Mitigating Language-Level Performance Disparity in mPLMs via Teacher Language Selection and Cross-lingual Self-Distillation

Haozhe Zhao[*1,2], **Zefan Cai**[*1,2], **Shuzheng Si**[*1,2], **Liang Chen**[1],
Yufeng He[1,2], **Kaikai An**[1,2], **Baobao Chang**[†1,3]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Software and Microelectronics, Peking University, China
[3]Collaborative Innovation Center for Language Ability, Xuzhou, 221009, China
{mimazhe55360,zefncai}@gmail.com, sishuzheng@stu.pku.edu.cn
chbb@pku.edu.cn

Abstract

Large-scale multilingual Pretrained Language Models (mPLMs) yield impressive performance on cross-language tasks, yet significant performance disparities exist across different languages within the same mPLM. Previous studies endeavored to narrow these disparities by supervise fine-tuning the mPLMs with multilingual data. However, obtaining labeled multilingual data is time-consuming, and fine-tuning mPLM with limited labeled multilingual data merely encapsulates the knowledge specific to the labeled data. Therefore, we introduce **ALSACE** to leverage the learned knowledge from the well-performing languages to guide underperforming ones within the same mPLM, eliminating the need for additional labeled multilingual data. Experiments show that ALSACE effectively mitigates language-level performance disparity across various mPLMs while showing the competitive performance on different multilingual NLU tasks, ranging from full resource to limited resource settings. The code for our approach is available at https://github.com/pkunlp-icler/ALSACE.

1 Introduction

Recently, Multilingual Pre-trained Language Models (mPLMs) have attracted significant attention (Doddapaneni et al., 2021). These mPLMs, such as mBERT (Devlin et al., 2018) and mT5 (Xue et al., 2020), are pre-trained on extensive amounts of corpus across hundreds of different languages, which enables them to handle multiple languages within a single model and effectively perform cross-lingual tasks (Lewis et al., 2019; Zhang et al., 2020; Stickland et al., 2020; Mutuvi et al., 2020; Brown et al., 2020; Choudhury and Deshpande, 2021).

However, all mPLMs share a key limitation. Due to discrepancies in the quality and quantity of pre-training corpus available for different languages, there is a noticeable performance disparity among different languages for the same mPLM, especially when comparing the performance of high-resource languages to that of low-resource languages. For example, in Cross-lingual Natural Language Inference (XNLI) task (Conneau et al., 2018), high-resource languages such as English can achieve a performance advantage of approximately 15 points compared to low-resource languages like Swahili, even within the same mPLM.

Several works have been proposed to investigate the reason for the performance disparity. Kassner and Schütze (2019); Wallat et al. (2021); Kassner et al. (2021) demonstrate that mPLMs could learn language-specific knowledge from different languages' pre-training corpus, but the imbalance of the corpus for different languages leads to the knowledge disparity for different languages. Therefore, Kassner et al. (2021) suggests the observed language-level performance disparity can be attributed to the disparity of learned different languages knowledge during the pre-training stage. Therefore, Dong et al. (2021); Hu et al. (2021) attempts to narrow the knowledge disparity by involving additional supervised data in different languages to fine-tune the mPLM. However, obtaining such labeled multilingual data is time-consuming and expensive. Moreover, these labeled data mostly come from limited tasks and domains, which makes it hard to compensate for the large knowledge disparity during the pre-training stage, restricting the generalization performance of the low-resource languages on downstream tasks.

To utilize the different knowledge across different languages within the same mPLM and mitigate the need for the labeled data, we introduce te**A**cher **L**anguage **S**election **A**nd **C**ross-lingual s**E**lf-distillation (**ALSACE**), which leverages the knowledge from the selected teacher languages to reduce the performance disparity among the languages. Specifically, ALSACE mainly consists of

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2893–2907

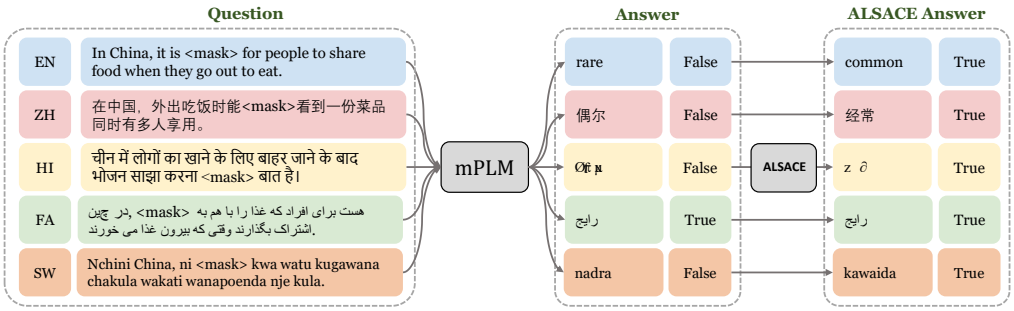

Figure 1: ALSACE can reduce language-level performance disparity via mitigating knowledge disparity across languages on GeoMLAMA benchmark (Yin et al., 2022).

two stages: Teacher Language Selection and Cross-Lingual Self-Distillation. For teacher language selection, the motivation is that high-resource languages may not be ideal for probing knowledge to supervise the other languages. For instance, although Persian is a relatively low-resource language, it may provide more precise answers for Kenya's cultural queries than English due to the closer linguistic proximity (Yin et al., 2022) between Persian and Swahili. Different from simply using the knowledge from high-resource languages (e.g., English) to improve the performance of low-resource languages (e.g., Swahili), we introduce Teacher Language Selection to choose reliable teacher languages for a specific task to supervise the student languages. Specifically, we employ a majority voting strategy to generate pseudo-labels derived from the consensus of the mPLMs' predictions across different languages in the given multilingual corpus. Then, we utilize the average confidence score of the different languages on the generated pseudo labels as the indicator to select the teacher languages automatically. As a result, we can select adaptive teachers for different tasks using the unlabeled sentences in the corpus. We further propose Cross-Lingual Self-Distillation to leverage the knowledge from each selected teacher language to supervise other languages, reducing the performance disparity. We further propose cross-lingual self-distillation to leverage the knowledge from each selected teachers languages to supervise other languages, reducing the performance disparity. It employs a consistency loss that encourages closer alignment between the model output distributions of each reliable teacher language and other languages. In this way, mPLMs can effectively mitigate the language-level performance disparity without relying on the supervised multilingual data.

Experiments show ALSACE consistently mitigates language-level performance disparity in various mPLMs and show the competitive performance on different multilingual benchmarks, including XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019) and XCOPA (Ponti et al., 2020). We also conduct knowledge probing experiments on the GeoMLAMA (Yin et al., 2022) as shown in Figure 1, demonstrating that ALSACE effectively mitigates language-level performance disparity by addressing knowledge disparity. Moreover, our experiments show that ALSACE improves performance not only in low-resource languages but also in high-resource languages. This finding illustrates that ALSACE enables effective knowledge transfer between different languages instead of only transferring knowledge from high-resource to low-resource languages. Further analysis shows that ALSACE can transfer both general knowledge across different languages and language-specific knowledge, i.e., some specific knowledge locally shared by people speaking the specific language, which is only present in the corpus of some specific languages.

2 Related Work

Knowledge Disparity Leads to Language-Level Performance Disparity in mPLMs. The mPLMs have shown strong capabilities in many NLP tasks including Natural Language Generation (NLG) (Si et al., 2022a, 2024; Zhao et al., 2023; Cai et al., 2023; Li et al., 2024; Liu et al., 2023b) and natural language understanding (NLU) (Si et al., 2022b, 2023; Liu et al., 2023a; An et al., 2023; Hu et al., 2023). However, there is a noticeable performance disparity across different languages in the same mPLM. Several works are proposed to investigate the reason of language-level performance disparity in mPLMs. Wallat et al. (2021); Kassner et al.

(2021) demonstrate that mPLMs could learn different knowledge from different languages data in the pre-training corpus, but imbalanced corpus might lead to knowledge disparity for different languages. Kassner et al. (2021) suggests that the performance disparities across different languages could be attributed to the imbalanced knowledge distribution of these languages acquired during the pre-training phase. Yin et al. (2022) further observe that different languages within a single mPLM can retain distinct knowledge that is locally shared by the people speaking the specific language. Therefore, we attempt to address language-level performance disparity from the knowledge disparity perspective.

Mitigating Language-Level Performance Disparity in mPLMs. Previous studies have utilized cross-lingual knowledge to mitigate the language-level performance disparity. He et al. (2021) employ lightweight adapters on the mPLMs to mitigate forgetting issues. InfoXLM (Chi et al., 2021a) designs a new pre-training task with 42GB parallel data to align the representation of multiple languages. XLE (Chi et al., 2022) pre-trains mPLMs with a generator and discriminator structure on 142B tokens. These methods attempt to incorporate multilingual resources to mitigate performance disparity. However, obtaining multilingual data can be time-consuming and restricts model performance on low-resource languages. Thus, Yang et al. (2022); Nguyen and Tuan (2021) attempt to enhance mPLMs by distilling knowledge from well-learned monolingual teachers. Qi et al. (2022) learn from different cross-lingual templates using consistency loss to enforce correspondence representation among languages. Different from distilling knowledge from other monolingual models, we aim to reduce the language-level performance disparity within mPLMs.

3 Method

3.1 Teacher Language Selection

To mitigate the language-level performance disparity within mPLMs, we utilize knowledge from the appropriate teacher language to supervise other languages. An intuitive idea is to transfer the knowledge from high-resourced to low-resourced languages to mitigate the disparity. However, due to the different linguistic proximity between different languages, the high-resource languages may not be ideal teachers for transferring knowledge to other languages in the specific task. For example, low-resourced Persian may provide more accurate responses to Kenya's cultural queries compared to high-resource English, which makes it a better teacher language for Swahili than English. Therefore, the proposed Teacher Language Selection aims to choose reliable teacher languages for a specific task to guide the student languages.

Considering the given corpus D for the specific multilingual task (e.g., Cross-lingual Natural Language Inference) that spans over T languages, we aim to utilize the proposed Teacher Language Selection to identify the teacher languages to mitigate language-level performance disparity efficiently. Precisely, we first fine-tune the mPLMs with an English training set D_{en} of the given task to obtain a better initialization. We secondly utilize the mPLMs to generate the prediction $\hat{y}_{t,i}$ of the given instance x_i from corpus D in language $t \in T$. Then, we employ a majority vote strategy on the predictions of different languages to generate the pseudo label y_i of the instance $x_i \in X$, as follows:

$$\begin{aligned} \hat{y}_{t,i} &= \underset{y \in Y}{\operatorname{argmax}} P(y \mid x_{t,i}) \\ y_i &= \underset{k}{\operatorname{argmax}} \sum_{t \in T} \mathbb{I}(\hat{y}_{t,i} = k) \end{aligned} \tag{1}$$

where $P(y \mid x_{t,i})$ denotes the predicted probability of the given mPLM on instance $x_{t,i}$ in language t. $\mathbb{I}$ is the indicator function, while k signifies the set of all possible results for the given task. The generated pseudo-labels reflect the collective understanding of the provided instance across various languages. Thus, it reduces the risk of incorrect pseudo-labeling compared to relying solely on the prediction from a single language (even a high-resource language like English).

We further employ the pseudo-labels to compute the average confidence score s_t for each language, which allows us to assess the capabilities of different languages in the mPLM. The average confidence score s_t indicates the level of agreement between each language and the common understanding of the mPLMs, i.e., languages with a higher average confidence score are more likely to make accurate predictions for a given instance. Ultimately, we normalize the confidence score and use the normalized score $\hat{s}_t$ to evaluate which lan-

guages demonstrate superior performance:

$$s_t = \frac{1}{|X|}\sum_{x_{t,i}}^{X} P(y_{t,i}|x_{t,i})$$
$$\hat{s}_t = \frac{e^{s_t}}{\sum_j^T e^{s_j}}, \quad t \in T \tag{2}$$

where the T refers to the collection of all languages involved in the given multilingual task. We set the threshold θ to be the average value of the normalized score $\hat{s}_t$ to select the teacher languages $T_{teacher}$ and student languages $S_{student}$, as follows:

$$T_{teacher} = \{t|t \in T, \hat{s}_t \geq \theta\}$$
$$T_{student} = \{t|t \in T, \hat{s}_t < \theta\} \tag{3}$$

In this way, we can automatically select appropriate teacher languages for the different multilingual tasks to mitigate language-level performance disparity efficiently. Moreover, we do not need any labeled multilingual data to improve the cross-lingual transfer ability of mPLMs (Chi et al., 2022, 2021a).

3.2 Cross-Lingual Self-Distillation

Having selected the appropriate teacher languages for the given multilingual task, we further introduce Cross-Lingual Self-Distillation to leverage the knowledge from each selected teacher language to supervise other languages. Specifically, we construct a parallel multilingual pair set $\hat{X}$ that consists of parallel sentence pairs between each two languages. To reduce the disturbance caused by student languages, we exclusively employ parallel pairs of teacher-student and teacher-teacher languages as potential candidates for self-distillation. Therefore, the instance pair $\hat{X}$ can be defined as:

$$\hat{X} = \{\ (x_{t1,i}, x_{t2,i}) \mid t_1 \in T,\ t_2 \in T_{teacher},\ x_i \in X\ \} \tag{4}$$

where $T_{teacher}$ is the selected teacher languages. We filter out student-student language pairs to prevent student languages from learning from each other.

For the selected candidate instance pairs, we use Kullback-Leibler divergence as a consistency loss to encourage closer alignment between the prediction distributions of the reliable teacher language and the target language. In this way, mPLMs can effectively transfer and distill the knowledge from the teacher language to the target language, mitigating the language-level performance disparity. The final consistency loss $\mathcal{L}$ can be formulated as follows:

$$\mathcal{L} = \frac{1}{|\hat{X}|}\sum_{\hat{x}_1,\hat{x}_2}^{\hat{X}} \mathrm{KL}(\mathrm{P}(\hat{x}_1)||\mathrm{P}(\hat{x}_2)) \tag{5}$$

where KL(P||Q) is the Kullback-Leibler divergence function. $P(\hat{x}_1)$ and $P(\hat{x}_2)$ are the prediction distributions of the given mPLM for the inputs $\hat{x}_1$ and $\hat{x}_2$ in different languages, respectively.

4 Experiment

4.1 Experimental Details

Datasets. As shown in Table 1, our experiments are conducted on various multilingual benchmarks: XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020) and GeoMLAMA (Yin et al., 2022).

Task	Dataset	Lang.	Metric
Natural Language Inference	XNLI	15	Acc.
Commonsense Reasoning	XCOPA	10	Acc.
Paraphrase Identification	PAWS-X	7	Acc.
Commonsense Probing	GeoMLAMA	5	Acc.

Table 1: The tasks involved in experiments.

Experimental Settings. We follow the cross-lingual transfer setting as Lauscher et al. (2020), first fine-tuning the model with an English training set and directly evaluating the model on multilingual test sets. We apply ALSACE to the fine-tuned model using unlabeled multilingual inputs X from T languages in order to address the language-level performance disparity across those languages. Specifically, We firstly use data generation methods, Supergen (Meng et al., 2022), which employ a language model to automatically generate text based on label-descriptive prompts, producing monolingual unlabeled data. Next, we use machine translation[1] to translate generated monolingual data and create unlabeled parallel multilingual pairs. By combining the data generation method and machine translation system, we establish an automated pipeline for generating unlabeled parallel corpora with minimal cost.

Baselines. We take the XLM-Align (Chi et al., 2021b), XLMR-adapter$_{256}$ (He et al., 2021), InfoXLM (Chi et al., 2021a), VECO (Luo et al., 2021), ERNIE-M (Ouyang et al., 2021) and XLE (Chi et al., 2022) as baselines.

Details can be found in Appendix A.1 and A.2.

[1]The translation API from http://api.fanyi.baidu.com/ is utilized for generating multilingual parallel data.

Method	Params	Perf.	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	**avg**
XLM-R-base	225M		84.23	77.39	78.20	76.45	75.97	77.80	75.35	73.27	71.84	74.93	71.88	74.23	69.22	64.55	65.77	74.07
		Δ ↓	\	6.84	6.03	7.78	8.26	6.43	8.88	10.96	12.39	9.30	12.35	10.00	15.01	19.68	18.46	10.88 (-)
XLM-Align	225M		86.70	80.60	81.00	78.80	77.40	78.80	77.40	75.20	73.90	76.90	73.80	77.00	71.90	67.10	66.60	76.30
		Δ ↓	\	**6.10**	**5.70**	7.90	9.30	7.90	9.30	11.50	12.80	9.80	12.90	9.70	14.80	19.60	20.10	11.24 (+0.36)
ALSACE -base	225M		84.11	77.80	78.30	77.50	76.51	78.28	76.01	74.19	72.12	75.50	72.81	74.90	70.12	65.55	66.37	74.67
		Δ ↓	\	6.31	5.81	**6.61**	**7.60**	**5.83**	**8.10**	**9.92**	**11.99**	**8.61**	**11.30**	**9.21**	**13.99**	**18.56**	**17.74**	**10.11 (-0.77)**
XLM-R-large	550M		86.45	80.90	81.84	81.22	79.36	80.74	78.78	77.23	77.03	77.82	75.53	77.82	74.55	69.62	70.86	77.98
		Δ ↓	\	5.45	4.51	5.13	6.99	5.61	7.57	9.12	9.32	8.53	10.82	8.53	11.80	16.73	15.49	8.97 (-)
XLM-R-adapter	567M		89.22	83.27	84.69	83.47	82.39	83.59	79.74	78.80	78.62	79.32	77.84	79.34	76.42	72.21	72.27	80.08
		Δ ↓	\	5.95	4.53	5.75	6.83	5.63	9.48	10.42	10.60	9.90	11.38	9.88	12.80	17.01	16.95	9.79(+0.82)
Info-XLM-large	550M		89.70	84.50	85.50	84.10	83.40	84.20	81.30	80.90	80.40	80.80	78.90	80.90	77.90	74.80	73.70	81.40
		Δ ↓	\	5.20	4.20	5.60	6.30	5.50	8.40	8.80	9.30	8.90	10.80	8.80	11.80	14.90	16.00	8.89 (-0.08)
VECO-large	550M		88.20	82.80	84.20	82.90	81.20	83.10	80.30	78.40	79.20	80.40	77.00	79.10	76.20	74.30	71.30	79.90
		Δ ↓	\	5.40	4.00	5.30	7.00	5.10	7.90	9.80	9.00	7.80	11.20	9.10	12.00	13.90	16.90	8.88(-0.09)
ERNIE-M-large	550M		89.30	85.10	85.70	84.40	83.70	84.50	82.00	81.20	81.20	81.90	79.20	81.00	78.60	76.20	75.40	82.00
		Δ ↓	\	4.20	3.60	4.90	5.60	4.80	7.30	8.10	8.10	7.40	10.10	8.30	10.70	**13.10**	13.90	7.86 (-1.11)
ALSACE -large	550M		86.65	82.61	83.21	82.16	81.34	83.09	80.98	79.50	79.60	79.98	78.18	79.74	77.13	72.71	73.58	80.03
		Δ ↓	\	**4.04**	**3.44**	**4.49**	**5.31**	**3.56**	**5.67**	**7.15**	**7.05**	**6.67**	**8.47**	**6.91**	**9.52**	13.94	**13.07**	**7.09 (-1.88)**
mT5-large	1.2B		88.42	82.44	83.49	81.68	81.14	81.96	79.90	77.33	76.87	78.52	75.31	77.74	75.31	72.63	70.88	78.91
		Δ ↓	\	5.98	4.93	6.74	7.28	6.46	8.52	11.09	11.55	9.90	13.11	10.68	13.11	15.79	17.54	10.19 (-)
XLE-large	840M		89.40	84.70	85.50	84.40	83.50	84.10	81.90	81.30	80.70	81.20	79.20	81.50	76.50	74.10	72.40	81.30
		Δ ↓	\	**4.70**	3.90	**5.00**	5.90	5.30	7.50	**8.10**	**8.70**	**8.20**	**10.20**	7.90	12.90	15.30	17.00	8.61 (-1.58)
ALSACE -mT5	1.2B		88.60	83.69	84.79	83.17	82.91	83.91	81.80	79.54	78.84	80.20	77.90	80.92	77.25	75.17	73.13	80.79
		Δ ↓	\	4.91	**3.81**	5.43	**5.69**	**4.69**	**6.80**	9.06	9.76	8.40	10.70	**7.68**	**11.35**	**13.43**	**15.47**	**8.37 (-1.82)**

Table 2: Main result of XLM-R-base, XLM-R-large, and mT5-large on XNLI dataset evaluated in accuracy. Δ represents the cross-lingual transfer gaps (Chi et al., 2021a), i.e., performance drop between English and other languages in zero-shot transfer. A smaller gap indicates better cross-lingual transferability. ALSACE achieves competitive results compared to the state-of-the-art methods and enhances the performance of most languages across all three mPLMs, simultaneously reducing the language-level performance disparity amongst all the languages.

4.2 Main Results

Overall Performance. The results presented in Table 2 demonstrate that ALSACE achieves the lowest cross-lingual transfer gaps across different baselines on XNLI for various mPLMs. ALSACE yields an improvement of up to 0.6 points, 2.05 points, and 1.88 points, respectively, in average accuracy compared with XLM-R-base, XLM-R-large, and mT5-large baselines. Importantly, we achieve competitive performance with state-of-the-art methods across different mPLMs while improving the cross-lingual transferability of mPLMs without introducing any extra information.

For example, InfoXLM (Chi et al., 2021a), which is also based on XLM-R, uses 42GB of multilingual parallel data for pretraining. In contrast, ALSACE depends solely on a small volume of unlabeled parallel data (500-shot), which can be automatically generated with minimal effort and and exhibits superior cross-lingual transferability compared to other baselines. While we also utilize parallel data to enhance cross-lingual transferability, our motivation diverges: Instead of aligning multilingual representations through parallel data, our goal is to leverage the knowledge from teacher languages within mPLMs to supervise others. The 500-shot unlabeled parallel data in ALSACE are exclusively used to distill the knowledge of other languages in mPLMs. As a result, Table 2 shows performance enhancement and cross-lingual transfer gap reduction for most languages across different models. In comparison to state-of-the-art methods, ALSACE does not mandate an extensive pre-training process or a large number of parallel corpora while achieving competitive performance and minimizing the cross-lingual transfer gaps.

Mitigating Languages-Level Performance Disparity. ALSACE effectively mitigates the language-level performance disparity of mPLMs and shows consistent improvements across different mPLMs in both high-resource and low-resource languages. Specifically, not only do the student languages achieve higher-than-average improvements, but teacher languages also benefit from the guidance of their peers. Through self-distillation, ALSACE facilitates cross-language knowledge transferring among both teacher and student languages. It also enables teacher languages to learn from each other. Even high-resource languages like French and Spanish have shown improvement across various mPLMs, which further supports this claim. Notably, low-resource languages such as Swahili and Urdu experience substantial gains with ALSACE , achieving improvements of 2.7 points and 2.4 points, respectively. These gains are particularly significant considering the relatively limited knowledge stored in multilingual

Method	avg(S) ↑	Δ(S) ↓	avg(T)↑	Δ(T) ↓	avg(A)↑	Δ(A) ↓
Excluding Weak in Stu.	76.70	10.08	82.56	4.93	79.44	7.35
Excluding Weak in Tea.	76.92	9.73	82.50	4.84	79.52	7.13
Random selection	76.93	9.87	82.67	4.83	79.61	7.20
No Selection	77.05	9.83	82.69	4.90	79.69	7.20
Scale-Based Selection	77.13	9.80	82.73	4.89	79.75	7.18
ALSACE	**77.55**	**9.10**	**82.86**	**4.42**	**80.03**	**6.62**

Table 3: Ablation Study of the Teacher Language Selection. Δ represents the cross-lingual transfer gaps, i.e., performance drop between English and other languages in zero-shot transfer. A smaller gap indicates better cross-lingual transferability. We report the average performance and cross-lingual transfer gaps of the student languages(**S**), teacher languages(**T**), and all languages(**A**), respectively.

Method	avg(S) ↑	Δ(S) ↓	avg(T)↑	Δ(T) ↓	avg(A)↑	Δ(A) ↓
XLM-R-base	70.71	13.52	77.91	7.37	74.07	10.88
E. Self-Train.	70.94	13.15	78.16	6.92	74.31	10.48
F. Self-Train.	71.10	13.03	78.27	6.84	74.45	10.37
ALSACE	**71.44**	**12.67**	**78.35**	**6.72**	**74.67**	**10.12**
XLM-R-large	75.06	11.39	81.33	5.98	77.98	9.07
E. Self-Train.	75.82	10.95	81.74	5.87	78.58	8.77
F. Self-Train.	75.89	10.92	82.10	5.50	78.79	8.60
ALSACE	**77.55**	**9.10**	**82.86**	**4.42**	**80.03**	**7.09**
mT5-large	75.57	12.85	82.72	6.65	78.91	10.19
E. Self-Train.	76.55	11.95	83.21	6.18	79.66	9.48
F. Self-Train.	76.81	11.83	83.32	6.21	79.85	9.42
ALSACE	**77.87**	**10.73**	**84.12**	**5.22**	**80.79**	**8.37**

Table 4: Comparison of self-distillation baselines with ALSACE . Δ represents the cross-lingual transfer gaps. **S**, **T**, and **A** stand for the set of student languages, teacher languages, and all languages, respectively.

pretrained language models (mPLMs) for these languages compared to other languages.

Compared with other baselines, ALSACE effectively reduces language-level performance disparities in mPLMs across various languages and minimizes the cross-lingual transfer gap. While some methods have enhanced overall performance, they have exacerbated the performance discrepancies between languages. They incorporated additional knowledge from the extensive parallel multilingual corpora into mPLMs. However, knowledge disparities persist and may even worsen, leading to increased cross-lingual transfer gaps. We also perform ALSACE across different tasks, such as PAWS-X and XCOPA. The result in Table 6 and Table 9 shows that ALSACE reduces the languages-level performance disparity of mPLMs.

4.3 Ablation Study

Ablation Study on Teacher Language Selection

To evaluate the effectiveness of Teacher Language Selection, we conduct an ablation study using XLM-R-large as backbone. We reported average performance and cross-lingual transfer gaps of different language groups in Table 3. It provides strong evidence for the effectiveness of our method.

Generally, the implementation of Teacher Language Selection in ALSACE significantly reduces the cross-lingual transfer gaps while improving performance across all languages, particularly for the student languages. It validates that, despite the efficacy of self-distillation, selecting adaptive teacher languages is crucial for boosting overall performance. With Teacher Language Selection, student languages achieve above-average improvements in both performance and cross-lingual transferability.

Specifically, when comparing ALSACE with other baselines, besides a performance improvement, there is a substantial reduction in the cross-lingual transfer gaps for all languages, particularly for student languages. ALSACE reduces the cross-lingual transfer gaps for student languages, ranging from 0.70 to 0.98 points and between 0.47 to 0.51 points for teacher languages.

Furthermore, excluding the teacher language selection diminishes the performance of student languages, limiting their ability to benefit from self-distillation. This results in an average performance decrease of 0.34 points and an increase of 0.73 points in cross-lingual transfer gaps for student lan-

Method	de	fr	ar	ru	zh	sw	en	vi	el	tr	bg	th	hi	ur	es	avg.
XLM-R-large.	69.84	69.52	65.18	69.00	66.43	61.65	71.81	67.75	68.76	66.27	70.28	63.17	63.94	61.85	69.24	66.98
E. Self-Train.	69.06	68.48	65.67	68.32	65.85	60.10	71.86	67.96	68.40	65.39	69.40	62.46	64.35	61.96	69.02	66.55
F. Self-Train.	69.04	68.46	65.69	68.30	65.83	60.10	71.82	67.98	68.38	65.37	69.40	62.50	64.33	61.94	69.00	66.54
ALSACE	**70.56**	**69.88**	**67.63**	**70.00**	**67.99**	**63.73**	**72.17**	**68.96**	**69.60**	**68.03**	**70.68**	**64.06**	**64.86**	**63.53**	**71.04**	**68.18**

Table 5: ALSACE performance on XLM-R-large in XNLI dataset under Limited-Resource Scenario. The metric in this table is accuracy. For each setting, we report the median scores among 5 runs.

Method	de	fr	zh	en	ko	ja	es	avg.
XLM-R-large.	82.35	82.75	77.05	85.60	73.24	72.70	83.40	79.58
E. Self-Train.	82.35	82.50	**78.05**	85.95	72.44	72.80	**83.70**	79.68
F. Self-Train.	82.35	**82.85**	77.20	86.15	**73.74**	72.85	83.40	79.79
ALSACE	**82.40**	82.75	77.70	**86.25**	73.29	**72.95**	83.55	**79.84**

Table 6: ALSACE performance on XLM-R-large in PAWS-X dataset under Limited-Resource Scenario. The metric in this table is accuracy. For each setting, we report the median scores among 5 runs.

guages. ALSACE still outperforms the random selection by 0.42 points in performance and reduces cross-lingual transfer gaps for student languages by 0.77 points. These comparisons underscore the importance of selecting adaptive teacher languages.

Additionally, we remove some languages from distillation. First, we removed languages that exhibited weak performance from student languages. As expected, without the guidance of teacher languages, the performance of student languages remained poorly, with an observed increase in cross-lingual transfer gaps by 0.98 points. Subsequently, excluding languages with weak performance from teachers also led to a decrease in performance for both teachers and students by 0.34 and 0.63 points, respectively. It underscores our hypothesis that the underperforming languages can serve as effective guidance for other languages due to closer linguistic proximity between languages. Further details of the ablation study can be found in Appendix A.6.

Ablation Study on Cross-lingual Self-Distillation To further investigate the source of performance increase and validate the effectiveness of the self-distillation, we conducted additional experiments with self-distillation methods as baselines with the following two settings:

English-Only Self-Training (Schick and Schütze, 2020): We utilize the model fine-tuned on an English training set to produce pseudo-labels for the unlabeled English data used in cross-lingual self-distillation. Then, we choose the top 50% of data with high confidence to fine-tune the model.

Full-Language Self-Training: We generate pseudo-labels for translated multilingual data in all languages and select the top 50% of multilingual data with high confidence to fine-tune the model.

We apply these two methods on mT5 (Xue et al., 2020) and XLM-R (Conneau et al., 2019) as baselines. As shown in Table 4, ALSACE outperforms all the self-distillation baselines on XNLI while improving the cross-lingual transferability of mPLMs, especially for the student languages. It validates our method and indicates that ALSACE's improved performance stems from our self-distillation rather than from the incorporation of multilingual data. We also compared our method with other state-of-the-art self-distillation methods in Appendix A.3.2.

4.4 Limited Resource Evaluation

In scenarios with limited resources, where acquiring training data is extremely difficult (even for English), mitigating language-level performance disparities in mPLMs can be more challenging and crucial. Therefore, to further evaluate the effectiveness of ALSACE , we performed experiments on both XNLI and PAWS-X datasets in such scenarios. Specifically, to simulate a limited resource scenario for XNLI, we fine-tune the mPLMs on 128-shot English labeled examples as the baseline. Similarly, for PAWS-X, we fine-tune the mPLMs on 512-shot English labeled examples. Further details can be found in Appendix A.1.

To minimize the impact of the unlabeled multilingual parallel data used in ALSACE , and thoroughly investigate the efficacy of self-distillation in ALSACE in limited resource situations, we also introduce two additional baselines: *English-Only Self-Training*(E. Self-Train) and *Full-Language Self-Training*(F. Self-Train). The results in Table 5

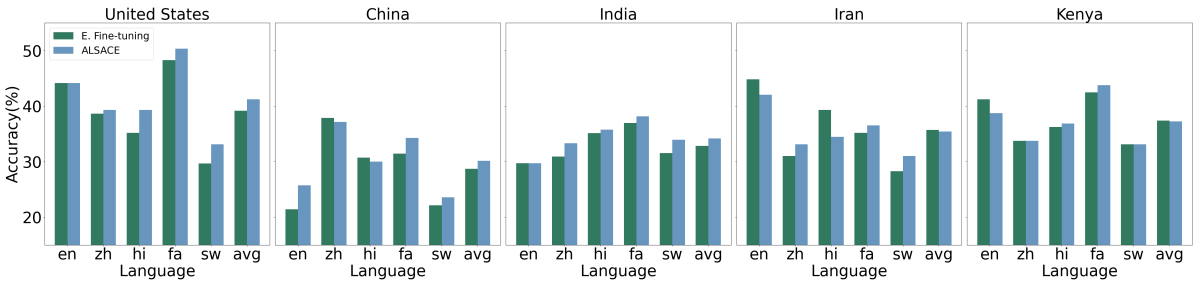

Figure 2: Result of ALSACE on XLM-R-large in GeoMLAMA dataset. The result shows that ALSACE utilizes the teacher languages to guide other languages and generally improves their languages-specific knowledge.

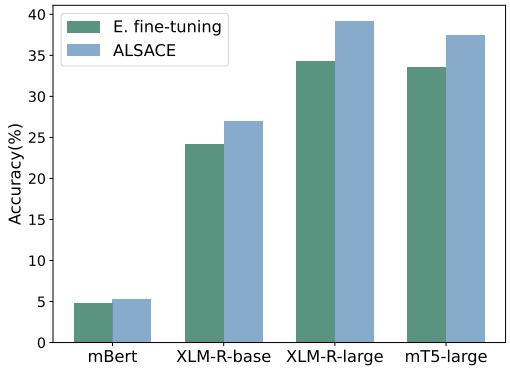

Figure 3: Accurately Answered Questions across All Languages in XNLI Baseline.

and Table 6 despite that ALSACE consistently improve the performance of all languages even when the training data is minimal. It underscores that ALSACE improves model performance not by relying on the parallel corpora but by leveraging the knowledge of teacher languages gained from the mPLM pre-training stage, hence proving its robustness and efficiency in limited-resource settings.

4.5 Analysis

The knowledge stored within mPLMs can be categorized into language-agnostic knowledge related to general tasks such as XNLI, which are based on logic and conceptual understanding, and language-specific knowledge related to specific linguistic and cultural factors. In order to evaluate the ALSACE 's ability to alleviate performance disparity by reducing knowledge disparity and thereby improving overall performance, we conducted knowledge probing in GeoMLAMA to evaluate the changes in language-specific knowledge of mPLMs. We use the accuracy of question answers grouped according to countries and languages to measure the knowledge of mPLMs.

We examined the changes in language-specific knowledge gains before and after applying ALSACE as shown in Figure 2. Results show that ALSACE improves the performance of mPLM on knowledge probing tasks over various languages. More details can be found in Table 12 in Appendix.

Notably, as shown in Figure 1, after applying Cross-lingual Self-Distillation, the specific knowledge of teacher languages can be transferred to other languages. It can be found out that under the guidance of teacher languages, other languages answer the geo-specific question correctly. For instance, as shown in the first sub-figure in Figure 2, English leverages its US-specific knowledge for other languages, leading to overall improvements for those respective languages. Similar results are observed in other sub-figures. This result strongly suggests that mPLMs capture far more knowledge than people previously believed, and language-specific knowledge remains a treasure for better alignment.

Furthermore, we explore whether ALSACE successfully enhances language-agnostic knowledge over languages. Therefore, as demonstrated in Figure 3, we evaluate the numbers of the accurately answered questions on the XNLI benchmark. This improvement demonstrates that the language-agnostic knowledge across different languages in mPLMs can mutually learn from each other. Our method reinforces the shared knowledge among the languages by bridging the knowledge disparity. As a result, we ensure that the efficacy of our method relies on alleviating the knowledge disparities across languages, including language-agnostic and language-special knowledge.

5 Conclusion

In this paper, we present ALSACE , a simple yet effective method to address the language-level per-

formance disparity in mPLMs. ALSACE mainly consists of two stages: Teacher Language Selection and Cross-Lingual Self-Distillation. ALSACE leverages the knowledge learned from the teacher languages to guide other languages and further improves the overall performance and cross-lingual transferability of mPLMs. Experiments show that ALSACE effectively mitigates language-level performance disparity and shows competitive performance on various multilingual datasets. In addition, we further analyze each part of the ALSACE to show the strengths of our proposed model. Overall, ALSACE is a promising approach to mitigating language-level performance disparity of mPLMs by utilizing self-distillation to reduce the performance disparity.

Limitation

Our work has three limitations:

1) We conduct experiments on a limited number of languages compared to the total number supported by mPLMs. Additionally, we only test other methods on the base and large model sizes of mT5 and XLM-R models. Therefore, in future work, we plan to extend our research to more languages and different mPLMs in different model sizes.

2) In the grand scheme of things, the languages we evaluate are relatively high-resource compared to some extremely low-resource languages such as Kaixana and Ainu. Improving our method on these extremely low-resource languages will be more exciting and meaningful. We plan to explore even more data-scarce settings in future work.

3) We use the cross-lingual transfer gap to measure mPLMs' cross-lingual transferability, aligning with prevailing research. However,if we reservedly enhances the performance of non-English languages while improving English greatly, the model's transfer gap could still be high despite the improvement in all languages. Hence, we advocate for the development of the metric that can better reflect the performance equity and utility in multilingual models.

Acknowledgements

We extend our heartfelt gratitude to the anonymous reviewers whose dedication and insightful feedback have significantly enhanced the caliber of this paper. Their constructive critiques and valuable suggestions were instrumental in refining our work. Additionally, we are deeply appreciative of the Program Chairs and Area Chairs for their meticulous handling of our submission and for their comprehensive and invaluable feedback. Their guidance has been pivotal in elevating the quality of our research. This work is supported by the National Science Foundation of China under Grant No.61936012 and 61876004.

References

Kaikai An, Ce Zheng, Bofei Gao, Haozhe Zhao, and Baobao Chang. 2023. Coarse-to-fine dual encoders are better frame identification learners. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13455–13466, Singapore. Association for Computational Linguistics.

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.

Yucheng Cai, Wentao Ma, Yuchuan Wu, Shuzheng Si, Yuan Shao, Zhijian Ou, and Yongbin Li. 2023. Unipcm: Universal pre-trained conversation model with task-aware automatic prompt.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment.

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.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12710–12718.

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Xin Dong, Yaxin Zhu, Zuohui Fu, Dongkuan Xu, and Gerard de Melo. 2021. Data augmentation with adversarial training for cross-lingual NLI. 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 5158–5167, Online. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 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 2208–2222, Online. Association for Computational Linguistics.

Helan Hu, Shuzheng Si, Haozhe Zhao, Shuang Zeng, Kaikai An, Zefan Cai, and Baobao Chang. 2023. Distantly-supervised named entity recognition with uncertainty-aware teacher learning and student-student collaborative learning.

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.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024. One shot learning as instruction data prospector for large language models.

Xinyu Liu, Yan Ding, Kaikai An, Chunyang Xiao, Pranava Madhyastha, Tong Xiao, and Jingbo Zhu. 2023a. Towards robust aspect-based sentiment analysis through non-counterfactual augmentations. *arXiv preprint arXiv:2306.13971*.

Yuliang Liu, Xiangru Tang, Zefan Cai, Junjie Lu, Yichi Zhang, Yanjun Shao, Zexuan Deng, Helan Hu, Zengxian Yang, Kaikai An, Ruijun Huang, Shuzheng Si, Sheng Chen, Haozhe Zhao, Zhengliang Li, Liang Chen, Yiming Zong, Yan Wang, Tianyu Liu, Zhiwei Jiang, Baobao Chang, Yujia Qin, Wangchunshu Zhou, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2023b. Ml-bench: Large language models leverage open-source libraries for machine learning tasks.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. Veco: Variable and flexible cross-lingual pre-training for language understanding and generation.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*.

Stephen Mutuvi, Emanuela Boroş, Antoine Doucet, Adam Jatowt, Gaël Lejeune, and Moses Odeo. 2020. Multilingual epidemiological text classification: a comparative study. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6172–6183.

Thong Nguyen and Luu Anh Tuan. 2021. Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaxing Lin, and Baobao Chang. 2023. Santa: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36.

Shuzheng Si, Shuang Zeng, and Baobao Chang. 2022a. Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4839–4847, Seattle, United States. Association for Computational Linguistics.

Shuzheng Si, Shuang Zeng, Jiaxing Lin, and Baobao Chang. 2022b. SCL-RAI: Span-based contrastive learning with retrieval augmented inference for unlabeled entity problem in NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2313–2318, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2020. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. *arXiv preprint arXiv:2004.14911*.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2021. Bertnesia: Investigating the capture and forgetting of knowledge in bert. *arXiv preprint arXiv:2106.02902*.

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

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Ziqing Yang, Yiming Cui, Zhigang Chen, and Shijin Wang. 2022. Cross-lingual text classification with multilingual distillation and zero-shot-aware training.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning.

A Appendix

A.1 Experiment Details

Implement Details. The unlabeled data used in ALSACE is constructed by Supergen (Meng et al., 2022), which uses PLM to generate text guided by label-descriptive prompts. We use machine translation[2] to generate unlabeled parallel multilingual text pairs based on the generated text. We leverage data generation methods(Supergen) and machine translation systems to construct an automatic pipeline for generating this valuable unlabeled parallel corpus at the lowest cost. We perform ALSACE on mPLMs using 500-shot unlabeled multilingual data with batch size 32 on each language corresponding to the tasks of XNLI, PAWS-X, and XCOPA. We set the learning rate to $3e-8$, and a dropout rate of 0.1. The thresholds θ in Equation 3 are used to select the teacher languages are 0.06, 0.2, and 0.2 for XNLI, PAWS-X, and XCOPA, respectively. We set the threshold θ to be the average value of the language score $\hat{s}_t$ across all languages.

To evaluate the effectiveness of ALSACE in limited resource scenarios, we fine-tune the mPLMs for 100 epochs with learning-rate of $1e-6$ on 128-shot English labeled examples as the baseline. Similarly, for PAWS-X, we fine-tune the mPLMs for 150 epochs with learning-rate of $1e-6$ on 512-shot English labeled examples.

A.2 Baselines

XLM-Align (Chi et al., 2021b) presents denoising word alignment as a new cross-lingual pre-training task with 310M instances. It self-labels word alignments for parallel sentences and haphazardly masks tokens in a bitext pair for mPLMs to predict.

[2]The translation API from http://api.fanyi.baidu.com/ is used to generate the multilingual parallel data.

Method	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
InfoXLM	**0.80**	0.90	0.70	1.00	1.00	0.50	0.60	1.70	1.40	0.40	1.10	1.10	1.10	2.10	0.40	1.00
ERNIE-M	0.20	1.00	0.60	0.50	0.80	0.50	0.80	1.60	1.40	1.10	1.10	0.80	1.70	2.30	1.60	1.10
ALSACE	0.20	**1.71**	**1.37**	.94	**1.98**	**2.35**	**2.20**	**2.27**	**2.57**	**2.16**	**2.65**	**1.92**	**2.58**	**3.09**	**2.72**	**2.05**

Table 7: Performance gain of each language compared with the initial XLM-R-large model.

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
PCT-XLM-R-large	88.30	84.20	85.10	83.70	83.10	84.40	81.90	81.20	80.90	80.70	78.80	80.30	78.40	73.60	75.60	81.30
$\Delta\downarrow$	\	4.10	3.20	4.60	5.20	3.90	6.40	7.10	7.40	7.60	9.50	8.00	9.90	14.70	12.70	7.45
ALSACE	88.30	84.37	85.59	83.71	83.33	84.67	82.16	80.28	80.84	81.80	79.24	81.94	79.12	73.29	75.11	**81.58**
$\Delta\downarrow$	\	3.93	2.71	4.59	4.97	3.63	6.15	8.02	7.46	6.51	9.06	6.37	9.18	15.01	13.19	**7.20**

Table 8: Comparison of PCT-XLM-R-large and ALSACE on XNLI benchmark across different languages. For a fair comparison, we report the performance of ALSACE under the same setting of PCT. Δ represents the cross-lingual transfer gaps. A smaller gap indicates better cross-lingual transferability.

InfoXLM (Chi et al., 2021a) implements on the basis of mPLMs and tries to align the representation of multiple languages by introducing parallel corpora with a new pre-training task. Initializes its parameters with XLM-R and employs contrastive learning using 42GB parallel corpora to encourage encoded representations of bilingual sentence pairs to be more similar than negative examples.

XLMR-adapter$_{256}$ (He et al., 2021) employs lightweight adapter modules on the XLM-R-large and achieves significant performances on low-resource and cross-lingual tasks.

ERNIE-M (Ouyang et al., 2021) is similar to InfoXLM and XLM-Align, which is implemented on the basis of XLM-R. It integrates back-translation into the pre-training process to encourage the model to align the representation of multiple languages with parallel corpora of about 68.8GB.

VECO (Luo et al., 2021) plug a cross-attention module into the transformer encoder to explicitly build the interdependence between languages to pretrain a variable cross-lingual language model for both NLU and NLG.

XLE (Chi et al., 2022) use ELECTRA-style tasks for pre-training mPLMs with a generator and discriminator structure using 142B tokens.

A.3 Compared with other state-of-art methods

A.3.1 Compared with pre-train based methods

InfoXLM (Chi et al., 2021a) initializes its parameters with XLM-R and employs contrastive learning using 42GB parallel corpora to encourage encoded representations of bilingual sentence pairs to be more similar than negative examples.

ERNIE-M (Ouyang et al., 2021) is implemented on the basis of XLM-R, and it integrates back-translation into the pre-training process to encourage the model to align the representation of multiple languages with parallel corpora of about 68.8GB.

While InfoXLM and ERNIE-M are built upon the basis of XLM-R by utilizing 42GB and 68.8GB data, respectively, our method only relies on a small amount of unlabeled parallel corpora (500-shot), which can be easily constructed with minimal effort. Despite this minimal requirement, our approach achieves substantial enhancements compared to the baseline XLM-R model. Table 7 illustrates the improvement of different methods across all languages on the XNLIdataset in comparison with the initial XLM-R-large baseline.

A.3.2 Compared with self-distillation-based methods

Qi et al. (2022) introduced PCT, a method that learns from various cross-lingual templates through a consistency loss, ensuring corresponding representations are aligned across languages. As indicated in Table 8, our ALSACE surpasses PCT-XLM-R-large in performance and demonstrates superior cross-lingual transferabilities. Thanks to the teacher language selection, ALSACE not only minimizes the performance disparities among the student languages but also enables the teacher languages to benefit from self-distillation. This approach yields improved overall performance and narrows the cross-lingual transfer gaps more effectively than PCT-XLM-R-large.

Method	Setup	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
XLM-R-base	CO-ZS	54.40	**49.80**	56.00	54.80	49.00	53.40	51.40	57.40	55.00	54.20	57.40	53.89
ALSACE	CO-ZS	**55.60**	49.40	**56.00**	**54.80**	**53.80**	**54.40**	**53.00**	**59.00**	**56.80**	**55.00**	**57.40**	**55.02**
XLM-R-base	SI-CO-ZS	**61.40**	51.60	66.60	64.40	**49.60**	55.80	62.00	61.60	60.20	**64.80**	68.20	60.56
ALSACE	SI-CO-ZS	58.40	**53.40**	**66.40**	**65.80**	49.00	**57.40**	**62.60**	**62.60**	**62.40**	64.60	**69.60**	**61.11**
XLM-R-large	CO-ZS	56.80	(50)	57.60	**58.60**	(50)	**52.20**	55.80	55.80	51.60	55.80	57.40	55.73
ALSACE	CO-ZS	**58.40**	(50)	**59.40**	57.60	(50)	51.80	**57.40**	**56.60**	**52.80**	**60.60**	**58.20**	**56.98**
XLM-R-large	SI-CO-ZS	72.00	(50)	77.00	77.20	(50)	61.60	67.20	76.40	74.40	76.60	77.40	73.31
ALSACE	SI-CO-ZS	**72.00**	(50)	**77.20**	**77.40**	(50)	**61.80**	**68.20**	**76.80**	**74.80**	**76.80**	**77.60**	**73.62**

Table 9: Accuracy scores of different models on the XCOPA test set when transferring from English. Models are either only fine-tuned on the COPA (Roemmele et al., 2011) training set and evaluated on different languages (CO-ZS) or fine-tuned first on SIQA (Sap et al., 2019) and then on COPA training set(SI-CO-ZS). Due to the inability of the XLM-R-large model to generate valid responses in Haitian Creole and Quechua, the scores for these languages are marked as (50) in the table.

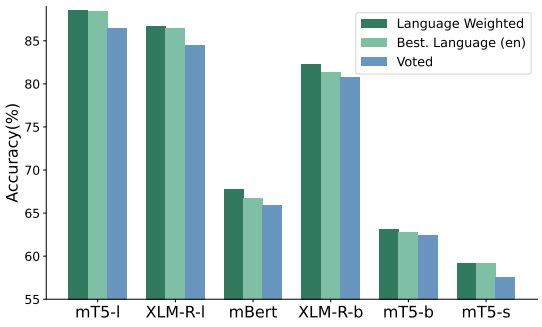

Figure 4: Performance of different Ensemble Methods.

A.4 Evaluation on XCOPA Benchmark

We evaluate ALSACE on the XCOPA benchmark, which is the causal commonsense reasoning benchmark across a range of typologically diverse languages, including both high and low-resource languages. Following the setting of Ponti et al. (2020), models are either fine-tuned solely on the COPA Roemmele et al. (2011) training set and then evaluated on XCOPA's multilingual test sets or sequentially fine-tuned—initially on the SIQA dataset Sap et al. (2019) followed by the COPA training set. Results in Table 9 show that our method achieves substantial performance gains in most languages under various settings across different model sizes. These outcomes underscore the robustness and overall effectiveness of our method.

A.5 Analysis

Why We Need to Select the Teacher Languages? To explore whether we need to select the teacher languages before transferring knowledge, we design an exploratory experiment on the XNLI dataset to demonstrate that selecting teacher language is necessary. We measure the contribution of different ensemble strategies to model performance. Specifically, ***language Weighted***: For predicted labels and confidence scores from different languages, we use the confidence score of each language as weights and calculate the final ensemble prediction.

Best Performing Language (en): We use the results predicted by English as the final prediction.

Voted: We give the same weight to the predicted labels for each language and get the final prediction result based on the voting result.

Figure 4 compares different multilingual models using different ensemble methods on the XNLI benchmark. *Voted* does not perform well due to noise from the under-performing student languages. On the other hand, by using the normalized language score $P(st)$ as weights for each language output in ensembling, it surpasses the performance of English, which is considered the best-performing high-resource language. This noteworthy discrepancy indicates that high-resource languages may not be suitable teacher languages. Besides high-resource languages, other languages also contribute to enhancing model performance. Figure 2 shows an experiment on GeomLAMA, demonstrating that high-resource languages may not be the most suitable for probing knowledge about a specific language condition. For instance, when addressing a query related to Chinese culture, Persian might yield a more accurate answer compared to English.

A.6 Ablation Study

We conducted an ablation study to investigate the impact of teacher language selection, with detailed results provided in Figure 10. A comparison of

Method		en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg.
XLM-R-large	Perf.	86.45	80.89	81.83	81.22	79.36	80.74	78.78	77.22	77.44	77.41	75.53	77.82	74.56	69.62	70.86	77.98
	Δ ↓	\	5.56	4.62	5.23	7.09	5.71	7.67	9.23	9.01	9.05	10.93	8.63	11.90	16.83	15.60	8.47
Excluding Weak Stu.	Perf.	86.79	81.70	83.19	82.32	80.86	82.55	80.54	79.08	78.36	79.22	77.07	79.12	76.45	71.62	72.71	79.44
	Δ ↓	\	5.09	3.59	4.47	5.93	4.23	6.25	7.70	8.42	7.56	9.72	7.66	10.34	15.17	14.07	7.35
Excluding Weak Tea.	Perf.	86.65	81.96	83.35	82.04	80.70	82.46	80.34	79.46	78.48	79.20	77.23	79.66	76.67	71.76	72.87	79.52
	Δ ↓	\	4.69	3.29	4.61	5.95	4.19	6.31	7.19	8.16	7.45	9.42	6.99	9.98	14.89	13.77	7.13
Random Selection	Perf.	86.81	82.15	83.15	82.18	80.94	82.66	80.78	79.40	79.14	79.08	77.78	79.34	75.93	71.89	72.91	79.61
	Δ ↓	\	4.66	3.66	4.63	5.87	4.15	6.03	7.41	7.67	7.73	9.03	7.47	10.88	14.92	13.90	7.20
No Selection	Perf.	86.89	82.29	83.13	82.22	80.96	82.66	80.70	79.24	79.06	79.02	77.50	79.62	76.81	72.09	73.09	79.69
	Δ ↓	\	4.60	3.76	4.67	5.93	4.23	6.19	7.65	7.83	7.87	9.39	7.27	10.08	14.80	13.80	7.20
Scale-Based Selection	Perf.	86.93	82.38	83.01	82.24	81.02	82.73	80.84	79.32	78.80	79.46	77.62	79.70	76.97	71.94	73.23	79.75
	Δ ↓	\	4.55	3.91	4.69	5.91	4.19	6.09	7.60	8.12	7.47	9.30	7.23	9.96	14.99	13.69	7.18
ALSACE	Perf.	86.65	82.61	83.21	82.16	81.34	83.10	80.98	79.50	79.60	79.98	78.18	79.74	77.13	72.71	73.57	**80.03**
	Δ ↓	\	4.04	3.44	4.49	5.31	3.55	5.67	7.15	7.05	6.67	8.47	6.91	9.52	13.94	13.08	**6.62**

Table 10: Ablation Study of the Teacher Language Selection. Δ represents the cross-lingual transfer gaps, i.e., performance drop between English and other languages in zero-shot transfer. A smaller gap indicates better cross-lingual transferability. We report the average performance and cross-lingual transfer gaps of all languages.

Method	Params	Perf.	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	**avg**(Δ ↓)
XLM-R-base	225M		84.23	77.39	78.20	76.45	75.97	77.80	75.35	73.27	71.84	74.93	71.88	74.23	69.22	64.55	65.77	74.07
		Δ ↓	\	6.84	6.03	7.78	8.26	6.43	8.88	10.96	12.39	9.30	12.35	10.00	15.01	19.68	18.46	10.88 (-)
E. Self-Train.	225M		84.09	77.96	78.28	76.73	76.25	78.14	75.65	73.33	72.12	75.27	71.78	74.35	69.54	64.85	66.27	74.31
		Δ ↓	\	6.13	5.81	7.36	7.84	5.95	8.44	10.76	11.97	8.82	12.31	9.74	14.55	19.24	17.82	10.48 (-0.40))
F. Self-Train.	225M		84.13	78.18	78.40	76.85	76.51	78.16	75.67	73.81	72.04	75.33	71.84	74.57	69.74	64.91	66.57	74.45
		Δ ↓	\	5.95	5.73	7.28	7.62	5.97	8.46	10.32	12.09	8.80	12.29	9.56	14.39	19.22	17.56	10.37 (-0.51))
ALSACE -base	225M		84.11	77.80	78.30	77.50	76.51	78.28	76.01	74.19	72.12	75.50	72.81	74.90	70.12	65.55	66.37	74.67
		Δ ↓	\	6.31	5.81	**6.61**	**7.60**	**5.83**	**8.10**	**9.92**	**11.99**	**8.61**	**11.30**	**9.21**	**13.99**	**18.56**	**17.74**	**10.11 (-0.77)**
XLM-R-large	550M		86.45	80.90	81.84	81.22	79.36	80.74	78.78	77.23	77.03	77.82	75.53	77.82	74.55	69.62	70.86	77.98
		Δ ↓	\	5.45	4.51	5.13	6.99	5.61	7.57	9.12	9.32	8.53	10.82	8.53	11.80	16.73	15.49	8.97 (-)
E. Self-Train.	550M		86.77	81.44	82.32	81.40	79.92	81.16	79.18	78.10	77.54	78.42	76.19	78.46	75.31	70.48	72.04	78.58
		Δ ↓	\	5.33	4.45	5.37	6.85	5.61	7.59	8.67	9.23	8.35	10.58	8.31	11.46	16.29	14.73	8.77 (-0.30))
F. Self-Train.	550M		86.81	81.54	82.69	81.68	80.30	81.96	79.70	78.40	78.12	78.90	76.45	78.06	74.93	70.34	71.90	78.79
		Δ ↓	\	5.27	4.12	5.13	6.51	4.85	7.11	8.41	8.69	7.91	10.36	8.75	11.88	16.47	14.91	8.60 (-0.47))
ALSACE -large	550M		86.65	82.61	83.21	82.16	81.34	83.09	80.98	79.50	79.60	79.98	78.18	79.74	77.13	72.71	73.58	80.03
		Δ ↓	\	**4.04**	**3.44**	**4.49**	**5.31**	**3.56**	**5.67**	**7.15**	**7.05**	**6.67**	**8.47**	**6.91**	**9.52**	13.94	**13.07**	**7.09 (-1.88)**
mT5-large	1.2B		88.42	82.44	83.49	81.68	81.14	81.96	79.90	77.33	76.87	78.52	75.31	77.74	75.31	72.63	70.88	78.91
		Δ ↓	\	5.98	4.93	6.74	7.28	6.46	8.52	11.09	11.55	9.90	13.11	10.68	13.11	15.79	17.54	10.19 (-)
E. Self-Train.	1.2B		88.50	82.46	84.33	82.02	81.84	82.34	80.96	78.04	78.06	79.70	76.81	78.44	75.73	73.57	72.04	79.66
		Δ ↓	\	6.04	4.17	6.48	6.66	6.16	7.54	10.46	10.44	8.80	11.69	10.06	12.77	14.93	16.46	9.48 (-0.72)
F. Self-Train.	1.2B		88.64	82.44	84.37	82.22	81.98	82.42	81.16	78.22	78.30	80.00	76.81	78.80	76.09	73.97	72.26	79.85
		Δ ↓	\	6.20	4.27	6.42	6.66	6.22	7.48	10.42	10.34	8.64	11.83	9.84	12.55	14.67	16.38	9.42 (-0.77)
ALSACE -mT5	1.2B		88.60	83.69	84.79	83.17	82.91	83.91	81.80	79.54	78.84	80.20	77.90	80.92	77.25	75.17	73.13	80.79
		Δ ↓	\	4.91	**3.81**	5.43	**5.69**	**4.69**	**6.80**	9.06	9.76	8.40	10.70	**7.68**	**11.35**	**13.43**	**15.47**	**8.37 (-1.82)**

Table 11: Comparison of self-distillation baselines with ALSACE . Δ represents the cross-lingual transfer gaps, i.e., performance drop between English and other languages in zero-shot transfer. A smaller gap indicates better cross-lingual transferability. We report the average performance and cross-lingual transfer gaps for all languages.

ALSACE's performance with and without including student-student pairs indicates that even though there is a performance improvement when student-student pairs are excluded, a significant performance gap remains compared to the complete ALSACE model. This is particularly evident for student languages, as detailed in Table 13. Additionally, when focusing on the student languages, such as Swahili and Urdu, the exclusion of student-student pairs results in comparatively diminished benefits from self-distillation.

The results clearly demonstrate that while the improvements persist, the performance of the ALSACE model employing randomly selected teacher languages still needs to catch up to the full ALSACE model across nearly all languages. This finding further underscores the efficacy of the teacher language selection strategy. ALSACE demonstrates competitive performance across various baselines, achieving notable results even with a limited amount of unlabeled parallel data. We successfully alleviated the performance disparities among different languages. As for the performance disparities, while there might still exist some gaps among different languages, ALSACE effectively mitigates these disparities, especially evident in languages like Swahili (sw), Urdu (ur), and Thai (th), as showcased in the performance comparison with English (en) in Table 2. This aligns with our motivation to enhance cross-lingual transferability.

Country	US		CN		IN		IR		KE		Avg.	
Method	XLM-R	ALSACE	XLM-R	ALSACE	XLM-R	ALSACE	XLM-R	ALSACE	XLM-R	ALSACE	XLM-R	ALSACE
en	0.4414	**0.4414**	0.2143	**0.2571**	0.2970	**0.2970**	**0.4483**	0.4207	**0.4125**	0.3875	0.3627	**0.3607**
zh	0.3862	**0.3931**	**0.3786**	0.3714	0.3091	**0.3333**	0.3103	**0.3310**	0.3375	**0.3375**	0.3443	**0.3533**
hi	0.3517	**0.3931**	**0.3071**	0.3000	0.3515	**0.3576**	**0.3931**	0.3448	0.3625	**0.3688**	**0.3532**	0.3529
fa	0.4828	**0.5034**	0.3143	**0.3429**	0.3697	**0.3818**	0.3517	**0.3655**	0.4250	**0.4375**	0.3887	**0.4062**
sw	0.2966	**0.3310**	0.2214	**0.2357**	0.3152	**0.3394**	0.2828	**0.3103**	0.3313	**0.3313**	0.2894	**0.3095**

Table 12: Detailed results of ALSACE on XLM-R-large in GeoMLAMA dataset. The result shows that ALSACE utilizes teacher languages to guide others and generally improves their language-specific knowledge.

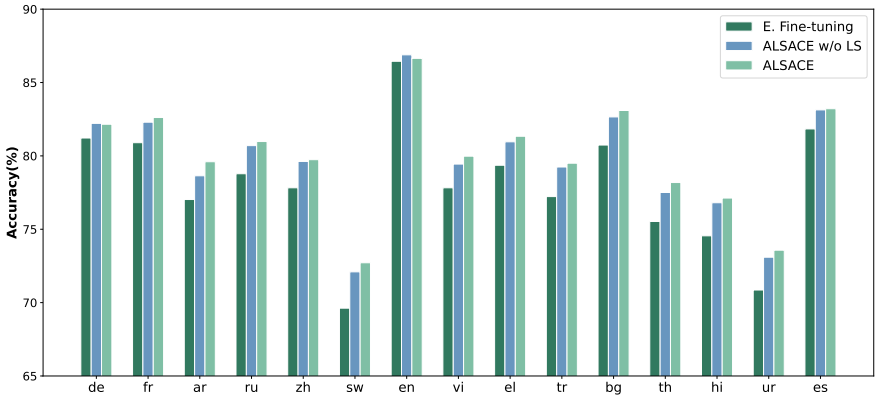

Figure 5: The Comparison of ALSACE Performance with and without Language Selection on XNLI dataset set. All results are based on XLM-R-large.

Lang.	Exc. Stu.(%)	ALSACE (%)	Change(△)(%)
en	0.44	0.20	-0.24
fr	1.40	1.72	0.32
es	1.30	1.38	0.08
de	1.00	0.94	-0.06
el	1.60	1.98	0.38
bg	1.92	2.36	0.44
ru	1.92	2.20	0.28
tr	2.02	2.28	0.26
zh	1.80	1.92	0.12
hi	2.26	2.57	0.32
vi	1.62	2.16	**0.54**
ar	1.62	2.57	**0.96**
th	1.98	2.65	**0.68**
sw	2.48	3.09	**0.62**
ur	2.24	2.71	**0.48**
avg	1.70	2.05	0.35

Table 13: Ablation Study Comparison

A.7 Geo-Diverse Commonsense across Countries

Figure 2 shows the detailed experiment results on GeomLAMA‘ (Yin et al., 2022), which demonstrates that ALSACE improves the performance of mPLM on knowledge probing tasks over various languages.