Layer-aware Dual-directional Modulation for Low-resource Language Machine Translation

Siqi Zhang[1,2], Ran Song[1,2]*, Shuting Jiang[1,2], Yuxin Huang[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
zhangsiqi.mo@foxmail.com,
{song_ransr, shuting_jiang22, huangyuxin2004}@163.com, ztyu@hotmail.com

Abstract

Although Large Language Models (LLMs) have achieved remarkable success in Machine Translation (MT), a significant performance gap persists between high-resource and low-resource languages due to imbalanced pre-training data. In this paper, we first investigate the internal mechanisms driving this performance disparity from a layer-wise perspective. We propose a metric termed *Activation Disparity* (ΔR) to quantify the activation divergence between high- and low-resource MT. Based on this metric, we distinguish between Task-Adaptive Layers (TAL, $\Delta R > 0$) that encode task-specific signals and Legacy-Inert Layers (LIL, $\Delta R < 0$) dominated by pre-trained bias. Leveraging this finding, we propose the **L**ayer-**a**ware **D**ual-directional **M**odulation (**LaDM**). Integrated with Low-Rank Adaptation (LoRA), LaDM employs a sparse strategy to bidirectionally modulate optimization dynamics. Specifically, it amplifies contributions from TAL to accelerate feature consolidation while inhibiting LIL to dampen misaligned legacy biases. Extensive experiments on Chinese-to-seven low-resource language translation using Llama-3.1, Qwen2.5, and Gemma-2 demonstrate that LaDM significantly outperforms standard LoRA fine-tuning, achieving an average improvement of 1.73 spBLEU. Code is available at https://github.com/zzssqqq/LaDM.

1 Introduction

Large Language Models (LLMs) have demonstrated excellent performance in Machine Translation (MT) (Jiao et al., 2023; Zhang et al., 2025). This success stems from the cross-lingual capabilities that emerge during the pre-training phase from massive data (Wang et al., 2024; Hua et al., 2024; Conneau et al., 2020). Based on this, LLMs can achieve high-quality results with parameter-efficient fine-tuning (PEFT) (Zhu et al., 2024; Chen et al., 2023a; Jiang et al., 2026).

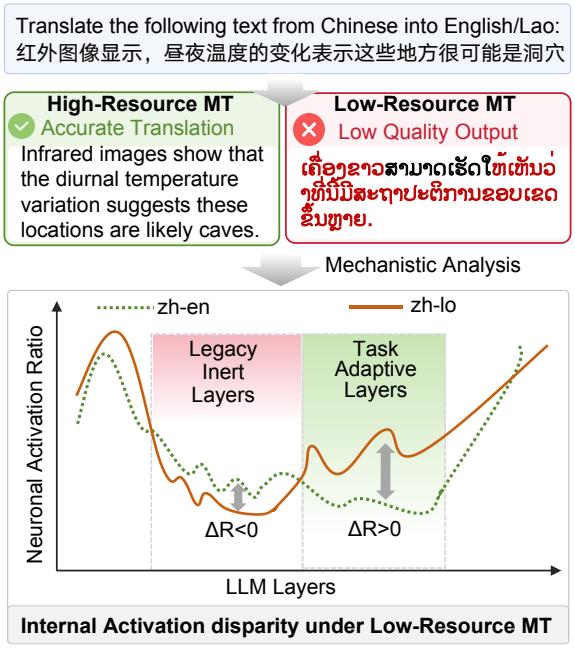

Figure 1: The Phenomenon of activation disparity in MT. Compared to the High-Resource Anchor (zh-en), the Low-Resource MT (zh-lo) exhibits a significant divergence. We term this gap Activation Disparity (ΔR). The intermediate layers show a negative disparity ($\Delta R < 0$), identifying them as Legacy-Inert Layers (LIL) dominated by pre-trained bias. Conversely, deeper layers show a positive disparity ($\Delta R > 0$), identifying them as Task-Adaptive Layers (TAL) that struggle to encode fragile task-specific signals.

Although LLMs have achieved remarkable performance in MT, a significant gap persists between high- and low-resource languages (Li et al., 2025; Hendy et al., 2023). This gap is attributed to imbalances within the pre-training datasets (Luo et al., 2025). From an interpretability perspective, such disparity may manifest as representational divergence across specific layers or components of LLMs. Consequently, a critical question arises: How to identify the divergence across LLM layers

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20988–21000

between high and low-resource languages MT?

Extensive works have established that different LLM layers exhibit distinct functional roles (Yang et al., 2025; Belrose et al., 2025; Song et al., 2025). These works reveal that layers specialize in capturing distinct hierarchical linguistic and semantic features. From the perspective of layer-wise specialization, we investigate the functional divergence between low- and high-resource MT. We conduct a preliminary layer-wise analysis and observe a clear divergence in activation patterns between low- and high-resource MT. As illustrated in Figure 1, we observe that the model exhibits a distinct internal activation disparity under low-resource settings. To quantify the activation divergence between high- and low-resource MT, we introduce a metric named Activation Disparity (ΔR). Based on this metric, we categorize layers into Task-Adaptive Layers (TAL, $\Delta R > 0$), which capture task-specific information, and Legacy-Inert Layers (LIL, $\Delta R < 0$), which retain a dominance of pre-trained bias.

Leveraging these findings, we propose the Layer-aware Dual-directional Modulation (LaDM). LaDM introduces learnable latent gating parameters to derive differentiable modulation factors, which adaptively regulate the magnitude of layer outputs. Integrated with Low-Rank Adaptation (LoRA), structurally differentiates optimization dynamics. LaDM amplifies TAL to consolidate critical features, while inhibiting LIL to dampen misaligned pre-trained biases. This distinct control strategy reduces interference, explicitly empowering the adapters to enhance task-specific representations. Extensive experiments on Chinese-to-Target translation across seven low-resource pairs using Llama-3.1, Qwen2.5, and Gemma-2 confirm LaDM's effectiveness, demonstrating an average improvement of 1.73 spBLEU points over the standard LoRA. Our method significantly outperforms PEFT baselines, effectively validating its capacity to overcome pre-training bias in data-scarce scenarios. Our contributions are summarized as follows:

- We propose a detection method based on Activation Disparity (ΔR) and observe a functional divergence across layers. We identify that while TAL exhibit high engagement, LIL show an activation deficit and are dominated by legacy biases.
- we propose the LaDM and integrate it into mainstream PEFT methods. Our framework strengthens task-relevant representations in TAL while implementing an inhibitory mechanism to suppress legacy biases in LIL.
- We conduct extensive experiments showing our method outperforms uniform and weight-decomposed fine-tuning baselines. Results on Llama-3.1, Qwen2.5, and Gemma-2 across extremely low-resource language pairs demonstrate that LaDM significantly enhances performance compared to standard PEFT methods.

2 Functional Localization of Layer in Low-Resource MT

This section locates the critical layers for the processing differences between high- and low-resource MT based on the analysis of layer divergence. We utilize neuron-level probing to pinpoint MT-sensitive components within attention and MLP blocks. And introduce Activation Disparity (ΔR) as a metric to quantify the relative difference in a layer's functional involvement between high- and low-resource MT.

2.1 Identifying MT-Sensitive Modules

To identify which neural components drive task adaptation during translation, we quantify the functional engagement of core sub-modules using the *Accumulated Activation Intensity* (I_m^l). Here, $m \in \{\text{MLP}, Q, K, V, O\}$ denotes the specific sub-module within layer l. Specifically, we employ a hook-based mechanism to extract the hidden states $h_m \in \mathbb{R}^{d_{\text{model}}}$. We calculate the accumulated intensity I_m^l by accumulating the proportion of activated neurons across the full translation context:

$$I_m^l = \sum_{t=1}^{T} \left(\frac{1}{d_{\text{dim}}} \sum_{i=1}^{d_{\text{dim}}} \mathbb{I}(h_{m,i,t} > 0) \right), \quad (1)$$

where T denotes the length of the generated sequence, d_{dim} represents the dimension of sub-module m, and $\mathbb{I}(\cdot)$ is the indicator function for active neurons. A higher I_m^l indicates that the component is persistently engaged in processing the linguistic information flow.

Based on this, we distinguish MT-sensitive modules by analyzing their activation patterns across the model's depth. We quantify this sensitivity by calculating the *Cross-Layer Activation Variance* (σ_m^2):

$$\sigma_m^2 = \frac{1}{L} \sum_{l=1}^{L} \left(\mathbb{E}_{x \sim \mathcal{D}} \left[I_m^l(x) \right] - \bar{I}_m \right)^2, \quad (2)$$

where $\mathbb{E}_{x\sim\mathcal{D}}[I_m^l(x)]$ denotes the expected activation intensity of module m at layer l over the dataset $\mathcal{D}$, and $\bar{I}_m$ represents the global average intensity across all L layers.

Finally, we identify the primary MT-sensitive module $\hat{m}$ by selecting the component that maximizes this cross-layer variance:

$$\hat{m} = \text{argmax}_{m\in\mathcal{M}}\sigma_m^2, \quad (3)$$

where $\mathcal{M}$ represents the set of candidate sub-modules defined previously. We target $\hat{m}$ for fine-tuning, identifying it as the most task-sensitive component across layers.

2.2 Locating Task-Adaptive and Legacy-Inert Layers

We propose a layer-wise categorization method based on neuron activation patterns. For each layer l, we measure the activation behavior of a target module $\hat{m}$ by defining the *activation ratio* $r^l(\mathbf{x})$ for a sequence $\mathbf{x}$ as the proportion of neurons in $\hat{m}$ that have positive cumulative activation:

$$r^l(\mathbf{x}) = \frac{1}{d_{\hat{m}}}\sum_{i=1}^{d_{\hat{m}}}\mathbb{I}\left(\sum_{t=1}^{T}h_{i,t,\hat{m}}^l > 0\right), \quad (4)$$

where $d_{\hat{m}}$ denotes the hidden dimension of module $\hat{m}$, and $h_{i,t,\hat{m}}^l$ represents the post-activation output of the i-th neuron at time step t in layer l. We then define the **Activation Disparity** (ΔR^l) to compare neuron engagement between low-resource MT ($\mathcal{D}_L$) and high-resource MT ($\mathcal{D}_H$):

$$\Delta R^l = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_L}\left[r^l(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_H}\left[r^l(\mathbf{x})\right]. \quad (5)$$

Based on ΔR^l, we categorize layers into two distinct functional types:

Task-Adaptive Layers (TAL, $\Delta R^l > 0$): These layers exhibit increased neuron engagement. The higher activation suggests that the model recruits more neurons to acquire new task-specific features required for the low-resource MT.

Legacy-Inert Layers (LIL, $\Delta R^l < 0$): These layers show reduced activation levels. This indicates a heavy reliance on general, pre-trained knowledge, suggesting that these layers fail to adapt effectively to the specific characteristics of the target task.

To validate this categorization, we further introduce the Entropy Gap (ΔH) based on the Singular Value Decomposition (SVD) of hidden states. Defined as $\Delta H = H(\mathcal{D}_L) - H(\mathcal{D}_H)$, a positive entropy gap ($\Delta H > 0$) suggests that representations in the low-resource setting are less concentrated than those in the high-resource setting. We take this as a sign that LIL are less effective at forming compact task-relevant representations.

2.3 Empirical Analysis

This section provides empirical evidence for the layer-wise differentiation of LLMs during low-resource MT.

Identification of MT-Sensitive Modules. As shown in Figure 2, different sub-modules exhibit distinct layer-wise activation profiles (I_m^l). The attention sub-modules (Q, K, V, O) remain relatively stable across layers. In contrast, the MLP sub-modules exhibit much larger variation. This difference suggests that MLPs are more sensitive to layer-wise functional changes, motivating our subsequent analysis of MLP sub-modules.

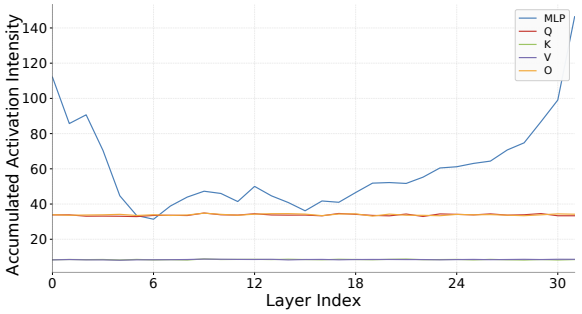

Figure 2: Layer-wise accumulated activation intensity (I_m^l) of Llama-3.1 sub-modules. The attention sub-modules (Q/K/V/O) remain relatively stable across layers, whereas the MLP sub-module exhibits much larger variation.

Activation and Entropy Patterns across TAL and LIL. Figure 3 shows the layer-wise variation in the neuron activation ratio (r^l), providing supporting evidence for the TAL/LIL categorization introduced in Section 2.2. As shown in Figure 3, TAL are associated with positive activation disparity ($\Delta R^l > 0$), indicating higher neuron engagement in the low-resource setting than in the high-resource anchor. In contrast, the middle layers identified as LIL show negative activation disparity ($\Delta R^l < 0$), reflecting lower neuron engagement in the low-resource setting.

As shown in Figure 4, ΔH in TAL remains relatively stable or slightly decreases. By contrast, the LIL region is associated with higher entropy gap values. This pattern suggests that representations in LIL are less concentrated than those in TAL under the low-resource setting.

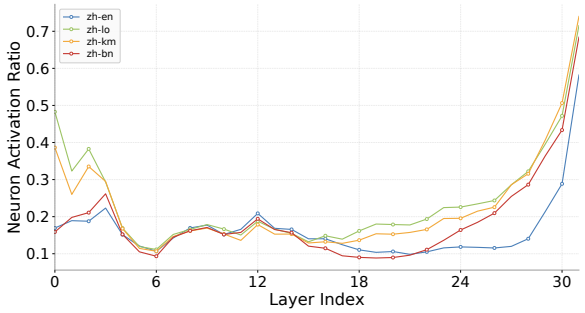

Figure 3: Layer-wise neuron activation ratio (r^l) across Llama-3.1 layers for several language pairs. Compared with the high-resource anchor (zh-en), low-resource settings tend to show lower activation in middle layers and higher activation in deeper layers.

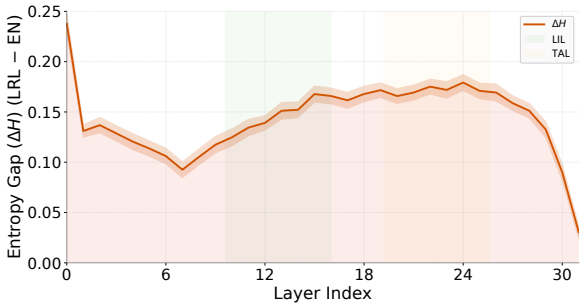

Figure 4: Layer-wise entropy gap (ΔH) across Llama-3.1 layers. Values represent the mean entropy difference between the low-resource language set (zh-lo, zh-km, zh-bn) and the high-resource anchor (zh-en). The LIL region (green) is associated with higher entropy gap values, whereas the TAL region (yellow) remains relatively stable.

3 Layer-aware Dual-directional Modulation Framework

Building upon the functional divergence identified in Section 2, we propose the **Layer-aware Dual-directional Modulation (LaDM)**. LaDM introduces a hierarchical control mechanism that adaptively regulates information flow based on layer-specific functional priors. We adopt a Top-k selection in which modulation is applied only to the k layers with the largest positive ΔR for amplification and the k layers with the most negative ΔR for inhibition, while intermediate layers remain unmodulated to preserve stability.

3.1 Latent Gating Parameters

To achieve differentiable control over layer engagement, we introduce a set of learnable latent gating parameters, denoted as $g_l \in \mathbb{R}$, for each selected layer l. Unlike direct scalar multipliers, g_l serves as an unconstrained control variable governing the potential of a layer's contribution.

We employ a category-aware initialization strategy for g_l based on the layer's functional grouping. Specifically, for layers in $\mathcal{S}_{\text{TAL}}$ (where $\Delta R^l > 0$), g_l is initialized to a near-neutral value. This positioning allows for flexible, bidirectional optimization, enabling the model to freely amplify or fine-tune these critical signals. In contrast, for layers in $\mathcal{S}_{\text{LIL}}$ (where $\Delta R^l < 0$), we adopt a conservative initialization strategy with a lower gate value. This differential setting functions as a soft constraint. This soft constraint dampens misaligned legacy biases, ensuring they do not interfere with the emergence of compensatory optimization dynamics.

3.2 Layer-wise Modulation Factors

The latent parameter g_l is transformed into an explicit modulation factor α^l. To prevent training instability, we bound the scaling factors using linear interpolation on the sigmoid activation $\sigma(g_l)$.

Let $\mathcal{R}_{\text{amplify}} = [b_{\min}, b_{\max}]$ be the amplification range and $\mathcal{R}_{\text{inhibit}} = [m_{\min}, m_{\max}]$ be the inhibition range. The modulation factor is defined as:

$$\alpha^l = \begin{cases} b_{\min} + (b_{\max} - b_{\min})\sigma(g_l) & l \in \mathcal{S}_{\text{TAL}}, \\ m_{\max} - (m_{\max} - m_{\min})\sigma(g_l) & l \in \mathcal{S}_{\text{LIL}}, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

For $\mathcal{S}_{\text{TAL}}$, higher gate activation drives α^l toward $b_{\max}$, boosting the layer's contribution. Crucially, for $\mathcal{S}_{\text{LIL}}$, we implement an inverse control mechanism: higher gate activation drives α^l toward $m_{\min}$, imposing stronger suppression. This allows the model to learn the optimal degree of dampening for these inert layers via gradient descent.

3.3 Integration with LoRA and Optimization Dynamics

We integrate the modulation factors into the LoRA architecture. LaDM acts as a structural gate for the output of the FFN sub-module. Formally, given the input $\mathbf{x}$, the frozen pre-trained weights W_0, and low-rank adapters B, A, the modulated forward pass is:

$$\mathbf{h}^l = \alpha^l \cdot \left(W_0\mathbf{x} + \frac{\gamma}{r} BA\mathbf{x} \right). \tag{7}$$

Optimization and Gradient Scaling. A critical advantage of LaDM lies in how the gating parameters reshape the optimization landscape. The scalar

α^l is jointly optimized with the adapter parameters $\theta_{\text{LoRA}} = \{A, B\}$. Considering the gradient dynamics:

$$\frac{\partial \mathcal{L}}{\partial \theta_{\text{LoRA}}} \propto \alpha^l \cdot \left(\frac{\partial \mathcal{L}}{\partial \mathbf{h}^l} \mathbf{x}^T \right). \tag{8}$$

This dependency establishes a self-regulating feedback loop that differentiates optimization strategies according to specific layer roles. For TAL where $\alpha^l > 1$, The modulation acts as signal amplification. By scaling up the forward pass, the effective learning rate for these layers is increased, accelerating the convergence of adapters to consolidate task-critical features. In contrast, For LIL where $0 < \alpha^l < 1$, The modulation imposes inertia inhibition. By scaling down the forward pass, LaDM explicitly suppresses the dominance of the frozen representations ($W_0\mathbf{x}$) which are identified as having negative functional divergence. This dampening effect reduces the interference of misaligned pre-trained biases on subsequent layers. Consequently, the optimization process is structurally encouraged to rely more heavily on the adaptive parameters θ_{LoRA} to reconstruct task-relevant representations, effectively shifting the layer's behavior from preserving legacy inertia to learning new task-specific alignments.

4 Experimental Analysis

4.1 Experimental Setup

Datasets and Benchmarks. We evaluate our method on Chinese-to-Target ($Zh \rightarrow X$) translation across seven language pairs. Training data is sourced from the Asian Language Treebank (ALT) (Riza et al., 2016) and NLLB (Costa-Jussà et al., 2022) corpora. To simulate a standardized low-resource setting, we subsample datasets to 20k sentence pairs per direction. Target languages are categorized into two groups based on resource scarcity: *Extreme Low-Resource* (Khmer [km], Lao [lo], Burmese [my]) and *Moderate Low-Resource* (Vietnamese [vi], Hindi [hi], Serbian [sr], Bengali [bn]). For evaluation, we employ the FLORES-200 (Costa-Jussà et al., 2022) as the test set.

Baselines and Backbones. To assess architectural generalization, we conduct experiments using three LLMs: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Team et al., 2024b), and Gemma-2-9b-it (Team et al., 2024a). We benchmark LaDM against standard PEFT methods, specifically LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024a). For LoRA, we examine two configurations: (1) **LoRA-MLP**, which adapts only the MLP modules ($W_{gate}, W_{up}, W_{down}$); and (2) **LoRA-All**, which applies adaptation to all linear projections, including attention ($W_{q,k,v,o}$) and MLPs. LaDM is added to the two LoRA configurations to isolate the gains attributed to our layer-aware modulation. For additional LoRA variants used in our experiments, we use the same training setup as LoRA-All unless otherwise specified, including the optimizer, learning rate, batch size, and target module scope.

Implementation Details. All models are fine-tuned on 8× NVIDIA A40 GPUs with a batch size of 4 and a learning rate of 1×10^{-5}. For LoRA-based methods, we use rank $r = 8$, LoRA scaling factor $\alpha = 16$, and dropout 0.05. DoRA follows the same target-module setting as LoRA-All. For LaDM, we apply modulation to the top-3 layers exhibiting the highest positive ΔR (for amplification) and the top-3 layers with the most negative ΔR (for inhibition). The latent gating parameters are initialized to $g = -0.3$ for amplification and $g = -1.0$ for inhibition. Given the sensitivity-tuned modulation ranges of $\mathcal{R}_{\text{amplify}} = [1.03, 1.07]$ and $\mathcal{R}_{\text{inhibit}} = [0.91, 0.96]$, this initialization sets the starting scaling factors to approximately $\alpha \approx 1.05$ and $\alpha \approx 0.95$, respectively. This ensures a conservative start closer to identity before optimization dynamics diverge.

Evaluation Metrics. Translation quality is assessed using spBLEU (Costa-Jussà et al., 2022) and ChrF (Popović, 2015). This combination provides a robust evaluation of both token-level lexical precision and character-level morphological fluency.

4.2 Main Results

The experimental results across seven language pairs and three model architectures consistently validate the efficacy of LaDM. As shown in Table 1, LaDM_{LA} achieves superior performance across almost all metrics, yielding average spBLEU gains of +3.51, +0.89, and +1.00 over the LoRA-All on Llama-3.1, Qwen2.5, and Gemma-2, respectively, and surpassing the stronger DoRA baseline. These consistent improvements suggest that LaDM's layer-aware modulation enables more effective task-specific adaptation across different backbones. A closer analysis on Llama-3.1 further shows that for language pairs where the base model is already strong (zh-hi and zh-bn), standard PEFT methods can underperform the zero-shot baseline,

Med	zh-lo	zh-my	zh-km	zh-vi	zh-sr	zh-hi	zh-bn	Avg
				Llama-3.1-8B-Instruct				
Base	1.35 / 17.2	2.73 / 18.66	5.29 / 23.61	27.29 / 45.98	13.97 / 31.67	**18.45 / 40.21**	**13.52 / 34.64**	11.80 / 30.28
LM	3.92 / 23.42	2.61 / 27.32	10.77 / 23.42	23.86 / 44.61	14.47 / 39.08	13.01 / 34.77	7.54 / 29.21	10.88 / 31.69
LA	5.55 / 28.06	3.29 / 28.68	10.08 / 31.47	28.23 / 48.31	15.67 / 40.12	13.89 / 35.40	10.40 / 31.98	12.44 / 34.86
DA	8.97 / 31.01	5.21 / 32.57	13.43 / 33.49	24.68 / 45.74	15.24 / 39.77	14.21 / 36.72	8.85 / 29.94	12.94 / 35.61
LaDM_{LM}	10.05 / 31.88	5.85 / 28.68	12.13 / 32.76	25.05 / 46.00	19.19 / 42.95	14.10 / 36.00	10.43 / 32.03	13.83 / 35.76
LaDM_{LA}	**13.83 / 35.33**	**8.67 / 39.63**	**13.99 / 34.65**	**29.15 / 49.13**	**20.16 / 43.76**	14.78 / 37.17	11.06 / 32.44	**15.95 / 38.87**
				Qwen2.5-7B-Instruct				
Base	1.84 / 15.73	1.96 / 23.29	3.93 / 21.93	19.42 / 42.26	7.60 / 29.96	2.53 / 14.80	4.49 / 20.94	5.97 / 24.13
LM	11.18 / 32.69	6.69 /32.88	10.57 / 29.17	27.85 / 48.42	14.16 / 38.18	13.27 /34.75	8.99 / 28.04	13.24 / 34.88
LA	10.78 / 32.40	6.03 / 32.41	12.70 / 33.21	28.49 / 48.66	14.07 / 38.09	13.29 / 34.65	8.82 / 27.20	13.45 / 35.23
DA	11.94 / 31.01	6.90 / 32.57	13.49 / 33.49	27.42 / 45.74	14.08 / 39.77	**13.71 / 36.72**	8.10 / 29.94	13.66 / 35.61
LaDM_{LM}	12.29 / 34.15	7.23 / 39.06	14.01 / 34.88	29.01 / **48.96**	14.47 / **38.85**	13.70 / 35.34	**9.41** / 28.28	14.30 / 37.07
LaDM_{LA}	**12.43 / 34.24**	**7.37 / 39.09**	**14.02 / 34.97**	**29.12** / 48.88	**14.53** / 38.49	13.50 / 35.16	9.40 / **28.85**	**14.34 / 37.10**
				Gemma-2-9b-it				
Base	4.08 / 19.29	5.71 / 27.99	8.02 / 27.48	28.57 / 49.73	16.31 / 35.86	16.88 / 34.80	12.07 / 27.66	13.09 / 31.83
LM	14.89 / 36.68	9.49 / 40.43	16.30 / 36.97	29.82 / 49.48	20.06 / 43.57	20.27 / 41.16	16.97 / 37.34	18.26 / 40.80
LA	14.96 / 36.86	9.39 / 40.27	16.14 / 36.67	30.26 / 49.91	19.79 / 43.21	20.61 / 41.75	16.72 / 36.88	18.27 / 40.79
DA	15.31 / 37.44	9.24 / 40.64	16.83 / 37.70	31.15 / 51.27	21.01 / 45.57	21.37 / 42.04	17.53 / 38.27	18.92 / 41.85
LaDM_{LM}	15.66 / 37.54	9.13 / 40.96	**17.23 / 37.75**	**31.36 / 51.75**	21.33 / 45.80	21.94 / 42.97	**17.93 / 38.44**	19.23 / 42.17
LaDM_{LA}	**15.83 / 37.57**	**9.45 / 41.44**	17.13 / 37.67	31.32 / 51.72	**21.41 / 45.86**	**22.01 / 42.91**	17.76 / 38.42	**19.27 / 42.23**

Table 1: Translation performance (spBLEU / ChrF) across seven language pairs. Methods are denoted as: Base (Zero-shot), LM (LoRA-MLP), LA (LoRA-All), DA (DoRA), LaDM_{LM} (LaDM + LoRA-MLP), and LaDM_{LA} (LaDM + LoRA-All). **Bold** denote the best performance among fine-tuned models.

likely due to forgetting caused by fine-tuning on limited low-resource data. In contrast, LaDM substantially mitigates this regression and consistently outperforms the standard PEFT baselines in such sensitive settings; for example, on zh-km, LaDM_{LA} achieves a +3.91 spBLEU gain over LoRA-All. Overall, LaDM's layer-aware structural modulation acts as an effective regularizer, leading to more stable and effective adaptation.

4.2.1 Comparison with Other LoRA Variants

To further compare LaDM with recent parameter-efficient fine-tuning methods, we evaluate it against two representative LoRA variants, VeRA (Kopiczko et al., 2023) and AFLoRA (Liu et al., 2024b), on three low-resource translation directions: zh-lo, zh-km, and zh-bn. Table 2 reports the results on Qwen2.5-7B-Instruct. VeRA exhibits limited effectiveness in these settings, indicating that fixed random projections may be inadequate for the fine-grained representational adjustments required in low-resource machine translation. In contrast, AFLoRA provides a substantially stronger baseline and consistently improves over standard LoRA-All. LaDM_{LA} achieves the best overall performance across the three language directions. In particular, it remains competitive on zh-lo and further improves over AFLoRA on zh-km and zh-bn. These results show that LaDM provides additional gains over existing PEFT methods in severely low-resource settings.

Method	zh-lo	zh-km	zh-bn
Base	1.84 / 15.73	3.93 / 21.93	4.49 / 20.94
LoRA-All	10.78 / 32.40	12.70 / 33.21	8.82 / 27.20
VeRA	0.71 / 10.26	3.01 / 20.10	2.45 / 18.78
AFLoRA	12.41 / 34.14	13.64 / 34.64	8.85 / 27.63
LaDM_{LA}	**12.43 / 34.24**	**14.02 / 34.97**	**9.40 / 28.85**

Table 2: Translation performance (spBLEU / ChrF) across three low-resource translation directions (zh-lo, zh-km, and zh-bn) on Qwen2.5-7B-Instruct. Methods include Base (Zero-shot), LoRA-All, VeRA, AFLoRA, and LaDM_{LA} (LaDM + LoRA-All). **Bold** denotes the best performance in each translation direction.

4.3 Further Analysis

4.3.1 Impact of Randomized and Reversed Modulation

To verify whether the efficacy of LaDM stems from its precise layer-wise targeting, we conduct

Model	Pair	LaDM_{LA}	$\text{LaDM}_{\text{Rand}}$	LaDM_{Rev}
L3	zh-lo	**13.83**	0.07 (↓99.5%)	4.83 (↓65.1%)
	zh-km	**13.99**	0.06 (↓99.6%)	0.07 (↓99.5%)
	zh-bn	**11.06**	10.18 (↓8.0%)	0.05 (↓99.5%)
Q2	zh-lo	**12.43**	11.44 (↓8.0%)	11.99 (↓3.5%)
	zh-km	**14.02**	13.13 (↓6.3%)	13.72 (↓2.1%)
	zh-bn	**9.40**	8.98 (↓4.5%)	8.65 (↓8.0%)
G2	zh-lo	**15.83**	15.46 (↓2.33%)	15.54 (↓1.83%)
	zh-km	**17.13**	17.06 (↓0.41%)	16.98 (↓0.86%)
	zh-bn	**17.76**	17.47 (↓1.63%)	17.23 (↓2.98%)

Table 3: Effect of randomized and reversed variants of LaDM_{LA}. L3, Q2, and G2 denote Llama-3.1, Qwen2.5, and Gemma-2, respectively. Each cell reports spBLEU, followed by the relative drop from LaDM_{LA}.

Method	zh-lo	zh-km	zh-bn
Base	1.35 / 17.20	5.29 / 23.61	13.52 / 34.64
LoRA-All	5.55 / 28.06	10.08 / 31.47	10.40 / 31.98
Layer-wise LR	12.18 / 33.72	13.23 / 33.36	10.77 / 32.23
Init-α	10.60 / 31.77	12.98 / 32.57	8.46 / 28.11
Extreme-α	10.70 / 31.63	12.03 / 32.62	9.58 / 30.33
Mid-α	10.74 / 31.50	13.56 / 33.67	10.70 / 31.77
LaDM_{LA}	**13.83 / 35.33**	**13.99 / 34.65**	**11.06 / 32.44**

Table 4: Comparison with static gradient-control baselines on Llama-3.1-8B. Layer-wise LR uses TAL/LIL learning-rate multipliers of $\times 2.0/\times 0.5$. Init-α, Extreme-α, and Mid-α denote fixed TAL/LIL scaling settings of 1.05 / 0.95, 1.07 / 0.91, and 1.05 / 0.935. **Bold** denotes the best performance in each translation direction.

a sensitivity analysis by introducing randomized ($\text{LaDM}_{\text{Rand}}$) and reversed ($\text{LaDM}_{\text{Rev}}$) variants of LaDM_{LA}. As shown in Table 3, the results reveal clear differences across model architectures in their sensitivity to perturbation. While Qwen2.5 and Gemma-2 exhibit robustness, Llama-3.1 suffers a catastrophic collapse, confirming its extreme sensitivity to layer-wise intervention. Crucially, on Llama-3.1, $\text{LaDM}_{\text{Rand}}$ results in near-zero performance, even worse than LaDM_{Rev}. This suggests that random selection disrupts critical backbone layers that LaDM leaves unmodulated. Therefore, precise layer targeting is strictly necessary to prevent model collapse, rather than being an optional improvement.

4.3.2 Comparison with Static Gradient Reweighting

We further complement our evaluation with comparisons against static gradient-control baselines across three translation directions: zh-lo, zh-km, and zh-bn. Specifically, we consider two types of static baselines. The first uses layer-wise learning-rate tuning, assigning a larger learning rate to TAL layers and a smaller one to LIL layers. The second applies fixed TAL/LIL scaling to LoRA updates, using several preset coefficient configurations derived from Eq. (6), including Init-α, Extreme-α, and Mid-α.

As shown in Table 4, static gradient-control baselines consistently outperform standard LoRA-All, suggesting that non-uniform update allocation across layers is beneficial for translation adaptation. However, LaDM consistently outperforms all static baselines. This suggests that its gains cannot be explained solely by fixed reweighting of gradients or LoRA updates. Instead, the gains of LaDM appear to stem from dynamic layer-wise modulation during training. Rather than relying on fixed scaling rules, LaDM adjusts layer contributions throughout training.

4.3.3 Layer-wise Representation Change

We employ Linear Centered Kernel Alignment (CKA) (Nakai et al., 2025) to measure representational similarity between the fine-tuned model and the frozen base model. As shown in Figure 5, LaDM maintains lower CKA similarity than LoRA across most layers, indicating greater representational deviation from the base model. The gap is most evident in the deeper layers, where Vanilla LoRA stays relatively close to the base representation. This pattern is consistent with the view that standard PEFT tends to preserve more of the original representation, whereas LaDM encourages stronger representational adaptation in later layers.

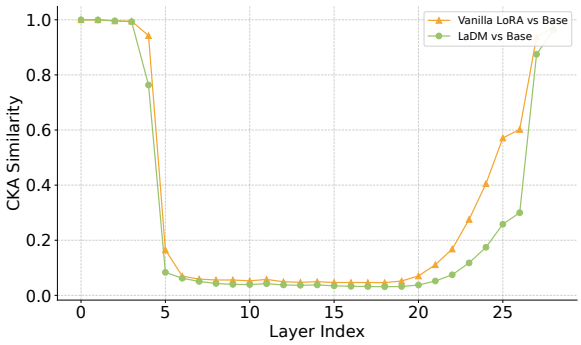

Figure 5: Layer-wise CKA similarity for Llama-3.1-8B on zh-km. Lower values indicate larger representational changes relative to the base model.

4.3.4 Parameter Update Intensity Analysis

To examine how modulation affects parameter updates, we measure update intensity using the Frobenius norm $\|\Delta W\|_F$ of the LoRA weights. As shown in Figure 6, LaDM produces larger parameter updates than Vanilla LoRA in the selected layers, with the strongest increase observed in the LIL layers. The model-level average update magnitude is also higher under LaDM. This pattern suggests that layer-wise modulation increases the magnitude of parameter updates in the selected TAL/LIL layers, with the strongest effect observed in the LIL layers.

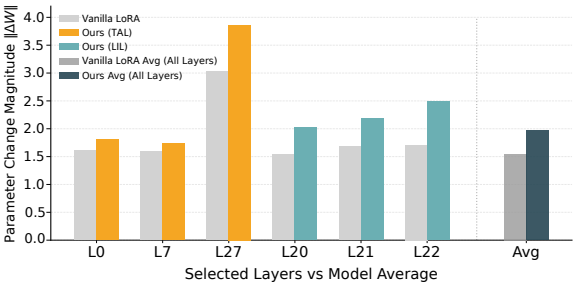

Figure 6: Parameter update magnitude ($\|\Delta W\|_F$) for Llama-3.1-8B on zh-km. Compared with Vanilla LoRA, LaDM shows larger update magnitudes in the selected layers, with the largest differences appearing in the LIL layers.

4.3.5 Training Dynamics of Latent Gates

We track the evolution of the gating parameters $\sigma(g^l)$ to visualize how the model re-regulates its representational stream during training. As shown in Figure 7, we observe a sharp divergence in learning dynamics. Gates in $\mathcal{S}_{\text{TAL}}$ rapidly saturate to maximum amplification, accelerating the consolidation of essential translation features. In contrast, gates in $\mathcal{S}_{\text{LIL}}$ exhibit a steady rising trend. Under our inverse control mechanism (Eq. 6), this increase in $\sigma(g)$ corresponds to a progressive tightening of the inhibition constraint, driving α^l toward its minimum bound. This suggests that the model autonomously identifies and further suppresses the inertia of these layers as training proceeds. By initially dampening entrenched biases and subsequently intensifying this inhibition, LaDM prevents these layers from being locked in their pre-trained states. This approach drives a systemic restructuring of internal states, compelling stagnant layers to break from pre-trained biases and contribute to the translation performance.

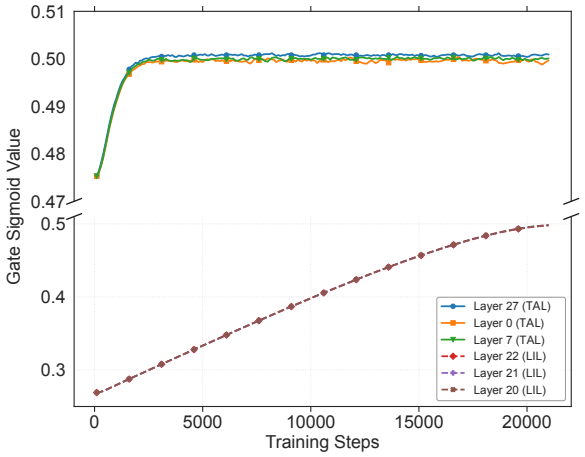

Figure 7: Training Dynamics of Latent Gates in Qwen2.5-7B on zh-km. Evolution of $\sigma(g)$ reveals distinct strategies: TAL gates saturate quickly for signal enhancement, while LIL gates exhibit a steady rise.

4.3.6 Hyperparameter Sensitivity Analysis

We evaluate the impact of modulation intensity and layer scope (k) to identify the optimal balance between task adaptation and structural stability. As illustrated in Figure 8, translation performance is highly sensitive to the modulation range. The configuration R2 ($\mathcal{R}_{\text{boost}}$: 1.03–1.07, $\mathcal{R}_{\text{mute}}$: 0.91–0.96) consistently yields the peak performance across both scaling and scope dimensions. This suggests that low-resource MT requires precise recalibration rather than aggressive weight modification. While minor scaling (R1) is insufficient to overcome functional inertia, excessive scaling (R3) triggers representational drift, leading to a sharp decline in BLEU. Furthermore, we observe that the optimal layer scope remains sparse (T3, $k = 3$). Performance degrades as k increases (T5–T10), confirming that modulating a sparse set of polarized layers is sufficient to trigger compensatory updates without destabilizing the model.

5 Related Work

5.1 Layer-wise Specialization in LLMs

Extensive research has established that LLM layers exhibit distinct functional specializations during information processing (Chen et al., 2023b; Liu et al., 2024c; Jin et al., 2025). Early probing studies suggest a bottom-up hierarchy, where lower layers capture surface-level linguistic features and deeper layers encode complex semantic mappings (He et al., 2024; Marks and Tegmark, 2023; Mao et al., 2025; Hatua, 2025). More recently, interpretability research has highlighted that model layers do not

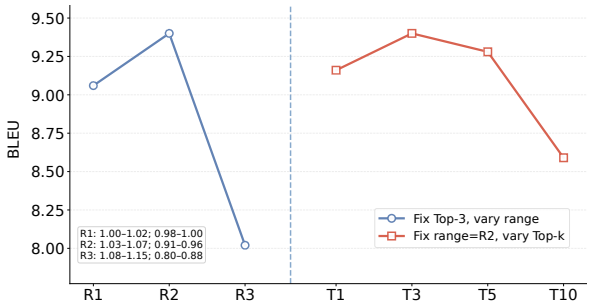

Figure 8: Sensitivity Analysis in Qwen2.5-7B on zh-bn. Performance peaks at intensity R2 ($k = 3$) and scope T3 (Range=R2). BLEU scores decline when scaling is reduced (R1) or excessively increased (R3), and similarly drop as the modulation scope expands from T5 to T10.

contribute uniformly to specific tasks (Zhang et al., 2024; Fan et al., 2024; Ikeda et al., 2025). These findings have motivated growing interest in moving beyond uniform model updates toward layer-aware adaptation strategies. We locate the exact structural bottlenecks in cross-lingual transfer, providing a functional map to guide our dual-directional modulation.

5.2 PEFT Advancements in MT

PEFT has become the standard paradigm for adapting LLMs to MT (Alves et al., 2023; Liang et al., 2025; Aggarwal et al., 2024). LoRA and its variants, including DoRA (Liu et al., 2024a), which decouples direction and magnitude in weight updates, VeRA (Kopiczko et al., 2023), which adopts a vector-based reparameterization, and AFLoRA (Liu et al., 2024b), which improves efficiency through adaptive freezing during fine-tuning, have demonstrated strong performance in cross-lingual transfer (Acharya et al., 2025). However, conventional PEFT methods typically treat all layers uniformly, without accounting for their functional differences. Such uniform updates may be insufficient in low-resource settings. We instead introduce a layer-aware modulation framework that adaptively regulates layer outputs based on layer-wise differences.

6 Conclusion

In this paper, we identify functional polarization as a key bottleneck in adapting LLMs for low-resource translation. We propose LaDM, a layer-aware framework that leverages functional divergence to re-regulate internal representation streams. Unlike layer-agnostic updates, LaDM performs targeted structural recalibration by amplifying fragile task-specific signals while Extensive experiments show that LaDM consistently outperforms strong PEFT baselines. These results suggest that effective low-resource adaptation requires not only acquiring new task knowledge but also overcoming entrenched pre-trained bias.

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), and the Science and Technology Projects of Yunnan Universities Serving Key Industries (Grant No. FWCY-ZD2025006).

Limitations

Although LaDM significantly improves performance in extremely low-resource settings, it is still sensitive to the predefined modulation bounds. While the gating scalars are learnable, their effectiveness depends on being constrained within a suitable range rather than left fully unconstrained. In addition, LaDM is designed primarily for data-scarce scenarios. In moderate-resource settings, where useful pre-trained representations remain more stable, strong inhibition may bring smaller gains or interfere with knowledge preservation. Future work may explore adaptive strategies for learning these bounds more flexibly under different resource conditions.

Ethics Statement

This work focuses on parameter-efficient adaptation of large language models for low-resource machine translation. All experiments use publicly available datasets and pretrained models, and no new personal data are collected. The study does not involve human subjects or real-world deployment.

References

Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. Junlp: Improving low-resource indic translation system with efficient lora-based adaptation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1201–1209.

Divyanshu Aggarwal, Ashutosh Sathe, Ishaan Watts, and Sunayana Sitaram. 2024. Maple: Multilingual evaluation of parameter efficient finetuning of large language models. *ArXiv*, abs/2401.07598.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023b. Beyond surface: Probing llama across scales and layers. *arXiv preprint arXiv:2312.04333*, 1(3).

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6022–6034.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. 2024. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*.

Amartya Hatua. 2025. Mechanistic interpretability of gpt-2: Lexical and contextual layers in sentiment analysis. *arXiv preprint arXiv:2512.06681*.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. *arXiv preprint arXiv:2403.17299*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tianze Hua, Tian Yun, and Ellie Pavlick. 2024. mothello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? *arXiv preprint arXiv:2404.12444*.

Wataru Ikeda, Kazuki Yano, Ryosuke Takahashi, Jaesung Lee, Keigo Shibata, and Jun Suzuki. 2025. Layerwise importance analysis of feed-forward networks in transformer-based language models. *arXiv preprint arXiv:2508.17734*.

Shuting Jiang, Ran Song, Yuxin Huang, Yan Xiang, Yantuan Xian, Shengxiang Gao, and Zhengtao Yu. 2026. Consensus-aligned neuron efficient fine-tuning large language models for multi-domain machine translation. *arXiv preprint arXiv:2602.05694*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, and 1 others. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st international conference on computational linguistics*, pages 558–573.

Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. 2023. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Xiao Liang, Yen-Min Jasmina Khaw, Soung-Yue Liew, Tien-Ping Tan, and Donghong Qin. 2025. Towards low-resource languages machine translation: A language-specific fine-tuning with lora for specialized large language models. *IEEE Access*.

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

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.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024c. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics. *ArXiv*, abs/2403.01509.

Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, Murun Yang, Dingyang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, Tong Xiao, and Jingbo Zhu. 2025. Beyond english: Toward inclusive and scalable multilingual machine translation with llms. *arXiv preprint arXiv:2511.07003*.

Cunli Mao, Xiaofei Gao, Ran Song, Shizhu He, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2025. Multilingual knowledge graph completion via efficient multilingual knowledge sharing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10882–10896.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Toshiki Nakai, Ravi Kiran Chikkala, Lena Sophie Oberkircher, Nicholas Jennings, Natalia Skachkova, Tatiana Anikina, and Jesujoba Oluwadara Alabi. 2025. Treplina: Layer-wise cka+ repina alignment improves low-resource machine translation in aya-23 8b. *arXiv preprint arXiv:2510.06249*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, and 1 others. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Xinyuan Song, Keyu Wang, Pengxiang Li, Lu Yin, and Shiwei Liu. 2025. Demystifying the roles of llm layers in retrieval, knowledge, and reasoning. *arXiv preprint arXiv:2510.02091*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024a. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during llm training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173.

Zhipeng Yang, Junzhuo Li, Siyu Xia, and Xuming Hu. 2025. Internal chain-of-thought: Empirical evidence for layer-wise subtask scheduling in llms. *arXiv preprint arXiv:2505.14530*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are llms for literary translation, really? literary translation evaluation with humans and llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988.

Yang Zhang, Yanfei Dong, and Kenji Kawaguchi. 2024. Investigating layer importance in large language models. *arXiv preprint arXiv:2409.14381*.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Finetuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? *arXiv preprint arXiv:2404.14122*.

A Mechanistic Analysis Details

A.1 Representational Entropy and Entropy Gap

All mechanistic analyses in this work are conducted on the FLORES-200 devtest, which contains 1,012 sentence pairs for each translation direction. This benchmark is used consistently for identifying MT-sensitive modules, localizing TAL/LIL layers, and computing the entropy gap. We base these analyses on devtest examples rather than training data so that the resulting observations reflect the model's behavior on unseen inputs rather than training-specific artifacts.

To characterize layer-wise representational structure, we adopt a spectral-entropy view of hidden states. For a given layer, let $X \in \mathbb{R}^{N \times D}$ denote the centered hidden-state matrix collected from N tokens, where D is the hidden dimension. We apply singular value decomposition (SVD) to X and obtain singular values $\{\sigma_i\}$. These values are normalized into a spectral distribution,

$$p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \tag{9}$$

which reflects how representational energy is distributed across latent directions.

Based on this distribution, we define the representational entropy as

$$H_{\text{svd}} = -\sum_i p_i \log p_i. \tag{10}$$

A lower value indicates that the representation is concentrated in a smaller number of dominant directions, whereas a higher value corresponds to a more diffuse spectrum.

We further compare this quantity between low-resource and high-resource settings by defining the entropy gap

$$\Delta H = H_{\text{svd}}^{\text{low}} - H_{\text{svd}}^{\text{high}}. \tag{11}$$

When $\Delta H > 0$, the low-resource setting exhibits a flatter spectral profile than the corresponding high-resource setting. In our analysis, this pattern suggests that the layer is less effective at organizing task-relevant information into a compact low-rank structure, indicating weaker representational specialization.

A.2 Stability of TAL/LIL Partition

We further analyze the stability of the TAL/LIL partition with respect to anchor choice, model architecture, and translation setting.

Target	Anchor	Top-3 TAL	Top-3 LIL
zh-lo	zh-fr	0, 1, 2	11, 12, 14
	zh-de	0, 1, 2	11, 12, 14
	en-fr	0, 1, 2	11, 12, 14
zh-km	zh-fr	0, 2, 30	11, 12, 13
	zh-de	0, 2, 30	11, 12, 13
	en-fr	0, 2, 30	11, 12, 13
zh-bn	zh-fr	27, 28, 29	0, 5, 6
	zh-de	27, 28, 29	0, 5, 12
	en-fr	27, 28, 29	0, 5, 12

Table 5: Top-3 TAL and Top-3 LIL identified under different high-resource anchors for zh-lo, zh-km, and zh-bn.

Stability across anchor choices. To verify that the partition does not depend on a single high-resource reference pair, we recompute the activation disparity ΔR using alternative anchors, including zh-fr, zh-de, and en-fr. As shown in Table 5, the resulting Top-3 TAL and Top-3 LIL layers remain highly consistent across anchors. In particular, for zh-lo and zh-km, the identified TAL/LIL layers are identical across all tested anchors. For zh-bn, the TAL set remains unchanged, while the LIL set shows only minor variation.

Stability across architectures and language directions. We also examine whether the observed structural pattern transfers across model families and translation settings. To this end, we analyze Qwen2.5-7B on en-tg (with en-fr as the anchor) and Qwen2.5-3B on zh-lo (with zh-en as the anchor). As shown in Table 6, although the exact layer indices vary across architectures, the regional organization remains consistent: TALs concentrate near the early or late parts of the network, whereas LILs are consistently located in the middle layers.

B Additional Experimental Analyses

B.1 Activation Intensity in TAL

To examine how LaDM affects activation patterns in the identified Task-Adaptive Layers (TAL), we measure the mean activation level within TAL across three translation directions. As shown in Figure 9, LaDM consistently yields slightly higher TAL activation than standard LoRA for all evaluated language pairs, with an overall increase of 0.8%. This pattern suggests that LaDM is associated with stronger activation in TAL during adaptation.

Model	#Layers	Top-3 TAL	TAL Region	Top-3 LIL	LIL Region	Structural Pattern
Qwen2.5-7B-Instruct	28	25, 26, 27	Late (3/3)	20, 21, 22	Middle (3/3)	TAL late / LIL middle
Qwen2.5-3B-Instruct	36	0, 34, 35	Early+Late (1+2)	26, 28, 30	Middle (3/3)	TAL edge / LIL middle

Table 6: Top-3 TAL and Top-3 LIL under alternative high-resource anchors. The identified layer sets are highly consistent across anchors for all evaluated translation directions.

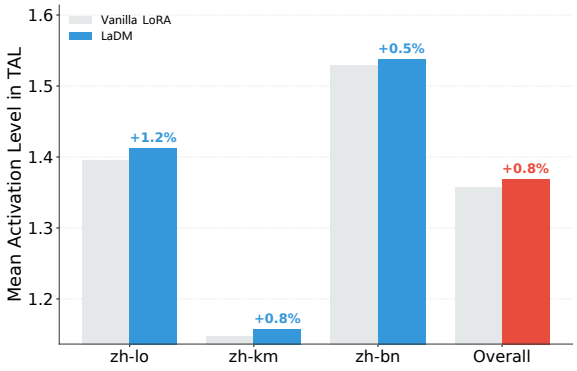

Figure 9: Mean activation in TAL across three translation directions. LaDM shows slightly higher activation in TAL for all language pairs.

Model	Pair	Amplify-only	Inhibit-only	LaDM
L3	zh-lo	7.13	7.41	13.83
	zh-km	0.88	0.98	13.99
	zh-bn	10.68	10.86	11.06
Q2	zh-lo	11.07	11.32	12.43
	zh-km	13.22	13.39	14.02
	zh-bn	8.67	9.10	9.40
G2	zh-lo	15.53	15.68	15.83
	zh-km	16.75	16.63	17.13
	zh-bn	17.53	17.69	17.76

Table 7: Ablation results for dual-directional modulation(spBLEU). L3, Q2, and G2 denote Llama-3.1, Qwen2.5, and Gemma-2.

B.2 Ablation on Dual-directional Modulation

To examine the contribution of the two modulation directions, we compare full LaDM with two single-direction variants: Amplify-only, which modulates only layers with positive ΔR, and Inhibit-only, which modulates only layers with negative ΔR. As shown in Table 7, LaDM consistently achieves the best performance across all evaluated backbones and language pairs. The gap is particularly large on Llama-3.1-8B for zh-km, where the full model substantially outperforms either single-direction variant. These results suggest that amplification and inhibition are complementary, and combining them is more effective than either alone.

B.3 Implementation Details of Additional LoRA Variants

For the additional LoRA variant comparisons in Section 4.2.1, we implement VeRA and AFLoRA under the same overall training framework as the main experiments. Both methods use the same target modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. VeRA uses its vector-based parameterization with $r = 128$, dropout 0.05, and $d_{\text{initial}} = 0.1$. AFLoRA uses a LoRA configuration with rank 8, scaling factor 16, and dropout 0.05, and performs adaptive freezing from step 1000 every 500 steps by freezing 10% of the active LoRA parameters with the lowest gradient scores.

B.4 Case Study

Source (Zh)	红外图像显示，昼夜温度的变化表示这些地方很可能是洞穴。
Reference	*Infrared images show that the temperature variations from night and day show that they are likely caves.*
LoRA (Base)	ຮູບຖ່າຍທີ່ສີແດງ[Literal Hallucination] ໃຫ້ເຫັນວ່າການປ່ຽນແປງຂອງອຸນຫະພູມໃນເຊົ້າແລະຄືນ (missing)[Information Omission]ວ່າເປັນ[Confused modal verbs]ເຫມືອງ. *Analysis: The baseline decomposes "Infrared" into "Red Photo" (literal hallucination), omits the subject "these locations", and generates confused modal verbs due to weak semantic activation.*
Ours	ໃນຮູບພາບອິນຟາເຣັດໄດ້ສະແດງໃຫ້ເຫັນການປ່ຽນແປງຂອງອຸນຫະພູມລະຫວ່າງຕອນເຊົ້າແລະຕອນກາງຄືນເຊິ່ງສະແດງໃຫ້ເຫັນວ່າສະຖານທີ່ນີ້ອາດເປັນຫມືອງ. *Analysis: Our method retrieves the correct technical term "Infrared Image", restores the missing subject via targeted layer modulation, and accurately aligns probabilistic markers for native-like fluency.*

Figure 10: Case study on zh-lo with translation outputs from the reference, LoRA, and LaDM.

Figure 10 presents a qualitative comparison on zh-lo. LoRA shows typical errors, including mistranslating the technical term *infrared* and omitting part of the source meaning. In contrast, LaDM is more consistent with the reference, translating the term more accurately and preserving the omitted content. This example shows that LaDM improves translation fidelity in low-resource settings.