Consensus-Aligned Neuron Efficient Fine-Tuning Large Language Models for Multi-Domain Machine Translation

Shuting Jiang[1,2], Ran Song[1,2]*, Yuxin Huang[1,2], Yan Xiang[1,2], Yantuan Xian[1,2], Shengxiang Gao[1,2], Zhengtao Yu[1,2]*

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
{shuting_jiang22, song_ransr, huangyuxin2004}@163.com, sharonxiang@126.com, xianyt@kust.edu.cn, {gaoshengxiang.yn, ztyu}@hotmail.com

Abstract

Multi-domain machine translation (MDMT) aims to build a unified model capable of translating content across diverse domains. Despite the impressive machine translation capabilities demonstrated by large language models (LLMs), domain adaptation still remains a challenge for LLMs. Existing MDMT methods such as in-context learning and parameter-efficient fine-tuning often suffer from domain shift, parameter interference and limited generalization. In this work, we propose a neuron-efficient fine-tuning framework for MDMT that identifies and updates consensus-aligned neurons within LLMs. These neurons are selected by maximizing the mutual information between neuron behavior and domain features, enabling LLMs to capture both generalizable translation patterns and domain-specific nuances. Our method then fine-tunes LLMs guided by these neurons, effectively mitigating parameter interference and domain-specific overfitting. Comprehensive experiments on three LLMs across ten German-English and Chinese-English translation domains evidence that our method consistently outperforms strong PEFT baselines on both seen and unseen domains, achieving state-of-the-art performance.

Code — https://github.com/fortunatekiss/CANEFT

Introduction

Multi-domain machine translation (MDMT) aims to build a unified model capable of accurately translating domain-specific terminology and context across diverse domains such as law, medicine, and subtitles (Pham, Crego, and Yvon 2021; Saunders 2022; Moslem et al. 2023). Conventional encoder-decoder approaches heavily rely on large amounts of parallel domain data, which are often scarce and costly (Li, Wang, and Yu 2020; Saunders 2022). In contrast, Large Language Models (LLMs), pretrained on extensive unlabeled corpora, acquire strong cross-lingual capabilities and show promising performance in general-domain translation (Zhao et al. 2023; Huang et al. 2024) but still facing challenges when translating domain-specific content (Pang et al. 2025). LLMs can notably improve their domain-specific translation capabilities via In-Context Learning (ICL) with a few examples (Ghazvininejad, Gonen, and Zettlemoyer 2023; He et al. 2024; Aycock and Bawden 2024; Li et al. 2025). However, the performance of ICL depends heavily on the quality of in-domain examples and often degrades for MDMT (Vilar et al. 2023).

Existing studies attempt to address these challenges

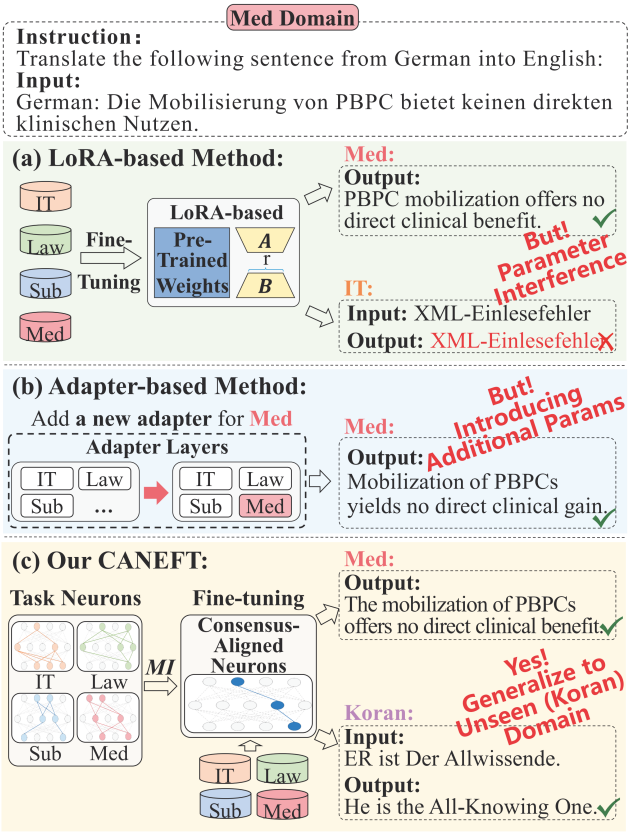

Figure 1: (a) LoRA-based fine-tuning causes parameter interference, while (b) adapter-based methods introduce additional parameters. (c) Our proposed CANEFT addresses these issues by only updating consensus-aligned neurons.

*Corresponding author

through parameter-efficient fine-tuning (PEFT) such as LoRA (Alves et al. 2023; Zheng et al. 2024b) and adapters (Hu et al. 2023; Wu et al. 2024; Eschbach-Dymanus et al. 2024). However, LoRA-based methods often face parameter interference. As shown in Figure 1(a), this interference can cause the model to overfit specific domains (Medical) while degrading performance on others (IT). Adapter-based methods typically introduce separate modules for each domain. As shown in Figure 1(b), these methods increase training and memory costs as the number of domain grows, and lacks generalization to unseen domains. In summary, these limitations motivate a central question: *Can we design a robust PEFT method for multi-domain machine translation without introducing additional parameters?*

Recent research has investigated neuron behavior in LLMs for tasks such as multilingual machine translation (Zhu et al. 2024), arithmetic reasoning (Hersche et al. 2024; Rai and Yao 2024) and knowledge editing (Jiang et al. 2024; Li et al. 2024). These studies indicate that neuron subsets spontaneously encode language- or task-specific functions, suggesting the potential of neuron-based approaches to enhance both performance and efficiency. Inspired by these findings, we delve into the correlation between neurons behavior and MDMT task, aiming to disentangle MDMT capabilities from a neuronal perspective in LLMs. Previous studies on task-specific neuron selection and fine-tuning have primarily focused on identifying neurons based on activation patterns or gradient variations (Song et al. 2024; Tang et al. 2024). However, these methods frequently select neurons either irrelevant to MDMT or overly specialized for single domains, which hinders generalization and causes domain-specific overfitting.

Intriguingly, recent neuroscience research shows that consensus-building dialogues enhance neural alignment among group members, with synchronized brain activity even generalizing to novel, unseen stimuli (Sievers et al. 2024). Motivated by this, we hypothesize that within LLMs, certain neurons might consistently encode knowledge relevant across multiple domains. We term these neurons *consensus-aligned neurons*. As illustrated in Figure 1(c), fine-tuning on consensus-aligned neurons achieves enhanced translation performance and generalization across multiple domains, contrasting with strategies that isolating domain-specific neurons. Consequently, accurately identifying the consensus-aligned neurons is a crucial prerequisite for improving MDMT performance.

In this work, we propose **C**onsensus-**A**ligned **N**euron **E**fficient **F**ine-**T**uning (CANEFT) for MDMT. This framework identifies and leverages consensus-aligned neurons to improve translation performance on both seen and unseen domains. Specifically, we first detect MDMT task-related neurons through activation-gradient analysis during inference. We then compute the mutual information (MI) between each task neuron and each domain to pinpoint consensus-aligned neurons. Finally, our approach significantly enhance LLM multi-domain translation performance by masking irrelevant neurons and fine-tuning on the identified consensus-aligned neurons. We conducted extensive experiments and analysis for German-English and Chinese-English translation across 10 domains. The results show that our method surpasses the full fine-tuning baseline by an average of 1.3 BLEU on De⇒En and 1.4 BLEU on Zh⇒En, also demonstrating strong generalization to unseen domains.

The main contributions of this work are as follows:

- We introduce consensus-aligned neurons critical for MDMT through MI-based strategy. These neurons effectively mitigate parameter interference and reduce the need for extensive fine-tuning domain data.
- We propose a neuron-efficient fine-tuning framework for MDMT, which selectively updates multi-domain consensus-aligned neurons to enhance both translation quality and cross-domain generalization.
- We validated our method on 3 instruction-tuned LLMs across 10 domain translation tasks in German-English and Chinese-English. Our method achieves an average BLEU improvement of 1.3 on De⇒En and 1.4 on Zh⇒En over the best-performing baseline, and demonstrates robust generalization to unseen domains.

Related Work

Domain Machine Translation

Recent research on domain-specific machine translation with LLMs has explored both inference-time adaptation and fine-tuning. Aycock and Bawden (2024) propose a topic-guided demonstration retrieval method to enhance translation performance of ICL without fine-tuning. Li et al. (2025) compare retrieval- and generation-based domain prompting, showing retrieval provides better grounding while generation enables flexibility. Hu et al. (2024) build a MDMT benchmark and enhance cross-domain performance via domain-aware chain-of-thought fine-tuning. Zheng et al. (2024a) introduce dictionary- and retrieval-augmented fine-tuning to bridge terminology gaps in domain translation.

Neuron Analysis in LLMs

Recent studies have revealed that neurons in LLMs exhibit modular behavior, with certain neurons responsible for specific languages (Tan, Wu, and Monz 2024; Tang et al. 2024; Zhu et al. 2024), tasks (Song et al. 2024; Leng and Xiong 2025), or knowledge (Dai et al. 2022; Chen et al. 2024; Niu et al. 2024; Mao et al. 2025). This intrinsic structure has motivated a growing works on neuron-level analysis (Voita, Ferrando, and Nalmpantis 2024) and selective fine-tuning (Xu et al. 2025). Tan, Wu, and Monz (2024) show that neuron activations correlate with language typological proximity, while Tang et al. (2024) introduce LAPE to identify language-specific neurons and control output via targeted activation. Zhu et al. (2024) achieve strong multilingual translation performance by routing updates through language-specific and general neurons. For tasks, Song et al. (2024) and Leng and Xiong (2025) identify task-specific neurons using activation patterns and gradient attribution, respectively. In terms of factual knowledge, Dai et al. (2022) define "knowledge neurons" that activate for specific facts. Chen et al. (2024) shedding light on the mechanisms of cross-lingual factual knowledge storage in LLM neurons.

Unlike previous work, we identify consensus-aligned neurons for MDMT by measuring mutual information with domain context, and propose a neuron-efficient fine-tuning framework that enhances cross-domain translation while mitigating PEFT limitations.

Methodology

In this section, we exhaustively introduce our neuron-efficient fine-tuning framework, designed to enhance LLMs performance and generalization on MDMT through identifying and selectively updating multi-domain consensus-aligned neurons. The framework consists of three key steps: *(i)* ***MDMT task-relevant neuron identification*** identifies MDMT task related neurons in feed-forward network (FFN) through calculating neuron activation sensitivity and gradient magnitude; *(ii)* ***MI-based multi-domain consensus-aligned neuron selection*** measures MI between importance of these task-relevant neurons and domain features, then select a small critical subset as MDMT consensus-aligned neurons that capture cross-domain translation knowledge as well as domain-specific nuances; *(iii)* ***Neuron-efficient fine-tuning*** only updates the MDMT consensus-aligned neurons using domain translation examples, enabling multi-domain adaptation and robust generalization.

MDMT Task-Relevant Neuron Identification

Neurons that consistently exhibit strong gradient and activation responses to task-specific inputs are likely to encode MDMT-relevant features. To identify such neurons, we compute gradient-activation importance scores for each neuron in the FFN layers of an LLM f_θ during MDMT inference.

For each domain d, we utilize parallel data $D_d = (\mathcal{T}^d, \mathbf{x}^d, \mathbf{y}^d)$, where $\mathbf{x}^d = \{x_1^d, ..., x_K^d\}$ is the source sequence and $\mathbf{y}^d = \{y_1^d, ..., y_T^d\}$ is the target sequence. $\mathcal{T}^d$ is the domain-specific instruction designed to inform the LLM of the domain context. For example, a De⇒En IT domain instruction is: "*You are a translation specialist who specializes in translating texts from German to English in the IT domain. Translate the following content into English and only reply to the translated sentence without line breaks or special symbols.*"

Inspired by studies on importance-based neuron selection for multilingual machine translation (Xie et al. 2021), we assess a neuron's importance by multiplying its activation during the forward pass by the loss gradient with respect to that activation during the backward pass. Neurons with high importance scores are identified as strongly associated with the MT task for a given domain d.

Specifically, let $A_{l,j}^{(d)}$ denote the activation of the j-th neuron in the l-th FFN layer when processing an input $\mathbf{x}^d$ from domain d, and $G_{l,j}^{(d)}$ be the gradient of the loss with respect to that activation:

$$G_{l,j}^{(d)} = \frac{\partial \mathcal{L}^{(d)}(\mathbf{x}^d, \mathbf{y}^d)}{\partial h_{l,j}^{(d)}}, \tag{1}$$

where $h_{l,j}^{(d)}$ is the output of the j-th neuron in the l-th FFN layer for domain d. And the token-level cross-entropy loss $\mathcal{L}^{(d)}$ between the model's prediction and the reference translation $\mathbf{y}^d$ is:

$$\mathcal{L}^{(d)}(\mathbf{x}^d, \mathbf{y}^d) = -\sum_{t=1}^{T} \log p_\theta(y_t^d \mid y_{<t}^d, \mathbf{x}^d, \mathcal{T}^d). \tag{2}$$

The neuron importance score for d is then computed as:

$$I_{l,j}^{(d)} = \mathbb{E}(\mathbf{x}^d, \mathbf{y}^d) \sim D_d \left[\left| A_{l,j}^{(d)} \cdot G_{l,j}^{(d)} \right| \right]. \tag{3}$$

To further prove the proposed gradient-based activation importance metric, we approximate a neuron's contribution by evaluating the change in loss when the neuron is removed. Thus, we apply a first-order Taylor Expansion (Molchanov et al. 2017) to estimate the impact of ablating individual neurons. Let $H^{(d)}$ represent neurons in layer l excluding the j-th neuron. Assuming the output of each neuron contributes independently to the loss, the change in loss due to removing the j-th neuron can be expressed as:

$$\begin{aligned} \left| \Delta \mathcal{L}^{(d)}(h_{l,j}^{(d)}) \right| = \Big| & \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)} = 0) \\ & - \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)}) \Big|, \end{aligned} \tag{4}$$

where $\mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)} = 0)$ is the loss on domain d when the j-th neuron is removed, and $\mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)})$ is the loss when it is retained. Then, applying a first-order Taylor approximation, this change in the loss for domain d can be estimated as:

$$\begin{aligned} \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)}) = & \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)} = 0) \\ & + \frac{\partial \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)})}{\partial h_{l,j}^{(d)}} + R_1(h_{l,j}^{(d)}), \end{aligned} \tag{5}$$

where $R_1(h_{l,j})$ is the Lagrange remainder term associated with the approximation for domain d:

$$R_1(h_{l,j}^{(d)}) = \frac{\partial^2 \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)})}{\partial^2 \delta h_{l,j}^{(d)}} (h_{l,j}^{(d)})^2, \tag{6}$$

where $\delta \in (0, 1)$. The first derivative of the loss function with respect to the neuron's output for domain d tends to become constant. Consequently, during the final stages of training for domain d, the second-order term approaches zero. Therefore, by neglecting the remainder term, the importance evaluation function for a neuron with respect to domain d can be approximated as:

$$|\Delta \mathcal{L}^{(d)}(h_{l,j}^{(d)})| \approx \left| \frac{\partial \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)})}{\partial h_{l,j}^{(d)}} \cdot h_{l,j}^{(d)} \right|. \tag{7}$$

Thus, the importance score $I_{l,j}^{(d)}$ serves as an approximation to the contribution of neuron j in layer l to the MDMT loss in domain d:

$$\begin{aligned} I_{l,j}^{(d)} &= \mathbb{E}(\mathbf{x}^d, \mathbf{y}^d) \sim D_d \left[\left| A_{l,j}^{(d)} \cdot G_{l,j}^{(d)} \right| \right] \\ &\approx \left| \frac{\partial \mathcal{L}^{(d)}(H^{(d)}, h_{l,j}^{(d)})}{\partial h_{l,j}^{(d)}} \cdot h_{l,j}^{(d)} \right|, \end{aligned} \tag{8}$$

which serves as the foundation for multi-domain consensus-aligned neuron selection, as detailed in the next subsection.

MI-based Multi-Domain Consensus-Aligned Neuron Selection

After computing MDMT task-relevant importance scores $I_{l,j}^{(d)}$, we identify consensus-aligned neurons that consistently exhibit high relevance to translation across all domains. These neurons capture domain-invariant translation mechanisms while retaining sensitivity to domain-specific nuances, thereby enabling generalization to unseen domains and efficient fine-tuning for seen ones.

To quantify the relationship between a neuron's importance and the domain identity, we employ MI measurement. MI can precisely quantify the statistical dependencies between variables, which enables the assessment of whether neurons are multi-domain consensus-aligned. Specifically, we first discretize the continuous importance scores into a set of fixed bins indexed by i to facilitate probability estimation over domain feature. For the j-th neuron in the l-th FFN layer, we then estimate its MI with the domain label d as follows:

$$\begin{aligned} \mathrm{MI}_{l,j} &= \sum_{i} \sum_{d \in D} H(I_{l,j}^{(d)}) + H(d) - H(I_{l,j}^{(d)}, d) \\ &= \sum_{i} \sum_{d \in D} p(I_{l,j}^{(d)} = i, d) \log\Big(\frac{p(I_{l,j}^{(d)} = i, d)}{p(I_{l,j}^{(d)} = i) p(d)}\Big), \end{aligned} \tag{9}$$

where term $p(I_{l,j}^{(d)} = i, d)$ denotes the joint probability that the neuron's importance score $I_{l,j}^{(d)}$ falls into importance bin i for a data sample belonging to domain d. This captures the joint behavior between neuron importance and domain identity, reflecting how frequently a neuron exhibits a particular level of importance within a specific domain. The marginal probability $p(I_{l,j}^{(d)} = i)$ represents the overall distribution of the neuron's importance score across all domains, indicating how frequently its score falls into bin i regardless of the domain. The term $p(d)$ is the marginal probability of encountering a data sample from domain d.

To select our multi-domain consensus-aligned neurons $\mathcal{N}_{\text{MDCA}}$, we avoid selecting neurons with high MI for individual domains, as this can introduce domain-specific noise and lead to overfitting. Instead, we identify neurons that consistently exhibit high MI across all domains:

$$\mathcal{N}_{\text{MDCA}} = \{(l, j) \mid \min \mathrm{MI}_{l,j} \geq \gamma\}, \tag{10}$$

where γ is a threshold, and ensures that a neuron is selected only if its MI with the domain label is at least γ in every domain $d \in D$. This allows us to identify neurons that are robustly aligned with multi-domain characteristics across the entire set of domains D.

Neuron-Efficient Fine-Tuning

We propose a Neuron-Efficient Fine-Tuning strategy that only updates the multi-domain consensus-aligned neurons.

Formally, let $\mathcal{M}_\theta$ denotes a LLM with parameters θ, we freeze all parameters except those directly associated with neurons in $\mathcal{N}_{\text{MDCA}}$. The FFN layer contains three modules, up projection, down projection and gate projection. For each module m, let $W_m \in \mathbb{R}^{in \times out}$ denotes the weight matrix, where in is the input dimension to the module and out is the output dimension. The number of neurons is identical to out. To enable selective gradient updates, we construct a binary mask $M \in \mathbb{R}^{in \times out}$, such that:

$$M_{:i} = \begin{cases} 1, & \text{if } i \in \mathcal{N}_{\text{MDCA}} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This mask is applied during the backward pass to suppress gradients for all non-selected neurons. Specifically, the gradient update for each weight matrix W_m is modified as:

$$\nabla W_m \leftarrow \nabla W_m \odot M, \tag{12}$$

where $\odot$ denotes element-wise multiplication. Consequently, only the parameters corresponding to the multi-domain consensus-aligned neurons in $\mathcal{N}_{\text{MDCA}}$ are updated.

Experiments

Experimental Setups

Dataset We conducted experiments in 2 translation directions: German-English (De⇒En) and Chinese-English (Zh⇒En). For De⇒En, we used a multi-domain dataset from Aharoni and Goldberg (2020). To validate the generalization capability of the proposed method, we set 4 seen domains (IT, Law, Medical and Subtitles) and 1 unseen domain (Koran). Similarly, For Zh⇒En, we use UM-Corpus (Tian et al. 2014), and also set 3 seen domains (Education, Spoken and Thesis) and 2 unseen domains (Science and Microblog). For neuron-based methods, we randomly sampled 10k data for neuron selection and further sampled 2k data from this set for neuron-efficient fine-tuning. And other baseline methods were fine-tuned on the full 10k data to ensure a fair comparison in terms of data volume.

Backbone Models We take LLaMA2-7B-Chat[1], LLaMA3.1-8B-Instruct[2], and Qwen2.5-7B-Instruct[3] as the backbone models for training.

Implementation Details For neuron-based methods, we select 1% neurons for fine-tuning. And for our CANEFT, the threshold γ (as defined in Eq. 10) is dynamically determined by selecting the top 1% of neurons with the highest MI scores across all domains as consensus-aligned neurons. For LoRA-based baselines, including LoRA, LoRA-MLP, DoRA, we set rank to 8. For LLaMA Pro, we add two adapter layers after layers 16 and 32. All experiments are executed on 8 NVIDIA A40 GPUs.

Baselines We compared our approach with several representative baselines. **Base Inference** performs zero-shot inference using the original LLMs without any fine-tuning or domain adaptation. **Full Fine-tuning** fine tunes all parameters of LLMs. **LoRA** and **LoRA-MLP** (**L-MLP**) (Hu et al. 2022) apply low-rank adaptation to enable PEFT. While LoRA adds low-rank matrices to all modules and layers, LoRA-MLP restricts this to only the FFN modules, further reducing the number of trainable parameters. **DoRA** (Liu

[1]https://huggingface.co/meta-llama/Llama-2-7b-chat

[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Methods	TP/AP	De ⇒ En												Zh ⇒ En									
		Seen								Unseen				Seen						Unseen					
		IT		Law		Med		Sub		Kor		Avg		Edu		Spo		The		Sci		Blog		Avg	
		B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C
LLaMA2-7B-Chat																									
Base	0/0	38.7	81.8	30.0	82.8	44.1	87.4	35.5	85.0	16.0	73.1	32.8	82.0	25.7	83.9	26.7	83.8	18.1	76.9	11.9	73.8	14.4	73.4	19.3	78.3
Full	7k/0	43.7	87.8	**47.9**	82.8	47.1	85.5	35.3	81.1	**20.5**	73.8	38.9	82.2	28.2	79.9	29.6	78.9	22.7	78.4	14.9	72.2	18.2	73.9	22.7	76.6
LoRA	20/20	46.7	87.4	35.8	80.1	49.9	87.7	33.5	82.5	15.3	69.6	36.2	81.4	28.2	82.0	28.6	82.1	21.6	77.1	16.3	74.3	15.3	73.8	22.0	77.8
L-MLP	12/12	46.4	86.9	40.8	82.5	49.7	88.0	33.7	82.5	14.4	70.2	37	82.0	27.3	82.7	28.3	82.4	21.6	77.7	15.5	74.2	18.7	76.7	22.2	78.7
DoRA	22/22	45.9	87.8	42.4	84.7	46.7	87.7	37.5	84.3	17.6	73.2	38.0	83.5	28.7	81.9	29.0	82.2	21.6	77.0	**16.6**	75.2	19.2	76.8	23.0	78.6
Pro	405/405	40.4	83.9	46.0	**84.9**	43.5	87.2	31.3	81.3	16.4	72.4	35.5	81.9	**29.3**	84.1	28.9	83.0	**23.5**	**80.0**	15.1	75.3	18.0	77.2	22.9	79.9
LAPE	91/0	39.3	80.6	30.4	80.5	40.5	85.8	29.0	76.2	15.5	73.1	30.9	79.2	21.4	82.8	22.0	82.0	15.4	76.5	13.2	73.9	14.5	73.5	17.3	77.7
RCN	91/0	25.5	84.1	30.8	82.0	33.6	85.8	40.4	86.1	14.6	72.7	28.9	82.1	25.1	76.9	25.0	74.0	17.7	75.1	10.7	70.1	12.2	70.8	18.1	73.3
CANEFT	91/0	**48.8**	**88.7**	43.8	84.4	**50.0**	**88.3**	**43.7**	**86.8**	19.2	**74.1**	**40.5**	**84.5**	28.9	**84.4**	**30.3**	**84.2**	22.6	78.5	16.4	**75.7**	**20.3**	**77.8**	**23.7**	**80.1**
LLaMA3.1-8B-Instruct																									
Base	0/0	48.1	87.8	38.3	84.4	46.4	87.4	44.4	87.4	19.3	75.5	39.3	84.5	30.7	85.2	31.1	86.0	21.6	80.6	15.4	77.4	17.1	79.1	23.1	81.6
Full	8k/0	47.2	84.3	**56.3**	**88.0**	51.4	88.7	42.1	87.2	**21.0**	75.5	43.6	84.7	31.3	85.3	33.4	84.6	25.5	**81.6**	16.2	**78.1**	19.4	79.2	25.1	81.7
LoRA	42/42	45.8	85.9	45.8	85.3	50.8	88.5	34.6	82.4	18.8	72.2	39.1	82.8	30.4	84.4	30.6	83.5	24.3	80.2	17.9	77.1	21.7	78.0	24.9	80.6
L-MLP	29/29	43.2	85.4	48.4	85.5	47.5	87.9	34.9	82.9	18.9	73.8	38.5	83.1	30.3	84.5	30.2	83.5	**25.8**	80.6	16.6	77.1	21.5	78.1	24.8	80.7
DoRA	44/44	44.3	85.5	50.0	85.9	49.3	88.4	35.7	83.1	19.5	73.9	39.7	83.3	30.4	84.4	30.5	83.5	24.5	80.1	16.7	76.1	21.1	78.2	24.6	80.4
Pro	436/436	44.7	85.5	50.9	86.1	49.8	88.4	34.8	82.6	20.2	74.2	40.0	83.3	29.3	84.1	28.9	83.0	23.5	80.0	17.0	77.5	20.3	78.4	23.8	80.6
LAPE	146/0	46.2	87.6	42.9	85.3	44.9	87.6	41.8	87.3	18.1	73.9	38.7	84.3	28.8	86.3	28.6	86.2	21.0	80.5	15.9	76.4	20.1	77.9	22.8	81.4
RCN	146/0	41.7	83.8	40.6	84.5	47.7	87.5	42.6	86.2	20.6	75.8	38.6	83.5	31.2	85.9	30.9	85.7	22.2	81.0	14.2	75.8	15.0	76.6	22.7	81.0
CANEFT	146/0	**54.8**	**90.7**	50.9	87.1	**52.3**	**90.3**	**46.3**	**88.0**	20.9	**75.9**	**45.0**	**86.4**	**34.1**	**86.5**	**34.8**	**86.5**	25.6	81.5	**18.9**	77.7	**22.1**	**79.4**	**27.1**	**82.3**
Qwen2.5-7B-Instruct																									
Base	0/0	54.3	90.5	42.6	86.2	50.3	89.5	44.0	87.8	19.4	75.1	42.1	85.8	33.9	86.4	34.2	86.4	23.4	81.1	18.3	77.9	19.3	79.1	25.8	82.1
Full	7k/0	53.2	87.8	**53.5**	85.9	**53.1**	89.9	42.2	87.1	21.2	75.6	44.6	85.2	35.1	**87.2**	36.8	87.3	26.6	81.3	20.6	78.3	21.5	79.2	28.1	82.6
LoRA	21/21	46.0	85.9	47.1	85.5	47.4	87.8	36.1	83.4	18.6	74.4	39.0	83.4	31.7	84.9	31.8	84.1	25.8	80.6	21.1	78.0	22.4	78.8	26.5	81.2
L-MLP	12/12	45.9	85.9	47.0	85.4	47.6	87.8	35.8	83.3	18.6	74.3	38.9	83.3	31.2	85.0	31.5	83.9	25.8	80.7	21.1	77.9	22.7	78.8	26.4	81.2
DoRA	22/22	45.4	85.8	47.2	85.4	47.9	87.8	36.1	83.3	18.8	74.4	39.0	83.3	31.5	84.9	31.8	84.0	25.7	80.6	20.9	77.9	21.8	78.8	26.3	81.2
Pro	466/466	45.7	85.7	50.6	86.7	50.0	88.4	35.0	82.9	19.7	74.2	40.2	83.5	31.0	84.8	31.3	83.7	25.6	80.6	18.5	78.1	21.0	79.1	25.4	81.2
LAPE	127/0	52.7	89.7	45.0	85.8	49.0	89.9	42.9	88.0	17.3	74.7	41.3	85.6	32.9	87.0	32.7	87.2	23.9	81.3	17.8	77.5	19.4	79.3	25.3	82.4
RCN	127/0	50.5	89.1	45.1	86.3	50.3	88.9	40.9	85.4	17.6	73.9	40.8	84.7	33.7	86.1	35.0	86.1	26.9	82.1	17.5	76.2	18.6	78.2	26.3	81.7
CANEFT	127/0	**56.5**	**91.4**	49.8	**87.2**	52.3	**90.1**	**47.2**	**88.3**	**21.5**	**76.1**	**45.5**	**86.6**	**36.9**	87.1	**38.0**	**87.4**	**27.2**	**82.2**	**22.7**	**78.9**	**23.2**	**79.9**	**29.6**	**83.1**

Table 1: Translation performance of 3 distinct models across IT, Law, Medical (Med), Subtitles (Sub), Koran (Kor), Education (Edu), Spoken (Spo), Thesis (The), Science (Sci), and Microblog (Blog) domains, evaluated using BLEU (B) and COMET (C) scores. "TP" denotes Trainable Parameters and "AP" refers to Additional Parameters, both measured in **millions**. The **best** results for each metric and domain are bolded, and the second-best are underlined.

et al. 2024) decouples model parameters into magnitude and direction components, updating only the directional component to achieve efficient domain adaptation. **LLaMA Pro** (**Pro**) (Wu et al. 2024) extends LLM depth by inserting identity-initialized transformer blocks, and fine-tuning only these additional layers using domain-specific data. For neuron-based method, we adopt **LAPE** (Tang et al. 2024) as a representative baseline. This method leverages neuron activation probability entropy statistics to detect and fine-tune domain-specific neurons. **Random Chosen Neurons** (**RCN**) fine-tunes a randomly selected subset of neurons, comparable in number to those selected by our CANEFT.

Main Results

Table 1 presents the performance of our method against several PEFT baselines across both seen and unseen domains, evaluated using *SacreBLEU*[4] (B) and COMET[5] (C).

Overall Performance Our CANEFT consistently delivers the best or second-best performance across all models and domains, underscoring its effectiveness in multi-domain translation. With Qwen2.5, it achieves 45.5 BLEU and 86.6 COMET on De⇒En and 29.6 BLEU and 83.1 COMET on Zh⇒En, surpassing the best-performing baseline by +1.2 BLEU and +0.7 COMET. Comparable gains are observed on LLaMA2 and LLaMA3.1, where CANEFT achieves an average improves by +1.4-1.6 BLEU and +1-1.7 COMET on De⇒En, and +0.7-2 BLEU and +0.2-0.6 COMET on Zh⇒En over the strongest competing baselines. Importantly, while full fine-tuning is the strongest baseline on average and achieves results comparable to CANEFT across most domains, CANEFT further surpasses it while updating only 1% of the parameters.

CANEFT also demonstrated substantial gains on unseen domains. For example, in Zh⇒En, CANEFT showed enhanced robustness when handling linguistically diverse domains like Science (formal, technical) and Microblog (informal, colloquial). These domains represent significant textual distribution divergences from the training data, underscoring the enhanced robustness of our method in handling linguistic style variations and cross-domain adaptation.

Notably, LLaMA Pro slightly outperforms CANEFT in the Law domain (e.g., +0.8 BLEU on Qwen2.5), likely due to its 4 times additional parameters capturing legal-specific

[4]BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a

[5]https://huggingface.co/Unbabel/wmt22-cometkiwi-da

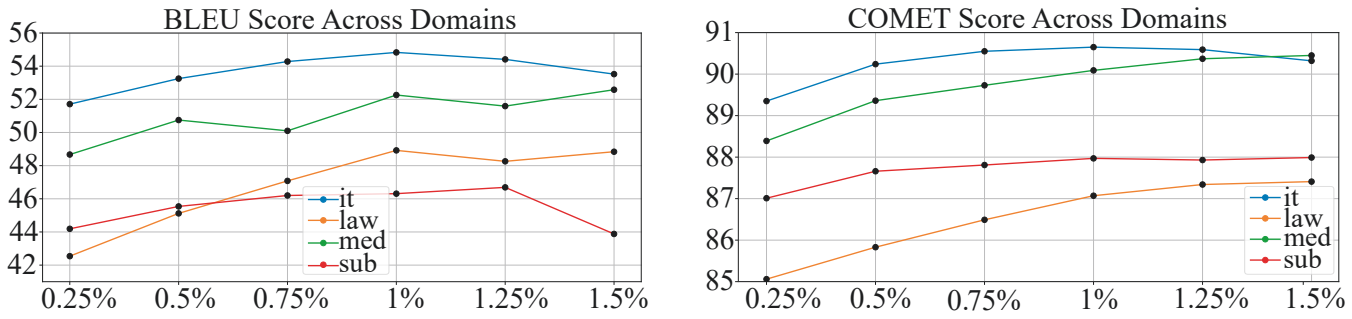

Figure 2: The impact of fine-tuning different multi-domain consensus-aligned neuron ratios on BLEU (left) and COMET (right) values with LLaMA3.1-8B-Instruct. In both plots, the x-axis shows neuron ratios, and the y-axis shows evaluation scores.

patterns. However, this comes at the expense of generalization, as LLaMA Pro suffers substantial drops in IT (-8.6 BLEU) and Subtitles (-9 BLEU) compared to Base Inference. Similarly, most PEFT baselines show uneven domain gains, improving in some domains while degrading in others. This reflects parameter interference during multi-domain fine-tuning and underscores the risk of catastrophic forgetting. In contrast, our method maintains robust performance across all domains, highlighting its advantage in balancing domain specialization with generalization.

Table 1 also shows that CANEFT introduces no additional parameters and updates only a minimal set of critical neurons. Although PEFT methods like LoRA and DoRA involve fewer trainable parameters, their performance across domains is significantly inferior, with some even underperforming base inference. In general, CANEFT strikes a superior balance between efficiency and translation quality. Moreover, CANEFT significantly outperforms the RCN and LAPE across all domains, demonstrating that the selected neurons capture meaningful multi-domain consensus-aligned information rather than arbitrary features. Furthermore, the consistent gains across 3 distinct backbones further demonstrate that our method is model-agnostic and applicable to various LLM-based translation frameworks.

Ablation Study

***w/o* MDMTN**: This variant computes MI between neurons and domain labels without identifying MDMT task-relevant neurons. As a result, MI is estimated over a larger and noisier neurons, making it harder to isolate meaningful consensus-aligned signals. As shown in Table 2, this leads to a noticeable performance drop across domains. These results underscore the importance of task-specific filtering prior to MI computation. In the absence of this filtering, the selected neurons include features not relevant to the task, which degrades the quality of adaptation.

***w/o* MDCAN**: This variant omits the multi-domain consensus-aligned neuron selection and directly fine-tunes the top 1% of neurons based solely on importance scores, resulting in a clear performance drop. While MDMTN ensures task relevance, not all identified neurons could reach a consensus alignment across domains. Without MDCAN, fine-tuning a less-refined neuron set increasing risks of parameter interference across domains and dilutes domain-invariant signals. The MI-based MDCAN step is thus essential for identifying a harmonized neuron subset that enables robust and generalizable adaptation across domains.

Method	IT	Law	Medical	Subtitles
CANEFT	54.8	50.9	52.3	46.3
w/o MDMTN	48.4	44.2	47.3	40.2
w/o MDCAN	46.2	40.6	44.9	41.8

Table 2: This table show ablation study and report BLEU scores in De⇒En with LLaMA3.1-8B-Instruct.

Impact of different neuron ratio

We investigate how varying the proportion of selected multi-domain consensus-aligned neurons (from 0.25% to 1.5%) influences performance using LLaMA3.1-8B-Instruct.

As shown in Figure 2, increasing the ratio of selected neurons consistently improves translation quality across both BLEU and COMET metrics. Notably, the IT and Law domains exhibit the most substantial improvements, with BLEU gains of 3–6 points, suggesting that incorporating more consensus-aligned neurons facilitates more effective cross-domain knowledge transfer. COMET scores follow a similar upward trend, reflecting enhanced semantic adequacy and fluency. However, performance plateaus beyond 0.75% and even degrades when over 1.25% in most domains, suggesting that most informative neurons have already been leveraged, and additional parameters contribute marginal returns. These results confirm that fine-tuning only 0.5%–1% of well-chosen neurons is sufficient to achieve strong multi-domain translation performance while avoiding unnecessary parameter updates.

Gradient changes between randomly selected and consensus-aligned neurons

To assess the relevance of the selected consensus-aligned neurons in MDMT, we analyze their behavior from a gradient perspective. We compare the mean absolute parameter changes in LLaMA3.1-8B's FFN components (*gate_proj*, *up_proj*, *down_proj*) under two neuron selection strategies:

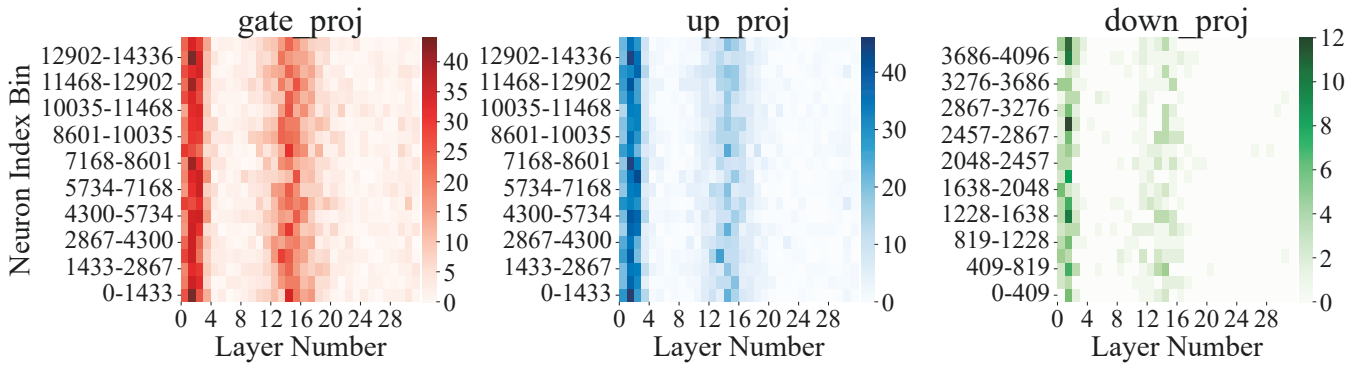

Figure 3: Distribution of multi-domain consensus-aligned neurons across layers within the FFN's *gate_proj*, *up_proj* and *down_proj* modules of LLaMA3.1-8B-Instruct. In each plots, the x-axis denotes the layer number, and the y-axis corresponds to neuron index bins, derived by segmenting the full range of neuron indices into 15 divisions.

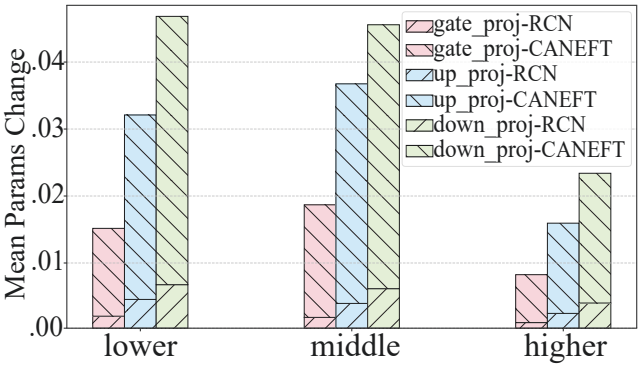

Figure 4: Gradient changes in consensus-aligned neurons (CANEFT) and randomly chosen neurons (RCN) within the FFN's *gate_proj*, *up_proj* and *down_proj* modules of LLaMA3.1-8B-Instruct. Layers are grouped by depth into lower, middle, and higher sections. In each bar, the lower segment represents the gradient changes of randomly chosen neurons, while the upper segment corresponds to those of consensus-aligned neurons.

random selection and multi-domain consensus-aligned neuron selection. Model layers are grouped into three ranges: lower (layers 0–10), middle (layers 11–20), and higher (layers 21–32). As shown in Figure 4, consensus-aligned neurons consistently exhibit larger gradient updates across all components and layers.

These substantial gradient differentials suggest that consensus-aligned neurons possess higher optimization potential for MDMT. Our selection method effectively identifies parameters most responsive to multi-domain adaptation, enabling the LLM to refine these neurons for improved translation performance. Such neurons undergo more meaningful, task-specific transformations, whereas randomly selected neurons display weaker and less focused changes. Moreover, the pronounced shifts in *up_proj* and *down_proj* highlight their pivotal role in domain adaptation, underscoring their importance in facilitating deeper integration of domain knowledge into the model's representations.

Distribution of consensus-aligned neurons

To visually demonstrate the distribution of multi-domain consensus-aligned neurons across different layers in LLaMA3.1-8B-Instruct. Figure 3 provides heatmaps of the selected neurons across the 32 layers for each of the three FFN projections. The selected neurons are unevenly distributed, with higher density in lower and middle layers particularly in *gate_proj* and *up_proj*. In contrast, *down_proj* neurons are selected sparsely, indicating a relatively limited role in cross-domain generalization. This pattern aligns with the understanding that lower layers capture general syntactic and semantic features beneficial for transfer, while higher layers encode more domain-specific information thus contributing fewer consensus-aligned neurons.

The salience of *gate_proj* and *up_proj* points to their critical role in regulating information flow and enriching representations for multi-domain processing. *gate_proj* modulates the activation gate, while *up_proj* expands hidden representations. The high density of selected neurons in these components suggests that consensus knowledge are more effectively encoded and propagated during the gating and expansion phases, rather than in the dimensionality reduction stage (*down_proj*). This is further supported by Figure 4, where these components show stronger fine-tuning signals, indicating their structural and functional significance in encoding transferable knowledge.

Conclusion

We propose a neuron-efficient fine-tuning framework for MDMT that selectively updates consensus-aligned neurons, identified by maximizing MI between neuron behavior and domain features. This method improves translation quality and mitigates parameter interference and domain-specific overfitting. Unlike existing PEFT methods that require additional parameters, our method generalizes well to unseen domains with no extra parameters, and achieves SOTA performance across 10 domains on 3 LLMs, highlighting the promise of neuronal domain adaptation in LLMs.

Ethics Statement

This research adheres to a strict ethical framework as it does not involve any ethical issues. The data constructed for this research is derived solely from open-source data, and the large language models employed in this study follows their declared licenses. I have fully informed the participants of all instructions, to ensure they are fully aware and consenting to participate in this work.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. U24A20334, 62366027, U21B2027, 62266027), the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202303AP140008, 202203AA080004, 202302AD080003, 202401BC070021), the General Projects of Basic Research in Yunnan Province (Grant Nos. 202201BE070001-021).

References

Aharoni, R.; and Goldberg, Y. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7747–7763. Online: Association for Computational Linguistics.

Alves, D.; Guerreiro, N.; Alves, J.; Pombal, J.; Rei, R.; de Souza, J.; Colombo, P.; and Martins, A. 2023. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11127–11148. Singapore: Association for Computational Linguistics.

Aycock, S.; and Bawden, R. 2024. Topic-guided example selection for domain adaptation in llm-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 175–195.

Chen, Y.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17817–17825.

Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2022. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8493–8502.

Eschbach-Dymanus, J.; Essenberger, F.; Buschbeck, B.; and Exel, M. 2024. Exploring the effectiveness of LLM domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, 610–622.

Ghazvininejad, M.; Gonen, H.; and Zettlemoyer, L. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

He, Z.; Liang, T.; Jiao, W.; Zhang, Z.; Yang, Y.; Wang, R.; Tu, Z.; Shi, S.; and Wang, X. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12: 229–246.

Hersche, M.; Camposampiero, G.; Wattenhofer, R.; Sebastian, A.; and Rahimi, A. 2024. Towards Learning to Reason: Comparing LLMs with Neuro-Symbolic on Arithmetic Relations in Abstract Reasoning. *arXiv preprint arXiv:2412.05586*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, T.; Zhang, P.; Yang, B.; Xie, J.; Wong, D.; and Wang, R. 2024. Large Language Model for Multi-Domain Translation: Benchmarking and Domain CoT Fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5726–5746.

Hu, Z.; Wang, L.; Lan, Y.; Xu, W.; Lim, E.-P.; Bing, L.; Xu, X.; Poria, S.; and Lee, R. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5254–5276. Singapore: Association for Computational Linguistics.

Huang, K.; Mo, F.; Zhang, X.; Li, H.; Li, Y.; Zhang, Y.; Yi, W.; Mao, Y.; Liu, J.; Xu, Y.; et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Jiang, H.; Fang, J.; Zhang, T.; Zhang, A.; Wang, R.; Liang, T.; and Wang, X. 2024. Neuron-level sequential editing for large language models. *arXiv preprint arXiv:2410.04045*.

Leng, Y.; and Xiong, D. 2025. Towards Understanding Multi-Task Learning (Generalization) of LLMs via Detecting and Exploring Task-Specific Neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2969–2987.

Li, B.; Luo, J.; Briakou, E.; and Cherry, C. 2025. Leveraging Domain Knowledge at Inference Time for LLM Translation: Retrieval versus Generation. *arXiv preprint arXiv:2503.05010*.

Li, R.; Wang, X.; and Yu, H. 2020. MetaMT, a Meta Learning Method Leveraging Multiple Domain Data for Low Resource Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8245–8252.

Li, Y.; Zhu, Y.; Yan, T.; Fan, S.; Wu, G.; and Xu, L. 2024. Knowledge Editing for Large Language Model with Knowledge Neuronal Ensemble. *arXiv preprint arXiv:2412.20637*.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Mao, C.; Gao, X.; Song, R.; He, S.; Gao, S.; Liu, K.; and Yu, Z. 2025. Multilingual Knowledge Graph Completion via Efficient Multilingual Knowledge Sharing. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and

Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 10882–10896. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; and Kautz, J. 2017. Pruning Convolutional Neural Networks for Resource Efficient Inference. In *International Conference on Learning Representations*.

Moslem, Y.; Haque, R.; Kelleher, J.; and Way, A. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 227–237.

Niu, J.; Liu, A.; Zhu, Z.; and Penn, G. 2024. What does the Knowledge Neuron Thesis Have to do with Knowledge? In *The Twelfth International Conference on Learning Representations*.

Pang, J.; Ye, F.; Wong, D. F.; Yu, D.; Shi, S.; Tu, Z.; and Wang, L. 2025. Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. *Transactions of the Association for Computational Linguistics*, 13: 73–95.

Pham, M.; Crego, J. M.; and Yvon, F. 2021. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9: 17–35.

Rai, D.; and Yao, Z. 2024. An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7174–7193.

Saunders, D. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75: 351–424.

Sievers, B.; Welker, C.; Hasson, U.; Kleinbaum, A. M.; and Wheatley, T. 2024. Consensus-building conversation leads to neural alignment. *Nature Communications*, 15(1): 3936.

Song, R.; He, S.; Jiang, S.; Xian, Y.; Gao, S.; Liu, K.; and Yu, Z. 2024. Does Large Language Model Contain Task-Specific Neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7101–7113.

Tan, S.; Wu, D.; and Monz, C. 2024. Neuron Specialization: Leveraging Intrinsic Task Modularity for Multilingual Machine Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6506–6527.

Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, W. X.; Wei, F.; and Wen, J.-R. 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)*, 5701–5715.

Tian, L.; Wong, D. F.; Chao, L. S.; Quaresma, P.; Oliveira, F.; Lu, Y.; Li, S.; Wang, Y.; and Wang, L. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1837–1842. Reykjavik, Iceland: European Language Resources Association (ELRA).

Vilar, D.; Freitag, M.; Cherry, C.; Luo, J.; Ratnakar, V.; and Foster, G. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15406–15427.

Voita, E.; Ferrando, J.; and Nalmpantis, C. 2024. Neurons in Large Language Models: Dead, N-gram, Positional. In *Findings of the Association for Computational Linguistics ACL 2024*, 1288–1301.

Wu, C.; Gan, Y.; Ge, Y.; Lu, Z.; Wang, J.; Feng, Y.; Shan, Y.; and Luo, P. 2024. LLaMA Pro: Progressive LLaMA with Block Expansion. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6518–6537. Bangkok, Thailand: Association for Computational Linguistics.

Xie, W.; Feng, Y.; Gu, S.; and Yu, D. 2021. Importance-based Neuron Allocation for Multilingual Neural Machine Translation. 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)*, 5725–5737.

Xu, H.; Zhan, R.; Ma, Y.; Wong, D. F.; and Chao, L. S. 2025. Let's Focus on Neuron: Neuron-Level Supervised Fine-tuning for Large Language Model. In *Proceedings of the 31st International Conference on Computational Linguistics*, 9393–9406.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, J.; Hong, H.; Liu, F.; Wang, X.; and Su, J. 2024a. DragFT: Adapting Large Language Models with Dictionary and Retrieval Augmented Fine-tuning for Domain-specific Machine Translation. *arXiv preprint arXiv:arXiv:2402.15061v2*.

Zheng, J.; Hong, H.; Liu, F.; Wang, X.; Su, J.; Liang, Y.; and Wu, S. 2024b. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

Zhu, S.; Pan, L.; Li, B.; and Xiong, D. 2024. 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)*, 12135–12148.