Analysis of Multi-Source Language Training in Cross-Lingual Transfer

Seong Hoon Lim[†], **Taejun Yun**[†], **Jinhyeon Kim**[†], **Jihun Choi**[‡], **Taeuk Kim**[*†]

[†]Hanyang University, [‡]Sony AI

{dmammfl,tj1616,kimjinhye0n,kimtaeuk}@hanyang.ac.kr, jihun.a.choi@sony.com

Abstract

The successful adaptation of multilingual language models (LMs) to a specific language-task pair critically depends on the availability of data tailored for that condition. While cross-lingual transfer (XLT) methods have contributed to addressing this data scarcity problem, there still exists ongoing debate about the mechanisms behind their effectiveness. In this work, we focus on one of the promising assumptions about the inner workings of XLT, that it encourages multilingual LMs to place greater emphasis on language-agnostic or task-specific features. We test this hypothesis by examining how the patterns of XLT change with a varying number of source languages involved in the process. Our experimental findings show that the use of multiple source languages in XLT—a technique we term Multi-Source Language Training (MSLT)—leads to increased mingling of embedding spaces for different languages, supporting the claim that XLT benefits from making use of language-independent information. On the other hand, we discover that using an arbitrary combination of source languages does not always guarantee better performance. We suggest simple heuristics for identifying effective language combinations for MSLT and empirically prove its effectiveness.

1 Introduction

There has been recent interest in multilingual language models (LMs), where multiple languages are used in training an LM, due to its capability of accepting multilingual inputs and achieving similar or even better performance compared to monolingual LMs at natural language processing (NLP) tasks. However, despite all these advantages, fine-tuning multilingual LMs using a specific language and an NLP task hinges on acquiring labeled data tailored to a specific purpose, which is not always possible.
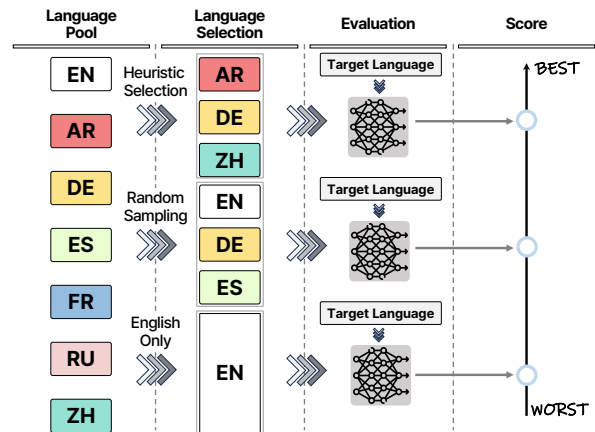
[*]Corresponding author.

Figure 1: Overview of the effectiveness of Multi-Source Language Training (MSLT) in cross-lingual transfer. As we adopt more sophisticated approaches for MSLT, we can expect improved performance (from bottom to top).

To address this challenge amid linguistic resource disparities, cross-lingual transfer (XLT) approaches have been proposed. In the context of fine-tuning multilingual LMs, XLT leverages resource-rich source languages in fine-tuning to boost performance on low-resource target languages.

Although there is still active debate regarding the inner workings of XLT, a significant body of previous research indicates that multilingual LMs are capable of separating *language-specific* and *language-agnostic* information from text (Muller et al., 2021), and that XLT favors the enhancement of language-agnostic features while reducing the emphasis on language-specific ones (Qi et al., 2022; Tu et al., 2022; Wang et al., 2022; *inter alia*).

While the standard practice in XLT is to employ only one language as a source for transfer, there exist attempts that have investigated the use of multiple source languages (Singh et al., 2019; Roy et al., 2020; Kew et al., 2023; Chai et al., 2024; Shaham et al., 2024). These studies reveal that employing multiple source languages in XLT—the concept we refer to as **Multi-Source Language**

Training (MSLT)—leads to performance improvement. However, despite it is evident that MSLT has clear advantages, several research questions remain unexplored, including: (1) the specific changes that it induces in the internal states of LMs, (2) the conditions under which it achieves its optimal performance, and (3) the breadth of applications where it can be effectively applied, among others.

In this work, we introduce a series of comprehensive analyses exploring the inner workings of MSLT, along with offering guidance for its effective utilization. Specifically, we first attempt to illustrate the utility of MSLT with a couple of intuitive visualizations, to clarify the underlying reasons for its effectiveness. In addition, our experimental results discover various intriguing findings. Regarding the number of languages involved in MSLT, we observe its positive correlation with performance in general, although the trend tends to plateau as the number of languages exceeds a certain threshold.

Our empirical discovery also suggests that the enhanced diversity of source languages in MSLT does not invariably lead to improved performance. This underscores the importance of carefully selecting an effective combination of source languages (see Figure 1 for an intuitive illustration). Therefore, we examine various heuristics that can identify a suitable selection of source languages from an exponentially large pool of possible combinations. Based on our tests, we propose several competent strategies, some of which are derived from the statistical properties of languages in terms of pretraining, while others utilize linguistic characteristics of languages. Lastly, we discuss intriguing patterns evident in the group of languages selected as sources for transfer, providing insights into the interplay of languages in MSLT.

2 Related Work

Multilingual LMs (Conneau et al., 2019; Scao et al., 2022; *inter alia*) are widely recognized for their ability to process inputs from diverse languages in an integrated manner. They have recently been the subject of extensive research for uncovering their working mechanisms, and this effort has converged on the robust hypothesis that the internal states of multilingual LMs can be categorized into language-sensitive and language-agnostic components (Choenni and Shutova, 2020; Chang et al., 2022; Zhao et al., 2020; Muller et al., 2021).

Meanwhile, cross-lingual transfer (XLT) aims to improve the performance of multilingual LMs on specific tasks in languages with limited resources (i.e., target languages), by utilizing the support from resource-rich languages (i.e., source languages). Previous research on XLT has explored diverse implementations of the concept. For instance, Yang et al. (2022) attempt to achieve XLT by blending representations from source and target languages, while Zheng et al. (2021) introduce data augmentation techniques such as subword sampling and code-switching. Wang et al. (2022) propose a contrastive learning framework aimed at reducing the differences in sentence embeddings between languages. Concurrently, Qi et al. (2022) present a prompt-based approach that facilitates information transfer from a source language to prompts composed in the target language.

Note that the majority of existing methods for XLT (Libovický et al., 2020; Tiyajamorn et al., 2021; Yang et al., 2022), including the aforementioned studies, adhere to a common theme: they are crafted to facilitate XLT by prompting the enhancement of language-agnostic or universal features within multilingual LMs. In this work, we focus on a simple strategy that yields a similar effect, namely, introducing multiple source languages in XLT instead of just one as is customary, deliberately steering clear of complex algorithms.

In the literature, there is empirical acknowledgment that increasing the diversity of source languages in XLT can boost its performance. Singh et al. (2019) suggest that pairing sentences written in different languages (for example, one sentence in one language and another in a different language) is beneficial for improving performance on the cross-lingual natural language inference (XNLI; Conneau et al., 2018) task. They additionally observe that the performance fluctuates based on the languages from which the sentences are sampled. Roy et al. (2020) demonstrate a similar trend in question-answering tasks. Further, Kew et al. (2023) and Shaham et al. (2024) reveal that MSLT is also advantageous for instruction-tuned LMs.

Despite the recent surge in attention towards MSLT, it is important to highlight the lack of a comprehensive analysis or tailored strategies in this area so far. For instance, Kew et al. (2023) consider only pre-defined language sets in their experiments, ignoring the potential of discovering the best combination of source languages. On the other hand, Singh et al. (2019) and Shaham et al. (2024) utilize

all available languages without specifically seeking to determine the ideal number of languages for the optimal performance of MSLT. Our work tackles these limitations, by presenting a series of well-supported illustrations that clarify how and why MSLT operates, along with providing guidelines for enhanced utilization of the MSLT framework.

3 Terminology

We first clarify the core concepts and their corresponding acronyms in this section.

Cross-Lingual Transfer (XLT) XLT is a technique applicable to multilingual LMs, particularly when obtaining a satisfactory amount of data for the language-task pair of interest is challenging. This is built on the assumption that fine-tuning a multilingual model for a task $\mathcal{T}$ in a resource-rich language $\mathcal{A}$ can lead to a synergistic improvement in performance for another language $\mathcal{B}$, especially if $\mathcal{B}$ has limited data available for the task $\mathcal{T}$.

In this work, we focus on the zero-shot XLT setting, under the assumption that no suitable data is available for the target language.

Single-Source Language Training (SSLT) SSLT represents an environment for XLT where multilingual LMs are trained on a single language. From now on, we adopt the notation 'SSLT(L)' to signify the setting where the language L is exclusively utilized as a source language in XLT.

Multi-Source Language Training (MSLT) MSLT refers to an environment in which two or more source languages are considered for XLT. We use the notation 'MSLT($L_1, L_2, \ldots, L_n$)' to denote the setting where MSLT is conducted with languages $L_1, L_2, \ldots, L_n$.

MSLT, when feasible, enables the possibility of leveraging an exponentially higher number of data points by allowing the use of instances from multiple languages. However, as the purpose of this study is to analyze the impact of language diversity, we decide to keep consistent the total number of data points used in experiments. For example, if SSLT(en) utilizes 1,000 data points in English, the corresponding MSLT(en, es) incorporates 500 data points each from English and Spanish, respectively.

4 Verifying the Effectiveness of MSLT

We present a thorough analysis to confirm the effectiveness of MSLT, and additionally provide insights into its inner workings. We describe the
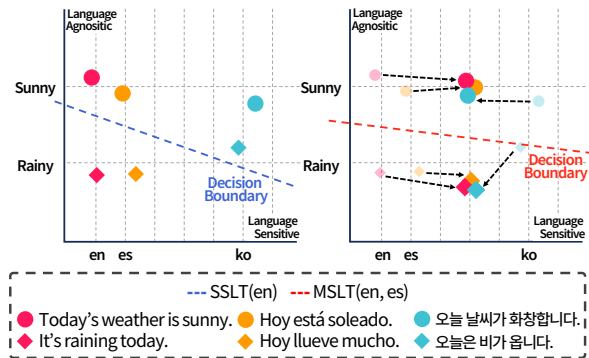

Figure 2: A conceptual illustration of the advantages of MSLT over SSLT. The left illustrates the training process of an LM using only English (en) (i.e., SSLT(en)), while the right represents MSLT with English (en) and Spanish (es) (i.e., MSLT(en, es)). Incorporating more source languages enhances language-agnostic features and blurs language-specific ones, potentially improving effectiveness for unseen languages such as Korean (ko).

intuition behind how MSLT enhances XLT performance (§4.2) and present visualizations related to our intuition (§4.3). We also provide empirical findings with respect to the ideal number of source languages in MSLT (§4.4). Section 4.1 presents the configurations used throughout the section.

4.1 Experimental Settings

To investigate the inner workings of MSLT, we conduct experiments using XLM-RoBERTa_Base (XLM-R_Base; Conneau et al., 2019), a compact multilingual encoder that covers around 100 languages. We select three tasks from the XTREME benchmark (Hu et al., 2020) as our testbed for analysis: named entity recognition (WikiANN; Pan et al., 2017), cross-lingual natural language inference (XNLI; Conneau et al., 2018), and cross-lingual paraphrase identification (PAWS-X; Yang et al., 2019). More details on the datasets and training processes are specified in Appendix A and B, respectively.

4.2 Advantages of MSLT over SSLT

Rooted in the widely accepted notion within the literature that multilingual LMs can categorize language-specific and language-agnostic concepts (Muller et al., 2021), there is a prevalent belief that XLT can be more effectively activated by encouraging the influence of language-agnostic features while diminishing the role of language-specific ones. We draw on this wisdom to intuitively explain why MSLT generally outperforms SSLT.

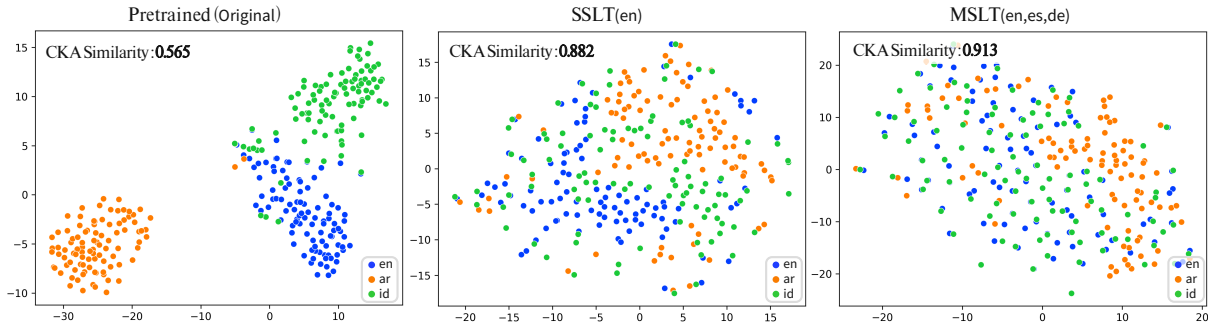Figure 2 illustrates the potential advantages of

Figure 3: Visualization of embeddings and corresponding CKA similarities (Kornblith et al., 2019) for 3 languages: English (en), Arabic (ar), and Indonesian (id). Note that English is used in both SSLT & MSLT, whereas Arabic and Indonesian are not. Therefore, we can observe the impact of SSLT & MSLT on both languages seen and unseen during training. **Left**: the original XLM-R. **Center**: XLM-R after SSLT(en). **Right**: XLM-R after MSLT(en, es, de). We find that while SSLT promotes language-agnostic alignment in the semantic space, MSLT enhances this further, leading to a more integrated space for languages.

MSLT over SSLT. When data from a single language is used in training for XLT, as in the case of SSLT, it may be challenging to guide a model to establish a decision boundary for classification (indicated by blue in Figure 2) that can robustly function across different languages. Conversely, with MSLT, the model is exposed to signals from a diverse range of languages, encouraging more frequent exploitation of language-agnostic features. This leads to representations that are comparatively language-independent, enabling the identification of a more robust decision boundary applicable across different languages (red in Figure 2). In summary, we speculate that MSLT provides a more advantageous environment for XLT compared to SSLT, by underscoring the utility of language-agnostic features.

4.3 Visualization of Embeddings after MSLT

To better understand the impact of MSLT and empirically validate the hypothesis presented in §4.2, we conduct a comparative study by visualizing embeddings before and after applying SSLT and MSLT. To be specific, we first train the XLM-R$_{Base}$ model with data instances from the XNLI dataset, resulting in three variants: (1) the original XLM-R, (2) XLM-R with SSLT(en), and (3) XLM-R with MSLT(en,es,de).[2] We then visualize the embeddings computed by each model variant,[3] using a set of sentences from the Parallel Universal Dependencies (PUD) treebanks (De Marneffe et al., 2021). This decision stems from our goal to evaluate the model's general capabilities, not just its

performance on fine-tuning tasks, but also its effectiveness with out-of-domain data instances. We make use of t-SNE (van der Maaten and Hinton, 2008) for dimension reduction.

In Figure 3, we observe that the XLM-R model with MSLT (the rightmost) demonstrates significantly better language-agnostic alignment for unseen languages (ar and id), compared to both (1) the original XLM-R and (2) XLM-R after SSLT. This implies that the diversification of source languages prevents a model from learning language-specific features and encourages it to be more language-independent, which is beneficial for achieving better performance in XLT.

4.4 On the Optimal Number of Languages

In the previous part, we qualitatively demonstrated the effectiveness of MSLT through graphical illustrations. Here, we present a supplementary quantitative analysis on the impact of the number of source languages on XLT, addressing two research questions: (1) Is MSLT consistently superior to SSLT, at least in the evaluated environments? (2) What is the optimal number of languages for MSLT? We select English, Spanish, German, Chinese, and French as candidates for source languages and evaluate the performance of different language groups on cross-lingual datasets, calibrating their size from 1 (SSLT) to 5 (MSLT with five languages).[4]

Figure 4 reveals that regardless of the task type or data quantity, task performance remarkably improves as the number of source languages increases from 1 to 3, demonstrating the effectiveness of

---

[2]en: English, es: Spanish, de: German.

[3]Embeddings are derived by mean-pooling the hidden states from the last layer of the language model.

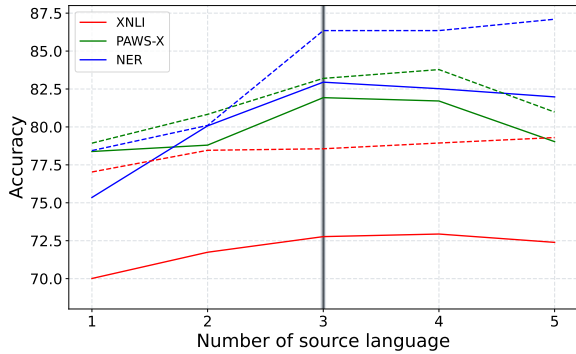[4]The scores of groups with the same size are averaged.

Figure 4: Performance of XLT can vary depending on the number of source languages. The solid lines correspond to XLM-R$_{\text{Base}}$ and dotted lines to XLM-R$_{\text{Large}}$.

Figure 5: Relative performance gaps vary with source language combinations, reaching as high as a 10-point difference between the best and worst options. The worst combinations of MSLT in XCOPA and XWinograd even harm performance compared to SSLT, highlighting the need for careful source language selection.

MSLT. However, when the size of the group exceeds 3, the performance exhibits minimal improvement or even declines in some cases.[5] Taking into account the outcomes of our experiments and the expenses involved in creating datasets across various languages, we propose opting for three source languages as a pragmatic and sensible selection for MSLT. Consequently, we make use of three-language groups in the forthcoming experiments.

5 Language Set Composition in MSLT

In the previous section, we demonstrated that MSLT exhibits markedly superior performance on XLT compared to SSLT. Assuming the selection of k source languages from a total of n languages, there exist $\binom{n}{k}$ possible combinations. Given that n can be significantly larger than k, fine-tuning with every possible combination of source languages to identify the optimal mix is not time- and resource-efficient. Therefore, it is necessary to select an appropriate combination more efficiently. In this section, we first establish the importance of selecting an effective set of source languages and then proceed to examine various criteria for choosing an optimal source language combination.

5.1 Experimental Settings

We perform experiments utilizing the BLOOM-7B model (Scao et al., 2022), in conjunction with the previously utilized XLM-R. We fix the size of the source language set at 3, and its elements are chosen from a pool of 7 language candidates,[6] which are commonly considered resource-rich languages.

When applying MSLT to XLM-R, we reuse the WikiAnn and XNLI datasets for fine-tuning and evaluation, as was done in the previous section.[7] For experiments with BLOOM-7B, we leverage separate datasets for instruction-tuning and evaluation, respectively. By introducing this instruction-tuning setting, we aim to evaluate whether MSLT is also effective in this new paradigm, which would further extend the applicability of MSLT. BLOOM-7B is instruction-tuned using the Bactrian-X dataset (Li et al., 2023). For evaluation, we employ another suite of datasets, which includes XCOPA (Ponti et al., 2020), XWinograd (Muennighoff et al., 2023), and XStoryCloze (Lin et al., 2022). Also note that we fine-tune BLOOM-7B with QLoRA (Dettmers et al., 2023) for computational efficiency and to ascertain how well MSLT operates in a parameter-efficient fine-tuning setup. Details on the experimental settings can be again found in Appendix A and B.

5.2 The Harmony of Languages Matters

We compare the performance of MSLT across various source language combinations. According to our experiments, the magnitude of performance improvement varies greatly depending on the combination used, and in some cases, a poor choice may lead to performance deterioration. Figure 5 represents the upper and lower bounds of the performance of MSLT, along with that of SSLT as a baseline. A notable performance gap is observed between the optimal and worst combinations in most

---

[5]In the case of XLM-R$_{\text{Large}}$, performance shows a slight improvement even with more than three languages.

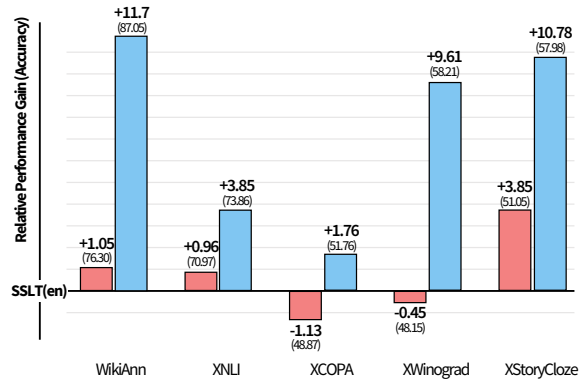[6]Arabic (ar), German (de), English (en), Spanish (es), French (fr), Russian (ru), Chinese (zh).

[7]PAWS-X is omitted owing to its limited linguistic variety.

Method	Encoder		Decoder		
	WikiAnn	XNLI	XCOPA	XWinograd	XStoryCloze
MIN	76.30 (35)	70.97 (35)	48.87 (35)	48.15 (35)	51.05 (35)
Size of Pretraining Data	78.52 (31)	72.87 (16)	50.35 (17)	54.77 (12)	54.05 (26)
Vocab Coverage	78.52 (31)	72.46 (26)	50.32 (22)	58.21 (2)	54.05 (25)
Embedding	85.58 (4)	72.26 (31)	48.85 (34)	53.33 (19)	56.98 (7)
Lang2Vec → Syntax	85.69 (3)	73.76 (3)	51.57 (4)	55.92 (7)	**57.42** (2)
→ Phonology	84.82 (8)	73.39 (7)	49.88 (27)	**58.21** (1)	55.12 (18)
→ Inventory	**86.07** (2)	**73.77** (2)	50.50 (11)	58.20 (3)	56.45 (10)
→ Family	84.82 (8)	73.39 (7)	49.88 (27)	**58.21** (1)	55.12 (18)
→ Geometry	85.58 (4)	73.24 (9)	**51.72** (2)	58.12 (6)	56.62 (9)
MAX	87.05 (1)	73.86 (1)	51.76 (1)	58.21 (1)	57.98 (1)

Table 1: Results of the proposed criteria for source language selection, evaluated across five different test datasets. The numbers in parentheses indicate the ranking of the specific language set chosen by each method among the total of 35 possible combinations, with higher ranks (indicating better performance) closer to blue and lower ranks (indicating inferior performance) closer to red. We confirm that Lang2Vec-based methods are proficient in proposing useful source language sets.

experiments. In certain tasks, e.g., XCOPA and XWinograd, MSLT with the least effective combination fails to outperform SSLT. These findings imply that the combination of source languages in MSLT has a substantial impact on performance, highlighting the need to establish criteria for selecting appropriate source languages.

5.3 Criteria for Source Language Selection

We further explore the criteria for selecting source language combinations by empirically testing several hypotheses related to them.

5.3.1 Size of Pretraining Data

We examine whether using a combination of the most frequent languages in the pretraining data leads to improved performance in MSLT. The underlying motivation is that the proportion of a language in the pretraining data might correlate with how much a model acquires the knowledge specific to that language during pretraining. Based on the volume of pretraining data for each language, as reported in the original papers (Conneau et al., 2019; Scao et al., 2022), we select three languages from the source language pool for fine-tuning: English, Russian, and German for XLM-R, and English, Simplified Chinese, and French for BLOOM-7B.

5.3.2 Vocabulary Coverage

Inspired by the findings of Pires et al. (2019), which suggest that XLT depends on vocabulary overlap between source and target languages, we test a heuristic that chooses source languages with extensive vocabulary coverage. Suppose there

are n source languages available, represented as $\mathcal{L} = \{L_1, \cdots, L_n\}$, and let the vocabulary set for each language L_i be denoted by V_{L_i}. The coverage of a candidate set $\mathcal{L}_C \subseteq \mathcal{L}$ is then defined by

$$\text{Coverage}(\mathcal{L}_C) = \frac{\left| \bigcup_{L_i \in \mathcal{L}_C} V_{L_i} \right|}{\left| \bigcup_{L_i \in \mathcal{L}} V_{L_i} \right|}.$$

We anticipate that language combinations possessing extensive lexical breadth will encompass a broader spectrum of the target language's vocabulary, ultimately leading to better XLT scores.

5.3.3 Linguistic Diversity

The last criterion we examine is the selection of source languages that exhibit distinct linguistic properties. This criterion shares a similar motivation with the previous ones, aiming to use languages that can generalize to other unseen languages. However, it uniquely posits that utilizing as diverse a range of source languages as possible will lead to a more stable and comprehensive modification of the model's internal mechanism.

To achieve this, we first represent each language as language vectors and then identify a candidate set whose language vectors have the *least* similarity to each other. We introduce Lang2Vec (Littell et al., 2017), a framework offering a suite of vectors for each language that mirrors its linguistic properties. In particular, we test five variants of Lang2Vec, i.e., Syntax, Phonology, Inventory, Family, and Geometry vectors. Formally, given a set of languages $\mathcal{L} = \{L_1, \ldots, L_n\}$ and their language vectors $\{v_1, \ldots, v_n\}$, the diversity score of a com-

Method	WikiAnn	XNLI	XCOPA	XW	XSC
Min	15	15	15	15	15
PT Data Size	12	6	**5**	6	11
Vocab Coverage	12	9	10	**2**	10
Embedding	**1**	13	9	9	7
L2V-Syn	**1**	**2**	11	**1**	6
L2V-Pho	**1**	**2**	11	**1**	6
L2V-Inv	6	12	8	5	8
L2V-Fam	**1**	**2**	11	**1**	6
L2V-Geo	6	12	8	5	8
Max	1	1	1	1	1

Table 2: Ranking of the results from different methods of selecting source languages in MSLT, especially in scenarios where English is consistently incorporated into the set. The patterns are similar; Lang2Vec-based criteria demonstrate reasonable performance.

Datasets	Top1	Top2	Top3	Top4	Top5
WikiAnn	de	de	ar	ar	ar
	fr	es	de	es	de
	zh	zh	zh	zh	ru
XNLI	ar	de	ar	ar	ar
	de	es	de	ru	fr
	ru	zh	zh	zh	zh
XCOPA	de	ar	fr	ar	ar
	en	es	ru	de	de
	zh	zh	zh	zh	es
XWinograd	ar	de	de	en	es
	en	en	es	es	fr
	zh	ru	zh	ru	ru
XStoryCloze	ar	ar	ar	es	ar
	de	de	de	fr	fr
	ru	zh	es	zh	zh

Table 3: The five most effective combinations of source languages for five specific tasks.

bination $\mathcal{L}_C \subseteq \mathcal{L}$ is defined by the sum of pairwise cosine similarity of language vectors:

$$\text{Diversity}(\mathcal{L}_C) = \sum_{\{L_i, L_j\} \subseteq \mathcal{L}_C} (1 - \text{sim}(v_i, v_j)).$$

Since our objective is to find a combination $\mathcal{L}_C$ having the maximum diversity, i.e. the minimal uniformity, thus the desired combination is obtained by $\arg\max_{\mathcal{L}_C} \text{Diversity}(\mathcal{L}_C)$.

On the other hand, we consider an extra baseline called just 'Language Embedding from LMs' or just 'Embedding'. This process is similar to that of Lang2Vec, but the vectors here are derived directly from pretrained LMs, specifically XLM-R or BLOOM-7B, depending on the case. The language vector for each language is calculated by averaging the embeddings of sentences in that language.

5.4 Evaluation on Selection Criteria

Table 1 represents the results from experiments on source language selection criteria presented in §5.3. The scores presented in the table are the averages of all scores for each task across target languages. Target languages will have the same language set when the selection criterion applied is identical.

The findings suggest that selection criteria based on the statistical properties of language models, referenced in Sections §5.3.1 and §5.3.2, do not substantially improve the performance of XLT. In contrast, selecting source languages based on linguistic diversity demonstrates a significant performance gain. This implies that in MSLT, prioritizing diversity within the source language combination is more important than relying on the characteristics of corpora used in pretraining LMs.

Furthermore, although the 'Embedding' method, which measures diversity directly based on embed-

dings from LMs, led to a fairly reasonable performance gain, measuring diversity using Lang2Vec vectors demonstrated the best outcomes. This further reinforces the importance of properly considering the inherent properties of languages in NLP.

5.5 Language Selection Sticking to English

This section features a case-study assuming that one of the source languages is necessarily English, a common scenario in practical applications. Table 2 reports that even in cases where English is always included in combinations, selecting other source languages based on diversity (or least similarity) according to Lang2Vec still exhibits reasonable performance.

6 Discussion

6.1 Which Language is Suitable for MSLT?

To identify the necessary conditions of the languages comprising the optimal set, we manually examine top-ranked language combinations for each task in practice. Table 3 showcases the top five language combinations that exhibit the best performance for each task. The results reveal that, across all tasks, Chinese (zh), Arabic (ar), and German (de) are the most commonly used languages, with frequencies of 17, 15, and 15, respectively. Furthermore, we find that each of the top 5 combinations includes two or more distinct writing systems, such as Arabic scripts for Arabic, Latin scripts for European languages, Chinese characters for Chinese, and Cyrillic scripts for Russian. This suggests that
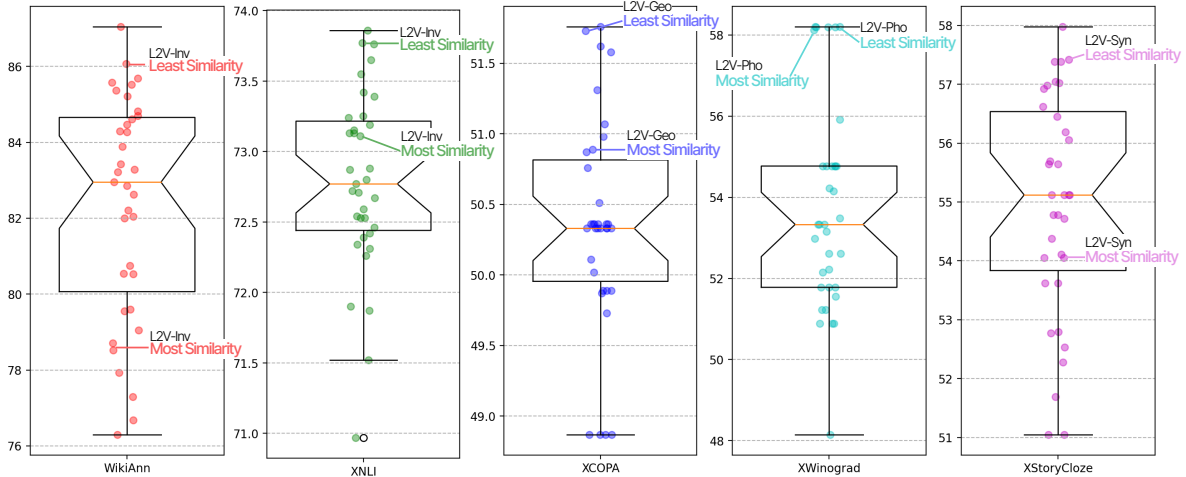
Figure 6: Visualization of performance for all (35) possible language sets. We observe that language combinations with the **least** similarity (**high** diversity) yield better performance than those with the **most** similarity (**low** diversity).

Dataset	Case 1	Case 2	Case 3
	$a = b = c$	$a = b \neq c$	$a \neq b \neq c$
WikiAnn	72.26	72.69	**73.07**
XNLI	80.36	81.54	**84.02**
XCOPA	52.82	53.06	**53.10**
XWinograd	52.33	52.89	**53.40**
XStoryCloze	52.54	53.32	**53.62**

Table 4: Test accuracy results based on the diversity of writing systems. Writing system diversity in language combinations is categorized into three levels: Case 1, with all three languages sharing the same system; Case 2, with two languages sharing a system; and Case 3, with each language having a different system. The outcomes show that combinations of language with distinct writing systems always outperform other cases.

the diversity of writing systems can act as a key factor in constructing good language sets for MSLT.

To further investigate this perspective, we conduct an extra experiment whose outcomes are reported in Table 4. In this experiment, we categorize source language sets (in total, 35) into three types based on the extent of diversity in their writing systems, and then compare the performance of each group. The outcomes show that combinations of language with distinct writing systems always outperform other cases, implying the importance of diversity in writing systems among languages. We believe that Lang2Vec-based criteria can properly exploit similar linguistic information.

6.2 Is Increasing Diversity Really Helpful?

In §5.3.3, we follow the intuition that selecting language combinations with elements exhibiting distinct characteristics is beneficial. In fact, we

can also consider the opposite direction, grouping languages with relatively high similarities rather than low ones. To determine which direction is more helpful for MSLT, we conduct an experiment using the two contrastive approaches.

Experimental results in Figure 6 reveal that constructing combinations with maximum diversity (as in §5.3.3) yields better performance than combinations having minimum diversity (i.e., high similarity) across the majority of tasks.[8] Notably, there is a significant performance gap between the two contrasting strategies for WikiAnn. We also observe similar patterns for other datasets, confirming the desirability of our decision to consider the diversity of source languages in §5.3.3.

7 Conclusion

This work demonstrates the positive impact of Multi-Source Language Training (MSLT) in crosslingual transfer (XLT). By leveraging multiple source languages, MSLT facilitates the learning of language-agnostic features in multilingual LMs. We provide qualitative and quantitative evidence suggesting that this approach promotes crosslingual semantic alignment, leading to better XLT performance compared to using a single source language. Furthermore, we identify the importance of selecting the optimal size and composition of source language sets for MSLT. To guide this selection efficiently, we propose diverse criteria based on the statistical and linguistic properties of candidate languages, such as Lang2Vec. Importantly,

---

[8]Full experimental results are listed in Appendix, Table 6.

MSLT demonstrates effectiveness across various configurations, including different tasks, architectures, and training paradigms. We hope this study serves as a springboard for further exploration of factors influencing the development of effective cross-lingual transfer methodologies.

Limitations

Limited source language pools   Given the practical availability of language-specific datasets and the ease of experimentation, this paper designates the top 7 languages that appeared most frequently in various multilingual datasets as source language candidates. By analyzing the permutations of selecting three out of these 7 languages, we aim to provide insights based on a wide range of scenarios to establish relevance in a generalized setting. However, it is important to note that further experiments with a greater number of languages as candidates for source languages are left as future work.

Comparison with other source language selection methods   Our study shares many similarities with other research on source language selection for cross-lingual transfer, aiming to find the optimal source languages. However, many selection methods tend to focus on the correlation between the source and target languages, making the selection process target-specific and often favoring single-language choices, which makes it challenging to quantitatively compare them with our approach.

Ethics Statement

In this study, we utilize models and datasets publicly available on Huggingface, ensuring that no ethical issues are associated with their usage. Additionally, we endeavor to minimize content related to social biases. However, we acknowledge the possibility that some biases may have been infused during the pretraining process of language models, highlighting the need for sufficient consideration in this regard.

Acknowledgements

References

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. 2022. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint arXiv:2009.12862*.

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

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *arXiv preprint arXiv:2312.12683*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pretrained multilingual representations. *arXiv preprint arXiv:2004.05160*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. *arXiv preprint arXiv:2101.11109*.

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.

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.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool. *arXiv preprint arXiv:2004.05484*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter openaccess multilingual language model. *arXiv preprint arXiv:2211.05100*.

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

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Languageagnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. *arXiv preprint arXiv:2205.04182*.

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

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2020. Inducing language-agnostic multilingual representations. *arXiv preprint arXiv:2008.09112*.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. *arXiv preprint arXiv:2106.08226*.

Appendix

A Dataset

WikiAnn is a named entity recognition dataset sourced from Wikipedia, comprising 282 language data segments. In our experimental setup, we utilize 7 languages (Arabic (`ar`), German (`de`), English (`en`), Spanish (`es`), French (`fr`), Russian (`ru`), Chinese (`zh`)) as the source language and 8 languages (Indonesian (`id`), Greek (`el`), Hebrew (`he`), Finnish (`fi`), Thai (`th`), Turkish (`tr`), Japanese (`ja`), Korean (`ko`)) as the target language for cross-lingual transfer.

PAWS-X comprises 23,659 human-translated evaluation pairs and 296,406 machine-translated training pairs across six typologically distinct languages: French (`fr`), Spanish (`es`), German (`de`), Chinese (`zh`), Japanese (`ja`), and Korean (`ko`). These translated pairs originate from examples within PAWS-Wiki. Two languages that are not included in the source language are used as target languages.

XNLI is an evaluation corpus for language transfer and cross-lingual sentence classification in 15 languages. XNLI cover 15 languages: Arabic (`ar`), Bulgarian (`bg`), Chinese (`zh`), English (`en`), French (`fr`), Greek (`el`), German (`de`), Hindi (`hi`), Russian (`ru`), Spanish (`es`), Swahili (`sw`), Thai (`th`), Turkish (`tr`), Urdu (`ur`), and Vietnamese (`vi`). Among the aforementioned 15 languages, all except the one employed as the source language in our experiment serve as target languages.

XCOPA, known as Cross-lingual Choice of Plausible Alternatives, is a benchmark designed to evaluate machine learning models' capability to transfer commonsense reasoning across different languages. This dataset comprises translations and re-annotations of the English COPA, encompassing 11 languages (Estonian (`et`), Haitian (`ht`), Indonesian (`id`), Italian (`it`), Quechua (`qu`), Swahili (`sw`), Tamil (`ta`), Thai (`th`), Turkish (`tr`)) from diverse language families and regions worldwide. All languages in XCOPA are evaluated as target languages by instruction-tuned models.

XWinograd is a multilingual benchmark designed for assessing commonsense reasoning, consisting of Winograd Schema Challenge problems presented in six languages. The task entails selecting the most plausible sentence from slightly varied options. We used Japanese (`ja`) and Portuguese (`pt`) as target languages for our experiments.

XStoryCloze comprises the English StoryCloze dataset professionally translated into 10 non-English languages. Out of these, we chose six languages as target languages, ensuring they did not overlap with the source language set. The selected target languages are as follows: Basque(`eu`), Hindi (`hi`), Indonesian (`id`), Burmese (`my`), Swahili (`sw`), and Telugu (`te`).

B Training Details

We report the hyperparameters used for training each model in Table 5. For XLM-R$_{\text{Base}}$ and XLM-R$_{\text{Large}}$, we conducted a hyperparameter search to find an appropriate learning rate for training. As for BLOOM-7B, we attempted to follow the settings from the original LoRA-tuned model,[9] but due to the lack of computation resources, we were unable to replicate the original settings such as batch size 128. Instead, we adjusted the training steps the same as the original LoRA-tuning setting. For the BLOOM model, prompt strategies employed for instruction-tuning and inference are listed in Table 7 and Table 8, respectively.

Model	XLM-R$_B$	XLM-R$_L$	BLOOM$_{7B}$
Common	cutoff_length: 512 & weight decay: 0.01		
Learning rate	2e-5	5e-6	3e-4
LoRA$_r$	-	-	64
LoRA$_m$	-	-	q,k,v
Task	WikiAnn	XNLI	PAWS-X	Bactrian-X
Batch	16	16	16	4
Step	243K	12.5K	30K	5K
GA	1	1	1	3

Table 5: Best hyperparameters for each task and model. LoRA$_r$ indicates the rank of LoRA, LoRA$_m$ represents the position of LoRA's module, and GA stands for Gradient Accumulation.

---

[9]https://huggingface.co/MBZUAI/bactrian-x-bloom-7b1-lora

Method	WikiAnn		XNLI		XCOPA		XWinograd		XStoryCloze	
	Most	Least	Most	Least	Most	Least	Most	Least	Most	Least
Embedding	77.93	85.58	72.88	72.26	48.85	48.85	52.61	53.33	55.12	56.98
L2V-Syn	78.52	85.69	72.87	73.76	50.32	51.57	**58.21**	55.92	54.05	**57.42**
L2V-Pho	76.68	84.82	72.88	73.39	49.88	49.88	58.20	**58.21**	51.05	55.12
L2V-Inv	78.71	**86.07**	73.13	**73.77**	50.85	50.50	52.98	58.20	56.92	56.45
L2V-Fam	82.95	84.82	71.90	73.39	50.75	49.88	58.20	**58.21**	53.62	55.12
L2V-Geo	78.71	85.58	73.13	73.24	50.85	**51.72**	52.98	58.12	56.92	56.62
AVG	78.92	85.43	72.80	73.30	50.25	50.40	55.53	57.00	54.61	56.29

Table 6: Performance comparison of language combinations with high and low inter-lingual similarity. Accuracy is used as the evaluation metric for all tasks. The best results for each task are in **bold**.

Only Instruction:	### Input: What is the capital of France? ### Output: The capital city of France is Paris.
Instruction with input:	### Input: Evaluate this sentence for spelling and grammar mistakes. He finished his meal and left the restaurant. ### Output: There are two spelling errors in the sentence. The corrected sentence should be: "He finished his meal and left the restaurant."

Table 7: Prompting strategies for instruction tuning.

XCOPA	### Input: Answer the number of options which is more plausible for the effect of the situation "Mees keeras kraani lahti." 1. Tualett täitus veega. 2. Tilast voolas vett. ### Output:
XWinograd	### Input: Answer the number of option which is more plausible for the blank. "チンパンジーはリナックスを使えなかった。_が動物だからだ。" 1. チンパンジー 2. リナックス ### Output:
XStoryCloze	### Input: Answer the number of option which is more natural for the end of story. "Rick tumbuh di keluarga bermasalah. Dia tidak pernah menerima dukungan dari keluarga, dan menjadi anggota geng. Tak butuh waktu lama sampai Rick tertembak dalam sebuah perampokan. Peristiwa itu membuatnya insaf." 1. Kini dia bahagia. 2. Dia menjadi anggota geng. ### Output:

Table 8: Prompting strategies for inference.