KDA: Knowledge Distillation Adapter for Cross-Lingual Transfer

Ta-Bao Nguyen, Nguyen-Phuong Phan, Huy Tien Nguyen, Tung Le
Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
Vietnam National University, Ho Chi Minh city, Vietnam
{21120205, 21120312}@student.hcmus.edu.vn, {ntienhuy, lttung}@fit.hcmus.edu.vn
Corresponding author: Tung Le – lttung@fit.hcmus.edu.vn

Abstract

State-of-the-art cross-lingual transfer often relies on massive multilingual models, but their prohibitive size and computational cost limit their practicality for low-resource languages. An alternative is to adapt powerful, task-specialized monolingual models, but this presents challenges in bridging the vocabulary and structural gaps between languages. To address this, we propose **KDA**, a **K**nowledge **D**istillation **A**dapter framework that efficiently adapts a fine-tuned, high-resource monolingual model to a low-resource target language. KDA utilizes knowledge distillation to transfer the source model's task-solving capabilities to the target language in a parameter-efficient manner. In addition, we introduce a novel adapter architecture that integrates source-language token embeddings while learning new positional embeddings, directly mitigating cross-lingual representational mismatches. Our empirical results on zero-shot transfer for Vietnamese Sentiment Analysis demonstrate that KDA significantly outperforms existing methods, offering a new, effective, and computationally efficient pathway for cross-lingual transfer. To facilitate reproducibility and future research, we release the adapter weights on Hugging Face[1].

1 Introduction

Cross-lingual transfer (CLT) is a critical subfield of Natural Language Processing (NLP) dedicated to leveraging knowledge from high-resource languages, typically English, to perform tasks in low-resource languages. The primary goal is to circumvent the expensive data annotation process required for each new language. The dominant and most successful paradigm for CLT has been large-scale multilingual pre-training. Although these models naturally develop some degree of unified multilingual representations (Pires et al., 2019; Conneau et al., 2020; Muller et al., 2021), a dedicated line of work has focused on further adapting them to languages with different scripts or morphological structures not well-represented in the shared vocabulary, using methods like language-specific adapters (Pfeiffer et al., 2020; Parović et al., 2022; Zhao et al., 2025; Borchert et al., 2025). Despite their effectiveness, these approaches are all constrained by a core limitation of the multilingual backbone: their massive parameter count leads to prohibitive computational costs, creating a substantial barrier for many researchers and practitioners.

These challenges motivate the exploration of more flexible, resource-efficient alternatives, leading to a compelling research question: Can we achieve effective CLT without relying on a massive, pre-trained multilingual model? A promising avenue is to adapt high-performing, readily available monolingual models. Prevailing approaches in this area include Vocabulary Adaptation, which modifies a model to use a new vocabulary (Liu et al., 2024; Han et al., 2024; Minixhofer et al., 2024; Remy et al., 2024), and representation alignment methods like Monolingual Embedding Transfer (Artetxe et al., 2020b; Minixhofer et al., 2022; Zeng et al., 2023; Liu and Niehues, 2025). However, these methods share a common shortcoming: they are not directly optimized for the final task in the target language. Instead, the adaptation phase optimizes a general objective such as masked language modeling, lexicon mapping, or an auxiliary alignment loss. Even when a downstream task loss is used (Liu and Niehues, 2025), direct supervision is only applied on the source language data. Consequently, these methods primarily endow the model with a general cross-lingual ability, rather than tailoring it for optimal performance on a specific end task.

To address this gap, we propose KDA, a novel Knowledge Distillation Adapter framework for direct, task-specific cross-lingual transfer.

[1]The adapter weights are publicly available at https://huggingface.co/haiimphuong/kda-roberta-twitter

Proceedings of the 18th International Natural Language Generation Conference, pages 122–133
October 29–November 2, 2025.

KDA transfers knowledge from a high-performing source-language teacher model to a student model that retains the same architecture but incorporates a new target-language embedding layer and a lightweight adapter, while reusing all other pretrained components. The distillation is further guided by a small parallel corpus to align cross-lingual representations effectively. As illustrated in Figure 1, only the adapter's parameters are updated during training. The adapter is optimized to align the student's output with the teacher's. Specifically, for a given source sentence fed to the teacher, the adapter learns to make the student produce an identical task-specific output distribution when given the corresponding target sentence. This approach efficiently adapts the model to the new language and task by updating only a small fraction of its total parameters.

To validate our approach, we demonstrate the effectiveness of KDA on a cross-lingual sentiment analysis task. Specifically, we transfer knowledge from a fine-tuned English sentiment model to perform sentiment analysis in Vietnamese without requiring any annotated Vietnamese data. Our experiments show that KDA outperforms both large multilingual models and recent monolingual adaptation methods. Notably, KDA achieves this superior performance while using a smaller backbone language model, highlighting the efficiency and effectiveness of optimizing directly on the downstream task in the target language.

2 Related Work

2.1 Monolingual Adaptation Methods

A widely used approach for cross-lingual adaptation is Machine Translation (MT), which involves either translating the test inputs into the source language (translate-test) or translating the source-language training data into the target language (translate-train) (Conneau et al., 2018b; Hu et al., 2020). While effective in certain settings, this strategy heavily depends on the quality of the translation system and often suffers from translation artifacts and additional computational overhead (Artetxe et al., 2020a), potentially limiting its robustness and scalability.

As an alternative, recent research has shifted toward parameter-efficient methods that avoid translation altogether by adapting model components directly for the target language. Parameter-efficient alternatives avoid these issues by modifying only a small parameter subset, typically the embedding layer, to incorporate a new language. Methodologies include retraining the embedding layer with a Masked Language Modeling (MLM) objective (Artetxe et al., 2020b), initializing new vocabularies from external resources like static embeddings or lexicons (Minixhofer et al., 2022; Zeng et al., 2023), or factorizing the embedding matrix (Liu et al., 2024). More advanced methods explicitly align token-level representations across languages using statistical translation models (Remy et al., 2024) or hyper-networks (Minixhofer et al., 2024) to generate new embeddings. Despite differing in their use of external resources, these parameter-efficient methods share the same core goal as ours: extending monolingual models to new languages with minimal architectural changes. Our approach builds on this principle while introducing a task-specific cross-lingual transfer mechanism that remains both efficient and adaptable.

2.2 Knowledge-Distillation Methods

Knowledge distillation has proven effective for cross-lingual transfer, with prior work extending it to multilingual sentence embeddings (Reimers and Gurevych, 2020) and cross-lingual information retrieval (Li et al., 2022), often relying on translated data or large unlabeled corpora. Other variations include minimal-resource approaches that use small lexicons to induce weak teachers for seed supervision (Karamanolakis et al., 2020), or adopt multi-stage pipelines that distill general cross-lingual knowledge before task-specific adaptation (Ansell et al., 2023). In contrast, our method introduces a novel, resource-minimal perspective that eliminates the need for external multilingual models, lexicons, or pre-aligned embeddings. It relies solely on the target language's embeddings and a lightweight adapter to enable direct, task-specific knowledge transfer, providing a simple yet effective solution for low-resource cross-lingual adaptation.

2.3 Adapter-Based Methods

Adapter-based frameworks enable modular cross-lingual transfer by inserting specialized modules into a multilingual model. The MAD-X framework, for example, uses separate language and task adapters (Pfeiffer et al., 2020), while subsequent work improved performance by using bilingual adapters (Parović et al., 2022) or by exposing task adapters to target-language modules during train-

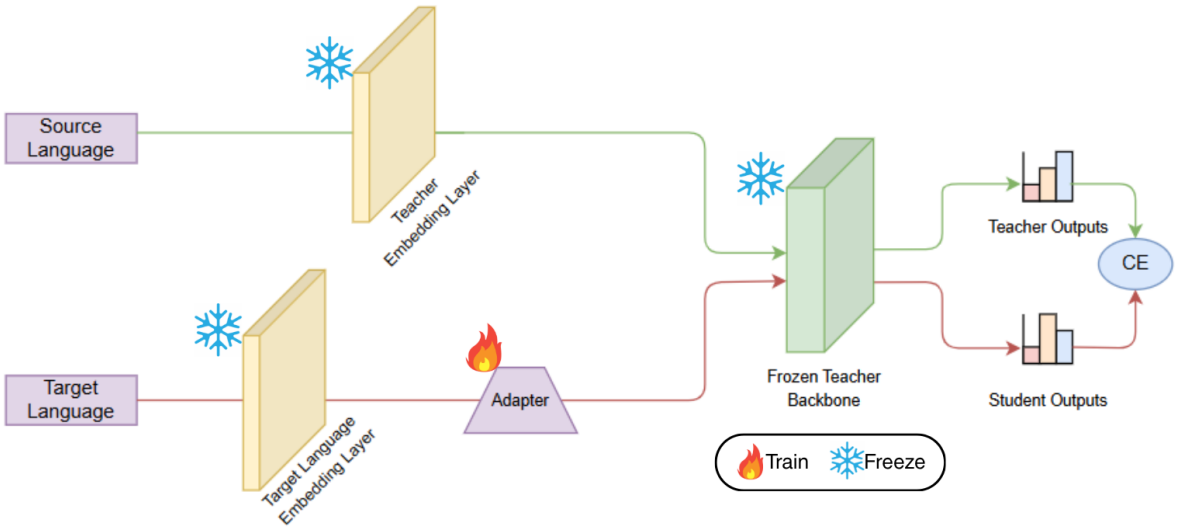

Figure 1: The KDA framework. The "teacher" model generates predictions for a source sentence. In parallel, a "student" model, which shares the teacher's backbone but has its own target-language embedding and a trainable adapter, processes the corresponding target sentence. The lightweight adapter is the sole component optimized during training, tasked with bridging the linguistic gap between target-source language.

ing to boost zero-shot capabilities (Parovic et al., 2023).

Recent advances focus on adapter composition, such as fusing language representations within LoRA bottlenecks (Borchert et al., 2025) or adaptively merging task and language adapters based on structural alignment (Zhao et al., 2025), both surpassing standard fusion baselines. Inspired by these approaches, our work introduces a new adapter architecture that bridges both vocabulary and structural inter-language gaps to create a more efficient cross-lingual pipeline.

3 Methodology

3.1 Task-Specific Distillation

This section details our framework for adapting a pre-trained, source-language model to perform a downstream task in a new target language. The primary challenge is bridging the representational gap between the two languages. This requires transforming target-language inputs into a format that the monolingual model can meaningfully comprehend, as even minor representational discrepancies can lead to a complete misinterpretation and incorrect output.

While prior work has addressed this challenge by evaluating intermediate cross-lingual alignment using metrics such as embedding similarity or representation space overlap (Conneau et al., 2018a,b; Artetxe and Schwenk, 2019; Ham and Kim, 2021), such metrics are only indirect indicators of transfer quality. In contrast, our approach sidesteps reliance on intermediate alignment and instead focuses on directly optimizing for task-specific performance in the target language.

Specifically, as illustrated in Figure 1, we propose a knowledge distillation framework to adapt a pre-trained monolingual model (referred to as the teacher) to the target language. This is accomplished using a parallel corpus of source-target sentence pairs (s_i, t_i), allowing the model to learn directly from task-specific outputs while preserving the architecture of the original model.

For each source sentence s_i, the teacher model - a language model with a conventional embedding layer and backbone - processes the input to generate a prediction distribution $\mathbf{y}_i^t$, which captures the model's task-specific knowledge in the source language. Unlike prior approaches that rely on intermediate representation alignment, our method directly distills this final output distribution. This enables the student model to learn both linguistic and task-level behavior, allowing for more precise and effective cross-lingual transfer.

Concurrently, for each corresponding target sentence t_i, the student model utilizes a pre-existing target-language embedding layer and a lightweight adapter module, sharing the frozen teacher backbone. The target sentence t_i is first embedded, then passed to the adapter. The adapter's function is to map the target-language representation

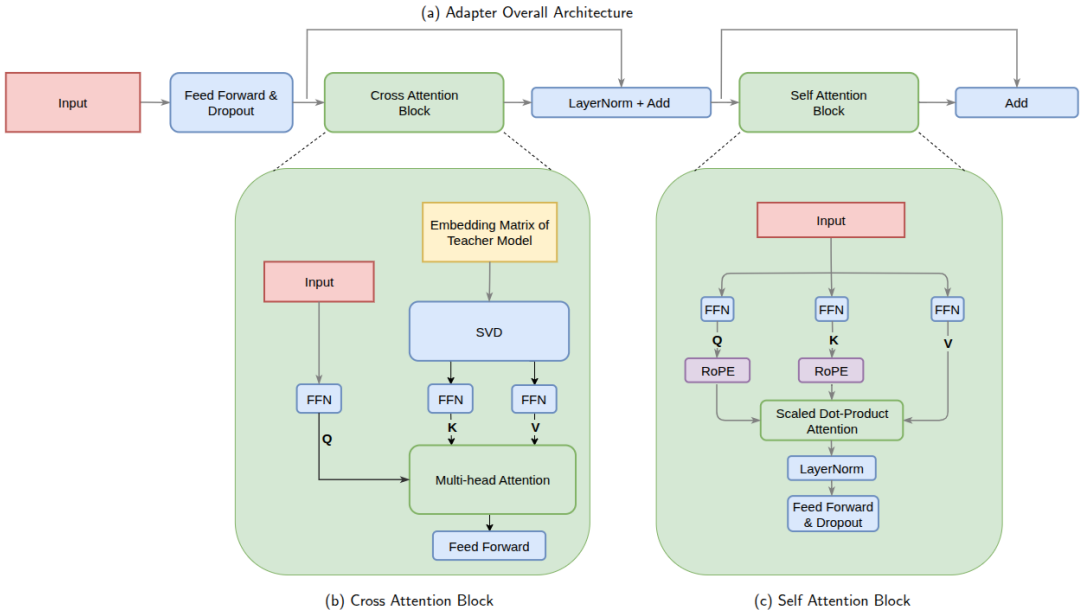

Figure 2: The KDA adapter architecture. Its function is to bridge the linguistic gap between the source and target languages via two key components: (b) a cross-attention block that integrates the teacher's token embeddings to align the target representation with the source vocabulary space, and (c) a self-attention block that injects relative positional information using Rotary Position Embeddings (RoPE).

into the latent space of the teacher's backbone. This adapted representation is then fed through the teacher frozen backbone to generate the student's prediction $\mathbf{y}_i^s$, which is trained to align with the teacher's output $\mathbf{y}_i^t$.

Training Protocol: The adapter parameters θ are optimized by minimizing the cross-entropy loss between the teacher and student output distributions in Equation 1.

$$\mathcal{L} = -\sum_{i=1}^{N} \mathbf{y}_i^t \log \mathbf{y}_i^s \tag{1}$$

While knowledge distillation often employs a combination of $\mathcal{L}_{CE}$ and Kullback-Leibler (KL) divergence loss (as in (Hinton et al., 2015)), our preliminary experiments indicated that utilizing solely $\mathcal{L}_{CE}$ led to better performance on downstream evaluation benchmarks. Therefore, we adopted $\mathcal{L}_{CE}$ as the sole optimization objective. During training, all components of the teacher model and the student's embedding layer are kept frozen. The adapter is the only trainable module and is implicitly guided to transform target-language embeddings into a latent representation compatible with the frozen teacher backbone, achieving functional alignment through end-to-end supervision.

3.2 Adapter Architecture

After the target-language token embeddings are generated, we introduce an adapter to transform these embeddings into a representation compatible with the teacher model's backbone input space. Traditionally, conventional adapter architectures used in cross-lingual transfer typically consist of a down-projection, nonlinearity, and up-projection combined with residual connections (Houlsby et al., 2019; Pfeiffer et al., 2020; Parović et al., 2022). These approaches, however, are insufficient for our specific cross-lingual transfer scenario. Firstly, it lacks information about the source language, which is essential for accurately mapping the target representation to the source representation. Secondly, it fails to explicitly model the distinct positional dependencies inherent to different languages - a critical aspect for language models.

To address these limitations, our proposed adapter architecture, illustrated in Figure 2, incorporates two key modifications. We introduce a cross-attention mechanism to dynamically integrate the teacher-model's token embedding matrices during the alignment process. Furthermore, a self-attention block, enhanced with Rotational Positional Embeddings (RoPE) (Su et al., 2024), is included to effectively encode positional informa-

tion through rotation matrices.

Specifically, the input is initially processed through a feed-forward layer and a subsequent dropout layer. The resulting tensor, denoted as x_0, then passes through a Cross-Attention block and a LayerNorm layer. A residual connection is employed around this operation, yielding an intermediate output $x_1 = x_0 + \text{LayerNorm}(\text{CrossAttentionBlock}(x_0))$. This output x_1 is subsequently processed by a Self-Attention block, where a second residual connection is applied to produce the adapter's final output, calculated as $x_1 + \text{SelfAttentionBlock}(x_1)$.

Cross-Attention Block: This module integrates token-embedding information from the source (teacher) model. The query ($\mathbf{Q}$) vector is derived from the input of the preceding layer, while the key $\mathbf{K}$ and value $\mathbf{V}$ are obtained from the teacher model's embedding matrix.

To mitigate the computational cost associated with the large teacher embedding matrix, dimensionality reduction techniques were employed. An empirical comparison between Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) was conducted, and SVD was selected due to its substantially faster compression time while maintaining comparable downstream task performance.

Consequently, the teacher embedding matrix undergoes SVD; the top r singular components are retained to form a low-rank basis. This compressed matrix is projected to produce $\mathbf{K}$ and $\mathbf{V}$. The resulting $\mathbf{K}$ and $\mathbf{V}$, along with $\mathbf{Q}$, are input to a multi-head attention mechanism, followed by a feed-forward network.

Self-Attention Block: Positional information is encoded via a self-attention mechanism augmented with Rotary Position Embedding (RoPE). The input is linearly projected into query, key, and value vectors through parallel feed-forward layers. RoPE is applied to the query and key vectors to capture positional dependencies, after which attention weights are computed. The resulting output is passed through a sequence comprising Layer Normalization, a feed-forward transformation, and Dropout to produce the final representation.

4 Experiments

4.1 Datasets

Task and Language Setting: We focus on zero-shot cross-lingual sentiment analysis, using English as the source language with labeled training data and Vietnamese as the target language without sentiment annotations. Cross-lingual transfer is performed using a parallel English-Vietnamese corpus, with no annotation in Vietnamese sentiment.

The parallel training corpus was constructed from two sources: the PhoMT dataset (Doan et al., 2021), a large-scale Vietnamese-English parallel corpus, and the Vietnamese Hate Speech Detection (VOZ-HSD) dataset. For the VOZ-HSD dataset, we utilized only the Vietnamese text and generated corresponding English translations using the DeepSeek-V3 model (DeepSeek-AI et al., 2024).

To address label imbalance, sentiment predictions were first generated on the parallel corpus using the teacher model. The data was then sampled to ensure a balanced distribution of positive, neutral, and negative classes, reducing bias during student training. Detailed statistics of the resulting dataset are shown in Table 1.

Dataset	Negative	Neutral	Positive
PhoMT	75,000	75,000	75,000
VOZ-HSD	25,000	25,000	25,000

Table 1: Sentiment label distribution of the training data. Note: These are pseudo-labels generated by the teacher model. The dataset was then sampled to mitigate potential training bias from the teacher's predictions.

To assess the effectiveness of our approach, we evaluate it on five Vietnamese sentiment analysis datasets including UIT-VSFC (Nguyen et al., 2018), ViOCD (Nguyen et al., 2021), VLSP (Nguyen et al., 2019), AIVIVN (Cocoz, 2019) and NTC-SCV (Nghia, 2020). These datasets encompass a variety of domains, text lengths, and contexts, allowing for a comprehensive assessment of our model's robustness.

Dataset	Negative	Neutral	Positive
UIT-VSFC	1,409	167	1,590
ViOCD	279	–	270
VLSP	350	350	350
AIVIVN	4,796	–	5,298
NTC-SCV	5,000	–	5,000

Table 2: Distribution of sentiment labels in the evaluation dataset.

4.2 Baselines

We compare our proposed method against several competitive approaches that fall into four main cat-

egories of cross-lingual transfer.

Machine Translation Strategy: As a competitive baseline, we adopt the translate-test approach (Ponti et al., 2021; Artetxe et al., 2023), where Vietnamese test data is translated into English using the Google Translate API. The translated inputs are then evaluated using two models: $\text{RoBERTa}_{\text{Tweet}}$ (Barbieri et al., 2020), denoted as MT^R, and $\text{GPT2}_{Twitter}$ (Pandey, 2024), denoted as MT^G. This setup enables direct comparison between our embedding-level adaptation and sentence-level translation-based methods.

Multilingual Model Fine-tuning: A foundational approach in cross-lingual transfer involves fine-tuning a massively multilingual pretrained model (MMPM) on a downstream task. In this setting, we evaluate two such baselines as comparative references.

- **XLM-R$^{\text{twitter}}$:** To establish a zero-shot baseline, we use the XLM-R model (Conneau et al., 2020), pre-trained on a 100-language CommonCrawl dataset (Wenzek et al., 2020). The model is then fine-tuned on an English-only Twitter sentiment dataset (Barbieri et al., 2022) and evaluated on the Vietnamese test set. This final step is performed in a zero-shot manner to assess its baseline cross-lingual transfer capabilities.

- **mDeBERTa$^{\text{NLI}}$:** We utilize mDeBERTa (He et al., 2020), a powerful multilingual model fine-tuned on large-scale Natural Language Inference (NLI) datasets (Laurer et al., 2024). Following a zero-shot classification setup, the model is adapted for sentiment analysis without further training. For each input Vietnamese sentence (the premise), we frame the sentiment labels (Positive, Negative, Neutral) as hypotheses and use the model to predict which hypothesis is entailed by the premise.

Adapter-based Multilingual Transfer: This category includes methods that employ adapters for cross-lingual transfer, similar in structure to our approach. The key difference is that these baselines utilize a multilingual pretrained model (MMPM) already exposed to the target language, whereas our method adapts a monolingual model. We use XLM-R (Conneau et al., 2020) as the multilingual backbone for the following approaches:

- **MAD-X$^{\text{XLM-R}}$:** Following the framework proposed by (Pfeiffer et al., 2020), we utilize a pretrained Vietnamese Language Adapter from AdapterHub. Since a suitable task adapter for 3-label sentiment analysis was unavailable, we reproduced a new one on the English Dynasent dataset (Potts et al., 2021). Zero-shot transfer is then performed by combining the Vietnamese language adapter with the English sentiment task adapter.

- **AdaMergeX$^{\text{XLM-R}}$:** As proposed by (Zhao et al., 2025), this method requires three adapters for its merging strategy. We configure its setup with: 1) an English language adapter trained on 200,000 samples from the cc-news dataset (Hamborg et al., 2017), 2) a Vietnamese language adapter trained on 200,000 samples from the cc-100 corpus (Wenzek et al., 2020), and 3) a task adapter trained on 40,000 English sentiment samples from the TweetEval benchmark (Barbieri et al., 2020).

- **FLARE$^{\text{XLM-R}}$:** We implement the FLARE framework (Borchert et al., 2025), which integrates translation components. The English sentiment fine-tuning is performed on the Dynasent (Potts et al., 2021) dataset, and the NLLB model (Team et al., 2022) is used for all translation operations.

Tokenizer Replacement: Finally, we evaluate against ZeTT (Minixhofer et al., 2024), a method that adapts a pretrained language model to a new language by replacing its tokenizer. To create a strong baseline for our Vietnamese experiments, we apply this methodology to XLM-R$^{\text{twitter}}$ (Barbieri et al., 2022), a multilingual model that has been trained on approximately 198M tweets and fine-tuned for sentiment analysis. Specifically, we replace the original tokenizer of XLM-R$^{\text{twitter}}$ with one derived from PhoBERT (Nguyen and Tuan Nguyen, 2020), a powerful monolingual BERT model for Vietnamese. We refer to this baseline as $ZeTT^{\text{XLM-R}^{\text{twitter}}}$.

4.3 KDA: Experimental Setup

Model and Components Unless otherwise specified, our KDA leverages a $\text{RoBERTa}_{\text{Tweet}}$ model as the English-language backbone (Barbieri et al., 2020). The Vietnamese embedding layer is initialized from the token embedding layer of PhoBERT, a robust monolingual model for Vietnamese. This configuration is referred to as KDA^R. The central

Parameter	Value
Architecture	
Input Embedding Dimension	768
Output Embedding Dimension	768
Intermediate FFN Dimension	768
Attention Heads (Cross and Self Attention)	8
Positional Encoding (Self-Attn)	RoPE
PFFN Activation Function	ReLU
Linear Layer Initialization	Xavier Uniform
Bias Initialization	0.0
Mapper Dropout Rate	0.1
Training	
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	128
Max Sequence Length	100
Max Epochs	20
Early Stopping Patience	3 (epochs)
Gradient Clipping Norm	1.0
Loss Function	Cross Entropy

Table 3: Hyperparameter configuration for the KDA architecture and training process.

component of our method is a lightweight adapter module trained on a parallel corpus (Table 1). Crucially, the proposed KDA framework is model-agnostic, allowing for its application to various pre-trained architectures. We demonstrate this versatility by integrating it with GPT2$_{Twitter}$ to create the KDA^G variant, with empirical results presented in Section 5. In the KDA^R configuration, the original RoBERTa embedding matrix (50257×768) is compressed to a fixed-size 768×768 representation. Both the English and Vietnamese embedding layers share a hidden size of 768, ensuring consistent dimensionality at the adapter's input and output. The adapter incorporates both self-attention and cross-attention mechanisms, each with 8 attention heads. Overall, the adapter contains approximately 6.4 million trainable parameters. A complete summary of architectural and training hyperparameters is provided in Table 3.

Training Procedure The adapter module was trained for 15 epochs with a batch size of 128. We used the AdamW optimizer with a learning rate of 1×10^{-4}.

Evaluation Strategy Accuracy and F1-score are reported across all five evaluation datasets. A key challenge lies in label set mismatch, as the backbone model produces three-way predictions (positive, negative, neutral), while some evaluation datasets are binary (positive and negative only), as shown in Table 2. To ensure consistency, the logit corresponding to the neutral class is removed during inference, and the final prediction is assigned based on the higher logit between the positive and negative classes.

5 Results and Discussion

The comprehensive performance of our KDA method in comparison to all baselines is summarized in Table 4. We discuss these findings below.

5.1 Performance of KDA in Cross-lingual Transfer

For clarity, all subsequent results for KDA are based on the RoBERTa$_{\text{Tweet}}$ backbone, unless stated otherwise. This primary configuration is labeled as KDA^R in Table 4.

KDA outperforms translation methods When compared to approaches that rely on machine translation, KDA^R demonstrates a substantial average improvement of 8% in accuracy and 7% in F1-score across the five Vietnamese sentiment analysis benchmarks. This result strongly suggests that performing adaptation directly at the embedding level is a more robust strategy than sentence-level translation. By bypassing an intermediate translation step, our method avoids the risk of propagating translation errors and better preserves the semantic nuances critical for sentiment analysis.

KDA outperforms multilingual-based methods KDA^R also establishes a new level of performance over conventional multilingual models. It achieves an average improvement of 4% in accuracy and 2% in F1-score over XLM-R^{twitter} and a more significant 6% accuracy and 5% F1-score gain over mDeBERTa^{NLI}. This outcome supports the hypothesis that large multilingual models, despite their broad language coverage, may suffer from the 'curse of multilinguality' (Wu and Dredze, 2020), where model capacity is diluted across many languages. In contrast, our approach, which specializes a strong monolingual backbone for a specific language pair, yields a more potent and task-focused representation.

KDA outperforms adapter-based methods Within the family of adapter-based methods, KDA demonstrates clear advantages. It surpasses strong baselines including MAD-X^{XLM-R} (by 8% accuracy and 7% F1), AdaMergeX^{XLM-R} (2% accuracy and 2% F1), and FLARE^{XLM-R} (2% accuracy and 4% F1). We attribute this superior performance to our adapter's architecture, which incorporates more sophisticated mechanisms for knowledge transfer. Specifically, the use of cross-attention allows for a richer integration of syntactic

Method	Accuracy						F1					
	VSFC	ViOCD	VLSP	AIVIVN	NTC-SCV	Avg	VSFC	ViOCD	VLSP	AIVIVN	NTC-SCV	Avg
Translation Methods												
MT^{G}	0.51	0.73	0.59	0.86	0.72	0.68	0.61	0.73	0.59	0.86	0.70	0.70
MT^{R}	0.61	0.77	0.60	0.87	0.75	0.72	0.69	0.77	0.60	0.87	0.74	0.73
Multilingual-based methods												
XLM-R^{twitter}	0.62	0.84	0.62	0.90	0.80	0.76	0.70	0.84	0.62	0.90	0.80	0.78
mDeBERTa^{NLI}	0.57	0.79	0.57	0.89	**0.86**	0.74	0.63	0.79	0.57	0.89	**0.86**	0.75
Adapter-based methods												
MAD-X^{XLM-R}	0.54	0.82	0.57	0.87	0.80	0.72	0.62	0.82	0.56	0.87	0.79	0.73
AdaMergeX^{XLM-R}	0.71	0.84	0.61	0.91	0.83	0.78	**0.76**	0.84	0.58	0.91	0.83	0.78
FLARE^{XLM-R}	0.67	**0.85**	**0.66**	0.89	0.83	0.78	0.57	**0.85**	**0.66**	0.89	0.83	0.76
Tokenizer Transfer												
ZeTT$^{\text{XLM-R}^{\text{twitter}}}$	0.63	0.84	0.60	0.91	0.81	0.76	0.70	0.84	0.60	0.91	0.81	0.77
Proposed KDA methods												
KDA^{G}	0.59	0.80	0.63	**0.93**	0.83	0.76	0.64	0.80	0.63	**0.93**	0.83	0.78
KDA^{R}	**0.72**	**0.85**	0.62	**0.93**	0.84	**0.80**	**0.76**	**0.85**	0.60	**0.93**	0.84	**0.80**

Table 4: Performance comparison of KDA against baseline models on the cross-lingual transfer task, with results reported in F1 and Accuracy. The best score is in bold while the second-best is underlined. Note that our KDA framework utilizes a monolingual model, whereas all baselines (except for the translation method) are built upon larger, multilingual models. The **Avg** column shows the average performance across all 5 datasets.

	Accuracy						F1					
	VSFC	ViOCD	VLSP	AIVIVN	NTC-SCV	Avg	VSFC	ViOCD	VLSP	AIVIVN	NTC-SCV	Avg
Linear	0.66	0.79	0.59	0.87	0.75	0.73	0.65	0.78	0.58	0.87	0.76	0.73
Linear + Self-Attention Block	0.70	0.81	0.60	0.88	0.77	0.75	0.70	0.79	0.59	0.87	0.77	0.74
Linear + Cross-Attention Block	0.70	0.82	0.60	0.91	0.80	0.77	0.72	0.81	0.59	0.91	0.80	0.77
KDA^{R}	0.72	0.85	0.62	0.93	0.84	0.80	0.76	0.85	0.60	0.93	0.84	0.80

Table 5: Ablation study for KDA^{R} adapter. "Linear" indicates a single linear layer used for projection; self-attention and cross-attention blocks follow Section 3.2.

information from the English teacher model, while RoPE provides a more effective way to encode positional information.

KDA outperforms Tokenizer Transfer methods On average, across five datasets, KDA^{R} achieves 4% accuracy and 3% F1 improvement compared with $ZeTT^{\text{XLM-R}^{\text{twitter}}}$. This proves that our proposed method can leverage the capabilities of pre-trained language models, which have been trained on the target task, much more efficiently than methods that apply tokenizer transfer techniques.

Parameter Efficiency A significant practical advantage of KDA is its parameter efficiency. Our complete model consists of approximately 130 million parameters. In contrast, the multilingual backbones used by many competing methods, such as XLM-R and mDeBERTa are substantially larger at 279 million parameters. KDA^{R} achieves superior performance while utilizing less than half the parameters of these large models. This computational efficiency is a crucial benefit, particularly for deployment in low-resource environments.

5.2 Robustness of KDA Across Backbone Variants

To demonstrate that the KDA framework is model-agnostic, we employed another strong pre-trained sentiment model, GPT2$_{Twitter}$ as the foundation for our method. We denote this variant as KDA^{G}, In this new setup, KDA^{G} continues to perform exceptionally well, particularly when compared against a translation baseline that also leverages the GPT2$_{Twitter}$ (MT^{G}). Our method, KDA^{G}, achieved a significant improvement of 8% in both accuracy and F1-score over the translation-based approach. This consistent outperformance with a different underlying architecture suggests that the benefits of the KDA framework are not tied to a specific pre-trained model.

5.3 Ablation Study

We ablate the adapter design in KDA^{R} by progressively adding a self-attention block and a cross-

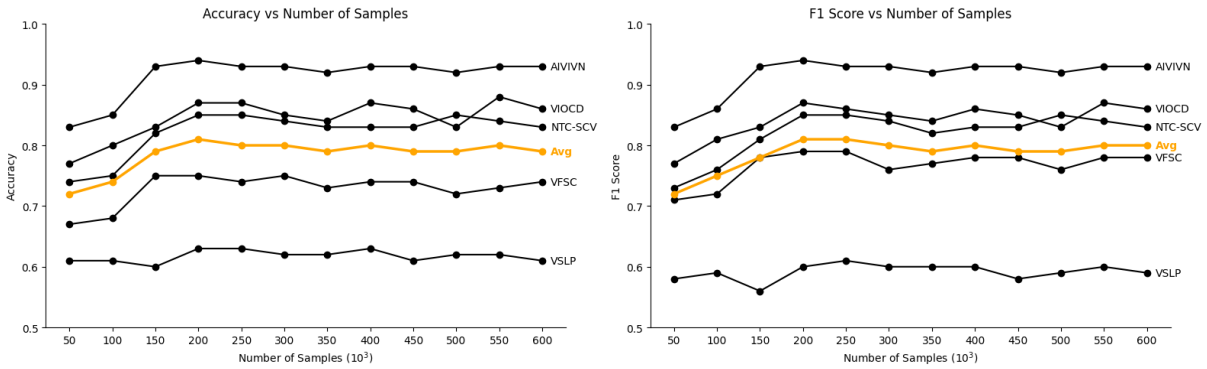

Figure 3: Effect of training-set size on KDA^R performance across five evaluation datasets. Models are trained on subsets of PhoMT (Doan et al., 2021); at each size, label balance is preserved as in Section 4.1. The Avg line shows the average performance across all 5 datasets.

attention block on top of a single linear layer (results are illustrated in Table 5). Across datasets, every component contributes. A linear-only adapter is already competitive (averaging 0.73 accuracy and 0.73 F1), making it attractive when latency or cost is the primary constraint. Adding a self-attention block yields a modest but consistent gain (+2% accuracy and +1% F1 on average vs. linear), indicating that modeling positional interactions within the adapter helps. Replacing that with a cross-attention block provides a larger boost (+4% accuracy and +4% F1 on average), highlighting the value of conditioning on the teacher model's embedding matrix. The full adapter (linear + self + cross) achieves the best results overall (+7% accuracy and +7% F1 on average vs. linear), with particularly notable improvements on AIVIVN and NTC-SCV.

5.4 Corpus Size Experiment

We examine how training corpus size affects KDA^R. Figure 3 plots performance versus the number of training examples. KDA^R learns quickly and plateaus by ∼300k samples; beyond this point, additional data yields only marginal gains. This indicates strong sample efficiency but an early saturation.

We posit three likely causes: (i) domain homogeneity-PhoMT is dominated by formal sources (Wikipedia, TED talks, news) and underrepresents informal, everyday language (e.g., slang, social media); (ii) a teacher ceiling-KDA^R may already distill most of the useful signal available from the teacher at this scale; and (iii) limitations of the current training recipe may blunt returns from larger datasets. A systematic follow-up: broadening domain coverage, evaluating stronger teachers, and revisiting scaling is left to future work.

6 Conclusion

In this work, we propose KDA, a novel and parameter-efficient framework for cross-lingual transfer that enables the use of monolingual pretrained models in new target languages without requiring large multilingual backbones or extensive cross-lingual resources. By combining a knowledge distillation process with a novel, embedding-aware adapter architecture, KDA offers a parameter-efficient pathway for adapting high-resource models to low-resource languages. Through comprehensive experiments on Vietnamese sentiment analysis, KDA demonstrates substantial improvements over multilingual fine-tuning and translation-based baselines, achieving competitive performance with a fraction of the trainable parameters. These results underscore the novelty and practicality of KDA as a scalable solution for low-resource language adaptation.

Acknowledgements

This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM) under grant number DS.C2025-18-10. The authors thank Prof. Sakriani Sakti and Prof. Hiroki Ouchi for valuable feedback. We acknowledge the Human–AI Interaction Laboratory, Nara Institute of Science and Technology, for providing computational resources during the research internships of Ta-Bao Nguyen and Nguyen-Phuong Phan.

Limitations

Our experiments are conducted primarily on models with approximately 150 million parameters, re-

flecting practical computational constraints. While this setup demonstrates the efficiency and effectiveness of KDA in resource-constrained environments, further evaluation on larger-scale models remains an important avenue for future research. Such experiments may provide deeper insights into the scalability and upper-bound performance of the framework.

Additionally, KDA is currently designed to operate in a per-language-pair setting, requiring a dedicated adapter for each source-target pair. This design introduces a trade-off between scalability and task specialization. Although less scalable than approaches that fine-tune a single multilingual model across multiple languages, KDA offers a more focused and optimized solution for specific transfer directions. This aligns with real-world scenarios where maximizing performance for a particular low-resource language is the primary goal.

References

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2023. Distilling efficient language-specific models for cross-lingual transfer.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. pages 4623–4637.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Philipp Borchert, Ivan Vulić, Marie-Francine Moens, and Jochen De Weerdt. 2025. Language fusion for parameter-efficient cross-lingual transfer. ArXiv preprint.

Minh Cocoz. 2019. Aivivn 2019 sentiment analysis dataset. Kaggle.

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. pages 8440–8451.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. ArXiv preprint; v3 revised 30-Jan-2018.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. Deepseek-v3 technical report.

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiyeon Ham and Eun-Sol Kim. 2021. Semantic alignment with calibrated similarity for multilingual sentence embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1781–1791, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. 2024. Adapters for altering llm vocabularies: What languages benefit the most? *CoRR*, abs/2410.09644. ArXiv preprint.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. Revised Jan 2021 (v2–v6).

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network.

Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2020. Cross-lingual text classification with minimal resources by transferring a sparse teacher. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3604–3622, Online. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning cross-lingual IR from an English retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. *Preprint*, arXiv:2502.14830.

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.

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.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. *arXiv preprint arXiv:2405.07883*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Nghia. 2020. Ntc-scv: Vietnamese sentiment classification dataset. GitHub.

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.

Huyen T M Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2019. Vlsp shared task: Sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24. IEEE.

Nhung Thi-Hong Nguyen, Phuong Phan-Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vietnamese complaint detection on e-commerce websites. *Preprint*, arXiv:2104.11969.

Rituraj Pandey. 2024. gpt2-sentiment-analysis-tweets. `https://huggingface.co/riturajpandey739/gpt2-sentiment-analysis-tweets`.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 176–193, Toronto, Canada. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *Preprint*, arXiv:2107.11353.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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. *CoRR*, abs/2408.04303.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. 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*, IJCAI '23.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.