TalkLoRA: Communication-Aware Mixture of Low-Rank Adaptation for Large Language Models

Lin Mu[1], Haiyang Wang[1], Li Ni[1]*, Lei Sang[1], Zhize Wu[2], Peiquan Jin[3], Yiwen Zhang[1]
[1]Anhui University, [2]Hefei University,
[3]University of Science and Technology of China,
{mulin, nili, sanglei, zhangyiwen}@ahu.edu.cn {wanghaiyang}@stu.ahu.edu.cn
wuzz@hfuu.edu.cn jpq@ustc.edu.cn

Abstract

Low-Rank Adaptation (LoRA) enables parameter-efficient fine-tuning of Large Language Models (LLMs), and recent Mixture-of-Experts (MoE) extensions further enhance flexibility by dynamically combining multiple LoRA experts. However, existing MoE-augmented LoRA methods assume that experts operate independently, often leading to unstable routing, expert dominance. In this paper, we propose **TalkLoRA**, a communication-aware MoELoRA framework that relaxes this independence assumption by introducing expert-level communication prior to routing. TalkLoRA equips low-rank experts with a lightweight **Talking Module** that enables controlled information exchange across expert subspaces, producing a more robust global signal for routing. Theoretically, we show that expert communication smooths routing dynamics by mitigating perturbation amplification while strictly generalizing existing MoELoRA architectures. Empirically, TalkLoRA consistently outperforms vanilla LoRA and MoELoRA across diverse language understanding and generation tasks, achieving higher parameter efficiency and more balanced expert routing under comparable parameter budgets. These results highlight structured expert communication as a principled and effective enhancement for MoE-based parameter-efficient adaptation. Code is available at https://github.com/why0129/TalkLoRA.

1 Introduction

Large language models (LLMs), pre-trained on massive corpora (Team, 2024; Yang et al., 2024; OpenAI, 2023), have achieved remarkable performance across a wide range of natural language processing tasks (Qin et al., 2023; Mu et al., 2024). However, adapting such models to domain-specific downstream tasks via full fine-tuning requires updating all parameters, incurring prohibitive memory and computational costs. To address this challenge, parameter-efficient fine-tuning (PEFT) methods (Houlsby et al., 2019) such as low-rank adaptation (LoRA) (Hu et al., 2022) has become a standard approach for adapting LLMs. By injecting low-rank updates into pre-trained weights, LoRA enables efficient adaptation while largely preserving the original model's capabilities. However, standard LoRA employs a single global low-rank update shared across all inputs, which may limit its flexibility under highly heterogeneous prompt and reasoning distributions. (Feng et al., 2024; Ma et al., 2024).

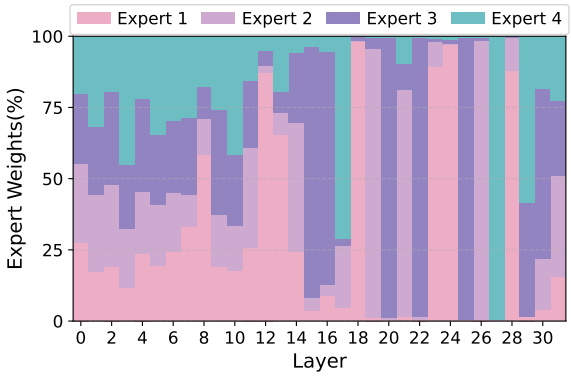


Figure 1: Average expert weights of different experts during inference in the MoELoRA architecture on the OBQA (Mihaylov et al., 2018) dataset.

A natural extension is to introduce multiple low-rank adapters within a Mixture-of-Experts (MoE) framework (Cai et al., 2025), where each LoRA module is treated as an expert and a trainable router dynamically combines experts for each input token (Zhang et al., 2025; Luo et al., 2024; Lin et al., 2025). While such MoE-augmented LoRA (MoELoRA) architectures improve flexibility and enable partial disentanglement of task-shared and task-specific knowledge, they implic-

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18436–18451

itly assume that experts operate independently. In practice, this independence amplifies routing noise, induces sharp and low-entropy gating distributions, and causes the routing mass to concentrate on a small subset of experts (Zuo et al., 2022). As illustrated in Figure 1, this effect intensifies with network depth: a few experts consistently dominate the routing decisions, while others receive negligible gradient signals, resulting in an ineffective contribution to the model's overall capacity. Moreover, because experts are trained independently under identical supervision signals and without any explicit coordination mechanism, they tend to learn highly overlapping representations (Liu et al., 2024a). This representational redundancy substantially reduces the effective expressivity per parameter, ultimately undermining the goal of parameter-efficient adaptation.

To address these limitations, we draw inspiration from communication mechanisms that relax independence assumptions among model components (Misra et al., 2016; Shazeer et al., 2020). In particular, talking-head attention (Shazeer et al., 2020) demonstrated that enabling controlled information exchange among otherwise independent components can significantly improve expressivity and stability. Motivated by this insight, we advocate expert-level communication as a principled extension to MoELoRA. By allowing LoRA experts to exchange compact, task-relevant information during adaptation, such communication facilitates coordination among experts, encourages meaningful specialization while reducing representational redundancy, and smooths parameter updates across expert boundaries.

Building on this idea, we propose **TalkLoRA**, a communication-aware MoELoRA framework that explicitly relaxes the independence assumption among LoRA experts. TalkLoRA introduces a **Talking Module** that enables information exchange across low-rank experts. Specifically, we use their internal low-rank projected features as input to the Talking Module, which aggregates global expert information prior to routing. A dense router then assigns weights to all experts and combines their outputs to form the final adaptation added to the frozen pre-trained weights. In addition, TalkLoRA incorporates a parameter-sharing mechanism that shares a subset of low-rank factors across layers to reduce the number of trainable parameters. Together, these designs enable TalkLoRA to achieve more expressive and better-balanced routing, resulting in more efficient parameter-efficient adaptation.

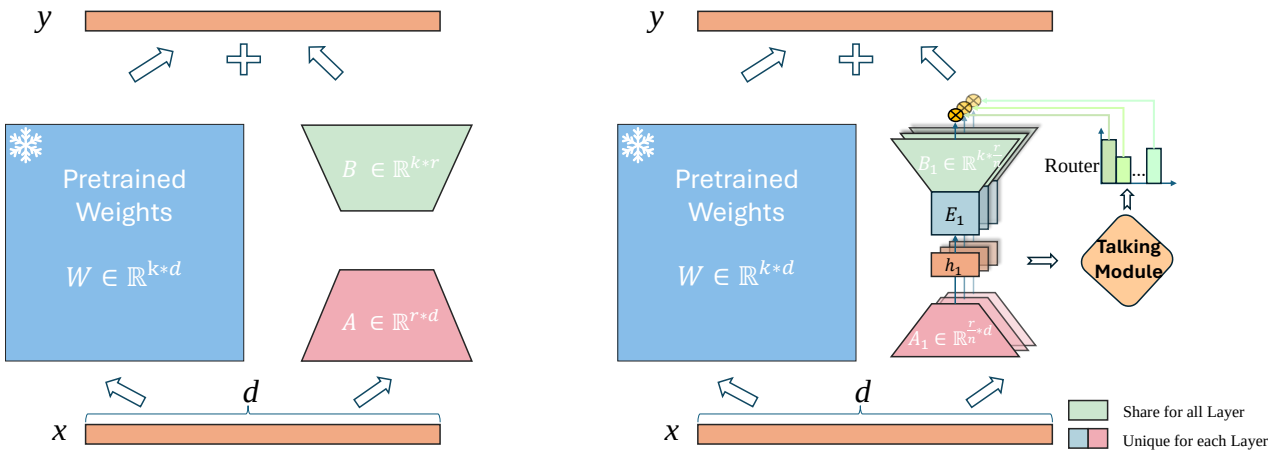
Our main contributions can be summarized as follows:

- We propose TalkLoRA, a communication-aware MoELoRA framework that introduces a lightweight Talking Module to enable information exchange among low-rank experts, effectively relaxing the independence assumption in MoELoRA.
- We provide a theoretical analysis of TalkLoRA, showing that expert communication strictly enlarges the function class of MoELoRA by enabling cross-expert interactions, and formally demonstrating that such communication promotes more balanced expert routing under mild assumptions.
- We conduct extensive experiments on multiple datasets and LLMs, confirming the effectiveness of TalkLoRA. In particular, it achieves 87.8% accuracy on commonsense reasoning with LLaMA3-8B, while demonstrating improved parameter efficiency and more balanced expert routing.

2 Related Works

LoRA and its variants. LoRA (Hu et al., 2022) freezed the pre-trained weights and injects trainable low-rank decomposition matrices into each transformer layer, effectively approximating weight updates through low-dimensional adaptations ($\Delta W = BA$). This design preserves inference efficiency (as the low-rank matrices can be merged into the original weights during deployment) while maintaining competitive downstream performance. Building on this, numerous variants emerge. For example, DoRA (Liu et al., 2024b) decomposed pre-trained weights into magnitude and direction components, enabling LoRA to focus solely on directional updates. Additionally, DenseLoRA (Mu et al., 2025) compressed LoRA updates into a compact dense matrix to mitigate redundancy in LoRA matrix pairs. HiRA (Huang et al., 2025) addressed this by using a Hadamard product to retain high-rank update parameters, improving the model capacity.

MoELoRA. Methods that integrate the Mixture-of-Experts (MoE) paradigm with LoRA have recently garnered considerable research attention. LoRAMoE (Dou et al., 2024) freezed



(a) LoRA

(b) TalkLoRA

Figure 2: Framework comparison of LoRA(left) and TalkLoRA(right).

the backbone model and forces a portion of LoRAs to focus on leveraging world knowledge to solve downstream tasks, to alleviate world knowledge-edge forgetting. MoSLD's (Zhao et al., 2025) core idea was to share the matrix A as the general-feature matrix and keep matrix B as specific-feature matrix and applies dropout to mitigate the imbalance in parameter updates. TeamLoRA (Lin et al., 2025), consisting of a collaboration and competition module for experts, thus achieving the right balance of effectiveness and efficiency. MTL-LoRA (Yang et al., 2025) introduced additional task-adaptive parameters to distinguish task-specific information and captures shared knowledge across tasks in a low-dimensional space.

Unlike existing MoELoRA architectures where experts operate in isolation, TalkLoRA introduces an expert-level communication mechanism. This facilitates inter-expert collaboration and effectively resolves the load imbalance issue.

3 Methodology

In this section, we elaborate on the technical details of TalkLoRA. An overview of the proposed architecture is presented in Figure 2.

3.1 Background

Low-Rank Adaptation. The core idea of LoRA (Hu et al., 2022) is to freeze the original model parameters and inject a low-rank decomposition into the weight updates. Specifically, a pretrained weight matrix $W_0 \in \mathbb{R}^{k\times d}$ is frozen, and two trainable low-rank matrices $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{k\times r}$ in LoRA handle the parameter updates. The rank r is much smaller than d and k (i.e. $r << min(d, k)$). Given an input $x \in \mathbb{R}^d$ to the LLMs, the output $y \in \mathbb{R}^k$ after LoRA is expressed as:

$$y = (W_0 + \Delta W)x = W_0x + BAx, \quad (1)$$

where matrix A undergoes Kaiming initialization (He et al., 2015), while matrix B receives zero initialization, ensuring that fine-tuning initially preserves the original output. During inference, ΔW merges with W_0 (i.e. $W^{'} = W_0 + \Delta W$), eliminating additional latency in the adapted model.

Mixture-of-Experts LoRA (MoELoRA). MoE (Jacobs et al., 1991; Shazeer et al., 2017) models constitute a neural network architecture designed to enhance model capacity and computational efficiency. The core principle activates a subset of expert for a given input or employs dense routing to activate all experts with assigned weights.

LoRA integrates with MoE by treating the product of matrices $A_i \in \mathbb{R}^{\frac{r}{n}\times d}$ and $B_i \in \mathbb{R}^{k\times\frac{r}{n}}$ as a single expert. Each MoELoRA layer contains n LoRA experts. The forward process of the layer is expressed as:

$$y = W_0x + \Delta Wx = W_0x + \sum_{i=1}^{n} g_i(x)B_iA_ix. \quad (2)$$

To balance the contribution of these experts, MoELoRA use a gate function g, which acts as a router network. This network is a fully connected layer with trainable weights $W_g \in \mathbb{R}^{n \times d}$. It is followed by a $softmax$ function that takes a x as input.

$$g(x) = softmax(W_g x). \tag{3}$$

3.2 TalkLoRA Architecture

To address experts operating in isolation and the load imbalance in traditional MoELoRA architectures. TalkLoRA incorporates a Talking Module to enable information exchange among experts and direct the assignment of routing weights. Additionally, inspired by DenseLoRA (Mu et al., 2025), we adopt a parameter sharing strategy for a subset of trainable parameters to reduce redundancy and refine experts. The following details the integration of TalkLoRA into LLMs.

Expert Component: We decompose the original LoRA into n sub-LoRA experts. We further parameterize the up-projection matrix of each LoRA expert as the product of matrix $E_i \in \mathbb{R}^{\frac{r}{n} \times \frac{r}{n}}$ and matrix $B_i \in \mathbb{R}^{k \times \frac{r}{n}}$. Each $A_i \in \mathbb{R}^{\frac{r}{n} \times d}$ and E_i learns domain-specific knowledge. Subsequently, B_i restores the expert dimension to match the original weight matrix output dimension, ensuring compatibility with the LLMs. We use $x \in \mathbb{R}^d$ as the input to each expert, and $y_i \in \mathbb{R}^k$ denotes the corresponding expert output. This process is formulated as:

$$y_i = B_i E_i A_i x, \tag{4}$$

where i ranges from 1 to n. Here, r denotes the total rank of TalkLoRA, and n represents the number of experts.

Talking Module: This module guides the router in weight allocation and relaxes the independence assumption among experts. It enables information exchange among experts prior to routing. Formally, we define:

$$\tilde{h_i} = \sum_{j=1}^{n} C_{ij} h_j, \tag{5}$$

where $C \in \mathbb{R}^{n \times n}$ is a learnable communication matrix and $h_j = A_j x \in \mathbb{R}^{\frac{r}{n}}$ serves as the internal representation of expert j.

This operation allows each expert to integrate compact, task-relevant signals from other experts while preserving its own specialization. The Talking Module is lightweight, adding only $O(n^2)$ parameters.

Routing: Unlike traditional routing, which relies solely on the original input x for decision-making, TalkLoRA performs routing decisions based on the communicated representations $\tilde{h} \in \mathbb{R}^{\frac{r}{n}}$. The process is formulated as follows:

$$g([\tilde{h_1}, \tilde{h_2}, ..., \tilde{h_n}]) = softmax(W_g[\tilde{h_1}, \tilde{h_2}, ..., \tilde{h_n}]), \tag{6}$$

where $W_g \in \mathbb{R}^{n \times r}$ denotes the router parameters. By conditioning routing on globally informed expert features, TalkLoRA mitigates routing overconfidence and reduces sensitivity to local noise.

The overall adaptation process in TalkLoRA can be mathematically formulated as follows, combining the frozen pre-trained weights $W_0 \in \mathbb{R}^{k \times d}$ with the improved experts and router:

$$\begin{aligned} y &= W_0 x + \Delta W x \\ &= W_0 x + \sum_{i=1}^{n} g([\tilde{h_1}, \tilde{h_2}, ..., \tilde{h_n}]) y_i. \end{aligned} \tag{7}$$

Initialization Strategies: Similar to LoRA's initialization strategy, we apply Kaiming initialization (He et al., 2015) to matrices $A_i \in \mathbb{R}^{\frac{r}{n} \times d}$ and $E_i \in \mathbb{R}^{\frac{r}{n} \times \frac{r}{n}}$ in the expert component of TalkLoRA, and zero initialization to $B_i \in \mathbb{R}^{k \times \frac{r}{n}}$, ensuring that initial training preserves the original output. For the routing component, matrices $C \in \mathbb{R}^{n \times n}$ and $W_g \in \mathbb{R}^{n \times r}$ also receive Kaiming initialization.

Notably, B_i is shared across different adaptation layers, whereas A_i and E_i remain unique to each layer. This strategy substantially reduces computational cost while preserving overall performance.

3.3 Analysis of TalkLoRA

Expressive Power Analysis. In MoELoRA, each expert operates within an independent low-rank subspace, which limits the expressivity of the resulting parameter updates to isolated expert contributions. By contrast, TalkLoRA introduces expert-level communication that enables linear interactions across experts prior to routing, thereby allowing parameter updates to span mixed expert subspaces.

- When $C_{ij} = 0$ for all $i \neq j$, TalkLoRA degenerates to MoELoRA, where experts remain fully independent..
- When $C_{ij} \neq 0$ for some $i \neq j$, the representation of expert h_i can incorporate in-

LLM	Method	#Param(%)	BoolQ	PIQA	SIQA	ARC-c	ARC-e	OBQA	HellaS.	WinoG.	Avg.
Qwen2.5-7B	LoRA† ($r = 16$)	0.4	60.0	73.6	70.0	71.7	85.9	74.4	78.6	75.8	73.8
	HiRA† ($r = 16$)	0.4	69.0	88.3	80.8	88.7	95.4	88.0	92.3	81.0	85.4
	TeamLoRA† ($r = 16$)	0.4	74.6	90.0	82.3	88.5	95.9	92.2	95.4	89.0	88.5
	TalkLoRA ($r = 16$)	**0.2**	73.6	90.9	83.0	89.6	96.5	92.8	95.5	89.9	89.0
LLaMA2-7B	LoRA ($r = 32$)	0.8	69.8	79.9	79.5	64.7	79.8	81.0	83.6	82.6	77.6
	DoRA ($r = 32$)	0.8	71.8	83.7	76.0	68.2	83.7	82.4	89.1	82.6	79.7
	HiRA ($r = 32$)	0.8	71.2	83.4	79.5	73.8	86.7	84.6	88.1	84.0	81.4
	MixLoRA ($r = 16$)	2.9	72.7	83.2	78.0	58.1	77.7	81.6	93.1	76.8	77.6
	MoELoRA ($r = 16$)	0.3	68.0	83.5	70.4	61.5	86.8	83.2	90.6	82.5	78.3
	TeamLoRA† ($r = 32$)	0.9	70.6	82.8	79.0	72.4	86.0	81.4	86.2	83.3	80.2
	TalkLoRA ($r = 16$)	**0.2**	72.6	83.9	81.5	73.5	88.0	86.0	89.5	85.2	82.5
	TalkLoRA ($r = 32$)	**0.4**	73.1	84.8	80.9	75.7	87.9	84.8	89.2	86.5	82.9
LLaMA3-8B	LoRA ($r = 32$)	0.7	70.8	85.2	79.9	71.2	84.2	79.0	91.7	84.3	80.8
	DoRA ($r = 32$)	0.7	74.6	89.3	79.9	80.4	90.5	85.8	95.5	85.6	85.2
	HiRA ($r = 32$)	0.7	75.4	89.7	81.2	82.9	93.3	88.3	95.4	87.7	86.7
	MixLoRA ($r = 16$)	3.0	75.0	87.6	78.8	79.9	86.5	84.8	93.3	82.1	83.5
	MoELoRA† ($r = 32$)	0.7	74,6	89.1	82.3	82.8	92.7	87.6	95.3	88.5	86.6
	TeamLoRA† ($r = 32$)	0.7	74.3	88.2	81.8	81.7	92.8	88.0	95.4	89.0	86.4
	TalkLoRA ($r = 16$)	**0.2**	75.3	88.8	82.3	84.3	93.2	89.2	96.2	89.6	87.4
	TalkLoRA ($r = 32$)	**0.4**	76.1	89.6	82.3	84.5	93.9	89.4	96.0	89.2	87.6
	TalkLoRA* ($r = 32$)	**0.4**	75.1	89.2	83.6	84.9	93.9	88.8	96.4	90.1	87.8

Table 1: Accuracy(%) comparison of various methods fine-tuning Qwen2.5-7B, LLaMA2-7B and LLaMA3-8B on the commonsense reasoning tasks. Results for LoRA, DoRA, and HiRA are sourced from (Huang et al., 2025), MixLoRA and MoELoRA are sourced from (Li et al., 2024) and (Yang et al., 2025), respectively. † indicates results reproduced using the same configuration as TalkLoRA. * denotes the use of 8 experts.

formation from expert h_j, resulting in cross-expert interactions that are inaccessible to MoELoRA.

Since MoELoRA is recovered as a degenerate case of TalkLoRA when expert communication vanishes, TalkLoRA strictly subsumes the function class of MoELoRA, yielding strictly greater expressive power.

Routing Sensitivity Analysis: Let $\tilde{h} = (C \otimes I)h$ denote the communicated expert representations used as input to the router. When the expert communication matrix C is non-expansive (the experimental verification shown in the Appendix C.), perturbations in the input are not amplified through the communication module, resulting in bounded changes in the router input.

Specifically, an input perturbation is first projected into the low-rank adaptation space and distributed across experts. The communication module then aggregates expert representations via a linear transformation. If this transformation is non-expansive, it smooths local variations by sharing information across experts rather than amplifying noise. Consequently, the router operates on more stable and less noisy representations, leading to smoother and more robust expert selection.

In contrast, MoELoRA corresponds to the degenerate case without expert communication, where routing decisions depend solely on isolated expert signals and are therefore more sensitive to input perturbations. A formal analysis under standard boundedness assumptions is provided in Appendix A.

4 Experiments

First, we fine-tune the Qwen2.5-7B (Yang et al., 2024), LLaMA2-7B (Touvron et al., 2023) and LLaMA-8B (Team, 2024) models on commonsense reasoning tasks and compare the performance of TalkLoRA against LoRA and its variants. Next, we evaluate smaller LLM, specifically RoBERTa-base (125M) (Liu et al., 2019). We compare different methods for fine-tuning RoBERTa-base on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). We then identify the optimal placement of TalkLoRA within the Transformer architecture. Subsequently, we conduct a robustness analysis of TalkLoRA, examining its performance variation across different expert ranks and numbers of experts. Finally, we verify the contribution of the core module in TalkLoRA.

Method	#Param	SST-2	MRPC	CoLA	QNLI	RTE	STS-B	Avg
BitFit	0.1M	$94.0_{\pm 0.87}$	$88.1_{\pm 1.57}$	$54.0_{\pm 3.07}$	$91.0_{\pm 0.05}$	$69.8_{\pm 1.51}$	$89.5_{\pm 0.35}$	81.1
IA3	0.06M	$93.4_{\pm 0.00}$	$86.4_{\pm 0.00}$	$57.8_{\pm 0.00}$	$91.1_{\pm 0.00}$	$73.5_{\pm 0.00}$	$88.5_{\pm 0.00}$	81.8
LoReFT	0.02M	$93.4_{\pm 0.64}$	$89.2_{\pm 2.62}$	$60.4_{\pm 2.60}$	$91.2_{\pm 0.25}$	$\mathbf{79.0}_{\pm 1.76}$	$90.0_{\pm 0.29}$	83.9
RED	0.02M	$93.9_{\pm 0.31}$	$89.2_{\pm 0.98}$	$61.0_{\pm 2.96}$	$90.7_{\pm 0.35}$	$78.0_{\pm 2.06}$	$90.4_{\pm 0.32}$	83.9
LoRA	0.3M	$93.9_{\pm 0.49}$	$88.7_{\pm 0.76}$	$59.7_{\pm 4.36}$	$92.6_{\pm 0.10}$	$75.3_{\pm 2.79}$	$90.3_{\pm 0.54}$	83.4
Adapter	0.4M	$93.3_{\pm 0.40}$	$88.4_{\pm 1.54}$	$60.9_{\pm 3.09}$	$92.5_{\pm 0.02}$	$76.5_{\pm 2.26}$	$90.5_{\pm 0.35}$	83.7
DeLoRA	0.3M	$94.1_{\pm 0.70}$	$89.0_{\pm 0.96}$	$63.6_{\pm 1.52}$	$92.8_{\pm 0.51}$	$77.1_{\pm 3.65}$	$\mathbf{90.9}_{\pm 0.31}$	84.6
TalkLoRA	0.3M	$\mathbf{94.2}_{\pm 0.37}$	$\mathbf{89.3}_{\pm 1.37}$	$\mathbf{64.2}_{\pm 2.51}$	$\mathbf{93.0}_{\pm 0.32}$	$77.6_{\pm 0.15}$	$\mathbf{90.9}_{\pm 0.52}$	**84.9**

Table 2: Performance comparison of RoBERTa-base model fine-tuned by TalkLoRA and other PEFT baseline methods on the GLUE benchmark (without MNLI and QQP). We report Matthew's correlation for CoLA,Pearson correlation for STS-B, and accuracy for the remaining tasks. All baseline results are sourced from (Bini et al., 2025).

4.1 Commonsense Reasoning

To demonstrate the performance of TalkLoRA on commonsense reasoning tasks, which include BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaS. (Zellers et al., 2019), WinoG. (Sakaguchi et al., 2021), ARC-c and ARC-e (Clark et al., 2018), and OBQA (Mihaylov et al., 2018). Detailed descriptions of these datasets appear in the Appendix B.1.

Experimental Details. For the commonsense reasoning dataset, we fine-tune LLMs for 2 epochs, with evaluation on the validation set every 80 steps. The best checkpoint is selected for the final testing. We use AdamW optimizer (Loshchilov and Hutter, 2019), detailed hyperparameter settings appear in the Table 6. We establish LoRA and its variants as baselines, including DoRA and HiRA, and also compare TalkLoRA with related MoELoRA-based methods: MixLoRA (Li et al., 2024) and TeamLoRA (Lin et al., 2025). All the experiments are conducted using 4 Nvidia 24GB 3090 GPU.

Main Results. Table 1 presents the commonsense reasoning performance of various parameter-efficient fine-tuning (PEFT) methods on the Qwen2.5-7B, LLaMA2-7B and LLaMA3-8B base models, reporting accuracy across eight standard benchmarks and their average (Avg.). The proposed TalkLoRA comprehensively outperforms most of the baselines under extremely low parameter budgets (0.2%–0.4%), demonstrating substantial performance gains.

On Qwen2.5-7B: The most significant observation is that TalkLoRA achieves the highest average accuracy of 89.0%, surpassing the strong baseline TeamLoRA (88.5%) and significantly outperforming the standard LoRA (73.8%).

On LLaMA2-7B: TalkLoRA achieves 82.9% average accuracy with only 0.4% trainable parameters, outperforming the DoRA by 3.2% and standard LoRA by 5.3%. Even in $r = 16$ (0.2% parameters), it reaches 82.5%, surpassing the HiRA (81.4%).

On LLaMA3-8B: TalkLoRA achieves an average accuracy of 87.6%, surpassing HiRA by 0.9% and DoRA by 2.4%. TalkLoRA with $r = 16$ attains 87.4%, outperforming all $r = 32$ baselines while doubling parameter efficiency. More remarkably, with total rank $r = 32$ and 8 experts—yielding a per-expert rank of only 4 (half the conventional minimum)—it further improves the best result (87.6%) by 0.2%.

4.2 Natural Language Understanding

To evaluate TalkLoRA on smaller language models, we fine-tune RoBERTa-base (Liu et al., 2019) and use GLUE (Wang et al., 2019) as the evaluation benchmark. The benchmark comprises eight widely used tasks covering syntactic acceptability (CoLA), sentiment classification (SST-2), paraphrase detection (MRPC, QQP), semantic similarity measurement (STS-B), and natural language inference (MNLI, QNLI, RTE). Due to computational constraints, we exclude the resource-intensive MNLI and QQP tasks. Detailed descriptions of these datasets appear in the Appendix B.1.

Experimental Details. First, we partition the validation set into two subsets. Detailed dataset sizes appear in the Table 5. We then construct hyperparameter groups comprising different learning rates, training epochs, and batch sizes. After each training epoch, we evaluate on the first subset; only when the highest validation score is achieved do we assess performance on the sec-

#Param(%)	LoRA	TalkLoRA	Avg.
0.74	QKVUD	-	80.8
0.17	-	QKV	87.0
0.24	-	UD	87.3
0.41	-	QKVUD	**87.6**

Table 3: Accuracy(%) comparison of TalkLoRA with several different tuning granularity fine-tuning LLaMA3-8B. Each module is represented by its first letter as follows: (Q)uery, (K)ey, (V)alue, (U)p, (D)own.

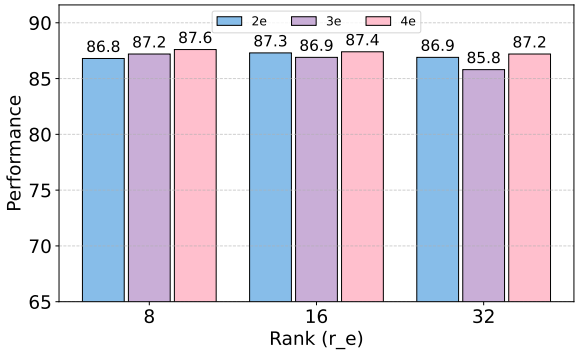


Figure 3: Robustness analysis of TalkLoRA. "2e, 3e, 4e" denotes different numbers of experts.

ond subset. Finally, we apply the hyperparameter configuration yielding the best result across multiple random seeds and report the average to ensure experimental reliability. Consistent with prior work (Bini et al., 2025), we apply TalkLoRA to the Query and Value projections, setting the total rank to 16 and the number of experts to 4 to ensure fairness in trainable parameters. We use BitFit (Ben Zaken et al., 2022), IA3 (Liu et al., 2022), LoReFT (Wu et al., 2024b), RED (Wu et al., 2024a), LoRA, Adapter (Houlsby et al., 2019) and DeLoRA (Bini et al., 2025) under the identical experimental protocol as controls. Detailed hyperparameter settings appear in the Table 7.

Main Results. Table 2 presents performance comparisons of TalkLoRA and other PEFT methods on GLUE, reporting the mean and standard deviation across five random seeds for each task. Compared with representative PEFT baselines, TalkLoRA obtains the best results on SST-2, MRPC, CoLA, QNLI, and overall average score, reaching 94.2, 89.3, 64.2, 93.0, and 84.9, respectively. In addition, it delivers competitive performance on the remaining tasks, demonstrating strong generalization capability. Overall, under a comparable parameter budget, TalkLoRA substantially outperforms existing methods and exhibits consistent robustness across tasks.

Method	# Param(%)	Avg.
TalkLoRA	0.4	87.6
w/o Sharing	0.7	87.7
w/o Talking	0.4	86.5

Table 4: Accuracy (%) comparison of several variants of TalkLoRA.

4.3 Tuning Granularity Analysis

In this section, we analyze the effects of applying TalkLoRA to different adaptation layers in LLMs. We specifically investigate performance in the self-attention module and the feed-forward network, targeting the (Q)uery, (K)ey, (V)alue, (U)p, and (D)own weight matrices. We fine-tune LLaMA3-8B on commonsense reasoning tasks with rank 32 and 4 experts. The result, shown in Table 3, highlight several key observations:

When tuning only the QKV (0.17%) or UD (0.24%), TalkLoRA achieves average accuracies above 87.0%, substantially higher than standard LoRA's 80.8% accuracy with a much larger parameter budget of 0.74%. This indicates that TalkLoRA is significantly more parameter-efficient. Furthermore, when applying the full QKVUD configuration, TalkLoRA reaches the highest accuracy of 87.6% with only 0.41% parameters, demonstrating both superior performance and compact parameter usage. Overall, TalkLoRA delivers more stable and effective improvements across different tuning granularities, showcasing its stronger parameter utilization capability and better scalability.

4.4 Robustness of Expert Rank and Expert Count

We evaluate the performance of TalkLoRA under varying configurations. Specifically, we explore combinations of three per-expert ranks $r_e \in \{8, 16, 32\}$ (where $r_e = r/n$) and three expert counts $n \in \{2, 3, 4\}$, and fine-tune LLaMA3-8B on commonsense reasoning tasks.

As illustrated in Figure 3, TalkLoRA consistently outperforms existing approaches across all tested configurations. At the lowest rank $r_e = 8$, it already attains 86.8% to 87.6% accuracy depending on the number of experts. The strongest

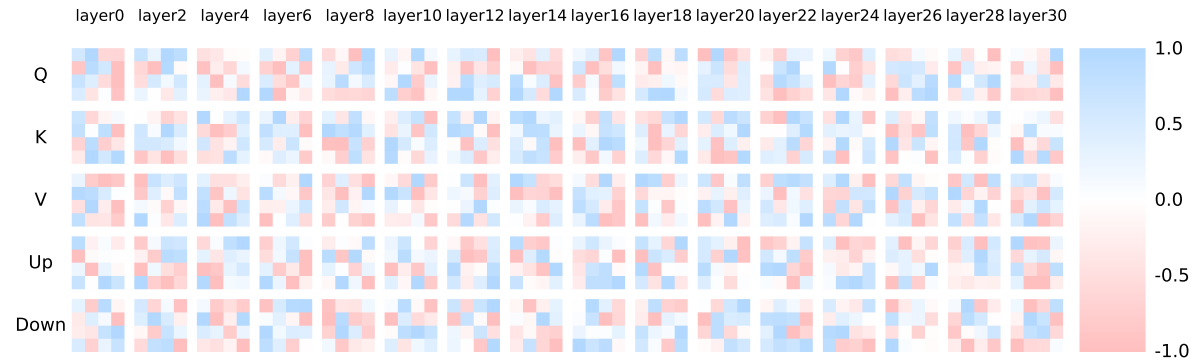
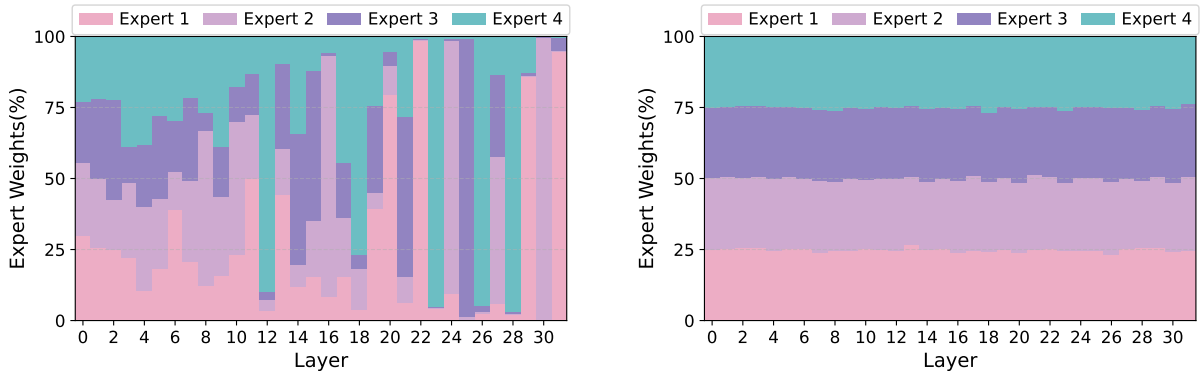


Figure 4: Visualization of the learned communication matrix C. All entries in each matrix are normalized to [-1, 1]. Due to space limitations, we select only even-numbered layers for visualization.



(a) Without Talking Module

(b) TalkLoRA

Figure 5: Router load visualization of without Talking Module (left) and TalkLoRA (right).

configuration (4 experts, $r_e = 8$) reaches 87.6%, improving over HiRA by 0.9% and over standard LoRA by 6.8%. Notably, under identical total rank, increasing the number of experts yields higher performance. For instance, at total rank $r = 32$, using 4 experts achieves 87.6%, surpassing 2 experts (87.3%) by 0.3%. Similarly, at $r = 64$, 4 experts attain 87.4%, outperforming 2 experts (86.9%) by 0.5%.

4.5 Understanding the TalkLoRA

Having demonstrated the superiority of TalkLoRA through extensive experiments, we further conduct studies to quantify the contribution of its internal modules and deep understanding of the architecture. Regarding this, we evaluate the importance of the parameter sharing strategy and the Talking Module through the following ablation settings: 1) We remove the combination of layer-unique E_i and shared B_i, and instead use unshared B_i matrices. 2) We eliminate the Talking Module, feeding the expert low-dimensional information inputs directly to the router. The result is shown in Table 4.

Effectiveness Analysis: TalkLoRA achieves comparable performance to the unshared strategy while using only half the trainable parameters. This substantially reduces redundancy within experts, enabling the MoE mechanism to operate effectively under resource-constrained settings while retaining performance advantages. Removing the Talking Module causes a significant performance drop from 87.6% to 86.5%. Despite contributing a negligible number of trainable parameters, the module yields substantial gains.

Stability Analysis: Figure 4 further reveals that the learned C matrices are neither diagonal nor sparse, confirming extensive information exchange across expert. Furthermore, we examine the routing load distribution on the Up matrix for the OBQA dataset in the commonsense reasoning task. Figure 5 clearly shows that models equipped with the Talking Module effectively mitigate overconfident routing. This prevents any single expert from dominating the selection process. The

Talking Module significantly improves load balancing, thereby ensuring higher overall model performance.

5 Conclusion

In this paper, we propose TalkLoRA, a communication-aware MoELoRA framework that explicitly relaxes the independence assumption among LoRA experts. By introducing a Talking Module, TalkLoRA enables structured information exchange across low-rank experts prior to routing, facilitating coordinated adaptation and more balanced expert utilization. Both theoretical analysis and empirical results show that TalkLoRA enhances the effective expressivity per parameter and improves expert routing, leading to more efficient and robust parameter-efficient adaptation. We believe TalkLoRA offers a principled approach to communication-aware adaptation and demonstrates the potential of structured expert interaction in LLMs.

6 Limitations

While we demonstrate improvements on multiple language understanding and reasoning benchmarks, the effectiveness of expert-level communication in broader tasks—such as multi-modal processing, extremely long-context modeling, mathematical reasoning, or code generation—remains to be explored. In addition, due to hardware constraints, we do not evaluate TalkLoRA on very large-scale models. Finally, the incorporation of a MoE-based Talking Module inevitably introduces additional inference latency, which may limit deployment in latency-sensitive applications.

7 Ethics Statement

First, as a parameter-efficient fine-tuning (PEFT) technique, TalkLoRA relies on LLMs; consequently, it may inherit or even amplify biases and toxic behaviors present in the base model or the fine-tuning datasets. Users must exercise caution regarding data selection and model evaluation to prevent the propagation of harmful content. Second, the increased accessibility of high-performance fine-tuning on consumer-grade hardware could potentially be exploited by malicious actors to adapt models for generating misinformation or offensive content at a lower cost. We encourage the community to prioritize responsible deployment and incorporate safety alignment measures when utilizing this architecture in real-world applications.

8 Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62206004, No.62572002, No.62272001, No.624065095), and the Natural Science Foundation of Anhui Province (No.2208085QF199, No.2508085MF159, No.2308085MF213).

References

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Massimo Bini, Leander Girrbach, and Zeynep Akata. 2025. Decoupling angles and strength in low-rank adaptation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):3896–3915.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-loras: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11371–11380. ELRA and ICCL.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qiushi Huang, Tom Ko, Zhan Zhuang, Lilian Tang, and Yu Zhang. 2025. Hira: Parameter-efficient hadamard high-rank adaptation for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, et al. 2024. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Tianwei Lin, Jiang Liu, Wenqiao Zhang, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Jiannan Guo, Hao Jiang, Siliang Tang, and Yueting Zhuang. 2025. TeamLoRA: Boosting low-rank adaptation with expert collaboration and competition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13622–13637, Vienna, Austria. Association for Computational Linguistics.

Boan Liu, Liang Ding, Li Shen, Keqin Peng, Yu Cao, Dazhao Cheng, and Dacheng Tao. 2024a. Diversifying the mixture-of-experts representation for language models with orthogonal optimizer. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 2966–2973. IOS Press.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *CoRR*, abs/2402.12851.

Yufei Ma, Zihan Liang, Huangyu Dai, Ben Chen, Dehong Gao, Zhuoran Ran, Wang Zihan, Linbo Jin, Wen Jiang, Guannan Zhang, Xiaoyan Cai, and Libin Yang. 2024. MoDULA: Mixture of domain-specific and universal LoRA for multi-task learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2770, Miami, Florida, USA. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct elec-

tricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003.

Lin Mu, Xiaoyu Wang, Li Ni, Yang Li, Zhize Wu, Peiquan Jin, and Yiwen Zhang. 2025. DenseLoRA: Dense low-rank adaptation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10198–10211, Vienna, Austria. Association for Computational Linguistics.

Lin Mu, Wenhao Zhang, Yiwen Zhang, and Peiquan Jin. 2024. Ddprompt: Differential diversity prompting in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 168–174. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. 2020. Talking-heads attention. *CoRR*, abs/2003.02436.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024a. Advancing parameter efficiency in fine-tuning via representation editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13445–13464, Bangkok, Thailand. Association for Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024b. Reft: Representation fine-tuning for language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and

Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. 2025. Mtl-lora: Low-rank adaptation for multi-task learning. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 22010–22018. AAAI Press.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. 2025. MoRE: A mixture of low-rank experts for adaptive multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1311–1324, Vienna, Austria. Association for Computational Linguistics.

Lulu Zhao, Weihao Zeng, Shi Xiaofeng, and Hua Zhou. 2025. MoSLD: An extremely parameter-efficient mixture-of-shared LoRAs for multi-task learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1647–1659, Abu Dhabi, UAE. Association for Computational Linguistics.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Jianfeng Gao, and Tuo Zhao. 2022. Taming sparsely activated transformer with stochastic experts. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

A Analysis of TalkLoRA

We provide a formal analysis supporting the claim that expert-level communication stabilizes routing decisions in TalkLoRA.

Setup

Let $x \in \mathbb{R}^d$ denote the input to a transformer layer. The low-rank projection is given by $h = Ax$, where $A \in \mathbb{R}^{r \times d}$ is a bounded linear operator. The projected representation is split into n expert subspaces $h = [h_1, \ldots, h_n]$. TalkLoRA applies expert communication before routing:

$$\tilde{h} = (C \otimes I)h, \tag{8}$$

where $C \in \mathbb{R}^{n \times n}$ is the expert communication matrix.

The router produces routing probabilities via

$$g(x) = \text{Softmax}(W_g \tilde{h}), \tag{9}$$

with router parameters W_g.

Assumptions

We make the following mild assumptions:

1. **Bounded projection.**
   $\|A\|_{\text{op}} \leq \alpha$.
2. **Non-expansive communication.**
   $\|C\|_{\text{op}} \leq 1$.
3. **Smooth router.**
   The router logits are Lipschitz-continuous:

$$\|W_g u - W_g v\| \leq \beta \|u - v\|.$$

These assumptions are standard and satisfied by common parameterizations used in practice.

Theorem 1 (Routing Stability)

Under the above assumptions, the routing function $x \mapsto g(x)$ in TalkLoRA is Lipschitz-continuous. Specifically, for any perturbation δx,

$$\|g(x + \delta x) - g(x)\| \leq \alpha\beta \|\delta x\|. \tag{10}$$

Discussion. MoELoRA corresponds to the special case where $C = I$, whereas TalkLoRA allows non-trivial but non-expansive expert communication ($\|C\|_{\text{op}} \leq 1$), which smooths perturbations across experts and leads to more stable routing decisions.

Proof

A perturbation δx propagates through the low-rank projection and expert split, yielding

$$\|\delta h\| \leq \|A\|_{\text{op}} \|\delta x\| \leq \alpha \|\delta x\|. \tag{11}$$

Applying expert communication,

$$\|\delta \tilde{h}\| = \|(C \otimes I)\delta h\| \leq \|C\|_{\text{op}} \|\delta h\| \leq \|\delta h\|. \tag{12}$$

By the Lipschitz continuity of the router logits,

$$\|W_g \tilde{h}(x + \delta x) - W_g \tilde{h}(x)\| \leq \beta \|\delta \tilde{h}\|. \tag{13}$$

Finally, since the Softmax function is Lipschitz on bounded domains, the same bound applies to the routing probabilities. Combining the above inequalities yields the stated result.

B Experimental Setting

B.1 Dataset

Commonsense Reasoning: comprising a total of 170,420 question-answer pairs and 120 random entries as the validation set, and consist of 8 benchmarks and the details are described as follows:

- BoolQ (Clark et al., 2019): This dataset comprises a collection of yes/no question examples, totaling 15942 examples. These questions are naturally occurring and generated in unprompted and unconstrained settings;
- PIQA (Bisk et al., 2020): This dataset consists of questions with two solutions that require physical commonsense to answer;
- SIQA (Sap et al., 2019): This dataset focuses on analyzing people's actions and their social implications;
- HellaS. (Zellers et al., 2019): This dataset consists of commonsense Natural Language Inference (NLI) questions, each featuring a context and multiple endings that complete the context;
- WinoG. (Sakaguchi et al., 2021): This dataset presents a fill-in-a-blank task with binary options. The goal is to select the appropriate option for a given sentence that requires commonsense reasoning;
- ARC-c and ARC-e (Clark et al., 2018): These two datasets are the Challenge Set and Easy Set of ARC dataset, which contains genuine grade-school level, multiple-choice science questions;
- OBQA (Mihaylov et al., 2018): This dataset comprises questions that require multi-step reasoning, the use of additional common sense knowledge, and thorough text comprehension.

GLUE: benchmark (Wang et al., 2019) consists of multiple tasks that target different aspects of natural language understanding, and this study adopts six commonly used datasets among them.

- SST-2 focuses on sentence-level sentiment classification and is evaluated using Accuracy.

Splits Sizes	SST-2	MRPC	CoLA	QNLI	RTE	STS-B
Training Set	67K	3.7K	8.5K	105K	2.5K	5.7K
New Validation Set	436	204	522	1K	139	750
New Test Set	436	204	521	4.5K	138	750

Table 5: GLUE dataset sizes.

HyperParmaters	Qwen2.5-7B	LLaMA2-7B		LLaMA3-8B	
Rank r	16	16	32	16	32
α	16	16	32	16	32
Dropout	0.05				
Optimizer	AdamW				
LR	1e-4	3e-4			
LR Scheduler	Linear				
Batch Size	32				
Warmup Steps	100				
Epochs	2				
Where	Q, K, V, Up, Down				

Table 6: The hyperparameters for TalkLoRA on the commonsense reasoning tasks.

- MRPC is a paraphrase detection task that measures whether a model can determine whether two sentences are semantically equivalent, with Accuracy as its evaluation metric.
- CoLA addresses grammatical acceptability judgment, requiring the model to determine whether a sentence is linguistically acceptable, and is evaluated using the Matthews Correlation Coefficient (MCC).
- QNLI is a natural language inference task converted from a question-answering setting, where the goal is to judge whether a sentence contains evidence that answers the question, and is evaluated using Accuracy.
- RTE focuses on recognizing textual entailment, determining whether a premise supports a hypothesis, and is likewise evaluated using Accuracy.
- STS-B is designed for semantic similarity assessment, where the model outputs a similarity score from 0 to 5, and its evaluation metrics include Pearson and Spearman correlation coefficients.

Together, these tasks cover essential language understanding abilities, including sentiment analysis, syntactic judgment, semantic similarity modeling, and natural language inference. As specified by (Wu et al., 2024a), for each benchmark task, we split the public validation set into two parts, as detailed in Table 5.

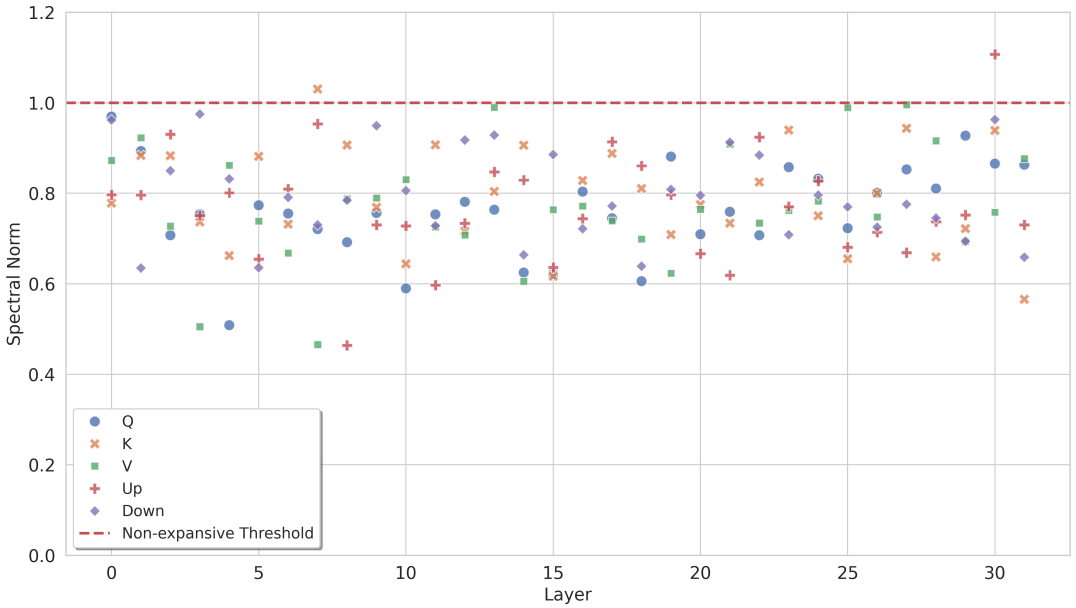


Figure 6: Distribution of spectral norms for the communication matrix across different layers and modules. The red dashed line ($y = 1.0$) denotes the non-expansive threshold. The results demonstrate that the learned matrices predominantly satisfy the non-expansive property, ensuring signal stability throughout the network.

HyperParmaters	SST-2	MRPC	CoLA	QNLI	RTE	STS-B
Rank r	16					
α	32					
Seed	42,43,44,45,46					
Optimizer	AdamW					
LR Schedule	Linear					
Warmup Ratio	6e-2					
Max Seq.Len.	512					
Epochs	40	50	50	40	60	80
Learning Rate	2e-3	2e-4	5e-4	1e-3	7e-4	3e-4
Batch Size	128	32	32	32	32	32

Table 7: The hyperparameters for TalkLoRA on the Natural Language Understanding.

B.2 Hyperparameters

Table 6 shows the detailed hyperparameters for commonsense reasoning tasking when fine-tuning the Qwen2.5-7B, LLaMA2-7B and LLaMA3-8B. Table 7 shows the detailed hyperparameters for GLUE benchmark when fine-tuning the RoBERTa-base.

C Additional Experimental

C.1 Non-expansive validation

As discussed in the main text, a non-expansive communication matrix C ensures that perturbations in the input are not amplified during transmission. This property guarantees bounded changes in the router's input, which is critical for the numerical stability of the Mixture-of-Experts (MoE) architecture during inference. To empirically verify this property in TalkLoRA, we extract the learned communication matrices from the Talking Module across all layers of the fine-tuned model. For each communication matrix $C \in \mathbb{R}^{n\times n}$, we computed its spectral norm, defined as the largest singular value of the matrix:

$$\|C\|_2 = \sigma_{\max}(C) = \max_{x\neq 0} \frac{\|Cx\|_2}{\|x\|_2}. \tag{14}$$

A matrix is non-expansive in the Euclidean space if its spectral norm satisfies $\|C\|_2 \leq 1$. We evaluate this quantity for all projection types (Q, K, V, Up, and Down) across all transformer layers. The resulting distributions of spectral norms are shown in Figure 6. where the x-axis denotes the layer index and the y-axis denotes the corresponding spectral norm. We observe that the spectral norms of the learned communication matrices consistently remain at or below 1 across layers and projection types. These results provide empirical evidence that the Talking Module in TalkLoRA learns non-expansive communication operators in practice, supporting the stability assumptions underlying our theoretical analysis of routing robustness.

C.2 Tuning Granularity Analysis

This section is a detail experiment result of adapting different weight modules using TalkLoRA.

#Param(%)	TalkLoRA	BoolQ	PIQA	SIQA	ARC-c	ARC-e	OBQA	HellaS.	WinoG.	Avg.
0.17	QKV	74.9	89.4	81.2	83.3	93.4	89.4	96.2	88.4	87.0
0.24	UD	74.0	90.2	82.7	83.2	93.1	89.4	96.4	89.2	87.3

Table 8: Accuracy(%) comparison of several different tuning granularity of TalkLoRA fine-tuning LLaMA3-8B. Each module is represented by its first letter as follows: (Q)uery, (K)ey, (V)alue, (U)p, (D)own.

#Param(%)	r_e	n	BoolQ	PIQA	SIQA	ARC-c	ARC-e	OBQA	HellaS.	WinoG.	Avg.
0.2	8	2	76.1	88.4	82.6	81.5	93.7	88.0	96.0	88.5	86.8
0.3	8	3	75.7	89.0	82.4	82.6	93.4	89.6	95.9	89.0	87.2
0.4	8	4	76.1	89.6	82.4	84.5	93.9	89.4	96.0	89.2	87.6
0.4	16	2	75.6	89.9	82.5	83.4	93.7	88.8	96.1	88.2	87.3
0.6	16	3	74.7	88.5	81.5	82.7	94.0	89.0	95.8	89.0	86.9
0.8	16	4	76.3	89.2	82.5	83.5	93.6	88.8	95.8	89.7	87.4
0.8	32	2	74.3	88.3	82.0	84.0	93.2	90.0	95.7	87.6	87.0
1.2	32	3	74.2	88.4	82.3	81.6	91.7	86.0	94.4	87.5	85.8
1.6	32	4	74.2	88.3	82.7	84.4	93.5	88.6	95.9	89.8	87.2

Table 9: Accuracy(%) comparison of robustness for TalkLoRA fine-tuning LLaMA3-8B.

#Param(%)	Method	BoolQ	PIQA	SIQA	ARC-c	ARC-e	OBQA	HellaS.	WinoG.	Avg.
0.7	w/o Sharing	75.4	89.8	83.0	84.5	93.7	89.4	96.0	89.8	87.7
0.4	w/o Talking	72.6	88.8	82.2	83.4	93.0	88.4	95.7	88.2	86.5

Table 10: Accuracy(%) comparison of variant of TalkLoRA fine-tuning LLaMA3-8B.

Each module is represented by its first letter as follows: (Q)uery, (K)ey, (V)alue, (U)p, (D)own. We conduct experiments using LLaMA3-8B with a rank of 32 and expert count of 4 on commonsense reasoning training samples. The result, shown Table 8

C.3 Robustness of TalkLoRA

This section presents detailed experimental results examining the robustness of TalkLoRA. We explore combinations of three per-expert ranks $r_e \in \{8, 16, 32\}$ (where $r_e = r/n$) and three expert counts $n \in \{2, 3, 4\}$, and fine-tune LLaMA3-8B on commonsense reasoning tasks. The result is shown in Table 9

C.4 Ablation Study of TalkLoRA

This section presents detailed experimental results investigating the two variants of TalkLoRA. We conduct experiments using LLaMA3-8B with a rank of 32 and expert count of 4 on commonsense reasoning training samples. The result is shown in Table 10.