Rethinking what Matters: Effective and Robust Multilingual Realignment for Low-Resource Languages

Quang Phuoc Nguyen[1*], **David Anugraha**[2*], **Felix Gaschi**[3*], **Jun Bin Cheng**[1], **En-Shiun Annie Lee**[1,4]
[1]Ontario Tech University [2]Stanford University [3]SAS Posos [4]University of Toronto
quangphuoc.nguyen@ontariotechu.net, david.anugraha@stanford.edu, felix@posos.fr

Abstract

Realignment is a promising strategy to improve cross-lingual transfer in multilingual language models. However, empirical results are mixed and often unreliable, particularly for typologically distant or low-resource languages (LRLs) compared to English. Moreover, word realignment tools often rely on high-quality parallel data, which can be scarce or noisy for many LRLs. In this work, we conduct an extensive empirical study to investigate whether realignment truly benefits from using all available languages, or if strategically selected subsets can offer comparable or even improved cross-lingual transfer, and study the impact on LRLs. Our controlled experiments show that realignment can be particularly effective for LRLs and that using carefully selected, linguistically diverse subsets can match full multilingual alignment, and even outperform it for unseen LRLs. This indicates that effective realignment does not require exhaustive language coverage and can reduce data collection overhead, while remaining both efficient and robust when guided by informed language selection.[1]

1 Introduction

Multilingual pre-trained language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) enable cross-lingual transfer, where models fine-tuned to a certain task with an English dataset can be generalized to the same task in other languages (Pires et al., 2019; Wu and Dredze, 2019). However, their performance often degrades for typologically distant languages, such as low-resource languages (LRLs) (Pires et al., 2019). A promising strategy to address this issue is to perform realignment, which explicitly retrains models to produce similar representations for translated sentence pairs using objectives inspired by multilingual word embeddings (Conneau et al., 2017; Artetxe et al., 2018).

Despite a strong correlation between alignment and cross-lingual transfer (Gaschi et al., 2023), results from realignment methods remain mixed. While some studies report benefits of realignment (Cao et al., 2020; Zhao et al., 2020), others observe limited or even negative effects (Wu and Dredze, 2020; Efimov et al., 2023). These findings align with previous observations, where multilingual models exhibit good alignment for closely related languages, but remain more misaligned for distant or LRLs (Dou and Neubig, 2021).

In addition, realignment is not always feasible for all languages. It requires high-quality translation data, which may be unavailable for many LRLs (Gu et al., 2018; Liu et al., 2021; Anugraha et al., 2024). Even when resources exist, alignment quality can vary significantly across languages, potentially degrading downstream performance. This raises a central question: **Do we need to use all available languages for a better realignment, or could a carefully selected subset of languages offer similar or improved cross-lingual transfer performance?**

In this work, we conduct an extensive empirical study to investigate whether realignment truly benefits from using all available languages, or if strategically selected subsets can offer comparable or even improved cross-lingual transfer. In summary, our key contributions are:

1. **We conduct the first large-scale, systematic evaluation of realignment across 65 languages**, including 29 LRLs, 3 tasks, 4 seeds, and 2 models (with a strong focus on low-resource scenarios). By introducing a sentence-level averaging and contrastive objective that removes the need for word aligners, we show significant gains of up to 10 points in cross-lingual transfer, especially for

[*]Equal contribution.
[1]Our code can be found at https://github.com/felixgaschi/multilingual-alignment-and-transfer.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1877–1905

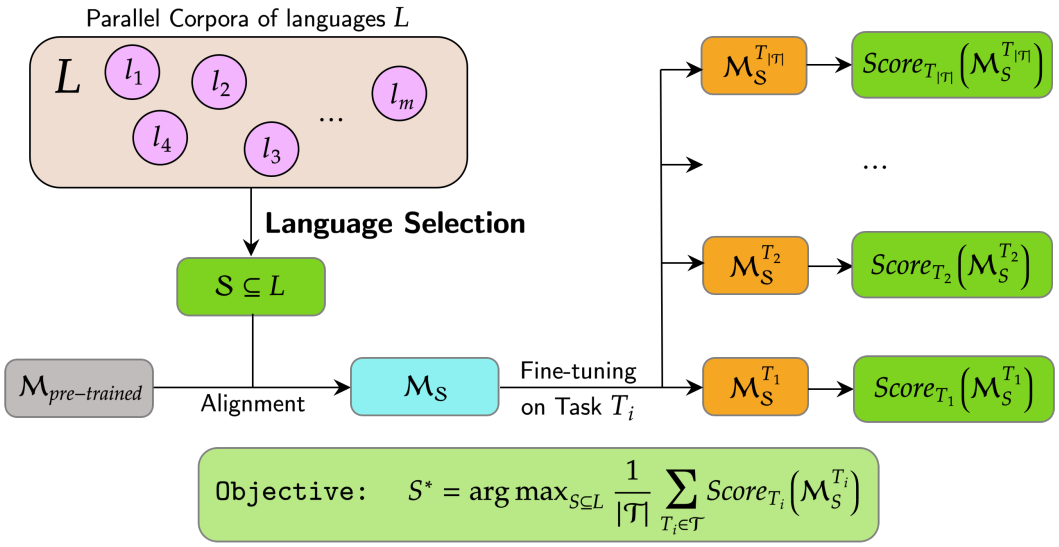

Figure 1: Overall diagram of the realignment process. Our goal is to empirically investigate how language selection within the realignment dataset impacts overall downstream task performance.

LRLs unseen during pre-training.

2. **We systematically investigate language subset selection for efficiency**, demonstrating that informed subsets chosen via heuristics (like URIEL featural diversity) can match or surpass full multilingual realignment. This shows that linguistic diversity matters more than the sheer number of languages.

3. **We perform comprehensive ablation studies, including out-of-distribution robustness**. We evaluate on unseen, out-of-distribution benchmarks (e.g., AmericasNLI) to show that diverse subset selection generalizes effectively. We also conduct ablations by scaling the number of languages and varying initial language pools to reflect realistic resource constraints, showcasing the importance of including LRLs in realignment.

To the best of our knowledge, we are the first to evaluate realignment massively on truly LRLs.

2 Methodology

Recall that we perform realignment to explicitly retrain multilingual encoders to produce similar representations for translated sentence pairs. In particular, we first perform realignment as a separate training phase, which is then followed by full-model fine-tuning on a downstream task, following previous work Wu and Dredze (2020); Gaschi et al. (2023); Bakos et al. (2025). For the realignment phase, we adopt the method proposed by Wu and Dredze (2020), which modifies the encoder to produce similar representations for semantically equivalent words across languages. This is achieved using a contrastive loss applied to word-level alignment pairs extracted from parallel corpora.

Prior work has typically relied on extracting word pairs from parallel sentences using word aligners such as FastAlign (Dyer et al., 2013) or bilingual dictionaries Gaschi et al. (2023). However, these alignment resources are often unreliable or entirely unavailable, especially for LRLs, and their use typically requires substantial computational resources. Thus, we propose a simple alternative that removes the dependency on word aligners while requiring significantly less time and computational resources. Our method instead averages the representations of words in each sentence of a translation pair and directly minimizes the distance between these sentence-level representations.

Formally, let B denote the batch size, and let $H = \{(h_i, \tilde{h}_i)\}_{i=1}^{B}$ represent a batch of B aligned sentences, where h_i is the averaged embedding of the words in a source (e.g., English) sentence and $\tilde{h}_i$ is the embedding of its aligned counterpart in the target language. The goal is to bring h_i and $\tilde{h}_i$ closer together in the embedding space while pushing h_i away from all other unaligned sentences in the batch. This is achieved via the following contrastive loss:

$$\mathcal{L}(\theta) = \frac{1}{2B} \sum_{h \in H} \log \frac{\exp\left(\text{sim}(h, \text{aligned}(h))/T\right)}{\sum_{h' \in H,\, h' \neq h} \exp\left(\text{sim}(h, h')/T\right)} \tag{1}$$

where $\text{sim}(h, h')$ denotes cosine similarity between two representations and T is a temperature hyperparameter, set to 0.1 in our experiments.

Note that the contrastive loss defined above implicitly depends on the translation data used, particularly the set of languages involved. Prior work typically performs realignment using all available languages for their parallel data (Wu and Dredze, 2020; Gaschi et al., 2023; Bakos et al., 2025). In contrast, we hypothesize that a carefully selected *subset* of languages may suffice to achieve comparable or even improved downstream cross-lingual generalization.

Formally, let $L = \{\ell_1, \ell_2, \ldots, \ell_m\}$ be the full set of languages for which parallel corpora with English are available. Let $\mathcal{D}_S$ denote the parallel data involving English and the languages in subset $S \subseteq L$, and let $\mathcal{M}_S$ denote the model after realignment using this data. Let $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ be a set of downstream tasks, and for each task $T_i \in \mathcal{T}$, we fine-tune the realigned model $\mathcal{M}_S$ on task-specific supervision to obtain the fine-tuned model $\mathcal{M}_S^{T_i}$. We then compute the corresponding evaluation score $\text{Score}_{T_i}(\mathcal{M}_S^{T_i})$. Overall, our goal is to find the subset $S^* \subseteq L$ that maximizes the macro-average across all downstream tasks:[2]

$$S^* = \arg\max_{S \subseteq L} \frac{1}{|\mathcal{T}|} \sum_{T_i \in \mathcal{T}} \text{Score}_{T_i}\left(\mathcal{M}_S^{T_i}\right) \tag{2}$$

Since the downstream evaluation metric is non-differentiable and trying all possible subsets of L is expensive, we do not optimize this objective directly. Instead, we construct subsets using linguistic-motivated heuristics as will be described in Section 3.1.

3 Experimental Setup

3.1 Language Subsets

We construct and evaluate subsets of languages based on heuristics designed to capture different dimensions of cross-lingual diversity and coverage. These heuristics consider factors such as linguistic feature diversity, language family affiliation, and script variation. To assess the effectiveness of these heuristics, we also compare their performance against randomly selected realignment languages, thereby evaluating the significance of each heuristic.

All subsets are drawn from the same pool of 65 languages, which we denote as L_{65}. This pool consists of 47 languages from XTREME-R (Ruder et al., 2021) together with 21 additional African languages, with some overlap between the two groups. The sets of 21, 47, and 65 languages serve as our baseline subsets. Details of the languages in L_{65} are provided in Table 5. For each heuristic, we evaluate subsets of size $n \in \{5, 10, 20, 40\}$, corresponding to increasing coverage when available.

Baselines. We include two types of baselines to contextualize the performance of our subset selection strategies. The first baseline uses the fixed language sets of size 21, 47, and 65 as mentioned above. The second baseline consists of random subsets sampled uniformly from L_{65} with $n \in \{5, 10, 20, 40\}$. These baselines help distinguish the effect of informed linguistic heuristics from arbitrary selection. All random subsets are generated using fixed random seeds for reproducibility.

Language Featural Diversity. This heuristic aims to compute diversity for languages based on their structural linguistic features. These features are obtained using the URIEL+ database (Khan et al., 2025), which is a language vector resource that encodes languages based on typological, geographic, phonological, syntactic, and phonetic inventory feature vectors. These representations allow for the computation of pairwise distances between languages, using angular distance over their vectorized representations. We compare two types of subsets: (1) subsets where we have the most diverse set of languages from set L_{65}, and (2) subsets where we have the least diverse set of languages from set L_{65}. To construct our most diverse subsets, we select languages that maximize their pairwise featural distance from English, with English included in the subset calculation but not considered during realignment. We also constructed the least diverse subsets to contrast with the diverse case by minimizing the total pairwise featural distance. The formal definition of the objective can be found in Section A.1.

[2] This performance evaluation setup follows prior work in multi-task multilingual learning, such as by XTREME-R (Ruder et al., 2021).

Language Family Diversity. This heuristic investigates whether diversity in genetic lineage contributes to effective realignment. We compare two types of subsets: (1) subsets where each language comes from a distinct language family other than the Indo-European family, and (2) subsets that are restricted to a single language family, specifically, Indo-European languages, to contrast with the diverse case.

Script Diversity. This heuristic investigates whether diversity in language scripts contributes to effective realignment. We compare three types of subsets: (1) subsets where each language is drawn from a distinct script other than Latin, (2) diverse subsets (as defined in Language Featural Diversity) but restricted to languages in L_{65} that use only the Latin script, and (3) least diverse subsets restricted to the Latin script, serving as a contrast to the diverse case.

All the language subsets and their languages are listed in the Appendix.

3.2 Models and Datasets

Realignment dataset We use OPUS-100 (Zhang et al., 2020) and NLLB (Costa-Jussà et al., 2022), which contain parallel corpora with sentence pairs across 100 and 200 languages, respectively. Whenever a language is not covered by OPUS-100, we fall back to the NLLB dataset.

Training and Downstream Task Datasets To evaluate cross-lingual transfer, we fine-tune all models exclusively on the English subset and evaluate them directly on other languages without additional fine-tuning. Our evaluation is mostly focused on *in-distribution* datasets, where evaluation languages are part of the realignment language set. However, we also included an *out-of-distribution (OOD)* dataset scenarios, which contain languages not seen during pre-training or realignment. Detailed dataset statistics are provided in Table 7.

In-Distribution Datasets. We consider three downstream tasks: Part-of-Speech (PoS) tagging, Named Entity Recognition (NER), and Natural Language Inference (NLI), evaluated on datasets covering both the XTREME-R language set and their African counterparts. For each task, we evaluate on datasets covering both the XTREME-R language set and African counterparts:

- For PoS tagging, we fine-tune on the UDPOS dataset (De Marneffe et al., 2021) and evaluate on both UDPOS and MasakhaPOS (Dione et al., 2023).
- For NER, we fine-tune on the English subset of WikiANN (Pan et al., 2017) and evaluate on WikiANN and MasakhaNER (Adelani et al., 2022).
- For NLI, we fine-tune on the English subset of XNLI (Conneau et al., 2018) and evaluate on four datasets: XNLI, IndoNLI (Mahendra et al., 2021), Myanmar-XNLI (Htet and Dras, 2025a), and AfriXNLI (Adelani et al., 2024). For AfriXNLI, we restrict evaluation to its African languages.

Out-of-Distribution Datasets. To study generalization beyond the languages present during pre-training or realignment, we evaluate NLI performance on AmericasNLI (Ebrahimi et al., 2021), which covers 10 typologically diverse languages absent from both the pre-training and realignment language sets. This dataset serves as a challenging out-of-distribution benchmark to assess zero-shot cross-lingual transfer.

Models We use mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) as our multilingual pre-trained language models. All experiments are run on 4 different seeds. More details about the hyperparameters can be found in Section A.6.

4 Results and Analysis

4.1 Results Overview

Figure 2 presents a comparison of the best average performance across different tasks for both XLM-R and mBERT, evaluated under different language subset heuristics. Detailed per-task results are reported in Tables 8 and 10.

First, Figure 2 shows that realignment provides significantly better overall results than the fine-tuning baseline. Except for two selection methods, realignment provides at minimum a one-point improvement, indicating the benefits of having to perform realignment.

However, performing realignment using all languages is not necessary to achieve comparable or even better performance. Figure 2 shows that the realignment strategy based on URIEL featural diversity and URIEL featural diversity within languages with Latin script consistently yields the best results, achieving performance comparable to using the full set of 65 realignment languages. This demonstrates

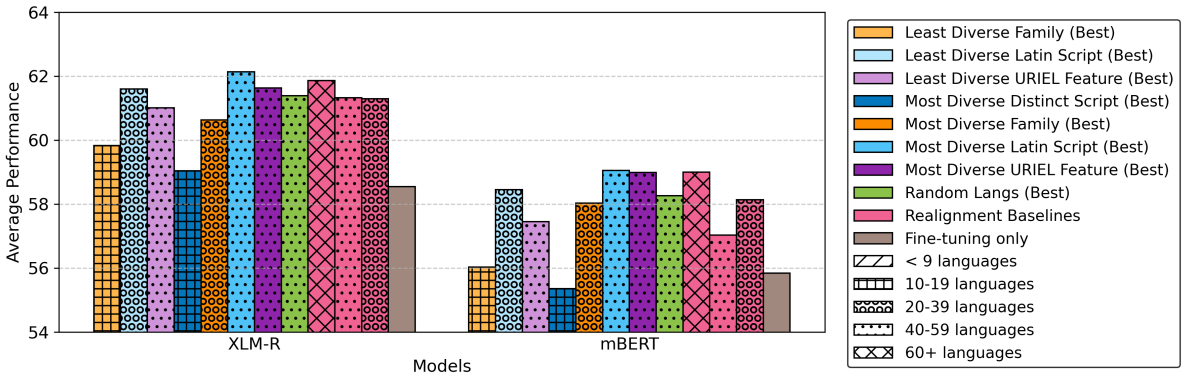

Figure 2: Average performance across PoS Tagging, NER, and NLI for XLM-R and mBERT. The baselines are compared against the best-performing configuration from each language subset heuristic.

that carefully selecting a smaller but linguistically diverse subset of languages can be as effective as, or even better than, using all languages. These two heuristics also outperform other baselines, including random selection, XTREME-R only, and African-only subsets.

Our results further highlight, across all heuristics, the least diverse subsets consistently underperform compared to their more diverse counterparts. Language subsets selected to maximize diversity in featural space outperform those that minimize such diversity across both models. Likewise, selecting languages from distinct families offers clear benefits over limiting realignment to a single family. These results show that realignment benefits from the inclusion of languages that provide diverse linguistic signals, probably because such signals help to anchor multilingual representations more robustly.

Finally, diversity affects realignment performance in different ways depending on the dimension of diversity considered. For example, diversity based on genetic lineage does not yield strong results, while selecting languages with distinct scripts, excluding Latin, produces the worst performance. This suggests that script diversity can be beneficial, but the absence of Latin script hurts alignment performance, likely due to English being the pre-training language.

4.2 Results Based on Language Resource Level

To assess how different realignment methods impact different languages in terms of the level of resources, we categorize the evaluation languages into four groups: high-resource languages (HRLs), medium-resource languages (MRLs), low-resource languages (LRLs) that are seen during pre-training, and LRLs that are unseen during pre-training[3]. Figure 3 provides the detailed breakdown of overall performance for XLM-R and mBERT across different subsets of evaluation languages.

Realignment yields substantial gains for LRLs, particularly for languages unseen during pre-training. For both models, the best realignment configuration improves LRL-unseen performance by up to 10 points over standard fine-tuning. These results demonstrate that representation alignment is especially effective when cross-lingual transfer is weakest.

For HRLs and MRLs, the trends differ. Fine-tuning alone remains competitive on these languages, and applying realignment leads to slight performance drops. This pattern is consistent with prior findings that realignment benefits do not always extend to higher-resource languages (Wu and Dredze, 2020; Gaschi et al., 2023), which have stronger initial cross-lingual representations.

We also compare different strategies for selecting realignment languages. While language choice has little influence on HRLs or MRLs, it noticeably affects LRL performance. The advantage of diversity-based over random selection observed in aggregate results primarily stems from improvements on LRLs.

Overall, these results highlight a key insight: even though realignment may offer limited gains for HRLs and MRLs, it provides consistent and substantial improvements for LRLs, especially those absent from pre-training. This makes realignment a promising direction for extending multilingual

[3]HRLs = Joshi class 5, MRLs = 3 and 4, LRLs = 0, 1, and 2 (Joshi et al., 2020)

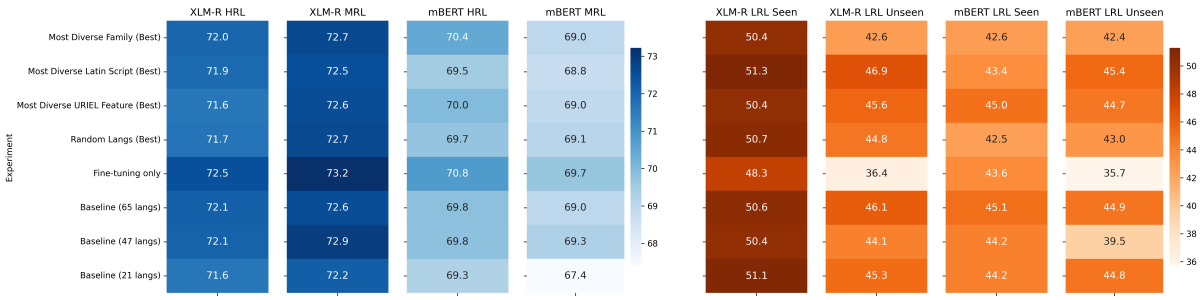

Figure 3: Heatmaps showing overall performance (averaged across four seeds) for different language subsets - HRLs, MRLs, and LRLs - seen and unseen during pre-training of XLM-R and mBERT. The fine-tuning only baseline remains strong for HRLs and MRLs, while realignment significantly improves performance on LRLs. Diversity-based language selection further amplifies these gains for LRLs.

encoders to truly underrepresented languages.

4.3 Results on Out-of-Distribution Languages

To complement our in-distribution analysis, we further evaluate different realignment approaches on AmericasNLI, which contains LRLs that were not used for realignment.

Figure 4 shows that the results on LRLs unseen during realignment do not differ much from languages used for realignment. Similarly to Figure 2, realignment significantly outperforms the fine-tuning only baseline. Diversity-based language selection outperforms the random baseline, and their homegenous counterparts, with the exception of maximizing URIEL diversity within Latin-scripted languages, which suggests that diversity should be enforced in all aspects (featural, script, and family).

One key difference with in-distribution results is that diversity-based selection, namely when using URIEL features, outperforms realignment on the entire set of available languages. Thus, when it comes to improving results across the board, including languages unseen during realignment, diversity might become more important than the number of languages involved.

As shown in Figure 4, realignment again substantially outperforms the fine-tuning baseline, mirroring the in-distribution trends. Between language subsets used for realignment, diversity-based selection continues to outperform both random selection and homogeneous subsets, with one exception: maximizing URIEL diversity within Latin-script languages does not provide the same advantage, suggesting that meaningful diversity must span features, scripts, and families rather than being constrained to a single script group. Furthermore, URIEL-based selection outperforms realignment on the full set of languages, demonstrating that when the goal is broad cross-lingual improvement, including languages never seen during realignment, the type of diversity in the realignment set matters more than the number of languages it contains.

5 Language-Scaling Behavior of Realignment Methods

Figure 5 shows how performance changes as we scale the number of languages used in each subset-selection strategy for realignment, while keeping the total computational budget fixed.

Across the board, every realignment strategy improves over simple fine-tuning, even with only five languages, indicating that cross-lingual realignment is beneficial even at very small scales. Among selection strategies, subsets based on distinct families or distinct scripts generally lag behind random sampling, whereas URIEL-diverse and diverse Latin-script language subsets provide stronger gains. Interestingly, the diverse Latin-script language subsets exhibit a non-monotonic trajectory, dipping from 10 to 20 languages before rising again at larger scales, suggesting that mid-scale expansions can occasionally introduce detrimental interactions before recovering.

For XLM-R, most strategies plateau around 20 languages, implying that the model absorbs most of the transferable signal once moderate coverage is reached. The diverse Latin-script language selection strategy is an exception, since the performance increases again at 40 languages and reversing its earlier dip. This suggests that additional gains still exist at large scales for strategies other than Latin-script diversity, but other optimal subsets may exist given a different language selection strategy.

For mBERT, the scaling behavior is more grad-

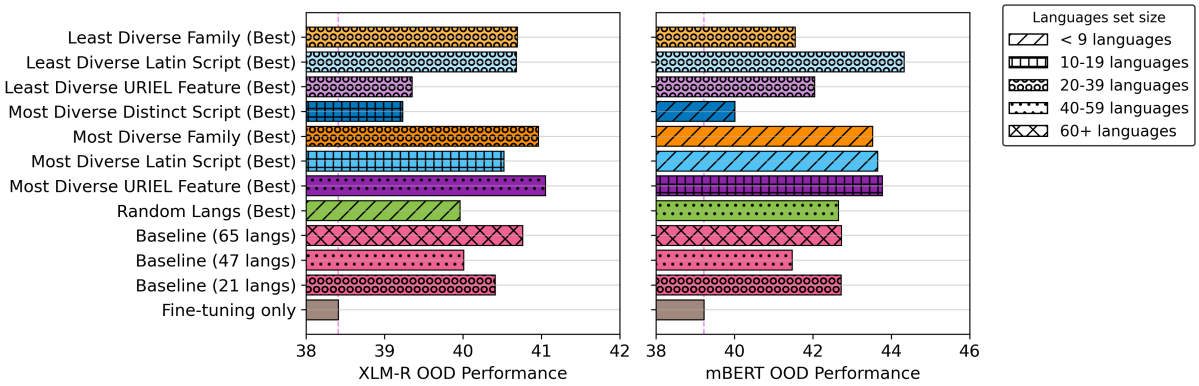

Figure 4: Averaged out-of-distribution performance of XLM-R and mBERT on the AmericasNLI dataset, comparing different language selection heuristics against three realignment baselines and a fine-tuning-only baseline. Realignment with diversity-based language subsets outperforms both the realignment and fine-tuning-only baselines.

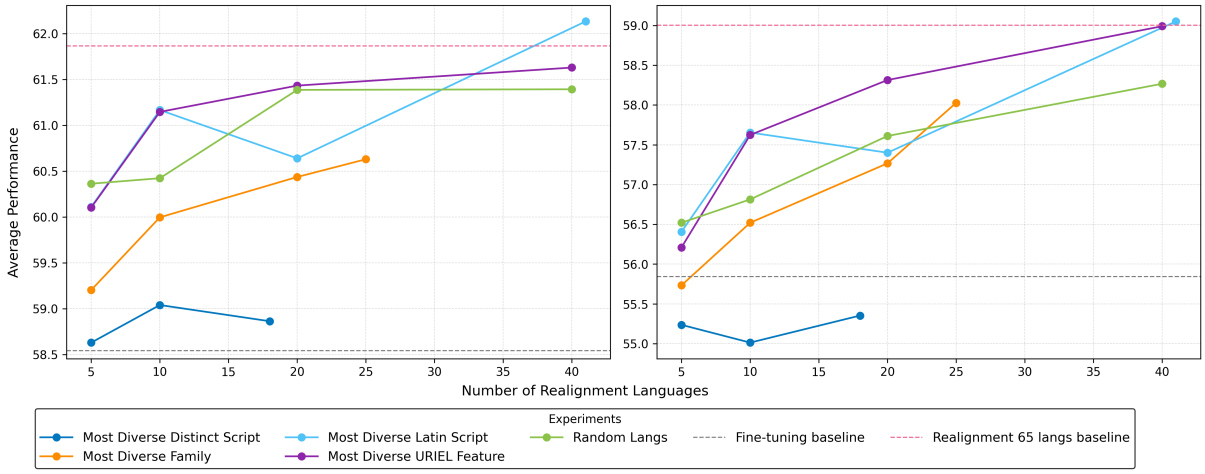

Figure 5: Scaling of average cross-lingual transfer performance with the number of languages used for realignment for XLM-R (left) and mBERT (right).

ual and nearly linear. Several strategies, such as URIEL-based and diverse Latin-script language selection, even surpass the full 65-language baseline at intermediate scales. This indicates that mBERT continues to benefit from expanded cross-lingual supervision over a wider range than XLM-R, and that its saturation point occurs later.

Overall, realignment is consistently beneficial, and that URIEL-based diversity and diverse Latin-script language selection are the most reliable and data-efficient approaches across different number of languages. We also find that different models follow distinct scaling dynamics: XLM-R saturates early, whereas mBERT accumulates gains more steadily across larger language sets.

6 Ablation study

In this ablation study, we focus specifically on analyzing the impact of including languages of different resource levels in the realignment mix. Specifically, we consider the case where only limited resources are available to collect high-quality parallel data for realignment. Our goal is to to determine whether, under such constraints, lower-resource or unseen languages remain necessary for improving cross-lingual transfer, or if higher-resource languages can serve as effective substitutes. For the sake of clarity, rather than reporting results for all heuristics, we include only the random selection heuristic, alongside two baselines: fine-tuning only, and realignment using the entire L_{65} set. Random language selection helps isolate the effect of different language pools on realignment performance, which is the main objective of this ablation study. Results for other selection heuristics can be found in Tables 9 and 11 in the Appendix.

We compare performance when randomly sampling 10 languages from different pools for realign-

Language Pool	POS	NLI	NER	Avg.
XLM-R				
Joshi 4 and 5	67.5	58.6	54.4	60.2
Joshi 3, 4, and 5	67.2	58.8	53.8	59.9
Joshi 3	67.0	58.9	51.0	59.0
Joshi 2	68.8	**59.9**	**54.5**	**61.1**
Unseen Languages	**69.1**	59.8	53.5	60.8
Seen Languages	67.1	58.8	53.3	59.7
Fine-tuning only	66.0	58.6	51.1	58.6
65 langs baseline	69.1	59.4	57.1	61.9
mBERT				
Joshi 4 and 5	63.7	53.3	50.7	55.9
Joshi 3, 4, and 5	63.5	53.3	51.5	56.1
Joshi 3	62.9	53.2	50.3	55.5
Joshi 2	**65.1**	**55.0**	**52.5**	**57.5**
Unseen Languages	64.5	54.4	**52.5**	57.1
Seen Languages	63.7	53.4	51.2	56.1
Fine-tuning only	62.2	53.1	52.2	55.8
65 langs baseline	66.9	55.4	54.8	59.0

Table 1: Performance of different realignment strategies for XLM-R and mBERT under a 10-language constraint. Only the random language selection strategy is shown. Bold indicates the highest result per task and model (excluding baselines). Standard deviation and results for other selection strategies are shown in Tables 9 and 11 in Appendix.

ment: higher-resource languages only (Joshi 4 and 5), MRLs only (Joshi 3), a mixed set of MRLs and HRLs (Joshi 3, 4 and 5), and languages either seen or unseen during pre-training. Our results from Table 1 show that in most cases, performing realignment on only 10 languages leads to improved cross-lingual transfer performance across tasks compared to fine-tuning alone. While using a reduced set of 10 languages for realignment does result in a performance drop relative to the 65-language baseline, the decrease is modest, ranging from 0.8% to 1.5%. This is a reasonable trade-off given the over sixfold reduction in the number of languages involved. On the other hand, comparisons among the ablation experiments reveal that language pools composed of Joshi Class 2 languages and unseen pretraining languages tend to yield better performance than other configurations. This highlights the importance of including LRLs and unseen languages in the realignment process to improve transfer on these same categories.

Our ablation results also indicate that other language pools - such as mid- to high-resource languages from Joshi Classes 4–5 in the case of XLM-R, as well as seen pretraining languages - can serve as practical substitutes for LRLs when the latter are not applicable. Although there is a performance drop, it remains relatively minor (less than 1%), while the availability of high-quality parallel data from these higher-resource language pools is considerably more likely.

7 Related Work

Realignment Strategies Realignment typically involves two components: the *alignment tool*, which identifies word correspondences between languages, and the *training strategy*, which updates model parameters to enforce alignment (Hämmerl et al., 2024). Different alignment tools can be used, such as the statistical FastAlign (Dyer et al., 2013) and the neural AwesomeAlign (Dou and Neubig, 2021). For our training strategy, we adopt the method proposed by Wu and Dredze (2020), which performs realignment to all layers by using a contrastive loss to word-level alignment pairs extracted from parallel corpora. Alternative training strategies include contrastive frameworks with different loss formulations (Chen et al., 2020), or architectural choices such as selectively realigning specific model layers (Bakos et al., 2025).

Data Selection Strategies Multiple works across machine learning research have been able to show that strategic data selection, rather than using all available data, can lead to better generalization, efficiency, and robustness (Wang and Neubig, 2019; Albalak et al., 2024; Liu et al., 2024; Anugraha et al., 2025b). In the context of cross-lingual transfer, prior work has shown that linguistic similarity between source and target languages is a strong predictor of transfer success, often outperforming naive or pivot-based strategies (Duong et al., 2015; Eronen et al., 2023). However, the optimal set of source languages depends heavily on the task, due to divergences in features like morphology, syntax, and script (Philippy et al., 2023). In contrast to methods like LangRank (Lin et al., 2019), which identify the best single transfer language per target language and downstream task, our work seeks to identify combinations of languages that optimize the average downstream performance across multiple target languages.

8 Conclusion

In this paper, we investigate whether realigning multilingual models with carefully selected language subsets can match or even surpass alignment using the full language set using linguistic-motivated heuristics. Our large-scale experiments demonstrated that realignment with a smaller language subsets often match the full set across models and tasks, especially when chosen based on their linguistic diversity, and evaluated on out-of-distribution languages. Moreover, our analysis shows that realignment most benefits LRLs, suggesting that realignment is particularly effective for languages whose embeddings are not yet well aligned. Our ablation studies further reveal that when the number of languages is limited, having LRLs in the language subset yields the strongest improvements, although HRLs and MRLs can still help enhance cross-lingual transfer. These results demonstrate that strategic language selection not only reduces computational and data overhead but can also strengthen multilingual generalization, pointing toward more efficient and targeted approaches to cross-lingual realignment.

Limitations

In this paper, we demonstrate the importance of linguistic diversity as a more inclusive and effective approach to improving cross-lingual generalization in multilingual language models. Rather than simply collecting all available data, our work shows that carefully selecting diverse subsets of languages can enhance cross-lingual transfer, indicating that our work is both more efficient in terms of data collection overhead and more effective overall.

A limitation of this study is the absence decoder-only models, which are increasingly prevalent in current research. While the Appendix Section A.3 presents our tentative realignment results on Llama 3.1 8B using LoRA adapters, these results indicate that the same realignment method does not straightforwardly transfer to decoder-only architectures, highlighting the need for careful adaptation in such settings. Moreover, encoder-only architectures remain highly relevant for cross-lingual classification due to their efficiency, stable representations, and strong transfer capabilities. Recent developments, such as ModernBERT (Warner et al., 2025) and LLM2Vec (BehnamGhader et al., 2024), which adapt decoder-only models into encoder-style architectures, further highlight their enduring importance. As shown in our small-scale experiment in Section A.4, encoder-based models can even outperform much larger multilingual decoder-only models on several classification tasks. Future work could also explore applying our algorithm-agnostic heuristics to decoder-only model fine-tuning for multilingual tasks (Anugraha et al., 2025a) or to reinforcement learning setups in multilingual contexts (Dang et al., 2024).

Another is that although averaging does not explicitly target word-level alignment, we empirically find that it provides results that are slightly lower but still comparable to FastAlign (Appendix A.2). Therefore, we opt for the averaging approach for practical reasons, as FastAlign is significantly more resource- and time-intensive.

Our exploration is also limited to heuristic-based language selection, although we have shown that such subsets do exist. A promising direction for future work is to move beyond heuristics by developing predictive algorithms that estimate downstream performance and dynamically determine both the optimal languages and the appropriate subset size (Anugraha et al., 2024).

We acknowledge that our full realignment language set, L_{65}, does not cover the entire spectrum of global linguistic diversity, despite covering many LRLs. We hope that our approach encourages the creation of parallel corpora for underrepresented languages, enabling greater diversity in alignment sets and fostering more inclusive multilingual models. Ultimately, we aim for our work to contribute toward broader and fairer access to language technologies, especially in the context of cross-lingual NLP research and deployment.

Acknowledgements

This research on "Multilingual multicultural NLP and LLMs" was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2024-06887) and Discovery Launch Supplement (DGECR-2024-00008). The authors also acknowledge the computational resources and support provided by the Digital Research Alliance of Canada (formerly Compute Canada) through grant RRG no. 5397.

This project was provided with HPC computing and storage resources by GENCI at IDRIS thanks to the grant 2025-AD010316268 on the supercomputer Jean Zay's H100 partition.

References

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, and 26 others. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, and 1 others. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, and 1 others. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

David Anugraha, Shou-Yi Hung, Zilu Tang, Annie En-Shiun Lee, Derry Tanti Wijaya, and Genta Indra Winata. 2025a. mr3: Multilingual rubric-agnostic reward reasoning models. *arXiv preprint arXiv:2510.01146*.

David Anugraha, Zilu Tang, Lester James V Miranda, Hanyang Zhao, Mohammad Rifqi Farhansyah, Garry Kuwanto, Derry Wijaya, and Genta Indra Winata. 2025b. R3: Robust rubric-agnostic reward models. *arXiv preprint arXiv:2505.13388*.

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, and En-Shiun Annie Lee. 2024. Proxylm: Predicting language model performance on multilingual tasks via proxy models. *arXiv preprint arXiv:2406.09334*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the ACL 2019*.

Steve Bakos, David Guzmán, Riddhi More, Kelly Chutong Li, Félix Gaschi, and En-Shiun Annie Lee. 2025. AlignFreeze: Navigating the impact of realignment on the layers of multilingual models across diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 562–586, Albuquerque, New Mexico. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. In *Transactions of the Association of Computational Linguistics*.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe

Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. *arXiv preprint arXiv:2407.02552*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, and 25 others. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China. Association for Computational Linguistics.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza-Ruiz, and 1 others. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. The impact of cross-lingual adjustment of contextual word representations on zero-shot transfer. In *European Conference on Information Retrieval*, pages 51–67. Springer.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. *arXiv preprint arXiv:2306.02790*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

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

Aung Kyaw Htet and Mark Dras. 2025a. Myanmar xnli: Building a dataset and exploring low-resource approaches to natural language inference with myanmar. *arXiv preprint arXiv:2504.09645*.

Aung Kyaw Htet and Mark Dras. 2025b. Myanmar xnli: Building a dataset and exploring low-resource approaches to natural language inference with myanmar. *Preprint*, arXiv:2504.09645.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of ACL 2020*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2024. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. *arXiv preprint arXiv:2105.03953*.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

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

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248. University of Helsinki.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. *arXiv preprint arXiv:1905.08212*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? *arXiv preprint arXiv:2010.02537*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

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

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

A Detailed Methodology and Experimental Setup

A.1 Language Featural Diversity Calculation

Let $L = \{\ell_1, \ell_2, \ldots, \ell_m\}$ be the list of all available languages, vec(ℓ) denote the URIEL featural vector of language ℓ, and let $d(u, v)$ denote the angular distance between vectors $u, v \in \mathbb{R}^d$. The angular distance is defined as:

$$d(u, v) = \frac{\arccos\left(\frac{u \cdot v}{\|u\| \, \|v\|}\right)}{\pi}$$

where:

- $u \cdot v$ is the dot product of u and v,
- $\|u\|$ is the Euclidean norm of vector u,
- $d(u, v) \in [0, 1]$ (normalized angle between vectors).

In order to find the most diverse subset size n, we try to maximize the total pairwise angular distance:

$$S^* = \underset{S \subseteq L, |S|=n}{\arg\max} \sum_{\{\ell_i, \ell_j\} \in \binom{S}{2}} d\left(\text{vec}(\ell_i), \text{vec}(\ell_j)\right)$$

Similarly, to find the least diverse subset size n, we try to minimize the total pairwise angular distance:

$$S^* = \underset{S \subseteq L, |S|=n}{\arg\min} \sum_{\{\ell_i, \ell_j\} \in \binom{S}{2}} d\left(\text{vec}(\ell_i), \text{vec}(\ell_j)\right)$$

A.2 Realignment using average words' representations vs FastAlign

Model	Task	Avg. Tokens	FastAlign
mBERT	NER	54.89 ± 0.93	**55.39 ± 0.52**
	NLI	55.34 ± 0.28	**56.58 ± 0.25**
	POS	66.85 ± 0.24	**70.31 ± 0.15**
XLM-R	NER	**57.19 ± 0.79**	56.72 ± 1.62
	NLI	59.5 ± 0.27	**60.65 ± 0.33**
	POS	69.1 ± 0.22	**71.29 ± 0.25**
Avg. Time (h)		**0.35 ± 0.003**	1.93 ± 0.039

Table 2: Comparison of Average and FastAlign performance across 5 random seeds. Avg. Time reflects the average time taken by realignment step only across all tasks (in hours); each seed reused the same realignment checkpoint.

Realignment methods introduced before this work take a batch of pairs of translated sentences, extract pairs of corresponding words across those pairs using alignment tools like FastAlign and AwesomeAlign, and compute an in-batch contrastive loss on those pairs of words. Our work introduces a simple "averaging trick": instead of computing the contrastive loss on word pairs extracted with an aligner, we compute the average of all tokens in a sentence and align sentences instead of words. This change is not made to improve cross-lingual transfer but rather to alleviate the need for a word aligner, which considerably reduces the time necessary to perform realignment.

We compare the time efficiency and performance of the two realignment methods using a different GPU type from that used in our main experiments, with the results presented in Table 2. Overall, FastAlign performs slightly better than the token averaging method on most tasks. However, we adopt the averaging approach for practical reasons, as FastAlign is considerably more resource- and time-intensive - even without accounting for the additional preprocessing required to prepare datasets for FastAlign across 65 languages - making it unsuitable for large-scale experiments. It is important to emphasize that we employ the averaging method **not to enhance the baseline performance**, but rather **to reduce computational overhead**, thereby enabling large-scale comparisons across different language selection strategies.

A.3 Tentative Decoder-only Results

Table 3: Comparison of Llama model performance with and without realignment on downstream tasks.

Metric	Llama w/o realignment	Llama w/ realignment
PoS target acc.	38.8	**41.9**
NER target F1	**31.6**	30.2
NLI accuracy	56.5	**57.8**

We experimented with Llama 3.1 8B using LoRA adapters to overcome computational limitations. The adapters were trained for realignment and then fine-tuned for the downstream task, following a similar procedure as for encoder-only models. The results are shown in Table 3.

While realignment improved PoS tagging and NLI, we observed that it negatively impacted NER performance for LLaMA, which suggests that **realignment does not transfer straightforwardly**

to decoder-only architectures.. Our preliminary experiments are not conclusive on whether realignment could ultimately benefit decoder-only models. We plan to investigate this further in future work.

More broadly, applying realignment to decoder-only LLMs for generative tasks raises unique challenges. In particular, **making such models entirely language-agnostic could exacerbate issues such as language confusion** (Marchisio et al., 2024), where the model generates text in an unintended language. This highlights an important avenue for future work and motivates careful consideration of how realignment should be adapted for decoder-only settings.

A.4 Encoder-only models vs Decoder-only models on cross-lingual transfer classification

Task	XLM-R	Llama 3.1	Gemma 2
POS en	**96.2**	90.9	93.5
POS XL	**62.0**	38.8	49.1
NER en	**82.0**	72.5	73.9
NER XL	**49.1**	31.6	35.9
NLI en	83.5	90.2	**91.3**
NLI XL	54.6	**56.5**	55.0

Table 4: Performance comparison of XLM-R, Llama 3.1 8B, and Gemma 2 9B across various three downstream tasks: POS Tagging, NER and NLI under the same experimental settings. **XL** indicates cross-lingual.

We fine-tuned LLaMA 3.1 8B on PoS, NER, and NLI under the same setup. Our results on Table 4 show that XLM-R (encoder-only) not only significantly outperforms Llama and Gemma on cross-lingual transfer for some tasks, but can even surpass them in English. Interestingly, this seems to be true for word-level tasks (POS and NER), but not sentence-level ones (NLI). This finding aligns with prior work suggesting that small fine-tuned encoder-only models often outperform prompted LLMs on classification tasks (Ahuja et al., 2023).

In conclusion, adapting realignment methods to encoder-only models and generative tasks is a modern and exciting direction for future research. Nevertheless, we focus on encoder-only models, which remain a relevant contribution for cross-lingual classification, especially for multilinguality.

A.5 Languages

Table 5 contains the full list of the 65 languages used in Gaschi et al. (2023). Table 6 contains the list of languages used in each experiment.

Table 6 shows the languages chosen within each experiment. All experiments are run with seeds of 17, 23, 42, and 66, including the selection of languages in the Random Subsets (Seeded) setting.

Figures 6, 7, and 8 show the language trees of the 65 languages used.

A.6 Hyperparameters and Resources

For both tasks, we follow the experimental setup used in Gaschi et al. (2023); Bakos et al. (2025). All experiments are conducted using NVIDIA H100 GPUs and run with 4 random seeds to account for variability. Realignment is performed for 24,544 steps. This is followed by task-specific fine-tuning: 5 epochs for PoS tagging and 2 epochs for NLI. We use a learning rate of `2e-5`, a batch size of 32 for both training and evaluation, and a maximum input length of 200 tokens for source and target sequences. During the realignment stage, we use a reduced maximum sequence length of 96 and a smaller batch size of 16.

A.7 Statistics about the datasets used

The size of the datasets used for training and evaluating are reported in Table 7.

A.8 Licenses for artifacts used

Below is a list of the datasets under study:

- The XNLI corpus (Conneau et al., 2018) has the CC BY-NC 4.0 license.
- The AfriXNLI dataset (Adelani et al., 2024) has the Apache 2.0 license.
- The IndoNLI dataset (Mahendra et al., 2021) has the CC-BY-SA 4.0 license.
- The Myanmar-XNLI dataset (Htet and Dras, 2025b) has the Apache 2.0 license.
- The UDPOS dataset (De Marneffe et al., 2021) has the CC0-1.0 license.
- The MasakhaPOS dataset (Dione et al., 2023) has the MIT license.
- The WikiANN dataset (Pan et al., 2017) has the Apache 2.0 2.0 license.

- The MasakhaNER 2.0 dataset (Adelani et al., 2022) has the AFL 3.0 license.
- The OPUS-100 dataset (Zhang et al., 2020) has no explicit license; it is a filtered subset of OPUS (Tiedemann, 2009), which aggregates translation corpora that is generally considered redistributable.
- The NLLB dataset (Team et al., 2022) has the ODC-By license.
- The XTREME-R benchmark suite (Ruder et al., 2021) does not have a unified license; it aggregates multiple datasets, each with its own license or terms of use:
  - The XNLI corpus (Conneau et al., 2018) has the CC BY-NC 4.0 license.
  - The PAWS-X dataset (Yang et al., 2019) is free to use for any purpose.
  - The UDPOS dataset (De Marneffe et al., 2021) has the CC0-1.0 license.
  - The WikiANN dataset (Pan et al., 2017) has the Apache 2.0 2.0 license.
  - The XQuAD dataset (Artetxe et al., 2020) has the CC BY-SA 4.0 license.
  - The MLQA dataset (Lewis et al., 2020) has the CC BY-SA 3.0 license.
  - The TyDiQA-GoldP dataset (Clark et al., 2020) has the Apache 2.0 license.
  - The BUCC 2018 dataset for the shared task on bitext mining (Zweigenbaum et al., 2018) is available for academic research use only; redistribution may be restricted.
  - The Tatoeba dataset (Artetxe and Schwenk, 2019) has the CC BY 2.0 license.

Below is a list of the other artifacts under study:

- The code for realignment comes from Gaschi et al. (2023) and has the MIT license.
- The URIEL+ knowledge base and distance calculation functions (Khan et al., 2025) have the CC BY-SA 4.0 license.
- The weights of XLM-R Base (Conneau et al., 2020) have the MIT license.
- The weights of mBERT Base (Devlin et al., 2019) have the Apache 2.0 license.

All artifacts were thus used in accordance with their open-source or non-commercial licenses.

A.9 Use of AI

For the writing of this paper, AI was solely used to reformulate some text, and as an autocompletion tool for writing the code used in the experiments.

B More Detailed Results

This section contains the full results of the experiments in this paper.

Language	Language Code‡	Script	Language Family	Joshi Class†	Vitality
Afrikaans	afr	Latin	Germanic	3	MRL
Akan	twi	Latin	Atlantic-Congo	1	LRL
Amharic	amh	Amharic	Semitic	2	LRL
Arabic	ara	Arabic	Semitic	5	HRL
Azerbaijani	aze	Latin	Oghuz	1	LRL
Bambara	bam	Latin	Mande	1	LRL
Basque	eus	Latin	N/A	4	MRL
Bengali	ben	Bengali	Indo-Iranian	3	MRL
Bulgarian	bul	Cyrillic	Balto-Slavic	3	MRL
Burmese	mya	Burmese	Sino-Tibetan	1	LRL
Chinese	zho	Chinese	Sino-Tibetan	5	HRL
Dholuo	luo	Latin	Nilo-Saharan	0	LRL
Dutch	nld	Latin	Germanic	4	MRL
Eastern Punjabi	pan	Gurmukhi	Indo-Iranian	2	LRL
Estonian	est	Latin	Finnic	3	MRL
Ewe	ewe	Latin	Atlantic-Congo	1	LRL
Finnish	fin	Latin	Finnic	4	MRL
Fon	fon	Latin	Atlantic-Congo	0	LRL
French	fra	Latin	Romance	5	HRL
Ganda	lug	Latin	Atlantic-Congo	1	LRL
Georgian	kat	Georgian	Kartvelian	3	MRL
German	deu	Latin	Germanic	5	HRL
Greek	ell	Greek	Hellenic	3	MRL
Gujarati	guj	Gujarati	Indo-Iranian	1	LRL
Hausa	hau	Latin	Chadic	2	LRL
Hebrew	heb	Hebrew	Semitic	3	MRL
Hindi	hin	Devanagari	Indo-Iranian	4	MRL
Hungarian	hun	Latin	Ugric	4	MRL
Igbo	ibo	Latin	Atlantic-Congo	1	LRL
Indonesian	ind	Latin	Malayic	3	MRL
Italian	ita	Latin	Romance	4	MRL
Japanese	jpn	Japanese	Japonic	5	HRL
Javanese	jav	Latin	Javanic	1	LRL
Kazakh	kaz	Cyrillic	Kipchak	3	MRL
Kinyarwanda	kin	Latin	Atlantic-Congo	1	LRL
Korean	kor	Korean	Koreanic	4	MRL
Lingala	lin	Latin	Atlantic-Congo	1	LRL
Lithuanian	lit	Latin	Balto-Slavic	3	MRL
Malay	msa	Latin	Malayic	3	MRL
Malayalam	mal	Malayalam	Southern Dravidian	1	LRL
Marathi	mar	Devanagari	Indo-Iranian	2	LRL
Mossi	mos	Latin	Gur	0	LRL
Nyanja	nya	Latin	Benue-Congo	1	LRL
Oromo	orm	Latin	Cushitic	1	LRL
Persian	fas	Arabic	Indo-Iranian	4	MRL
Polish	pol	Latin	Balto-Slavic	4	MRL
Portuguese	por	Latin	Romance	4	MRL
Romanian	ron	Latin	Romance	3	MRL
Russian	rus	Cyrillic	Balto-Slavic	4	MRL

Shona	sna	Latin	Atlantic-Congo	1	LRL
Spanish	spa	Latin	Romance	5	HRL
Swahili	swa	Latin	Atlantic-Congo	2	LRL
Tagalog	tgl	Latin	Philippine	3	MRL
Tamil	tam	Tamil	Southern Dravidian	3	MRL
Telugu	tel	Telugu	South-Central Dravidian	1	LRL
Thai	tha	Thai	Kra–Dai	3	MRL
Tswana	tsn	Latin	Atlantic-Congo	2	LRL
Turkish	tur	Latin	Oghuz	4	MRL
Ukrainian	ukr	Cyrillic	Balto-Slavic	3	MRL
Urdu	urd	Arabic	Indo-Iranian	3	MRL
Vietnamese	vie	Latin	Austroasiatic	4	MRL
Wolof	wol	Latin	Atlantic-Congo	2	LRL
Xhosa	xho	Latin	Atlantic-Congo	2	LRL
Yoruba	yor	Latin	Atlantic-Congo	2	LRL
Zulu	zul	Latin	Atlantic-Congo	2	LRL

Table 5: The 65 languages used for the realignment phase with their vitality class mapping. The language codes follow [‡]ISO639-3 coding. Languages are mapped to their [†]rarity taxonomy based on Joshi et al. (2020) vitality classes: Low Resource Language (LRL, 0-2), Medium Resource Language (MRL, 3-4), and High Resource Language (HRL, 5).

Method	#	Languages‡
Baseline		
All 65 languages	65	afr, amh, ara, aze, bul, ben, deu, ell, spa, est, eus, fas, fin, fra, guj, heb, hin, hun, ind, ita, jpn, kat, kaz, kor, lit, mal, mar, msa, mya, nld, pan, pol, por, ron, rus, tam, tha, tur, ukr, urd, vie, zho, bam, ewe, fon, hau, ibo, kin, lin, lug, luo, mos, nya, gaz, sna, swh, tsn, twi, wol, xho, yor, zul, jav, tgl, tel
Present in XTREME-R	47	afr, ara, aze, bul, ben, deu, ell, spa, est, eus, fas, fin, fra, guj, heb, hin, hun, ind, ita, jpn, jav, kat, kaz, kor, lit, mar, mal, msa, mya, nld, pan, pol, por, ron, rus, swh, tam, tel, tgl, tha, tur, ukr, urd, vie, wol, yor, zho
Present in Africa	21	amh, bam, ewe, fon, hau, ibo, kin, lin, lug, luo, mos, nya, gaz, sna, swh, tsn, twi, wol, xho, yor, zul
Featural Diversity		
Most diverse from English	5	fon, kat, kaz, lin, gaz
	10	afr, ara, fon, kat, jpn, kaz, lin, gaz, sna, vie
	20	afr, ara, aze, eus, zho, fon, lug, kat, ell, hau, heb, ibo, jpn, kaz, kor, lin, luo, gaz, sna, twi, vie, yor
	40	afr, ara, aze, eus, mya, zho, ewe, fon, fra, lug, kat, ell, hau, heb ibo, jpn, kaz, kin, kor, lin, msa, mal, mar, nya, gaz, fas, rus, sna spa, tgl, tam, tel, tha, tur, twi, urd, vie, xho, yor, zul
Least diverse from English	5	ita, por, ron, spa, ukr
	10	bul, deu, ell, spa, fra, ita, nld, por, ron, ukr
	20	bul, nld, est, fin, fra, deu, ell, guj, hin, hun, ita, lit, fas, pol por, pan, ron, rus, spa, ukr
	40	amh, ara, aze, bam, eus, ben, bul, nld, est, fin, fra, deu, ell, guj hau, heb, hin, hun, ind, ita, jav, lit, luo, mal, mar, mos, fas, pol por, pan, ron, rus, spa, tgl, tam, tel, tur, ukr, urd, wol
Phylogenetic Diversity		
Most diverse families	5	kat, kaz, lin, gaz, vie
	10	ara, zho, kat, jpn, kaz, lin, msa, gaz, tam, vie

	20	ara, aze, eus, zho, fra, kat, ell, hau, jpn, kaz, kor, lin, luo, msa, mar, gaz, rus, tam, tha, vie
	25	ara, aze, bam, eus, zho, fin, fra, kat, ell, hau, hun, jpn, kaz, kor lin, luo, msa, mar, mos, gaz, rus, tam, tel, tha, vie
Most diverse families within Indo-European	5	afr, nld, deu, ita, por
	10	afr, bul, nld, fra, deu, ita, por, ron, spa, ukr
	20	afr, ben, bul, nld, fra, deu, ell, guj, hin, ita, lit, mar, pol, por pan, ron, rus, spa, ukr, urd
Script Diversity		
Most diverse using distinct scripts	5	ara, kat, jpn, kaz, tha
	10	ara, mya, zho, kat, ell, heb, jpn, kaz, tam, tha
	18	amh, ara, ben, mya, zho, kat, ell, guj, heb, hin, jpn, kaz, kor, mal, pan, tam, tel, tha
Most diverse using Latin scripts	5	aze, fon, lin, gaz, tgl
	10	afr, aze, eus, fon, lin, gaz, sna, tgl, twi, vie
	20	bam, nld, est, fin, fra, deu, hau, hun, ind, ita, jav, lit, luo, msa, pol, por, ron, spa, tgl, wol
	41	afr, aze, bam, eus, nld, est, ewe, fin, fon, fra, lug, deu, hau, hun ibo, ind, ita, jav, kin, lin, lit, luo, msa, mos, nya, gaz, pol, por ron, sna, spa, swh, tgl, tsn, tur, twi, vie, wol, xho, yor, zul
Least diverse using Latin scripts	5	nld, fra, deu, ita, por
	10	nld, est, fin, fra, deu, hun, ita, por, ron, spa
	20	afr, aze, eus, ewe, fon, fra, lug, lin, lit, msa, gaz, pol, sna, spa tgl, tur, twi, vie, yor, zul
Ablation Languages		
Joshi Class = 2 (Random)	10	amh, aze, bam, ewe, fon, gaz, guj, hau, ibo, jav
Joshi Class = 2 (Most URIEL)	10	aze, mya, fon, kin, lin, mar, gaz, sna, tel, yor
Joshi Class = 2 (Most Family)	10	aze, mya, hau, jav, lin, luo, mar, mos, gaz, tel
Joshi Class = 2 (Most Script)	10	amh, mya, guj, lin, mal, mar, gaz, pan, tel, yor
Joshi Class = 3 (Random)	17	afr, bul, ben, ell, est, heb, ind, kat, kaz, lit, msa, ron, tam, tgl, tha, ukr, urd
Joshi Class = 3 (Most URIEL)	10	afr, kat, ell, heb, kaz, msa, tgl, tam, tha, urd
Joshi Class = 3 (Most Family)	10	est, kat, ell, heb, kaz, lit, msa, tam, tha, urd
Joshi Class = 3 (Most Script)	10	ben, bul, kat, ell, heb, kaz, tam, tha, ukr, urd
Joshi Class = 3,5 (Random)	23	afr, ara, bul, ben, deu, ell, spa, est, eus, fas, fin, fra, heb, hin, hun, ind, ita, jpn, kat, kaz, kor, lit, msa,

		nld, pan, pol, por, ron, rus, tam, tgl, tha, tur, ukr, urd, vie, zho
Joshi Class = 3,5 (Most URIEL)	10	afr, ara, eus, zho, kat, jpn, kaz, msa, tam, vie
Joshi Class = 3,5 (Most Family)	10	ara, zho, kat, ell, jpn, kaz, msa, tam, tha, vie
Joshi Class = 3,5 (Most Script)	10	ara, zho, kat, ell, heb, jpn, kaz, kor, tam, tha
Joshi Class = 4,5 (Random)	20	ara, deu, spa, eus, fas, fin, fra, hin, hun, ita, jpn, kor, nld, pol, por, rus, tam, tur, vie, zho
Joshi Class = 4,5 (Most URIEL)	10	ara, eus, zho, fra, jpn, kor, fas, rus, tur, vie
Joshi Class = 4,5 (Most Script)	10	ara, eus, fra, hin, jpn, kor, fas, rus, tur, vie
Seen by mBERT (Random)	47	afr, ara, aze, bul, ben, deu, ell, spa, est, eus, fas, fin, fra, guj, heb, hin, hun, ind, ita, jpn, jav, kat, kaz, kor, lit, mar, mal, msa, mya, nld, pan, pol, por, ron, rus, swh, tam, tel, tgl, tha, tur, ukr, urd, vie, yor, zho
Seen by mBERT (Most URIEL)	10	afr, ara, zho, kat, jpn, kaz, swh, tam, vie, yor
Seen by mBERT (Most Family)	10	ara, aze, zho, kat, jpn, kaz, msa, tam, vie, yor
Seen by mBERT (Most Script)	10	ara, mya, zho, kat, ell, heb, jpn, kaz, tam, tha
Seen by XLM-R (Random)	51	afr, amh, ara, aze, bul, ben, deu, ell, spa, est, eus, fas, fin, fra, gaz, guj, hau, heb, hin, hun, ind, ita, jpn, jav, kat, kaz, kor, lit, mar, mal, msa, mya, nld, pan, pol, por, ron, rus, swh, tam, tel, tgl, tha, tur, ukr, urd, vie, xho, zho
Seen by XLM-R (Most URIEL)	10	afr, ara, zho, kat, jpn, kaz, msa, gaz, swh, vie
Seen by XLM-R (Most Family)	10	ara, zho, kat, jpn, kaz, msa, gaz, tam, vie, xho
Seen by XLM-R (Most Script)	10	ara, mya, zho, kat, ell, heb, jpn, kaz, tam, tha
Unseen by mBERT (Random)	34	amh, bam, ewe, fon, gaz, hau, ibo, kin, lin, lug, luo, mos, nya, sna, tsn, twi, wol, xho, yor, zul
Unseen by mBERT (Most URIEL)	10	amh, ewe, fon, hau, lin, luo, gaz, sna, twi, xho
Unseen by mBERT (Most Family)	10	amh, bam, fon, hau, lin, luo, mos, gaz, sna, twi
Unseen by XLM-R (Random)	30	bam, ewe, fon, ibo, kin, lin, lug, luo, mos, nya, sna, tsn, twi, wol, yor, zul
Unseen by XLM-R (Most URIEL)	10	ewe, fon, lin, luo, mos, sna, twi, wol, yor, zul
Unseen by XLM-R (Most Family)	10	bam, fon, lin, luo, mos, sna, twi, wol, yor, zul

Table 6: List of languages used for each experiment and selection strategy. Language codes follow ‡ISO639-3 coding .

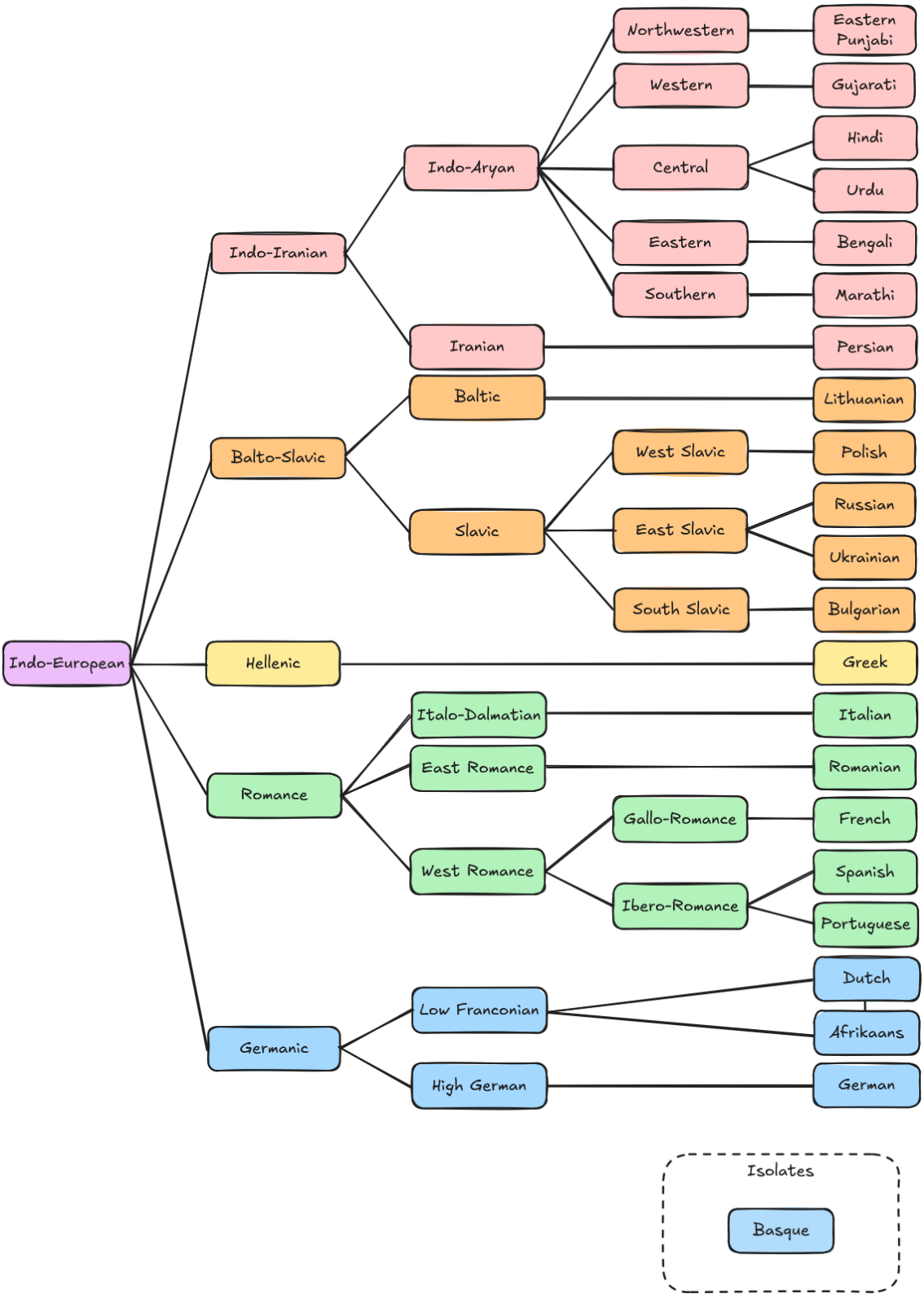

Figure 6: Phylogenetic tree for Indo-European languages (including Basque as an isolate) used in realignment.

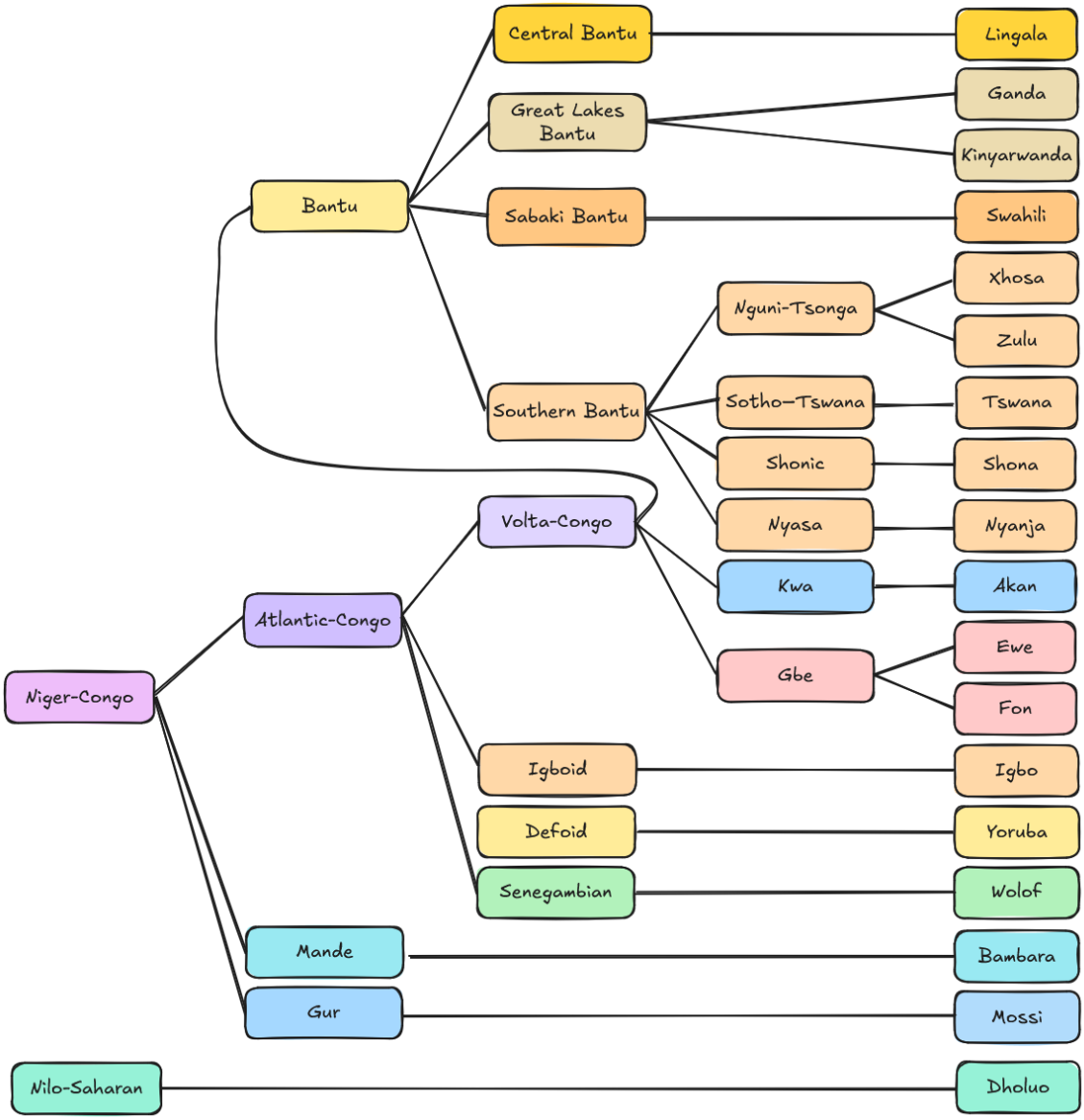

Figure 7: Phylogenetic trees for Niger-Congo languages, and Dholuo (from the Nilo-Saharan family, due to its proximity) used in realignment.

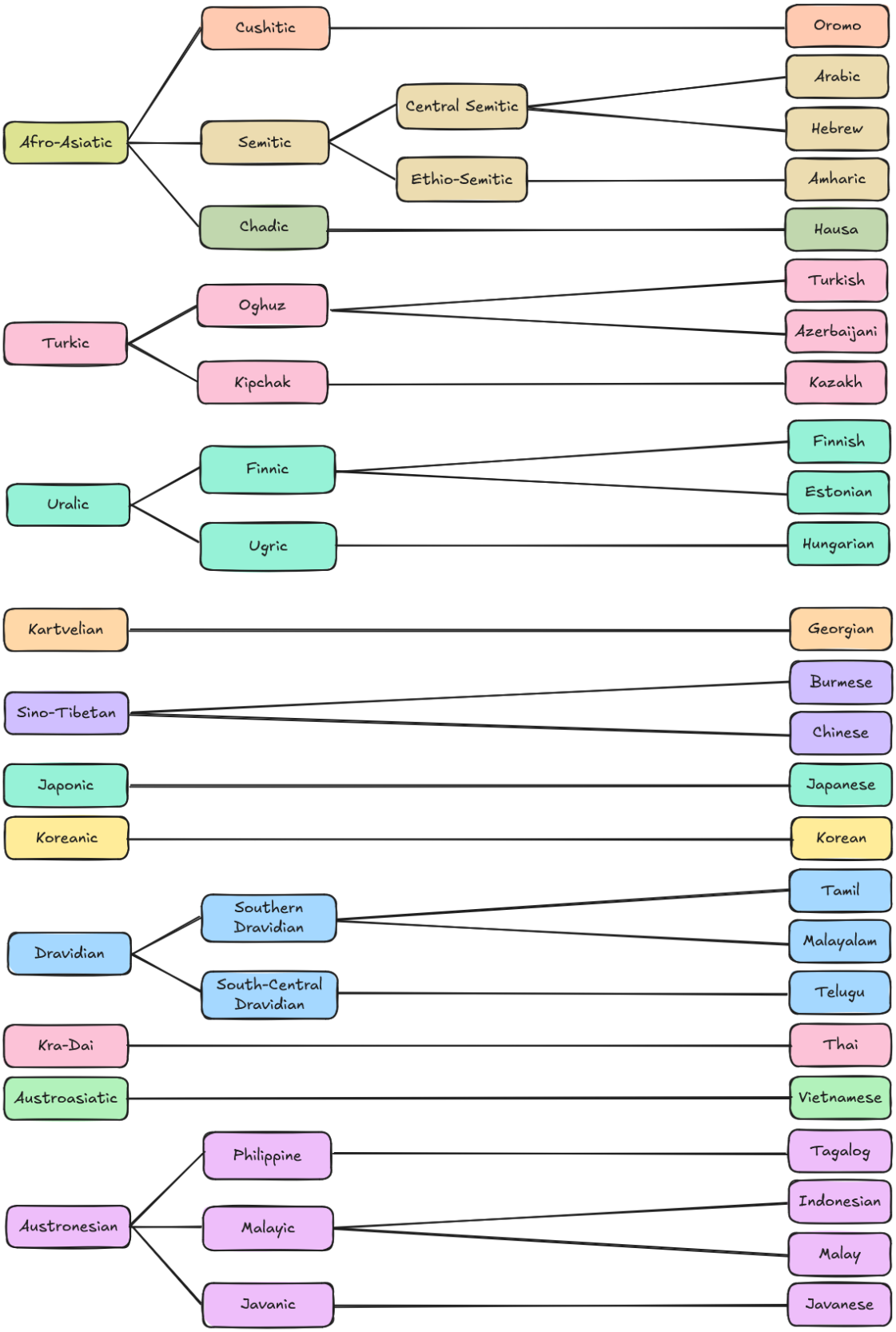

Figure 8: Residual phylogenetic trees for all other languages languages used in realignment (Non-Indo-European, Non-Niger-Congo, Non-Nilo-Saharan, Non-Basque).

Language	NLI	PoS-tagging	NER	Seen by			
				XLM-R	mBERT	OPUS-100	NLLB
English (training)	392,702	21,787	20,000	✓	✓	✓	✓
Afrikaans	-	425	1000	✓	✓	✓	✓
Arabic	5010	1680	10000	✓	✓	✓	✓
Azerbaijani	-	-	1000	✓	✓	✓	✓
Basque	-	1799	10000	✓	✓	✓	✓
Bengali	-	-	1000	✓	✓	✓	✓
Bulgarian	5010	1116	10000	✓	✓	✓	✓
Burmese	5010	-	100	✓	✓	✓	✓
Chinese	5010	3455	10000	✓	✓	✓	✓
Czech	-	10159	-	✓	✓	✓	✓
Dutch	-	1471	10000	✓	✓	✓	✓
English (evaluation)	5010	5440	10000	✓	✓	✓	✓
Estonian	-	4127	10000	✓	✓	✓	✓
Finnish	-	6544	10000	✓	✓	✓	✓
French	5010	7542	10000	✓	✓	✓	✓
Georgian	-	-	10000	✓	✓	✓	✓
German	5010	22358	10000	✓	✓	✓	✓
Greek	5010	456	10000	✓	✓	✓	✓
Gujarati	-	-	100	✓	✓	✓	✓
Hebrew	-	491	10000	✓	✓	✓	✓
Hindi	5010	2684	1000	✓	✓	✓	✓
Hungarian	-	449	10000	✓	✓	✓	✓
Indonesian	2984	1931	10000	✓	✓	✓	✓
Italian	-	3518	10000	✓	✓	✓	✓
Japanese	-	2365	10000	✓	✓	✓	✓
Javanese	-	-	100	✓	✓	✗	✓
Kazakh	-	1047	1000	✓	✓	✓	✓
Korean	-	4276	10000	✓	✓	✓	✓
Lithuanian	-	739	10000	✓	✓	✓	✓
Malay	-	-	1000	✓	✓	✓	✓
Malayalam	-	-	1000	✓	✓	✓	✓
Marathi	-	47	1000	✓	✓	✓	✓
Persian	-	2055	10000	✓	✓	✓	✓
Polish	-	4942	10000	✓	✓	✓	✓
Portuguese	-	2680	10000	✓	✓	✓	✓
Eastern Punjabi	-	-	100	✓	✓	✓	✓
Romanian	-	2272	10000	✓	✓	✓	✓
Russian	5010	8973	10000	✓	✓	✓	✓
Spanish	5010	3147	10000	✓	✓	✓	✓
Tagalog	-	-	1000	✓	✓	✗	✓
Tamil	-	654	1000	✓	✓	✓	✓
Telugu	-	146	1000	✓	✓	✓	✓
Thai	5010	1000	10000	✓	✓	✓	✓
Turkish	5010	6647	10000	✓	✓	✓	✓
Ukrainian	-	892	10000	✓	✓	✓	✓
Urdu	5010	535	1000	✓	✓	✓	✓
Vietnamese	5010	800	10000	✓	✓	✓	✓
Akan	-	628	1211	✗	✗	✗	✓

(continued)

Language	NLI	PoS-tagging	NER	XLM-R	mBERT	OPUS-100	NLLB
Amharic	600	-	-	✓	✗	✓	✓
Bambara	-	619	1274	✗	✗	✗	✓
Chichewa	-	-	1785	✗	✗	✗	✓
Dholuo	-	606	1474	✗	✗	✗	✓
Ewe	600	582	1001	✗	✗	✗	✓
Fon	-	646	1228	✗	✗	✗	✓
Ghomala	-	599	966	✗	✗	✗	✗
Hausa	600	601	1633	✓	✗	✓	✓
Igbo	600	642	2181	✗	✗	✓	✓
Kinyarwanda	600	604	2235	✗	✗	✓	✓
Lingala	600	-	-	✗	✗	✗	✓
Ganda	600	586	1412	✗	✗	✗	✓
Mossi	-	604	1294	✗	✗	✗	✓
Naija	-	-	1613	✗	✗	✗	✗
Oromo	600	-	-	✓	✗	✓	✗
Setswana	-	602	996	✗	✗	✗	✓
Shona	600	596	1773	✗	✗	✗	✓
Southern Sotho	600	-	-	✗	✗	✗	✓
Swahili	600	553	1883	✓	✓	✗	✓
Wolof	600	625	1312	✗	✗	✗	✓
Xhosa	600	601	1633	✓	✗	✓	✓
Yoruba	600	713	1964	✗	✓	✓	✓
Zulu	600	601	1670	✗	✗	✓	✓
Aymara	750	-	-	✗	✗	✗	✓
Asháninka	-	-	-	✗	✗	✗	✗
Bribri	750	-	-	✗	✗	✗	✗
Guaraní	750	-	-	✗	✗	✗	✓
Nahuatl	750	-	-	✗	✗	✗	✗
Otomí	750	-	-	✗	✗	✗	✗
Quechua	750	-	100	✗	✗	✗	✓
Rarámuri	750	-	-	✗	✗	✗	✗
Shipibo-Konibo	750	-	-	✗	✗	✗	✗
Wixárika	750	-	-	✗	✗	✗	✗

Table 7: The size of the combined datasets. The table is split into 3 sections: 1) The original 44 languages used for realignment 2) African languages exclusive to AfriXNLI and MasakhaPOS, 3) South American languages exclusive to AmericasNLI.

Method	#Languages	NLI	PoS-Tagging	NER
Baseline				
All 65 languages	65	**55.43** ± **0.29**	**66.87** ± **0.27**	**54.75** ± **1.02**
Present in XTREME-R	47	53.51 ± 0.27	65.06 ± 0.38	52.54 ± 1.41
Present in Africa	21	54.94 ± 0.38	65.79 ± 0.18	53.72 ± 0.86
Fine-tuning only	0	53.12 ± 0.25	62.20 ± 0.78	52.25 ± 1.05
Featural Diversity				
Most diverse from English	5	53.86 ± 0.24	64.08 ± 0.30	50.73 ± 0.73
	10	54.40 ± 0.31	64.87 ± 0.27	53.64 ± 0.59
	20	54.94 ± 0.31	65.66 ± 0.06	54.38 ± 1.05
	40	**55.94** ± **0.24**	**66.24** ± **0.07**	**54.82** ± **0.45**
Least diverse from English	5	53.22 ± 0.17	62.57 ± 0.33	51.43 ± 1.41
	10	53.26 ± 0.52	63.14 ± 0.60	52.70 ± 1.14
	20	53.26 ± 0.12	63.33 ± 0.39	51.58 ± 0.69
	40	53.69 ± 0.28	65.83 ± 0.08	52.85 ± 0.49
Phylogenetic Diversity				
Most diverse families	5	54.03 ± 0.30	63.53 ± 0.06	49.69 ± 1.06
	10	53.86 ± 0.25	63.85 ± 0.51	51.92 ± 0.82
	20	53.98 ± 0.34	63.87 ± 0.18	53.98 ± 0.84
	25	**54.25** ± **0.19**	**65.13** ± **0.37**	**54.73** ± **0.85**
Most diverse families within Indo-European	5	52.86 ± 0.35	61.89 ± 0.48	50.20 ± 0.68
	10	53.18 ± 0.45	62.40 ± 0.49	52.52 ± 1.56
	20	53.06 ± 0.29	63.59 ± 0.87	51.08 ± 0.80
Script Diversity				
Most diverse scripts	5	52.80 ± 0.25	61.74 ± 0.66	51.19 ± 1.55
	10	52.54 ± 0.34	62.11 ± 0.87	50.41 ± 1.14
	18	52.69 ± 0.13	62.82 ± 0.74	50.59 ± 0.73
Most diverse using Latin script	5	53.94 ± 0.31	63.92 ± 0.38	51.38 ± 0.50
	10	54.75 ± 0.13	64.95 ± 0.40	53.28 ± 0.73
	20	53.89 ± 0.12	65.25 ± 0.15	53.10 ± 0.89
	41	**55.96** ± **0.15**	**67.23** ± **0.20**	**54.00** ± **0.48**
Least diverse using Latin script	5	53.20 ± 0.56	61.36 ± 0.13	49.22 ± 0.47
	10	53.14 ± 0.31	62.89 ± 0.51	51.04 ± 0.33
	20	55.73 ± 0.19	66.05 ± 0.25	53.60 ± 0.91
Random Selection				
Random Seeded	5	54.21 ± 0.86	63.61 ± 0.55	51.76 ± 1.90
	10	54.24 ± 0.88	64.78 ± 0.34	51.45 ± 0.51
	20	54.64 ± 0.76	65.27 ± 0.42	52.92 ± 1.14
	40	**55.36** ± **0.24**	**66.06** ± **0.63**	**53.42** ± **0.48**

Table 8: Accuracy of XLM-R on NLI, PoS-Tagging, NER tasks. Results are averaged across 4 seeds along with the standard deviation.

Method	#Languages	NLI	PoS-Tagging	NER
Most featural diversity from English				
Joshi Class = 2	10	54.88 ± 0.18	**64.72 ± 0.26**	51.42 ± 0.30
Joshi Class = 3	17	53.06 ± 0.36	63.30 ± 0.64	50.17 ± 0.81
Joshi Class = 3,4,5	37	53.20 ± 0.47	63.35 ± 1.02	51.13 ± 1.74
Joshi Class = 4,5	20	53.24 ± 0.30	63.84 ± 0.52	50.66 ± 0.41
Seen by mBERT	47	53.85 ± 0.32	64.50 ± 0.88	**52.96 ± 1.14**
Seen by XLM-R	51	54.27 ± 0.28	64.39 ± 0.26	52.71 ± 0.87
Unseen by mBERT	34	**55.14 ± 0.40**	64.02 ± 0.23	52.63 ± 0.44
Unseen by XLM-R	30	54.88 ± 0.34	64.70 ± 0.58	52.80 ± 0.73
Most Phylogenetic Diversity				
Joshi Class = 2	10	54.11 ± 0.60	63.96 ± 0.18	51.23 ± 0.68
Joshi Class = 3	17	52.96 ± 0.39	63.60 ± 0.49	50.73 ± 1.91
Joshi Class = 3,4,5	37	53.10 ± 0.29	63.68 ± 0.23	52.53 ± 1.67
Seen by mBERT	47	53.57 ± 0.25	64.06 ± 0.67	52.52 ± 1.12
Seen by XLM-R	51	54.42 ± 0.58	64.26 ± 0.17	53.34 ± 0.80
Unseen by mBERT	34	**54.59 ± 0.37**	64.02 ± 0.33	**53.37 ± 0.82**
Unseen by XLM-R	30	**54.59 ± 0.38**	**65.04 ± 0.47**	52.54 ± 0.97
Most Script Diversity				
Joshi Class = 2	10	**54.20 ± 0.17**	**63.92 ± 0.28**	50.50 ± 0.66
Joshi Class = 3	17	52.67 ± 0.18	61.73 ± 1.00	49.46 ± 0.31
Joshi Class = 3,4,5	37	52.60 ± 0.19	62.46 ± 0.57	**52.24 ± 1.45**
Joshi Class = 4,5	20	53.16 ± 0.13	63.54 ± 0.35	51.61 ± 0.35
Seen by mBERT	47	52.73 ± 0.16	61.95 ± 0.37	49.73 ± 0.79
Seen by XLM-R	51	52.73 ± 0.16	61.95 ± 0.37	49.73 ± 0.79
Random Seeded				
Joshi Class = 2	10	**55.03 ± 0.65**	65.06 ± 0.53	52.46 ± 0.97
Joshi Class = 3	17	53.28 ± 0.25	62.94 ± 0.49	50.31 ± 2.05
Joshi Class = 3,4,5	37	53.36 ± 0.21	63.54 ± 0.40	51.51 ± 0.70
Joshi Class = 4,5	20	53.30 ± 0.15	63.69 ± 0.14	50.67 ± 1.50
Seen by mBERT	47	53.43 ± 0.37	63.67 ± 0.81	51.25 ± 0.40
Seen by XLM-R	51	53.86 ± 0.48	63.84 ± 0.62	51.08 ± 0.65
Unseen by mBERT	34	54.45 ± 0.87	64.55 ± 0.25	52.48 ± 1.19
Unseen by XLM-R	30	54.87 ± 0.20	**65.11 ± 0.29**	**52.62 ± 0.53**

Table 9: Ablation studies: Accuracy of mBERT on NLI, POS-Tagging, NER tasks. Results are averaged across 4 seeds along with the standard deviation.

Method	#Languages	NLI	PoS-Tagging	NER
Baseline				
All 65 languages	65	59.43 ± 0.17	69.14 ± 0.24	57.07 ± 0.86
Present in XTREME-R	47	**59.44 ± 0.39**	67.32 ± 0.59	**57.24 ± 0.73**
Present in Africa	21	59.90 ± 0.33	**69.92 ± 0.15**	54.10 ± 1.14
Fine-tuning only	0	58.61 ± 0.10	65.98 ± 0.73	51.09 ± 0.96
Featural Diversity				
Most diverse from English	5	58.97 ± 0.28	67.99 ± 0.37	53.39 ± 1.17
	10	58.87 ± 0.19	68.48 ± 0.43	56.11 ± 0.77
	20	59.27 ± 0.07	68.57 ± 0.40	56.51 ± 1.16
	40	**59.64 ± 0.24**	**68.89 ± 0.26**	**56.39 ± 0.98**
Least diverse from English	5	58.59 ± 0.17	66.04 ± 0.79	52.80 ± 0.58
	10	58.53 ± 0.09	65.99 ± 0.80	52.87 ± 1.47
	20	58.74 ± 0.11	67.10 ± 0.25	53.96 ± 1.34
	40	58.75 ± 0.17	68.10 ± 0.36	56.21 ± 0.42
Phylogenetic Diversity				
Most diverse families	5	58.81 ± 0.20	67.20 ± 0.37	51.63 ± 0.72
	10	58.91 ± 0.29	67.31 ± 0.22	53.81 ± 1.04
	20	**59.15 ± 0.31**	66.64 ± 0.10	**55.53 ± 0.77**
	25	59.06 ± 0.12	**67.87 ± 0.26**	54.99 ± 0.66
Most diverse families within Indo-European	5	58.74 ± 0.30	66.39 ± 0.94	51.90 ± 1.23
	10	58.59 ± 0.24	67.14 ± 0.64	53.81 ± 2.65
	20	58.92 ± 0.10	67.15 ± 0.33	52.35 ± 1.42
Script Diversity				
Most diverse scripts	5	58.76 ± 0.03	67.26 ± 0.46	49.89 ± 1.18
	10	58.70 ± 0.27	67.04 ± 0.73	51.40 ± 1.97
	18	58.77 ± 0.20	67.04 ± 0.92	50.83 ± 1.50
Most diverse using Latin script	5	58.75 ± 0.35	67.88 ± 0.26	53.74 ± 0.80
	10	59.15 ± 0.30	68.11 ± 0.15	56.27 ± 0.34
	20	58.75 ± 0.23	67.75 ± 0.09	55.47 ± 1.45
	41	**59.88 ± 0.06**	**69.62 ± 0.30**	**56.94 ± 0.44**
Least diverse using Latin script	5	58.86 ± 0.28	65.49 ± 1.16	51.00 ± 1.02
	10	58.77 ± 0.22	65.82 ± 1.30	53.27 ± 1.10
	20	59.75 ± 0.04	68.65 ± 0.29	56.44 ± 0.22
Random Selection				
Random Seeded	5	**59.49 ± 0.44**	67.46 ± 1.51	54.18 ± 1.44
	10	59.13 ± 0.50	67.91 ± 0.78	54.29 ± 1.59
	20	59.27 ± 0.54	68.26 ± 0.46	**56.66 ± 1.18**
	40	**59.49 ± 0.20**	**68.69 ± 0.32**	56.04 ± 0.73

Table 10: Accuracy of mBERT on NLI, PoS-Tagging, NER tasks. Results are averaged across 4 seeds along with the standard deviation.

Method	#Languages	NLI	PoS-Tagging	NER
Most featural diversity from English				
Joshi Class = 2	10	59.49 ± 0.25	**68.83 ± 0.40**	52.75 ± 1.66
Joshi Class = 3	17	59.00 ± 0.21	67.46 ± 0.36	51.90 ± 2.04
Joshi Class = 3,4,5	37	58.79 ± 0.15	67.13 ± 0.12	53.95 ± 1.32
Joshi Class = 4,5	20	58.75 ± 0.24	67.27 ± 0.37	54.26 ± 1.83
Seen by mBERT	47	59.04 ± 0.38	68.04 ± 0.55	**55.39 ± 0.98**
Seen by XLM-R	51	59.00 ± 0.38	67.58 ± 0.54	52.55 ± 1.75
Unseen by mBERT	34	**59.86 ± 0.22**	68.25 ± 0.30	54.70 ± 0.57
Unseen by XLM-R	30	59.68 ± 0.44	68.62 ± 0.22	54.78 ± 1.39
Most Phylogenetic Diversity				
Joshi Class = 2	10	59.01 ± 0.12	67.59 ± 0.19	53.02 ± 1.29
Joshi Class = 3	17	58.88 ± 0.21	67.62 ± 0.34	51.15 ± 0.83
Joshi Class = 3,4,5	37	58.63 ± 0.20	67.91 ± 0.43	52.44 ± 0.68
Seen by mBERT	47	58.79 ± 0.20	68.38 ± 0.25	**56.23 ± 0.43**
Seen by XLM-R	51	59.07 ± 0.17	67.99 ± 0.22	55.08 ± 1.52
Unseen by mBERT	34	**59.77 ± 0.10**	68.58 ± 0.27	54.10 ± 1.09
Unseen by XLM-R	30	59.34 ± 0.37	**68.73 ± 0.13**	53.34 ± 0.98
Most Script Diversity				
Joshi Class = 2	10	**59.11 ± 0.23**	67.48 ± 0.15	**54.33 ± 0.30**
Joshi Class = 3	17	58.47 ± 0.21	67.48 ± 0.35	50.89 ± 1.63
Joshi Class = 3,4,5	37	58.66 ± 0.22	**67.52 ± 0.28**	51.64 ± 0.85
Joshi Class = 4,5	20	58.65 ± 0.16	67.07 ± 0.29	53.26 ± 0.61
Seen by mBERT	47	58.67 ± 0.20	66.91 ± 1.18	50.46 ± 0.65
Seen by XLM-R	51	58.67 ± 0.20	66.91 ± 1.18	50.46 ± 0.65
Random Seeded				
Joshi Class = 2	10	**59.93 ± 0.05**	68.84 ± 0.68	**54.54 ± 1.80**
Joshi Class = 3	17	58.95 ± 0.19	67.05 ± 0.79	51.00 ± 0.98
Joshi Class = 3,4,5	37	58.80 ± 0.24	67.16 ± 0.43	53.80 ± 1.23
Joshi Class = 4,5	20	58.69 ± 0.29	67.47 ± 0.70	54.38 ± 1.82
Seen by mBERT	47	58.74 ± 0.13	67.39 ± 0.90	52.39 ± 2.13
Seen by XLM-R	51	58.88 ± 0.21	67.10 ± 0.44	53.26 ± 2.65
Unseen by mBERT	34	59.78 ± 0.38	69.01 ± 0.23	53.26 ± 1.33
Unseen by XLM-R	30	59.82 ± 0.26	**69.09 ± 0.14**	53.54 ± 1.42

Table 11: Ablation studies: Accuracy of XLM-R on NLI, PoS-Tagging, NER tasks. Results are averaged across 4 seeds along with the standard deviation.