	**73.3**	**63.1**	**65.1**	**73.4**

Table 2: Cross-lingual transfer results across 5 tasks on the **Mediterranean-Amharic-Farsi** group. Columns represent the script of the source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each source-target language pair, the best score is **bolded**. For each target language, we underline the best source transfer score for each task (for both Glot500 and our method).

language as the queries and retrieve the most similar sentences in the target languages. For tasks that require additional fine-tuning, we report the results averaged over five different seeds. The downstream tasks are as follows (see details in §B):

SR-B A sentence retrieval dataset where the parallel sentences are from the Bible. We compute the top-10 retrieval accuracy on 500 parallel sentences following ImaniGooghari et al. (2023).

Taxi1500 A multilingual text classification dataset covering more than 1500 languages with sentences from 6 topics (Ma et al., 2023).

SIB200 A multilingual text classification dataset covering more than 200 languages for 7 topics (Adelani et al., 2024).

NER A multilingual sequence labeling dataset for named entity recognition (Pan et al., 2017) that consists of articles annotated with 7 different tags, e.g., location, person, etc.

POS A multilingual sequence labeling dataset for part-of-speech (POS) tagging (de Marneffe et al., 2021) consisting of sentences annotated with 17 universal POS tags, e.g., NOUN, ADJ, etc.

5 Results and Analysis

We report the results of Glot500 and our post-trained aligned models on the downstream tasks in Table 2 for Mediterranean-Amharic-Farsi and in Table 3 for South+East Asian Languages. Overall, our aligned models outperform Glot500 across different tasks for both language groups, occasionally with a slight performance drop for certain source-target language combinations. In the following, we highlight our essential findings from the results.

Per-group performances differ slightly. Starting with **Mediterranean-Amharic-Farsi**, we observe that the post-trained aligned model generally outperforms Glot500 on all downstream tasks. The SR-B task shows the most significant improvement, with the aligned model achieving, on average, more than 20% higher accuracy than Glot500 for all source languages. For other tasks, the aligned model also demonstrates a consistent, albeit smaller, improvement, with Glot500 occasionally outperforming the post-trained aligned model for specific source-target language pairs. However, we observe a more mixed performance for the **South+East Asian Languages**, especially for sequence labeling tasks, namely NER and POS, with

	SR-B				Taxi1500				SIB200				NER				POS		
	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani
Glot500																			
tha_Thai	**45.4**	47.2	42.6	src	64.1	63.8	**73.6**	src	82.0	83.1	**83.1**	src	**4.3**	**2.8**	7.5	src	**55.0**	29.8	**49.0**
kor_Hang	**61.0**	src	64.6	51.2	68.6	src	63.5	65.5	82.7	src	**83.0**	82.5	51.8	src	40.2	**10.1**	52.6	src	39.2
yue_Hani	24.0	31.6	65.8	44.4	64.0	62.8	68.0	62.1	84.7	84.4	src	86.9	**24.0**	**37.4**	69.1	**16.4**	**38.8**	**49.3**	78.5
wuu_Hani	-	-	-	-	-	-	-	-	-	-	-	-	35.1	**58.3**	62.4	**16.1**	-	-	-
zho_Hani	**44.4**	46.4	src	39.6	65.0	61.9	src	62.0	-	-	-	-	**23.6**	32.7	src	15.0	**40.1**	**49.9**	src
lzh_Hani	**63.4**	**65.8**	**75.8**	**43.2**	57.3	**60.5**	**61.5**	**56.1**	-	-	-	-	**12.0**	29.8	60.1	**22.3**	**19.4**	**27.2**	50.7
lao_Laoo	49.6	59.8	48.6	64.6	**72.0**	68.5	**73.9**	**74.8**	80.4	80.0	80.0	82.4	-	-	-	-	-	-	-
lhu_Latn	5.0	6.0	6.6	7.2	**27.0**	34.1	**27.8**	26.4	-	-	-	-	-	-	-	-	-	-	-
mya_Mymr	29.4	37.8	29.0	33.6	61.8	63.2	56.8	60.5	80.1	80.7	78.6	79.5	54.1	**65.2**	**49.7**	**9.3**	-	-	-
bod_Tibt	33.2	49.4	44.4	49.8	-	-	-	-	70.0	68.8	65.1	**72.7**	36.5	42.8	**50.0**	**25.5**	-	-	-
Average	39.4	43.0	47.1	41.7	60.0	60.7	**62.1**	59.9	79.3	79.4	78.0	80.8	30.2	38.4	**48.4**	**16.4**	**41.2**	**39.1**	54.4
Ours																			
tha_Thai	45.2	**85.6**	**55.2**	src	**66.3**	**66.7**	67.8	src	**86.6**	**85.9**	82.9	src	3.5	2.6	**8.6**	src	50.6	**30.3**	48.6
kor_Hang	58.8	src	**79.6**	**76.6**	**71.3**	src	**65.8**	**68.2**	**83.1**	src	82.4	**83.8**	**55.5**	src	**43.2**	3.1	**52.8**	src	**45.0**
yue_Hani	**71.4**	**91.8**	**98.8**	**89.8**	**64.5**	**66.4**	**69.2**	**67.9**	**87.2**	**84.7**	src	**89.0**	21.0	36.1	**70.1**	15.8	29.3	38.6	**79.3**
wuu_Hani	-	-	-	-	-	-	-	-	-	-	-	-	**45.5**	56.7	**64.7**	1.7	-	-	-
zho_Hani	41.4	**75.0**	src	**46.4**	**65.8**	**65.0**	src	**67.7**	-	-	-	-	21.5	**35.9**	src	**15.6**	32.2	44.1	src
lzh_Hani	37.4	48.6	67.6	37.0	**63.4**	58.3	61.4	55.2	-	-	-	-	11.7	**33.7**	**61.4**	20.0	13.9	14.3	**52.1**
lao_Laoo	**56.4**	**90.8**	**67.0**	**77.4**	70.2	**69.5**	69.4	67.2	**82.6**	**81.6**	**81.0**	**83.2**	-	-	-	-	-	-	-
lhu_Latn	**15.8**	**26.6**	**21.0**	**27.4**	23.9	**36.5**	24.9	**32.1**	-	-	-	-	-	-	-	-	-	-	-
mya_Mymr	**37.8**	**60.4**	**54.4**	**59.6**	**62.2**	**68.3**	**59.2**	**63.0**	**80.9**	**81.8**	**80.0**	**81.0**	**56.0**	62.4	45.0	5.8	-	-	-
bod_Tibt	**60.8**	**88.8**	**83.4**	**80.2**	-	-	-	-	**70.5**	**73.5**	**71.4**	71.4	**40.3**	**43.1**	41.6	1.1	-	-	-
Average	**47.2**	**70.9**	**65.8**	**61.8**	**61.0**	**61.5**	59.7	**60.2**	**81.8**	**81.5**	**79.6**	**81.7**	**31.9**	**38.6**	47.8	9.0	35.8	31.8	**56.3**

Table 3: Cross-lingual transfer results across 5 tasks on the **South+East Asian Languages** group. Columns represent the script of the source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each source-target language pair, the best score is **bolded**. For each target language, we underline the best source transfer score for each task (for both Glot500 and our method)

the aligned models performing worse than Glot500 when transferring from more than half of the source languages. We hypothesize that this performance drop is primarily due to the transliteration process, which loses the semantic and contextual nuances and induces more token-level ambiguity for most of the languages (Amrhein and Sennrich, 2020; Liu et al., 2024a). This ambiguity makes token-level alignment more difficult. In contrast, the aligned model achieves consistent improvements in NER and POS for Mediterranean-Amharic-Farsi, where less token-level ambiguity is introduced as the languages are originally written in phonetic scripts, to which Latin also belongs.

Source languages matter. We observe that the performance can vary significantly for both groups of languages depending on the source language used for transfer. This phenomenon occurs due to the script and language similarity between specific source and target languages. In general, transferring from in-group high-resource languages performs better than transferring from English. Taking the SR-B task for example, for Mediterranean-Amharic-Farsi, the best performance is achieved when transferring from Farsi. For the South+East Asian Languages, the best performance is achieved when transferring from Korean. The text classification tasks generally show less variation in performance, though transferring from Hebrew achieves the worst performance for Mediterranean-Amharic-Farsi in Taxi1500. For the NER and POS tasks, transferring from Arabic and Hebrew achieves the best performance for Mediterranean-Amharic-Farsi while transferring from Chinese achieves the best performance for both tasks in South+East Asian Languages. Nevertheless, comparing our aligned models against Glot500, the performance generally improves for most source languages. This indicates that our proposed transliteration-based post-training method effectively improves the alignment between related languages and further boosts performance when a proper source language is chosen.

5.1 Ablation Study

We perform an ablation study to investigate the impact of different training objectives on the performance of the post-trained aligned models. Starting from the base Glot500 model, we apply different combinations of the training objectives: masked language modeling (MLM), sentence-level alignment (SEQ), and token-level alignment

	SR-B				Taxi1500				SIB200				NER				POS			
	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr
Glot500	42.2	44.5	48.6	37.0	45.7	45.8	43.8	**40.1**	78.9	77.5	81.3	80.4	59.3	65.1	60.8	62.0	72.8	61.6	<u>65.0</u>	72.0
MLM	50.2	53.2	55.4	40.2	47.7	43.8	46.9	31.8	**82.6**	**82.4**	<u>82.5</u>	80.4	<u>61.7</u>	**68.5**	63.3	<u>66.5</u>	71.9	<u>62.4</u>	63.7	72.0
MLM+SEQ	**62.9**	<u>71.9</u>	<u>69.8</u>	**57.1**	49.0	<u>46.0</u>	**48.5**	36.1	<u>82.4</u>	**82.4**	**83.5**	<u>82.4</u>	61.1	<u>67.4</u>	**64.6**	66.1	72.5	62.1	64.7	72.7
MLM+TLM	50.4	54.7	55.7	41.6	**50.4**	**46.5**	<u>47.3</u>	<u>37.2</u>	**82.6**	<u>81.7</u>	82.4	81.8	**61.9**	67.1	64.0	**66.8**	<u>73.2</u>	**63.1**	64.5	**73.5**
MLM+SEQ+TLM	<u>62.8</u>	**72.7**	**70.9**	<u>55.7</u>	<u>50.1</u>	<u>46.0</u>	46.8	35.5	<u>82.4</u>	**82.4**	82.2	**82.7**	61.2	<u>67.4</u>	<u>64.3</u>	66.3	**73.3**	**63.1**	**65.1**	<u>73.4</u>

Table 4: Ablation study results for **Mediterranean-Amharic-Farsi**. The columns represent the script of the source language. The results are averaged over all target languages. **Bold** (<u>underlined</u>): best (second-best) result.

	SR-B				Taxi1500				SIB200				NER				POS		
	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani
Glot500	39.4	43.0	47.1	41.7	60.0	59.2	**60.7**	58.2	79.3	79.4	78.0	80.8	30.2	38.4	**48.4**	**16.4**	**41.2**	**39.1**	54.4
MLM	45.8	54.3	54.7	51.4	58.5	60.2	<u>60.4</u>	<u>60.3</u>	80.9	<u>80.3</u>	79.4	<u>81.6</u>	28.7	37.9	46.7	7.2	29.5	27.3	54.2
MLM+SEQ	<u>47.0</u>	<u>63.8</u>	<u>61.0</u>	<u>56.1</u>	60.6	59.9	60.0	**61.0**	<u>81.8</u>	79.3	79.3	80.3	30.9	**39.3**	46.5	7.7	31.4	30.9	<u>55.4</u>
MLM+TLM	45.4	55.7	57.8	54.5	**61.4**	<u>60.6</u>	59.5	60.0	**81.9**	**81.5**	<u>79.5</u>	81.5	<u>31.0</u>	<u>39.0</u>	47.6	<u>9.1</u>	34.6	<u>32.1</u>	54.8
MLM+SEQ+TLM	**47.2**	**70.9**	**65.8**	**61.8**	<u>61.0</u>	**61.5**	59.7	60.2	<u>81.8</u>	**81.5**	**79.6**	**81.7**	**31.9**	38.6	<u>47.8</u>	9.0	<u>35.8</u>	31.8	**56.3**

Table 5: Ablation study results for **South+East Asian Languages**. The columns represent the script of the source language. The results are averaged over all target languages. **Bold** (<u>underlined</u>): best (second-best) result.

(TLM). There are four different combinations in total (MLM+SEQ is the training objective of TRANSLICO (Liu et al., 2024a)). Note that we do not consider the variant where the MLM objective is missing since MLM is important to preserve the language modeling capability. We report the average performance of using four different source languages across all target languages for each language group in Table 4 and Table 5.

The lone MLM objective already provides some slight improvement over the base Glot500 model. We hypothesize this is due to the benefit of specializing the model to a small group of languages. When the sequence-level alignment (SEQ) objective is included, the performance is generally further improved compared with the MLM variant, especially for the retrieval task. This is not surprising as the SEQ improves the sequence-level alignment. When instead the token-level alignment (TLM) objective is included, there is a slight improvement in the performance for most tasks compared to the MLM+SEQ objective, except for the retrieval task. We also observe that text classification tasks show the least variation in performance for most source languages. This is especially the case for SIB200, which seems to be the easiest task for the models.

Our whole objective (MLM+SEQ+TLM) generally performs better than the other models by combining sequence and token-level alignment benefits. For Mediterranean-Amharic-Farsi, though MLM+SEQ+TLM performs on par with the MLM+SEQ objective on NER, it achieves better performance on POS. Similarly, our complete training objective outperforms MLM+SEQ on both NER and POS for South+East Asian Languages. When comparing the performance of the full training objective against MLM+TLM, we observe an equal performance across different tasks, except for the retrieval task, where the full training objective noticeably outperforms MLM+TLM. Even though SEQ objective is the most critical for improving the retrieval task, we also observe a performance increase in the retrieval task when TLM is included. This indicates that our post-training alignment method, with TLM, effectively improves both the sequence and token-level alignment.

6 Conclusion

In this work, we propose a transliteration-based post-training method that contains both sequence and token-level objectives to improve the cross-lingual/script alignment of mPLMs and thus boost their zero-shot cross-lingual transfer performance. We apply our post-training method to fine-tune Glo500 on two language groups that share areal features and have extensive lexical overlap. Our extensive experiments using different source languages show that our aligned models consistently outperform the original Glot500 model. In particular, our method enhances the alignment between related languages and, therefore, improves cross-lingual transfer between these languages. We also analyze the impact of different training objectives and show that the sequence and token-level alignment objectives are both critical for achieving the best performance across different tasks.

Limitations

Even though the mPLM is fine-tuned with our method, where the transliterated text is used as an auxiliary input, the mPLM has only seen the Latin transliterations during its pre-training phase. This can limit the performance of the post-trained aligned models, especially for languages with complex scripts. An extension of this work could expand the vocabulary to include the subwords from Latin transliterations as done by Liu et al. (2024b) before fine-tuning or continued-pre-train the model on the transliterated text so that the models can be more effective in modeling the transliterated data.

We are further limited by the transliteration process, which only partially captures the phonetic and semantic information of the original text, especially for languages with significantly different scripts from the Latin script. This leads to a loss of information during the alignment process, which can negatively impact the performance of the post-trained aligned models. Future work could improve the transliteration process to better capture the linguistic properties of the original text.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

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

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020b. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. 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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovickỳ, and Alexander Fraser. 2024. Understanding cross-lingual alignment–a survey. *arXiv preprint arXiv:2404.06228*.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal Romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. 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.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schütze. 2024a. Translico: A contrastive learning framework to address the script barrier in multilingual pretrained language models. *arXiv preprint arXiv:2401.06620*.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schütze. 2024b. Transmi: A framework to create strong baselines from multilingual pretrained language models for transliterated data. *arXiv preprint arXiv:2405.09913*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

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.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Kees Versteegh. 2001. Linguistic contacts between arabic and other languages. *Arabica*, 48(4):470–508.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hans H Wellisch, Richard Foreman, Lee Breuer, and Robert Wilson. 1978. The conversion of scripts, its nature, history, and utilization.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Training Details

For the MLM objective, we use the normal masking probability of 15%. We use the AdamW optimizer (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) with an initial learning rate of $2e-5, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$, weight decay of 0.01, and a linear learning rate scheduler with no warm-up steps. We use FP16 mixed precision training (Micikevicius et al., 2018). Each batch contains sentence pairs in the original and Latin scripts, with a maximum sequence length of 512 tokens. We set the per-device batch size to 16, gradient accumulation steps to 8, and use four NVIDIA A100 80GB GPUs. This leads to an effective batch size of 512. The Mediterranean-Amharic-Farsi model is trained for 2 epochs, while the South+East Asian Languages model is trained for 8 epochs. We use the HuggingFace Transformers library (Wolf et al., 2020) for all experiments. The trainings take around 30 hours for both groups. We store the model checkpoints every 2000 steps.

For the sequence-level contrastive learning objective, we unify all the per-device batch sentence embeddings into a global batch in order to have a larger amount of negative samples. For the token-level alignment objective, differently from the original TLM, we do not reset the positional embeddings for the second sentence in a given pair.

B Downstream Tasks Fine-Tuning

For the downstream tasks that require fine-tuning, we further fine-tune the post-trained aligned models on the training set of a given source language, select the best checkpoint with early stopping based on the f1 score on the source language's validation set, and evaluate the macro F1 score on the test sets of the remaining languages in each group. The fine-tuning results are averaged over five different seeds. All models are fine-tuned on a single GPU. Unless otherwise stated, the same optimizer and scheduler settings are used as in the post-training alignment phase.

SR-B We use the same setup as ImaniGooghari et al. (2023), by calculating the top-10 retrieval accuracy on 500 parallel sentences from the Bible. This task does not require any fine-tuning, and the retrieval is done by computing the cosine similarity between the average 8th-layer contextual embeddings of the source and target sentences.

Taxi1500 A multilingual text classification dataset covering more than 1500 languages formed by classifying 1077 bible verses into six topics (Ma et al., 2023). The learning rate is set to $1e-5$, and we fine-tune the models for 40 epochs with a batch size of 32 and a maximum sequence length of 100 tokens.

SIB200 A multilingual text classification dataset covering more than 200 languages formed by classifying 1004 article sentences into 7 topics (Adelani et al., 2024). We use the same fine-tuning setup as the Taxi1500 task, except for the maximum sequence length, which is set to 160 tokens.

NER We evaluate the models on the WikiAnn named entity recognition dataset (Pan et al., 2017), a multilingual dataset consisting of articles annotated with 7 different tags. We set the learning rate to $2e-5$, the batch size to 32, gradient accumulation steps to 2, and the maximum sequence length to 256 tokens. We fine-tune the models for 5 epochs.

POS We evaluate the models on the Universal Dependencies (UD) v2.11 part-of-speech tagging dataset (de Marneffe et al., 2021), a multilingual dataset consisting of sentences annotated with 17 universal POS tags. We use the same fine-tuning setup as the NER task, except for number of epochs, which is set to 10.

C Vocabulary Analysis

We compare the coverage of the Glot500 vocabulary in the original and transliterated corpora by tokenizing the fine-tuning datasets and counting the number of **unique** tokens. The results are shown in Table 6. As expected, the transliterated text is represented by a smaller part of the vocabulary leading to text being broken down into smaller subwords. This result suggests that the performance of the post-trained aligned models could be further improved by extending the vocabulary of the pre-trained model based on the transliterated corpus.

Language Group	Original Tokens	Transliterated Tokens
Mediterranean-Amharic-Farsi	209K	125K
South+East Asian Languages	120K	88K

Table 6: Number of **unique** tokens in the Glot500 vocabulary covered in the original and transliterated corpora for each language group. A smaller set of unique tokens is used after transliterating the corpora into the common Latin script.

D Full Ablation Results

We provide the full results of the other models trained with different combinations of training objectives for both language groups in Table 7, Table 8, Table 9, Table 10, Table 11, and Table 12.

	SR-B				Taxi1500				SIB200				NER				POS			
	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr
tur_Latn	78.6	89.4	67.2	38.6	65.1	65.1	58.9	36.2	86.8	86.2	84.8	82.9	77.2	73.1	76.5	73.9	73.2	51.1	66.5	68.7
mlt_Latn	76.2	78.6	80.2	47.2	61.3	56.6	59.5	35.6	84.1	85.8	84.8	83.0	75.2	73.3	71.6	74.7	81.6	62.6	77.5	75.8
ell_Grek	57.2	64.2	src	36.2	61.5	60.5	src	32.4	82.9	81.3	src	79.9	73.8	74.1	src	75.0	85.3	57.0	src	63.0
heb_Hebr	25.6	28.0	32.2	src	43.8	43.3	45.8	src	78.6	77.2	79.2	src	50.8	60.5	55.9	src	65.8	70.8	57.7	src
amh_Ethi	62.8	77.8	60.8	38.2	4.8	5.1	12.5	11.6	77.8	79.4	76.6	76.0	43.8	53.2	52.7	49.8	64.5	64.4	60.5	74.7
ara_Arab	-	-	-	-	-	-	-	-	-	-	-	-	58.2	src	59.6	66.6	62.9	src	59.8	77.2
arz_Arab	30.4	42.4	59.4	51.2	39.5	38.4	47.8	37.3	84.4	src	82.9	80.3	58.9	77.6	65.3	67.4	-	-	-	-
ary_Arab	19.4	23.4	34.0	36.0	35.2	37.8	39.2	36.0	82.9	81.7	85.3	81.5	-	-	-	-	-	-	-	-
arb_Arab	12.6	22.4	30.8	33.4	-	-	-	-	83.0	85.0	83.4	79.2	-	-	-	-	-	-	-	-
fas_Arab	89.8	src	79.2	40.8	70.1	src	64.6	33.8	-	-	-	-	55.6	67.8	61.7	58.0	69.6	68.6	60.2	72.5
Average	50.2	53.2	55.4	40.2	47.7	43.8	46.9	31.8	82.6	82.4	82.5	80.4	61.7	68.5	63.3	66.5	71.9	62.4	63.7	72.0

Table 7: Results for the **Mediterranean-Amharic-Farsi** model with the MLM objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.

	SR-B				Taxi1500				SIB200				NER				POS			
	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr
tur_Latn	80.6	90.0	77.2	52.4	68.1	70.5	62.2	40.0	86.8	87.4	86.8	85.9	78.0	73.0	77.2	72.7	73.5	50.0	66.7	68.2
mlt_Latn	83.2	90.2	87.8	62.8	64.1	59.2	59.2	40.3	84.8	85.2	85.4	84.3	74.5	71.3	72.1	75.0	82.4	61.9	76.8	75.9
ell_Grek	67.0	83.0	src	47.4	61.6	61.8	src	38.9	83.1	82.0	src	82.2	73.1	73.6	src	75.1	85.9	57.5	src	68.6
heb_Hebr	32.0	37.4	45.4	src	45.9	46.0	45.9	src	79.8	77.5	80.1	src	50.4	60.6	57.0	src	66.6	71.1	58.8	src
amh_Ethi	57.4	71.0	57.8	46.6	4.8	6.2	14.6	13.7	76.4	79.2	78.8	77.5	41.9	52.3	54.1	49.5	65.1	63.6	63.1	74.1
ara_Arab	-	-	-	-	-	-	-	-	-	-	-	-	56.5	src	61.9	65.8	63.9	src	61.7	78.0
arz_Arab	59.0	79.2	82.6	71.4	39.1	40.3	49.3	43.0	83.6	src	83.9	81.8	59.1	75.6	66.9	67.3	-	-	-	-
ary_Arab	52.6	68.4	68.4	68.0	36.9	38.2	39.5	38.8	82.3	81.5	84.8	82.8	-	-	-	-	-	-	-	-
arb_Arab	47.6	56.0	56.2	52.2	-	-	-	-	82.6	84.0	85.0	82.6	-	-	-	-	-	-	-	-
fas_Arab	87.0	src	83.0	56.2	71.2	src	68.6	38.0	-	-	-	-	55.1	65.2	63.0	57.4	69.9	68.2	60.7	71.3
Average	62.9	71.9	69.8	57.1	49.0	46.0	48.5	36.1	82.4	82.4	83.5	82.4	61.1	67.4	64.6	66.1	72.5	62.1	64.7	72.7

Table 8: Results for the **Mediterranean-Amharic-Farsi** model with the MLM+SEQ objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.

	SR-B				Taxi1500				SIB200				NER				POS			
	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr	Latn	Arab	Grek	Hebr
tur_Latn	78.2	89.4	66.2	40.0	64.0	66.2	59.9	41.2	86.8	86.2	84.9	85.2	76.7	72.6	76.4	73.3	73.3	52.4	66.9	68.9
mlt_Latn	76.8	82.0	82.2	48.0	66.8	61.5	60.2	45.0	86.3	85.1	85.7	83.6	75.8	72.0	72.6	75.6	83.1	64.0	77.3	78.7
ell_Grek	56.6	64.8	src	34.0	65.1	66.5	src	40.0	82.5	81.5	src	81.8	73.6	73.7	src	75.5	85.8	58.3	src	67.4
heb_Hebr	26.2	27.8	33.2	src	47.1	43.9	41.4	src	77.6	77.6	80.7	src	51.9	60.4	56.6	src	67.5	70.8	59.2	src
amh_Ethi	60.0	78.2	57.0	37.4	7.5	8.5	17.0	10.6	77.1	77.6	76.7	76.4	46.2	51.7	54.0	46.6	65.7	63.4	62.5	75.3
ara_Arab	-	-	-	-	-	-	-	-	-	-	-	-	60.1	src	59.3	66.8	65.7	src	60.0	77.9
arz_Arab	30.2	45.4	61.2	54.4	42.8	40.9	49.4	42.1	83.9	src	82.2	82.6	57.6	76.0	66.2	68.9	-	-	-	-
ary_Arab	21.8	27.0	35.6	39.6	38.3	37.8	39.6	39.0	83.6	81.1	82.5	81.5	-	-	-	-	-	-	-	-
arb_Arab	13.8	23.6	31.2	36.2	-	-	-	-	83.1	82.9	84.0	81.6	-	-	-	-	-	-	-	-
fas_Arab	90.0	src	79.6	43.4	72.0	src	63.5	42.2	-	-	-	-	53.5	63.2	63.0	60.7	71.6	69.6	60.9	73.0
Average	50.4	54.7	55.7	41.6	50.4	46.5	47.3	37.2	82.6	81.7	82.4	81.8	61.9	67.1	64.0	66.8	73.2	63.1	64.5	73.5

Table 9: Results for the **Mediterranean-Amharic-Farsi** model with the MLM+TLM objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.

	SR-B				Taxi1500				SIB200				NER				POS		
	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani
tha_Thai	39.6	47.8	46.4	src	65.2	65.1	70.6	src	85.5	84.3	83.8	src	1.9	0.4	8.7	src	47.1	29.2	48.2
kor_Hang	74.4	src	70.0	65.4	63.3	src	62.0	65.4	82.8	src	83.0	83.8	56.2	src	40.7	2.9	52.9	src	38.2
yue_Hani	38.2	55.8	82.4	67.6	68.6	66.4	70.2	69.4	86.0	85.5	src	88.3	16.8	33.7	70.4	11.8	20.3	30.7	78.6
wuu_Hani	-	-	-	-	-	-	-	-	-	-	-	-	32.2	54.2	62.8	1.3	-	-	-
zho_Hani	46.0	49.8	src	44.4	64.8	67.8	src	67.8	-	-	-	-	17.7	33.1	src	11.0	20.6	33.9	src
lzh_Hani	62.2	70.4	74.8	44.6	60.0	58.5	64.4	57.6	-	-	-	-	8.2	33.3	62.0	14.7	6.7	15.2	51.6
lao_Laoo	53.6	78.4	53.8	72.2	67.9	67.4	72.6	69.3	83.0	80.8	81.2	83.9	-	-	-	-	-	-	-
lhu_Latn	6.4	8.8	7.6	10.4	20.2	31.1	21.7	29.6	-	-	-	-	-	-	-	-	-	-	-
mya_Mymr	32.8	43.2	32.8	43.2	57.8	65.2	61.3	62.9	79.6	79.6	80.5	79.6	56.6	68.6	38.7	3.6	-	-	-
bod_Tibt	59.8	80.4	70.2	64.0	-	-	-	-	68.6	71.4	68.5	72.1	39.7	42.0	43.2	5.5	-	-	-
Average	45.8	54.3	54.7	51.4	58.5	60.2	60.4	60.3	80.9	80.3	79.4	81.6	28.7	37.9	46.7	7.2	29.5	27.3	54.2

Table 10: Results for the **South+East Asian Languages** model with the MLM objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.

	SR-B				Taxi1500				SIB200				NER				POS		
	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani
tha_Thai	49.0	75.6	55.6	src	66.4	67.1	71.2	src	86.3	83.6	84.3	src	2.7	1.7	8.3	src	47.8	27.6	48.3
kor_Hang	58.6	src	72.2	70.8	67.3	src	61.6	68.3	82.9	src	83.3	83.4	55.8	src	41.2	3.6	53.1	src	41.8
yue_Hani	60.2	84.4	97.4	81.0	66.2	61.9	67.7	67.2	86.9	83.7	src	88.0	19.3	35.0	70.5	12.2	23.0	37.9	79.4
wuu_Hani	-	-	-	-	-	-	-	-	-	-	-	-	44.2	58.6	63.9	2.0	-	-	-
zho_Hani	45.2	67.4	src	44.4	66.2	64.5	src	68.0	-	-	-	-	19.9	36.2	src	13.2	23.5	41.2	src
lzh_Hani	36.6	37.2	55.8	30.0	65.0	59.8	60.1	58.4	-	-	-	-	9.3	31.4	62.4	18.8	9.6	16.8	52.1
lao_Laoo	55.2	81.6	60.2	72.6	70.9	66.2	72.9	67.9	82.8	80.9	82.2	83.5	-	-	-	-	-	-	-
lhu_Latn	19.4	23.4	18.0	22.4	23.1	36.7	27.6	30.7	-	-	-	-	-	-	-	-	-	-	-
mya_Mymr	40.2	58.0	49.0	50.6	59.5	63.4	58.5	66.3	81.8	79.6	78.0	78.7	57.2	65.6	43.2	2.4	-	-	-
bod_Tibt	58.6	83.4	79.8	77.4	-	-	-	-	70.2	68.9	68.9	68.1	38.9	46.2	36.3	1.7	-	-	-
Average	47.0	63.8	61.0	56.1	60.6	59.9	60.0	61.0	81.8	79.3	79.3	80.3	30.9	39.3	46.5	7.7	31.4	30.9	55.4

Table 11: Results for the **South+East Asian Languages** model with the MLM+SEQ objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.

	SR-B				Taxi1500				SIB200				NER				POS		
	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani	Thai	Latn	Hang	Hani
tha_Thai	37.6	46.8	45.6	src	64.6	63.7	69.4	src	86.5	84.5	85.0	src	1.7	2.2	9.4	src	50.4	30.6	47.4
kor_Hang	72.8	src	75.0	67.6	66.9	src	61.4	66.0	83.6	src	82.8	83.7	56.0	src	43.4	4.0	52.8	src	41.2
yue_Hani	43.0	67.2	88.8	76.6	69.1	62.7	66.3	65.2	87.1	86.0	src	89.1	19.4	34.7	70.6	13.4	27.4	38.9	79.4
wuu_Hani	-	-	-	-	-	-	-	-	-	-	-	-	44.5	56.6	63.0	3.6	-	-	-
zho_Hani	45.0	49.4	src	44.0	68.3	63.2	src	69.2	-	-	-	-	19.8	34.0	src	12.4	30.3	43.6	src
lzh_Hani	57.6	69.6	80.0	48.4	64.0	59.7	61.6	55.4	-	-	-	-	12.1	35.4	61.1	17.2	12.1	15.3	51.3
lao_Laoo	57.2	77.4	54.2	71.8	72.2	68.9	72.6	70.4	81.9	82.5	81.4	83.6	-	-	-	-	-	-	-
lhu_Latn	8.0	10.4	9.4	13.2	22.6	38.4	25.0	31.9	-	-	-	-	-	-	-	-	-	-	-
mya_Mymr	33.0	44.2	39.4	47.4	63.6	67.9	60.1	61.6	80.9	81.2	80.0	80.3	54.3	66.3	40.5	5.4	-	-	-
bod_Tibt	54.6	80.6	70.0	67.2	-	-	-	-	71.2	73.5	68.5	70.7	40.4	44.2	44.9	8.0	-	-	-
Average	45.4	55.7	57.8	54.5	61.4	60.6	59.5	60.0	81.9	81.5	79.5	81.5	31.0	39.0	47.6	9.1	34.6	32.1	54.8

Table 12: Results for the **South+East Asian Languages** model with the MLM+TLM objective. Columns represent the script of the transfer source language (denoted with "src"), while rows represent the target languages. Results are averaged over 5 random seeds. For each target language, we underline the best source transfer score for each task.