Language Fusion for Parameter-Efficient Cross-lingual Transfer

Anonymous ACL submission

Abstract

Limited availability of multilingual text corpora for training language models often leads to poor performance on downstream tasks due to undertrained representation spaces for languages other than English. This 'under-representation' has motivated recent cross-lingual transfer methods to leverage the English representation space by e.g. mixing English and 'non-English' tokens at the input level or extending model parameters to accommodate new languages. However, these approaches often come at the cost of increased computational complexity. We propose **F**usion for **La**nguage **Re**presentations (FLARE) in adapters, a novel method that enhances representation quality and downstream performance for languages other than English while maintaining parameter efficiency. FLARE integrates source and target language representations within low-rank (LoRA) adapters using lightweight linear transformations, maintaining parameter efficiency while improving transfer performance. A series of experiments across representative cross-lingual natural language understanding tasks, including natural language inference, question-answering and sentiment analysis, demonstrate FLARE's effectiveness. FLARE achieves performance improvements of 4.9% for Llama 3.1 and 2.2% for Gemma 2 compared to standard LoRA fine-tuning on question-answering tasks, as measured by the exact match metric.[1]

1 Introduction

Representation degradation for 'non-English' languages poses a challenge in the context of pretrained multilingual language models (mPLMs)[2]. Large-scale English text corpora are widely available for self-supervised pretraining, resulting in superior representation quality and downstream task performance when compared to low(er)-resource languages (Lauscher et al., 2020; Yang et al., 2022). Despite the substantial improvements, the imbalance in pretraining resources still substantially reduces downstream performance (Winata et al., 2022).

Cross-lingual transfer (termed XLT henceforth) aims to narrow this performance gap by transferring task-specific knowledge acquired in high-resource languages to lower-resource languages (Ruder et al., 2019). Given the dominance of English in pretraining corpora, machine translations (MT) are frequently utilized to avoid processing non-English data (Shi et al., 2010; Artetxe et al., 2020, 2023; Ansell et al., 2023). However, translation can result in information loss, including the loss of cultural nuances, which can negatively impact downstream task performance (Conia et al., 2024). Various XLT techniques address this issue by leveraging both source and target language representation spaces, such as language mixup (Yang et al., 2022) and concatenating multilingual input sequences for in-context XLT (Kim et al., 2024; Tanwar et al., 2023; Cueva et al., 2024). These approaches, while improving XLT, typically focus on representations in a specific mPLM layer or require extensive training and computational resources by extending the input length.

Parameter-efficient fine-tuning (PEFT) methods are designed to acquire new knowledge and specialize general-purpose models for specific tasks or domains while minimizing the number of extra parameters required and keeping the large underlying mPLM frozen (Hu et al., 2022). In particular, bottleneck-style adapters, such as low-rank adapters (LoRA), extract relevant features from new data by compressing model representations with the assumption that task information

[1] Our code repository is available at https://anonymous.4open.science/r/FLARE-241E

[2] The domination of the English representation space is observed independent of model architectures, including encoder-only, decoder-only and encoder-decoder transformer (Wu and Dredze, 2020; Lee et al., 2022a; Yang et al., 2022; Wendler et al., 2024; Tang et al., 2024).

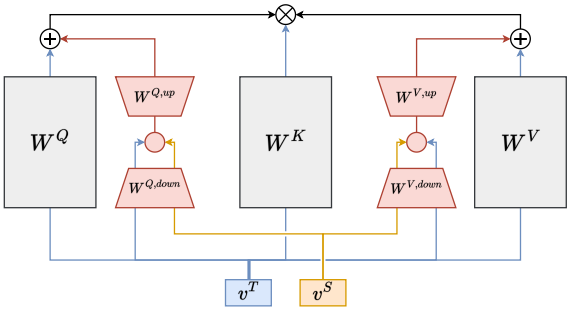

Figure 1: Fusion of source and target representations in LoRA adapters inserted within the query and value matrices. The representations are fused in the adapter bottlenecks and the outputs are added ⊕ to the query and value outputs before softmax ⊗ activation.

can be captured in a lower-dimensional space (Houlsby et al., 2019; Hu et al., 2022). This directly aligns with the XLT objectives, providing resource-efficient language and task adaptation capabilities. In XLT, adapters are widely used for acquiring task and language knowledge (Pfeiffer et al., 2020). Yet, the extent of knowledge transfer across languages within adapters remains underexplored.

In this work, we introduce **F**usion for **La**nguage **Re**presentations (FLARE), a novel approach that merges latent representations from different languages *within lower-dimensional adapter bottlenecks* to enable parameter-efficient XLT. By merging representations from high-resource languages like English into target language representations through lightweight fusion functions, such as addition or multiplication, FLARE facilitates effective cross-lingual information transfer with minimal computational overhead. As illustrated in Figure 1, FLARE performs token-wise fusion of source and target language representations within each transformer block, without adding additional parameters to LoRA and maintaining computational efficiency.

Our experiments demonstrate FLARE's effectiveness across tasks like natural language inference, sentiment classification, and question answering, using encoder-only, encoder-decoder, and decoder-only multilingual pre-trained language models (mPLMs). It is particularly beneficial for downstream tasks that involve text generation, such as question answering. For instance, FLARE improves the exact match performance for Llama 3.1 and Gemma 2 on the TyDiQA dataset by 4.9% and 2.2%, respectively. Further experiments illustrate that computational efficiency can be further enhanced by using *latent translations* as source language inputs in FLARE, and demonstrate the versatility of the method, which is orthogonal to the choice of mPLMs and MT systems.

Contributions. **1)** We introduce FLARE, a novel method that fuses language representations within adapter bottlenecks for parameter-efficient cross-lingual transfer. **2)** Our approach improves performance across diverse multilingual downstream tasks, particularly benefiting text generation tasks like question answering. **3)** We demonstrate the adaptability of our approach by incorporating machine translation encoder representations directly into the mPLM.

2 Related Work

Cross-lingual Representation Transfer. Improving performance for underrepresented languages mPLMs often involves aligning and combining latent representations from different languages (Oh et al., 2022). Several methods have been proposed to achieve this, including concatenating multilingual input sequences to leverage a shared representation space (Kim et al., 2024; Tanwar et al., 2023; Cueva et al., 2024). Another line of work focuses on projection-based methods, where target language representations are projected onto high-resource languages, such as English, to enhance feature extraction (Xu et al., 2023). Yang et al. (2022) introduced X-Mixup, which combines source and target representations in one specific mPLM layer using cross-attention during downstream task adaptation. Building on this idea, Cao et al. (2023) proposed using cross-attention with additional semantic and token-level alignment loss terms. In contrast, our FLARE method provides a more parameter-efficient approach by directly merging latent source and target language representations within adapter bottlenecks, thereby contributing to the stream of *parameter-efficient XLT*.

Representation fusion has also been applied to integrate information across different modalities, such as vision and language (Fang et al., 2021; Ramnath et al., 2021). For instance, Qu et al. (2025) used feature routing in cross-modal vision-language tasks, guiding language model representations through LoRA bottlenecks using the last hidden state of a vision model. Our work differs in its scope and fusion methodology: FLARE extracts significantly richer representations from the source and target languages by capturing layer-wise representations for each transformer block

in the mPLMs. Moreover, by ensuring dimensional alignment, we perform token-wise representation fusion within adapter bottlenecks, thereby transferring finer-grained information across languages.

PEFT in Multilingual Language Models and Cross-Lingual Transfer. PEFT aims to incorporate task or language-specific knowledge into mPLMs without updating all model weights (Pfeiffer et al., 2020). Most prominent techniques include sparse fine-tuning, which selectively updates model parameters (Ansell et al., 2022), and inserting adapter modules that reduce trainable parameters to a small fraction of total weights of the underlying mPLM (Houlsby et al., 2019). Furthermore, PEFT modules are composable, allowing for the combination of information from multiple modules (Wang et al., 2022; Lee et al., 2022b). Bottleneck adapters, such as LoRA (Hu et al., 2022) and its variants (Liu et al., 2024), are widely used for fine-tuning language models. These adapters project model representations into a lower-dimensional space, creating a bottleneck that regulates information flow (Houlsby et al., 2019). In XLT, mixtures of task and language adapters are used to merge language representation spaces effectively (Lee et al., 2022b). For instance, AdaMergeX combines the weights of adapters trained on task data in English with adapters trained on self-supervised data in the target language (Zhao et al., 2024). In contrast, our approach modifies the adapter architecture to process and combine inputs from multiple languages, enabling cross-lingual transfer of task-specific knowledge without requiring self-supervised data or additional model parameters.

3 Methodology

3.1 Language Representation Fusion

Our methodology is based on the hypothesis that incorporating English with target language representations enhances cross-lingual knowledge transfer and distills task-relevant information into the target language. We assume (MT-created) parallel corpora $\mathcal{P} = \{(x^S, x^T)\}$ during task fine-tuning, where x are instances in the respective source and target language. Our methodology particularly focuses on employing machine-translated 'silver' parallel data, akin to *translate-train* and *translate-test* settings, as we believe this approach is the most realistic in practice.

Yang et al. (2022) introduced cross-lingual manifold mixup (X-Mixup), aligning multilingual representations within a specific transformer layer using consistency loss terms and a cross-attention module. However, this method introduces additional model parameters and shows performance variability depending on the choice of the mixup layer. Another effective method for aligning multilingual representations is to concatenate source and target language input sequences $x^{S,T} = [x^S; x^T]$ where $x \in \mathbb{R}^{2m}$, with m representing the sequence length of both source and target languages. This so-called *input-level fusion* enables cross-lingual knowledge transfer across all layers of the mPLM, facilitating in-context learning, which typically does not require additional training (Cueva et al., 2024). However, this approach is computationally expensive due to increased input sequence lengths and encounters scalability issues related to the context length limitations in mPLMs.

To address these limitations, we propose FLARE, a method for *representation-level language fusion* within bottleneck adapters, as illustrated in Figure 1. Instead of extending the input, FLARE processes source and target language representations independently and fuses them only within the adapters, thus preserving computational efficiency. Source language representations v_i^S, extracted from the frozen mPLM without adapters, and target language representation v_i^T at transformer block i are down-projected using W^{down} and combined with fusion function ϕ (see Section 3.2) to create a fused representation $h = \phi\left(v_{i+1}^S W^{down}, v_i^T W^{down}\right)$, where $h \in \mathbb{R}^{m \times r}$ with sequence length m and bottleneck dimensions r. We utilize the source representation v_{i+1}^S, which has been processed by the subsequent transformer block, to leverage task-specific information extracted from the source language. Following a standard LoRA procedure, this fused low-rank representation is then up-projected and added to the frozen attention outputs v^0 to form the target language output representation $v_{i+1}^T = hW^{up} + v^0$ of the attention block.

This enhances model performance during task adaptation in the target language by directing the model's attention to task-relevant information. Thereby, the adapter bottleneck is used for cross-lingual knowledge transfer, as well as task and language adaptation. A key advantage of FLARE is the reduction in computational complexity, thereby enhancing parameter efficiency for both task and language adaptation. By processing multilingual in-

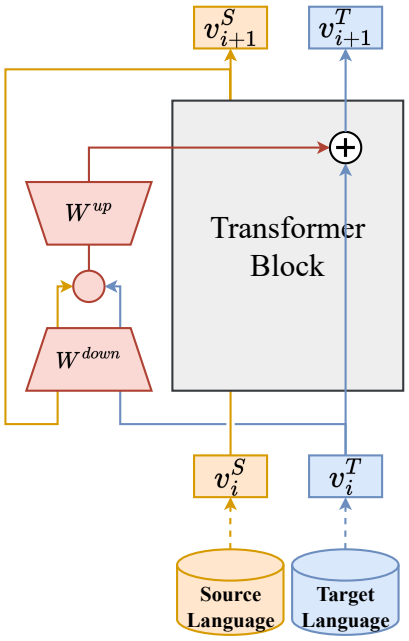

Figure 2: During the forward pass with FLARE, source language representations x^S are processed by transformer block i and before fusion with target language representations x^T. Source representations are obtained by inferencing the mPLM without the fusion adapters.

puts separately and only fusing highly compressed representations within adapter bottlenecks, our method avoids the computational overhead associated with quadratic scaling in attention computations for model dimensions d, thus enhancing resource efficiency. Furthermore, the memory requirements are limited to the last hidden states obtained from the output of each transformer block.

Moreover, our fusion approach is agnostic regarding the source language representation. We exploit this flexibility in the FLARE MT variant, which explores the impact of reducing computational resources for processing the source language on cross-lingual transfer performance. Specifically, FLARE MT utilizes representations from a MT encoder $\mathcal{M}$ as 'latent translations' that serve as source language representations. This avoids discretizing the translation as text through the MT decoder. FLARE MT further enhances resource efficiency compared to regular FLARE by bypassing the forward pass of the source language in the mPLM. We extract a single representation (latent translation) from the MT encoder by processing the target language input $v^T = \mathcal{M}\left(x^T\right)$, where $v^T \in \mathbb{R}^{m \times d_{\mathcal{M}}}$. To ensure compatibility between the dimensionality of the MT encoder outputs and the mPLM, we utilize a linear projection layer W^{proj}. This projection is jointly trained during the adaptation to the downstream task, ensuring resource efficiency. The up-projected representation $v^T W^{proj}$ is fused with the target language representation within the adapter bottlenecks of each mPLM layer, as displayed in Figure 7.

3.2 Fusion Functions

To fuse cross-lingual representations in bottleneck adapters, we evaluate both linear and non-linear transformations that do not require additional model parameters, alongside cross-attention. We extract token-wise representations from source and target language sequences, capturing rich contextual information at the token level.

The down-projected representations in the adapter bottlenecks for source and target languages are denoted as $S = v^S W^{down}$ and $T = v^T W^{down}$, where S and T are representations of dimensions $\mathbb{R}^{m \times r}$. These representations are subsequently combined at the token level through the following fusion functions:

1. element-wise addition (*add*): $S + T$
2. element-wise multiplication (*mul*): $S \circ T$
3. *cross-attention*:[3] $\text{softmax}\left(\frac{W_a^Q S \left(W_a^K T\right)'}{\sqrt{r}}\right) W_a^V T$

W_a^Q, W_a^K and W_a^V are the weight matrices of the query, key and value projections in the adapter a, respectively, and $'$ denotes the matrix transpose. We focus on lightweight linear transformations to maintain parameter and computational efficiency.

Additionally, linear fusion functions are extended with non-linear transformations through rectified linear units $ReLU\left(S\right)$ and $ReLU\left(T\right)$ (Qu et al., 2025). This allows for selective information flow in token representations, which can be particularly beneficial for multilingual input sequences that may be misaligned at the token level. By introducing non-linear transformation functions, we can restrict the propagation of misaligned information, potentially leading to improved downstream task performance.

3.3 Training

To adapt the mPLM to downstream tasks in the target language, we insert LoRA fusion adapters into the query and value weight matrices of the mPLM

[3] Although cross-attention modules add parameters to the adapters, the low bottleneck dimensions r, typically smaller than 64, minimize the parameter count in comparison to the model's internal dimensions d. Specifically, we utilize a single cross-attention head to maintain efficiency.

that has been previously fine-tuned on English task data, referred to as the *base model*. These adapters implement fusion function ϕ that combines source and target language input representations into a single fused representation. Consistent with standard PEFT training, only the task head and LoRA parameters are trainable, while all other parameters remain frozen.

During the forward pass, illustrated in Figure 2, we extract representations from both the source and target languages at each transformer block. Source language representations are obtained from the base model without fusion adapters. These layer-wise representations are stacked in matrix $V^S \in \mathbb{R}^{l \times m \times d}$, where l represents the number of layers in the mPLM. Target language representations are obtained during the forward pass through the base model with fusion adapters. In our FLARE approach, the source and target language representations are compressed to lower dimensions $r \ll h$ using the adapter's down-projection W^{down}. The compressed representations are then combined through the fusion function, and decompressed in the up-projection. By sharing the down-projection layers for both source and target language representations before fusion, we hypothesize that the model's reliance on the English representation space is reduced.

4 Experimental Setup

4.1 Underlying Models and Baselines

mPLMs. Our experiments are based on various mPLMs including the encoder-only XLM-R Large (550M) (Conneau et al., 2020), the encoder-decoder mT5-XL (3.7B) (Xue et al., 2021), the decoder-only Llama 3.1 (8B) (Grattafiori et al., 2024), and the decoder-only Gemma 2 (9B) (Gemma Team et al., 2024).

Fine-Tuning Setup. We follow a modular XLT approach where the mPLM is fine-tuned on English task data and subsequently adapted using task data in the target language (Zhao et al., 2021). For decoder-only models like Llama and Gemma, we use a causal language modeling objective for fine-tuning, and the models generate predictions as text accordingly. We employ the QLoRA fine-tuning approach with 4-bit quantization and insert LoRA adapters in all linear layers for Llama and Gemma models (Dettmers et al., 2023). Importantly, we apply representation fusion in FLARE only in the attention modules, ensuring a consistent experimental setup across different transformer architectures. The LoRA configurations use $r = 64$ and $\alpha = 128$, and the hyperparameter configurations for each model are detailed in Table 8 in the appendix.

Baselines. We evaluate FLARE against several baselines, including zero-shot cross-lingual transfer, translate-test, as well as translate-train methods such as regular LoRA fine-tuning, X-Mixup, and input-level fusion. All translate-train models are trained with the same LoRA configurations. Unless otherwise specified, FLARE models are trained using the *add+relu* fusion function, with a detailed comparison of fusion functions presented in Table 2. Model checkpoints are selected based on validation data that was machine-translated from English to the respective target languages.

X-Mixup aligns source and target language representations through cross-attention in one specific transformer layer and further aligns model outputs using consistency loss terms (Yang et al., 2022). In contrast, input-level fusion combines source and target language texts directly in the input prompt of the mPLM, doubling the sequence length (Kim et al., 2024; Cueva et al., 2024). More details on the baselines below:

Zero-Shot XLT. The base model fine-tuned on English task data is directly evaluated on test data in the target languages without further training.

Translate-Test. Test sets in each target language are translated into English using NLLB (NLLB Team et al., 2022). Subsequently, the base model is evaluated on these machine-translated test sets.[4]

Translate-Train. The base model is fine-tuned on machine-translated task data in the respective target languages. The training data comprises instances translated from English to the target language using NLLB. For fusion methods and X-Mixup, we obtain the required 'silver' parallel data also through MT (using NLLB). The training set consists of parallel sets of English and MT-ed instances, whereas the validation and test sets consist of parallel target language instances and corresponding machine translations into English. We posit that the assumed absence of gold translations both during training and during inference is the most realistic evaluation of FLARE models.

[4]Although monolingual English-only PLMs can process machine-translated text, they fail to outperform multilingual models, particularly when evaluating low-resource languages or culturally sensitive content (Ebing and Glavaš, 2024).

4.2 Evaluation Tasks and Datasets

XNLI consists of machine-translated sentence pairs that are translated from English to 15 languages (Conneau et al., 2018). The task involves determining whether a sentence entails, contradicts, or is neutral to a given premise.

NusaX is a human-annotated sentiment classification dataset that spans 11 Indonesian languages, including low-resource languages (Winata et al., 2023). With 500 labeled instances for each language, the dataset evaluates few-shot adaptation.

TyDiQA-GoldP is a human-annotated extractive QA dataset covering 8 languages (Clark et al., 2020). The task is to extract the answer spans from context passages.

Additional information on evaluation languages and datasets used for source language fine-tuning are available in Table 9 in the appendix.

4.3 Machine Translations

We utilize the NLLB 3.3B variant (NLLB Team et al., 2022) as the main MT model, employing greedy decoding to obtain translations (Artetxe et al., 2023). Additionally, FLARE MT utilizes the encoder of the NLLB 600M variant to generate latent translations. To maintain consistency in our experimental setup, we also translate languages that are not directly supported by NLLB. Specifically, Madurese (mad) and Ngaju (nij) are translated using the Indonesian language identifier, as these languages are not supported by NLLB[5] (Winata et al., 2023). For translating extractive QA datasets, we employ EasyProject (Chen et al., 2023), which involves enclosing answer spans within marker tokens prior to translation with NLLB. This method allows us to determine the position of the translated answer spans by locating these marker tokens in the translated text. Instances that fail to retain the marker tokens in the translated output are excluded from evaluation.

5 Results and Discussion

Main Results The results displayed in Table 1 confirm our hypothesis that task-specific knowledge can be efficiently transferred from English to other languages within adapter bottlenecks. Our proposed approach, FLARE, consistently surpasses all baselines across various tasks, demonstrating robust performance and validating the effectiveness of our method. It improves the average performance, averaged across metrics for all tasks, by 2.14% and 1.27% for Llama and Gemma, respectively, when compared to standard LoRA fine-tuning. The most substantial performance gains are observed on the TyDiQA dataset, particularly for text generation tasks with decoder-only models. FLARE significantly improves performance on this dataset, with the largest gains achieved on Indonesian, Russian, and Swahili. This suggests that latent representation fusion with FLARE works best for text generation when the target languages have a similar word order to the source language, in this case, subject-verb-object. However, we also observe substantial performance gains for the Llama model on Telugu, which has a different word order than English, indicating that FLARE can still achieve significant improvements even when the word order differs. The results on the XNLI and NusaX classification tasks do not exhibit a clear correlation between performance benefits for languages with subject-verb-object word order. Furthermore, the results on NusaX demonstrate that FLARE can consistently provide performance improvements for lower-resourced languages, even when only a few training data is available. This highlights the potential of FLARE to support language adaptation in low-resource settings, where data is scarce. Compared to all benchmarked models, FLARE provides consistent performance benefits, demonstrating its effectiveness in transferring knowledge from English to other languages, and its potential to improve the performance of downstream tasks in low-resource languages. Beyond performance benefits, FLARE reduces the average training time on TyDiQA by more than 40% when compared to input-level fusion.

Impact of Translation Quality.

Figure 3 presents averaged performance results for XLM-R Large with FLARE on TyDiQA and NusaX, comparing the use of different-sized machine translation (MT) models, specifically NLLB 3.3B and NLLB 600M. The results demonstrate that FLARE is robust to lower-quality machine translations. Although utilizing the larger NLLB 3.3B model yields performance improvements of 1.27% and 1.77% on NusaX and TyDiQA, respectively, the FLARE models trained on lower-quality machine translations still achieve competitive performance with standard LoRA fine-tuning based on higher-quality machine translations. This demon-

[5]We note that Toba Batak (bbc) is unsupported by NLLB and excluded from the evaluation due to translation artifacts resulting in random classification performance.

Model	XNLI	TyDiQA	NusaX	Avg.
Zero-Shot Cross-Lingual Transfer *(models are trained on English data)*				
XLM-R Large	76.95 ± 0.3	36.31 ± 2.3	75.26 ± 1.0	62.84
mT5-XL	77.92 ± 1.2	45.90 ± 0.2	74.72 ± 1.6	66.18
Llama 3.1 8B	77.40 ± 0.2	2.36 ± 0.2	71.74 ± 2.8	50.50
Gemma 2 9B	80.47 ± 0.1	2.46 ± 0.2	71.61 ± 3.4	51.51
Translate-Test *(test data is translated to English)*				
XLM-R Large	77.13 ± 0.2	41.06 ± 1.6	74.85 ± 1.0	64.35
mT5-XL	79.03 ± 0.2	47.92 ± 0.2	75.77 ± 0.3	67.57
Llama 3.1 8B	79.43 ± 0.5	2.53 ± 0.4	72.67 ± 2.4	51.54
Gemma 2 9B	79.99 ± 0.9	2.28 ± 0.2	71.61 ± 3.4	51.29
Translate-Train *(models are trained on training data translated to the target language)*				
XLM-R Large w/ LoRA	80.49 ± 1.3	40.14 ± 0.4	77.00 ± 0.8	65.88
w/ X-Mixup	79.47 ± 0.2	38.24 ± 3.2	76.37 ± 2.8	64.69
w/ input-level fusion	77.24 ± 0.8	40.45 ± 0.5	78.53 ± 0.3	65.41
w/ FLARE MT	81.60 ± 0.3	38.88 ± 1.3	77.18 ± 0.2	65.89
w/ FLARE	80.99 ± 0.9	40.93 ± 0.2	79.18 ± 1.4	**67.03**
mT5-XL w/ LoRA	79.79 ± 2.1	46.76 ± 0.7	80.41 ± 0.2	68.99
w/ X-Mixup	79.63 ± 1.0	48.23 ± 0.5	78.61 ± 0.2	68.82
w/ input-level fusion	78.81 ± 0.2	47.58 ± 0.2	80.12 ± 0.2	68.84
w/ FLARE MT	80.80 ± 1.4	48.48 ± 0.2	81.37 ± 0.8	70.22
w/ FLARE	81.00 ± 1.2	49.34 ± 0.3	80.54 ± 0.2	**70.29**
Llama 3.1 8B w/ LoRA	80.74 ± 0.4	42.84 ± 0.7	74.76 ± 1.4	66.11
w/ X-Mixup	80.22 ± 0.2	17.47 ± 1.6	75.91 ± 0.7	57.87
w/ input-level fusion	80.70 ± 0.5	46.09 ± 0.9	74.60 ± 1.6	67.13
w/ FLARE MT	80.83 ± 0.2	38.95 ± 0.2	74.52 ± 1.6	64.77
w/ FLARE	80.92 ± 0.2	47.74 ± 1.2	76.08 ± 1.1	**68.25**
Gemma 2 9B w/ LoRA	84.89 ± 0.4	49.93 ± 0.7	79.37 ± 1.2	71.40
w/ X-Mixup	84.62 ± 0.5	35.45 ± 2.0	79.94 ± 1.2	66.67
w/ input-level fusion	80.53 ± 0.2	51.29 ± 0.3	77.98 ± 1.1	69.93
w/ FLARE MT	84.84 ± 0.3	49.63 ± 0.9	78.09 ± 0.9	70.85
w/ FLARE	85.01 ± 0.4	52.14 ± 0.7	80.86 ± 0.5	**72.67**

Table 1: Average performance (with standard deviation) on natural language understanding datasets. Metrics used are: Accuracy for XNLI, Exact Match for TyDiQA, and Macro F1 for NusaX. The best-performing results for each XLT model are highlighted in **bold**.

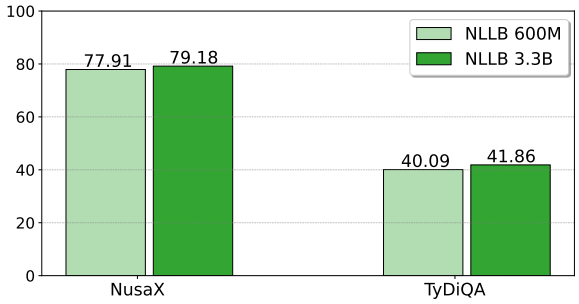

Figure 3: Average performance differences on NusaX and TyDiQA for XLM-R Large using FLARE with MT models of different size.

Fusion Function	TyDiQA	NusaX
Translate-Train (models are trained on data translated to the target language)		
add	40.76	79.56
mul	40.44	78.81
add+relu	40.93	79.18
cross-attention	39.63	78.11

Table 2: Average performance of different fusion functions using XLM-R Large with FLARE, evaluated on TyDiQA with Exact Match and on NusaX with Macro F1.

strates how FLARE can further enhance resource efficiency by effectively leveraging smaller MT models, thereby reducing computational requirements without compromising performance relative to its benchmarks.

On Latent MT Fusion.

For encoder-only models like XLM-R Large and encoder-decoder models like mT5, latent MT fusion provides notable performance benefits compared to standard LoRA fine-tuning, X-Mixup, and input-level fusion, as shown in Table 2. However, for decoder-only models like Llama and Gemma, we do not observe significant performance benefits from latent MT fusion. This suggests that reducing the computational resources for processing the source language representations in regular FLARE can negatively impact cross-lingual transfer performance, particularly for larger models. Nonetheless, it provides a resource-efficient alternative to regular FLARE for smaller mPLMs by avoiding the need for decoding in the MT and eliminating the forward pass for the source language representations.

Impact of Fusion Function.

Our study on the impact of fusion functions, presented in Table 2, shows that adding non-linearity

Model	r	TyDiQA	NusaX
Translate-Train (models are trained on training data translated to the target language)			
XLM-R Large w/ FLARE MT	8	40.86	77.84
w/ FLARE		42.37	79.52
XLM-R Large w/ FLARE MT	64	38.88	77.18
w/ FLARE		40.93	79.18
XLM-R Large w/ FLARE MT	128	40.21	77.18
w/ FLARE		40.88	78.32

Table 3: Average performance for varying adapter bottleneck size r in LoRA; based on XLM-R Large, using FLARE. Evaluation metrics include Exact Match for TyDiQA and Macro F1 for NusaX.

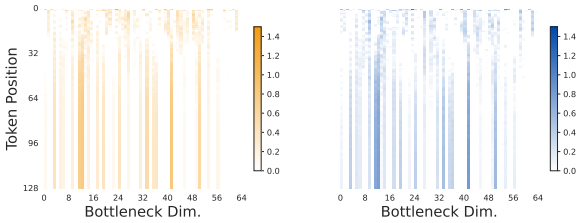

Figure 4: Average activation values for English and Acehnese in the first bottleneck query layer in XLM-R Large for the NusaX test set; *add+relu* fusion.

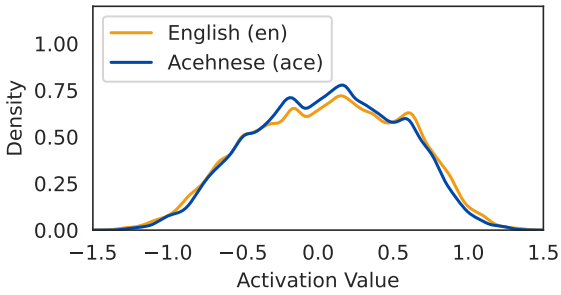

Figure 5: Average activations in the adapters across all XLM-R Large layers for the NusaX test set.

to the fusion functions does not necessarily provide decisive performance benefits over simpler linear transformations. Notably, the functions *add* and *add+relu* demonstrate the best performance. Despite the additional parameters available in cross-attention, this technique does not yield superior downstream performance, consistent with the low performance of X-Mixup in Table 1. These findings suggest that the optimal fusion function is task-dependent and can be regarded as a hyperparameter that can be fine-tuned based on validation data.

Impact of Adapter Capacity.

Our ablation study, presented in Table 3, investigates the impact of adapter capacity on FLARE's performance. The results reveal that small bottleneck sizes ($r = 8$) yield optimal performance for XLM-R Large on the TyDiQA and NusaX datasets. This finding is consistent with the observations in the original LoRA paper (Hu et al., 2022), indicating that the introduction of our fusion adapter does not affect the intrinsic rank of the tasks.

Layer-wise Language Activation.

Figure 5 shows that the magnitudes of source and target language activations across the entire XLM-R Large are comparable. This indicates that FLARE does not overly rely on either source or target representations during fusion, but instead integrates both sources of information in a balanced manner. Further, Figure 4 displays the average activations for English and Acehnese in the first adapter bottleneck: this confirms that both source and target languages maintain similar activation magnitudes. Hence, subsequent Acehnese representations are infused with the English representations from this initial transfer, integrating balanced source and target language information. Detailed activations for individual instances are illustrated in Figure 6, which show positional activation differences and demonstrate the alignment of source and target languages for information transfer.

6 Conclusion

In this paper, we introduced Fusion for Language Representations (FLARE), a parameter-efficient method for cross-lingual transfer (XLT) that enhances representation quality and downstream performance for languages other than English. Our experimental results demonstrate that FLARE consistently outperforms strong XLT baselines, including target language fine-tuning with LoRA adapters, X-Mixup, and input-level fusion, on various natural language understanding tasks. FLARE demonstrates robust performance, even for lower-quality machine translations. A key takeaway is that FLARE remains more parameter-efficient compared to benchmarked baseline approaches, while yielding superior performance. Furthermore, FLARE provides most substantial performance benefits for multilingual questions answering with decoder-only language models.

7 Limitations

Our work demonstrates that highly compressed English language representations can be effectively transferred to other languages within adapter bottlenecks. However, our experiments focus on bilin-

gual transfer settings. Extending fusion adapters to integrate multiple target languages is non-trivial, as it requires adapters to extract language-agnostic information across multiple languages.

The proposed FLARE method by design relies data availability for both source and target languages. Consequently, the application of FLARE is dependent upon the availability of machine translation models. Furthermore, our evaluation exclusively employs English as the high-resource source language for representation fusion. While English is predominantly used in mPLM pretraining corpora, exploring other high-resource languages that share linguistic similarities, with the target languages could potentially yield similar or improved cross-lingual transfer performance.

Finally, our choice of base multilingual LMs has been motivated by the current state-of-the-art (SotA) in the field of multilingual NLP and XLT to low-resource languages for NLU tasks. The main models are SotA encoder-only (XLM-R) and encoder-decoder mPLMs (mT5), and decoder-only LLMs (Llama 3, Gemma 2). However, we note that the LLM technology and its adaptation to XLT for NLU in lower-resource languages has not been proven to be fully mature yet (Lin et al., 2024; Razumovskaia et al., 2024).

References

Alan Ansell, Marinela Parović, Ivan Vulić, Anna Korhonen, and Edoardo Ponti. 2023. Unifying cross-lingual transfer across scenarios of resource scarcity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3995, Singapore. Association for Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. 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.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Tingfeng Cao, Chengyu Wang, Chuanqi Tan, Jun Huang, and Jinhui Zhu. 2023. Sharing, teaching and aligning: Knowledgeable transfer learning for cross-lingual machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 455–467, Singapore. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Emilio Cueva, Adrian Lopez Monroy, Fernando Sánchez-Vega, and Thamar Solorio. 2024. Adaptive cross-lingual text classification through in-context

one-shot demonstrations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8317–8335, Mexico City, Mexico. Association for Computational Linguistics.

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

Benedikt Ebing and Goran Glavaš. 2024. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5325–5344, Mexico City, Mexico. Association for Computational Linguistics.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12776–12784.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer,

Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai

Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2024. Cross-lingual QA: A key to unlocking in-context cross-lingual performance. In *ICML 2024 Workshop on In-Context Learning*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022a. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Jaeseong Lee, Seung-won Hwang, and Taesup Kim. 2022b. FAD-X: Fusing adapters for cross-lingual transfer to low-resource languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 57–64, Online only. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. MaLA-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jaehoon Oh, Jongwoo Ko, and Se-Young Yun. 2022. Synergy with translation artifacts for training and inference in multilingual tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6754, Abu Dhabi, United Arab Emirates. 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.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5.

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2025. Introducing routing functions to vision-language parameter-efficient fine-tuning with low-rank bottlenecks. In *Computer Vision – ECCV 2024*, pages 291–308, Cham. Springer Nature Switzerland.

Kiran Ramnath, Leda Sari, Mark Hasegawa-Johnson, and Chang Yoo. 2021. Worldly wise (WoW) - cross-lingual knowledge fusion for fact-based visual spoken-question answering. In *Proceedings of the 2021 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 1908–1919, Online. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual NLU: are we there yet? *Preprint*, arXiv:2403.01929.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6724–6743, Miami, Florida, USA. Association for Computational Linguistics.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, Cambridge, MA. Association for Computational Linguistics.

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

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. 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 5751–5767, Online. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *Preprint*, arXiv:2402.18913.

A Detailed Evaluation Results

Figure 6 displays average activations within the first adapter bottlenecks in the XLM-R Large model using FLARE and the *add+relu* fusion function. This visualization highlights the positional alignment process between English and Acehnese token representations, with varying activation values across different sequence positions reflecting the dynamics of language representation fusion.

Table 4 presents the performance of FLARE and input-level fusion when using gold translations for fusion, as opposed to machine translations generated by NLLB. The results demonstrate that input-level fusion performance is sensitive to the quality of English input provided. Notably, when gold translations are available, input-level fusion replicates English performance, indicating that it heavily relies on the quality of English inputs. In contrast, FLARE balances the fusion of source and target language information, as evident from the findings in Figure 5. While input-level fusion outperforms FLARE when gold translations are available, FLARE achieves significantly higher performance in the more realistic setting using machine-translated data.

Table 5 shows the results for the XNLI dataset for each language in zero-shot XLT, translate-test, translate-train settings, including translate-train with gold translations in the source language. The results confirm that FLARE consistently improves XTL performance in the translate-train setting across different languages without particular bias towards typological relatedness to English or frequency in pretraining corpora.

Table 6 details the results for the TyDiQA dataset for each language in the zero-shot XLT, translate-test, and translate-train settings. The outcomes demonstrate that FLARE performance extends to tasks including positional information, such as extractive question-answering.

Table 7 outlines the performance for the NusaX dataset for each language in zero-shot XLT, translate-test, translate-train, and translate-train settings with gold translations in the source language. Even with few training samples, our FLARE method demonstrates consistent performance improvements across the low-resource languages included in the NusaX dataset.

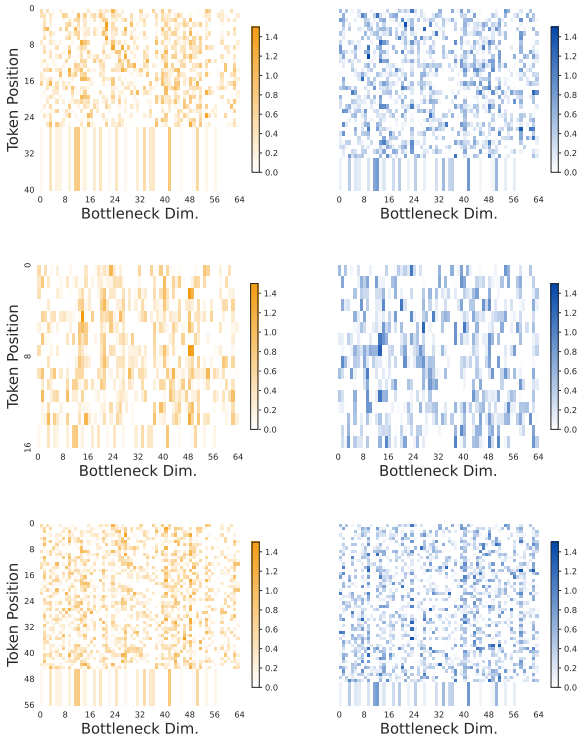

Figure 6: Activation values for individual instances included in the NusaX test set. English and Acehnese activation values are extracted from the first bottleneck query layer in XLMR-Large, which is trained with the *add+relu* fusion function.

B Training Details

Our evaluation results are averaged across *five random seeds*. Initially, we fine-tune the language models on English task data using LoRA adapters set with $r = 64$ and $\alpha = 128$, which are subsequently integrated into the model's weights prior to task fine-tuning in the target languages. Hyperparameter configurations for each mPLM are provided in Table 8.

The total computation time for the experimental results exceeds 5,000 GPU hours. All models are trained using half-precision.

C Implementation Details of FLARE MT

We introduce FLARE MT as variant of FLARE aimed at further enhancing resource efficiency. FLARE MT improves efficiency in two key ways. Firstly, it leverages latent translations generated by the machine translation (MT) encoder, thereby reducing the computational resources required to produce full text translations. Secondly, it eliminates the

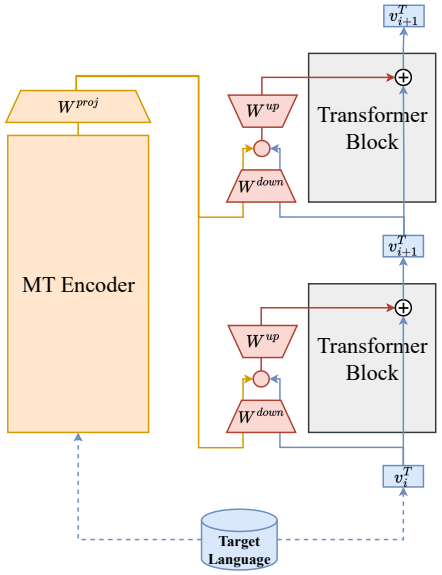

Figure 7: Illustration of the FLARE MT variant where projected encoder representations from an MT model are directly fused with target language representations within the fusion adapters in the mPLM. Encoder representations from the MT model serve as latent translations, avoiding discretization in the decoder.

need for a forward pass through the source language representations in the mPLM, resulting in significant computational savings. As a results, a single source language representation, namely the latent translation, is fused with the target language representation in the fusion adapters for each transformer layer. To enable this fusion, a projection module is introduced to align the dimensions of the MT encoder with those of the mPLM. Although this module adds additional parameters, it is essential for ensuring compatibility between the two models. Notably, related work suggests that extending the single projection layer to a MLP and training it on additional self-supervised data can yield substantial performance benefits (Liu et al., 2023; Schmidt et al., 2024). This provides a promising direction for future research and potential improvements to the FLARE MT approach.

D Practical Implications

The practical implementation of bilingual cross-lingual transfer methods, such as FLARE, requires an additional step of language identification to determine bilingual adapter for model inference. While this introduces a preprocessing stage, language identification systems are widely accessible and highly accurate. For example, NLLB achieves a 95% F1 score across 193 FLORES languages,

Model	XNLI	NusaX
*Translate-Train (fusion models are trained on data translated into the target language and evaluated using **gold translations** from the target language to the source language)*		
XLM-R Large		
w/ input-level fusion	87.19	90.93
w/ FLARE	88.15	84.66
mT5-XL		
w/ input-level fusion	89.67	90.57
w/ FLARE	86.57	80.72

Table 4: Average performance for the *translate-train* setting with gold English translations during inference across languages included in the XNLI, and NusaX datasets, representing optimal translation quality. Evaluation metrics include accuracy for XNLI and Macro F1 for NusaX.

including many low-resource languages (Burchell et al., 2023), ensuring that this step can be seamlessly integrated into real-world applications.

E Another Ablation: Representation Fusion during Training Only

To investigate the importance of utilizing source language representations during inference, we modified FLARE to restrict representation fusion to the training phase only. Specifically, we limited the fusion with source language representations to 50% of the training instances and excluded source language data during inference. This evaluates cross-lingual transfer capabilities based on instance-independent patterns learned from source language representations during training. Our findings reveal that fusion adapters struggle to learn patterns that are independent of specific instances from source language representations during training. As a result, when implemented in the XLM-R Large model on the NusaX test set, the performance of the *train-only* FLARE variant decreased by 30%. Crucially, this significant drop underscores the importance of incorporating source language representations during inference to achieve effective cross-lingual adaptation.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
Zero-Shot Cross-lingual Transfer																
XLM-R Large	87.81	76.70	81.37	80.27	80.24	82.58	81.56	73.52	78.31	65.48	76.01	76.04	69.43	77.69	78.07	76.95
mT5-XL	89.04	77.13	82.27	81.34	81.00	83.43	82.74	74.58	80.22	70.19	75.93	77.22	70.35	76.33	78.15	77.92
Llama 3.1 8B	91.47	79.56	80.56	83.38	80.73	85.06	84.07	72.07	81.66	60.64	75.38	76.51	62.64	81.39	80.01	77.40
Gemma 2 9B	93.05	80.38	83.81	84.42	83.91	85.05	85.52	77.59	82.23	73.96	76.92	79.27	71.31	81.54	80.73	80.47
Translate-Test (translate test data to English using NLLB 3.3B)																
XLM-R Large	87.81	76.52	81.46	81.31	81.03	82.37	81.57	74.47	77.80	71.96	72.65	78.24	67.87	78.13	74.43	77.13
mT5-XL	89.04	79.04	83.15	83.07	82.56	83.73	83.30	76.69	80.47	73.02	74.69	79.54	69.80	80.26	77.15	79.03
Llama 3.1 8B	91.47	78.89	83.61	84.00	82.86	85.92	83.97	76.25	80.00	73.30	73.34	79.84	69.03	79.49	76.55	79.08
Gemma 2 9B	93.05	79.76	84.55	84.99	83.98	87.04	84.82	77.01	81.24	73.75	74.04	81.31	69.24	80.57	77.58	79.99
Translate-Train (models are trained on training data translated to the target language)																
XLM-R Large w/ LoRA	87.81	79.33	84.13	82.64	83.21	84.50	83.06	77.78	81.42	74.44	79.73	80.72	74.71	81.68	79.44	80.49
w/ X-Mixup	87.81	78.33	82.48	82.16	80.12	82.57	81.23	76.06	80.54	74.11	79.48	79.15	73.67	81.48	81.18	79.47
w/ input-level fusion	87.81	77.29	81.41	81.07	81.36	82.40	81.07	74.89	77.79	72.13	72.76	78.47	68.46	78.11	74.18	77.24
w/ FLARE MT	87.81	80.91	84.76	84.12	83.97	85.03	83.80	79.04	82.07	76.71	80.32	81.70	76.29	81.88	81.76	81.60
w/ FLARE	87.81	81.04	84.12	83.35	83.44	83.95	83.53	79.26	79.58	75.58	80.40	80.15	75.50	81.30	82.71	80.99
mT5-XL w/ LoRA	89.04	79.44	83.37	83.23	81.65	84.17	83.76	76.84	81.31	75.97	76.93	77.68	73.00	79.13	79.80	79.73
w/ X-Mixup	89.04	80.14	82.21	82.37	82.73	82.87	82.54	77.16	79.87	76.10	79.03	78.00	73.42	79.97	78.43	79.63
w/ input-level fusion	89.04	79.03	82.76	82.36	82.14	83.43	82.89	76.37	80.28	72.97	75.11	79.01	69.73	79.61	77.70	78.81
w/ FLARE MT	89.04	80.41	83.73	83.45	82.91	83.81	83.57	78.44	81.35	77.02	78.62	81.13	75.46	80.49	80.75	80.80
w/ FLARE	89.04	81.32	83.72	83.46	82.23	84.87	83.47	79.00	81.28	77.43	79.54	80.50	74.27	81.30	81.66	81.00
Llama 3.1 8B w/ LoRA	91.47	80.27	82.88	84.23	83.16	86.85	85.49	79.82	83.97	67.39	77.58	79.62	76.94	81.91	80.37	80.74
w/ X-Mixup	91.47	79.58	82.94	84.11	81.61	86.23	85.39	79.15	83.16	66.49	76.51	79.06	77.20	81.44	80.18	80.22
w/ input-level fusion	91.47	79.17	85.63	85.39	83.61	87.10	86.00	77.55	81.91	74.72	74.75	82.32	71.78	82.07	77.79	80.70
w/ FLARE MT	91.47	80.27	83.17	84.87	82.95	86.75	85.67	80.35	82.70	67.19	77.41	79.95	77.30	81.92	81.15	80.83
w/ FLARE	91.47	80.00	83.09	84.92	82.90	86.55	86.04	80.80	83.14	67.08	77.21	79.33	77.95	82.33	81.55	80.92
Gemma 2 9B w/ LoRA	93.05	85.19	87.87	88.03	87.78	89.06	87.13	82.49	86.10	79.52	83.20	84.25	78.07	85.09	84.69	84.89
w/ X-Mixup	93.05	84.70	87.76	87.84	87.32	88.61	87.83	82.44	85.27	80.12	82.60	83.51	77.35	84.50	84.86	84.62
w/ input-level fusion	93.05	79.98	84.84	85.19	84.16	86.65	85.12	77.39	81.74	74.48	75.07	81.98	70.70	81.74	78.34	80.53
w/ FLARE MT	93.05	85.07	87.73	87.89	87.70	88.93	87.90	82.69	84.97	80.52	82.82	84.18	77.36	84.88	85.19	84.84
w/ FLARE	93.05	84.67	87.93	88.14	87.77	89.23	88.10	82.86	85.97	79.73	83.15	84.08	78.13	85.23	85.19	85.01
Translate-Train (fusion models are trained on data translated into the target language and evaluated using gold translations from the target language to the source language)																
XLM-R Large w/ input-level fusion	87.81	88.41	88.54	88.46	88.36	88.28	88.02	88.38	85.91	86.23	85.91	85.85	86.05	85.85	86.45	87.19
w/ FLARE	87.81	88.10	88.06	88.04	88.12	88.02	88.08	88.40	88.12	88.46	88.16	88.14	88.22	88.04	88.16	88.15
mT5-XL w/ input-level fusion	89.04	90.04	89.80	89.54	89.70	89.78	89.50	89.80	89.52	89.56	89.84	89.66	89.38	89.52	89.70	89.67
FLARE	89.04	88.62	88.74	88.80	85.34	87.83	86.19	84.31	86.12	89.66	88.49	89.56	79.22	85.33	83.73	86.57

Table 5: Average scores per language in the XNLI dataset. Model performance is evaluated using the Accuracy metric.

Model	en	ar	ben	fi	ind	ko	ru	sw	tel	Avg.
Zero-Shot Cross-lingual Transfer										
XLM-R Large	49.05	24.27	32.78	30.15	44.51	29.92	30.15	40.27	58.42	36.31
mT5-XL	55.23	30.94	43.89	35.56	49.59	41.59	41.47	50.63	73.54	45.90
Llama 3.1 8B	55.61	2.41	0.00	5.22	1.64	0.58	5.49	2.52	1.06	2.36
Gemma 2 9B	60.08	1.63	2.46	3.20	0.88	0.24	2.77	1.87	6.61	2.46
Translate-Test (translate test data to English using NLLB 3.3B)										
XLM-R Large	49.05	28.15	52.78	30.54	47.79	36.23	34.92	52.76	45.30	41.06
mT5-XL	55.23	34.01	49.00	36.08	51.97	42.61	39.87	54.95	74.89	47.92
Llama 3.1 8B	55.61	1.35	5.00	2.21	0.99	0.58	2.71	4.68	2.75	2.53
Gemma 2 9B	60.08	0.68	3.33	1.74	0.66	0.00	1.68	1.08	9.04	2.28
Translate-Train (models are trained on training data translated to the target language)										
XLM-R Large w/ LoRA	49.05	31.10	38.97	30.35	46.61	37.11	24.54	47.13	65.34	40.14
w/ X-Mixup	49.05	26.25	36.67	26.65	43.70	34.19	24.40	45.85	68.21	38.24
w/ input-level fusion	49.05	31.56	40.00	30.04	45.63	33.33	28.12	50.29	64.62	40.45
w/ FLARE MT	49.05	30.21	35.00	33.02	44.82	35.90	25.38	48.46	58.26	38.88
w/ FLARE	49.05	30.83	41.67	33.44	44.61	35.33	26.17	48.47	66.93	40.93
mT5-XL w/ LoRA	55.23	33.48	46.71	38.90	49.84	46.77	33.20	51.46	73.73	46.76
w/ X-Mixup	55.23	32.55	54.49	38.15	52.17	49.05	33.75	52.04	73.68	48.23
w/ input-level fusion	55.23	34.75	49.74	39.29	51.74	45.29	30.69	52.79	76.32	47.58
w/ FLARE MT	55.23	46.08	48.59	39.31	53.94	44.90	30.14	49.16	75.75	48.48
w/ FLARE	55.23	47.86	49.55	40.98	54.23	46.05	30.41	50.85	74.82	49.34
Llama 3.1 8B w/ LoRA	55.61	39.69	26.11	44.27	56.61	53.56	37.23	53.47	31.75	42.84
w/ X-Mixup	55.61	23.44	16.11	37.26	30.69	0.00	10.96	0.00	21.32	17.47
w/ input-level fusion	55.61	37.48	21.03	45.86	56.68	62.61	37.03	61.82	46.20	46.09
w/ FLARE MT	55.61	38.33	26.11	37.90	48.78	53.85	34.04	44.94	27.63	38.95
w/ FLARE	55.61	44.48	26.66	48.09	62.80	56.98	42.44	59.73	40.70	47.74
Gemma 2 9B w/ LoRA	60.08	43.75	46.67	44.69	57.52	59.83	38.82	59.72	48.42	49.93
w/ X-Mixup	60.08	37.71	20.00	46.92	54.07	39.32	14.33	25.71	45.54	35.45
w/ input-level fusion	60.08	45.74	50.14	40.45	59.97	61.87	41.06	61.91	49.14	51.29
w/ FLARE MT	60.08	43.23	45.00	44.05	59.45	57.83	40.05	58.82	48.60	49.63
w/ FLARE	60.08	44.79	46.67	47.35	65.24	60.68	41.82	60.75	49.83	52.14

Table 6: Average scores per language in the TyDiQA dataset. Model performance is evaluated using the Exact Match metrics.

Model	en	ace	ban	bjn	bug	ind	jav	mad	min	nij	sun	Avg.
Zero-Shot Cross-lingual Transfer												
XLM-R Large	92.04	68.34	75.37	80.37	51.90	90.76	84.69	69.01	80.06	69.23	82.89	75.26
mT5-XL	91.77	72.26	76.42	79.79	49.51	90.61	87.49	61.38	77.71	65.31	86.73	74.72
Llama 3.1 8B	89.75	70.50	72.00	80.33	39.92	89.75	77.25	64.75	77.75	65.42	79.75	71.74
Gemma 2 9B	91.15	66.42	71.58	82.08	31.92	91.67	86.25	64.00	80.75	64.33	77.08	71.61
Translate-Test (translate test data to English using NLLB 3.3B)												
XLM-R Large	92.04	73.20	73.88	82.09	60.47	88.85	84.27	61.24	81.19	59.35	83.97	74.85
mT5-XL	91.77	76.27	73.43	81.72	69.29	86.86	83.50	60.63	82.47	60.86	82.68	75.77
Llama 3.1 8B	89.75	70.83	73.00	80.75	39.92	89.58	78.00	65.25	81.58	67.33	80.42	72.67
Gemma 2 9B	91.15	66.42	71.58	82.08	31.92	91.67	86.25	64.00	80.75	64.33	77.08	71.61
Translate-Train (models are trained on training data translated to the target language)												
XLM-R Large w/ LoRA	92.04	74.19	74.55	81.84	60.99	89.40	85.90	70.75	81.15	67.35	83.87	77.00
w/ X-Mixup	92.04	73.10	73.08	81.18	62.22	88.38	85.90	65.79	82.30	68.97	82.74	76.37
input-level fusion	92.04	77.77	75.89	82.67	69.96	89.44	87.92	66.66	79.55	68.47	87.01	78.53
w/ FLARE MT	92.04	73.33	75.95	81.13	57.20	90.76	86.59	69.77	83.42	68.90	84.73	77.18
w/ FLARE	92.04	76.47	77.27	80.71	70.18	90.54	87.42	71.33	85.15	70.16	82.59	79.18
mT5-XL w/ LoRA	91.77	80.66	81.92	85.83	65.36	89.78	90.40	69.85	82.30	69.27	88.76	80.41
w/ X-Mixup	91.77	80.34	74.60	83.76	68.87	88.52	88.75	68.25	83.66	65.60	83.76	78.61
input-level fusion	91.77	81.00	79.48	85.54	71.44	89.75	87.58	66.33	83.28	68.02	88.78	80.12
w/ FLARE MT	91.77	81.19	84.12	85.19	66.59	90.14	89.67	71.16	84.80	71.87	88.94	81.37
w/ FLARE	91.77	81.03	82.03	85.88	66.95	89.55	89.80	68.63	84.20	69.31	88.05	80.54
Llama 3.1 8B w/ LoRA	89.75	76.26	73.71	78.10	62.82	88.66	84.29	62.91	82.20	58.04	80.64	74.76
w/ X-Mixup	89.75	77.25	76.58	79.00	64.17	89.92	85.08	64.58	82.17	59.83	80.50	75.91
input-level fusion	89.75	74.83	66.17	80.17	66.67	89.17	85.50	59.63	82.63	57.25	84.00	74.60
w/ FLARE MT	89.75	78.21	72.00	74.29	64.21	87.96	83.04	64.38	80.75	62.38	77.96	74.52
w/ FLARE	89.75	78.88	75.25	80.25	64.25	91.17	85.88	65.13	81.38	57.88	80.75	76.08
Gemma 2 9B w/ LoRA	91.15	77.89	79.14	82.30	66.83	91.71	87.56	69.81	85.72	67.61	85.18	79.37
w/ X-Mixup	91.15	80.94	77.92	82.82	68.22	91.46	88.29	69.41	87.64	67.26	85.48	79.94
input-level fusion	91.15	77.67	76.88	83.50	70.58	89.92	87.17	63.92	84.50	60.71	84.92	77.98
w/ FLARE MT	91.15	79.88	80.67	81.88	62.50	91.04	85.67	66.04	84.71	64.42	84.08	78.09
w/ FLARE	91.15	82.75	81.00	83.17	65.08	92.83	86.83	73.08	87.75	70.58	85.50	80.86
Translate-Train (fusion models are trained on data translated into the target language and evaluated using gold translations from the target language to the source language)												
XLM-R Large input-level fusion	92.04	91.24	91.08	90.55	90.69	91.99	90.88	91.23	91.07	90.07	90.52	90.93
w/ FLARE	92.04	89.24	88.98	82.55	90.07	90.22	88.15	71.20	87.58	72.93	85.71	84.66
mT5-XL input-level fusion	91.77	91.39	90.39	91.47	91.54	90.88	89.49	88.87	90.86	89.20	91.60	90.57
w/ FLARE	91.77	83.80	80.55	84.06	64.70	88.32	90.50	74.36	83.64	69.29	88.00	80.72

Table 7: Average scores per language in the NusaX dataset. Model performance is evaluated using the Macro F1 metric.

Model	Hparam	XNLI	TyDiQA	NusaX
XLMR-Large	epochs	10	10	20
	batch size	64	64	64
	sequence length	128	512	128
	learning rate	2e-5	2e-4	2e-4
mT5-XL	epochs	10	10	20
	batch size	64	64	64
	sequence length	128	512	128
	learning rate	2e-5	2e-4	2e-4
Llama 3 8B	epochs	3	3	5
	batch size	64	64	64
	sequence length	128	512	128
	learning rate	2e-5	2e-4	2e-4
Gemma 2 9B	epochs	3	3	5
	batch size	64	64	64
	sequence length	128	512	128
	learning rate	2e-5	2e-4	2e-4

Table 8: Hyperparameter configurations for each mPLM across the XNLI, TyDiQA, and NusaX datasets.

Task	Language	ISO Code	Source
XNLI	Arabic	ar	Crowd-sourced (Williams et al., 2018)
	Bulgarian	bg	
	Chinese	zh	
	French	fr	
	German	de	
	Greek	el	
	Hindi	hi	
	Russian	ru	
	Spanish	es	
	Swahili	sw	
	Thai	th	
	Turkish	tr	
	Urdu	ur	
	Vietnamese	vi	
TyDiQA	Arabic	ar	Wikipedia (Clark et al., 2020)
	Bengali	ben	
	Finnish	fi	
	Indonesian	ind	
	Korean	ko	
	Russian	ru	
	Swahili	sw	
	Telugu	tel	
NusaX	Acehnese	ace	SmSA (Purwarianti and Crisdayanti, 2019)
	Balinese	ban	
	Banjarese	bjn	
	Buginese	bug	
	Indonesian	ind	
	Javanese	jav	
	Madurese	mad	
	Minangkabau	min	
	Ngaju	nij	

Table 9: Overview of languages and corresponding source data used in the experiments, categorized by task.