Semantic Aware Linear Transfer by Recycling Pre-trained Language Models for Cross-lingual Transfer

Seungyoon Lee, Seongtae Hong, Hyeonseok Moon, Heuiseok Lim[†]
Korea University, Republic of Korea
{dltmddbs100, ghdchlwls123, glee889, limhseok}@korea.ac.kr

Abstract

Large Language Models (LLMs) are increasingly incorporating multilingual capabilities, fueling the demand to transfer them into target language-specific models. However, most approaches, which blend the source model's embedding by replacing the source vocabulary with the target language-specific vocabulary, may constrain expressive capacity in the target language since the source model is predominantly trained on English data. In this paper, we propose **S**emantic **A**ware **L**inear **T**ransfer (**SALT**), a novel cross-lingual transfer technique that recycles embeddings from target language Pre-trained Language Models (PLMs) to transmit the deep representational strengths of PLM-derived embedding to LLMs. SALT derives unique regression lines based on the similarity in the overlap of the source and target vocabularies to handle each non-overlapping token's embedding space. Our extensive experiments show that SALT significantly outperforms other transfer methods, achieving lower loss and faster convergence during language adaptation. Notably, SALT achieves remarkable performance in cross-lingual understanding setups compared to other methods. Furthermore, we highlight the scalable use of PLMs to enhance the functionality of contemporary LLMs by conducting experiments with varying architectures.

1 Introduction

As Large Language Models (LLMs) continue to demonstrate remarkable performance across various knowledge-based benchmarks and sub-tasks, the demand for robust linguistic capabilities in multilingual or language-specific contexts has surged (Wei et al., 2022a,b; Peng et al., 2023; Taori et al., 2023; Zhou et al., 2024). However, most multilingual-considered LLMs show limited practicality in target languages due to their English-centric training and reliance on common vocabulary designed to incorporate a wide range of languages (Puttaparthi et al., 2023; Le Scao et al., 2023; Lai et al., 2023a; Zhao et al., 2024; Team et al., 2024; Dubey et al., 2024). This drawback also correlates with the large portion of their parameters, which are the embeddings of irrelevant language tokens to target language.

To address this challenge, a promising line of studies has focused on cross-lingual transfer (Artetxe et al., 2020; Tran, 2020; Gee et al., 2022; Dobler and De Melo, 2023; Remy et al., 2024; Liu et al., 2024; Mundra et al., 2024; Ye et al., 2024). These efforts include replacing a source vocabulary with a target language vocabulary, simply initializing embeddings with newly manipulated embeddings stem from source model (Minixhofer et al., 2022; Dobler and De Melo, 2023; Liu et al., 2024; Mundra et al., 2024).

However, a newly initialized embedding that relies solely on the source model inevitably experiences limited expressiveness in the target language due to its immature learning process and less consideration of the target language. Moreover, the majority of approaches adopt small encoder architectures such as BERT-like models (Devlin et al., 2019; Liu, 2019; Conneau et al., 2020) as a source model, leaving their applicability on contemporary decoder-based LLMs unexplored (Gogoulou et al., 2022; Gee et al., 2022; Zeng et al., 2023; Dobler and De Melo, 2023; Ye et al., 2024).

In this work, we propose **S**emantic **A**ware **L**inear **T**ransfer (**SALT**), a novel cross-lingual transfer method that leverages the rich representational power of target language embeddings in conventional Pre-trained Language Models (PLMs)[1] to convert English-centric LLMs into target language–specialized LLMs. To facilitate the transfer,

[†]Corresponding author

[1]In this work, we define PLMs to small-scale language models with only millions of parameters, developed prior to the advent of LLMs.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16180–16193

we identify the semantically closest tokens for each non-shared token from a shared vocabulary space. Based on these similar tokens, we fit a unique linear regression to transfer the embeddings from the PLM's space to the LLM's space. This procedure enables non-shared vocabulary to be transferred to the LLM space while retaining the semantic representation embedded in the PLM embedding.

In our experiments, we focus on investigating the potential of transferred embeddings from PLMs in three aspects: (i) whether the transferred embeddings can be well-aligned with inherent knowledge in the source model, (ii) the influence on better initialization to the target language and convergence during continual pre-training, and (iii) the cross-lingual understanding ability between target language and mainstream language.

We empirically demonstrate that embeddings from target language PLMs can be more helpful for cross-lingual transfer in LLMs than existing methods. SALT not only surpasses various strong baselines in downstream tasks but also provides a better initialization for target language adaptation with faster convergence during continual pre-training. Notably, we find that SALT preserves English capability more effectively while maintaining alignment between English and the target language in a cross-lingual setup. Furthermore, by conducting additional studies on various PLM architectures (Encoder, Decoder, and Encoder-Decoder), we discover that SALT can be extended to PLMs with various architectures and may become a valid strategy for basic models such as BERT (Devlin et al., 2019). Our contributions are as follows:

- We propose SALT, a novel embedding transfer method to effectively project the deep representation capabilities of PLM embeddings onto LLMs based on semantic information.
- We empirically verify that SALT has superiority in cross-lingual environments as well as downstream tasks and language modeling for the target language.
- We further show the versatility of SALT in experiments with various types of PLMs.
- By recycling previously prominent PLMs as target language embeddings, we demonstrate the potential scalability and propose a new approach to leverage PLMs in the era of LLMs.

2 Related Work

Cross-lingual transfer to a target language includes approaches that manipulate the vocabulary or initialize embedding for the new target vocabulary. In vocabulary manipulation, most research has considered adding target language relevant vocabulary (Wang et al., 2020; Chau et al., 2020; Cui et al., 2023; Larcher et al., 2023; Fujii et al., 2024; Mundra et al., 2024; Yamaguchi et al., 2024b; Zhao et al., 2024). Vocabulary expansion is widely adopted for developing language-specific models from a source model (Balachandran, 2023; Cui et al., 2023; Fujii et al., 2024). This line of work requires access to a large amount of target language data. In light of this, Yamaguchi et al. (2024a) delve into cross-lingual vocabulary expansion in low-resource settings across initialization approaches and training strategies. Also, Zhao et al. (2024) investigate training scales required for vocab extension and the influence of transfer on capabilities of language generation and following instructions.

On the other hand, given that overall cost increase derived from extended vocabulary, numerous studies have pursued replacing the original vocabulary with a new target vocabulary (Zeng et al., 2023; Ostendorff and Rehm, 2023; Dobler and De Melo, 2023; Ye et al., 2024; Liu et al., 2024; Remy et al., 2024). They focus on initializing new target embeddings by manipulating source embeddings according to semantic similarities between vocabularies. For instance, Gee et al. (2022) compresses the source model's vocabulary using an averaging-based technique to accommodate domain-specific terms.

More recently, lines of work use source embedding and well-aligned external word embeddings to initialize new subword embeddings for a target-specialized vocabulary (Minixhofer et al., 2022; Ostendorff and Rehm, 2023; Ye et al., 2024). Notably, FOCUS (Dobler and De Melo, 2023) computes a weighted mean from a multilingual language model, guided by token similarities from fastText (Bojanowski et al., 2017), to obtain new embeddings. Similarly, OFA (Liu et al., 2024) adopts a comparable strategy to build multilingual models and proposes a factorization–based dimension reduction embedding transfer method for efficiency. As a hybrid method, Remy et al. (2024) combines the source model's embedding under a statistical translation scheme across languages.

However, initializing embedding relying on the

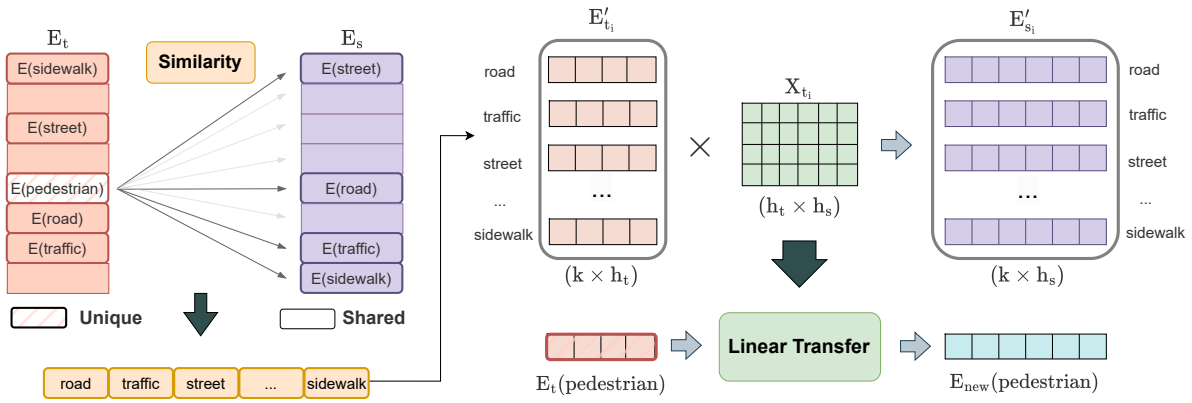

Figure 1: Summary of SALT. By using the paired embeddings of semantically similar tokens for each non-shared token, we create a unique least square matrix X_{t_i} to transfer from PLM to LLM. k denotes the number of selected nearest tokens among overlapping tokens for linear transfer, and h_t and h_s refer to the hidden dimensions of the target and source models.

source model's embedding inherently lacks crucial information specialized for the target language since the source embedding does not prioritize target language adaptation. Given this concern, we adopt embeddings from PLMs dedicated to the target language to convey richer semantic features that the source model fails to capture during the transfer. Based on this approach, we propose a new cross-lingual transfer method that recycles target language PLMs to elicit LLMs for better adaptation in target languages.

3 SALT

Our design is based on the assumption that embeddings from target language-specialized PLMs, mainly pre-trained on a target language corpus, may have richer semantic information than those from LLMs trained in a multilingual context with imbalanced language consideration. We explore the way of transferring embeddings from the PLM to the source LLM and aim to enhance generalization on various tasks within the target language. We show the summary of SALT framework in Figure 1 and describe the process step by step as follows:

Step 0: Objective Given source LLM (vocabulary embedding $\boldsymbol{E}_s$) with its vocabulary V_s and target language PLM (vocabulary embedding $\boldsymbol{E}_t$) with its vocabulary V_t, we want to make newly initialized target language embedding for V_t to replace source LLM's embedding while minimizing contextual misalignment and maintaining semantically rich information inherent in $\boldsymbol{E}_t$.

Step 1: Subword Embedding Extraction For each V_s and V_t, we use external static embeddings from fastText (Bojanowski et al., 2017) to extract auxiliary embeddings. We select fastText trained in each target language, as it can provide embeddings of various words using combinations of multiple subwords. For rare tokens that do not exist in fastText or cannot be formed via subword combinations, we initialize them from a normal distribution with mean and standard deviation from $\boldsymbol{E}_s$.

Step 2: Estimating Similarity between Shared and Non-shared Vocabulary As a next step, we identify the shared vocabulary set between the V_s and V_t. We copy the shared vocabulary embedding from $\boldsymbol{E}_s$. We provide further details about token overlap in Appendix A. We assume that overlapping tokens are not specific to the target language and have likely been sufficiently learned during the source model's training stage, which is also done by previous works (Dobler and De Melo, 2023; Liu et al., 2024). Hence, our focus is on the non-shared vocabulary set.

The most challenging task is to transfer the remaining non-shared vocabulary embeddings from $\boldsymbol{E}_t$ into the source model's space while preserving semantic components. We find semantically similar tokens among shared vocabulary, $V_{shared} = V_s \cap V_t$, for each non-overlapping token using extracted auxiliary embeddings. Then we calculate cosine similarity scores to identify the semantically nearest token in V_{shared} for each v_{t_i}. We denote similarity score as in Equation 1, where f is a fastText and $v_o \in V_{shared}$:

$$sim(v_{t_i}, v_o) = \frac{f(v_{t_i}) \cdot f(v_o)}{\|f(v_{t_i})\| \|f(v_o)\|} \quad (1)$$

Then we can get semantic similarity set C_{t_i} for

each $v_{t_i} \notin V_{\text{shared}}$:

$$C_{t_i} = \{sim(v_{t_i}, v_o) \,|\, \forall v_o \in V_{shared}\} \quad (2)$$

Step 3: Building Nearest Vocabulary Set From C_{t_i}, we identify the top k nearest tokens. To determine the criteria of k, we use Sparsemax (Martins and Astudillo, 2016). Sparsemax is a variant of the softmax function that eliminates less relevant elements by assigning them values of zero, which has been adopted in prior studies (Tran, 2020; Dobler and De Melo, 2023).

Sparsemax on C_{t_i} assigns weights to each paired token's similarity score. Most tokens assigned zero according to the similarity distribution in C_{t_i} are excluded. Through this process, we extract dynamic top k nearest vocabulary subsets to build a linear regression that transcribes non-shared tokens.

Step 4: Linear Least Square Transform Our primary goal is to transform non-shared vocabulary embeddings into the source model space by leveraging semantically similar shared tokens' paired embeddings. We employ the cheap and fast Least Squares Transform approach. To this end, we stack the source and target embeddings of the selected nearest tokens, which are $\boldsymbol{E}'_{t_i}$ and $\boldsymbol{E}'_{s_i}$ respectively. Note that selected tokens have embeddings from $\boldsymbol{E}_s$ and $\boldsymbol{E}_t$ since we only extract the nearest tokens from V_{shared}.

We find the transformation matrix X_{t_i} for a non-shared token v_{t_i} by using $\boldsymbol{E}'_{s_i}$ as a gold label for $\boldsymbol{E}'_{t_i}$:

$$\underset{X \in \mathbb{R}^{h_t \times h_s}}{\arg\min} \|\boldsymbol{E}'_{t_i} X_{t_i} - \boldsymbol{E}'_{s_i}\| \quad (3)$$

where h_s and h_t are the dimension of source and target model. The solution for each X_{t_i} can be derived as $X_{t_i} = \boldsymbol{E}'^{+}_{t_i} \cdot \boldsymbol{E}'_{s_i}$ where $\boldsymbol{E}'^{+}_{t_i}$ is a pseudo-inverse of $\boldsymbol{E}'_{t_i}$ (Peters and Wilkinson, 1970).

Using X_{t_i}, we project each non-shared vocabulary embedding from $\boldsymbol{E}_t$ into $\boldsymbol{E}_s$ space. As we consider the similarity between tokens in fitting individual least square matrix for each target token, this semantic-aware approach enables PLM's embedding to maintain the richness of representation in the target language during the transfer and facilitates better alignment with the source model.

Step 5: Target Language Adaptation As a final step, we perform additional training with an unlabeled target language corpus, called Language Adaptive Continual Pre-training, to align the weights between the transferred embedding and source model layers in the target language. This stage is crucial after the initialization or transformation of the embedding since the new embedding lacks alignment with the upper layers (Chau et al., 2020; Dobler and De Melo, 2023; Liu et al., 2024; Mundra et al., 2024).

4 Experimental Setup

4.1 Models

Source Models For source models, we employ the models that consider multilingual tokens in their vocabulary and have a large size of vocabulary. We employ Gemma-2b (Team et al., 2024) and XGLM-1.7b (Lin et al., 2022b), which both have vocab sizes 256k. They include vocabulary for multiple languages and consider the multilingual environment.

Target Models For the target models, we select PLMs with published references of various architectures for each language to ensure reliability, focusing exclusively on base-level models. By considering these widely used standard PLMs, we ensure the generalizability of our proposed method.

To be specific, we use PLMs with BERT (Devlin et al., 2019), GPT (Brown et al., 2020), and T5 (Raffel et al., 2020) architectures for each language to verify the scalability. Unless otherwise specified, our experiments use decoder-based PLMs (GPT) as the target model. Further details about selected target models are provided in Appendix B.

4.2 Baseline

To evaluate SALT, we compare ours against multiple strong baselines which are widely adopted for language transfer.

FOCUS We establish FOCUS (Dobler and De Melo, 2023) as our strong baseline due to its effectiveness and success in language transfer. FOCUS trains a new tokenizer for the target language and builds target language-specific embeddings by weighted averaging the source model's embedding weights through semantic mapping with auxiliary embeddings. Given its superior performance compared to Wechsel (Minixhofer et al., 2022) and the standard random initialization commonly used in cross-lingual transfer tasks, we consider FOCUS a reliable baseline for our experiments.

OFA The One For All (OFA) Framework (Liu et al., 2024) uses multilingual static word vectors to inject alignment knowledge into the new subword

embeddings. With the embedding factorization approach via Singular Value Decomposition, OFA initializes the embedding by leveraging a weighted average of the source embedding.

Multivariate For a robust baseline, we copy the original source embeddings for the shared vocabulary like other methods and sample the non-shared vocabulary embeddings from a multivariate Gaussian distribution whose mean and variance come from the original source embedding. The only difference is in how we initialize the non-shared vocabulary.

To ensure a fair comparison, we adopt the same fastText [2] as an external static embedding for each language in OFA and FOCUS. Across all methods, we use the same tokenizer as the target model's tokenizer.

4.3 Language Adaptive Continual Pre-training

Following the source models' pre-training objective, we use Causal Language Modeling (CLM) on an unlabeled corpus for language adaptation. We evaluate all benchmarks after the training to verify the generalizability across target languages and report changes in training loss over steps to evaluate convergence speed.

We use the Wikipedia dump (Wikimedia), version dated 20231101, extracted from Wikipedia for training. We employ NLTK sentence tokenizer to segment the corpus into individual sentences, extracting 8M sentences for each language. Subsequently, the data is split sequentially in a 9:1 ratio into training and validation sets for model training. Hyper-parameters and training details are provided in Appendix C.

4.4 Evaluation

In our experiments, we perform the transfer on three languages with distinct language families: German, Arabic, and Vietnamese, due to the availability of diverse architectures of open-source PLMs.

Knowledge-based Benchmark We evaluate the alignment between the inherent knowledge in the model's upper layers and the newly initialized embedding using three standard knowledge-based benchmarks: ARC (Clark et al., 2018), TruthfulQA (Lin et al., 2022a), and HellaSwag (Zellers et al., 2019). Due to limited support for non-English languages, we use the multilingual version developed by Lai et al. (2023b), which includes translations in several languages. All three benchmarks follow a multiple-choice question-answering (MCQA) format, and we evaluate the model by measuring the log-likelihood of each answer candidate and selecting the most probable one. Accordingly, we report accuracy for HellaSwag and normalized accuracy for ARC and TruthfulQA.

Machine Reading Comprehension We employ the Machine Reading Comprehension (MRC) task to determine whether the transferred model can accurately understand contextual meaning and implications. We assess the ability to integrate complex information to generate proper answers to questions based on a given context. We employ the MLQA (Lewis et al., 2020), XQuAD (Artetxe et al., 2020), and Belebele (Bandarkar et al., 2024) benchmarks, each containing a context and questions. We report Exact Match (EM) and F1-score for MLQA and XQuAD, which require the model to generate answers. We report accuracy similar to the knowledge benchmark evaluation for Belebele. All models are evaluated in a three-shot setting.

Cross-lingual Question Answering We also evaluate the model's linguistic ability in a cross-lingual setup where English and the target language coexist. Among the datasets employed in our MRC evaluation, MLQA encompasses instances where the context and question are in different languages. We report the performance of transferred models under both English-target language and target language-English configurations. Through comparisons with other approaches, we assess the inner alignment between English and the target language and understanding ability.

5 Results

We present downstream task performance for knowledge-based benchmarks in Section 5.1. In Section 5.2, we show the training loss curves during language adaptive continual pre-training with each transfer method. We also report MRC performance in a monolingual and cross-lingual setup in Section 5.3. Additionally, we verify the extensibility of our approach by examining the performance when using various types of target language-specialized PLMs beyond the decoder type in Section 5.4.

[2] https://fasttext.cc/docs/en/crawl-vectors.html

Source Model	Method	German				Arabic				Vietnamese			
		ARC	HS	TQA	Avg	ARC	HS	TQA	Avg	ARC	HS	TQA	Avg
XGLM	Multivariate	24.72	29.66	26.27	26.88	**26.69**	26.37	26.78	26.61	25.64	32.61	28.41	28.89
	FOCUS	23.95	29.89	26.90	26.91	25.15	26.74	**28.72**	26.87	24.87	33.13	28.54	28.85
	OFA	24.12	30.47	26.90	27.16	24.55	28.36	27.94	26.95	24.87	32.50	28.54	28.64
	SALT (Ours)	**25.49**	**31.73**	**27.28**	**28.17**	25.24	**29.27**	28.07	**27.53**	**25.73**	**33.69**	**28.54**	**29.32**
Gemma	Multivariate	29.43	39.26	24.37	31.02	26.43	31.68	28.98	29.03	27.61	42.60	29.43	33.21
	FOCUS	26.95	38.27	24.24	29.82	25.66	29.86	28.98	28.17	27.86	42.15	28.28	32.76
	OFA	28.74	38.74	24.62	30.70	26.52	33.82	28.07	29.47	26.58	41.80	29.17	32.52
	SALT (Ours)	**30.80**	**42.58**	**24.75**	**32.71**	**27.03**	**35.90**	**28.98**	**30.64**	**30.77**	**43.77**	**29.43**	**34.66**

Table 1: Performance on knowledge-based benchmarks. HS and TQA stand for HellaSwag and TruthfulQA. We **bold** the best result in each section under the same condition.

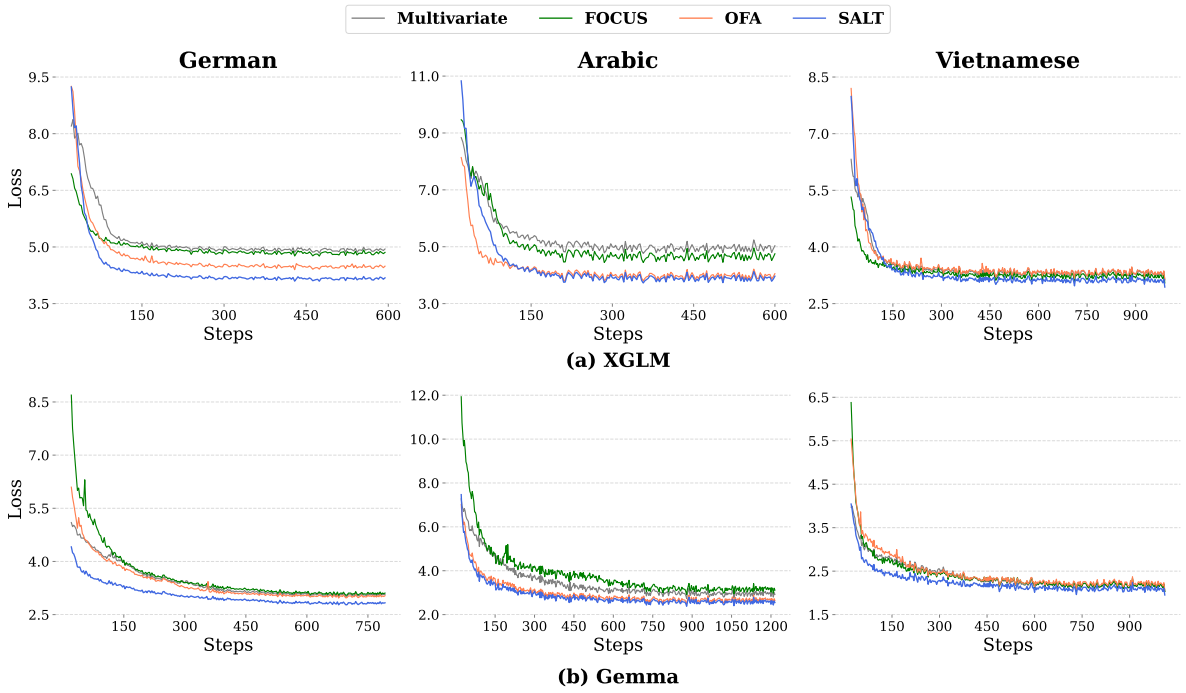

Figure 2: Causal Language Modeling (CLM) loss of language adaptive continual pre-training. The x-axis represents the progression of training steps, while the y-axis denotes the training loss. The first data point is logged at 20 steps.

5.1 Inherent Knowledge Alignment

Table 1 shows the effectiveness of SALT in aligning inherent knowledge in the source model. Regardless of the source model, SALT consistently outperforms all baselines across tasks and languages on average, delivering significant performance gains.

Notably, the major difference is evident in HellaSwag. In Gemma, SALT obtains 42.58 points on German HellaSwag, exceeding the second-best multivariate initialization by more than three points. In HellaSwag, models are required to select a sentence that logically follows a given context. This shows that SALT is superior in capturing the linguistic context and logical flow in the target language. The gains mainly arise from whether the models utilize PLM embedding trained on the target language. Since PLM embedding has already experienced the alignment with target language knowledge during pre-training, SALT further promotes alignment between the embedding layer and knowledge in the upper layers, which is also observed in Figure 2.

In contrast, FOCUS and OFA often yield similar or lower scores than the multivariate. Although both FOCUS and OFA consider semantics during the transfer, their reliance on the source model's shallow representation of the target language may harm effective alignment. This finding stresses the need for new transfer methods that differ from the weighted average of the source model's embedding.

Language	Method	MLQA			XQuAD			Belebele		
		German	Arabic	Vietnamese	German	Arabic	Vietnamese	German	Arabic	Vietnamese
Target	Multivariate	19.70 / 31.68	10.35 / 25.37	23.44 / 42.16	22.23 / 34.62	21.67 / 31.50	31.01 / 51.74	33.22	**27.89**	26.22
	FOCUS	25.66 / 37.97	12.63 / 25.63	27.17 / 44.77	27.21 / 39.00	17.09 / 25.01	35.71 / 54.65	31.00	26.67	31.33
	OFA	23.22 / 37.05	16.49 / 33.52	25.06 / 45.14	26.52 / 38.99	22.27 / 32.58	35.88 / 55.07	28.67	27.67	31.56
	SALT (Ours)	**28.07 / 41.47**	**18.33 / 36.36**	**29.70 / 49.63**	**31.28 / 44.41**	**27.07 / 39.19**	**36.30 / 56.90**	**34.11**	27.67	**35.78**
English	Multivariate	30.22 / 46.12	36.52 / 51.42	33.38 / 47.50	33.70 / 48.25	40.17 / 53.51	36.05 / 49.27	32.00	**32.56**	32.33
	FOCUS	35.66 / 50.26	38.00 / 52.57	36.82 / 50.56	40.34 / 52.43	40.25 / 53.45	41.34 / 53.06	31.00	28.44	32.00
	OFA	31.94 / 46.94	35.28 / 49.75	35.32 / 50.46	35.63 / 49.97	37.98 / 49.80	40.25 / 54.34	28.67	26.33	32.56
	SALT (Ours)	**40.02 / 55.01**	**40.24 / 54.96**	**43.64 / 58.54**	**43.78 / 57.25**	**42.10 / 54.35**	**48.24 / 61.85**	**33.67**	30.67	**35.22**

Table 2: Performance of Gemma initialized with SALT and baselines in monolingual MRC tasks. Exact Match / F1-score for MLQA and XQuAD, and accuracy for Belbele are reported.

Method	English - Target				Target - English			
	German	Arabic	Vietnamese	Avg	German	Arabic	Vietnamese	Avg
Multivariate	22.23 / 34.62	21.67 / 31.50	22.58 / 33.06	22.16 / 33.06	18.51 / 29.08	11.68 / 23.89	22.57 / 40.17	17.59 / 31.05
FOCUS	27.21 / 39.00	17.09 / 25.01	25.02 / 33.54	23.11 / 32.52	24.51 / 35.66	14.26 / 25.69	26.99 / 42.76	21.92 / 34.70
OFA	26.52 / 38.99	22.17 / 32.58	25.30 / 36.41	24.66 / 35.99	19.46 / 30.96	16.16 / 29.44	24.53 / 41.38	20.05 / 33.93
SALT (Ours)	**31.28 / 44.41**	**27.07 / 39.19**	**29.30 / 40.59**	**29.22 / 41.40**	**26.59 / 38.79**	**17.11 / 30.92**	**29.17 / 46.15**	**24.29 / 38.62**

Table 3: Results of MLQA in cross-lingual setup in English with the target language. On the top of the Table, the first language denotes the context's language, while the second language denotes the question's language.

5.2 Generalization from Continual Pre-training

To analyze how different initialization methods influence continual pre-training, we visualize the training loss of models initialized with SALT and all baselines. In Figure 2, models initialized with SALT generally converge faster and yield the lowest training loss at the last step in every scenario (Appendix D). In settings with limited computing resources, SALT may be even more beneficial. Even for Gemma, it maintains the lowest loss from the beginning to the end. Unexpectedly, FOCUS consistently exhibits higher loss in German and Arabic compared to SALT and OFA. This demonstrates that initializing embeddings from target language PLMs while considering semantic features can be superior during the adaptation period, underscoring the benefits of adopting previous PLMs for language transfer.

5.3 Effects in Language Understanding

Table 2 and Table 3 show the results in two different setups: the monolingual setup and a cross-lingual setup.

Target Language MRC At the top of Table 2, SALT shows its superiority in MRC tasks in the target language. The gap from the baselines is especially large in MLQA and XQuAD, both of which adopt generation-based evaluations. Unlike other baselines that rely on embedding from the source model, SALT may benefit from the richer semantic information specialized in the target language inherited from the PLM's embedding. This approach allows transferred models to capture contextual cues better and generate correct answers by leveraging necessary linguistic information more accurately in language comprehension tasks.

However, despite achieving performance on par with other baselines in previous results, multivariate initialization records the lowest scores on MLQA and XQuAD. An initialization method ignoring the target language's semantic information struggles to capture the deeper meanings needed for high-level reasoning based on external context.

Influence on Cross-lingual Understanding We also find that SALT is more effective in understanding the mutual context in a cross-lingual environment. In Table 3, SALT outperforms other approaches when English and the target language coexist within the same sample. In the English context with the target language question setup, SALT obtains 29.22 EM and 41.40 F1-score on average, surpassing the second-best baseline OFA by about 4 and 5 points, respectively—a 13% improvement in F1-score. We observe similar trends when the languages of the context and question are reversed. SALT outperforms FOCUS by approximately 2.4 points in EM and 4 points in F1-score.

In general, since most also consider English with the target language during the training stage, they

Source Model	Target Model	German				Arabic				Vietnamese			
		ARC	HS	TQA	Avg	ARC	HS	TQA	Avg	ARC	HS	TQA	Avg
XGLM	BERT	25.15	32.13	**28.05**	**28.44**	23.78	27.95	28.05	26.59	25.13	31.83	**30.45**	29.14
	GPT	**25.49**	31.73	27.28	28.17	**25.24**	29.27	**28.07**	**27.53**	**25.73**	**33.69**	28.54	**29.32**
	T5	24.47	**32.57**	27.92	28.32	24.38	**29.48**	27.92	27.26	25.73	33.67	28.41	29.27
Gemma	BERT	30.37	42.70	25.00	32.69	28.57	34.29	25.00	29.29	27.35	37.43	22.29	29.02
	GPT	**30.80**	42.58	24.75	32.71	27.03	35.90	**28.98**	**30.64**	**30.77**	43.77	**29.43**	**34.66**
	T5	30.71	**43.13**	**25.51**	**33.12**	**28.66**	**36.52**	25.51	30.23	30.17	**43.82**	27.77	33.92

Table 4: Performance across different target model architectures in knowledge benchmarks.

often undergo the alignment between English and the target language. By projecting this shared space to a new embedding, SALT successfully blends the source model's English ability with newly initialized embeddings for the target language, leading to notable benefits in cross-lingual understanding.

Likewise, it is evidenced by the result of the English evaluation at the bottom of Table 2. Our direct evaluation in English confirms that SALT preserves English proficiency better than other transfer methods, even though each model is further trained only on the target language. This highlights the greater effectiveness of SALT in cross-lingual environments and its utility in general situations that require both the target language and English.

5.4 Extension to Target Models with Different Architectures

To demonstrate the scalability of SALT, we compare performance using target models whose architectures differ from the source model (e.g., BERT and T5). Their embedding formations differ based on learning objectives. Our results indicate that SALT exhibits high compatibility with target models, even when their architectures differ from the source model.

As illustrated in Table 4, the decoder-based target model mostly achieves the best average results. However, employing PLMs with different structures as the target model performs better in some cases. For German, using T5 generally outperforms GPT in both source models. In addition, for German and Arabic, BERT's performance gap with GPT remains within about one point on average.

We also report the evaluation loss according to various types of target models in SALT in Table 5. In some cases, BERT achieves lower evaluation loss compared to GPT. Although learning patterns may vary depending on the underlying target model, these observations suggest that other types of PLMs can also be integrated into LLMs from a transfer standpoint. Given that various kinds of PLMs exist, SALT can be a flexible method that is not restricted to decoder-type target models and can be extended to various architectures.

Source	Target	German			Arabic			Vietnamese		
		30%	60%	90%	30%	60%	90%	30%	60%	90%
XGLM	BERT	3.89	3.82	3.80	3.76	3.25	3.24	3.49	3.40	3.40
	GPT	4.08	4.00	3.99	3.83	3.68	3.67	3.25	3.18	3.17
	T5	3.77	3.71	3.70	4.27	4.14	4.13	3.07	3.02	3.01
Gemma	BERT	2.81	2.77	2.73	2.39	2.24	2.22	2.50	2.45	2.37
	GPT	3.09	2.88	2.81	2.76	2.54	2.51	2.37	2.32	2.24
	T5	3.07	2.90	2.85	3.15	2.91	2.86	2.32	2.29	2.22

Table 5: Evaluation loss at various percentages of total steps for each target model in SALT.

6 Conclusion

We propose SALT, a novel cross-lingual transfer method that recycles the embeddings of PLMs for transferring from English-centric LLMs to target language-specialized LLMs. By deriving a unique linear regression line that takes into account the semantic similarity of each non-overlapping vocabulary embedding, we map the rich semantic features from the PLM's embedding space onto that of the LLM. A series of experiments reveals that SALT benefits from the rich expressiveness of PLM's embedding in the target language. Furthermore, SALT promotes better initialization and faster convergence for language adaptation, enhancing generalization to the target language. Notably, among existing initialization methods, SALT preserves English capabilities more effectively and demonstrates greater superiority in cross-lingual setups. We also investigate expandability by experimenting with multiple types of target models, indicating that SALT can be applied to various PLM architectures. For future work, we plan to explore methods that facilitate the transfer of language-specific knowledge embedded within model layers.

Limitations

Our methodology is based on target language-specialized PLMs, making their presence essential. Our selection of languages in this paper is based on this consideration. For our experiments in diverse types of PLMs, low-resource languages where these PLMs are unavailable are out of our scope. Given that SALT benefits from target language PLMs' rich representation ability, SALT is likely to be an even more effective method in low-resource languages if PLMs exist.

In the evaluation, we did not investigate the impact of instruction tuning. Our focus lies in leveraging foundational models for transfer and evaluating the benefits derived from PLM embeddings and the linguistic ability of the transferred model, since our approach is only related to the initialization stage. Our transfer method provides a robust starting point for embeddings, laying the groundwork for future instruction tuning to build models that perform better in the target language.

Also, we evaluate SALT only for models with large vocabulary sizes accounting for a relatively heavy portion of embedding in the model's total parameters. Given that one of the objectives of cross-lingual transfer is mitigating overload caused by irrelevant tokens to the target language and pursuing effectiveness, this aligns with the goal. In principle, SALT can also be applied to larger models (>7b), and language-specialized LLMs can be used for target models instead of PLMs. However, our focus is on the usability of the previous PLMs, and we would call for exploration in this direction.

Ethical Statement

In this study, we utilized publicly available LLMs and PLMs. These models were chosen from open sources that meet ethical standards. The static embeddings used for cross-lingual transfer and the data used in continual pre-training were sourced from publicly accessible resources and applied in accordance with appropriate licenses to align with the research objectives. Furthermore, we are aware of the potential biases that might be present in the foundation model, which could be inadvertently transferred to target languages during the language transfer process, and such biases may lead to differential performance or unintended outcomes across various languages and societal groups.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2025-R2408111) grant funded by the Korea government (MSIT). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence star fellowship support program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government (MSIT).

References

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Aragpt2: Pre-trained transformer for arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207.

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.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. 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.

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. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

dbmdz. 2021. dbmdz/german-gpt2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Konstantin Dobler and Gerard De Melo. 2023. Focus: Effective embedding initialization for monolingual specialization of multilingual models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). 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. *arXiv preprint arXiv:2404.17790*.

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.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. Cross-lingual transfer of monolingual models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

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

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. 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.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR.

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.

Nandini Mundra, Aditya Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024. An empirical comparison of vocabulary expansion and initialization approaches for language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 84–104.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

NlpHUST. 2022. Nlphust/gpt2-vietnamese.

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning. *arXiv preprint arXiv:2301.09626*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Gwen Peters and James Hardy Wilkinson. 1970. The least squares problem and pseudo-inverses. *The Computer Journal*, 13(3):309–316.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142. Association for Computational Linguistics.

Poorna Chander Reddy Puttaparthi, Soham Sanjay Deo, Hakan Gul, Yiming Tang, Weiyi Shang, and Zhe Yu. 2023. Comprehensive evaluation of chatgpt reliability through multilingual inquiries. *arXiv preprint arXiv:2312.10524*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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. *arXiv preprint arXiv:2408.04303*.

Stefan Schweter, Philip May, and Philipp Schmid. 2024. Germant5/t5-efficient-gc4-german-base-nl36.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Zihan Wang, K Karthikeyan, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual bert to low-resource languages. In *Findings of the Association*

for Computational Linguistics: EMNLP 2020, pages 2649–2656.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wikimedia. Wikimedia downloads.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024a. How can we effectively expand the vocabulary of llms with 0.01 gb of target language text? *arXiv preprint arXiv:2406.11477*.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024b. Vocabulary expansion for low-resource cross-lingual transfer. *arXiv preprint arXiv:2406.11477*.

Haotian Ye, Yihong Liu, Chunlan Ma, and Hinrich Schütze. 2024. MoSECroT: Model stitching with static word embeddings for crosslingual zero-shot transfer. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 1–7, Mexico City, Mexico. Association for Computational Linguistics.

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.

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigt, and Jie Yang. 2023. Greenplm: cross-lingual transfer of monolingual pre-trained language models at almost no cost. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6290–6298.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Further Details of SALT

Token Overlap SALT determines token overlap to transfer the target model's non-overlapping vocabulary into the source model's space. In this paper, we consider various target models with distinct tokenizers; GPT uses Byte-Level BPE, BERT uses WordPiece, and T5 uses SentencePiece. The main difference between tokenizers is how they handle whitespace tokens and non-ASCII character tokens. For example, HuggingFace's Byte-Level BPE implementation prefixes whitespace tokens with 'Ġ', SentencePiece uses '_', and BERT's WordPiece uses '##'. To avoid these discrepancies, we remove all white spaces and only consider meaningful characters in each token, which is similar to 'fuzzy search' in FOCUS (Dobler and De Melo, 2023).

Vocabulary Coverage We report the number of overlapping tokens for each language and source model, as well as the number of tokens transferred from PLMs for initialization in Table 6. Coverage refers to the proportion of the target PLM's vocabulary contained in the source vocabulary. Coverage varies according to the source model and language. Vietnamese generally shows the highest overlap, while Arabic exhibits the lowest.

Insufficient overlap potentially hinders the formation of a linear regression line conveying the capabilities of target-language specialized PLMs. Consequently, comparatively lower performance in Arabic than in other languages in our experiments may stem from the low coverage rate of overlapping tokens.

Implementation We apply the same mechanism to both the embedding layer and the LM head layer. Accordingly, both the embedding layer and the head layer are initialized with the vocabulary of the target language model. If the LM embedding layer has dimensions of (vocab_size × hidden_size), then the head layer adopts dimensions of (hidden_size × vocab_size), which is the transpose of the LM embedding matrix. However, this may vary depending on whether the model uses tied embedding weights. In our case, we used a model where the LM head and the embedding layer share the same weights; therefore, we initialized the new embedding matrix and then assigned its transpose to the LM head. If the weights are not shared, it would be necessary to apply SALT to the LM head weight matrix separately.

Lang	Model	Copied	Initialized	Coverage (%)
De	XGLM	25,254	24,853	50.2
	Gemma	29,369	20,852	58.4
Ar	XGLM	10,758	53,070	21.4
	Gemma	7,851	56,006	15.6
Vi	XGLM	30,992	18,736	61.7
	Gemma	34,114	15,934	67.9

Table 6: The number of tokens being initialized or copied from the original embeddings.

B Selected Target Models

We list the target PLMs used in our experiments in Table 7. We prefer PLMs with published references to ensure reliability. If no relevant paper exists, we select widely used pre-trained PLMs that have not been fine-tuned for downstream tasks, which have a large number of downloads on the HuggingFace Model Hub.

Lang	Target Model
De	- deepset/gbert-base (Chan et al., 2020) - dbmdz/german-gpt2 (dbmdz, 2021) - GermanT5/t5-efficient-gc4-german-base-nl36 (Schweter et al., 2024)
Ar	- aubmindlab/bert-base-arabertv2 (Antoun et al.) - aubmindlab/aragpt2-medium (Antoun et al., 2021) - UBC-NLP/AraT5v2-base-1024 (Elmadany et al., 2023)
Vi	- vinai/phobert-base-v2 (Nguyen and Tuan Nguyen, 2020) - NlpHUST/gpt2-vietnamese (NlpHUST, 2022) - VietAI/vit5-base (Phan et al., 2022)

Table 7: Target model (PLMs) used for our experiments.

C Training Details

We conduct all experiments using a uniform setup across all languages, employing two NVIDIA A100 GPUs, each with 80GB of memory, to perform language-adaptive continual pre-training. For Gemma, we set a maximum length of 1024 and utilized a batch size of 32 with gradient accumulation step to 4, resulting in a global batch size of 512 (32 x 4 x 4) and trained single epoch. The learning rate is established at 1e-5, using a cosine scheduler and a warmup ratio of 0.05. All other hyper-parameters remain consistent across the models. We adopt the Pytorch framework (Paszke et al., 2019) and the Huggingface Transformers library (Wolf et al., 2020). For knowledge benchmark evaluation, we utilize the Language Model Evaluation Harness framework [3] for evaluation.

D Training loss

We report the training loss at the final step in Table 8. As mentioned in the paper, SALT records the lowest training loss in the last step.

Model	Method	German	Arabic	Vietnamese
XGLM	Multivariate	4.9465	5.0360	3.1243
	FOCUS	4.8564	4.7522	3.0472
	OFA	4.4922	4.0532	3.1390
	SALT	**4.1823**	**3.9648**	**2.9373**
Gemma	Multivariate	3.0726	2.9218	2.0330
	FOCUS	3.1088	3.1154	2.0319
	OFA	3.0326	2.6474	2.0671
	SALT	**2.8334**	**2.5489**	**1.9509**

Table 8: The training loss at the final step.

E Length & Parameter Footprint Analysis

We also report the efficiency benefits of language transfer from a practical standpoint. Table 9 presents a comparison of tokenized lengths for randomly extracted 100,000 samples per language from the training dataset.

The result shows a notable reduction in tokenized length for the identical sentences when employing SALT compared to the original LLMs. For Arabic, the decrease in Gemma is about 25%, highlighting the lack of consideration for language-specific vocabulary in the original LLM. This reduction is closely related to computational efficiency. The generalized tokenization strategy of original LLMs, which neglects language-specific traits, can lead to a considerable computational burden during training or inference.

Similarly, SALT results in a substantial decrease of parameters: 24% in XGLM and 17% in Gemma. The reduction is due to a decrease in embedding parameters by eliminating vocabulary that is unnecessary in the target language within LLMs. SALT reduces unnecessary loads for the target language and enables the use of a language-specific tokenization strategy that accounts for the unique characteristics

[3] https://github.com/EleutherAI/lm-evaluation-harness

Model	Language		
	German	Arabic	Vietnamese
	Avg. Tokenized Length		
XGLM-1.7b	30.95	44.47	40.54
Gemma-2b	31.62	51.18	43.33
SALT	**28.21** (9% / 11%)↓	**38.61** (13% / 25%)↓	**38.64** (5% / 11%)↓
	# of parameters		
SALT	**1.31B / 2.08B** (24% / 17%)↓	**1.34B / 2.11B** (23% / 16%)↓	**1.31B / 2.08B** (24% / 17%)↓

Table 9: Reduction in average tokenized length and the number of total parameters according to original LLMs and after SALT. We report the percentage reduction compared to the original LLMs with 'XGLM / Gemma'.

of the target language derived from a considerable amount of monolingual corpus.