MLAS-LoRA: Language-Aware Parameters Detection and LoRA-Based Knowledge Transfer for Multilingual Machine Translation

Tianyu Dong[1], Bo Li[2,3†], Jingsong Liu[4], Shaolin Zhu[1*], Deyi Xiong[1*]
[1]TJUNLP Lab, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Software, Tsinghua University, Beijing, China
[3]Baidu APP Technology and Platform R&D Department, Baidu Inc, Beijing, China
[4]NewTranx Co., Ltd, Beijing, China
{tydong, zhushaolin, dyxiong}@tju.edu.cn
{li-b19}@mails.tsinghua.edu.cn
{jinsong.liu}@newtranx.com

Abstract

Large language models (LLMs) have achieved remarkable progress in multilingual machine translation (MT), demonstrating strong performance even with limited parallel data. However, effectively fine-tuning LLMs for MT is challenging due to parameter interference, which arises from the conflicting demands of different language pairs and the risk of overwriting pre-trained knowledge. To address this issue, we propose **MLAS-LoRA**, a novel multiple language-aware LoRA knowledge transfer framework. MLAS-LoRA efficiently adapts LLMs to MT by selectively transferring knowledge from a large teacher to a small student model. Our approach first evaluates the awareness of neurons and extracts linguistic knowledge in the teacher model to both the general MT task and specific language pairs. We then propose a multiple language-specific LoRA architecture to inject the extracted knowledge into the student model. During fine-tuning, only the parameters of the relevant language-general and language-specific LoRA modules are updated. Experimental results on diverse multilingual language pairs demonstrate that MLAS-LoRA significantly outperforms strong baselines by +1.7 BLEU on average, including standard fine-tuning and other parameter-efficient methods.

1 Introduction

Large language models (LLMs) have attained notable advancements in natural language processing, particularly in challenging tasks such as multilingual machine translation (MT) (Zhu et al., 2024d; Siu, 2024), with ongoing work exploring diverse strategies for further improvement (Zhu et al., 2024c; Cui et al., 2024; Zhu et al., 2024a). As the scale of LLMs increases, these models demonstrate increasingly powerful capabilities, including improved zero-shot and few-shot learning for translation (Schaeffer et al., 2024; Zhu et al., 2024e). However, the training and inference of massive LLMs are computationally expensive, requiring substantial resources and hindering their deployment in applications (Wang and Li, 2024; Guo et al., 2025).

To address these challenges, some works propose various efficient fine-tuning methods. These methods aim to adapt LLMs to specific tasks with minimal parameter updates, such as adapter (Alves et al., 2023) and LoRA (Zhang et al., 2023b). While these methods have demonstrated that task-specific parameter adjustments are both detectable and editable within a single model, the broader question of whether such targeted knowledge is transferable across different LLMs remains an open and under-explored area (Zhong et al., 2024). Another research has focused on knowledge distillation (KD) to transfer knowledge from a large "teacher" model to a small "student" model (Gou et al., 2021). Compared to adapter and LoRA, it can leverage the knowledge encoded within a large and powerful LLM to improve the performance of a smaller model. However, current KD approaches often focus on mimicking the teacher's output distributions (e.g., through soft labels) and may overlook the rich, structured knowledge implicitly stored within the teacher's parameters themselves (Zhong et al., 2024).

Existing methods (Michel et al., 2019) have explored the saliency of attention heads in encoder-decoder and encoder-only architectures, aiming to identify task-relevant components and prune redundant ones to improve efficiency in single-task scenarios. In contrast, the present work focuses on decoder-only LLMs for multilingual machine translation, where understanding model behavior at the neuron level is particularly critical. Recent investigations for LLMs have revealed the existence of language-specific neurons (or parameters) (Cao

[†] Equal contribution.
[*] Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15645–15660
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024; Zhu et al., 2024b; Tang et al., 2024). These studies demonstrate that certain neurons appear to be specialized for particular languages, while others are language-agnostic, contributing to general linguistic processing (Liu et al., 2024a; Huang et al., 2024). This specialization can also lead to parameter interference, potentially causing two key issues for KD and LoRA approaches: (1) Distilling the entire teacher model's parameters to a student model often leads to performance degradation in some tasks within a multi-task setting (Zhong et al., 2024). (2) Applying LoRA indiscriminately across different languages during fine-tuning can result in a sub-optimal configuration (Cao et al., 2024; Tian et al., 2025). This is because the fine-tuning process may inadvertently disrupt the beneficial language-specific representations learned during pre-training.

To address these issues, we propose **MLAS-LoRA**, a novel multiple language-aware LoRA knowledge transfer framework for extracting language-specific parameters from a teacher LLM and subsequently injecting them into a student model via LoRA for multilingual machine translation. Specifically, we first evaluate the MT awareness of each neuron in the LLM's layers, identifying neurons that are significantly involved in the overall MT task. For those MT-relevant neurons, we further assess and extract the linguistic knowledge of each neuron to each individual language pair. Then, we propose a multiple language awareness LoRA architecture. This acts as a bridge to inject the knowledge extracted from both language awareness parameters of the teacher model into the student model. This facilitates efficient fine-tuning of the student on multilingual MT, achieving the knowledge transfer process. During fine-tuning, only the parameters of the language awareness LoRA modules corresponding to the current language pair are updated. This targeted approach minimizes parameter interference and maximizes the transfer of relevant knowledge.

To summarize, the key contributions of this paper are threefold: (1) We propose MLAS-LoRA that identifies and extracts language-specific and language-general knowledge from a teacher LLM's neurons, enabling more focused knowledge transfer to a student model. (2) We design a new LoRA-based method that selectively injects the extracted knowledge into the student, updating only relevant parameters for each language pair, minimizing interference and improving efficiency. (3) Experiments on ten language pairs show that our model achieves the state-of-the-art results compared to previous strong baselines and demonstrate the robustness of the proposed model under various settings.

2 Related Work

LLMs have shown remarkable success on a wide range of NLP tasks, including multilingual MT (Chowdhery et al., 2023; Touvron et al., 2023). Recent years have witnessed growing research interest in fine-tuning LLMs for domain-specific applications.

Prominent examples include adapter-based methods (Pfeiffer et al., 2021; Nguyen and Le, 2024). Houlsby et al. (2019a) introduce adapter modules, which are small bottleneck layers inserted between layers of a pre-trained Transformer, allowing for task-specific adaptation with minimal parameter updates. Pfeiffer et al. (2021) extend this with AdapterFusion, combining knowledge from multiple adapters trained on different tasks. Another approach is LoRA (Hu et al., 2022), which updates a low-rank decomposition of the weight matrices. Other approaches include prompt tuning (Lester et al., 2021), which prepends trainable soft prompts to the input, and prefix tuning (Li and Liang, 2021), which prepends trainable continuous prefixes to each layer's activations. Guo et al. (2021) find that fine-tuning multilingual LLMs do not always bring about improvements, and sometimes even undermine translation quality due to catastrophic forgetting.

Another significant body of work explores knowledge distillation (KD) for transferring capabilities and knowledge from a teacher LLM to a smaller student model (Hinton et al., 2015; Liao et al., 2025). Jain et al. (2023) develop a method for selecting the most informative multilingual data for distillation, focusing on high-resource languages to improve low-resource performance. Recentlly, Xu et al. (2024) explore weight selection for uniformly selecting parameters from a larger teacher model to initialize a smaller variant. Other studies concentrate on function-preserving methods (Honovich et al., 2023; Wang et al., 2023), ensuring the initialized large model replicates the behaviors of the original small model.

Recent work has explored language-agnostic components within multilingual LLMs (Bhattacharya and Bojar, 2023; Qin et al., 2024). Neu-

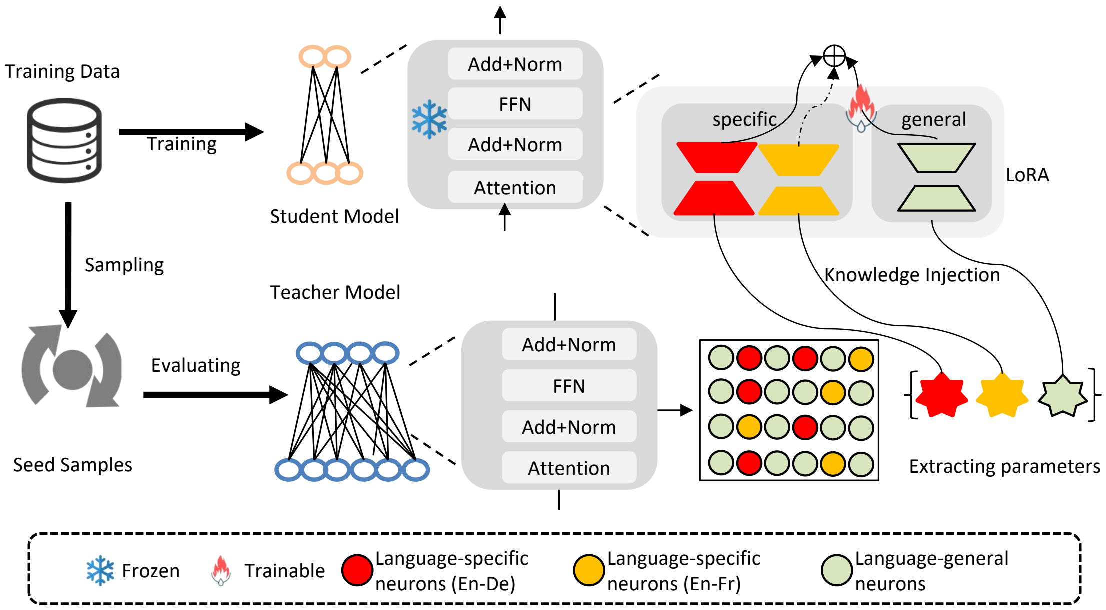

Figure 1: Illustration of the proposed MLAS-LoRA.

ron analysis, inspired by neurobiology (Patel et al., 2022), is gaining traction. Studies have shown that neurons can encode specialized contextual information (Gurnee et al., 2023), positional information (Voita et al., 2024), and linguistic properties (Marco and Fraser, 2024; Gurnee et al., 2024). However, much prior work on multilingual LLMs hasn't fully addressed negative language interaction. Assuming shared representations can be detrimental when languages require specialized processing. Fine-tuning or distilling without considering language-specific aspects can lead to interference. In this work, we explicitly identify and leverage both language-specific and language-general knowledge within a teacher LLM. By selectively extracting and transferring this knowledge using a multiple language awareness LoRA, we minimize negative language interaction and achieve more efficient and effective multilingual MT.

3 Methodology

The proposed MLAS-LoRA is illustrated in Figure 1. We first analyze which layers of the teacher LLM have strong relevance to a source-target language pair, and evaluate the strength of awareness of neurons at those layers to the given language pair. Then, we extract language-aware parametric knowledge from the teacher model. In order to use extracted language-aware knowledge, we propose a multiple language-aware LoRA as a bridge to inject the knowledge from the teacher model into the student.

3.1 Evaluating the Language Awareness of Neurons

As our goal is to transfer knowledge of a teacher model to a student model, we need to select a subset of layers from the teacher model that are most crucial for the MT. The number of selected layers should match the number of layers in the student model. We adapt a representation analysis approach, inspired by the concept of measuring activation differences between layers (Zhu et al., 2024b), to quantify the contribution of each layer in the teacher model. For a given language pair (s,t), we consider the forward pass of a source sentence x_s through the teacher model. Let $\mathbf{A}_i(x^{(n)})$ represent the activation vector of layer i during the n-th forward propagation, and N be the total number of forward propagations. We compute the layer relevance score R_i for each layer i as the L1-norm of the difference between the activations of consecutive layers:

$$\mathrm{R}_i(s,t) = \frac{1}{N}\sum_{n=1}^{N} ||\mathbf{A}_{i+1}(x^{(n)}) - \mathbf{A}_i(x^{(n)})||_1 \quad (1)$$

After calculating $\mathrm{R}_i(s,t)$ for all layers in the teacher model, we select the top L layers with the highest values. L is equal to the number of layers of the student model to be updated.

LLMs training enables knowledge transfer but also causes interference, largely due to optimization conflicts among various languages (Tan et al., 2024). Therefore, within the L selected relevant layers, we further analyze individual neurons to determine their language awareness. we divide neurons into language-general neurons and language-specific neurons. Language-general neurons capture knowledge that might include general linguistic principles across languages. Language-specific neurons are specialized for a particular language pair to address parameter interference.

To quantify neuron awareness, we design a sensitivity-based approach inspired by Taylor expansion methods. For a given neuron j in layer l, We define its awareness score $\Phi_{l,j}(s,t)$ for the language pair as follows:

$$\Phi_{l,j}(s,t) = \left| \frac{\partial \mathcal{L}}{\partial \mathbf{h}_{l,j}} \cdot \mathbf{h}_{l,j} \right| \quad (2)$$

$\mathcal{L}$ is the loss function. $\mathbf{h}_{l,j}$ is the output of neuron j in layer l. $\frac{\partial \mathcal{L}}{\partial \mathbf{h}_{l,j}}$ is the gradient of the loss with respect to the output of neuron j. This score approximates the change in the loss function if the output of neuron j is set to zero. We calculate $\Phi_{l,j}(s,t)$ by performing a forward and backward pass with seed sentences of the language pair.

To categorize neurons into language-general and language-specific neurons, we analyze the distribution of their awareness scores across different language pairs. For neuron j in layer l, we compute a set of awareness scores:

$$\Phi_{l,j} = \{\Phi_{l,j}(s_1,t_1), \Phi_{l,j}(s_2,t_2), ..., \Phi_{l,j}(s_n,t_n)\} \quad (3)$$

If $\Phi_{l,j}(s_i,t_i)$ has highest awareness score and $\Phi_{l,j}(s_i,t_i) < \lambda$, neuron j in layer L is classified as a language-general neuron. If the highest awareness score $\Phi_{l,j}(s_i,t_i) > \lambda$, neuron j in layer l is classified as a language-specific neuron. The threshold λ is a hyperparameter that distinguishes language-general neurons from language-specific neurons. We set the value of λ as 0.2 according to a previous evaluation (Zhu et al., 2024b). We examine the effect with different values of λ in Appendix A.6.

3.2 Extracting Parameters from Language-Aware Neurons

We face a practical challenge that the dimensionality of the teacher model's neurons typically exceeds that of the student. Therefore, we need a method to extract a relevant subset of parameters from the teacher model that is compatible with the student. For each selected layer l, and for each two-dimensional weight matrix $\mathbf{W}_T \in \mathbb{R}^{M\times N}$ in the teacher model, we aim to extract a submatrix $\mathbf{W}_T \in \mathbb{R}^{m\times n}$, where $M > m$ and $N > n$ are the corresponding dimensions in the teacher and student model. We use the neuron awareness scores $\mathbf{\Phi}_{l,j}$ to guide this extraction. Unlike simple dimensionality reduction, our goal is to preserve the structural integrity of the most relevant parts of the teacher's weight matrices. We perform this extraction separately for language-general and language-specific neurons.

Let $\mathbf{W}_T$ be a weight matrix in a selected layer. We define two sets of indices $\mathcal{I}_{\text{gen}}$ and $\mathcal{I}_{\text{spe}}$. $\mathcal{I}_{\text{gen}}$ is set of indices of language-general neurons. $\mathcal{I}_{\text{spe}}$ is set of indices of language-specific neurons associated with a language pair. For language-general neurons, we extract a submatrix $\mathbf{W}_s^{\text{gen}}$ as follows:

$$\mathbf{W}_s^{\text{gen}} = \text{Extract}(\mathbf{W}_T, \mathcal{I}_{\text{gen}}, m, n) \quad (4)$$

For language-specific neurons associated with a language pair, we extract a submatrix $\mathbf{W}_s^{\text{spe}}$ as follows:

$$\mathbf{W}_s^{\text{spe}} = \text{Extract}(\mathbf{W}_T, \mathcal{I}_{\text{spe}}, m, n) \quad (5)$$

Extract$(.)$ is awareness-based scoring function. For each submatrix, they compute an aggregate awareness score. For a language-general submatrix, this score is the sum of the average awareness scores of all the involved general neuron. For a language-specific matrix, this score is the sum of language-specific awareness values.

3.3 Multiple Language-Aware LoRA

To effectively transfer knowledge from the teacher to the student model and address the challenges of parameter interference in multilingual MT, we introduce a novel LoRA-based architecture MLAS-LoRA. This architecture builds upon the core idea of Low-Rank Adaptation (LoRA), which modifies a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ by adding a low-rank decomposition:

$$\mathbf{W}' = \mathbf{W} + \mathbf{BA} \quad (6)$$

where $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$and the rank $r <<$ $\min(d,k)$ Only $\mathbf{B}$ and $\mathbf{A}$ are updated during fine-tuning, significantly reducing the number of trainable parameters.

However, standard LoRA is language-agnostic and does not address the specific challenges of multilingual fine-tuning. Inspired by the concept of intrinsic language-specific subspaces (Voita et al., 2024), and the observation that some neurons are language-general and some are language-specific, we propose a dual-LoRA structure within each layer of the student model. This structure consists of language-general LoRA and language-specific LoRA. Language-general LoRA is a single LoRA module $(\mathbf{B}_{\text{gen}}, \mathbf{A}_{\text{gen}})$, that is shared across all language pairs. This module aims to capture and transfer general translation knowledge that is beneficial regardless of the specific languages. Language-specific LoRA is a set of LoRA modules $\{(\mathbf{B}_{\text{spe}}(s_1, t_1), \mathbf{A}_{\text{spe}}(s_1, t_1)), (\mathbf{B}_{\text{spe}}(s_2, t_2), \mathbf{A}_{\text{spe}}(s_2, t_2)), ..., (\mathbf{B}_{\text{spe}}(s_n, t_n), \mathbf{A}_{\text{spe}}(s_n, t_n))\}$, where each module $(\mathbf{B}_{\text{spe}}(s_i, t_i), \mathbf{A}_{\text{spe}}(s_i, t_i))$ is associated with a specific language pair. These modules aim to capture and transfer knowledge that is unique to the nuances of each language pair, mitigating parameter interference by isolating language-specific adaptations.

The key to transferring knowledge lies in how we initialize these LoRA modules. We first use the extracted language-general submatrix $\mathbf{W}_s^{\text{gen}}$ to initialize the language-general LoRA module. We perform Singular Value Decomposition (SVD) $\mathbf{W}_s^{\text{gen}} = U\Sigma V^T$. Then, we set:

$$\begin{aligned} \mathbf{B}_{\text{gen}} &= U[:, : r_{\text{gen}}] \cdot \Sigma[: r_{\text{gen}}, : r_{\text{gen}}] \\ \mathbf{A}_{\text{gen}} &= V^T[: r_{\text{gen}}, :] \end{aligned} \tag{7}$$

where r_{gen} is the rank of the language-general LoRA module. Similarly, we use the extracted language-specific submatrix $\mathbf{W}_s^{\text{spe}}$ to initialize the corresponding language-specific LoRA module using SVD $\mathbf{W}_s^{\text{spe}} = U\Sigma V^T$ as follows:

$$\begin{aligned} \mathbf{B}_{\text{spe}} &= U[:, : r_{\text{spe}}] \cdot \Sigma[: r_{\text{spe}}, : r_{\text{spe}}] \\ \mathbf{A}_{\text{spe}} &= V^T[: r_{\text{spe}}, :] \end{aligned} \tag{8}$$

During fine-tuning, we adopt a sparse activation mechanism to further enhance parameter efficiency and to strictly enforce the separation between language-general and language-specific knowledge. For given an input language pair i, The language-general LoRA module $(\mathbf{B}_{\text{gen}}, \mathbf{A}_{\text{gen}})$ is always activated. Only the language-specific LoRA module corresponding to the current language pair $(\mathbf{B}_{\text{spe}}(s_i, t_i), \mathbf{A}_{\text{spe}}(s_i, t_i))$ is activated. All other language-specific LoRA modules are frozen. The modified weight matrix $\mathbf{W}'$, given the input language pair i, is computed as:

$$\mathbf{W}' = \mathbf{W} + \mathbf{B}_{\text{gen}}\mathbf{A}_{\text{gen}} + (\mathbf{B}_{\text{spe}}(s_i, t_i)\mathbf{A}_{\text{spe}}(s_i, t_i)) \tag{9}$$

During backpropagation, only the parameters in the activated LoRA modules $(\mathbf{B}_{\text{gen}}, \mathbf{A}_{\text{gen}}, \mathbf{B}_{\text{spe}}(s_i, t_i), \mathbf{A}_{\text{spe}}(s_i, t_i))$ are updated. The original student model parameters and the parameters of the inactive language-specific LoRA modules remain frozen. This selective updating further reduces the number of trainable parameters and prevents interference between different language pairs.

4 Experiments

We conducted extensive experiments on a large number of language pairs on Gemma-2-2b-it (Team, 2024) and compared them with a series of strong baselines.

4.1 Datasets

In the fine-tuning stage, we selected 14 language pairs (as described in 3) to adjust the language-specific LoRA and language-general LoRA of LLM. All original training data are from the training set part of the WMT 18 dataset[1], and the data for calculating the awareness score come from the validation set of WMT 18. All data follow a license that can be used freely for research purposes. Specifically, we randomly extracted 200,000 sentence pairs for each translation direction, and randomly extracted 1,000 sentences for calculating the awareness score. In addition, we used the following translation instruction fine-tuning templates and applied them to our parallel data:

$$\text{Translate from [SRC] to [TGT]}: \tag{10}$$

where [SRC] and [TGT] denote the source and target language name of the language pair, respectively. We evaluated the performance of our model using the established test set (WMT18). Additionally, to demonstrate the generalization ability of our model across different datasets, we conducted evaluations on the FLORES-200 (Costa-Jussà et al., 2022) devtest set, with detailed results provided in Appendix A.5.

4.2 Settings and Baselines

Setting We executed a freezing operation on the parameters of non-current language-specific LoRA

[1] https://www.statmt.org/wmt18/

while exclusively fine-tuning the parameters within current language-specific LoRA and language-general LoRA. During the model fine-tuning stage, we configured the fintuning hytper-parameters as follows: the fine-tuning epoch was set to 3, the number of language pairs was specified as 14, the batch size was set 64, and the AdamW optimizer was employed. Additionally, the learning rate was set to 1e-4. Furthermore, we introduced a gradient accumulation operation, updating the model parameters every 8 iterations to enhance convergence. The teacher model used in our experiments is Gemma-2-9b-it, and the student model used is Gemma-2-2b-it.

Baselines We compared MLAS-LoRA with In-context Learning (Zhang et al., 2023a), which is a training-free approach that allows the LLMs to perform downstream tasks (we used 5 random shots as in-context demonstrations); P-tuning (Liu et al., 2022), which is a highly efficient prompt tuning method that achieves performance comparable to fine-tuning; LoRA (Hu et al., 2022), which fine-tunes a model for a downstream task by converting certain structures into low-rank matrices; and LoRA-Flow (Wang et al., 2024), which introduces dynamic fusion weights to combine LoRAs for generative tasks; MELoRA (Ren et al., 2024), improves parameter-efficient fine-tuning by using mini-ensemble low-rank adapters; AFLoRA (Liu et al., 2024b), is a parameter-efficient fine-tuning method that improves performance by incrementally freezing low-rank projection matrices; Full parameter fine-tuning, which fine-tunes all model parameters for the task; Adapter (Houlsby et al., 2019b), which facilitates the acquisition of new knowledge by incorporating additional adapter modules. LSLo (Cao et al., 2024), introduces language-specific LoRA for efficient fine-tuning of multilingual neural machine translation models. Detailed baselines are described in A.4. For evaluating translation performance, we used automatic evaluation metrics sacreBLEU.[2]

4.3 Main Results

Table 1 highlights the BLEU scores of various models and adaptation strategies on a multilingual MT across 10 language pairs. We also show the corresponding METEOR and COMET results in the Appendix A.3, and put the results of the remaining 4 language directions in Appendix A.7. The 0-shot row establishes the baseline performance of the untuned Gemma-2-2b-it model, revealing its inherent, yet limited, multilingual MT. BLEU range considerably (e.g., 24.26 for cs-en, but only 6.41 for en-et), indicating that while the LLM possesses some cross-lingual understanding, it's far from optimized for translation.

For fine-tuning methods, we observe a clear improvement across all approaches compared to the zero-shot and in-context baselines. However, significant differences emerge among the fine-tuning techniques themselves, revealing the importance of how knowledge is adapted. Methods like standard LoRA, MELoRA, AFLoRA, and Adapter do not explicitly distinguish between language-general and language-specific knowledge. They apply parameter updates within a shared parameter space for all languages. This can lead to parameter interference, where updates that benefit one language pair may degrade performance on others (we will demonstrate this in section 5.2). In contrast, ours consistently outperforms all baselines across every reported language pair and direction. This substantial and consistent improvement demonstrates the effectiveness of our core innovations: the explicit separation and targeted transfer of language-specific and language-general knowledge. By identifying neurons relevant to the overall MT task and further categorizing them based on their language awareness, ours avoids the pitfalls of indiscriminate parameter updates. The use of multiple language-specific LoRA modules allows for focused adaptation to each language pair, minimizing interference and maximizing the positive transfer of knowledge from the teacher model. The results strongly suggest that a language-aware approach, which considers both the shared and unique aspects of different languages, is crucial to achieve optimal performance in multilingual MT with LLMs.

4.4 Ablation Study

To validate the effectiveness of our method, we conducted an ablation study with ten experimental settings: -R-i, -R-a, -R-s, -M-i, -Ms-a, -Ms-spe, -Ms-gen, -Mr-a, -Mr-spe and -Mr-gen. These settings fall into four broad categories: (2)-M-i uses multiple randomly initialized LoRAs (one per language pair); (3) -Ms-* employs multiple LoRAs per layer, initialized with part parameters extracted from random layers of the teacher model; (4) and -Mr-* also uses multiple LoRAs per layer, but initialized using our proposed MLAS-LoRA extrac-

[2] BLEU+case.mixed+numrefs.1+smooth.none+tok.13a +version.2.2.1

Methods	cs-en	en-cs	de-en	en-de	et-en	en-et	fi-en	en-fi	ru-en	en-ru
Basic LLM Model										
0-shot	24.26	10.52	34.99	21.11	17.54	6.41	17.19	6.90	26.83	16.57
In-context	16.87	9.99	23.57	19.63	9.37	7.65	12.28	8.45	17.94	15.47
Prior Similar Studies										
Full fine-tune	21.63	11.03	30.14	18.15	17.98	11.10	15.18	8.80	22.18	13.16
P-tuning	22.10	12.35	34.42	23.16	20.37	11.09	17.71	9.29	24.95	16.23
Adapter	26.60	13.03	36.36	26.72	22.98	11.46	19.51	10.59	27.39	19.28
LoRA	26.03	13.17	36.16	25.10	22.16	11.23	19.03	9.55	26.73	19.44
MELoRA	26.75	13.86	36.45	26.69	22.84	11.47	19.47	10.63	27.31	19.59
AFLoRA	26.80	13.64	36.52	26.28	22.44	11.50	19.52	10.44	27.40	19.64
LSLo	27.07	13.49	36.66	25.83	22.06	11.19	19.53	10.26	27.31	19.74
LoRA-Flow	27.03	13.50	36.90	25.98	22.21	11.62	19.75	10.34	27.65	19.81
Ours										
MLAS-LoRA	**28.80**	**15.56**	**38.64**	**29.46**	**24.04**	**13.31**	**21.25**	**13.39**	**28.83**	**21.31**

Table 1: BLEU scores on the 10 language pairs for xx-to-English and English-to-xx translation. The highest score on each translation direction is highlighted in bold.

Methods	en-cs	en-de	en-et	en-fi
MLAS-LoRA	15.56	29.46	13.31	13.39
-R-i	13.17	25.10	11.23	9.55
-R-a	12.56	26.46	11.31	9.89
-R-s	13.03	26.14	11.61	10.20
-M-i	13.59	26.93	11.19	10.66
-Ms-a	13.29	26.16	10.61	9.72
-Ms-spe	13.33	26.22	10.80	9.93
-Ms-gen	13.63	26.39	11.02	10.37
-Mr-a	14.33	27.04	11.87	11.02
-Mr-spe	14.72	28.10	12.13	12.18
-Mr-gen	14.83	28.49	12.35	12.21

Table 2: Ablation Study

tion method.

The specific details for each setting are as follows:

- -R-i: Use a single LoRA for all language pairs, with random initialization.
- -R-a: Use a single LoRA for all language pairs, with LoRA parameters derived from related layers of the teacher model, without distinguishing between language-specific and language-general.
- -R-s: Use a single LoRA for all language pairs, with LoRA parameters derived from randomly selected layers of the teacher model, without distinguishing between language-specific and language-general.
- -M-i: Use a specific LoRA for each language pair, with random initialization.
- -Ms-a: Use a specific LoRA for each language pair, with LoRA parameters derived from randomly selected layers of the teacher model, without distinguishing between language-specific and language-general.
- -Ms-spe: Use a specific LoRA for each language pair, with LoRA parameters derived from randomly selected language-specific parameters of the teacher model.
- -Ms-gen: Use a specific LoRA for each language pair, with LoRA parameters derived from randomly selected general language parameters of the teacher model.
- -Mr-a: Use a specific LoRA for each language pair, with LoRA parameters derived from related layers of the teacher model, without distinguishing between language-specific and language-general.
- -Mr-spe: Use a specific LoRA for each language pair, with LoRA parameters derived from related language-specific parameters of the teacher model.
- -Mr-gen: Use a specific LoRA for each language pair, with LoRA parameters derived from related general language parameters of the teacher model.

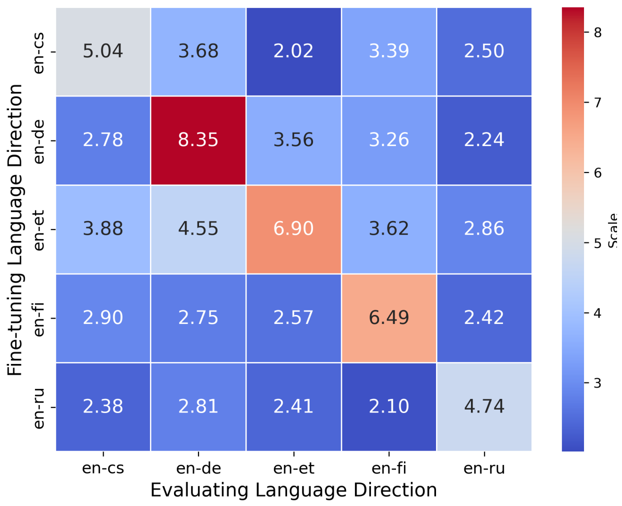

Figure 2: BLEU improvements achieved on other language pairs using the our method for fine-tuning only one language pair.

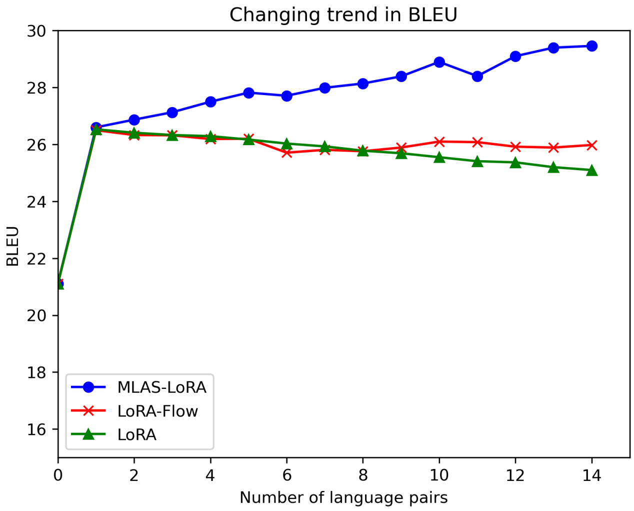

Figure 3: The effect of the number of training language pairs on en-de in terms of BLEU.

Table 2 presents the BLEU scores for each of these settings across the four language pairs (en-cs, en-de. en-et and en-fi).

Comparing to the single-LoRA (-R-*), the different multiple-LoRA (-Mr-* and -Ms-*) show improvement over the single-LoRA baselines, suggesting that simply providing separate parameter spaces for each language pair is beneficial. comparing -Ms-* and -Mr-* ,we can find that -Mr-* can get much better results. This shows the effectiveness of our method, explicitly separating language-specific and language-general parameters during extraction and injection, leads to the best overall performance. This separation allows for targeted fine-tuning that minimizes interference and maximizes positive knowledge transfer, resulting in substantial gains in translation quality. This demonstrates that carefully selecting and transferring knowledge from the teacher model, guided by neuron awareness and language specificity, is critical for effective multilingual fine-tuning.

5 Analysis

5.1 MLAS-LoRA Improves Transfer Learning across Languages

We examined the transfer learning ability of MLAS-LoRA in different translation directions. We fine-tuned the model using only parallel data from a particular language direction. In other words, we fine-tuned only the language-general and language-specific parameters for that language pair, and then observed the performance of the model in other language directions. The Y-axis of Figure 2 shows the single language direction that we have fine-tuned, and the X-axis shows the language direction of the test data, which is plotted as the improvement in the model's translation performance before and after the fine-tuning. Since LLM is a model that is not mainly trained on a parallel corpus, its translation performance before fine-tuning is poor, which is the reason for the large improvement in the model's translation performance. We observe that when fine-tuning one language direction, the results of other language directions can also be significantly improved, which proves that our method is effective in facilitating transfer learning between languages.

5.2 Effect of Increasing the Number of Language Pairs

In Figure 3, we investigate the impact of increasing the number of language pairs in the training data on the en-de translation performance. We first train the model using en-de data and then use other data. The X-axis represents the number of language pairs included in the training set, while the Y-axis shows the BLEU score on the en-de translation task. We compare three configurations: a single, shared LoRA (denoted as "LoRA"), a multi-LoRA setup where each language pair has its own dedicated LoRA module (implictly shown as "LoRA-Flow"), and our proposed method.

A key finding is the decreasing performance of the single shared LoRA as more languages are added, demonstrating the detrimental effect of pa-

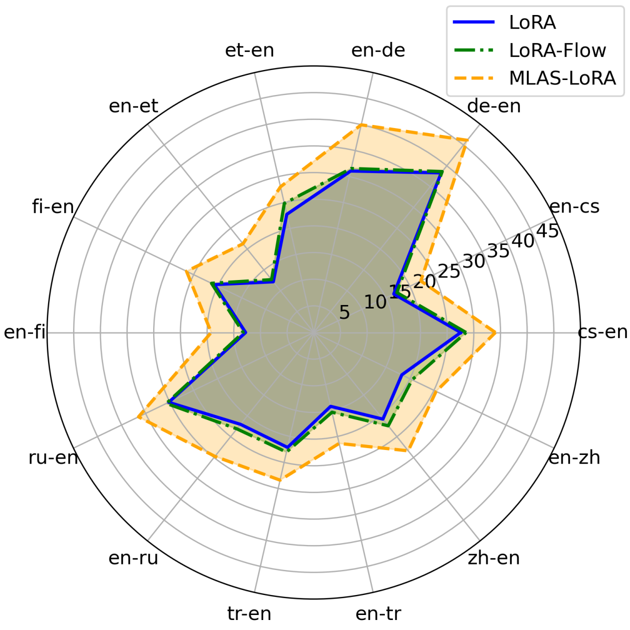

Figure 4: Comparison of BLEU scores on the WMT test set across ten language directions for fine-tuning the LLaMA3.1-8b model using LoRA, LoRA-Flow and our proposed method.

rameter interference. The multi-LoRA configuration, by providing separate modules for each language pair, mitigates this interference and maintains relatively stable performance. Crucially, our method, which combines multiple LoRAs with a mechanism for sharing language-general knowledge, exhibits a positive trend: performance improves as more language pairs are included. This highlights the ability to leverage both language-specific and language-general knowledge, demonstrating its scalability and robustness in a multilingual setting, and showcasing the benefits of cross-lingual transfer learning.

5.3 Results on Other LLMs

We also fine-tuned the LLaMA3.1-8b model (teacher model is LLaMA3.1-70b) using the our method and compared it with the LoRA and LoRA-Flow fine-tuning approach. Results are shown in Figure 4. We observe that across the 14 language directions selected our proposed method outperforms the LoRA and LoRA-Flow fine-tuning method. This demonstrates the applicability of the our method across different models, achieving optimal results not only in the gemma models but also in the LLaMA model.

6 Conclusion

In this paper, we have presented MLAS-LoRA, a novel framework for knowledge transfer in multilingual MT of LLMs. Our approach addresses the critical challenges of parameter interference, which commonly hinder the effective fine-tuning of LLMs for diverse language pairs. We propose a two-stage process: (1) identifying and extracting both language-general and language-specific knowledge from a pre-trained teacher model (2) injecting this extracted knowledge into a student model using a multiple language-specific LoRA architecture. Experimental results across a wide range of language pairs and benchmark datasets demonstrate that MLAS-LoRA consistently outperforms strong baselines.

Acknowledgments

The present research was supported by the National Key Research and Development Program (Grant No.2023YFE0116400), National Natural Science Foundation of China Youth Fund (Grant No.62306210) and the Tianjin Natural Science Foundation of Youth Fund (Grant No.23JCQNJC01690). We would like to thank the anonymous reviewers for their insightful comments.

Limitaion

Although MLAS-LoRA has shown excellent performance in multilingual MT tasks, with flexibility that it can be fine-tuned for different language directions, and can transfer translation capabilities between multiple languages, there is still room for improvement in the applicability of datasets. The model will adapt to the text style of a specific dataset during training, but when faced with a different text style of a new dataset, it may not be able to adjust quickly. Considering the complexity and diversity of datasets in actual application scenarios, how to further optimize the model to better adapt to the differences between different datasets is still an important direction that needs to be focused on in future research.

Ethics Statement

This study adheres to the ethical guidelines set forth by our institution and follows the principles outlined in the ACM Code of Ethics and Professional Conduct. All datasets used in our experiments are publicly available.

References

Duarte M Alves, Nuno M Guerreiro, João Alves, José Pombal, Ricardo Rei, José GC de Souza, Pierre Colombo, and André FT Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. *arXiv preprint arXiv:2310.13448*.

Sunit Bhattacharya and Ondrej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*, pages 120–126. Association for Computational Linguistics.

Zhe Cao, Zhi Qu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Exploring intrinsic language-specific subspaces in fine-tuning multilingual neural machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21142–21157.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. *arXiv preprint arXiv:2406.07081*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Junliang Guo, Zhirui Zhang, Linli Xu, Boxing Chen, and Enhong Chen. 2021. Adaptive adapters: An efficient way to incorporate BERT into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1740–1751.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *Trans. Mach. Learn. Res.*, 2024.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14409–14428. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019a. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019b. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5028–5041. Association for Computational Linguistics.

Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Madry. 2023. A data-based perspective on transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 3613–3622. IEEE.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Huanxuan Liao, Shizhu He, Yupu Hao, Xiang Li, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2025. Skintern: Internalizing symbolic knowledge for distilling better cot capabilities into small language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 3203–3221. Association for Computational Linguistics.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024a. Unraveling babel: Exploring multilingual activation patterns of llms and their applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11855–11881. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Zeyu Liu, Souvik Kundu, Anni Li, Junrui Wan, Lianghao Jiang, and Peter Beerel. 2024b. AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–167, Bangkok, Thailand. Association for Computational Linguistics.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1009–1020. ELRA and ICCL.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Tuc Nguyen and Thai Le. 2024. Adapters mixup: Mixing parameter-efficient adapters to enhance the adversarial robustness of fine-tuned pre-trained text classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21183–21203. Association for Computational Linguistics.

Gal Patel, Leshem Choshen, and Omri Abend. 2022. On neurons invariant to sentence structural changes in neural machine translation. In *Proceedings of the 26th Conference on Computational Natural Language Learning, CoNLL 2022, Abu Dhabi, United Arab Emirates (Hybrid Event), December 7-8, 2022*, pages 194–212. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 487–503. Association for Computational Linguistics.

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, and Heng Ji. 2024. Why does new knowledge create messy ripple effects in llms? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12602–12609. Association for Computational Linguistics.

Pengjie Ren, Chengshun Shi, Shiguang Wu, Mengqi Zhang, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Jiahuan Pei. 2024. MELoRA: Mini-ensemble low-rank adapters for parameter-efficient fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3064, Bangkok, Thailand. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Sai Cheong Siu. 2024. Revolutionising translation with ai: Unravelling neural machine translation and generative pre-trained large language models. In *New*

Advances in Translation Technology: Applications and Pedagogy, pages 29–54. Springer.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6506–6527. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5701–5715. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Yimin Tian, Bolin Zhang, Zhiying Tu, and Dianhui Chu. 2025. Adapters selector: Cross-domains and multi-tasks lora modules integration usage method. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 593–605. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1288–1301. Association for Computational Linguistics.

Hanqing Wang, Bowen Ping, Shuo Wang, Xu Han, Yun Chen, Zhiyuan Liu, and Maosong Sun. 2024. LoRA-flow: Dynamic LoRA fusion for large language models in generative tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12871–12882, Bangkok, Thailand. Association for Computational Linguistics.

Renzhi Wang and Piji Li. 2024. Lemoe: Advanced mixture of experts adaptor for lifelong model editing of large language models. *arXiv preprint arXiv:2406.20030*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Zhiqiu Xu, Yanjie Chen, Kirill Vishniakov, Yida Yin, Zhiqiang Shen, Trevor Darrell, Lingjie Liu, and Zhuang Liu. 2024. Initializing models with larger ones. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Zhong Zhang, Bang Liu, and Junming Shao. 2023b. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, Toronto, Canada. Association for Computational Linguistics.

Ming Zhong, Chenxin An, Weizhu Chen, Jiawei Han, and Pengcheng He. 2024. Seeking neural nuggets: Knowledge transfer in large language models from a parametric perspective. In *12th International Conference on Learning Representations, ICLR 2024*.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024b. Landermt: Dectecting and routing language-aware neurons for selectively finetuning llms to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.

Shaolin Zhu, Leiyu Pan, and Deyi Xiong. 2024c. Fedsicl: Enhancing translation ability and efficiency of large language model by optimizing demonstration selection. *Information Processing & Management*, 61(5):103825.

Shaolin Zhu, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024d. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024e. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

A Appendix

A.1 Task Formulation

Our primary goal is to improve the performance of a student model on a multilingual MT task by transferring language awareness knowledge from a teacher model. The teacher model is parameterized by Γ_T, and the student model is parameterized by Γ_S. Typically, $|\Gamma_T| > |\Gamma_S|$, meaning the teacher model has significantly more parameters than the student model. Let τ represent the multilingual MT task, which encompasses a set of language pairs $\Phi = \{(s_1, t_1), (s_2, t_2), ..., (s_n, t_n)\}$, where s_i denotes the source language and t_i denotes the corresponding target language. We focus on transferring knowledge awareness to specific source-target language pairs, which can effectively avoid parameter interference. In order to achieve this, we identify and extract parameters from the teacher model, that are particularly relevant to the specific language pair. We represent this extraction process formally as a function:

$$\Gamma_T^{(S,T)} = \text{Extract}(\Gamma_T, \Gamma_S, \Gamma_T) \qquad (11)$$

The extracted parameters $\Gamma_T^{(S,T)}$ are then injected into the student model. This injection modifies the student model's parameters, resulting in a new parameter set Γ'_S. We represent this injection process as:

$$\Gamma'_S = \text{Inject}(\Gamma_T^{(S,T)}) \qquad (12)$$

After the injection, the student model Γ'_S can be optionally fine-tuned on a training dataset specific to the language pair. This fine-tuning step allows the student model to further adapt the injected knowledge to the specific characteristics of the language pair. During the fine-tuning, we only adjust a small part from Γ'_S.

A.2 Detail Languages

Each language and its ISO 639 code are shown in Table 3.

ISO 639	Language
cs	Czech
de	German
en	English
et	Estonian
fi	Finnish
ru	Russian
tr	Turkish
zh	Chinese

Table 3: ISO 639 Language Codes and Names

A.3 Results on METEOR and COMET

In addition to BLEU, we use two machine translation quality metrics, METEOR and COMET, to evaluate the translation results generated by the Gemma-2-2b-it to more fully explain the superiority of our method. When we use COMET to evaluate translation quality, the model we use is Unbabel/wmt22-comet-da.[3] The results are shown in Table 4 and 5, respectively.

A.4 Detail Baselines

- In-context (Zhang et al., 2023a) refers to the ability of large language models (LLMs) to use contextual information around the input text to improve performance when performing tasks. In the study of machine translation, it is reflected in the specific application of few-shot prompting. In few-shot prompting, the model will receive a small number of labeled examples, which are presented to the model as contextual information to help the model better understand the task. For example, in a translation task, these examples will be connected to the test input according to a specific template, and the model will improve the translation quality by learning the input and output patterns of these examples.

- P-tuning (Liu et al., 2022) is a type of prompt tuning. In the case where the parameters of the pre-trained language model are frozen, it completes specific tasks by adjusting continuous prompts (i.e., adding trainable continuous embeddings), rather than updating the entire model parameter set like fine-tuning. P-tuning includes techniques such as reparameterization (although its effectiveness varies depending on tasks and datasets), adjusting the prompt length (simple classification tasks prefer shorter prompts, while hard sequence tagging tasks prefer longer prompts), multi-task learning (which can jointly optimize multiple tasks before fine-tuning a single task to provide better initialization), and using a randomly initialized classification head instead

[3] https://huggingface.co/Unbabel/wmt22-comet-da

Methods	cs-en	en-cs	de-en	en-de	et-en	en-et	fi-en	en-fi	ru-en	en-ru
Basic LLM Model										
0-shot	54.03	42.94	61.66	48.30	47.42	33.75	47.02	33.68	56.02	55.16
In-context	41.28	39.34	47.12	45.62	31.08	34.23	36.81	35.94	40.44	50.54
Prior Similar Studies										
Full fine-tune	52.20	44.87	58.71	48.85	48.81	43.45	45.58	38.58	52.49	53.79
P-tuning	53.08	45.39	59.12	50.02	49.73	44.05	46.94	38.79	53.98	54.22
Adapter	53.74	46.85	60.38	57.96	52.83	42.70	48.25	40.53	54.20	59.55
LoRA	54.23	47.03	60.68	57.66	52.62	42.68	48.59	40.90	54.82	59.76
MELoRA	54.28	47.08	60.74	57.72	52.67	42.72	48.64	40.94	54.88	59.82
AFLoRA	54.45	47.13	60.89	57.85	52.75	42.83	48.77	41.09	55.05	59.97
LSLo	54.31	47.10	60.81	57.74	52.70	42.75	48.66	40.97	54.90	59.84
LoRA-Flow	54.27	47.97	61.41	58.09	52.86	42.97	48.71	41.33	55.19	59.80
Ours										
MLAS-LoRA	**56.12**	**48.91**	**63.99**	**59.16**	**53.12**	**44.93**	**49.18**	**42.11**	**56.40**	**60.45**

Table 4: METEOR scores on the 10 language pairs for xx-to-English and English-to-xx translation. The highest score on each translation direction is highlighted in bold.

Methods	cs-en	en-cs	de-en	en-de	et-en	en-et	fi-en	en-fi	ru-en	en-ru
Basic LLM Model										
0-shot	79.97	69.99	83.63	71.70	75.63	56.28	79.30	65.20	82.07	74.47
In-context	67.40	66.84	69.62	68.54	57.89	59.16	67.40	70.45	68.71	71.87
Prior Similar Studies										
Full fine-tune	79.43	68.53	83.23	72.95	77.28	67.08	78.46	70.26	80.84	76.64
P-tuning	80.53	70.45	84.71	73.84	78.24	68.32	79.59	72.74	81.31	77.49
Adapter	83.03	78.56	86.34	84.06	81.56	69.35	82.03	78.93	83.85	86.91
LoRA	82.94	78.40	86.14	84.22	81.53	68.96	82.78	78.64	83.22	86.73
MELoRA	83.02	78.49	86.32	84.30	81.69	69.03	82.86	78.76	83.38	86.82
AFLoRA	83.11	78.56	86.23	84.39	81.61	69.10	82.95	78.80	83.30	86.91
LSLo	83.08	78.48	86.26	84.29	81.65	69.06	82.93	78.72	83.35	86.84
LoRA-Flow	83.19	78.64	86.40	84.47	81.77	69.17	83.03	78.87	83.47	87.00
Ours										
MLAS-LoRA	**84.38**	**79.78**	**87.52**	**85.60**	**83.08**	**71.18**	**84.90**	**81.55**	**84.78**	**88.02**

Table 5: COMET scores on the 10 language pairs for xx-to-English and English-to-xx translation. The highest score on each translation direction is highlighted in bold.

of a language modeling head for prediction.

- LoRA (Low-Rank Adaptation) (Hu et al., 2022) is a method designed to efficiently adapt large pre-trained models to specific downstream tasks without the need for full fine-tuning. Instead of retraining all model parameters, LoRA freezes the pre-trained weights and inserts trainable low-rank matrices (decompositions) into each layer of the Transformer architecture. This significantly reduces the number of trainable parameters, lowering both memory requirements and computational cost.

- Adapter tuning (Houlsby et al., 2019b) is a parameter - efficient transfer learning method for NLP. Adapter modules are added between layers of a pre - trained network. In the Transformer architecture, two serial adapters are inserted after each of the attention and feed - forward sub - layers in each Transformer layer. They are applied to the output of the sub - layer, after the projection back to the input size but before the skip connection, and the output of the adapter is passed to the following layer normalization.

- MELoRA (Mini-Ensemble Low-Rank Adapters) (Ren et al., 2024) is an extension of the LoRA method designed to improve parameter-efficient fine-tuning (PEFT) of large pre-trained language models. While LoRA reduces the number of trainable parameters by using low-rank matrices, it can sometimes suffer from generalization errors on specific tasks. MELoRA addresses this by training an ensemble of mini LoRAs, each with a small number of parameters, while maintaining a higher rank than LoRA.

- AFLoRA (Adaptive Freezing of Low-Rank Adaptation) (Liu et al., 2024b) is a parameter-efficient fine-tuning (PEFT) method designed to improve the efficiency of adapting pre-trained models. It builds upon LoRA by adding parallel paths of trainable low-rank matrices (down-projection and up-projection), each followed by a feature transformation vector. During fine-tuning, AFLoRA uses a "freezing score" to incrementally freeze these projection matrices, reducing the number of trainable parameters and helping to mitigate overfitting.

- LoRA-Flow (Wang et al., 2024) is an extension of the LoRA method designed to improve the combination of multiple LoRAs for generative tasks. While previous approaches to combining LoRAs use task-level weights, which apply the same weights to all tokens or examples, LoRA-Flow introduces dynamic weights that adjust the contribution of different LoRAs depending on the specific tokens or parts of the task.

- LSLo (Cao et al., 2024) introduces language-specific LoRA for efficient fine-tuning of multilingual neural machine translation (NMT) models. Instead of fine-tuning all model parameters, which can lead to inefficiency and negative interactions among languages, the proposed method isolates each language's intrinsic subspace, fine-tuning only a small fraction of parameters specific to that language. The approach includes architecture learning techniques and a gradual pruning schedule during fine-tuning to explore the optimal settings and minimal subspaces for each language.

A.5 Results on Other Datasets

In addition to the main evaluation on the WMT18 test set, we also conducted tests on the FLORES-200 devtest set to assess the generalization capability of our model on a different dataset. Note that all fine-tuning procedures, including language-specific and language-general LoRA adjustments, as well as neuron identification processes, were exclusively performed using the WMT18 training and validation data. The evaluation on FLORES-200 was purely for testing purposes and no further fine-tuning was done on this dataset. The results on FLORES-200 devtest in Table 7 demonstrate that our model maintains strong performance across diverse datasets, indicating good generalization beyond the original fine-tuning data.

A.6 The effect of λ

We define λ_i as the normalized Φ, which is the value of $\Phi_{l,j}(s_i, t_i)$ for a specific language pair (s_i, t_i) divided by the sum of Φ values for all language pairs. We assume that the number of language pairs is N. Specifically, λ_i is given by:

Methods	tr-en		en-tr		zh-en		en-zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Basic LLM Model								
0-shot	17.65	78.33	6.24	61.71	18.21	77.56	8.59	49.04
In-context	10.36	62.38	9.18	69.08	14.68	69.44	22.25	72.07
Prior Similar Studies								
Full fine-tune	19.39	80.68	11.59	76.92	14.68	77.79	21.92	79.65
P-tuning	21.54	81.63	11.98	78.85	17.05	78.96	26.69	81.02
Adapter	23.50	84.13	13.74	83.31	19.92	81.16	28.41	84.53
LoRA	23.63	83.33	13.34	82.73	19.75	80.93	28.34	84.85
MELoRA	23.47	83.42	13.44	82.82	19.81	81.07	28.58	84.94
AFLoRA	23.36	83.50	13.47	82.90	19.85	81.07	28.62	85.02
LSLo	23.27	83.45	13.54	82.86	19.36	81.04	27.93	84.98
LoRA-Flow	23.78	83.58	13.59	82.98	19.84	81.13	28.30	85.11
Ours								
MLAS-LoRA	**24.14**	**85.06**	**14.19**	**84.94**	**20.47**	**82.36**	**29.48**	**86.02**

Table 6: BLEU and COMET scores on the 4 language pairs for xx-to-English and English-to-xx translation. The highest score on each translation direction is highlighted in bold.

Methods	fi-en	en-fi	tr-en	en-tr
Basic LLM Model				
0-shot	23.23	8.99	24.76	10.56
In-context	14.08	8.20	10.95	8.96
Prior Similar Studies				
Full fine-tune	21.63	8.80	22.15	9.60
P-tuning	22.10	9.29	23.45	10.22
Adapter	22.98	10.11	24.66	11.05
LoRA	24.39	10.05	26.59	13.19
MELoRA	24.97	10.47	27.15	13.65
AFLoRA	25.60	10.80	27.85	14.10
LSLo	25.15	10.26	27.09	13.50
LoRA-Flow	26.12	10.97	28.10	14.40
Ours				
MLAS-LoRA	**27.25**	**11.50**	**29.60**	**15.14**

Table 7: BLEU scores on the 4 language pairs for xx-to-English and English-to-xx translation. The highest score on each translation direction is highlighted in bold.

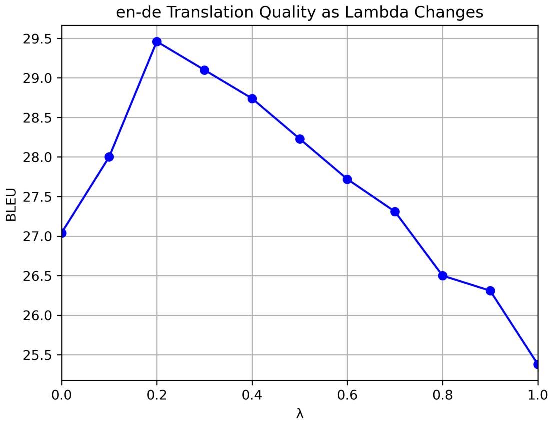

Figure 5: En-de Translation Quality as λ Changes.

$$\lambda_i = \frac{\Phi_{l,j}(s_i, t_i)}{\sum_{n=1}^{N} \Phi_{l,j}(s_n, t_n)}$$

We tested the translation quality in the en-de language direction with different λ settings, as shown in the Figure 5.

A.7 More Results

In addition to the language directions mentioned above, we also used Gemma-2-2b-it to conducted experiments on the remaining four language directions of WMT18. The experimental results are shown in Table 6.