Optimizing LoRA Allocation of MoE with the Alignment of Topic Correlation

Hengyuan Xu[1], Wenjun Ke[2,3*], Yao He[4], Jiajun Liu[2], Dong Nie[5], Peng Wang[2,3], Ziyu Shang[2], Zijie Xu[2]

[1]College of Software Engineering, Southeast University, China
[2]School of Computer Science and Engineering, Southeast University, China
[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4]Institute of Collaborative Innovation, University of Macau, Macau, China
[5]Meta Inc.
{xuhengyuan, kewenjun, jiajliu, pwang, ziyus1999, zijiexu}@seu.edu.cn, mc46477@um.edu.mo, dongnie@cs.unc.edu

Abstract

Mixture of experts (MoE) dynamically routes inputs to specialized expert networks to scale model capacity with low inference overhead. However, the excessive parameter growth in MoE models poses challenges in low-resource settings. To address these issues, MoE with parameter-efficient fine-tuning (PEFT) methods have emerged as a lightweight adaptation paradigm that distributes knowledge among experts via multiple LoRA blocks. Existing MoE-PEFT methods can be broadly categorized into External and Internal PEFT methods. External PEFT methods incorporate lightweight models into existing MoE architectures without modifying their routing, which limits the model's parameter efficiency. To overcome these issues, Internal PEFT methods integrate MoE architectures into PEFT, enabling minimal parameter overhead. However, they still face two major challenges: (1) lack of expert functional differentiation, resulting in overlapping specialization across modules, and (2) absence of a structured attribution mechanism to guide expert selection based on semantic relevance. To alleviate these challenges, we propose TopicLoRA, a novel three-stage framework that leverages topic knowledge as semantic anchors to guide expert allocation. Specifically, (1) to address expert redundancy, we construct a topic-level prior graph using Graph Neural Network-enhanced representation learning over Big-Bench categories, enforcing structural separation among expert embeddings, and (2) to introduce semantic attribution, we design a dual-loss training mechanism that softly aligns input-query relevance with topic-guided routing distributions via KL divergence. Extensive experiments on representative datasets (e.g., MMLU, GSM8K, Flanv2) demonstrate that TopicLoRA outperforms state-of-the-art PEFT baselines by 2.40% on average in accuracy. Notably, the maximum improvement is 4.21%. Furthermore, ablation studies demonstrate that our framework's robustness to intricate topics and input sequence variations, which stems from the dual-loss training mechanism.

Code —
https://anonymous.4open.science/r/TopicLoRA-00AF/

*Corresponding authors.

Introduction

Mixture of experts (MoE) (Lepikhin et al. 2020; Zhu et al. 2024; Li et al. 2025; Ning et al. 2025; Truong et al. 2025; Le et al. 2024) has emerged as a scalable architecture to enhance the capability of large language models (LLMs) on various downstream tasks, including question answering (Dai et al. 2024), instruction following (Zhu et al. 2024; Xu et al. 2024) and multimodal reasoning (Lin et al. 2024). MoE activates a limited subset of experts to enable high-capacity models while reducing inference costs. However, MoE-based models still retain substantial total parameter counts, resulting in considerable storage demands and fine-tuning overhead. (He et al. 2025; Kong et al. 2024) For example, Mixtral-8×7B (Jiang et al. 2024) uses 46.7 billion active parameters per inference, yet retains over 100 billion in total, posing challenges for deployment in resource-limited environments.

Several methods have been proposed to alleviate these limitations, such as Expert Pruning (Lu et al. 2024; Chowdhury et al. 2024), Expert Distillation (Xie et al. 2024) and MoE-Based parameter-efficient fine-tuning (PEFT) (Tian et al. 2024; Wu et al. 2024; Dou et al. 2024; Dettmers et al. 2023; Xu et al. 2023). Among them, MoE-Based PEFT enables efficient adaptation by injecting lightweight modules into expert networks, allowing limited parameters to be updated during fine-tuning. Existing MoE-Based PEFT methods can be categorized into two major paradigms: External PEFT (Liu et al. 2024; Wang et al. 2024) methods and Internal PEFT (Luo et al. 2024; Tian et al. 2024; Wu et al. 2024; Dou et al. 2024) methods. External PEFT methods incorporate lightweight models into existing MoE architectures without modifying their routing. However, these methods rely on massive training parameters, which constrains their efficiency. As illustrated in Figure 1(a), the dense clustering of points for External PEFT in the t-SNE plot indicates its reliance on extensive parameter tuning. Compared to External PEFT, Internal PEFT (Tian et al. 2024; Wu et al. 2024; Dou et al. 2024; Luo et al. 2024) methods integrate MoE architectures into PEFT, enabling lightweight adaptation with minimal parameter overhead. Despite their advantage in parameter efficiency, existing Internal PEFT methods still suffer from two disadvantages (Jiao et al. 2024; Lin et al. 2025;

Bai et al. 2024). On the one hand, Internal PEFT methods lack explicit guidance for expert specialization, resulting in overlapping expert representations and limited divergence in their activation patterns. As shown in Figure 1(b), Internal PEFT methods yield entangled expert clusters across different inputs, indicating a lack of clear functional specialization. On the other hand, Internal PEFT methods lack a structured attribution mechanism, which makes the expert selection less aligned with the actual semantics of the input tasks. As shown in Figure 1(c), different tasks exhibit highly similar expert activation patterns in the radar plots, reflecting indistinguishable relevance distributions and undermining the model's ability to perform task-specific adaptation. To overcome these limitations, we argue that expert allocation in MoE can be aligned with explicit topic priors, leading to improved generalization. Guided by this principle, we propose TopicLoRA, a novel three-stage framework that leverages topic knowledge as semantic anchors to guide expert allocation. Specifically, we first optimize the knowledge partition provided by Big-Bench (bench authors 2023) using a combination of LLMs vs LLMs (Cohen et al. 2023) and graph representation learning (Scarselli et al. 2009; Hamilton 2020), which enhances the reliability of the prior to avoid error propagation, resulting in a knowledge prior graph. Second, the knowledge prior graph is used to guide the learning process. We align model-learned relevance scores with prior-based scores to ensure routing mechanism consistency with the topic. Third, TopicLoRA performs forward propagation based on model-learned relevance scores, reducing dependence on topic quality. We conduct experiments on single-topic datasets and multi-topic datasets. The results indicate that our strategy surpasses the baselines in MMLU(Hendrycks et al. 2021), GSM8K (Cobbe et al. 2021) and Big-Bench-Hard (BBH) (Suzgun et al. 2022), achieving accuracy enhancements of 4.21%, 3.89% and 3.2%, respectively. This work makes three primary contributions:

- We argue that synchronizing LoRA expert modules with constructed topics enables more effective correlation detection than model-learned expert allocation.
- We provide a knowledge topic framework based on priors to improve LoRA's training efficacy by enhancing the expert specialization of MoE.
- We demonstrate that TopicLoRA enhances performance across single-topic and multi-topic datasets, and remains robust to low-quality training data.

Method

The framework of TopicLoRA is illustrated in Figure 2. In Stage #1, we construct a topic-level knowledge prior graph from a coarse classification of input topics derived by Big-Bench. The initial partition is refined through LLMs vs LLMs comparison and then embedded into a graph structural via prompt-based encoding and graph representation learning. In Stage #2, expert selection is guided by the topic graph via KL divergence-based alignment of routing decisions. In Stage #3, we use the model-learned relevance scores to propagate information, lessening topic quality dependence. The pseudocode of TopicLoRA is provided in supplementary material. For preliminary details, see the supplementary material.

Prior-Driven Topic Modeling

To enable structure-aware expert selection, we first construct a topic knowledge graph that captures semantic relationships between task clusters. This graph serves as the external prior to guide routing behavior in later stages.

Prior Topic Initialization. Let $\mathcal{D}_{\text{train}=\{(Q^{(i)},y^{(i)})\}_{i=1}^{N}}$ be the multi-task training dataset, where each query $Q^{(i)} \in \mathcal{X}$ is associated with a label $y^{(i)} \in \mathcal{Y}$. A set of topic categories $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_k\}, |\mathcal{T}| = 9$ is derived from coarse metadata provided by Big-Bench, forming an initial partition $\mathcal{P}_0 : \mathcal{X} \to \mathcal{T}$.

LLMs Competition Refinement. Given the initial topic $\mathcal{T}$, we refine topic-level descriptions by leveraging multi-agent LLMs collaboration. Each topic $\tau_i \in \mathcal{T}$ is associated with a textual description d_i. To unify semantically similar topics and reduce redundancy, we employ GPT-4o to derive a condensed topic set $\mathcal{T}' = \tau'_1, \tau'_2, \ldots, \tau'_k$, where $\tau' \subset \tau$ and $|\mathcal{T}'| = 4$. Then, we perform multi-round optimization through cross-model consensus. Let $E_{\text{DS}}(\cdot)$ and $E_{\text{QW}}(\cdot)$ denote the embedding functions of DeepSeek-v3 and Qwen-2.5, respectively. At round t, we define the consensus discrepancy score between the two models as:

$$S_{i,j}^{(t)} = \frac{1}{2}\left(\|E_{\text{DS}}(d_{i,j}^{(t)}) - E_{\text{QW}}(d_{i,j}^{(t)})\|_{\text{idf}} + \|E_{\text{QW}}(d_{i,j}^{(t)}) - E_{\text{DS}}(d_{i,j}^{(t)})\|_{\text{idf}}\right) \quad (1)$$

The optimized description at round $(t+1)$ is the updated by GPT-4o with feedback from the discrepancy score gradient:

$$d_{i,j}^{(t+1)} = \text{GPT-4o}(d_{i,j}^{(t)}; \nabla S_{i,j}^{(t)}) \quad (2)$$

This process is repeated until convergence, resulting in a refined description matrix $\mathbf{D}^* = \{d_{i,j}^*\}$ for all $i \in [1,4], j \in [1,4]$, which serve as an input for downstream semantic encoding. As shown in Figure 2 Stage #1, we initialize and optimize the topic from [*Language Understanding and Generation NLP Tasks,Understand the Commonsense World Knowledge,Scientific and Technical Knowledge Understanding*]. Full implementation details, including the prompting scheme and multi-model consensus-based validation, are provided in supplementary material.

Semantic Text Encoding. We encode the refined topic descriptions into dense vector representations $v_{i,j}$ to initialize the topic node features. Specifically, we use a frozen DeBERTa-v3 (He, Gao, and Chen 2021) encoder $f_{\text{enc}} : \text{text} \to \mathbb{R}^d$ to convert each optimized description $d_{i,j}^* \in \mathbf{D}^*$ into a semantic embedding:

$$v_{i,j} = f_{\text{enc}}(d_{i,j}^*) \in \mathbb{R}^d \quad (3)$$

For each topic $\tau'_i \in \mathcal{T}'$, we concatenate the embeddings from the $j = 1$ to $j = 4$ dimensions to obtain its feature matrix:

$$\mathbf{F}_i = \text{CAT}(v_{i,1}, v_{i,2}, v_{i,3}, v_{i,4}) \in \mathbb{R}^{4\times d} \quad (4)$$

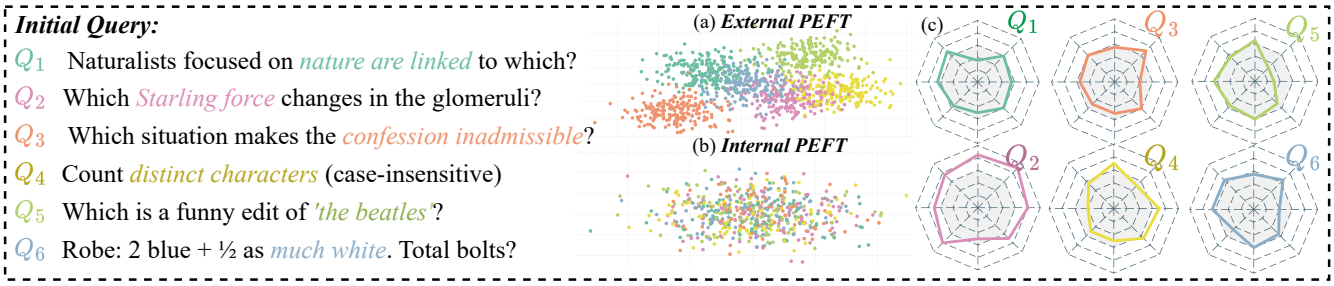

Figure 1: Initial query Q_1 to Q_6 are representative data of MMLU(Hendrycks et al. 2021), Medical(Hendrycks et al. 2021), Law(Hendrycks et al. 2021), HumanEval (Chen et al. 2021), GSM8K (Cobbe et al. 2021) and Big-Bench-Hard (Suzgun et al. 2022), respectively. (a) and (b) shows the t-SNE of External PEFT and Internal PEFT experts, respectively. The clustering of points in (a) and (b) reflects the distribution and grouping of experts, where each point represents an expert and points with the same color correspond to the same task category. (c) shows the relevance scores of Q_1 to Q_6 across different experts in Internal PEFT, indicating how each query aligns with various task-specific experts. The 8 dimensions in (c) correspond to 8 distinct experts within Internal PEFT.

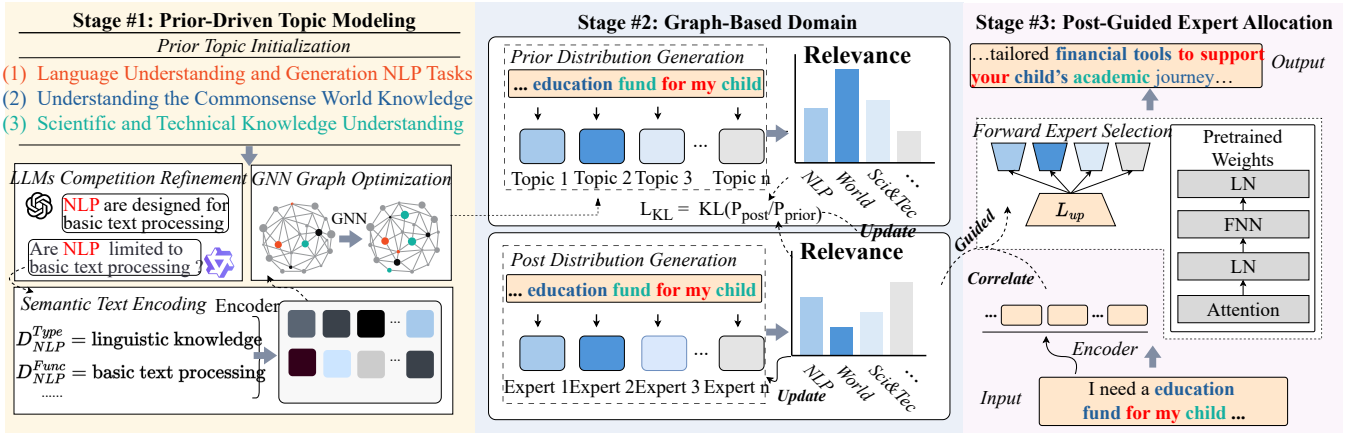

Figure 2: An overview of TopicLoRA framework. The arrows indicate the direction of data flow.

We then compute the pairwise affinity between any two topics τ_i' and τ_j' using a Gaussian kernel over matrix-level Euclidean distance:

$$\mathcal{R}_{i,j} = \exp\left(-\frac{\|\mathbf{F}_i - \mathbf{F}_j\|_F^2}{2\sigma^2}\right) \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and σ is a bandwidth parameter that controls the sensitivity of similarity. The resulting relevance matrix $\mathcal{R} \in \mathbb{R}^{k\times k}$ serves as the weighted edge set of a fully-connected graph $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$, where $\mathcal{V} = \{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k\}$ denotes the set of topic node features and $\mathcal{E}_0 = \{\mathcal{R}_{i,j}\}$ represents semantic relationships between them.

GNN Graph Optimization. To capture higher-order dependencies among topic nodes, we refine the initial graph $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ using GNN. Let $\mathbf{F}_i^l$ denote the representation of node i at layer l of the GNN. The message-passing rule is defined as:

$$\mathbf{F}_i^{l+1} = \delta\left(\sum_{j\in\mathcal{N}(i)} \mathcal{E}_{ij} \cdot \mathbf{F}_j^l \cdot \mathbf{W}^l\right) \tag{6}$$

where $\mathcal{N}i$ denotes the neighborhood of node i, $\mathbf{W}^l \in \mathbb{R}^{d\times d}$ is the learnable projection matrix at layer l, and $\delta(\cdot)$ is the activation function. After t^{th} iteration, we obtain the optimized node embeddings $\mathbf{V} = \{\mathbf{F}_1^t, \mathbf{F}_2^t, \ldots, \mathbf{F}_k^t\}$, which define the updated topic prior graph $\mathcal{G}_{\text{prior}} = (\mathcal{V}', \mathcal{E})$, where $\mathcal{V}' = \mathbf{V}$ and $\mathcal{E}$ retains the original edge weights based on semantic similarity. Supplementary material provides a spectral analysis of prior graph effect on LoRA stability.

Graph-Based Topic Alignment

To ensure structure-aware routing during adaptation, we incorporate the topic-level prior graph $\mathcal{G}_{\text{prior}} = (\mathcal{V}', \mathcal{E})$ into the model via a probabilistic alignment mechanism. Specifically, we guide expert selection by aligning the model-learned expert distribution with prior topic relevance.

Prior Distribution Generation. Given the prior topic graph $\mathcal{G}_{\text{prior}=(\mathcal{V}',\mathcal{E})}$ and training queries $\mathcal{Q} = \{Q_{i=1}^N\}$, we compute a prior expert distribution for each input based on its semantic affinity to the topic nodes. Each query $\mathcal{Q}^i$ is encoded via DeBERTa-v3 f_{enc} to obtain a contextual vector $\mathbf{h}^i \in \mathbb{R}^d$. For each topic node $\tau_j' \in \mathcal{T}$, the final node

representation $\mathbf{F}_j^t \in \mathbb{R}^{4\times d}$ is pooled to a vector $\bar{\mathbf{v}}_j = \text{AvgPool}(\mathbf{F}_j^t) \in \mathbb{R}^d$. The affinity between $\mathbf{h}^i$ and each topic vector $\bar{\mathbf{v}}_j$ is computed as:

$$\alpha_{i,j}^{\text{prior}} = \frac{\exp(\mathbf{h}^i \cdot \bar{\mathbf{v}}_j/\tau)}{\sum_{l=1}^{k}\exp(\mathbf{h}^i \cdot \bar{\mathbf{v}}_l/\tau)}, \quad \mathbf{P}_i^{\text{prior}} = [\alpha_{i,1}^{\text{prior}}, \ldots, \alpha_{i,k}^{\text{prior}}] \tag{7}$$

Where $\tau > 0$ is a temperature parameter that controls the confidence of the prior. This yields a prior distribution vector $\mathbf{P}_i^{\text{prior}} \in \Delta^k$ for the i^{th} query. Aggregating across the dataset forms the prior expert distribution matrix $\mathbf{P}^{\text{prior}} \in \mathbb{R}^{N\times k}$, which will be used in subsequent alignment.

Post Distribution Generation. Given the encoded input representation $\mathbf{h}^i \in \mathbb{R}^d$, we derive a post-hoc expert distribution that reflects the model's implicit routing behavior:

$$\alpha_{i,j}^{\text{post}} = \frac{\exp(\mathbf{h}^i \cdot \bar{\mathbf{u}}_j)}{\sum_{l=1}^{k}\exp(\mathbf{h}^i \cdot \bar{\mathbf{u}}_l)}, \quad \mathbf{P}_i^{\text{post}} = [\alpha_{i,1}^{\text{post}}, \ldots, \alpha_{i,k}^{\text{post}}] \tag{8}$$

where $\bar{\mathbf{u}}_i \in \mathbb{R}^d$ is the Xavier-initialized vector associated with i^{th} expert $\mathcal{E}_i$. This gives the post distribution vector $\mathbf{P}_i^{\text{post}} = [\alpha_{i,1}^{\text{post}}, \ldots, \alpha_{i,k}^{\text{post}}] \in \Delta^k$. Collectively, the full post distribution matrix is:

$$\mathbf{P}^{\text{post}} = \{\mathbf{P}_1^{\text{post}}, \ldots, \mathbf{P}_N^{\text{post}}\} \in \mathbb{R}^{N\times k} \tag{9}$$

KL-Based Distribution Alignment. Given the prior expert distribution $\mathbf{P}^{\text{prior}} \in \mathbb{R}^{N\times k}$ and the post expert distribution $\mathbf{P}^{\text{post}} \in \mathbb{R}^{N\times k}$, we compute a KL-Divergence Regularization loss that aligns the model's expert routing behavior with prior topic guidance:

$$\mathcal{L}_{\text{KL}} = \text{KL}(\mathbf{P}^{\text{post}} \| \mathbf{P}^{\text{prior}}) \tag{10}$$

LoRA Allocation with Topic-aware Prior Knowledge

In the final stage, TopicLoRA performs modular adaptation by assigning low-rank parameter modules (LoRA) to experts selected based on topic-aware routing.

Post-Guided Allocation. Given the posterior expert distribution $\mathbf{Q}_i^{\text{post}} \in \Delta^k$, we activate all experts with soft routing. Each expert $\mathcal{E}_i$ is equipped with a trainable LoRA adapter, parameterized by a shared projection $A \in \mathbb{R}^{d\times r}$ and expert-specific matrix $B_i \in \mathbb{R}^{r\times d}$, forming:

$$\Delta W_i = AB_i \tag{11}$$

where $r \ll d$ denotes the rank of the adapter. As shown in Figure 2 Stage #3, we use $\mathbf{Q}_i^{\text{post}}$ as the routing score to perform forward propagation on *I need to plan an education fund for my child.*

Forward Expert Selection. The output representation is computed by aggregating expert-specific forward paths, weighted by the post distribution:

$$f(Q) = \sum_{i=1}^{k} \alpha_i^{\text{post}} \cdot \mathcal{E}_i(Q; \Delta W_i) \tag{12}$$

where α_i^{post} denotes the routing weight, and $\mathcal{E}_i(Q; \Delta W_i)$ is the expert-specific output with the associated LoRA adapter.

Training Objective. We define the total loss for this stage by integrating the alignment term into the main objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} \tag{13}$$

where $\mathcal{L}_{\text{task}}$ denotes the supervised task loss, and λ_{KL} is a scalar hyperparameter controlling the influence of prior alignment.

Experiment

Datasets and Settings

We perform studies on both single-topic and multi-topic datasets. Regarding single-topic datasets: In the general, medical, law, math and code field, we train the model in Databricks-Dolly-15k (Conover et al. 2023), the GenMedGPT and Clinic-10k from ChatDoctor (Li et al. 2023), the Lawyer-Instruct (Alignment-Lab-AI 2024) and US-Terms (Chalkidis et al. 2023), the GSM8K (Cobbe et al. 2021) training set, respectively. The evaluation is conducted on MMLU(Hendrycks et al. 2021) (medical and legal), GSM8K, and HumanEval (Chen et al. 2021). Regarding multi-topic datasets: we select a subset of the FlanV2 (Wei et al. 2021) dataset for fine-tuning and evaluated the model using the Big-Bench-Hard (BBH) (Suzgun et al. 2022). Comprehensive details regarding the datasets and evaluation metrics are outlined in supplementary material. DeBERTa-v3 (He, Gao, and Chen 2021) is used to generate question representations. During fine-tuning, the hyperparameter settings are as follows: *lora rank* = 8, *lora heads* = 4, training epochs = 3, batch size = 8, hidden dimension d=768, learning rate = 1e-5, λ_{KL} = 0.3, τ = 1.2, σ = 1.0 and dropout rate = 0.05. All experiments are done with 4*L40s (48G) GPUs. We adopt LLaMA2-7B (Touvron et al. 2023) as the base model.

Baselines

MoE-Free PEFT **Prompt Tuning** (Lester, Al-Rfou, and Constant 2021) incorporates adjustable cues into the job to direct the pre-trained model in comprehending the task specifications without altering the original model parameters. **P-Tuning** (Liu et al. 2022b) transforms the prompt into a trainable embedding layer and utilizes MLP and LSTM to process the prompt embedding. **Prefix Tuning** (Li and Liang 2021) incorporates an extra trainable prefix vector into every layer of the Transformer. **IA (Invisible Adapter)** (Liu et al. 2022a) is a specialized fine-tuning technique that enhances efficiency by including a compact model into the pre-trained model. **AdaLoRA** (Zhang et al. 2023) is a method for adaptive budget allocation in low-rank adapters that implements a methodology to optimize the parameter updating strategy during fine-tuning. **LoRA** (Hu et al. 2022) employs low-rank matrix decomposition to approximate the weight updates of fully linked layers, therefore diminishing the number of parameters requiring optimization.

MoE-Based PEFT **LoRAMoE** (Dou et al. 2024) proposes a MoE-style plugin architecture to address the issue of *catastrophic forgetting of world knowledge* during large-scale model fine-tuning. The backbone model is frozen

while multiple LoRA expert modules and a routing network are added. **HydraLoRA** (Tian et al. 2024) introduces an asymmetric LoRA architecture that shares a common A matrix across tasks while assigning multiple task-specific B matrices. This design captures both shared and task-specific patterns, thereby reducing task interference. **MoSLoRA** (Wu et al. 2024) integrates sparse gating mechanisms with LoRA to enable parameter-efficient multi-task fine-tuning by updating a small subset of expert modules, maintaining high performance with minimal parameters.

Main Results

The primary findings are shown in Table 1, from which we derive the subsequent conclusions.

Initially, compared to MoE-Free PEFT methods, MoE-Based PEFT approaches demonstrate superior performance across various datasets. Taking HydraLoRA as an example, it achieves improvements of 0.63%, 1.39%, 1.37%, 1.19% (P@1), and 1.60% over the best-performing MoE-Free LoRA method on the MMLU, Medical, Law, HumanEval, and GSM8K, respectively.

Furthermore, TopicLoRA exhibits markedly enhanced performance in single datasets tasks relative to current baseline approaches, especially in identifying latent relationships across knowledge subjects. On the MMLU, Medical, Law, HumanEval, and GSM8K benchmarks, TopicLoRA improves performance by 5.76%, 2.54%, 0.63%, 1.77%(P@1), and 3.89% against HydraLoRA , respectively; and by 4.21%, 0.67%, 5.30%, 2.36%(P@1), and 9.49% against MoSLoRA , respectively. This demonstrates the effectiveness of MoE-based designs in improving overall model performance. The observed performance pattern, which records a 5.76% gain on MMLU while the gains on Law and HumanEval are limited to 0.63% and 1.77% respectively, indicates that TopicLoRA is particularly effective in multi-topic scenarios, where its enhanced expert diversity and precision better support complex cross-topic reasoning.

Finally, compared with existing methods, TopicLoRA has achieved a performance improvement of 3.2% on multi-topic datasets. The experimental results are shown in Table 2, compared to MoE-Based LoRA methods (e.g., LoRAMoE (Dou et al. 2024), HydraLoRA (Tian et al. 2024) and MoSLoRA (Wu et al. 2024)), TopicLoRA improves on multi-topic by 4.4%, 3.2% and 3.7%, respectively. This phenomenon indicates that the prior knowledge-guided mechanism of TopicLoRA not only enhances the model's average performance but also improves its robustness across diverse task scenarios.

Ablation Experiments

To explore the contribution of each component in the proposed method, we conduct ablation experiments on three different datasets and the results are shown in Table 3. The performance decline follows the order: TopicLoRA *w/o* KLD > TopicLoRA *w/o* MCV > TopicLoRA *w/o* GRL.

About Multimodel Cross-Validation Mechanism (MCV). We conduct ablation studies on the full TopicLoRA model. As shown in Table 3, we observe that MCV makes a significant contribution to the reliability of knowledge topic priors. Specifically, using only a single model generated prior leads to performance drops of 4.56%, 9.43%, and 6.29% in the general, medical, and legal topics, respectively. This indicates that prior optimization is achieved through LLM vs. LLM comparison, thereby avoiding error propagation.

About Graph Representation Learning (GRL). We further analyze the impact of removing GRL from the full model. The results in Table 3 show that eliminating GRL leads to performance degradation of 2.80%, 6.19%, and 6.17% on the general, medical, and legal topics, respectively. These results demonstrate that GRL plays essential roles in constructing reliable prior correlations.

About Kullback-Leibler Divergence-based Dual Loss (KLD). The model is subjected to ablation experiments on model-prior alignment strategies. As shown in Table 3, compared to rigid topic-based models, KLD improves model performance significantly. Specifically, KLD-trained model outperforms rigidly trained model by 8.42% in general topic, 7.64% in medical topic, and 7.74% in legal topic. This shows that our KLD strategy maintains the model's autonomous learning ability during training and reduces the alignment prior routing method's dependence on prior quality.

Analysis Experiments

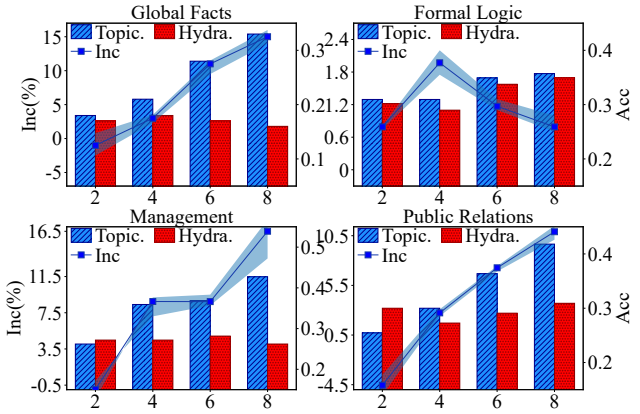

Figure 3: Performance of TopicLoRA and HydraLoRA under increasing experts. Bars (right y-axis) show accuracy (Acc); lines (left y-axis) show performance increase (Inc). Shaded areas indicate confidence intervals.

Effect of Different LoRA Expert Numbers To test TopicLoRA with different LoRA experts, we run experiments using between 2 and 8 LoRA experts. As shown in Figure 3, TopicLoRA is compared to HydraLoRA on *global facts*, *formal logic*, *management*, and *public relations* from MMLU. As the number of LoRA experts increases, the performance of the HydraLoRA model remains relatively stable, while the performance of TopicLoRA improves as the number of experts grows. For example, on the tasks of *management* and *public relations*, the performance improvements of TopicLoRA over HydraLoRA are 8.74% and 6.36% on average

Categories	Methods	MMLU	Medical	Law	HumanEval		GSM8K	#A	#B
					P@1	P@10			
Base Model	zero-shot	38.88†	35.98†	33.51†	13.10†	20.34†	10.38†	-	-
MoE-Free	Prompt Tuning	39.91†	37.59†	35.02†	13.66†	21.55†	13.18†	-	-
	P-Tuning$_{(256)}$	41.11†	39.81†	36.72†	13.60†	21.13†	15.56†	-	-
	Prefix Tuning	41.78†	40.28†	36.54†	13.23†	22.56†	16.89†	-	-
	$(IA)^3$	40.45†	37.12†	35.25†	13.54†	23.17†	13.98†	-	-
	AdaLoRA$_{(r=8)}$	44.32†	42.83†	39.36†	14.81†	23.78†	19.51†	-	-
	LoRA$_{(r=8)}$	43.22†	41.59†	37.85†	15.67†	22.95†	18.24†	1	1
	LoRA$_{(r=16)}$	45.45†	43.10†	39.64†	16.71†	25.60†	20.32†	1	1
	LoRA$_{(r=32)}$	46.59†	44.32†	40.81†	17.12†	25.89†	20.67†	1	1
MoE-Based	LoRA-Split$_{(4\times8)}$	46.94†	45.28†	41.35†	18.20†	<u>26.85</u>†	21.92†	4	4
	HydraLoRA$_{(r=8)}$	47.22†	45.71†	<u>42.18</u>†	<u>18.31</u>†	**27.43**†	<u>22.27</u>†	1	4
	MoSLoRA	<u>48.77</u>	<u>47.58</u>	37.51	17.72	25.16	16.67	4	4
Ours	**TopicLoRA$_{(r=8)}$**	**52.98**	**48.25**	**42.81**	**20.08**	26.28	**26.16**	1	4

Table 1: Comparison of performance on five single-topic datasets in accuracy (%). Bold and underline represent the best and second best score, respectively. The results with † are from the experiments of Tian et al. (Tian et al. 2024). MoSLoRA results are obtained from the authors' released code. #A and #B refer to the A and B numbers of the LoRA matrix, respectively.

Method	Result	A/B Train	A/B Infer
Base	31.6†	0/0	0/0
LoRA	36.8†	1/1	1/1
LoRAHub	39.7†	48/48	20/20
LoRAMoE	40.3†	48/48	48/48
HydraLoRA	<u>41.5</u>†	1/10	1/10
MoSLoRA	41.0	4/4	4/4
TopicLoRA	**44.7**	1/4	1/4

Table 2: Comparison of performance on multi-topic datasets in accuracy (%). Bold and underline represent the best and second best score, respectively. The results with † are from the experiments of Tian et al. (Tian et al. 2024). MoSLoRA results are obtained from the authors released code.

in accuracy, respectively. We argue that, more experts improve the topic refinement and expert differentiation, thus improving performance.

Effect of Different Topic Numbers To assess the efficacy of TopicLoRA in managing diverse quantities of knowledge topics, we amalgamate subsets with variable amounts of topics to create new training datasets. These new datasets contain 2, 4, and 6 categories of topics, respectively, and the total number of datasets is 4,000 in each case. Figure 4 compares models from two training methods on datasets with different topic quantities. The accuracy of models from both fine-tuning methodologies decreases as the number of topics increases, and TopicLoRA declines more gradually. Specifically, Baseline's accuracy ranged from 22.12% to 16.15% to 13.85%, and TopicLoRA's accuracy ranged from 22.89% to 20.26% to 17.95%.

Effect of Different Sampling Strategies We conduct experiments with different dataset ingest orders, such as the

Models	General		Medical		Law	
	ACR	↓(%)	ACR	↓(%)	ACR	↓(%)
TopicLoRA	52.98	-	52.25	-	42.81	-
w/o GRL	50.18	2.80	46.06	6.19	36.64	6.17
w/o MCV	48.42	4.56	42.82	9.43	36.52	6.29
w/o KLD	44.56	8.42	44.61	7.64	35.07	7.74

Table 3: Results of Ablation Experiments. Here, ↓ represents the declines of variant configurations in accuracy.

`Normal approach`, which trains data sequentially, the `Alternating approach`, which interleaves data, and the `Random approach`, which shuffles data from all topics before training. Figure 5 shows that data input order impacts model performance with the same samples. Specifically, on alternating and random approaches, TopicLoRA reduces the RMSE values by 0.165 and 0.563 compared to Baseline, respectively. This shows the robustness of our method in confusing training environments.

Case Study

Figure 6 presents four representative cases of model reasoning across different topics, comparing the relevance scores of TopicLoRA, TopicLoRA *w/o* KLD, HydraLoRA and GPT-4o. The four TopicLoRA experts in this scene are *Language Generation & Transformation*, *Semantic Understanding & Logic Reasoning*, *Reading Comprehension & Knowledge Retrieval*, and *Sentiment, Context & Interaction Modeling*. As shown in Table 4, *Example*: $\mathcal{C}_1$ *Input*: *Which of the following best describes the organ that collects urine in the body?[Bladder, Kidney, Ureter, Urethra]*. TopicLoRA rated Expert 3 as highly relevant(0.37) and Expert 4 as less relevant(0.13), which matches the scores given by model. Additionally, the relevance score of HydraLoRA has a distance of 0.177 (calculated using KL divergence) from the GT. How-

Case	Instance	Probability			
		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$
$\mathcal{C}_1$	Which of the following best describes the **structure** that collects **urine** in the body? [*Bladder, Kidney, Ureter, Ureth*]	25%	24%	**37%**	13%
$\mathcal{C}_2$	**Which entity** is **they** this sentence? **Medics** have a tough job of trying to heal the wounded since **they** normally have to fix them in **hostile environments**.	24%	**26%**	**27%**	23%
$\mathcal{C}_3$	**Ich wünsche mir genauere Angaben über die Beschäftigung der örtlichen Bevölkerung auf den Booten**. In English?	**38%**	28%	20%	14%
$\mathcal{C}_4$	Which of the following **statistics** provides the most information about how spread out a **distribution of scores** is? [*variance, mean, range, median*]	15%	20%	**44%**	21%

Table 4: Expert Relevance Scores of Four Representative Cases Evaluated by TopicLoRA, with Instances and Relevance Probabilities Assigned to Each Expert.

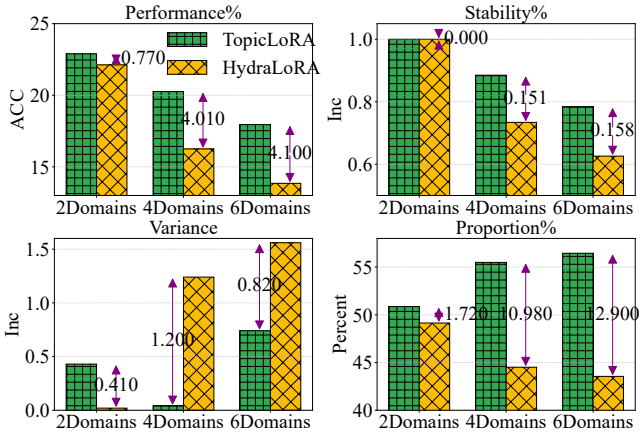

Figure 4: TopicLoRA vs. HydraLoRA across topic counts. Subplots: Accuracy, Stability (retention with more topics), Variance, and Proportion (accuracy share).

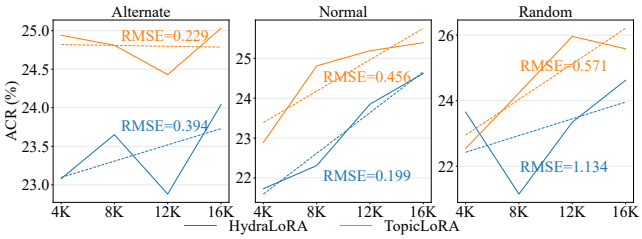

Figure 5: Performance of TopicLoRA and HydraLoRA under different orders of data input. Subplots (Alternate, Normal, Random) represent interleaved, sequential, and random inputs. Solid lines show model performance under different training strategies while dashed lines are linear fits. Residual RMSE indicates trend stability over epochs.

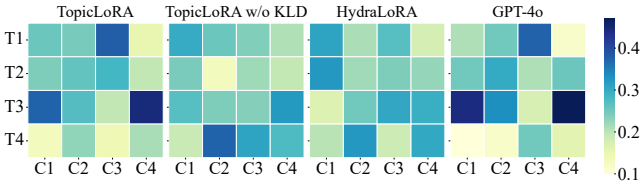

Figure 6: Relevance scores of expert modules across four representative cases. The figure shows four 4×4 heatmaps representing the relevance distributions of experts (Ti) over input cases (Ci) for TopicLoRA, TopicLoRA without KLD, HydraLoRA, and GPT-4o as ground truth.

ever, with the assistance of the prior knowledge framework, the distances for TopicLoRA and TopicLoRA *w/o* KLD are significantly reduced to 0.014 and 0.087, respectively.

Related Work

Mixture of Experts Mixture of Experts (MoE) (He et al. 2025; Kong et al. 2024; Jiang et al. 2024; Seo, Kim, and Shin 2025; Yu, Yang, and Yi 2025; Liu et al. 2025) scales LLMs by activating only selected subnetworks for efficient computation. However, expert overlap and parameter underuse—termed knowledge hybridity and redundancy—limit specialization. Recent works address these issues through improved structures and training strategies (Lu et al. 2024; Chowdhury et al. 2024; Xie et al. 2024). In summary, they promote more specialized and efficient MoE architectures.

MoE-Based PEFT Recently, a growing number of research studies have integrated low-rank adaptation with mixture-of-experts frameworks (Luo et al. 2024; Ning et al. 2025; Li et al. 2025). For instance, MoSLoRA (Wu et al. 2024) integrates sparse gating with LoRA to enable efficient multi-task fine-tuning by updating only a small subset of expert modules, keeping the backbone frozen. MoE-LoRA (Luo et al. 2024) employs modality-specific LoRA experts using MoE routing to mitigate cross-modal conflicts. LoRAMoE (Dou et al. 2024) enhances LLM stability by combining low-rank adaptation with expert routing in a MoE plugin. These methods advance the integration of PEFT and MoE architectures.

Conclusion

This paper presents TopicLoRA, a training framework that aligns MoE expert routing with topic priors derived from topic partitioning and refined through graph representation learning. The method strengthens expert specialization with explicit priors and improves robustness via a KL-divergence auxiliary loss. Experiments show that TopicLoRA consistently outperforms strong baselines.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by National Science Foundation of China (Grant Nos.62376057), the Start-up Research Fund of Southeast University (RF1028623234) and the Big Data Computing Center of Southeast University.

References

Alignment-Lab-AI. 2024. Lawyer-instruct.

Bai, T.; Yu, Y.; Huang, L.; and et al, X. Z. 2024. GraphLoRA: Empowering LLMs Fine-Tuning via Graph Collaboration of MoE. *arXiv preprint arXiv:2412.16216*.

bench authors, B. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Chalkidis, I.; Garneau, N.; Goanta, C.; and et al, K. D. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *ACL*.

Chen, M.; Tworek, J.; Jun, H.; and et al, Y. Q. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chowdhury, M. N. R.; Wang, M.; El Maghraoui, K.; and et al, W. N. 2024. A Provably Effective Method for Pruning Experts in Fine-tuned Sparse Mixture-of-Experts. In *ICML*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; and et al, C. M. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. In *EMNLP*.

Conover, M.; Hayes, M.; Mathur, A.; Xie, J.; et al. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. https://www.databricks.com/blog/free-dolly. Accessed: 2025-11-14.

Dai, D.; Deng, C.; Zhao, C.; Xu RX et al, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115.

Dou, S.; Zhou, E.; Liu, Y.; and et al, G. S. 2024. LoRAMoE: Alleviating World Knowledge Forgetting in Large Language Models via MoE-Style Plugin. In *ACL*.

Hamilton, W. L. 2020. *Graph representation learning*. Morgan & Claypool Publishers.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

He, Y.; Liu, Y.; Liang, C.; and Awadalla, H. H. 2025. Efficiently Editing Mixture-of-Experts Models with Compressed Experts. *arXiv preprint arXiv:2503.00634*.

Hendrycks, D.; Burns, C.; Basart, S.; and et al, A. Z. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hu, E. J.; Shen, Y.; Wallis, P.; and et al, Z. A.-Z. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; and et al, M. A. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiao, P.; Wu, X.; Zhu, B.; and et al, C. J. 2024. Rode: Linear rectified mixture of diverse experts for food large multi-modal models. *arXiv preprint arXiv:2407.12730*.

Kong, R.; Li, Y.; Feng, Q.; and et al, W. W. 2024. SwapMoE: Serving Off-the-shelf MoE-based Large Language Models with Tunable Memory Budget. In *ACL*.

Le, M.; Nguyen, C.; Nguyen, H.; Tran, Q.; Le, T.; and Ho, N. 2024. Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts. *arXiv preprint arXiv:2410.02200*.

Lepikhin, D.; Lee, H.; Xu, Y.; and et al, C. D. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*.

Li, D.; Wang, N.; Zhang, Z.; and Yin, H. e. a. 2025. DynMoLE: Boosting Mixture of LoRA Experts Fine-Tuning with a Hybrid Routing Mechanism. *arXiv preprint arXiv:2504.00661*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL*.

Li, Y.; Li, Z.; Zhang, K.; and et al, D. R. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*.

Lin, B.; Tang, Z.; Ye, Y.; and et al, C. J. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Lin, T.; Liu, J.; Zhang, W.; and et al, L. Z. 2025. Teamlora: Boosting low-rank adaptation with expert collaboration and competition. In *ACL*.

Liu, H.; Tam, D.; Muqeeth, M.; and et al, M. J. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *NeurIPS*.

Liu, X.; Ji, K.; Fu, Y.; and et al, T. W. 2022b. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *ACL*.

Liu, Y.; Ma, Y.; Chen, S.; and et al, D. Z. 2024. PERFT: Parameter-Efficient Routed Fine-Tuning for Mixture-of-Expert Model. *arXiv preprint arXiv:2411.08212*.

Liu, Z.; Wu, H.; She, R.; and Fu, X. e. a. 2025. MoLAE: Mixture of Latent Experts for Parameter-Efficient Language Models. *arXiv preprint arXiv:2503.23100*.

Lu, X.; Liu, Q.; Xu, Y.; and et al, Z. A. 2024. Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models. In *ACL*.

Luo, T.; Lei, J.; Lei, F.; and et al, L. W. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

Ning, L.; Lara, H.; Guo, M.; and Rastogi, A. 2025. MoDE: Effective Multi-task Parameter Efficient Fine-Tuning with a Mixture of Dyadic Experts. In *NAACL*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; and et al, H. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*.

Seo, J.; Kim, J.; and Shin, H. 2025. MoFE: Mixture of Frozen Experts Architecture. In *NAACAL*.

Suzgun, M.; Scales, N.; Schärli, N.; and et al, G. S. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*.

Tian, C.; Shi, Z.; Guo, Z.; and et al, L. L. 2024. HydraLoRA: An Asymmetric LoRA Architecture for Efficient Fine-Tuning. In *Neurips*.

Touvron, H.; Martin, L.; Stone, K.; and et al, A. P. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Truong, T.; Nguyen, C.; Nguyen, H.; Le, M.; Le, T.; and Ho, N. 2025. RepLoRA: Reparameterizing Low-Rank Adaptation via the Perspective of Mixture of Experts. *arXiv preprint arXiv:2502.03044*.

Wang, Z.; Chen, D.; Dai, D.; and et al, X. R. 2024. Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models. In *EMNLP*.

Wei, J.; Bosma, M.; Zhao, V.; and et al, G. K. 2021. Finetuned Language Models are Zero-Shot Learners. In *ICLR*.

Wu, T.; Wang, J.; Zhao, Z.; and Wong, N. 2024. Mixture-of-Subspaces in Low-Rank Adaptation. In *EMNLP*.

Xie, Z.; Zhang, Y.; Zhuang, C.; and et al, S. Q. 2024. Mode: A mixture-of-experts model with mutual distillation among the experts. In *AAAI*.

Xu, Y.; Xie, L.; Gu, X.; Chen, X.; Chang, H.; Zhang, H.; Chen, Z.; Zhang, X.; and Tian, Q. 2023. Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717*.

Xu, Z.; Wang, P.; Ke, W.; Li, G.; Liu, J.; Ji, K.; Chen, X.; and Wu, C. 2024. Incorporating Schema-Aware Description into Document-Level Event Extraction. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6597–6605. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yu, B.; Yang, Z.; and Yi, X. 2025. MoKA:Parameter Efficiency Fine-Tuning via Mixture of Kronecker Product Adaption. In *COLING*.

Zhang, Q.; Chen, M.; Bukharin, A.; and et al, P. H. 2023. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *ICLR*.

Zhu, T.; Qu, X.; Dong, D.; and et al, R. J. 2024. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *EMNLP*.