Tokenizer-Aware Cross-Lingual Adaptation of Decoder-Only LLMs through Embedding Relearning and Swapping

Anonymous ACL submission

Abstract

Extending Large Language Models (LLMs) to support new languages is a challenging problem, with most methods proposed suffering from high computational cost and catastrophic forgetting of original model capabilities. Embedding relearning (Artetxe et al., 2020), a technique that creates new tokenizers and tunes embeddings on fixed model weights for target language adaptation, is both lightweight and performant. However, it has only been demonstrated to work for older generation encoder-only models and for high resource languages. In this paper, we extend this framework to decoder-only LLMs focusing on joint adaptation to many languages, including low-resource languages. We experiment in three language groups over 100 languages each. Our approach adapts a pre-trained model via switching to a customized tokenizer, and relearning the embedding layer. Across 96 diverse languages spanning both classification and generation tasks, we demonstrate embedding relearning improves `Gemma2` models (with up to 27B parameters) by up to 20%, being more effective than, or on par with, full-weight updating baselines while effectively mitigating English forgetting (1-3% regressions). Analysis reveals the critical role of customizing tokenizers in achieving effective language transfer, particularly for non-Latin script languages. We further show embedding relearning helps transfer reasoning abilities across languages, achieving a 14% improvement over a math-optimized LLM across 20 languages.

1 Introduction

Large Language Models (LLMs) have transformed the field of natural language processing through pre-training on extensive web-scale corpora (Brown et al., 2020; Anil et al., 2024). Despite these advancements, their success has been primarily centered on English, leaving the multilingual ability less explored. While the multilingual potential of LLMs has been demonstrated across multiple languages (Shi et al., 2023), their practical applications remain largely confined to a limited set of high-resource languages. This limitation reduces their utility for users speaking under-represented languages (Ahia et al., 2023).

Recently, many works focus on increasing the language support of LLMs. For instance, continued pre-training approaches further train LLMs on additional multilingual data by using either the original vocabulary (Zheng et al., 2024; Üstün et al., 2024) or introducing language-specific tokens (Fujii et al., 2024; Lu et al., 2024) (Figure 1 (b)). Despite the effectiveness in enhancing multilingual support, this paradigm typically requires full-parameter tuning on vast data, making adapting an LLM to accommodate new languages expensive. Moreover, such adaptation poses a significant risk of catastrophic forgetting, whereby the LLM loses previously acquired knowledge from the initial pre-training phase (Luo et al., 2024; Shi et al., 2024). Additionally, these methods usually focus on a smaller number of high-resource languages (Aryabumi et al., 2024; Alves et al., 2024; Xu et al., 2024), neglecting mid- and low-resource languages. Given these challenges, it is crucial to explore efficient and scalable methods for developing multilingual LLMs that can support diverse languages across varying resource levels.

Alternatively, *embedding relearning* – a technique that adapts models to new languages by learning new tokenizers and embeddings while keeping the transformer layers fixed – has shown to be effective in improving performance in target languages (Artetxe et al., 2020). However, existing embedding relearning approaches face several limitations: 1) the major focus on ‘outdated‘ encoder-only PLMs limits the applicability; 2) the monolingual transfer strategy (i.e., one embedding per language) reduces efficiency; 3) limited linguistic coverage and exclusive evaluation on classification

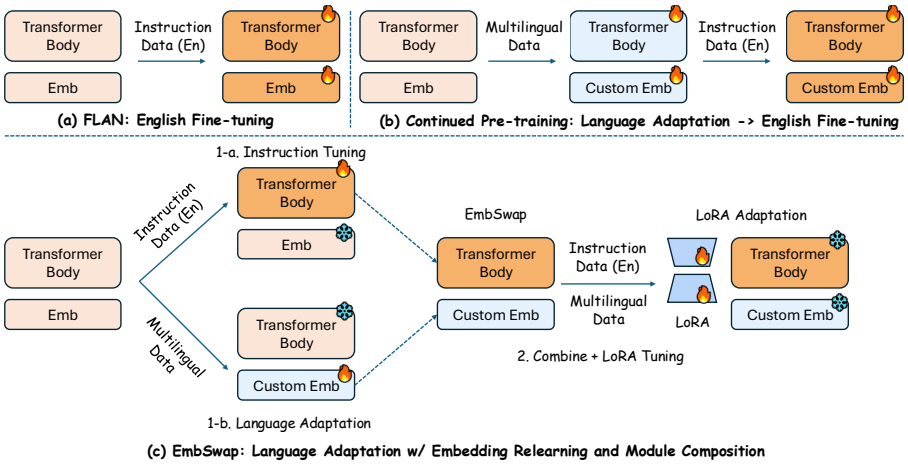

Figure 1: Methods for cross-lingual transfer of LLMs. (a): Use English instruction data for fine-tuning for vanilla cross-lingual transfer. (b): Continued pre-training: create customized tokenizers and do full-parameter tuning on multilingual data followed by English instruction-tuning. (c): EMBSWAP: 1-a: freeze the original embeddings of LLMs and instruction-tune the transformer body using English alignment data; 1-b: learn new multilingual embeddings by freezing the transformer body for target language adaptation of LLMs; 2) combine new embeddings with instruction-tuned transformer body as the EMBSWAP and further perform LoRA tuning to connect the combined components for enhanced cross-lingual transfer.

tasks leave the diverse linguistic and task settings underexplored (Liang et al., 2023). These limitations prompt important questions: 1) Can this approach be extended to modern decoder-only LLMs and how well it performs on generation tasks as compared to well-studied classification ones? 2) Can we achieve efficient cross-lingual transfer to diverse languages with different resource levels altogether? 3) What are the effective ways of building tokenizers for a large group of target languages that also help preserve LLM's existing abilities?

We revisit *embedding relearning* and extend it to decoder-only LLMs for cross-lingual transfer across many languages. Figure 1 (c) shows that the pipeline starts with two parallel training from LLMs primarily pre-trained on English data: 1) We adapt it to a target language group[1] by learning multilingual embeddings on new tokenizers while fixing the transformer body; 2) We employ English alignment data to instruction-tune the transformer body, keeping the original embeddings fixed. After this, the relearned multilingual embeddings are combined with the instruction-tuned transformer body for efficient zero-shot cross-lingual transfer, which we denote as EMBSWAP. We can further enhance the compatibility of the composed modules via a cost-effective LoRA-based adaptation.

We conduct experiments on diverse multilingual NLP tasks spanning classification, generation and mathematical reasoning. Results show that EMBSWAP surpasses the original `Gemma2` models at all scales and achieves results comparable to full-weight updating baselines, while effectively mitigating knowledge forgetting. Our findings indicate that carefully-curated tokenizer construction and embedding initialization methods are crucial, and non-Latin script languages benefit the most from customized tokenizers. We believe these insights equip practitioners with a recipe for achieving effective and efficient cross-lingual transfer in LLMs through tokenizer and embedding relearning.

2 Methodology

Prior embedding relearning methods create dedicated tokenizer and embedding layer for each target language, which is practical as the involved language model is monolingual. By contrast, most recent LLMs exhibit a degree of multilingualism in both their representations and tokenizers (Blevins and Zettlemoyer, 2022). A naive solution to applying embedding relearning to LLMs is to continue training the embedding layer on additional multilingual data. However, the tokenizers employed in these LLMs are biased towards English and several high-resource languages. This bias results in the over-fragmentation of text from long-tail languages (Figure 2), thereby degrading the performance and efficiency of processing such languages (Ahuja et al., 2023). In this paper, we demonstrate that having tokenizers that provide fairer representation

[1]We focus on three language groups: South East Asia (SEA), Indic (IND), and African (AFR).

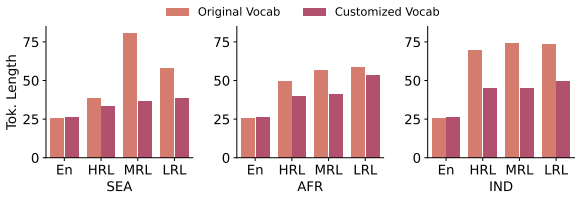

Figure 2: The tokenization comparison between using the vanilla and customized multilingual tokenizers on Gemma2. Tok. Length refers to the average number of tokens required to represent the same amount of texts.

across languages is critical to achieving effective cross-lingual transfer.

2.1 Customized Vocabulary Construction

Our strategy involves constructing distinct tokenizers for each language group (§3.2). Tailoring tokenizers to specific language groups enhances cross-lingual transfer among geographically related long-tail languages compared to using monolingual tokenizers. Moreover, this approach avoids the shortcomings of a universal tokenizer that treats all low-resource languages uniformly poorly. Based on this, we propose a *Prune-with-Extension* approach for developing tokenizers optimized for language adaptation while maintaining English ability. First, we prune the tokenizer of target LLMs by removing non-English tokens. Then the pruned tokenizer is extended through adding new tokens, which are obtained by training tokenizers for target languages using BPE (Gage, 1994; Sennrich et al., 2016).

Pruning the Tokenizer To preserve the pre-trained knowledge embedded in the language model, current approaches often expand the vocabulary by adding new tokens (Fujii et al., 2024; Cui et al., 2024). This, however, can substantially increase pre-training time due to the extra computational cost of the output embedding layer (Liang et al., 2023). To avoid this, we first prune the existing tokenizers by retaining only English tokens before adding those from low-resource languages. Given the predominant English training data for LLMs, we hypothesize that a significant portion of their knowledge is associated with English tokens, and reusing English tokens can effectively retain this knowledge (Garcia et al., 2021). In our implementation, for a given LLM, we identify English tokens by tokenizing a set of 20 million English sentences using its tokenizer, with further filtering through removing non-Latin script tokens.[2]

Training Multilingual Tokenizers To get the data for building a multilingual vocabulary for long-tail languages, we sample from the Next Thousand Languages (NTL) corpus (Caswell et al., 2020; Bapna et al., 2022). Our empirical analysis reveals that sampling data for each language up to a maximum of 500K lines from NTL effectively addresses the imbalance between high- and low-resource languages, outperforming temperature-sampling techniques. Subsequently, we train a BPE tokenizer using the sampled data to generate a vocabulary whose size aligns with that of the target LLMs.

Extending the Pruned Tokenizer We sequentially add the identified English tokens followed by tokens from the newly built multilingual tokenizer. Both types of tokens are added in the same preference order as in their respective tokenizers. The final vocabulary maintains the same size as the original tokenizer, with over 60% token overlap, resulting in negligible variations in English tokenizations (see Table 6 in Appendix). Figure 2 shows our final vocabulary achieves significant compression rate improvements by consistently producing shorter sequences across languages of a spectrum of resource levels while barely affecting English.

2.2 Training Recipe

Embedding Initialization To maximally inherit the pre-trained knowledge embedded in the target LLM's embedding layer, we adopt a strategy inspired by Gee et al. (2022). For tokens that overlap between the target LLM's vocabulary and our multilingual vocabulary, we directly copy the corresponding embeddings. For new tokens, we employ the LLM's original tokenizer to decompose them into subtokens and initialize their embeddings using the average of their subtoken embeddings.[3]

Language Adaptation We fine-tune the customized embeddings on 200B curated multilingual tokens $\mathcal{D}_{la}$ while keeping the transformer body frozen (Figure 1 (c) step 1-b), with the same training objective used for pre-training the base LLM. This is based on the assumption that the pre-trained transformer body encapsulates universal cross-lingual knowledge (Zhao et al., 2024b; Wendler et al., 2024; Tang et al., 2024), while the embedding layer encodes language-specific information, which suggests embedding tuning should be effective for language adaptation.

EMBSWAP We instruction-tune the transformer

[2] 40% of tokens are discarded: non-Latin scripts tokens from high-resource languages and very rare English tokens.

[3] We find initialization from original embeddings are crucial, while the exact method (averaging or max-pooling) makes minimal difference (See Appendix Figure 16).

body of base LLMs on a diverse range of English tasks $\mathcal{D}_{it}$ (Wei et al., 2022) with 20B tokens (Figure 1 (c) step 1-a). Notably, we employ the LLM's *original* embeddings and keep them frozen in this step.[4] We then integrate the customized embeddings obtained from the language adaptation stage into the instruction-tuned transformer body.

LoRA-Adaptation Since the transformer body and customized embeddings are independently trained, EMBSWAP may suffer from incompatible parameters. Our empirical findings indicate that EMBSWAP is effective for discriminative tasks but sometimes underperforms an instruction-tuned baseline on generative tasks. To mitigate this and ensure the assembled model's effectiveness across various tasks, we insert LoRA weights into the self-attention layer of the tuned transformer body (Figure 1 (c) step 2). These weights are then fine-tuned on a sub-sampled joint corpus $\mathcal{D}_{mix} = \mathcal{D}_{la} \cup \mathcal{D}_{it}$ with 4B tokens, while keeping both the body and embeddings frozen.

3 Experiment setup

3.1 Pre-training Data

The data $\mathcal{D}_{la}$ for embedding training is a mixed corpus with 65% sentence-level and 35% document-level texts. The sentence-level data is exclusively from the Next Thousand Languages (NTL) corpora (Caswell et al., 2020; Bapna et al., 2022), which provides web-crawled monolingual sentences and translation pairs for over 1000 languages. For document-level texts, we sample data from multilingual Wikipedia and mC4 (Xue et al., 2021). We use UniMax sampling (Chung et al., 2023) with $N = 5$ to up-sample low-resource languages. Please refer to Appendix Figure 14 for pre-training data ablations.

We take FLAN (Wei et al., 2022) as the instruction tuning data $\mathcal{D}_{it}$. The data $\mathcal{D}_{mix}$ used in LoRA-Adaptation consists of a 10% sample of $\mathcal{D}_{it}$, combined with an equal number of instances from $\mathcal{D}_{la}$.[5]

3.2 Languages

We select languages from three families based on geographic relations: South East Asian (SEA), African (AFR), and Indic (IND). This results in 212 languages from SEA, 392 from AFR, and 170 from IND. Each regional dataset is processed separately, with a tailored tokenizer, language-adapted embeddings and LoRA update parameters.

3.3 Models

Our evaluation is focused on `Gemma2` (2B, 9B, 27B) (Riviere et al., 2024). We also test the generalization ability of embedding relearning in two LLMs with varying degrees of multilinguality: `Aya23` (8B, 35B) (Aryabumi et al., 2024) and `PaLM2` (XXS, S) (Anil et al., 2023).

As shown in Figure 1, we end up having four types of models: (i) `FLAN` (step 1-a): models that undergo instruction tuning. (ii) `Lang-Adapt` (step 1-b): the LLMs after language adaptation with embedding tuning. (iii) EMBSWAP (step-2 Combine): model denoted as EMBSWAP is constructed by combining the transformer body of `FLAN` with the embeddings from `Lang-Adapt`. (iv): `LoRA-Adapt` (step-2 LoRA Tuning): EMBSWAP models with the LoRA-Adaptation process. Detailed training procedures for each model type are in Appendix A.1.

We compare EMBSWAP with three variants: 1) **Continued Pre-training (CPT)**: the most costly method which first fully tunes the base model on multilingual data $\mathcal{D}_{la}$ to obtain language-adapted model θ_{la}, then instruction tunes θ_{la} on English data $\mathcal{D}_{it}$. 2) **ChatVector (Huang et al., 2024)**: which avoids repeated instruction tuning, through obtaining a task vector (Ilharco et al., 2023) from a base model and IT model: $\theta_{\text{task}} = \theta_{\text{it}} - \theta_{\text{base}}$, which is then added to θ_{la}. 3) **IT+Lang-Adapt**: an embedding relearning variant that performs one-step embedding tuning on `Gemma2-IT` model with the combination of $\mathcal{D}_{la}$ and $\mathcal{D}_{it}$. For all three variants, we adopt the same customized tokenizer, initialized embeddings, and training data conditions (i.e., 200B $\mathcal{D}_{la}$ and 20B $\mathcal{D}_{it}$) for fair comparison.

3.4 Evaluation Tasks

We use *five-shot* prompting for evaluating language-adapted LLMs. In contrast, EMBSWAP is evaluated in a *zero-shot* setting, given it has been instruction tuned. We also evaluate EMBSWAP using a compiled English benchmark (Appendix B) to examine potential regressions in general English ability.

We perform evaluation with various multilingual tasks including machine translation (MT), multiple choice question answering (MCQA), topic classification, mathematical reasoning, cross-lingual question answering (QA) and summarization. For MT, we focus on EN-XX and select 23 SEA, 42 AFR,

[4] Freezing embeddings makes the instruction tuning and language adaptation processes symmetry. This enhances modularity and improves the parameter compatibility.

[5] This ensures the instruction-following ability isn't forgotten. Please refer to Appendix Figure 15 for detailed analysis.

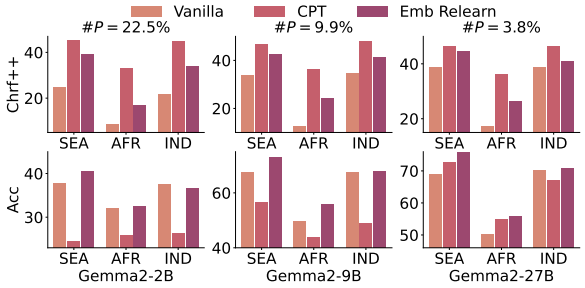

Figure 3: FLORES-200 EN-XX and BELEBELE Language Adaptation results across all sizes of Gemma2 models with 5-shot prompting. $\#P$: fraction of tuned parameters in embedding relearning.

Model	SEA	AFR	IND	Avg.
CALM (PaLM2-XXS-NTL+S)†	25.3	17.8	17.9	19.8
PaLM2-S-NTL	25.2	17.4	15.1	18.2
PaLM2-S	22.0	15.3	14.2	16.4
★ Lang-Adapt	**25.6**	**18.8**	**22.3**	**22.3**
Gemma2-9B	19.9	18.3	13.6	16.4
★ Lang-Adapt	**36.4**	**25.8**	**29.6**	**30.4**
Gemma2-27B	30.6	22.2	18.4	22.4
★ Lang-Adapt	**41.9**	**31.8**	**27.5**	**32.2**

Table 1: Language adaptation results on GSM8K-NTL, with the best in **bold**. †: from Bansal et al. (2024).

and 21 IND languages from FLORES-200 (Goyal et al., 2022). For MCQA, we use BELEBELE (Bandarkar et al., 2024) and evaluate on 15 SEA, 25 AFR, and 19 IND languages. For multilingual mathematical reasoning, we use GSM8K-NTL (Shi et al., 2023; Bansal et al., 2024) translated from GSM8K and select 5 SEA, 5 AFR, and 10 IND languages. For cross-lingual QA and summarization, the XORQA-IN and XSUM-IN datasets from INDICGENBENCH (Singh et al., 2024) are used and 29 IND languages are included for evaluation.

3.5 Main Results

We first present the results of Language Adaptation through embedding tuning on its own (§2.2 – Language Adaptation) in §3.5.1. Then we evaluate EMBSWAP alongside LoRA adaptation (§2.2 – EMBSWAP & LoRA-Adaptation) in §3.5.2.

3.5.1 Language Adaptation Results

Embedding relearning improves cross-lingual transfer. Figure 3 shows embedding relearning consistently outperforms vanilla Gemma2 across three language groups and achieves better cross-lingual transfer than CPT, as evident by BELEBELE results, a task that relies on skills learnt from English. While CPT excels on FLORES (likely due to its full model adaptation for MT with training on parallel data), it does not generalize well to other

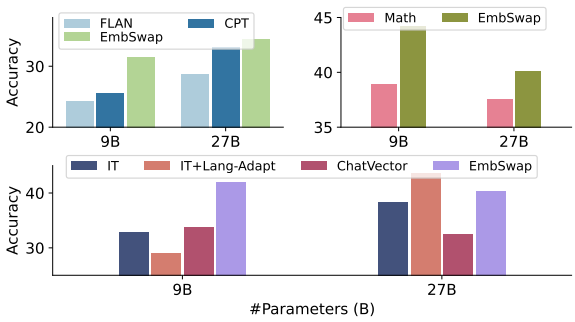

Figure 4: Zero-shot multilingual mathematical reasoning results on GSM8K-NTL. IT: LLMs are instruction-tuned without fixing embedding.

task types. See Appendix C.1 for more language adaptation results on Aya23 and PaLM2.

Embedding relearning helps preserve LLMs' reasoning knowledge. We evaluate how language adaptation affects English reasoning transfer to other languages. Table 1 shows embedding relearning consistently improves the mathematical reasoning ability across various low-resource languages. Compared to CALM (Bansal et al., 2024), a form of adapter enabling model composition, embedding relearning achieves more substantial improvements over PaLM2-S with only embedding tuning and incurring no extra inference costs.[6] Moreover, it shows superior performance (+4%) compared to PaLM2-S-NTL that was full-parameter tuned on NTL.[7] Overall, our results suggest that embedding relearning offers an effective alternative to CALM and full-parameter tuning. We observe performance improvements with more modern models, with particularly pronounced gains in Gemma2.

3.5.2 EMBSWAP Results

EMBSWAP transfers mathematical reasoning abilities across languages. Figure 4 shows that EMBSWAP outperforms the instruction-tuned baseline for zero-shot evaluation on mathematical reasoning tasks, with up to 8% gains across 20 low-resource languages. Moreover, EMBSWAP further advances performance by incorporating LLMs with enhanced mathematical reasoning ability. These include the IT models aligned via reinforcement learning and math models instruction-tuned on reasoning intensive data (Yue et al., 2024).

[6]Embedding relearning requires more training as it must be repeated for different linguistic regions.

[7]Both methods use the same training dataset. However, NTL training did not create region-specific models by splitting training data, potentially diminishing its effectiveness due to the curse of multilinguality (Conneau et al., 2020), which arises when a single model is trained on too many languages.

Task Type	% Tuned Params (Lang Adapt)	English	Classification						Generation							Avg.
Eval. Metric			BeleBele Accuracy			Sib-200 Accuracy			Flores-200 ChrF++			XorQA-In Token-F1		XSum-In ChrF		
Model			Sea	Afr	Ind	Sea	Afr	Ind	Sea	Afr	Ind	Ind	En	Ind	En	
Gemma2-2B																
FLAN	-	**73.0**	52.0	36.0	50.2	65.9	47.3	67.8	27.8	9.6	20.7	7.6	47.8	3.7	**36.6**	31.7
It+Lang-Adapt	22.5%	54.1	51.2	36.7	42.9	67.0	48.7	67.1	35.4	14.2	29.2	**12.3**	40.2	7.1	27.5	33.0
EmbSwap	22.5%	69.8	62.0	40.2	54.4	72.9	**57.9**	70.5	27.0	11.0	17.2	8.6	54.1	6.6	34.1	35.3
★ LoRA-Adapt	22.5%	69.7	**63.3**	**43.7**	**55.8**	**73.5**	47.8	**70.7**	**37.1**	**18.3**	**32.5**	9.8	**60.5**	**9.9**	31.9	**38.7**
ChatVector	100%	39.1	28.8	32.5	30.6	24.5	29.3	52.2	42.1	32.7	36.1	19.4	16.1	19.3	21.9	28.0
CPT	100%	64.3	55.4	49.0	58.9	66.9	65.9	78.4	44.2	24.4	36.0	11.2	68.2	14.8	34.0	43.0
Gemma2-9B																
FLAN	-	**83.5**	70.6	49.9	68.9	74.4	61.0	79.1	32.0	12.6	27.8	9.8	60.3	15.1	**37.5**	42.0
It+Lang-Adapt	9.9%	73.9	75.4	57.9	72.7	77.2	**69.2**	79.7	39.3	22.2	38.7	**23.1**	**69.6**	17.0	33.9	**48.3**
EmbSwap	9.9%	82.3	78.1	57.5	71.2	**78.6**	68.8	79.8	35.9	16.5	31.4	12.5	65.2	15.2	37.1	45.5
★ LoRA-Adapt	9.9%	82.5	**78.4**	**60.2**	**73.1**	78.5	**69.2**	**80.1**	**40.0**	**25.6**	**40.5**	12.1	64.1	**17.4**	37.0	47.3
ChatVector	100%	58.4	69.0	52.5	59.3	82.1	66.1	77.2	44.5	26.6	40.5	17.3	29.8	22.1	26.9	43.7
CPT	100%	74.6	76.3	65.4	74.1	83.0	64.8	80.0	45.2	34.1	44.6	19.5	72.7	22.6	37.0	51.7
Gemma2-27B																
FLAN	-	**84.3**	71.9	52.6	72.9	72.6	60.2	76.4	33.2	15.2	29.6	20.4	61.7	15.0	**37.6**	43.7
It+Lang-Adapt	3.8%	81.9	**81.7**	**62.7**	**76.0**	**82.1**	**70.0**	**81.9**	**42.5**	24.8	39.3	**26.4**	**73.0**	**18.3**	35.8	**51.1**
EmbSwap	3.8%	83.6	78.8	56.3	73.1	78.1	66.1	78.5	33.7	13.2	23.9	15.2	59.3	12.4	36.8	43.5
★ LoRA-Adapt	3.8%	83.4	79.5	60.4	74.3	79.5	68.3	80.0	**42.5**	**25.7**	**40.6**	16.4	70.1	18.0	36.9	49.1
ChatVector	100%	62.9	78.5	55.7	71.7	81.5	55.9	77.1	43.4	31.4	37.8	16.2	18.2	20.0	22.8	43.0
CPT	100%	79.5	81.2	68.6	77.6	72.6	72.3	78.2	45.4	34.4	45.0	21.1	72.4	23.7	38.7	52.5

Table 2: Zero-shot cross-lingual transfer results. **Bold** and underlined: best and second-best results for embedding relearning methods. Red values indicate instances where EmbSwap hurts FLAN. English results are excluded from the average. **% Tuned Params (Lang Adapt)**: proportions of trainable parameters during language adaptation.

EmbSwap is more effective for classification tasks. Table 2 shows that EmbSwap improves the performance on classification tasks across all model sizes with up to 10% gains. By contrast, for generation tasks, the method's behavior is inconsistent, often leading to performance degradation. We attribute this phenomenon to intrinsic difficulty: classification tasks are generally easier as the solution space is typically small compared to generation tasks, making them more robust to embedding changes. However, the embedding layer is used for both text encoding and decoding in generation, and the auto-regressive generation paradigm makes the model sensitive to embedding changes due to error propagation accumulating over time steps.

LoRA-Adaptation connects both worlds. In Table 2, we demonstrate that the two components within EmbSwap can cooperate better after LoRA-Adaptation, leading to significant gains on generation tasks. This results in an average improvement of 5.4% on the largest 27B variant. The results demonstrate the practical use of EmbSwap for efficient zero-shot cross-lingual transfer of instruction-tuned LLMs. See Appendix C.2 and Table 4 for additional results on Aya23 and PaLM2.

Embedding relearning on It LLMs represents a viable alternative to EmbSwap. Table 2 reveals that for smaller model sizes, EmbSwap outperforms It+Lang-Adapt in downstream tasks. However, as model size increases, It+Lang-Adapt becomes the more effective solution. This suggests that It+Lang-Adapt provides a highly effective alternative to EmbSwap for zero-shot cross-lingual transfer involving a single model. However, when handling multiple It models derived from a common base model, EmbSwap offers greater modularity by eliminating the need for costly embedding relearning across individual models.

EmbSwap mitigates catastrophic forgetting. CPT with full-weight updating boosts cross-lingual transfer but suffers from forgetting of English capabilities. ChatVector shows large drops both in English and multilingual tasks, demonstrating the technique in not well suited to our problem. In contrast, EmbSwap shows only minor regressions in English, indicating the benefits of its modular composition which reuses existing knowledge. This advantage is clear in multilingual mathematical reasoning (Figure 4), where EmbSwap, by preserving English knowledge, significantly outperforms full-weight updating methods. Similar findings are observed for Aya23 *etc.* (See Table 4).

3.6 Ablation Analysis

Customized vocabulary amplifies the benefits of training on multilingual data. We decouple

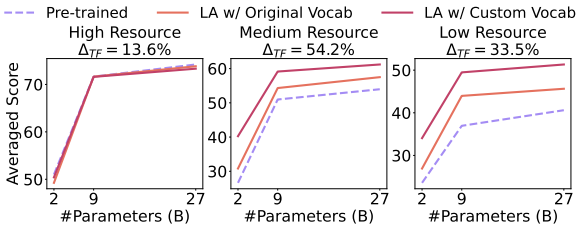

Figure 5: Tokenizer ablations for language adaptation. SEA averaged scores on FLORES-200 and BELEBELE are reported. Δ_{TF}: % tokenizer fertility reduction.

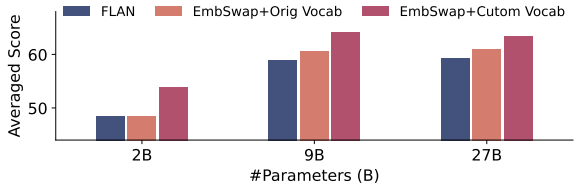

Figure 6: Tokenizer ablations under EMBSWAP. The macro-averaged scores on SEA subset of BELEBELE, SIB-200 and FLORES-200 are reported.

the effects of multilingual data in language adaptation from the change in vocabularies. Figure 5 shows that simply tuning embeddings on multilingual data can significantly improve the performance in medium and low-resource languages, indicating LLMs are under-fitting to these languages. Based on this, employing a new vocabulary customized for these languages can amplify the benefits of embedding tuning on multilingual data. This enhanced learning process facilitates better knowledge acquisition (Zhang et al., 2022; Hofmann et al., 2022). Moreover, the improvements are more pronounced in smaller models, highlighting the importance of effective tokenization for these models to adapt well to low-resource languages. Figure 6 indicates that the importance of customized vocabulary is also apparent in the EMBSWAP setting.

Tokenization fertility is inversely correlated to downstream performance. We study how tokenization fertility (the average number of tokens produced per word) affects the LLM's performance across languages. To ablate this effect, we generate several re-sampled replicates of our tokenizer training datasets with different levels of priority given to high *v.s.* lower resource languages. Specifically, we use temperature sampling to manipulate the training sentences of each language for building different tokenizer training data, and train tokenizers with varying fertilities. We then relearn embeddings for each of these tokenizers and analyze the downstream performance. As shown in

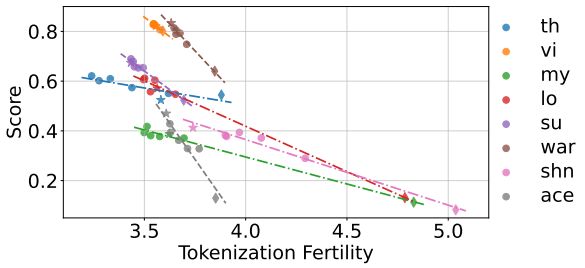

Figure 7: Correlation between the results of language adaptation on `Gemma2-2B` with tokenizer fertility. Normalized `ChrF++` on FLORES-SEA is reported. ♦ and ★ indicate the original and our customized tokenizers used in the other settings, respectively.

Figure 7, we find that the performance is inversely correlated to tokenization fertility, but the correlation is not uniform across languages. Notably, slight reductions in fertility lead to significant performance improvements in low-resource languages (e.g., ACE) while high-resource languages are less sensitive to fertility changes. Furthermore, Latin-script languages generally benefit more from fertility reductions compared to those in non-Latin scripts. Please refer to Appendix Figure 13 which reports similar findings for `PaLM2-XXS`.

Pruning with extension outperforms other tokenizer construction methods. We study how different ways of creating tokenizers affect how well LLMs adapt to new languages. We compared *Prune+Extension* to two variants: 1) *Scratch* trains tokenizers completely new for both English and target languages; and 2) *Extension* appends target language tokens to an existing tokenizer without removing any old ones, which adds 34% extra embedding parameters. Figure 8 shows *Prune+Extension* works the best overall. On a simpler task (FLORES), all methods perform similarly. We believe this is because FLORES is not as complex as BELEBELE, which needs reasoning skills likely learned from English. Supporting this, *Prune+Extension* greatly improves English performance on BELEBELE by keeping the original English tokens, and transfers this advantage to medium and low resource languages. By contrast, *Scratch* has far fewer original English tokens, suggesting that keeping these English tokens is important for retaining knowledge from pre-training.[8] *Prune+Extension* also outperforms *Extension*, showing that removing tokens

[8] It's an open question of whether other facets of the tokenizer need to be retained to preserve other behaviours, e.g., markup tokens to facilitate code understanding.

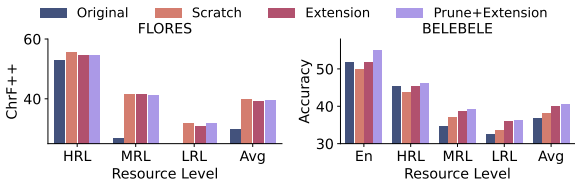

Figure 8: Ablations on tokenizer building methods. We report SEA language adaptation on Gemma2-2B.

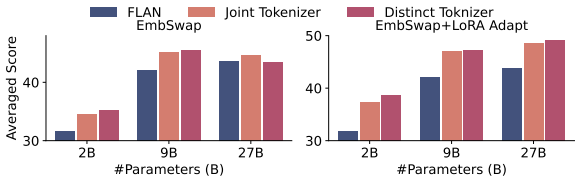

Figure 9: Distinct tokenizers per language group *v.s.* Joint tokenizer for all languages. Averaged results of five tasks included in Table 2 are reported.

irrelevant to target languages is beneficial.[9]

A joint tokenizer for all languages is inferior to language-group specific tokenizers. We test if using one joint tokenizer for SEA, AFR, and IND languages is better than using distinct tokenizers for each group. The joint tokenizer is trained on the combined tokenizer training corpora and has the same vocabulary size as the distinct ones. We relearn a single embedding layer with the joint tokenizer, followed by applying EMBSWAP with LoRA-Adaptation. As shown in Figure 9, although the joint tokenizer improves FLAN across all model scales, it is outperformed by distinct tokenizers. Smaller models show the biggest performance drop with the joint tokenizer. This happens because the shared vocabulary limits how much capacity is dedicated to each language, resulting in about 10% more tokens and making learning less effective.

4 Related Work

Language Adaptation of LLMs. Traditional ways of adapting LLMs to new languages involve continued pre-training on monolingual data (Cui et al., 2024; Zhao et al., 2024a) or multilingual instruction-tuning on synthetic data (Chen et al., 2024; Üstün et al., 2024; Aryabumi et al., 2024). These methods usually focus on monolingual transfer and a few high-resource languages. Moreover, their full-parameter tuning strategy is expensive, being prone to catastrophic forgetting. We achieve efficient adaptation to hundreds of languages through embedding relearning and prevent knowledge forgetting with reusing parts of existing models.

Tokenizer Adaptation and Embedding Relearning. Previous tokenizer adaptation methods introduce languages into models through embedding relearning on new vocabularies (Artetxe et al., 2020; de Vries and Nissim, 2021; Marchisio et al., 2023; Chen et al., 2023). This increases flexibility in handling linguistic diversity and avoids oversegmentation that impairs task performance (Ahuja et al., 2023). However, they only consider encoder-only PLMs and monolingual transfer. We reevaluate this method for large-scale decoder-only LLMs with extension to hundreds of languages.

Many works focus on how to initialize the new embeddings of new vocabularies. They involve complex combinations (Dobler and de Melo, 2023; Liu et al., 2024) of overlapping tokens using auxiliary embeddings and external resources (Minixhofer et al., 2022; Remy et al., 2024), which limits their use across many languages. We avoid external resources by initializing the new embeddings with averaging the corresponding original embeddings (Gee et al., 2022; Mosin et al., 2023), making it scalable for hundreds of languages.

Model Merging. Model merging combines different specialized LLMs to create new ones with all their abilities (Wortsman et al., 2022). This can be done with simple arithmetic on existing parameters (Ilharco et al., 2023). We compared with one such method for instruction following across languages, ChatVector (Huang et al., 2024). Layer Swapping (Bandarkar et al., 2025) retains specific layers from language-adapted models while merging others from task fine-tuned models. Similar to Layer Swapping, EMBSWAP merges embeddings relearned on new vocabulary into instruction-tuned transformer body for cross-lingual transfer.

5 Conclusion

We tackled the challenge of adapting LLMs to diverse languages through embedding relearning. Our empirical findings reveal that embedding tuning with customized tokenizers contributes to effective language adaptation of LLMs, especially for low-resource languages with severe fragmentation. We also demonstrate that these embeddings can be integrated into any instruction-tuned LLMs to enable cross-lingual transfer with minimal training costs, outperforming or matching other strong baselines with high computational cost.

[9] We suspect large vocabulary could increase ambiguity in output embeddings. Evidence also reveals that large vocabularies are not optimal for smaller LLMs (Tao et al., 2024).

Limitations

Although embedding relearning involves only embedding tuning while keeping the transformer body frozen, it still incurs a computational cost of performing a full forward pass and back-propagation through all transformer layers. To address this, more efficient approaches, such as embedding tuning on a subset of transformer layers (Marchisio et al., 2023), could be employed to accelerate the training process.

We achieve substantial language adaptation by learning shared embeddings targeted at groups of geographically related languages, thereby avoiding the monolingual adaptation requiring the creation of language-specific embeddings. In addition, we demonstrate that constructing distinct embeddings for each language group outperforms the approach of treating all languages uniformly. However, this group-specific strategy may block certain forms of cross-lingual transfer and raises the critical question of how to best define 'regions' to facilitate the best transfer.

Although embedding relearning effectively addresses catastrophic forgetting, we still observe a slight regression in English and several other high-resource languages (HRLs). This decline is primarily attributable to the fact that the original LLMs were already extensively pre-trained on these languages. Furthermore, the English and HRL data included in our multilingual dataset $\mathcal{D}_{la}$ is likely of lower quality compared to the original pre-training data. This limitation becomes particularly pronounced in `Gemma2`, where the more advanced `Gemini` models (Anil et al., 2024) are employed for knowledge distillation. We believe that such regressions in English and HRL performance could be alleviated with access to higher-quality data.

While we provide empirical results to demonstrate the effectiveness of embedding relearning and the importance of customized vocabularies tailored to target languages for cross-lingual transfer, future research should explore the conditions under which performing embedding relearning on LLMs is advantageous for cross-lingual transfer and the scenarios where it might be less effective.

There might also be safety concerns arising from the disruption of alignment abilities due to the changes to the embeddings, and how to align the model effectively and efficiently to reject unsafe queries after the EMBSWAP pipeline which would need to be carefully addressed.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

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.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Rohan Anil et al. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Rohan Anil et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. LLM augmented LLMs: Expanding capabilities through composition. In *The Twelfth International Conference on Learning Representations*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *Preprint*, arXiv:2205.03983.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

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.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual

specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. 2022. Fast vocabulary transfer for language model compression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 409–416, Abu Dhabi, UAE. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. 2024. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10943–10959, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2024. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of*

the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Vladislav Mosin, Igor Samenko, Borislav Kozlovskii, Alexey Tikhonov, and Ivan P. Yamshchikov. 2023. Fine-tuning transformers: Vocabulary transfer. *Artificial Intelligence*, 317:103860.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP. In *First Conference on Language Modeling*.

Morgane Riviere et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *Preprint*, arXiv:2404.16789.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. 2024. MAmmoTH2: Scaling instructions from the web. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7725–7738, Miami, Florida, USA. Association for Computational Linguistics.

Overview of Appendix

Our supplementary includes the following sections:

- Section A: Experimental Settings, including implementation details for training and evaluation.
- Section B: Detailed result comparison between instruction-tuned models and EMBSWAP in English tasks.
- Section C: Additional Language Adaptation and EMBSWAP results on PaLM2, Aya23 and Gemma2-IT models.
- Section D: Supplementary analysis including additional data ablations.
- Section E: Results by language for both language adaptation and EMBSWAP.

A Experimental Settings

A.1 Training Details

For language adaptation, we use a constant learning rate of 1×10^{-4} for PaLM2 and 1×10^{-5} for Gemma2 and Aya23 with the Adam optimizer (Kingma and Ba, 2014). The embeddings are trained on a total of 200B tokens,[10] with each batch consisting of examples packed to a maximum sequence length of 2K for PaLM2 and 8K for Gemma2 and Aya23. We pretrain the model using the UL2 objectives (Tay et al., 2023) for PaLM2 and causal language modeling objectives for Gemma2 and Aya23. Language adaptation consumes up to 256 TPU-v5 chips for the largest Gemma2-27B and Aya23-35B models. The batch size is selected based on the model size and computing resources we have, with batch sizes of 256, 128, and 64 assigned to the Gemma2-2B, 9B, and 27B models, respectively. A similar strategy is applied to the Aya23 models. For PaLM2 models, we use the same batch size of 2048 for all variants. We choose the best checkpoint based on the performance of FLORES-200 development sets corresponding to each target language group. The training time varies with model size, where smaller models complete training within 24 hours while larger models require up to 1 week to finish.

We instruction-tune the transformer body of LLMs on $\mathcal{D}_{it}$ using the same hyper-parameter setting to obtain XX-FLAN, where we sample up to 200M instances from the FLAN mixture to construct $\mathcal{D}_{it}$. We use early stopping to select the best model based on the performance on MMLU (Hendrycks et al., 2021) and assemble it with customized embedding to obtain EMBSWAP models. We observe all LLMs converge fast, as indicated by the average performance on MMLU. In most cases, training is completed within 24 hours.

For LoRA-Adaptation, we add LoRA weights to the self-attention module of all transformer layers with a LoRA rank of 64 and exclusively optimize these weights.[11] We use a learning rate of 5×10^{-6} for all models with 10% steps of warm-up. Analogous to embedding tuning, the FLORES-200 development set is used for model selection. The training process is computationally efficient, completing within 12 hours even for the largest model.

Dataset	Prompt
BELEBELE	The following are multiple choice questions (with answers). Passage: [Target Language Passage] Question: [Target Language Question] (A) [Choice A] (B) [Choice B] (C) [Choice C] (D) [Choice D] Answer:
SIB-200	[News Article in Target Language] Question: What label best describes this news article? (A) science/technology (B) travel (C) politics (D) sports (E) health (F) entertainment (G) geography Answer:
FLORES-200	Translate this from English to [Target Language Name]: English: [Sentence in English] [Target Language Name]:
XSUM-IN	I will first show a news article in English and then provide a one sentence summary of it in [Target Language Name]. Summarize the following article: [Article in English] Summary in [Target Language Name]:
XORQA-IN	Generate an answer in [Target Language Name/English] for the question based on the given passage: [Passage in English] Q: [Question in Target Language] A:
GSM8K-NTL	Q: [Question in Target Language] A: [Let's think step by step.]

Table 3: Prompt templates used in each of the evaluation dataset. For few-shot evaluation, n-shot examples have the same format as the last test instance, which comes after the preamble but before the test instances.

[10] English accounts for 30.8% of the tokens.

[11] This adds less than 1% parameters.

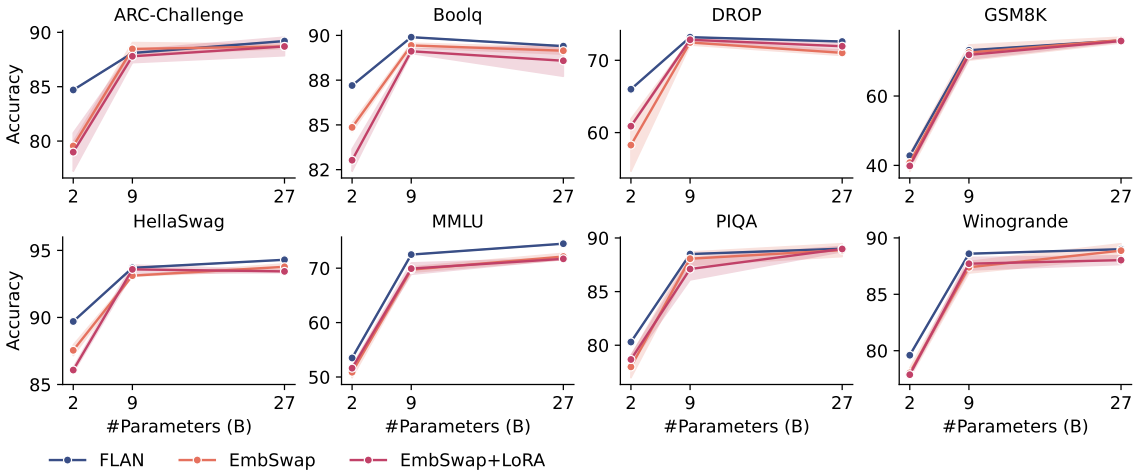

Figure 10: Performance on English tasks. For EMBSWAP methods, we present the averaged performance with variances across three embeddings (i.e., SEA, AFR, IND). Shading indicates the standard deviations measured over three embeddings.

A.2 Evaluation Details

We use the prompt formats listed in Table 3 for evaluation. For generation tasks, greedy decoding is employed with a maximum sequence length of 256 tokens. For classification tasks, we calculate the logits of each available option (e.g., (A), (B)) and select the option with the highest score as the predicted answer.

B Details on English Benchmarking

For benchmarking performance in English, we choose ARC-Challenge (Clark et al., 2018), Boolq (Clark et al., 2019), DROP (Dua et al., 2019), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), PIQA (Bisk et al., 2019), and Winogrande (Sakaguchi et al., 2021). We use different number of shots for evaluation following Riviere et al. (2024).

Figure 10 shows the comparison between the instruction-tuned FLAN model and EMBSWAP across various sizes of `Gemma2`. Both EMBSWAP variants exhibit performance regressions across all tasks compared to the FLAN model, although the performance gap closes as model capacity scales up. We believe that these minor regressions are justifiable in light of the substantial gains achieved in zero-shot cross-lingual transfer.

C Additional Main Results

C.1 Language Adaptation on `PaLM2` and `Aya23`

We evaluate the generalization ability of language adaptation on two LLMs with varying levels of multilingualism. Among them, `PaLM2-S` exhibits the strongest multilingual abilities while `Aya23` models demonstrate limited multilingual performance. Figure 11 shows that the performance gains from language adaptation decrease as the original multilingual scope of the LLMs expand (`Aya23` → `PaLM2-XXS` → `PaLM2-S`). Moreover, larger performance improvements are observed on `Aya23` models when scaling up their size, suggesting that language adaptation may be particularly effective for models with stronger English proficiency.

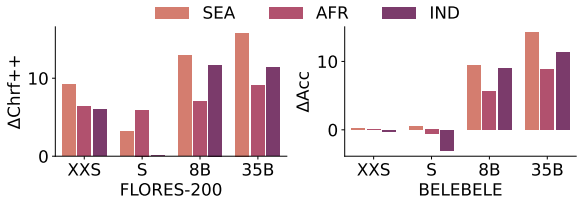

Figure 11: Language Adaptation on `PaLM2` (XXS, S) and `Aya23` (8B, 35B). Absolute gains over the pre-trained models are reported.

C.2 EMBSWAP Results on `PaLM2`, `Aya23` and `Gemma2-IT`

Employing the same EMBSWAP pipeline, we observe similar patterns of performance improvement across all sizes of the `PaLM2` and `Aya23` models, as shown in Table 4. This demonstrates the broad applicability of EMBSWAP to various types of LLMs. In addition, we also show the detailed results of `Gemma2-IT` models. EMBSWAP also performs effectively with off-the-shelf IT models that have undergone complex supervised fine-tuning and reinforcement learning. This further underscores the versatility of EMBSWAP in integrating pre-trained multilingual embeddings into LLMs that have been instruction-tuned using diverse methodologies for

Task Type		Classification						Generation							Avg.
Eval. Metric	English	BeleBele Accuracy			Sib-200 Accuracy			Flores-200 ChrF++			XorQA-In Token-F1		XSum-In ChrF		
Model		Sea	Afr	Ind	Sea	Afr	Ind	Sea	Afr	Ind	Ind	En	Ind	En	
PaLM2-XXS-FLAN	59.7	54.5	40.0	49.4	67.2	51.3	70.9	28.5	13.9	21.0	9.7	42.1	**4.8**	33.7	32.8
PaLM2-XXS-FA	**62.6**	58.0	40.2	50.9	**76.8**	62.2	**77.2**	27.7	12.2	17.9	10.4	59.4	2.1	**34.2**	35.9
★ LoRA-Adapt	58.9	**59.7**	**43.4**	**53.8**	73.7	**63.5**	76.9	**32.3**	**16.0**	**26.7**	**11.7**	**64.2**	1.0	30.7	**37.7**
PaLM2-S-FLAN	**86.3**	80.2	67.4	**82.1**	70.6	60.9	74.5	37.6	19.3	36.7	**21.3**	46.9	15.8	41.1	45.4
PaLM2-S-FA	85.1	83.2	67.1	79.9	77.0	66.9	74.6	36.1	23.1	34.6	21.1	51.8	**17.3**	40.1	47.1
★ LoRA-Adapt	85.7	**83.6**	**68.4**	81.8	**77.7**	**68.7**	**77.8**	**39.3**	**24.1**	**38.7**	18.9	**57.3**	15.2	**41.2**	**48.2**
Aya23-8B-FLAN	**74.3**	47.8	34.1	42.7	64.3	48.5	60.6	22.9	6.0	16.1	10.5	58.5	**11.7**	**34.5**	32.3
Aya23-8B-FA	71.1	56.6	38.4	48.0	**72.4**	**58.7**	**71.3**	28.8	9.1	11.5	7.8	**61.6**	3.9	25.7	34.2
★ LoRA-Adapt	71.3	**59.1**	**39.0**	**50.8**	65.6	55.3	65.2	**37.1**	**16.9**	**33.4**	9.6	56.5	9.4	30.7	**36.8**
Aya23-35B-FLAN	**83.6**	56.9	40.3	54.6	67.0	52.9	67.2	27.1	9.1	23.6	11.4	64.7	10.4	**37.9**	36.5
Aya23-35B-FA	79.2	57.6	40.2	53.0	70.8	53.8	68.5	27.7	9.5	13.9	6.8	62.3	7.3	22.1	35.2
★ LoRA-Adapt	82.0	**72.6**	**50.1**	**65.2**	**75.2**	**64.7**	**74.0**	**41.6**	**23.4**	**40.1**	**12.6**	**67.8**	**15.7**	37.1	**45.1**
Gemma2-2B-IT	55.5	47.5	35.4	45.8	57.4	42.8	61.2	24.2	7.3	17.7	14.9	53.0	10.1	31.9	31.6
+ Lang-Adapt	54.1	51.2	36.7	42.9	67.0	48.7	67.1	35.4	14.2	29.2	12.3	40.2	7.1	27.5	33.0
Gemma2-2B-FA	55.1	53.7	37.4	46.0	66.7	49.5	64.0	27.4	11.7	27.0	**16.7**	58.7	**13.4**	**32.0**	35.9
★ LoRA Adapt	**58.4**	**56.4**	**39.9**	**50.7**	**67.5**	**53.0**	**69.9**	**37.6**	**18.6**	**32.4**	11.7	**60.5**	9.3	27.7	**37.6**
Gemma2-9B-IT	**79.9**	71.5	51.4	72.5	72.5	57.8	79.3	32.2	14.0	31.1	**24.0**	64.4	16.6	**34.3**	44.5
+ Lang-Adapt	73.9	75.4	57.9	**72.7**	77.2	69.2	79.7	39.3	22.2	38.7	23.1	**69.6**	**17.0**	33.9	48.3
Gemma2-9B-FA	78.5	**77.4**	54.5	72.6	78.2	66.1	78.5	33.8	16.6	18.8	21.4	65.2	9.6	34.2	43.8
★ LoRA-Adapt	79.5	76.5	**58.1**	70.8	**81.3**	**71.2**	**82.8**	**41.5**	**25.1**	**39.8**	18.2	68.2	15.7	33.0	**48.4**
Gemma2-27B-IT	82.1	74.9	54.6	75.9	76.2	63.1	**83.0**	35.2	20.1	34.5	**27.7**	68.2	**19.0**	34.7	48.0
+ Lang-Adapt	81.9	**81.7**	**62.7**	**76.0**	**82.1**	70.0	81.9	**42.5**	24.8	39.3	26.4	**73.0**	18.3	**35.8**	**51.1**
Gemma2-27B-FA	79.9	80.8	59.5	75.9	81.4	68.9	**83.0**	35.1	6.2	29.9	20.0	68.8	15.0	34.6	46.7
★ LoRA-Adapt	**82.2**	78.8	60.1	74.3	81.7	**71.7**	82.6	42.3	**25.6**	**40.6**	24.2	67.9	15.6	34.0	49.6

Table 4: Additional EmbSwap results on PaLM2, Aya23 and Gemma2-IT models. The best and second-best results are marked in **bold** and underlined. Red values indicate EmbSwap hurts the performance.

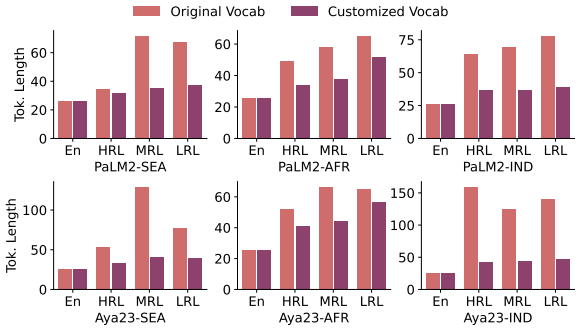

Figure 12: The tokenization comparison between using the vanilla and customized multilingual tokenizers on Gemma2. Tok. Length refers to the average number of tokens required to represent the same amount of texts.

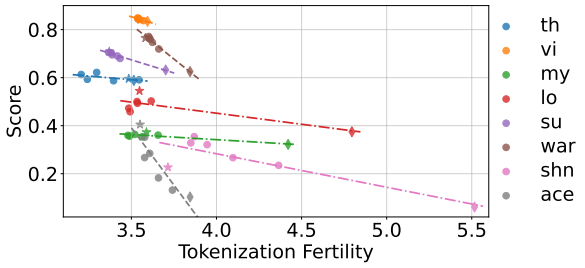

Figure 13: Correlation between the performance of language adaptation on PaLM2-XXS with tokenizer fertility. Normalized ChrF++ on Flores-Sea are reported. ♦ and ★ indicate the original and customized tokenizers in PaLM2.

efficient zero-shot cross-lingual transfer.

D Supplementary Analysis

Additional tokenization results on the customized vocabularies. We present additional results on tokenization fertility using customized vocabularies developed for PaLM2 and Aya23. Figure 12 shows similar patterns as those reported for Gemma2, where the fertility tokenization for both MRLs and LRLs shows a substantial decrease, while the English tokenization remains roughly unchanged. This effect is particularly pronounced in Aya23, where we observe over ×3 reduction in fertility for Sea and Ind languages. In Table 6, we also show a few tokenized examples for low-resource languages. We find that the customized tokenizer produces more meaningful tokens and avoids overtokenization.

Tokenizer fertility is also inversely correlated to downstream performance in PaLM2. We replicate the analysis shown in Figure 7 using PaLM2-XXS, with results shown in Figure 13. We observe patterns analogous to those reported for Gemma2-2B, wherein reduced tokenizer fertility is generally associated with improved downstream performance

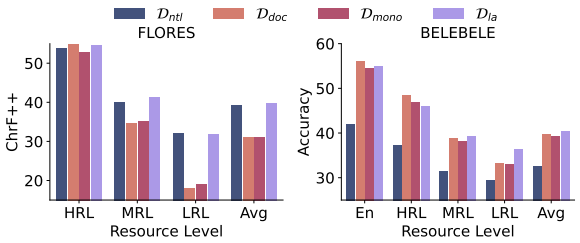

Figure 14: Pre-training data ablation for language adaptation. $\mathcal{D}_{ntl}$: multilingual data sampled from the NTL corpus; $\mathcal{D}_{doc}$: multilingual data sampled Wikipedia and mC4; $\mathcal{D}_{la}$: our final data mixture for language adaptation, i.e., $\mathcal{D}_{la} = \mathcal{D}_{ntl} \cup \mathcal{D}_{doc}$; $\mathcal{D}_{mono}$: monolingual data by excluding parallel sentences from $\mathcal{D}_{la}$. SEA languages results on `Gemma2-2B` are reported.

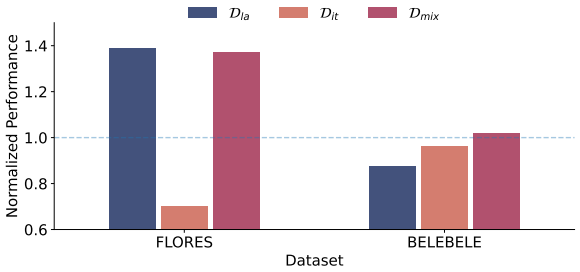

Figure 15: Training data ablation for Lora-Adaptation. $\mathcal{D}_{la}$: multilingual data for embedding tuning; $\mathcal{D}_{it}$: the FLAN mixture for instruction-tuning; $\mathcal{D}_{mix}$: a combination of equally sub-sampled $\mathcal{D}_{la}$ and $\mathcal{D}_{it}$. Averaged SEA language results (normalized score *v.s.* EMBSWAP) on `Gemma2-2B` are reported.

and this relationship is particularly pronounced in LRLs and languages written in non-Latin script.

Long-tail NTL and document-level data are both important for language adaptation. We perform language adaptation on `Gemma2-2B` in SEA languages with different data mixture. Figure 14 reveals that improved performance in long-tail languages primarily stems from the inclusion of NTL data, while document-level data plays a crucial role in preserving knowledge for high-resource languages. The significance of incorporating document-level data is further underscored by the results on BELEBELE, where the removal of $\mathcal{D}_{doc}$ leads to a substantial performance decline across all resource levels. This finding highlights the importance of $\mathcal{D}_{doc}$ in maintaining the ability of LLMs in processing various types of texts. In addition, excluding parallel data from the training mixture (i.e., $\mathcal{D}_{mono}$) leads to a significant decline in performance on translation tasks. Moreover, the performance of LRLs on BELEBELE also declines substantially, indicating the critical role of parallel data in enhancing tasks beyond translation through facilitated cross-lingual transfer (Anil et al., 2023).

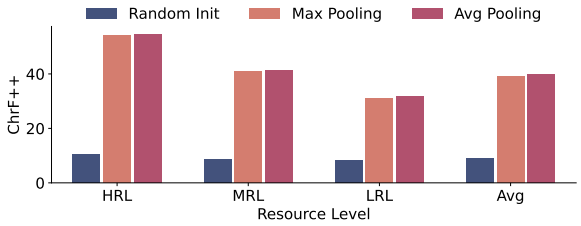

Figure 16: Ablations on embedding initialization methods. FLORES-SEA language performance of language adaptation on `Gemma2-2B` is reported. `Max Pooling`: for each new token in the customized vocabulary, we use the original tokenizer to tokenize it and apply max pooling over the embeddings of the corresponding subtokens as the initialization.

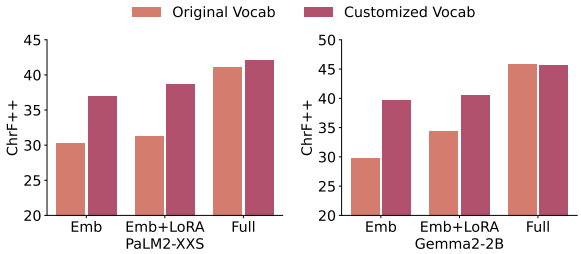

Figure 17: The effects of using customized vocabularies with different proportions of tuned parameters. The averaged score on SEA languages of FLORES-200 is reported.

Both multilingual and instruction data are important for LoRA-Adaptation We investigate the impact of employing multilingual $\mathcal{D}_{la}$ and instruction-tuning data $\mathcal{D}_{it}$ in LoRA-Adaptation. As shown in Figure 15, the removal of either $\mathcal{D}_{la}$ or $\mathcal{D}_{it}$ harms the performance compared to the vanilla EMBSWAP model on FLORES-200 or BELEBELE, while combining both datasets enables the adapted LLM to achieve the best overall results.

Employing the original embeddings for initialization is essential to language adaptation. In Figure 16, we show that without initializing the customized embeddings using the original embeddings from the LLM, the language adaptation does not perform well at all, yielding `ChrF++` scores below 10 even for HRLs. In contrast, initializing with the original embeddings significantly improves the effectiveness of language adaptation, where the average pooling method slightly outperforms the max pooling variant.

The benefits of employing customized vocabulary decrease with more tuned parameters. We study the effects of employing customized embeddings with varying ratios of tuned parameters. As shown in Figure 17, the benefits of using customized embeddings diminish as the number of tuned parameters increases (Emb → Emb+LoRA

Size	Vocab	HRL	MRL	LRL
2B	Original	59.1	55.6	57.1
	Custom	61.1 ↑3.38%	61.2 ↑10.07%	60.3 ↑5.60%
9B	Original	30.2	27.2	28.6
	Custom	33.8 ↑11.92%	33.3 ↑22.43%	32.1 ↑12.24%
27B	Original	17.1	15.0	16.0
	Custom	19.6 ↑14.62%	19.2 ↑28.0%	18.4 ↑15.0%

Table 5: Comparing latency for using original and customized vocabularies in EMBSWAP. The number of instances processed per second (i.e., prefilling) by Gemma2 are reported. We use the passages in BELEBELE for all SEA, AFR, and IND languages.

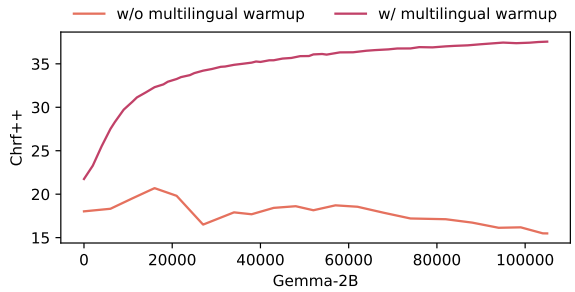

Figure 18: Learning curves of language adaptation on LLMs with limited multilingual abilities. Averaged results on SEA languages of FLORES-200 are reported.

→ Full). Specifically, on Gemma2-2B model, customized embeddings show no advantage when full-parameter tuning is employed. This phenomenon arises because increasing the number of tuned parameters allocates greater model capacity for language adaptation, which simplifies the adaptation process compared to relying solely on embedding tuning. Nonetheless, full-parameter tuning could exacerbate the problem of catastrophic forgetting, while embedding tuning provides a safer alternative. Furthermore, the use of customized embeddings amplifies the advantages of embedding tuning, making it a promising technique.

Customized vocabulary improves latency. We evaluate latency by measuring the number of texts processed per second by LLMs. We use the passages from all SEA, AFR, and IND languages in BELEBELE as test texts. For comparison, we test EMBSWAP models integrated with embeddings trained using either the original Gemma2 tokenizer or our customized one. Table 5 shows that employing customized tokenizer consistently improves latency, particularly in MRLs and LRLs. This trend becomes increasingly pronounced as model size scales, highlighting the importance of customized tokenizers in achieving low-latency processing for long-tail languages in larger LLMs.

Multilingual warmup is necessary for embedding relearning on LLMs with limited multilingual abilities. We investigate whether embedding tuning with customized embeddings is a universal technique for the language adaptation of any types of LLMs. As shown in Figure 18, simply performing embedding tunning on Gemma-2B, a model with very limited multilingual capabilities, does not successfully adapt it to various languages. In contrast, when a multilingual continued pre-training process is conducted prior to embedding tunning, where the document-level data $\mathcal{D}_{doc}$ is used to warm up the LLM, we observe consistent improvements throughout the training process. This suggests that, for LLMs, a good initial multilingual ability is essential for the success of embedding relearning.

E Results by Language

Table 7 presents an overview of languages available across our evaluation benchmarks. We show the per-language results on each task for both language adaptation (Tables 8 – 10) and EMBSWAP (Tables 11 – 16).

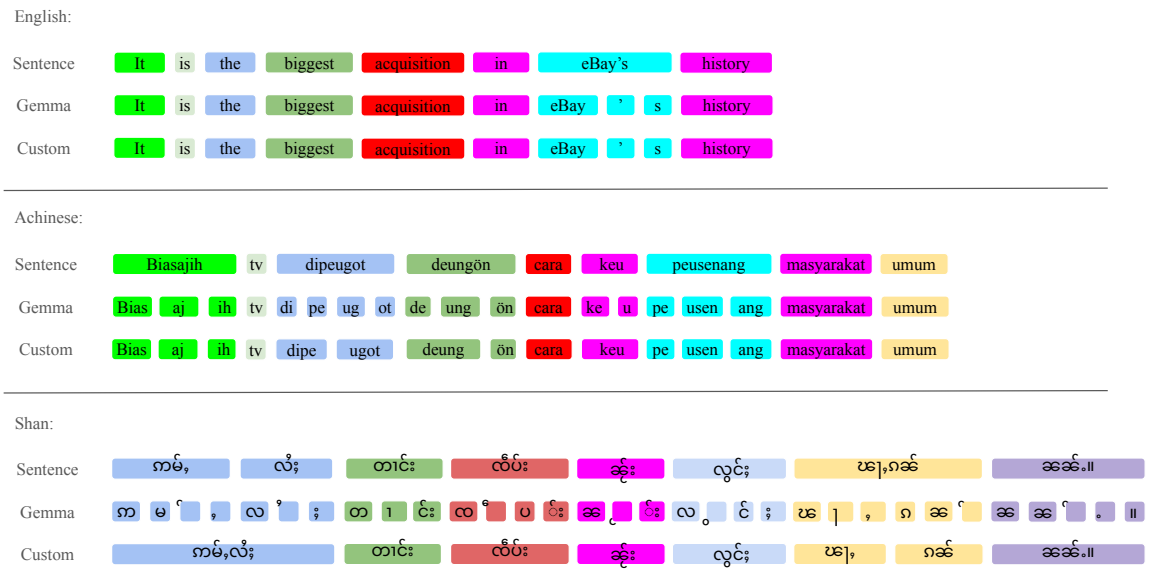

Table 6: Qualitative examples for comparing tokenization using the customized vocabulary against Gemma's original one. We find that the customized tokenizer reduces overtokenization without affecting English tokenization. Sentences have been word-segmented to simplify presentation.

Language name	ISO code	BELEBELE	SIB-200	FLORES-200	XORQA-IN	XSUM-IN	GSM8K-NTL
English	eng_Latn	✓	✓		✓	✓	
			SEA				
Achinese (Arabic)	ace_Arab^{L}		✓	✓			
Achinese	ace_Latn^{L}		✓	✓			
Balinese	ban_Latn^{L}		✓	✓			
Betawi	bew_Latn^{M}						✓
Banjar (Arabic)	bjn_Arab^{L}		✓	✓			
Banjar	bjn_Latn^{L}		✓	✓			
Buginese	bug_Latn^{L}		✓	✓			
Cebuano	ceb_Latn^{M}	✓	✓	✓			
Ilocano	ilo_Latn^{M}	✓	✓	✓			✓
Indonesian	ind_Latn^{H}	✓	✓	✓			
Javanese	jav_Latn^{M}	✓	✓	✓			
Kachin	kac_Latn^{L}	✓	✓				
Khmer	khm_Khmr^{M}	✓	✓	✓			
Lao	lao_Laoo^{M}	✓	✓	✓			
Kedah Malay	meo_Latn^{M}						✓
Pattani Malay	mfa_Arab^{L}						✓
Minangkabau (Arabic)	min_Arab^{L}		✓	✓			
Minangkabau	min_Latn^{L}		✓	✓			✓
Myanmar (Burmese)	mya_Mymr^{M}	✓	✓	✓			
Pangasinan	pag_Latn^{L}		✓	✓			
Shan	shn_Mymr^{L}	✓	✓	✓			
Sundanese	sun_Latn^{M}	✓	✓	✓			
Tagalog	tgl_Latn^{H}		✓	✓			
Thai	tha_Thai^{H}	✓	✓	✓			
Vietnamese	vie_Latn^{H}	✓	✓	✓			
Waray (Philippines)	war_Latn^{L}	✓	✓	✓			
Standard Malay	zsm_Latn^{H}		✓	✓			
			AFR				
Afrikaans	afr_Latn^{H}	✓	✓				
Twi	aka_Latn^{L}		✓				
Amharic	amh_Ethi^{H}	✓	✓				
Bambara	bam_Latn^{L}	✓	✓				
Bemba (Zambia)	bem_Latn^{L}		✓	✓			
Chokwe	cjk_Latn^{L}		✓	✓			

Language name	ISO code	BELEBELE	SIB-200	FLORES-200	XORQA-IN	XSUM-IN	GSM8K-NTL
Dinka	din_Latn^{L}			✓			
Dyula	dyu_Latn^{L}		✓				
Efik	efi_Latn^{L}						✓
Ewe	ewe_Latn^{L}		✓	✓			
Fon	fon_Latn^{L}		✓	✓			
Fulfulde	ful_Latn^{L}			✓			
Nigerian Fulfulde	fuv_Latn^{L}	✓	✓				
Hausa	hau_Latn^{H}	✓	✓				
Igbo	ibo_Latn^{H}	✓	✓	✓			
Kamba (Kenya)	kam_Latn^{L}		✓	✓			
Kabiyè	kbp_Latn^{L}		✓	✓			
Kabuverdianu	kea_Latn^{L}	✓	✓				
Kikuyu	kik_Latn^{L}		✓	✓			
Kinyarwanda	kin_Latn^{M}	✓	✓	✓			
Kimbundu	kmb_Latn^{L}		✓	✓			
Central Kanuri (Arabic)	knc_Arab^{L}			✓			
Central Kanuri	knc_Latn^{L}			✓			
Kongo	kon_Latn^{L}		✓	✓			
Krio	kri_Latn^{L}						✓
Lingala	lin_Latn^{L}	✓	✓	✓			
Tshiluba (Luba-Lulua)	lua_Latn^{L}		✓	✓			
Luganda	lug_Latn^{M}	✓	✓	✓			
Luo	luo_Latn^{L}	✓	✓	✓			
Mossi	mos_Latn^{L}		✓	✓			
Sepedi	nso_Latn^{L}	✓	✓	✓			✓
Nuer	nus_Latn^{L}		✓	✓			
Chichewa (Zambia)	nya_Latn^{L}	✓	✓	✓			
Oromo	orm_Latn^{M}			✓			
Nigerian Pidgin	pcm_Latn^{L}						✓
Rundi	run_Latn^{L}		✓	✓			
Sango	sag_Latn^{L}		✓	✓			
Shona	sna_Latn^{M}	✓	✓	✓			
Somali	som_Latn^{H}	✓	✓	✓			
Sesotho	sot_Latn^{H}	✓	✓	✓			
Swati	ssw_Latn^{L}	✓	✓	✓			
Swahili	swa_Latn^{H}			✓			
Tamasheq (Latin)	taq_Latn^{L}		✓	✓			
Tamasheq (Tifinagh)	taq_Tfng^{L}		✓	✓			
Tigrinya	tir_Ethi^{M}	✓	✓	✓			
Tswana	tsn_Latn^{M}	✓	✓	✓			
Tsonga	tso_Latn^{L}	✓	✓	✓			✓
Tumbuka	tum_Latn^{L}		✓	✓			
Umbundu	umb_Latn^{L}		✓	✓			
Wolof	wol_Latn^{L}	✓	✓	✓			
Xhosa	xho_Latn^{M}	✓	✓	✓			
Yoruba	yor_Latn^{H}	✓	✓	✓			
Zulu	zul_Latn^{H}	✓	✓	✓			
			IND				
Assamese	asm_Beng^{M}	✓	✓	✓	✓	✓	✓
Awadhi	awa_Deva^{L}		✓	✓	✓	✓	
Bengali	ben_Beng^{H}	✓	✓	✓	✓	✓	
Bengali (Latin)	ben_Latn^{M}	✓					
Haryanvi	bgc_Deva^{L}				✓	✓	
Bhojpuri	bho_Deva^{L}		✓	✓	✓	✓	✓
Tibetan	bod_Tibt^{M}	✓		✓	✓	✓	
Bodo (India)	brx_Deva^{L}				✓	✓	
Dhivehi	div_Thaa^{M}						✓
Dogri	doi_Deva^{L}						✓
Dzongkha	dzo_Tibt^{M}		✓				✓
Garhwali	gbm_Deva^{L}				✓	✓	
Goan Konkani	gom_Deva^{L}				✓	✓	✓
Gujarati	guj_Gujr^{H}		✓	✓	✓	✓	
Hindi	hin_Deva^{H}	✓	✓	✓	✓	✓	

Language name	ISO code	BELEBELE	SIB-200	FLORES-200	XORQA-IN	XSUM-IN	GSM8K-NTL
Hindi (Latin)	hin_Latn^{M}	✓					
Chhattisgarhi	hne_Deva^{L}		✓	✓	✓	✓	
Hadothi	hoj_Deva^{L}				✓	✓	
Kannada	kan_Knda^{H}	✓	✓	✓	✓	✓	
Kashmiri	kas_Arab^{L}		✓				
Kashmiri (Devanagari)	kas_Deva^{L}		✓				
Mizo	lus_Latn^{M}		✓				
Magahi	mag_Deva^{L}		✓				
Maithili	mai_Deva^{L}		✓	✓	✓	✓	✓
Malayalam	mal_Mlym^{H}	✓	✓	✓	✓	✓	
Marathi	mar_Deva^{H}	✓	✓	✓	✓	✓	
Meitei (Bengali)	mni_Beng^{L}		✓	✓	✓	✓	✓
Malvi	mup_Deva^{L}				✓	✓	
Marwari	mwr_Deva^{L}				✓	✓	
Mazanderani	mzn_Arab^{M}						✓
Nepali	npi_Deva^{M}			✓		✓	
Odia	ory_Orya^{M}	✓		✓	✓	✓	✓
Punjabi	pan_Guru^{M}	✓	✓	✓	✓	✓	
Southern Pashto	pbt_Arab^{M}			✓			
Pashto	pbu_Arab^{M}				✓	✓	
Sanskrit	san_Deva^{M}		✓	✓	✓	✓	
Santali (Ol Chiki)	sat_Olck^{L}		✓		✓	✓	
Sinhala	sin_Latn^{L}	✓					
Sinhala	sin_Sinh^{H}	✓	✓				
Sindhi (Perso-Arabic)	snd_Arab^{M}	✓	✓				
Tamil	tam_Taml^{H}	✓	✓	✓	✓	✓	
Telugu	tel_Telu^{H}	✓	✓	✓	✓	✓	
Urdu	urd_Arab^{H}	✓	✓	✓	✓	✓	
Urdu	urd_Latn^{L}	✓					

Table 7: Overview of languages included in each of our evaluation benchmark. H, M, L indicate high, medium, and low-resource languages, respectively.

Language Adaptation Results on FLORES-200 EN-XX

Model	PaLM2				Gemma2						Aya23			
Size	XXS		S		2B		9B		27B		8B		35B	
Variant	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA
SEA														
ind_Latn	65.9	65.0	70.7	69.1	64.4	64.6	68.9	68.3	69.9	68.5	73.2	69.5	71.9	68.6
tha_Thai	42.9	40.9	50.3	49.8	35.6	35.3	44.7	43.8	47.4	45.6	20.3	33.9	28.4	37.0
vie_Latn	56.4	56.4	61.7	61.6	52.1	54.2	58.7	58.3	60.0	59.8	65.7	63.5	64.9	62.3
zsm_Latn	63.0	64.6	68.0	67.4	58.4	63.1	64.6	65.6	65.6	65.9	54.2	64.6	57.4	64.9
tgl_Latn	54.2	55.1	62.1	61.4	48.4	55.2	57.8	58.3	60.4	59.7	36.7	55.3	47.4	57.9
mya_Mymr	22.4	26.9	40.9	37.4	12.9	27.5	25.4	33.5	30.1	33.3	6.5	23.4	7.3	29.5
lao_Laoo	30.6	37.8	48.8	47.9	9.5	39.0	26.7	42.9	33.7	42.1	4.0	27.0	10.2	41.1
khm_Khmr	24.7	29.0	36.7	36.4	12.7	25.9	24.3	28.0	27.4	32.0	2.2	20.6	8.7	26.3
ceb_Latn	41.7	56.3	58.5	60.6	31.8	54.0	50.6	58.2	54.9	59.1	25.8	55.6	35.1	57.2
jav_Latn	43.9	50.1	53.9	53.3	26.4	48.7	43.5	51.6	47.0	52.6	25.8	51.4	30.1	52.6
sun_Latn	38.9	48.1	51.1	52.0	21.8	45.1	39.4	44.5	41.7	47.8	29.2	43.7	29.2	48.7
ilo_Latn	19.3	42.9	45.1	51.8	19.0	43.3	32.0	49.0	42.6	52.2	18.0	45.5	20.0	49.9
war_Latn	29.9	51.6	54.6	59.2	33.7	54.1	49.0	58.2	52.4	59.3	25.6	52.5	35.2	55.5
bug_Latn	7.9	15.9	16.5	25.0	14.1	22.2	18.1	28.9	21.7	28.5	15.4	21.9	18.8	27.2
pag_Latn	14.5	32.8	28.2	43.6	25.1	38.6	28.2	43.3	34.2	43.8	17.9	39.2	20.5	43.0
shn_Mymr	4.4	17.7	3.0	21.0	11.6	29.6	14.9	32.4	16.2	32.7	4.6	9.2	5.0	28.4
min_Latn	25.0	46.9	49.8	53.6	23.8	47.7	38.4	51.3	45.9	52.8	35.1	47.6	36.0	50.8
ace_Latn	8.2	29.0	25.7	40.0	9.4	32.5	16.1	37.0	26.8	40.3	16.0	32.6	15.2	38.3
ban_Latn	13.7	41.3	39.4	42.4	21.7	44.4	30.1	46.8	37.1	47.5	21.8	45.1	24.7	47.2
bjn_Latn	21.6	26.1	47.2	44.5	24.9	29.9	36.3	38.0	42.7	40.6	36.8	32.6	36.7	35.5
ace_Arab	2.7	3.5	7.9	10.1	3.0	15.7	3.9	15.7	9.3	18.0	3.9	1.6	2.2	14.8
bjn_Arab	2.7	4.5	9.8	11.3	4.7	11.0	6.5	11.9	12.0	20.6	3.9	7.3	4.1	19.2
min_Arab	3.4	5.2	10.9	16.2	5.1	20.0	6.6	17.9	10.3	20.9	5.3	1.1	2.7	18.8
Avg.	27.7	**36.9**	40.9	**44.1**	24.8	**39.2**	34.1	**42.8**	38.7	**44.5**	23.8	**36.7**	26.6	**42.4**

Language Adaptation Results on FLORES-200 EN-XX														
Model	PaLM2				Gemma2						Aya23			
Size	XXS		S		2B		9B		27B		8B		35B	
Variant	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA
AFR														
swa_Latn	52.2	54.4	63.8	62.1	40.3	51.1	57.0	58.7	61.5	59.5	9.8	39.9	15.1	49.6
lin_Latn	9.0	25.7	20.6	40.8	7.9	30.2	14.7	42.2	21.0	42.1	9.9	30.8	12.1	35.1
yor_Latn	9.5	14.9	24.8	23.0	6.3	14.7	11.6	21.6	18.4	22.8	4.0	13.6	5.9	16.9
ful_Latn	4.0	6.2	5.5	13.4	8.0	10.0	5.3	11.0	6.0	16.0	6.4	11.2	7.8	11.0
ibo_Latn	23.2	30.1	39.0	38.9	8.7	23.8	20.5	34.4	29.6	35.9	5.0	22.6	7.0	27.3
orm_Latn	4.6	8.8	12.1	24.6	8.3	14.1	9.0	25.1	11.2	25.7	9.2	12.7	9.3	15.8
som_Latn	24.6	32.5	42.5	42.4	11.2	29.0	26.4	38.4	32.3	40.0	17.6	28.5	25.4	37.3
tso_Latn	9.8	20.3	16.4	38.4	9.3	20.7	16.0	37.0	26.3	40.7	7.6	17.3	9.2	18.4
nya_Latn	25.5	37.5	44.8	44.6	9.0	28.1	17.9	40.1	27.5	40.2	7.7	26.5	10.9	30.2
zul_Latn	27.2	39.0	49.0	47.9	10.3	25.2	24.4	39.5	32.4	40.6	8.3	22.1	7.8	26.1
kin_Latn	5.9	22.8	25.1	37.0	7.8	15.9	16.0	29.7	22.5	33.9	8.3	10.8	8.3	20.3
run_Latn	5.0	18.6	17.0	30.4	5.5	14.2	11.6	26.3	18.0	27.8	6.3	11.8	7.1	15.2
sna_Latn	25.3	33.5	40.3	39.9	8.3	22.8	18.0	33.6	26.5	35.5	6.9	18.2	9.4	22.3
xho_Latn	26.3	35.6	44.4	44.5	12.2	26.9	23.5	36.5	32.2	38.4	9.1	22.4	8.0	28.1
tsn_Latn	10.4	26.5	34.1	42.2	7.8	21.0	17.6	39.7	30.6	42.1	8.4	19.5	10.5	24.7
tir_Ethi	3.0	5.1	13.2	13.1	3.0	14.0	4.4	20.5	5.4	18.9	2.2	6.6	2.1	9.4
kik_Latn	5.9	9.0	9.6	12.6	9.1	11.0	9.5	9.0	12.0	11.2	7.1	7.8	7.0	8.2
kon_Latn	8.1	28.6	16.3	38.7	8.8	33.4	13.8	38.9	16.0	38.2	7.8	31.7	9.9	35.6
lua_Latn	10.2	19.1	7.3	20.9	8.4	19.7	8.8	27.5	11.8	30.5	7.1	11.8	9.4	18.7
umb_Latn	5.6	3.6	6.2	7.1	7.9	5.8	7.3	8.1	8.6	8.5	6.4	7.0	6.9	7.1
sot_Latn	21.1	35.6	50.0	50.5	7.9	30.2	21.0	42.5	34.3	42.9	8.1	29.3	10.9	31.7
mos_Latn	3.8	7.7	3.5	5.6	5.7	5.4	5.9	4.6	6.9	8.3	4.7	4.3	4.3	4.7
nso_Latn	10.0	29.9	33.5	46.2	7.5	20.5	16.7	42.3	31.3	45.0	8.4	15.8	10.8	25.9
knc_Latn	5.3	7.9	6.5	8.4	8.3	9.4	5.9	12.5	7.6	12.8	6.9	8.6	5.8	9.6
knc_Arab	7.9	4.5	6.1	5.5	6.2	5.9	4.4	3.9	7.8	8.9	8.2	3.5	9.0	5.4
luo_Latn	6.7	9.5	8.5	15.2	12.2	11.0	10.1	18.1	13.0	22.2	5.9	7.9	7.7	10.8
bem_Latn	6.2	8.6	19.3	23.4	7.0	13.0	11.3	17.5	17.3	22.2	7.7	12.1	9.8	10.0
lug_Latn	5.4	13.9	15.6	30.6	6.4	13.4	9.1	22.1	16.0	27.1	6.4	11.4	8.0	12.5
wol_Latn	3.3	7.4	11.9	18.9	5.5	9.9	7.3	15.0	9.2	17.0	6.9	6.1	8.0	8.9
kmb_Latn	7.7	7.4	7.5	11.6	8.0	10.8	8.8	14.9	10.8	21.3	6.9	9.4	8.2	11.3
kam_Latn	13.4	8.3	8.5	12.0	10.1	9.2	10.9	10.5	10.2	11.3	6.2	10.2	6.9	10.3
ewe_Latn	10.0	11.9	7.4	26.0	6.2	16.1	7.0	27.7	9.6	27.8	4.6	8.0	6.2	19.3
ssw_Latn	14.2	28.7	31.6	37.6	8.9	17.9	15.9	28.5	23.6	32.5	7.4	17.4	6.5	18.1
tum_Latn	9.0	24.1	23.4	33.2	7.7	18.8	11.9	27.8	16.9	31.5	6.6	17.4	8.9	18.5
fon_Latn	4.8	6.1	2.7	11.3	4.6	11.8	4.5	18.1	4.5	15.1	4.5	6.1	4.8	11.6
din_Latn	4.6	2.7	5.5	8.5	8.2	9.0	6.1	13.4	7.6	18.8	6.7	8.4	6.6	9.7
kbp_Latn	4.7	8.6	9.9	15.4	8.0	12.6	8.6	18.7	9.7	19.3	6.5	6.7	6.7	11.5
cjk_Latn	4.7	6.3	5.2	7.2	6.5	9.9	6.8	9.2	7.4	14.9	5.5	7.8	6.7	8.4
nus_Latn	3.4	5.2	4.3	8.7	8.6	13.1	6.0	17.1	6.6	16.4	5.1	7.8	6.8	11.4
taq_Latn	4.1	4.1	5.4	7.4	5.9	6.6	7.4	9.2	7.4	10.5	7.4	7.0	8.2	7.5
taq_Tfng	2.6	3.5	8.7	6.1	9.4	3.5	4.7	3.9	5.8	6.2	3.6	1.2	6.0	1.8
sag_Latn	8.0	11.3	10.4	15.1	10.2	15.5	7.9	23.0	11.3	27.8	8.0	11.5	10.4	20.8
Avg.	10.9	**17.3**	19.5	**25.4**	8.7	**16.8**	12.7	**24.2**	17.3	**26.4**	7.1	**14.1**	8.5	**17.6**
IND														
hin_Deva	50.2	50.3	60.2	58.3	47.0	50.2	54.8	54.1	55.8	54.4	60.2	48.7	59.9	53.1
ben_Beng	34.5	39.8	48.7	47.8	26.8	37.2	41.9	44.0	45.3	43.8	13.9	36.8	27.5	41.4
urd_Arab	31.8	37.9	47.8	48.3	26.8	37.0	40.8	45.5	44.8	45.5	11.3	38.7	19.5	42.4
tel_Telu	33.0	36.7	54.6	51.8	22.8	42.0	43.3	49.1	48.6	50.2	9.0	29.4	13.9	42.3
tam_Taml	33.7	37.0	54.4	52.4	27.0	38.3	45.7	48.9	49.8	48.9	19.0	31.7	32.6	41.6
mar_Deva	31.2	35.1	47.7	46.3	22.9	34.7	38.3	42.0	42.9	42.5	20.0	33.3	24.5	35.9
mai_Deva	24.8	38.1	47.2	49.4	23.4	38.5	35.7	44.6	41.8	46.2	32.3	35.5	37.7	41.8
bho_Deva	26.7	33.7	39.9	41.7	25.5	34.9	34.3	40.8	37.6	40.1	32.5	37.0	36.2	34.8
pbt_Arab	15.6	24.2	32.6	33.2	6.0	18.2	14.6	24.9	15.9	27.4	2.9	24.5	4.6	25.9
guj_Gujr	27.6	36.8	49.6	49.2	22.5	40.2	41.8	46.4	46.1	47.0	11.9	37.4	23.0	41.6
kan_Knda	30.6	34.0	50.8	49.2	21.7	40.4	40.9	47.2	45.6	46.2	9.1	32.7	18.9	39.9
awa_Deva	33.4	35.9	45.6	46.5	32.2	36.4	39.6	43.5	41.2	44.4	39.0	39.3	42.4	41.1
ory_Orya	14.8	19.8	46.3	44.5	10.1	24.9	21.3	39.6	27.7	35.7	4.6	5.8	9.1	14.3
mal_Mlym	27.3	29.9	52.4	49.1	24.1	35.2	42.1	45.5	47.1	45.4	3.3	23.8	5.1	34.4
pan_Guru	26.1	36.3	48.2	48.6	21.0	37.5	41.4	45.7	46.1	45.7	7.2	33.7	11.9	38.7
hne_Deva	30.2	40.5	48.2	50.4	29.8	40.5	40.8	47.9	44.0	46.9	39.8	39.5	42.4	43.7
npi_Deva	34.6	43.1	52.3	51.3	27.1	41.3	44.0	46.9	47.5	46.8	25.4	38.5	30.0	40.2

Language Adaptation Results on FLORES-200 EN-XX														
Model	PaLM2				Gemma2						Aya23			
Size	XXS		S		2B		9B		27B		8B		35B	
Variant	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA
asm_Beng	13.5	26.3	36.9	36.4	13.3	25.0	27.8	31.7	33.3	33.4	8.4	18.2	15.4	26.9
mni_Beng	3.9	6.8	8.1	20.0	7.3	19.4	8.8	31.7	10.7	27.3	7.0	14.4	8.6	18.2
bod_Tibt	12.7	12.1	26.8	24.6	9.0	13.8	14.4	18.8	16.3	17.0	9.0	6.9	8.3	9.6
san_Deva	11.2	20.1	27.6	28.8	11.3	20.8	21.5	26.5	26.2	27.3	14.0	20.0	19.4	23.2
Avg.	26.1	**32.1**	44.1	**44.2**	21.8	**33.6**	34.9	**41.2**	38.8	**41.0**	18.1	**29.8**	23.4	**34.8**

Table 8: Language Adaptation results on FLORES-200 EN-XX with 5-shot in-context learning. PT: Pre-training; LA: Language Adaptation.

Language Adaptation Results on GSM8K-NTL												
Model	PaLM2				Gemma2				Aya23			
Size	S				9B		27B		8B		35B	
Variant	PT	NTL	CALM	LA	PT	LA	PT	LA	PT	LA	PT	LA
asm_Beng	5.2	4.0	9.2	17.2	14.8	29.2	15.6	29.6	0.8	3.6	2.0	10.4
bew_Latn	33.6	33.6	34.8	33.6	33.6	42.8	48.4	50.4	33.6	31.6	45.6	42.4
bho_Deva	23.6	22.8	29.2	26.0	19.6	35.6	29.2	40.0	14.8	12.0	21.6	24.0
doi_Deva	17.2	21.6	22.4	26.8	12.0	33.2	17.2	30.0	9.6	9.6	18.8	24.0
div_Thaa	11.2	13.2	14.8	19.2	6.8	26.8	8.0	21.2	1.6	3.2	1.2	14.0
dzo_Tibt	0.8	0.0	0.4	7.6	1.2	15.2	2.8	8.8	0.4	0.0	0.0	1.2
efi_Latn	14.8	14.0	18.0	22.0	10.8	26.0	16.4	31.2	4.0	5.2	4.0	18.4
gom_Deva	22.4	22.8	25.2	30.0	18.8	36.4	26.8	33.6	8.4	8.0	9.2	20.4
ilo_Latn	14.8	14.0	16.8	18.8	18.0	27.6	23.6	34.8	5.6	9.6	9.2	16.0
kri_Latn	12.4	20.0	18.8	18.8	20.4	25.2	21.6	34.4	8.0	8.8	13.2	11.6
mai_Deva	22.8	21.2	24.8	25.6	19.6	29.2	25.2	30.4	15.6	12.8	29.2	26.4
meo_Latn	28.8	33.2	34.0	28.4	28.4	42.0	45.6	46.4	28.8	29.2	41.2	39.6
mfa_Arab	14.0	20.4	17.6	22.4	6.0	34.8	5.2	38.8	4.0	16.8	4.4	27.2
min_Latn	25.2	24.8	23.6	24.8	13.6	34.8	30.0	39.2	9.2	19.2	17.2	32.0
mni_Beng	2.8	6.0	4.4	9.6	3.2	19.6	2.8	15.6	1.6	2.8	1.2	7.6
mzn_Arab	31.6	27.6	36.4	36.8	33.2	40.8	44.0	38.4	30.4	18.8	41.2	34.0
nso_Latn	8.4	9.6	8.4	13.2	6.4	18.4	14.0	22.0	2.8	6.0	5.2	8.4
ory_Orya	9.6	12.0	12.4	24.0	6.4	30.4	12.0	27.6	1.2	1.6	2.8	10.0
pcm_Latn	34.4	31.6	33.6	30.0	43.2	44.0	47.6	51.2	28.8	26.0	41.2	38.8
tso_Latn	7.2	11.6	10.0	10.0	10.8	15.2	11.2	20.0	3.2	5.2	4.8	8.8
Avg.	17.0	18.2	19.7	**22.2**	16.3	**30.4**	22.4	**32.2**	10.6	**11.5**	15.7	**20.8**

Table 9: Language Adaptation results on GSM8K-NTL with 5-shot in-context learning. PT: Pre-training; LA: Language Adaptation.

Language Adaptation Results on BELEBELE														
Model	PaLM2				Gemma2						Aya23			
Size	XXS		S		2B		9B		27B		8B		35B	
Variant	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA
SEA														
eng_Latn	23.8	28.0	85.9	86.8	61.2	54.9	90.6	91.3	92.3	91.6	82.4	80.0	91.4	89.6
ind_Latn	25.9	26.4	80.7	80.0	49.3	45.6	84.2	84.1	87.8	87.0	73.7	71.7	84.2	80.8
tha_Thai	27.2	26.6	77.9	78.2	44.1	40.0	78.4	76.7	82.8	81.8	44.6	56.9	65.0	68.2
vie_Latn	25.9	25.8	82.1	80.6	46.3	43.9	84.3	84.7	88.1	86.6	73.3	71.3	84.2	81.1
mya_Mymr	25.8	25.9	75.7	71.2	32.6	38.1	68.0	72.6	65.7	75.2	27.9	48.8	36.2	61.7
lao_Laoo	29.1	27.0	70.3	67.7	28.7	36.7	60.6	69.6	59.1	73.4	27.4	47.1	29.4	61.6
khm_Khmr	27.1	26.9	77.4	70.1	30.6	38.6	64.7	72.1	65.2	75.6	25.9	47.0	33.8	64.4
ceb_Latn	25.2	25.8	74.7	75.1	38.2	41.2	76.0	79.4	78.7	81.6	44.1	55.8	55.0	71.2
jav_Latn	27.4	27.2	74.9	73.7	37.9	42.0	73.0	76.1	75.6	77.0	44.8	57.6	56.3	71.7
sun_Latn	26.3	26.2	71.0	70.1	35.9	38.8	68.6	73.2	70.1	76.0	37.0	59.2	47.6	68.1
ilo_Latn	27.9	28.0	64.6	67.3	34.7	39.0	60.9	69.9	63.8	74.6	36.9	45.2	39.9	56.1
war_Latn	26.4	27.2	77.0	77.3	36.9	41.4	70.7	79.3	73.3	81.4	43.9	56.3	50.1	69.3
shn_Mymr	27.4	26.2	28.9	40.6	25.2	31.2	32.0	47.9	33.0	49.8	23.6	32.4	26.4	39.2

Language Adaptation Results on BELEBELE														
Model	PaLM2				Gemma2						Aya23			
Size	XXS		S		2B		9B		27B		8B		35B	
Variant	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA	PT	LA
kac_Latn	24.9	27.1	33.4	43.2	27.6	35.3	33.1	48.1	29.3	48.7	31.0	33.9	28.6	43.4
Avg.	26.5	**26.7**	69.6	**70.1**	37.8	**40.5**	67.5	**73.2**	68.9	**75.7**	44.0	**54.5**	52.0	**66.2**
AFR														
eng_Latn	23.8	26.8	85.9	86.7	61.2	50.9	90.6	90.7	92.3	92.1	82.4	82.1	91.4	88.4
afr_Latn	26.8	27.6	82.8	81.0	46.4	40.9	87.4	86.4	88.6	88.0	61.3	71.9	78.4	81.3
hau_Latn	29.6	26.8	65.6	61.0	35.0	33.3	63.9	65.0	65.8	70.2	29.3	41.2	29.8	49.0
lin_Latn	26.6	25.8	44.7	47.4	30.0	33.8	36.6	51.8	39.8	51.6	30.9	37.6	30.4	40.7
yor_Latn	28.3	27.3	50.8	48.4	28.9	31.0	41.6	48.7	40.6	46.7	26.7	32.8	26.8	36.6
ibo_Latn	25.2	27.9	54.4	49.0	29.8	30.4	48.1	51.2	49.4	54.2	30.0	36.8	30.1	41.8
amh_Ethi	28.2	25.8	75.3	68.1	32.8	35.9	57.8	70.6	57.3	67.4	29.2	40.4	27.3	48.4
som_Latn	27.3	28.7	60.8	56.3	30.8	30.3	49.6	56.3	51.8	58.0	30.8	36.1	32.9	42.4
bam_Latn	27.9	27.4	35.7	38.6	27.4	28.6	34.9	40.4	33.0	39.2	32.4	33.7	30.3	36.2
tso_Latn	24.7	26.8	44.4	51.4	29.7	32.1	43.7	56.1	44.9	57.9	32.0	39.2	33.2	41.9
nya_Latn	28.2	26.8	54.7	53.4	29.9	30.2	47.0	54.1	50.1	54.9	28.6	33.2	31.4	40.2
zul_Latn	27.2	27.1	54.6	56.9	29.9	32.2	55.3	61.8	55.4	61.2	30.4	37.4	31.3	44.3
fuv_Latn	23.8	28.8	28.9	30.4	25.3	27.3	30.0	30.8	29.6	30.8	28.7	27.9	27.4	28.1
kin_Latn	25.9	24.3	54.3	54.0	32.6	33.8	51.0	60.0	55.2	59.9	30.7	37.7	33.9	38.3
sna_Latn	25.8	28.7	62.4	60.9	34.0	35.7	57.7	62.1	59.8	66.8	32.0	38.6	35.6	46.7
xho_Latn	28.7	26.9	59.8	56.3	32.4	31.0	56.0	63.3	54.3	61.9	29.8	37.3	32.8	41.1
tsn_Latn	28.2	27.4	54.6	56.8	31.0	32.7	46.9	58.3	46.8	54.4	28.2	35.6	33.3	42.3
tir_Ethi	27.6	26.3	56.8	52.6	26.9	31.1	37.6	55.8	35.3	51.9	28.4	33.3	27.0	36.2
sot_Latn	29.1	27.9	57.4	57.3	31.6	32.7	51.6	62.2	50.2	58.9	29.9	39.7	31.7	43.4
nso_Latn	27.1	25.4	47.6	53.3	28.8	33.2	45.8	55.9	45.4	51.3	29.0	36.2	32.6	42.1
luo_Latn	25.1	27.4	34.6	33.2	29.4	27.8	34.9	36.8	33.8	35.9	29.4	33.8	32.2	33.2
lug_Latn	27.1	25.7	42.9	41.7	27.0	27.6	38.8	44.1	38.3	42.9	27.4	33.0	31.3	32.7
wol_Latn	27.3	25.7	36.0	34.8	26.3	28.1	34.0	35.7	33.6	33.8	27.3	30.7	27.8	30.2
ssw_Latn	26.1	26.4	48.2	48.2	29.0	29.9	46.2	51.0	43.0	51.3	28.4	34.6	32.4	35.6
kea_Latn	27.3	27.1	65.1	63.4	36.4	31.3	54.8	47.3	58.6	52.4	43.6	40.0	50.2	41.9
Avg.	26.9	26.9	**54.3**	53.7	32.1	**32.5**	49.7	**55.9**	50.1	**55.7**	33.5	**39.2**	36.1	**43.3**
IND														
eng_Latn	23.8	27.6	85.9	86.0	61.2	49.8	90.3	90.8	92.3	92.0	82.4	84.2	91.4	89.9
hin_Deva	26.3	25.0	75.7	73.2	40.8	40.0	77.0	72.3	80.1	76.3	60.6	64.8	73.8	70.9
ben_Beng	27.1	28.1	76.0	73.6	41.1	39.2	76.7	73.6	78.6	77.6	36.1	54.1	48.7	67.7
urd_Arab	27.3	27.8	77.0	73.8	41.4	39.3	76.6	75.2	79.8	78.9	39.2	53.8	49.2	68.3
tel_Telu	27.2	27.7	71.7	69.6	38.1	36.1	71.3	69.4	75.2	73.0	31.6	43.1	34.6	58.8
tam_Taml	28.4	26.1	75.2	72.6	42.3	40.6	75.7	73.6	77.9	76.3	40.3	51.2	61.0	64.7
mar_Deva	25.6	24.8	77.1	73.3	38.1	36.0	76.7	73.1	79.2	75.7	39.1	51.8	52.3	65.4
kan_Knda	28.6	23.6	78.1	73.0	40.3	41.1	74.7	73.9	77.1	78.2	28.9	52.2	41.1	68.7
ory_Orya	26.9	27.0	77.6	72.3	30.9	36.6	60.9	70.4	63.1	74.0	30.6	30.0	40.8	49.6
mal_Mlym	26.8	29.2	79.7	75.9	42.6	39.6	76.1	75.2	80.1	77.1	35.3	44.2	51.6	60.9
pan_Guru	28.6	27.9	76.8	72.1	36.8	35.2	77.0	74.3	77.4	74.1	29.6	49.2	33.0	65.2
snd_Arab	26.6	26.2	68.4	68.4	32.4	34.9	59.0	66.4	64.9	69.4	32.3	44.1	39.6	59.7
sin_Latn	26.6	24.1	35.1	34.9	28.2	26.8	38.3	36.8	38.9	38.3	30.9	28.9	34.3	32.8
sin_Sinh	26.2	28.1	76.9	75.7	35.0	39.9	70.9	74.3	71.4	73.2	30.2	50.3	38.0	61.4
asm_Beng	28.7	24.4	75.9	72.3	36.0	32.8	67.9	70.8	73.3	71.1	34.0	41.3	40.9	58.8
hin_Latn	26.7	29.1	68.3	63.9	37.9	38.0	69.4	69.9	74.2	74.3	45.3	48.9	57.6	60.8
bod_Tibt	27.3	27.8	50.2	38.7	27.3	28.4	39.1	43.0	35.0	44.8	25.7	26.3	30.1	31.8
ben_Latn	28.6	26.4	52.4	51.9	29.8	30.6	49.8	54.1	55.6	57.8	32.2	34.7	38.9	40.7
urd_Latn	26.8	27.7	56.3	54.8	34.2	31.8	56.8	56.8	60.0	61.7	36.6	37.8	43.6	46.6
Avg.	**27.0**	26.8	**70.2**	67.2	**37.6**	36.7	67.6	**68.1**	70.2	**70.7**	37.9	**46.9**	47.4	**59.1**

Table 10: Language Adaptation results on BELEBELE with 5-shot in-context learning. PT: Pre-training; LA: Language Adaptation.

EMBSWAP Results on BELEBELE

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR

EMBSWAP Results on BELEBELE

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
SEA																					
eng_Latn	80.9	76.0	78.9	95.7	95.9	95.8	87.7	85.9	87.4	93.9	93.9	93.6	95.4	95.0	94.0	88.9	83.4	86.4	94.2	87.0	93.2
ind_Latn	74.0	72.3	73.4	91.8	91.9	92.0	78.7	76.1	75.8	88.8	87.6	88.2	90.4	89.7	88.3	82.0	74.3	74.3	89.6	72.7	86.3
tha_Thai	66.8	65.2	65.8	88.3	88.3	88.1	70.0	67.7	68.3	83.6	83.3	82.7	85.7	85.1	85.3	57.2	59.3	63.7	73.9	60.7	79.4
vie_Latn	74.1	72.0	72.0	92.6	92.7	92.6	75.6	74.0	74.9	89.7	89.2	88.4	90.0	89.7	88.8	81.9	75.9	74.6	90.2	75.6	87.6
mya_Mymr	52.0	56.1	56.7	86.8	85.1	85.7	36.3	62.1	60.8	69.7	76.1	77.9	65.2	78.2	78.2	34.0	51.1	56.2	38.1	49.2	69.0
lao_Laoo	44.9	55.9	57.3	81.2	82.3	83.1	43.3	60.9	61.4	61.4	77.3	75.9	59.7	77.1	77.3	31.8	50.6	51.9	38.8	49.2	66.4
khm_Khmr	51.9	52.6	55.8	86.3	84.6	86.7	45.1	65.0	62.6	69.2	78.6	78.4	68.6	81.1	79.4	26.8	46.9	51.4	43.6	51.7	71.2
ceb_Latn	52.2	61.8	62.6	88.8	90.6	90.0	52.9	64.2	65.8	79.0	82.4	84.4	82.2	84.1	85.9	45.1	59.8	60.9	57.9	59.1	77.8
jav_Latn	61.0	61.8	62.2	86.7	86.7	87.3	48.1	59.8	64.7	78.7	80.8	80.9	79.0	81.7	81.4	48.1	62.9	65.3	62.0	58.7	77.8
sun_Latn	55.2	56.4	58.3	86.3	84.3	85.4	41.1	56.8	61.7	70.9	79.6	78.8	71.2	79.6	79.9	41.0	57.7	61.1	51.2	53.7	75.4
ilo_Latn	41.9	50.7	53.4	79.2	84.0	84.3	43.1	53.0	56.8	63.8	75.9	76.9	68.6	74.6	78.2	34.4	45.7	45.3	42.1	47.6	65.0
war_Latn	50.6	60.8	63.4	89.3	90.6	90.8	49.3	65.3	64.8	75.1	84.7	83.7	79.1	84.1	87.9	42.4	58.8	64.2	53.9	59.0	77.2
shn_Mymr	26.1	33.9	36.8	28.8	52.1	50.8	28.0	39.8	40.0	33.4	53.3	53.8	35.7	54.4	54.1	26.8	33.7	34.6	26.4	30.1	46.2
kac_Latn	31.7	36.8	39.4	40.8	55.4	58.3	28.3	36.0	38.9	31.6	50.1	54.1	35.3	48.6	53.1	28.1	32.0	37.8	34.1	33.0	43.9
Avg.	54.5	58.0	**59.7**	80.2	83.2	**83.6**	52.0	61.9	**63.1**	70.6	78.1	**78.4**	71.9	78.8	**79.4**	47.8	56.6	**59.1**	56.9	56.2	**72.6**
AFR																					
eng_Latn	80.9	79.1	80.7	95.7	95.4	95.7	87.7	85.2	87.7	93.9	94.3	93.6	95.4	94.9	93.1	88.9	84.2	86.1	94.2	92.2	91.8
afr_Latn	70.4	69.8	71.6	93.3	93.8	93.1	72.4	74.9	75.3	90.1	90.6	89.4	92.1	91.1	89.4	69.6	74.9	74.4	84.0	81.9	87.3
hau_Latn	40.4	38.4	43.8	80.2	75.7	77.9	35.0	42.9	45.4	67.4	68.1	71.2	68.1	68.1	74.1	28.0	40.7	40.2	33.8	43.1	58.3
lin_Latn	34.2	36.7	40.7	57.1	62.9	65.7	30.2	34.4	42.3	36.6	55.0	57.3	42.0	47.6	58.8	28.1	34.0	37.2	35.7	35.7	47.0
yor_Latn	32.7	29.7	32.7	62.7	58.9	59.9	29.7	31.9	34.4	40.9	46.4	51.8	41.6	45.3	51.9	27.3	30.4	33.4	34.4	31.2	44.6
ibo_Latn	34.7	35.6	37.2	68.6	64.7	65.8	31.4	35.1	44.2	47.6	54.8	55.7	48.8	51.7	57.2	29.7	35.6	37.1	35.0	37.0	49.1
amh_Ethi	48.7	43.4	50.4	86.9	82.2	83.0	35.7	56.6	55.9	60.3	74.2	75.6	60.3	72.3	73.7	25.2	41.9	40.8	32.8	43.6	57.8
som_Latn	39.9	37.2	39.9	73.3	70.9	72.7	33.9	38.3	41.8	50.0	59.3	60.9	52.4	58.2	63.0	29.8	34.8	34.3	36.1	37.6	49.6
bam_Latn	34.7	34.3	37.4	43.7	45.7	46.2	30.2	30.6	35.6	34.2	40.8	41.1	36.8	40.7	43.7	32.6	32.6	30.8	36.2	32.6	38.3
tso_Latn	36.3	39.6	44.3	57.9	69.7	70.8	34.7	41.1	45.8	43.7	54.7	62.6	48.7	58.3	62.3	31.2	36.4	37.0	40.1	37.0	52.6
nya_Latn	38.4	38.8	40.7	70.8	67.1	70.1	30.1	34.3	39.2	46.7	55.3	58.1	50.6	53.4	59.6	27.1	34.7	34.3	32.1	35.6	46.3
zul_Latn	41.2	41.8	46.7	76.2	72.0	75.0	33.7	41.3	42.9	57.4	66.0	69.6	58.2	62.7	68.9	33.0	36.9	37.6	38.8	40.8	52.3
fuv_Latn	30.7	28.0	27.9	31.0	31.9	32.1	29.0	27.3	28.0	28.2	29.8	31.0	29.7	31.2	29.9	27.4	25.7	26.7	31.0	28.1	29.3
kin_Latn	37.1	41.3	44.3	68.9	71.8	73.8	35.3	37.2	41.0	51.6	64.2	67.2	58.6	59.6	68.4	32.0	35.3	34.9	37.3	36.0	46.8
sna_Latn	42.9	44.1	50.3	77.6	75.7	77.9	32.4	42.6	46.4	58.7	67.0	69.4	62.3	65.0	68.9	30.3	39.0	42.2	38.2	42.7	56.0
xho_Latn	42.9	42.4	47.0	76.6	74.1	76.2	35.4	40.3	44.1	58.4	64.7	68.3	60.2	65.3	70.2	30.3	36.9	38.2	38.2	39.9	52.0
tsn_Latn	35.4	38.8	44.6	68.3	69.9	70.8	31.8	37.1	43.1	48.8	57.4	63.4	53.3	55.8	61.4	32.3	35.4	36.6	36.6	35.7	49.2
tir_Ethi	32.2	38.1	40.9	70.2	71.9	71.6	29.3	41.6	44.3	33.2	59.1	60.4	37.6	51.9	57.7	27.4	33.3	32.9	31.6	29.7	45.1
sot_Latn	38.1	40.4	42.2	76.9	76.0	77.9	29.4	38.7	42.9	52.7	64.6	66.9	53.4	60.6	66.8	30.1	37.4	37.1	34.4	39.1	53.3
nso_Latn	35.4	38.3	41.7	65.3	69.8	70.7	31.3	35.7	42.4	44.0	57.6	63.1	51.4	58.1	63.3	30.0	34.2	35.7	35.8	35.8	47.2
luo_Latn	33.6	31.1	34.4	40.8	42.2	42.7	30.0	29.7	32.7	34.1	34.6	36.0	36.2	37.0	36.7	27.9	32.1	32.0	34.0	33.0	36.9
lug_Latn	31.6	32.0	34.8	51.8	54.9	57.9	29.9	29.7	33.0	38.8	45.0	47.0	40.0	43.9	47.4	29.8	32.2	32.1	35.2	28.9	39.3
wol_Latn	31.4	32.3	31.8	44.0	38.7	39.2	29.2	26.4	29.4	30.2	30.4	35.3	31.9	34.9	34.2	28.2	28.9	29.6	32.4	30.3	32.0
ssw_Latn	33.3	36.3	41.0	63.2	61.3	66.2	29.2	33.9	38.0	42.7	51.7	57.1	46.8	51.9	56.1	28.7	33.1	32.7	32.9	31.6	41.7
kea_Latn	41.9	37.7	38.8	84.9	79.3	80.6	42.4	37.1	37.4	56.3	51.3	52.3	59.4	47.0	52.8	46.6	40.1	40.2	57.4	44.9	48.4
Avg.	40.0	40.2	**43.4**	67.4	67.1	**68.5**	36.0	40.2	**43.7**	49.9	57.5	**60.2**	52.6	56.3	**60.4**	34.1	38.4	**39.0**	40.3	40.1	**50.1**
IND																					
eng_Latn	80.9	81.8	82.7	95.7	94.9	95.3	87.7	86.1	86.9	93.9	93.3	93.0	95.4	94.9	94.2	88.9	85.7	88.1	94.2	92.6	93.1
hin_Deva	62.3	59.9	61.3	86.7	86.4	86.9	64.0	64.1	63.8	80.6	78.1	79.0	81.4	80.1	81.1	69.1	65.9	66.9	79.1	66.3	76.2
ben_Beng	59.1	59.6	61.8	89.1	86.9	87.8	59.7	65.0	62.4	79.0	77.3	78.6	81.4	81.4	82.3	42.8	58.0	59.0	58.8	60.3	74.8
urd_Arab	57.8	54.9	59.4	89.4	87.4	88.7	60.7	62.7	62.7	81.1	78.2	79.0	82.1	80.0	79.9	47.2	61.3	65.0	65.3	65.6	76.7
tel_Telu	48.2	48.4	51.0	83.3	81.1	82.9	54.4	57.2	54.8	75.3	75.0	76.1	77.0	76.0	76.8	36.6	45.9	49.3	47.8	54.3	66.3
tam_Taml	57.0	59.0	59.1	85.3	84.2	85.8	58.7	59.7	60.6	76.6	77.4	79.8	79.8	78.3	79.8	49.4	53.0	56.4	68.1	55.0	71.9
mar_Deva	60.8	61.7	61.7	88.8	86.4	88.0	58.0	59.2	60.8	77.9	78.8	80.7	81.1	79.6	80.6	46.8	56.3	59.0	64.2	60.8	71.9
kan_Knda	52.7	55.1	59.2	89.6	88.0	88.6	53.1	61.3	61.9	77.1	79.2	80.1	82.0	79.9	80.8	36.9	50.3	55.9	45.7	60.7	74.3
ory_Orya	40.9	44.1	52.1	86.3	85.6	88.9	38.9	51.9	56.8	64.1	76.4	76.8	68.7	77.7	80.4	36.1	34.6	34.1	51.3	37.4	60.7
mal_Mlym	55.8	56.6	58.1	87.6	87.3	89.3	57.8	60.3	60.6	80.0	81.9	81.9	81.4	78.4	81.1	44.9	44.9	51.0	59.4	52.7	70.7
pan_Guru	46.6	50.6	56.8	87.4	85.3	86.7	55.1	57.3	58.8	75.0	76.6	79.7	81.6	78.8	78.8	35.4	49.1	53.7	46.1	56.3	71.1
snd_Arab	46.1	47.2	52.1	84.1	83.6	86.1	38.6	52.7	55.8	61.0	68.9	72.9	67.7	73.1	73.3	35.0	48.7	52.2	44.4	53.3	67.1
sin_Latn	32.0	32.2	33.7	45.9	42.0	43.2	31.4	30.1	32.7	37.0	36.2	36.0	41.2	39.8	38.2	29.6	26.0	27.1	37.7	32.1	35.0
sin_Sinh	53.9	55.7	60.3	89.8	87.3	89.4	45.7	57.8	59.7	71.0	77.1	79.9	75.7	75.9	79.3	34.2	47.6	52.2	47.3	53.2	69.7
asm_Beng	47.2	49.2	53.9	88.2	86.1	87.8	45.6	54.4	56.4	67.1	72.3	75.2	73.9	73.8	74.9	35.4	41.9	46.1	41.8	46.2	66.3
hin_Latn	41.6	47.2	50.3	81.3	80.2	81.4	47.7	50.3	52.2	69.9	70.9	72.9	75.8	75.6	76.4	48.7	49.8	49.8	67.1	57.6	65.8

EMBSWAP Results on BELEBELE

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
bod_Tibt	27.9	29.4	31.9	61.0	48.1	54.2	27.1	32.4	35.6	40.6	45.4	47.4	37.6	41.9	47.7	27.4	25.8	27.3	30.1	24.6	31.4
ben_Latn	35.3	38.3	38.6	68.7	66.9	70.7	33.6	34.9	38.7	44.6	53.1	58.6	56.9	60.7	63.1	30.6	31.7	34.9	38.9	37.2	44.9
urd_Latn	33.2	35.6	39.0	72.0	69.9	73.3	35.6	38.8	39.3	56.9	56.1	60.7	64.2	64.0	63.4	35.9	36.0	37.9	49.1	41.4	50.9
Avg.	49.4	50.9	**53.8**	**82.1**	79.9	81.8	50.2	54.5	**55.8**	68.9	71.2	**73.1**	72.9	73.1	**74.3**	42.7	48.0	**50.8**	54.6	53.0	**65.2**

Table 11: EMBSWAP results on BELEBELE with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; ES: EMBSWAP; +LR: EMBSWAP with LoRA Adaptation.

EMBSWAP Results on SIB-200

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
SEA																					
eng_Latn	85.3	83.8	82.4	76.0	76.5	77.9	82.4	82.4	79.9	83.3	81.4	80.4	83.3	81.9	82.4	82.8	80.8	79.8	82.8	80.8	79.8
ace_Arab	27.9	51.5	47.1	53.4	70.1	67.6	37.7	45.6	50.5	53.4	64.2	65.2	46.6	62.7	67.6	37.4	46.5	38.4	39.4	51.5	65.7
ace_Latn	63.2	78.4	71.6	71.1	80.9	77.0	62.3	68.6	74.0	73.5	76.5	79.9	71.1	73.0	81.9	63.6	74.7	61.6	68.7	73.7	76.8
ban_Latn	75.0	80.9	76.5	73.0	77.0	77.9	68.6	74.0	76.0	81.4	78.4	78.9	81.9	77.5	81.4	70.7	75.8	66.7	71.7	70.7	75.8
bjn_Arab	17.2	43.6	47.1	55.4	63.7	62.3	31.4	42.2	48.0	47.1	59.8	62.3	44.1	59.3	64.7	36.4	42.4	39.4	40.4	48.5	58.6
bjn_Latn	73.0	78.9	75.0	76.5	79.4	81.4	67.6	71.6	71.1	80.4	78.9	79.9	77.9	78.9	79.9	72.7	73.7	60.6	73.7	70.7	71.7
bug_Latn	57.8	67.2	71.1	63.7	74.0	72.5	57.4	62.3	64.2	62.7	69.6	69.1	63.2	68.6	70.1	55.6	66.7	54.5	65.7	69.7	75.8
ceb_Latn	79.4	84.3	79.4	79.9	77.9	77.0	75.5	79.4	77.9	85.3	84.3	79.9	83.8	84.3	85.3	73.7	77.8	75.8	77.8	76.8	79.8
ilo_Latn	68.6	81.9	80.4	72.5	80.9	81.9	73.5	79.4	81.4	78.9	83.8	82.4	79.9	85.3	82.4	68.7	78.8	73.7	66.7	76.8	81.8
ind_Latn	85.8	83.8	80.4	77.5	80.4	80.9	84.8	83.3	80.9	83.8	83.3	83.3	82.8	80.9	83.3	84.8	78.8	74.7	80.8	77.8	78.8
jav_Latn	75.5	81.4	77.9	75.5	76.5	76.5	73.0	76.0	75.5	82.4	81.4	79.9	79.9	81.4	81.4	76.8	76.8	67.7	79.8	73.7	77.8
kac_Latn	41.2	64.7	66.7	42.2	70.6	70.1	35.8	63.7	67.6	50.0	71.6	76.5	49.0	70.6	74.5	48.5	62.6	60.6	45.5	60.6	72.7
khm_Khmr	82.4	85.3	79.9	80.9	77.9	77.9	71.6	79.9	76.0	83.8	81.4	83.3	77.0	82.8	83.8	55.6	73.7	70.7	62.6	74.7	74.7
lao_Laoo	79.9	86.8	80.4	78.4	82.8	83.8	69.6	81.4	78.4	76.5	81.4	81.4	73.0	82.8	82.8	55.6	75.8	66.7	58.6	74.7	77.8
min_Arab	17.6	40.7	45.1	57.8	68.1	68.6	32.4	44.1	51.0	44.6	60.3	59.3	40.2	59.8	62.3	27.3	43.4	30.3	40.4	41.4	54.5
min_Latn	76.0	81.9	77.0	76.5	78.9	81.9	70.1	75.5	77.0	81.9	85.3	83.8	77.5	83.3	82.4	66.7	78.8	71.7	73.7	77.8	79.8
mya_Mymr	78.9	83.3	78.4	77.9	79.9	81.9	65.2	79.9	77.5	80.4	84.3	83.3	78.9	81.9	80.9	40.4	72.7	66.7	56.6	63.6	74.7
pag_Latn	67.6	83.3	78.9	76.0	80.4	82.4	67.6	76.0	77.5	77.0	80.4	81.9	72.5	85.3	83.8	65.7	73.7	69.7	68.7	71.7	75.8
shn_Mymr	39.7	76.5	79.4	42.2	78.9	80.4	42.6	76.0	80.9	56.9	79.9	81.4	56.4	78.4	78.9	53.5	70.7	66.7	51.5	71.7	74.7
sun_Latn	77.0	82.8	79.4	76.5	78.9	81.4	74.0	78.4	79.4	80.4	81.4	80.9	80.4	80.9	82.4	72.7	77.8	72.7	71.7	73.7	82.8
tgl_Latn	83.3	81.9	77.9	78.4	77.0	80.4	77.9	77.0	76.5	85.3	83.3	81.9	83.3	80.4	82.8	79.8	82.8	77.8	81.8	79.8	77.8
tha_Thai	84.8	85.8	79.4	76.5	78.4	79.9	82.8	81.4	77.0	80.4	83.3	82.4	81.9	82.4	81.9	72.7	73.7	66.7	75.8	75.8	75.8
vie_Latn	82.4	85.8	78.9	77.0	77.9	79.9	81.9	81.4	80.9	81.9	82.8	82.4	82.4	82.4	81.9	85.9	83.8	72.7	80.8	76.8	79.8
war_Latn	75.0	80.9	75.0	72.5	77.9	77.9	77.5	81.4	78.4	82.8	81.9	79.9	81.9	83.3	83.3	79.8	83.8	82.8	78.8	77.8	78.8
zsm_Latn	85.3	83.3	77.0	77.0	79.4	81.9	83.8	80.9	80.9	85.3	85.8	82.4	86.8	84.3	85.3	80.8	82.8	70.7	79.8	79.8	78.8
Avg.	67.2	**76.7**	73.7	70.6	77.0	**77.6**	65.9	72.9	**73.5**	74.4	**78.6**	78.5	72.6	78.1	**79.5**	64.3	72.4	**65.6**	66.9	70.8	**75.2**
AFR																					
eng_Latn	85.3	83.3	79.9	76.0	77.5	80.9	82.4	83.3	66.7	83.3	82.8	80.9	83.3	82.8	83.8	82.8	83.8	73.7	82.8	76.8	79.8
afr_Latn	81.9	82.4	79.9	76.0	79.4	80.4	80.4	81.4	65.2	82.8	82.4	79.4	82.4	82.4	84.3	75.8	80.8	71.7	78.8	73.7	76.8
aka_Latn	50.0	60.3	65.2	64.7	72.1	72.5	43.6	56.9	46.6	63.2	74.5	74.5	62.3	65.7	74.0	55.6	65.7	59.6	56.6	54.5	67.7
amh_Ethi	75.5	77.0	76.0	80.4	79.4	82.8	55.9	76.5	57.8	75.5	84.3	80.9	78.4	80.9	80.9	29.3	69.7	63.6	29.3	63.6	78.8
bam_Latn	38.7	51.5	52.9	51.0	58.3	58.3	35.3	49.0	40.7	53.9	61.3	63.7	50.5	57.8	61.3	43.4	48.5	48.5	51.5	40.4	51.5
bem_Latn	49.5	56.9	58.3	69.1	68.1	69.1	44.6	49.0	41.7	57.8	66.2	69.1	59.3	61.8	67.6	54.5	57.6	52.5	52.5	51.5	58.6
cjk_Latn	45.6	49.5	50.0	41.2	44.6	49.5	39.2	39.7	38.2	50.5	51.5	50.5	47.5	48.0	47.1	45.5	43.4	39.4	46.5	40.4	50.5
dyu_Latn	41.7	43.6	51.5	48.5	48.0	54.4	44.1	47.1	41.7	54.9	51.0	53.9	51.5	52.5	50.5	38.4	43.4	42.4	45.5	36.4	52.5
ewe_Latn	37.7	55.9	60.3	43.6	64.7	66.2	34.3	47.5	42.2	48.5	62.3	72.5	45.6	60.8	68.1	48.5	49.5	50.5	48.5	41.4	67.7
fon_Latn	35.3	45.1	51.5	38.2	53.9	50.0	36.8	43.1	39.2	51.0	60.3	61.3	46.1	55.9	55.9	48.5	53.5	53.5	50.5	39.4	58.6
fuv_Latn	47.1	52.9	56.9	50.5	57.4	58.8	45.1	50.0	40.7	52.5	60.3	59.3	47.5	52.9	52.5	53.5	57.6	57.6	54.5	52.5	67.7
hau_Latn	73.5	71.1	72.5	78.4	80.9	80.9	47.5	68.6	57.4	77.0	78.4	76.5	77.9	79.9	83.3	48.5	77.8	70.7	47.5	67.7	83.8
ibo_Latn	73.0	73.5	72.1	78.4	77.5	80.9	63.2	65.7	51.5	76.5	78.9	75.5	70.6	78.4	79.9	49.5	66.7	61.6	53.5	59.6	68.7
kam_Latn	43.1	50.5	50.0	50.0	54.4	58.8	41.7	42.6	40.7	51.0	52.9	54.9	53.4	51.0	51.5	46.5	49.5	45.5	52.5	50.5	49.5
kbp_Latn	41.2	58.8	58.3	53.4	65.7	67.2	36.8	58.3	49.5	56.9	69.6	73.0	52.0	61.3	65.2	43.4	49.5	48.5	46.5	41.4	57.6
kea_Latn	71.6	69.6	66.7	74.5	77.9	83.3	66.7	63.7	51.5	80.9	76.0	76.0	75.5	71.1	76.0	75.8	69.7	62.6	78.8	64.6	71.7
kik_Latn	51.5	52.5	52.5	59.8	61.3	63.7	48.0	53.9	41.7	55.9	59.8	55.4	53.9	52.0	54.4	46.5	44.4	45.5	48.5	41.4	48.5
kin_Latn	44.1	72.1	77.5	76.0	80.9	81.9	45.6	66.7	57.4	75.5	81.9	82.4	74.5	78.4	83.3	45.5	61.6	58.6	50.5	59.6	71.7
kmb_Latn	41.7	51.5	52.9	42.6	52.0	54.4	43.6	44.6	39.2	45.1	54.9	52.0	46.1	49.0	49.5	48.5	49.5	41.4	46.5	47.5	54.5
kon_Latn	52.0	76.0	74.5	57.8	75.0	76.5	53.4	67.2	56.4	57.4	79.4	81.9	59.8	77.0	77.9	44.4	66.7	60.6	54.5	63.6	75.8
lin_Latn	53.4	71.6	71.1	64.2	74.5	76.0	52.5	67.2	52.9	64.7	77.0	79.9	61.3	74.0	74.0	53.5	68.7	59.6	54.5	59.6	71.7
lua_Latn	43.6	61.8	65.7	52.0	67.6	68.1	44.1	49.5	46.1	55.4	72.1	70.6	51.0	62.7	68.6	44.4	56.6	52.5	53.5	57.6	72.7
lug_Latn	45.1	60.3	60.8	61.3	69.1	72.1	41.7	49.5	46.6	60.8	67.6	70.1	58.3	62.3	68.6	47.5	57.6	51.5	48.5	48.5	63.6
luo_Latn	45.6	55.9	62.3	52.0	61.3	64.7	47.1	52.0	40.7	49.5	57.8	63.7	52.5	57.4	63.2	47.5	48.5	45.5	46.5	45.5	59.6
mos_Latn	41.2	41.7	46.1	41.7	42.2	45.1	45.1	46.1	36.8	44.6	47.5	49.5	51.0	51.0	52.0	40.4	49.5	42.4	44.4	40.4	41.4

EMBSWAP Results on SIB-200																					
Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
nso_Latn	41.7	71.1	75.0	71.6	74.0	79.9	46.6	65.2	47.5	62.3	76.5	76.0	64.7	73.5	79.4	45.5	69.7	55.6	43.4	60.6	75.8
nus_Latn	24.0	41.2	44.6	34.8	54.9	56.9	34.3	41.7	36.8	42.2	60.3	62.7	43.6	50.5	54.4	44.4	42.4	42.4	52.5	36.4	55.6
nya_Latn	70.6	76.0	71.1	76.0	79.4	78.9	49.0	67.2	51.5	71.1	77.9	75.5	71.1	78.9	78.9	43.4	65.7	61.6	62.6	73.7	73.7
run_Latn	41.2	73.5	72.5	71.6	74.5	75.0	45.1	59.8	54.4	66.2	74.5	78.4	66.7	75.0	76.5	48.5	57.6	58.6	56.6	58.6	68.7
sag_Latn	44.6	54.4	60.3	46.6	62.3	62.7	41.7	53.4	40.7	54.4	66.2	68.6	50.0	61.8	66.7	50.5	56.6	58.6	53.5	48.5	60.6
sna_Latn	62.7	74.0	71.6	72.5	78.4	76.0	46.1	67.6	56.4	69.6	76.0	76.5	69.6	81.9	82.8	45.5	58.6	60.6	53.5	67.7	78.8
som_Latn	61.3	73.0	72.5	75.5	77.5	78.4	54.4	66.2	56.4	73.5	79.9	75.5	72.1	79.4	77.5	48.5	71.7	68.7	62.6	56.6	68.7
sot_Latn	57.4	72.5	74.5	73.0	74.0	76.5	49.5	69.1	52.9	68.6	77.5	77.0	67.2	78.9	79.9	51.5	60.6	61.6	44.4	63.6	69.7
ssw_Latn	52.5	74.0	73.5	70.1	76.0	77.0	48.0	67.2	52.0	68.6	77.5	76.5	63.7	72.1	75.5	42.4	64.6	59.6	48.5	56.6	69.7
taq_Latn	46.6	49.0	47.5	49.5	51.5	58.3	44.1	48.5	43.6	51.5	55.9	51.5	52.0	52.0	52.5	48.5	56.6	50.5	58.6	50.5	56.6
taq_Tfng	11.3	27.0	27.5	21.6	23.0	25.0	26.5	27.0	25.0	22.1	27.5	26.5	32.4	27.0	25.5	30.3	26.3	27.3	30.3	15.2	27.3
tir_Ethi	53.9	63.2	68.1	73.5	75.0	76.5	44.6	69.1	51.5	62.7	80.9	82.4	59.3	76.5	75.5	30.3	62.6	54.5	31.3	47.5	66.7
tsn_Latn	44.6	66.2	69.6	66.7	74.5	74.5	45.6	64.2	51.5	63.2	77.9	79.4	63.2	74.0	79.4	51.5	65.7	61.6	57.6	64.6	74.7
tso_Latn	39.7	61.8	70.1	64.2	76.0	77.0	46.1	59.8	47.5	56.9	74.0	75.5	58.3	72.5	77.9	45.5	60.6	60.6	53.5	60.6	72.7
tum_Latn	60.3	77.9	74.0	68.1	74.5	75.0	44.1	62.7	53.9	68.6	78.4	79.9	63.2	74.0	77.5	52.5	58.6	62.6	57.6	61.6	66.7
umb_Latn	38.7	48.0	47.1	40.2	45.6	46.1	37.3	38.7	37.3	50.5	49.0	50.0	50.0	47.1	45.1	44.4	45.5	45.5	51.5	40.4	48.5
wol_Latn	52.9	52.0	54.4	61.3	66.2	66.7	49.0	57.4	51.0	59.3	59.3	63.2	59.3	63.7	59.8	51.5	57.6	60.6	55.6	53.5	62.6
xho_Latn	69.6	78.4	77.9	76.0	77.0	80.4	54.9	71.1	54.4	73.5	81.9	79.4	77.0	79.4	83.3	47.5	67.7	58.6	57.6	62.6	74.7
yor_Latn	64.2	63.7	67.2	70.6	72.5	75.5	52.0	56.4	44.6	66.7	73.5	72.1	62.7	68.1	69.1	47.5	56.6	57.6	60.6	54.5	66.7
zul_Latn	63.2	78.9	73.0	75.5	78.9	79.9	52.5	73.0	53.9	70.6	81.9	78.4	70.6	82.4	81.4	53.5	66.7	61.6	63.6	68.7	72.7
Avg.	51.3	62.2	**63.5**	60.9	66.9	**68.7**	47.3	**57.9**	47.8	61.0	68.8	**69.2**	60.2	66.1	**68.3**	48.5	58.7	**55.3**	52.9	53.8	**64.7**
IND																					
eng_Latn	85.3	84.8	84.3	76.0	75.0	78.4	82.4	81.4	81.4	83.3	81.4	81.4	83.3	82.4	84.3	82.8	85.9	77.8	82.8	79.8	76.8
asm_Beng	73.0	80.9	79.4	77.0	78.4	79.9	69.6	73.0	68.6	82.8	83.8	82.4	79.4	82.4	79.4	56.6	65.7	62.6	61.6	68.7	72.7
awa_Deva	80.9	84.8	82.4	79.4	78.9	79.9	77.5	74.5	73.5	80.9	80.9	80.4	78.9	79.9	82.4	81.8	78.8	72.7	83.8	79.8	80.8
ben_Beng	81.9	84.8	79.9	80.9	78.9	81.4	77.9	80.4	76.0	83.8	84.3	82.4	84.3	82.4	84.3	61.6	80.8	72.7	70.7	76.8	77.8
bho_Deva	76.5	81.4	80.4	75.5	78.9	77.9	74.0	73.5	74.0	79.9	83.8	82.4	77.0	80.9	83.8	74.7	82.8	72.7	76.8	77.8	77.8
dzo_Tibt	40.2	71.6	68.1	63.7	75.5	77.0	23.5	60.8	56.9	60.8	74.5	77.5	49.0	60.8	67.2	26.3	33.3	32.3	28.3	38.4	59.6
guj_Gujr	77.5	82.4	82.4	78.4	79.9	81.9	74.5	79.4	77.5	81.4	85.8	84.8	81.9	81.4	85.3	55.6	80.8	69.7	67.7	74.7	77.8
hin_Deva	81.9	83.3	83.8	82.8	74.5	80.9	83.3	80.9	77.9	86.8	84.8	83.3	82.4	84.3	82.8	81.8	82.8	75.8	81.8	75.8	77.8
hne_Deva	78.9	84.3	83.8	77.9	75.0	80.4	74.5	73.0	77.5	82.4	81.4	79.9	78.4	81.4	80.4	77.8	83.8	72.7	78.8	79.8	79.8
kan_Knda	77.5	83.8	81.9	79.9	77.5	78.9	74.0	77.5	73.5	84.3	81.4	83.3	83.3	84.8	86.3	52.5	70.7	69.7	68.7	73.7	73.7
kas_Arab	64.7	69.6	70.1	77.0	79.4	82.4	63.7	63.7	66.2	74.5	75.5	74.0	70.6	75.0	77.5	52.5	68.7	64.6	63.6	67.7	70.7
kas_Deva	56.9	61.3	64.2	74.5	71.6	79.4	51.0	53.9	57.8	64.7	69.1	73.0	63.2	69.1	66.7	62.6	59.6	56.6	63.6	56.6	62.6
lus_Latn	58.3	78.4	75.5	59.8	72.1	77.9	61.8	71.1	70.6	77.9	80.9	80.9	78.4	82.4	81.9	54.5	65.7	67.7	54.5	66.7	76.8
mag_Deva	81.9	81.4	79.4	78.9	74.5	78.4	77.5	74.5	76.0	82.8	79.9	80.9	79.9	84.8	81.9	77.8	81.8	68.7	81.8	73.7	75.8
mai_Deva	83.8	83.8	82.8	77.0	77.5	81.9	76.0	74.5	76.0	80.4	83.8	82.8	78.9	82.4	81.4	76.8	81.8	71.7	79.8	78.8	79.8
mal_Mlym	77.0	79.4	83.8	80.9	77.9	80.9	73.5	72.1	76.5	83.8	81.4	80.9	80.4	83.3	84.8	66.7	70.7	66.7	73.7	71.7	76.8
mar_Deva	82.4	80.4	81.9	78.4	79.4	79.9	75.5	75.5	77.0	83.3	82.4	83.8	78.4	85.3	87.3	73.7	78.8	70.7	75.8	74.7	79.8
mni_Beng	34.8	55.9	64.2	57.4	72.5	77.0	44.6	52.5	54.9	56.4	76.0	75.5	50.5	69.1	77.9	27.3	59.6	51.5	40.4	52.5	72.7
pan_Guru	75.5	80.4	78.9	76.5	75.5	80.9	72.1	78.4	74.5	85.3	84.3	84.3	83.3	81.9	83.8	44.4	67.7	61.6	63.6	70.7	76.8
san_Deva	71.6	76.0	75.5	76.0	75.5	77.9	66.2	65.2	67.2	78.4	77.9	79.4	69.6	73.5	77.5	68.7	68.7	60.6	75.8	67.7	71.7
sat_Olck	13.7	23.0	24.5	37.7	26.5	28.4	34.8	24.5	29.9	69.6	43.6	48.0	67.6	36.3	41.2	26.3	26.3	27.3	26.3	14.1	35.4
sin_Sinh	78.9	81.9	82.4	77.0	74.0	75.5	61.3	73.5	76.0	83.8	82.4	83.8	77.5	81.4	81.4	46.5	74.7	70.7	65.7	67.7	78.8
snd_Arab	72.1	81.4	80.9	77.9	76.5	79.9	63.7	70.6	70.1	77.0	82.8	83.3	80.4	80.9	84.8	54.5	74.7	72.7	63.6	74.7	80.8
tam_Taml	79.4	82.8	79.4	80.4	78.9	80.9	77.9	76.5	77.0	82.4	83.8	83.8	82.4	83.3	83.3	69.7	70.7	66.7	76.8	72.7	73.7
tel_Telu	77.9	85.3	84.3	79.4	78.4	82.8	78.9	77.0	77.0	85.8	84.3	85.8	83.3	86.8	87.3	60.6	77.8	67.7	63.6	71.7	77.8
urd_Arab	79.9	83.3	84.8	77.0	77.5	80.9	73.5	74.0	74.5	84.3	85.3	84.3	83.8	85.3	84.8	61.6	80.8	71.7	76.8	73.7	79.8
Avg.	70.9	**77.2**	76.9	74.5	74.6	**77.8**	67.8	70.5	**70.7**	79.1	79.8	**80.1**	76.4	78.5	**80.0**	60.6	**71.3**	65.2	67.2	68.5	**74.0**

Table 12: EMBSWAP results on SIB-200 with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; ES: EMBSWAP; +LR: EMBSWAP with LoRA Adaptation.

EMBSWAP Results on FLORES-200																					
Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
SEA																					
ind_Latn	65.1	63.9	65.7	68.4	66.9	67.0	61.5	54.5	61.6	65.9	64.9	65.4	65.6	64.5	65.6	68.7	67.1	66.6	67.0	54.2	67.3
tha_Thai	44.0	38.4	42.3	48.2	40.6	44.6	37.4	37.6	35.1	42.4	42.4	41.9	43.6	41.8	42.4	23.4	29.0	35.1	29.3	18.9	39.2
vie_Latn	57.0	56.4	57.2	60.5	59.7	59.6	49.4	49.9	50.3	57.1	57.3	56.5	57.2	56.8	58.0	61.8	62.4	60.0	61.8	57.3	60.9
zsm_Latn	62.6	63.8	64.8	67.9	66.8	67.0	56.6	59.6	60.7	63.1	65.1	64.5	64.6	63.6	65.1	50.5	62.6	63.9	54.2	46.0	64.9
tgl_Latn	55.2	55.4	56.4	59.5	59.1	59.4	51.4	52.1	54.2	54.9	56.6	57.0	55.2	54.1	57.7	37.1	52.0	55.7	47.2	50.2	56.7
mya_Mymr	9.3	2.6	21.8	37.1	29.6	32.8	12.9	13.3	23.8	14.7	33.2	27.8	16.0	16.9	31.1	3.6	9.3	24.3	7.0	20.4	28.1
lao_Laoo	29.9	30.8	39.9	40.5	27.6	38.0	27.3	26.7	34.3	24.3	40.9	40.1	25.0	1.2	41.7	9.6	26.0	31.1	17.0	25.4	37.5

EMBSWAP Results on FLORES-200

Model	PaLM2						Gemma2												Aya23					
Size	XXS			S			2B			9B			27B			8B			35B					
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR			
khm_Khmr	24.9	25.0	31.0	33.8	32.5	33.1	22.6	19.5	28.5	21.5	32.7	31.3	20.7	26.5	31.3	8.6	15.2	26.0	15.1	17.4	30.1			
ceb_Latn	35.4	37.4	56.5	57.0	52.8	58.8	34.1	34.1	53.7	37.0	48.0	57.7	46.9	52.4	58.0	27.0	46.0	55.9	32.3	34.2	56.7			
jav_Latn	30.3	19.3	46.7	49.8	48.3	49.7	30.4	30.8	46.7	36.2	33.1	50.6	37.7	38.9	51.0	19.4	45.2	48.5	29.0	37.4	50.4			
sun_Latn	32.7	42.4	42.2	45.2	43.9	46.1	30.4	31.5	41.5	33.8	42.1	43.9	36.3	40.9	43.4	26.8	24.1	43.8	31.0	34.9	43.6			
ilo_Latn	19.6	18.7	22.1	43.5	50.5	51.6	24.3	26.8	45.6	26.8	39.4	51.3	34.7	45.2	52.3	19.2	31.0	47.1	22.7	26.1	51.4			
war_Latn	15.4	31.3	31.9	38.3	43.2	54.3	28.8	30.5	52.9	32.8	37.8	57.0	40.2	51.5	58.5	23.8	34.1	55.3	27.6	27.5	58.0			
bug_Latn	21.7	21.8	24.7	23.4	25.3	27.5	13.5	17.3	23.9	25.7	25.7	26.6	25.5	25.2	27.1	9.1	16.0	23.0	23.8	17.4	26.7			
pag_Latn	25.7	23.0	23.9	26.3	28.1	28.6	25.8	26.7	40.7	27.9	28.5	42.5	26.5	36.4	43.6	21.0	26.6	42.4	24.7	21.9	43.8			
shn_Mymr	4.5	5.8	1.6	3.7	5.0	6.4	4.0	2.2	29.5	4.8	6.3	30.0	5.5	0.7	31.4	1.9	2.7	26.0	3.3	16.5	28.7			
min_Latn	28.1	19.6	17.9	46.1	48.9	52.5	35.1	34.4	47.1	37.3	43.0	50.7	37.6	45.2	52.5	34.2	43.1	50.0	35.6	35.5	52.2			
ace_Latn	16.8	22.9	28.7	28.8	28.5	36.7	11.1	7.8	33.1	28.6	32.9	37.5	28.1	32.9	37.9	2.3	19.2	35.5	1.0	24.7	39.0			
ban_Latn	26.3	29.6	31.8	33.2	34.4	38.3	29.8	29.0	42.0	32.6	38.4	44.2	32.0	40.9	45.5	21.9	38.1	42.5	30.7	34.2	44.9			
bjn_Latn	20.4	9.4	16.0	36.5	30.9	35.1	35.2	30.9	23.5	37.0	35.8	26.6	36.6	34.0	32.9	28.8	9.1	7.7	34.3	30.3	32.0			
ace_Arab	10.2	6.2	4.6	8.8	1.8	5.5	8.3	1.7	11.7	10.6	1.6	6.8	10.2	1.6	13.4	9.9	0.9	2.3	9.4	1.4	13.5			
bjn_Arab	10.5	7.5	8.0	5.0	4.6	10.3	6.6	1.7	10.2	10.9	9.8	8.3	10.3	1.6	17.9	10.1	1.3	8.4	10.0	3.6	20.6			
min_Arab	9.9	5.5	7.5	3.5	1.1	1.1	2.7	2.6	1.9	10.4	9.2	1.2	6.9	0.9	18.2	8.3	1.0	2.6	9.4	1.4	11.2			
Avg.	28.5	27.7	**32.3**	37.6	36.1	**39.3**	27.8	27.0	**37.1**	32.0	35.9	**40.0**	33.2	33.6	**42.5**	22.9	28.8	**37.1**	27.1	27.7	**41.6**			
AFR																								
swa_Latn	51.1	47.0	52.7	61.0	58.4	59.0	30.8	38.1	45.0	53.8	55.2	56.6	55.7	52.2	57.9	9.0	30.7	41.7	11.0	29.4	52.0			
lin_Latn	16.4	10.2	10.9	17.0	32.3	31.2	8.4	10.6	34.0	13.5	19.9	39.5	13.6	17.7	41.7	7.4	16.0	37.5	10.4	11.7	41.8			
yor_Latn	16.6	8.6	12.1	25.9	22.8	23.1	18.9	10.8	16.9	10.5	12.1	18.7	10.1	8.3	18.6	7.0	5.9	16.4	10.7	4.9	18.5			
ful_Latn	13.1	16.5	11.2	6.1	13.4	12.3	8.3	9.5	13.4	8.1	9.9	14.4	11.3	8.8	15.7	4.8	8.3	13.0	10.3	6.2	15.4			
ibo_Latn	27.9	20.8	29.2	34.4	35.5	36.9	21.8	20.4	27.9	23.2	29.5	33.7	23.1	13.0	34.0	8.5	17.0	27.0	15.2	15.3	33.9			
orm_Latn	1.7	5.7	7.5	11.7	16.0	22.1	9.1	13.1	15.6	9.8	9.1	24.8	10.7	11.1	24.6	9.2	4.1	13.4	10.8	10.0	23.3			
som_Latn	27.0	30.6	34.4	40.7	40.3	40.6	19.3	24.2	29.7	21.0	31.3	36.3	20.1	23.6	36.2	16.6	26.0	27.4	20.4	26.4	35.3			
tso_Latn	15.3	12.7	11.5	17.1	29.1	36.7	10.2	12.1	25.6	11.1	14.8	37.7	13.0	14.4	40.7	6.2	8.4	24.1	8.7	8.9	34.0			
nya_Latn	14.6	14.4	34.8	43.5	43.3	43.1	9.4	13.0	28.5	12.3	30.7	37.2	18.7	25.5	37.9	5.4	13.5	26.9	9.2	14.7	34.5			
zul_Latn	14.2	15.3	36.4	48.6	47.5	48.3	13.9	17.0	27.4	14.6	25.9	36.0	18.9	25.2	38.1	7.3	10.1	26.8	9.7	11.6	37.5			
kin_Latn	12.4	11.7	17.2	18.4	33.3	34.9	8.8	11.3	15.7	12.2	22.1	26.8	15.5	19.6	29.7	7.3	10.4	13.5	8.2	10.2	24.6			
run_Latn	14.3	14.0	15.6	16.5	28.0	30.1	9.9	11.9	15.9	10.0	19.2	24.6	12.8	17.6	26.5	6.8	11.0	15.9	8.8	10.2	21.7			
sna_Latn	11.4	15.4	29.6	39.4	39.6	39.9	7.9	14.9	26.7	12.6	26.9	34.3	19.8	26.2	35.7	4.8	14.9	24.7	7.9	11.5	32.4			
xho_Latn	25.3	14.8	30.5	43.3	43.2	44.4	16.3	17.0	29.9	14.9	27.0	35.8	20.7	25.4	37.4	7.5	12.0	27.1	9.7	11.8	35.8			
tsn_Latn	10.1	5.6	17.4	30.2	40.9	43.2	8.6	12.6	29.5	10.5	21.1	40.0	17.3	24.0	41.9	5.5	8.8	29.5	7.9	9.4	38.0			
tir_Ethi	0.9	0.6	1.0	7.1	7.8	10.8	0.4	0.4	13.5	4.2	7.6	19.4	1.9	2.9	19.1	1.3	3.7	10.2	1.2	2.7	15.8			
kik_Latn	13.3	11.1	16.0	15.8	17.5	13.7	13.6	11.7	7.8	18.7	18.2	14.5	15.7	12.0	15.2	4.9	7.5	4.2	7.1	9.7	15.4			
kon_Latn	14.0	9.6	9.9	14.6	20.5	28.7	8.1	10.0	33.5	10.9	19.4	41.8	15.5	11.5	41.3	7.2	11.0	37.0	8.5	10.5	42.0			
lua_Latn	15.2	8.6	19.9	10.4	25.1	25.3	6.8	10.3	22.7	8.9	14.9	26.7	16.0	7.3	30.8	5.8	11.3	21.0	7.3	7.5	24.8			
umb_Latn	10.4	12.2	14.3	12.8	16.1	16.0	9.4	11.5	7.3	12.7	8.6	8.2	15.7	3.4	11.2	6.4	7.6	5.6	12.7	8.0	10.5			
sot_Latn	11.4	13.6	30.9	43.9	43.2	45.8	9.2	11.6	34.9	10.8	25.2	40.7	20.1	26.1	41.5	6.1	10.9	32.7	8.1	10.4	39.8			
mos_Latn	14.6	12.4	15.3	14.4	14.1	14.7	12.9	10.2	8.8	13.3	7.3	7.9	14.7	13.2	9.6	3.0	11.2	2.9	11.1	7.8	6.5			
nso_Latn	12.3	15.4	14.5	28.4	34.4	39.1	8.7	11.2	26.8	9.5	18.0	39.4	17.9	24.6	42.1	6.2	8.2	29.3	8.5	10.9	36.3			
knc_Latn	16.4	8.2	8.5	10.5	7.0	6.7	8.0	7.6	9.5	13.8	13.0	15.9	11.6	5.2	14.4	5.3	7.4	9.2	8.7	8.7	14.1			
knc_Arab	7.2	0.8	0.6	6.9	6.3	5.8	3.9	0.9	1.0	7.5	3.2	1.1	7.4	2.3	2.1	7.0	0.5	0.2	6.8	0.5	0.7			
luo_Latn	16.0	12.9	16.8	13.1	12.8	9.8	11.2	9.4	12.3	17.2	17.5	13.1	11.7	4.6	13.4	5.4	11.2	6.1	7.3	8.5	12.1			
bem_Latn	11.3	7.6	9.9	15.3	19.7	16.7	8.5	9.1	6.1	9.7	13.4	10.5	14.4	10.2	13.6	5.5	1.9	0.7	7.6	8.8	15.2			
lug_Latn	8.3	14.1	11.0	13.5	29.2	30.2	8.0	11.2	15.6	9.1	17.3	25.5	13.5	16.3	27.2	5.7	8.8	12.3	7.5	9.1	22.0			
wol_Latn	9.1	16.3	11.2	16.3	15.9	14.8	9.0	10.3	7.4	14.8	8.8	9.2	16.1	7.6	11.8	5.4	5.0	5.5	13.7	8.2	9.2			
kmb_Latn	12.7	2.3	12.0	14.7	15.6	15.6	7.8	7.1	14.4	14.0	11.3	21.2	14.7	3.8	23.6	7.3	9.7	15.6	11.6	7.7	20.2			
kam_Latn	6.9	10.3	15.1	19.9	19.4	14.3	11.5	9.9	12.3	22.2	22.4	18.1	21.0	9.7	19.3	5.7	2.5	0.3	8.5	12.3	14.0			
ewe_Latn	15.1	14.9	10.0	3.8	10.6	11.0	6.2	9.0	16.9	5.5	9.9	28.3	7.3	5.4	29.7	4.3	6.3	18.2	5.3	6.4	27.4			
ssw_Latn	16.1	16.5	21.0	14.9	32.5	34.1	1.7	12.9	21.5	9.2	19.9	28.1	16.5	20.4	31.3	2.9	9.1	17.6	8.5	9.4	27.2			
tum_Latn	9.0	12.3	25.3	24.2	32.0	33.0	8.4	10.8	23.1	13.9	18.6	29.4	14.1	15.8	30.2	5.6	11.4	20.8	8.6	10.7	26.2			
fon_Latn	10.9	10.2	5.1	10.0	9.2	12.7	8.4	5.3	12.1	7.7	7.0	18.3	11.1	3.7	19.2	6.6	5.2	14.4	6.9	3.6	18.1			
din_Latn	15.2	5.1	7.7	1.6	8.6	10.4	5.8	7.4	9.3	6.7	7.5	15.9	15.2	8.0	15.3	5.5	6.0	11.3	9.4	5.7	14.9			
kbp_Latn	9.3	4.7	7.5	10.7	11.1	15.6	6.7	6.1	13.7	5.8	7.1	19.4	10.4	3.7	21.3	3.6	5.1	12.2	6.8	6.6	19.2			
cjk_Latn	14.1	6.1	12.1	16.1	13.2	14.3	8.0	9.2	12.2	15.1	8.8	14.2	15.6	8.8	14.9	6.6	6.9	11.1	7.7	7.7	13.3			
nus_Latn	9.6	9.9	7.8	1.7	2.1	2.8	4.8	5.0	13.4	4.5	6.9	17.8	8.9	4.6	17.2	2.7	3.7	12.4	8.1	5.0	17.0			
taq_Latn	16.7	16.8	4.4	4.6	8.0	7.6	7.9	7.5	8.8	8.5	8.3	13.3	14.1	5.1	13.9	7.3	6.5	7.3	8.7	8.8	10.1			
taq_Tfng	2.1	2.5	1.0	1.1	1.1	0.8	0.3	0.3	0.2	0.8	0.8	0.9	1.2	0.2	1.3	0.7	0.4	0.1	3.5	0.9	0.9			
sag_Latn	15.8	15.7	13.0	9.5	13.7	15.0	7.4	7.5	22.0	16.4	16.5	31.6	15.0	6.4	33.2	3.7	7.8	28.2	14.4	10.4	32.1			
Avg.	13.9	12.2	**16.0**	19.3	23.1	**24.1**	9.6	10.9	**18.3**	12.6	16.5	**24.4**	15.2	13.2	**25.7**	6.0	9.1	**16.9**	9.1	9.5	**23.4**			
IND																								
hin_Deva	53.4	50.8	54.0	58.7	56.2	58.1	45.2	37.6	49.2	51.1	52.0	54.1	50.1	48.2	53.8	55.4	24.6	49.6	54.6	38.5	54.1			
ben_Beng	32.7	29.5	39.8	43.5	41.9	44.8	25.0	24.6	36.7	32.1	35.9	42.4	35.5	26.7	44.2	17.7	7.7	39.1	23.1	18.6	43.1			
urd_Arab	34.0	38.5	42.8	48.2	47.2	48.4	27.2	30.5	40.1	37.2	42.7	44.5	39.8	39.4	44.7	18.4	13.3	38.3	26.1	14.1	45.3			
tel_Telu	35.1	33.5	41.5	43.1	43.7	43.8	29.9	25.8	40.8	37.8	35.5	48.2	39.1	27.3	49.0	18.3	19.1	39.9	26.5	18.9	47.5			
tam_Taml	35.1	40.0	42.9	50.4	47.7	48.0	29.2	24.5	35.5	37.5	39.5	46.9	39.1	19.6	47.1	21.1	16.0	37.3	31.9	22.3	47.2			
mar_Deva	33.4	35.2	38.5	44.1	41.0	43.1	24.7	25.9	36.2	30.3	35.2	41.7	30.2	31.6	41.1	20.0	17.3	36.9	26.5	17.5	41.4			
mai_Deva	12.5	2.3	0.7	30.0	4.5	38.2	17.4	11.7	36.1	29.0	26.3	43.7	31.1	35.6	44.8	17.9	0.3	40.6	30.0	4.3	44.8			

EMBSWAP Results on FLORES-200																					
Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
bho_Deva	18.2	8.0	25.2	34.2	34.8	38.0	24.1	16.7	34.5	29.8	31.6	39.4	29.0	31.5	39.1	25.1	3.2	37.7	27.8	9.9	39.7
pbt_Arab	8.1	2.6	18.3	29.1	28.3	31.0	9.5	10.0	22.5	12.4	21.4	29.3	10.9	16.4	27.3	6.5	13.1	17.7	9.2	3.2	29.8
guj_Gujr	35.4	4.1	9.7	48.5	39.4	45.6	29.9	25.8	41.9	35.9	40.9	47.0	39.2	33.6	45.9	16.7	14.6	42.1	29.2	12.8	46.0
kan_Knda	28.1	25.2	36.1	47.1	44.6	46.5	10.5	9.5	37.1	27.8	36.4	45.6	37.6	22.7	46.0	9.2	24.1	39.3	15.4	20.4	46.4
awa_Deva	2.9	11.0	38.6	35.7	40.5	39.1	30.2	13.8	34.9	38.5	40.1	42.3	37.6	37.1	42.5	8.4	11.4	39.5	36.1	9.5	42.3
ory_Orya	11.7	3.2	8.5	34.4	21.7	25.8	14.2	7.7	17.9	13.0	26.4	37.2	15.3	6.6	35.7	12.1	3.8	11.0	11.0	4.4	31.0
mal_Mlym	25.6	21.1	28.4	40.8	38.0	40.1	23.8	15.9	30.7	29.4	30.8	41.6	31.0	7.5	44.3	18.0	3.6	34.3	25.0	16.8	44.0
pan_Guru	20.0	14.3	28.0	46.0	43.8	45.8	22.3	19.4	37.7	32.9	38.2	44.7	37.9	10.4	44.6	13.5	23.5	38.0	27.2	22.7	44.3
hne_Deva	2.5	21.0	36.9	33.4	40.1	43.4	16.2	15.9	37.8	34.6	29.7	47.2	37.2	33.3	46.5	18.4	20.4	41.4	33.7	13.2	46.3
npi_Deva	33.4	27.4	37.1	49.3	46.1	48.4	26.7	27.1	40.1	36.1	42.7	46.0	37.6	38.5	46.5	19.7	3.6	40.3	28.2	21.6	45.8
asm_Beng	2.3	6.3	21.7	32.9	31.7	35.0	11.0	8.5	23.6	15.1	24.5	32.6	21.7	17.0	33.7	2.5	6.1	26.7	11.2	10.0	33.4
mni_Beng	0.4	1.2	6.6	0.8	7.2	17.9	0.5	1.5	21.7	4.6	3.8	31.0	0.9	0.1	30.6	0.4	10.1	22.6	0.2	2.4	29.1
bod_Tibt	2.2	1.0	0.7	0.5	5.2	7.8	0.8	3.6	10.2	1.1	7.2	19.0	2.5	1.0	18.5	3.6	4.7	9.6	4.6	4.5	15.2
san_Deva	13.8	0.6	4.2	20.0	23.4	24.1	16.5	5.3	16.7	18.1	19.0	26.0	18.0	17.0	25.7	15.8	0.4	20.3	17.7	5.4	25.6
Avg.	21.0	17.9	**26.7**	36.7	34.6	**38.7**	20.7	17.2	**32.5**	27.8	31.4	**40.5**	29.6	23.9	**40.6**	16.1	11.5	**33.4**	23.6	13.9	**40.1**

Table 13: EMBSWAP results on FLORES-200 with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; ES: EMBSWAP; +LR: EMBSWAP with LoRA Adaptation.

EMBSWAP Results on GSM8K-NTL																		
Model	Gemma2						Gemma2						Gemma2					
Size	2B		9B		27B		2B		9B		27B		2B		9B		27B	
Variant	FL	ES	FL	ES	FL	ES	IT	ES	IT	ES	IT	ES	MH	ES	MH	ES	MH	ES
asm_Beng	11.2	14.4	30.8	26.4	34.8	35.2	15.6	12.4	48.0	42.4	53.2	48.0	20.4	29.2	46.0	28.0	43.2	36.0
bew_Latn	28.4	24.8	48.4	46.4	48.8	50.4	41.6	30.0	58.0	57.6	62.4	53.2	51.6	42.0	70.0	68.8	71.6	68.8
bho_Deva	16.4	17.2	29.2	35.6	38.0	35.6	23.2	19.2	48.4	48.0	50.8	51.6	29.2	36.4	51.2	48.8	49.6	49.6
doi_Deva	11.2	14.0	19.6	34.0	26.4	33.6	13.6	19.6	32.4	44.0	37.6	46.0	17.6	25.2	36.4	36.4	29.2	37.6
div_Thaa	6.8	10.4	9.6	23.2	18.8	26.8	2.8	12.4	15.6	36.8	22.0	32.8	2.8	15.6	17.2	26.4	13.2	12.0
dzo_Tibt	2.0	2.8	6.8	12.8	10.4	8.4	0.4	3.6	10.0	14.0	14.0	20.4	0.8	3.6	11.6	6.4	8.8	4.0
efi_Latn	8.0	10.4	13.2	27.6	16.0	30.8	5.2	5.6	17.2	31.6	24.8	20.8	6.8	20.0	21.6	48.0	19.6	38.0
gom_Deva	15.2	15.6	25.2	35.6	28.4	37.2	13.6	14.4	37.2	48.4	44.0	49.2	11.6	30.4	38.4	47.6	36.4	41.6
ilo_Latn	8.4	15.6	24.8	35.2	28.8	39.2	10.4	11.2	30.8	46.4	41.2	44.8	12.8	22.4	42.8	57.2	43.6	56.4
kri_Latn	10.8	8.0	23.2	26.0	28.8	29.6	12.0	8.4	32.0	34.8	36.0	36.4	19.2	22.8	43.2	43.6	40.0	42.8
mai_Deva	13.6	14.4	29.2	35.2	37.2	38.8	23.6	17.6	47.2	48.4	48.8	52.0	25.2	32.4	46.8	50.8	46.0	47.2
meo_Latn	24.4	24.4	46.0	45.2	45.2	47.2	37.6	30.0	58.4	56.4	60.0	56.4	58.4	40.8	74.8	71.6	72.0	68.0
mfa_Arab	6.4	10.8	6.8	37.2	9.2	42.8	3.6	18.0	12.0	52.0	8.8	49.2	3.2	25.2	7.2	54.8	7.2	48.0
min_Latn	12.0	18.8	32.4	38.0	35.6	44.8	15.6	22.0	42.8	52.8	49.6	44.4	22.0	29.2	45.2	58.0	54.8	53.6
mni_Beng	4.8	5.2	4.8	16.8	8.8	17.6	1.6	6.0	5.2	20.8	8.4	25.2	5.6	10.0	4.8	17.6	8.0	4.8
mzn_Arab	21.6	20.0	40.0	42.8	46.8	48.4	29.2	19.2	56.4	56.4	60.0	59.6	33.6	32.8	65.6	59.2	61.2	52.4
nso_Latn	4.4	10.8	14.8	26.0	21.2	24.0	4.4	7.2	19.2	26.8	25.6	13.6	7.2	10.8	24.4	36.4	22.4	23.2
ory_Orya	10.0	12.8	19.2	26.4	24.8	27.2	2.8	11.6	29.6	34.8	37.2	42.4	6.4	21.6	33.6	26.4	27.2	23.6
pcm_Latn	28.0	23.6	46.8	43.2	48.4	50.0	43.6	36.0	60.0	60.0	61.6	51.6	61.2	62.8	77.6	72.4	77.6	74.4
tso_Latn	8.8	9.6	12.4	16.4	17.2	20.4	4.0	5.2	14.4	24.8	22.4	10.0	7.6	8.8	19.6	24.8	17.6	20.0
Avg	12.6	**14.2**	24.2	**31.5**	28.7	**34.4**	15.2	**15.5**	33.7	**41.9**	38.4	**40.4**	20.2	**26.1**	38.9	**44.2**	37.5	**40.1**

Table 14: EMBSWAP results on GSM8K-NTL with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; IT: LLMs aligned with supervised fine-tuning and reinforcement learning with human feedback; MH: LLMs instruction-tuned on the WebInstruct math dataset; ES: EMBSWAP;

EMBSWAP Results on XSUM-IN																					
Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
eng_Latn	33.7	34.2	30.7	41.1	40.7	41.2	36.6	34.1	31.9	37.5	37.1	37.0	37.6	36.8	36.9	34.5	25.7	30.7	37.9	22.1	37.1
asm_Beng	0.3	0.3	0.3	25.5	16.1	6.3	2.7	1.3	7.6	13.7	16.5	20.1	17.8	11.3	21.3	5.0	1.3	4.6	6.7	6.7	19.4
awa_Deva	0.1	0.1	0.1	6.6	20.5	11.1	5.2	5.9	14.2	19.0	17.3	17.2	13.7	17.3	17.6	0.2	0.0	4.5	9.4	8.8	17.1
bgc_Deva	0.2	0.2	0.2	5.6	3.2	4.3	0.2	0.7	11.8	16.5	0.9	16.1	12.1	15.6	16.2	16.1	0.2	0.4	3.1	8.4	16.4
bho_Deva	0.1	0.1	0.1	19.3	21.6	18.5	2.0	6.9	9.9	19.1	18.0	16.7	14.2	15.8	18.1	16.3	0.0	7.5	3.7	12.8	18.1
ben_Beng	16.7	2.2	0.1	29.8	28.3	25.9	6.7	14.4	16.9	22.4	23.9	21.6	25.5	20.0	22.0	16.9	11.3	17.2	17.0	14.5	22.2
bod_Tibt	0.2	0.2	0.1	10.8	9.2	9.8	0.2	0.9	5.7	0.5	3.4	8.8	0.8	0.8	7.4	4.6	2.5	6.0	3.7	4.0	8.6
brx_Deva	0.2	0.2	0.2	0.3	1.0	3.5	0.2	0.7	1.5	0.2	1.3	5.2	0.3	11.4	14.8	4.5	0.0	0.7	0.2	0.4	8.4
gbm_Deva	4.5	3.8	3.7	7.7	19.5	18.0	4.6	4.0	6.8	14.1	8.2	12.8	16.6	14.3	14.8	14.8	0.2	6.5	8.2	7.9	12.3

EMBSWAP Results on XSUM-IN

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
gom_Deva	0.2	0.2	0.2	0.2	7.3	3.7	0.4	5.0	10.3	14.2	14.9	17.5	12.1	17.4	18.5	13.2	0.3	10.5	0.4	6.4	15.2
guj_Gujr	11.8	0.4	0.4	25.1	24.4	20.4	6.7	11.8	16.0	18.8	21.1	19.4	22.1	16.3	19.7	13.5	11.2	16.3	16.8	0.2	19.1
hin_Deva	22.1	1.3	2.4	31.2	30.5	27.6	4.7	14.3	19.2	24.2	24.4	24.9	23.6	18.6	23.7	26.4	11.2	18.2	27.3	17.7	26.3
hne_Deva	0.2	0.2	0.3	3.5	20.1	20.5	2.2	2.8	5.8	20.2	18.3	17.9	13.3	16.5	18.4	18.7	0.0	2.0	8.6	10.3	17.2
hoj_Deva	0.2	0.2	0.1	0.6	0.5	0.5	1.1	1.0	0.5	1.9	14.9	15.7	3.1	4.6	12.8	0.8	0.1	9.8	0.3	0.1	0.7
kan_Knda	4.4	2.7	0.7	28.6	28.3	28.9	7.9	4.9	14.9	17.9	21.1	23.4	24.0	8.5	24.3	8.2	7.2	15.8	13.4	12.4	22.1
mai_Deva	0.1	0.1	0.1	13.7	19.1	17.3	2.4	1.6	4.0	19.0	18.2	17.6	19.1	15.2	17.1	14.5	0.1	11.8	3.4	5.3	15.7
mal_Mlym	13.5	2.1	2.0	27.2	27.1	27.7	5.0	8.6	15.4	21.2	21.5	22.4	23.4	4.0	22.4	16.9	3.2	14.1	19.2	10.9	21.2
mni_Beng	0.2	0.2	0.2	0.2	8.3	4.9	0.1	1.7	3.1	4.9	1.4	8.7	0.1	0.0	10.6	0.8	1.8	3.0	1.3	0.3	8.6
mar_Deva	16.6	10.9	1.0	30.5	28.9	27.4	13.7	14.1	16.7	19.7	19.1	20.7	21.2	21.8	19.9	17.6	5.1	16.1	21.5	11.2	21.0
mup_Deva	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.4	2.8	2.5	9.6	14.9	9.3	14.5	16.0	8.5	0.0	0.3	2.3	3.0	0.9
mwr_Deva	0.2	0.2	0.2	0.3	1.4	1.9	0.8	1.5	2.7	19.2	8.3	18.8	9.3	17.4	18.6	9.2	0.1	0.2	3.4	1.0	12.7
npi_Deva	2.3	0.0	0.0	29.1	23.9	5.3	9.3	15.4	17.3	21.5	24.1	22.8	11.1	23.5	20.2	18.5	3.5	16.5	20.7	12.7	19.5
ory_Orya	0.6	0.6	0.6	23.8	18.8	22.2	3.4	8.5	12.9	13.6	14.8	17.7	24.6	5.7	16.8	12.8	2.6	3.7	14.2	5.0	13.8
pan_Guru	7.9	1.3	0.2	21.5	23.3	22.6	3.7	11.8	12.4	17.4	18.8	18.3	16.2	1.6	21.0	8.2	9.7	15.8	14.5	9.4	20.4
pbu_Arab	0.2	0.2	0.2	6.2	13.5	4.3	1.8	7.2	10.1	11.4	15.3	18.0	16.4	11.0	18.3	7.6	11.7	15.2	5.7	3.1	18.2
san_Deva	0.1	0.1	0.1	19.8	17.7	19.3	2.4	5.4	11.7	16.3	15.3	16.1	8.8	15.1	15.8	16.4	0.0	4.2	15.9	3.6	8.1
sat_Olck	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	15.3	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0
tam_Taml	7.0	12.7	1.0	35.5	34.5	35.3	4.3	10.0	9.0	27.2	28.0	26.7	0.0	12.4	27.2	21.7	11.5	18.9	25.2	16.8	26.6
tel_Telu	16.2	12.5	12.2	26.3	26.1	26.7	13.4	13.9	13.5	20.5	18.6	21.6	28.9	7.6	22.5	14.5	8.9	13.5	17.2	11.4	21.3
urd_Arab	12.8	6.3	2.3	29.1	28.8	26.5	1.9	17.1	14.9	20.5	24.1	24.0	22.6	21.6	24.8	13.6	8.8	19.7	16.9	7.4	25.0
Avg.	**5.8**	3.1	2.0	16.6	**18.1**	16.1	4.8	7.5	**10.6**	15.8	15.9	**18.1**	17.6	13.2	**18.6**	**12.5**	4.6	10.1	11.3	7.8	**16.4**

Table 15: EMBSWAP results on XSUM-IN with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; ES: EMBSWAP; +LR: EMBSWAP with LoRA Adaptation.

EMBSWAP Results on XORQA-IN

Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
ANSWER IN EN																					
eng_Latn	70.9	75.5	74.8	75.1	82.1	82.6	71.2	66.6	73.3	75.7	74.6	74.4	75.5	78.5	75.9	75.6	72.6	71.0	78.4	65.2	74.9
asm_Beng	47.9	63.0	68.4	52.0	51.1	58.3	54.3	61.3	67.7	71.6	72.4	71.2	65.7	64.3	74.1	56.7	67.2	57.9	64.8	69.2	73.9
awa_Deva	48.4	64.0	64.6	49.7	52.9	57.7	49.6	54.2	57.9	63.2	66.3	68.1	63.3	58.8	70.7	68.3	64.9	63.2	77.1	67.6	72.9
bgc_Deva	43.4	61.1	67.0	52.9	59.5	58.1	45.1	53.0	60.4	57.9	68.6	70.9	60.8	57.1	73.7	70.1	69.0	66.4	74.6	67.9	71.1
bho_Deva	50.8	64.8	69.7	47.4	58.5	58.5	50.4	57.5	59.4	63.8	67.2	67.0	62.0	60.8	72.8	68.8	67.2	65.8	73.2	65.1	69.4
ben_Beng	57.3	62.5	71.0	52.0	58.5	63.1	59.8	64.1	67.7	74.5	74.2	72.6	65.9	64.9	73.7	68.4	69.0	63.1	76.6	68.6	70.2
bod_Tibt	3.2	2.6	17.5	23.5	12.4	38.5	11.4	36.0	40.9	27.5	41.7	33.8	43.5	44.7	60.8	32.0	30.1	23.5	30.3	36.4	30.6
brx_Deva	5.3	45.6	58.1	25.6	57.6	61.7	12.4	40.9	49.5	19.6	56.5	61.2	25.2	48.1	60.5	28.2	47.4	52.4	28.0	46.3	63.4
gbm_Deva	40.9	59.1	66.1	53.0	53.6	60.1	46.3	51.1	57.6	56.0	66.0	68.4	56.0	56.6	71.6	61.3	64.6	64.5	66.3	62.8	68.8
gom_Deva	38.7	64.8	69.5	50.4	67.3	66.9	33.8	54.6	63.0	58.8	69.2	69.8	60.7	55.6	71.3	44.1	65.8	61.6	51.0	62.9	72.1
guj_Gujr	45.1	70.3	72.7	44.0	43.5	54.8	60.4	61.6	67.4	67.9	70.3	59.8	72.4	66.8	75.2	57.1	68.2	64.4	73.4	70.2	75.6
hin_Deva	61.8	72.5	73.2	55.0	47.5	52.1	68.8	66.7	68.7	72.6	71.8	64.9	65.5	57.7	75.4	74.7	70.4	67.0	76.8	71.5	76.8
hne_Deva	49.0	61.8	66.9	47.3	58.5	58.3	45.2	51.6	59.8	58.3	65.8	67.8	65.8	59.3	71.4	65.6	67.6	61.2	73.8	64.2	70.4
hoj_Deva	40.1	56.7	60.4	48.6	50.8	56.2	41.3	48.5	54.3	54.7	60.7	64.7	56.8	61.0	70.7	59.5	62.0	56.9	69.0	61.0	68.6
kan_Knda	45.0	66.7	69.6	44.9	53.9	57.9	65.0	63.3	68.2	69.8	71.2	63.4	70.1	63.3	70.1	60.6	68.3	61.6	66.6	71.4	75.1
mai_Deva	49.0	71.9	72.8	49.3	50.9	55.2	50.2	59.4	63.9	63.6	69.2	69.6	64.9	58.1	72.7	67.0	68.4	64.7	69.8	66.7	72.1
mal_Mlym	51.8	69.9	71.0	56.8	57.1	63.7	68.1	69.2	72.5	74.7	71.2	67.7	70.3	63.3	74.4	71.8	66.9	61.2	78.0	74.0	76.9
mni_Beng	9.8	28.6	40.2	37.0	46.6	56.5	20.6	27.3	31.6	29.4	43.2	48.9	33.6	47.8	50.2	30.8	39.3	34.7	32.1	36.8	44.7
mar_Deva	55.8	68.1	71.5	43.6	50.0	55.4	57.9	61.6	65.2	70.7	71.3	67.0	62.1	55.7	71.8	63.8	65.0	58.4	73.6	65.6	71.6
mup_Deva	47.4	66.4	68.8	50.9	50.4	57.5	56.5	52.7	59.2	65.0	67.7	65.2	63.7	59.1	69.6	66.3	63.3	60.3	68.4	63.7	69.7
mwr_Deva	46.5	64.8	65.6	51.6	52.8	57.0	42.0	51.3	58.1	57.0	63.6	66.8	61.3	59.1	70.1	63.4	62.3	58.1	72.6	59.0	67.2
ory_Orya	25.1	57.5	68.3	45.0	43.8	55.4	34.3	48.2	58.2	51.0	64.8	66.4	62.6	58.4	69.0	56.4	45.5	37.6	71.0	60.8	64.7
pan_Guru	41.4	69.4	71.4	42.7	39.8	51.8	61.7	62.8	70.1	67.3	69.2	60.3	70.1	67.3	73.7	59.9	70.6	63.0	71.1	69.4	75.1
pbu_Arab	46.9	60.4	68.1	71.3	67.3	64.6	31.1	49.0	61.1	59.3	67.8	62.3	63.6	68.2	74.0	39.9	65.5	60.0	43.2	61.8	70.8
san_Deva	43.9	61.3	67.3	56.4	58.5	64.5	41.0	50.8	61.1	62.5	66.8	66.1	59.4	51.1	71.6	52.2	63.7	52.0	61.3	63.3	66.3
sat_Olck	0.8	5.3	13.3	8.9	24.9	29.7	18.3	20.7	35.6	48.6	37.1	46.0	59.1	48.6	47.6	31.8	19.3	9.0	31.0	30.8	32.1
tam_Taml	51.4	69.1	73.3	41.4	56.0	57.5	64.2	63.1	66.3	69.9	65.0	61.6	69.2	56.1	73.3	72.9	69.2	61.5	78.1	71.6	75.4
tel_Telu	45.9	71.0	72.4	36.7	45.0	53.2	65.4	65.4	71.1	72.1	70.1	67.4	72.8	63.8	74.5	62.7	66.6	59.0	73.3	68.9	76.5
urd_Arab	56.1	67.2	68.3	48.2	53.1	56.1	61.2	55.4	63.6	65.7	68.3	66.4	68.4	66.6	71.6	66.0	66.6	58.6	72.2	63.5	68.7
Avg.	42.1	59.4	**64.2**	46.9	51.8	**57.3**	47.8	54.1	**60.5**	60.3	**65.2**	64.1	61.7	59.3	**70.1**	58.5	**61.6**	56.5	64.7	62.3	**67.8**
ANSWER IN IND																					
asm_Beng	2.6	1.9	3.1	15.0	6.8	5.6	0.6	0.2	1.7	1.0	4.3	3.6	15.3	7.9	13.4	1.8	2.6	5.8	0.8	0.7	8.0
awa_Deva	5.7	15.3	16.5	25.2	30.7	33.1	6.8	6.6	8.0	6.8	6.3	9.9	23.7	20.0	16.9	14.8	2.3	8.8	15.8	7.1	10.9
bgc_Deva	11.8	9.4	9.8	21.2	14.3	13.3	10.8	11.2	12.4	11.6	13.2	12.2	20.7	18.1	19.4	17.0	7.9	13.1	18.2	9.0	16.3

EMBSWAP Results on XORQA-IN																					
Model	PaLM2						Gemma2									Aya23					
Size	XXS			S			2B			9B			27B			8B			35B		
Variant	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR	FL	ES	+LR
bho_Deva	5.8	3.4	5.3	20.4	22.1	17.7	1.6	3.7	2.8	1.5	2.0	2.3	16.8	13.4	9.0	5.5	2.6	2.6	7.3	2.1	8.8
ben_Beng	7.4	2.5	0.6	7.5	16.6	9.6	0.8	1.1	1.8	1.8	1.8	1.4	13.5	9.3	5.0	0.7	6.2	4.3	0.7	1.8	3.2
bod_Tibt	0.2	0.9	0.7	7.2	4.6	7.3	1.1	5.3	6.3	3.8	5.7	4.0	5.2	4.9	5.3	1.2	0.6	1.1	2.7	1.1	4.3
brx_Deva	0.8	13.8	17.7	15.0	22.5	22.7	3.1	15.4	17.9	6.5	18.0	19.9	9.3	15.8	20.0	6.4	11.0	16.9	6.6	7.0	18.5
gbm_Deva	10.4	0.9	2.5	20.2	8.3	8.5	11.5	12.9	13.3	12.3	13.4	13.3	18.1	15.1	17.4	16.3	6.7	13.1	16.2	7.6	16.2
gom_Deva	1.4	1.6	3.1	9.2	7.8	2.5	1.1	2.9	2.2	1.6	6.0	5.9	14.9	11.2	8.8	5.7	2.6	3.7	2.3	1.8	5.4
guj_Gujr	15.6	12.3	20.4	33.8	25.6	26.2	9.8	11.6	13.1	17.0	18.7	17.2	26.6	16.8	19.5	10.2	12.9	10.6	15.5	7.1	13.4
hin_Deva	40.7	28.1	35.2	59.5	51.4	51.0	23.0	25.3	25.6	25.3	27.3	28.5	45.0	37.6	37.9	21.4	22.1	21.8	22.7	18.1	25.4
hne_Deva	15.9	20.6	20.7	29.3	30.2	28.2	13.5	13.5	13.9	16.0	18.3	17.6	31.9	25.8	25.3	21.0	9.0	14.1	23.5	10.3	19.6
hoj_Deva	14.3	16.4	17.6	28.3	28.6	25.4	14.8	17.9	18.4	15.5	20.8	20.6	33.4	30.4	29.3	20.0	17.7	18.4	21.8	13.8	20.6
kan_Knda	10.8	11.6	10.4	29.4	23.1	18.2	11.3	8.8	12.3	11.8	14.2	18.1	32.2	15.6	25.0	8.4	10.3	10.9	10.6	11.9	20.6
mai_Deva	9.4	1.6	4.3	30.0	26.9	24.1	10.8	10.8	11.6	12.2	18.9	13.7	27.1	23.0	18.8	18.0	7.1	13.8	14.1	8.0	16.3
mal_Mlym	21.8	25.5	27.4	35.8	41.1	37.7	20.6	18.6	21.8	30.3	26.3	29.9	35.1	19.7	32.7	21.6	19.2	19.3	26.1	15.1	25.7
mni_Beng	0.8	0.3	0.2	1.1	10.3	4.2	0.6	0.1	0.3	0.7	0.6	0.6	1.5	0.4	2.0	1.6	1.1	2.3	0.4	0.2	3.6
mar_Deva	31.1	30.6	32.9	38.5	43.2	36.0	22.3	22.4	25.6	24.6	27.2	23.5	40.1	31.0	30.3	24.8	17.7	20.8	34.5	19.6	25.9
mup_Deva	4.6	16.9	16.4	12.4	21.5	23.8	5.3	5.8	5.4	5.9	6.4	7.3	14.8	11.6	10.0	6.9	3.9	5.0	5.8	3.6	6.2
mwr_Deva	2.4	1.6	3.3	10.5	12.1	9.7	1.6	1.7	2.4	4.2	8.7	5.8	14.8	11.7	10.8	6.7	2.2	5.8	6.0	2.0	4.0
ory_Orya	5.7	5.6	6.1	15.4	14.1	11.6	6.6	6.5	9.9	11.1	12.0	17.0	17.7	12.4	17.4	10.9	6.4	8.1	16.0	8.9	12.4
pan_Guru	8.6	7.1	10.8	20.7	11.5	11.5	4.7	5.8	7.0	7.4	10.6	12.8	21.9	7.2	16.4	5.3	12.3	11.0	7.0	5.6	12.7
pbu_Arab	3.8	6.9	6.8	8.2	8.1	7.5	1.2	2.9	3.6	3.2	5.1	3.5	7.4	8.8	6.0	3.2	4.5	4.9	3.9	3.5	8.1
san_Deva	3.1	4.9	5.1	29.7	15.9	17.4	1.3	2.7	3.5	7.6	12.0	11.4	19.4	11.1	11.8	10.3	1.7	4.4	7.8	3.0	10.3
sat_Olck	1.3	0.2	0.1	7.2	0.2	0.4	0.0	0.0	0.1	0.0	0.0	2.2	0.0	0.0	0.6	0.8	0.6	1.0	0.3	0.0	0.8
tam_Taml	16.8	20.5	19.0	23.8	32.7	27.1	13.9	12.8	15.0	17.2	24.6	17.7	29.8	16.8	22.6	14.5	14.1	12.3	17.5	11.8	17.3
tel_Telu	10.7	23.6	23.1	23.6	38.6	31.7	12.4	12.3	14.5	13.9	15.5	15.3	21.2	12.9	18.1	14.0	11.0	12.6	14.0	8.8	15.2
urd_Arab	8.4	6.7	9.3	17.6	22.7	17.4	1.7	2.5	3.5	4.1	11.3	4.1	13.4	16.5	9.3	4.3	2.9	3.3	1.7	1.9	4.0
Avg.	9.7	10.4	**11.7**	**21.3**	21.1	18.9	7.6	8.6	**9.8**	9.8	**12.5**	12.1	**20.4**	15.2	16.4	**10.5**	7.8	9.6	11.4	6.8	**12.6**

Table 16: EMBSWAP results on XORQA-IN with zero-shot prompting. FL: LLMs instruction-tuned on FLAN mixture; ES: EMBSWAP; +LR: EMBSWAP with LoRA Adaptation.