D^{2}MoRA: Diversity-Regulated Asymmetric MoE-LoRA Decomposition for Efficient Multi-Task Adaptation

Jianhui Zuo[1], Xuemeng Song[2*], Haokun Wen[3,4], Meng Liu[5], Yupeng Hu[1], Jiuru Wang[6], Liqiang Nie[3*]

[1]School of Software, Shandong University
[2]Department of Computer Science and Engineering, Southern University of Science and Technology
[3]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[4]School of Data Science, City University of Hong Kong
[5]School of Computer and Artificial Intelligence, Shandong Jianzhu University
[6]School of Computer Science and Engineering, Linyi University
{zuojianhuisdu, sxmustc}@gmail.com, haokunwen@outlook.com, mengliu.sdu@gmail.com, huyupeng@sdu.edu.cn, wangjiuru@lyu.edu.cn, nieliqiang@gmail.com

Abstract

Low-Rank Adaptation (LoRA) has emerged as a powerful parameter-efficient fine-tuning method for adapting large language models to downstream tasks. Recent studies have leveraged Mixture-of-Experts (MoE) mechanism to effectively integrate multiple LoRA modules, facilitating efficient parameter adaptation for multi-task scenarios. It has been shown that fostering knowledge sharing across LoRA experts can greatly enhance parameter adaptation efficiency. However, the existing approach for LoRA expert knowledge sharing still faces two key limitations: constrained functional specialization and induced expert homogenization. To address these issues, we propose a novel diversity-regulated asymmetric MoE-LoRA decomposition framework, which achieves flexible knowledge sharing through asymmetric expert decomposition and guarantees expert diversity with a dual orthogonality regularization. Extensive experiments on eight public benchmarks, spanning both multi-task and single-task settings, demonstrate the superiority of our approach over existing methods.

Code — https://github.com/softwavec/D2MoRA

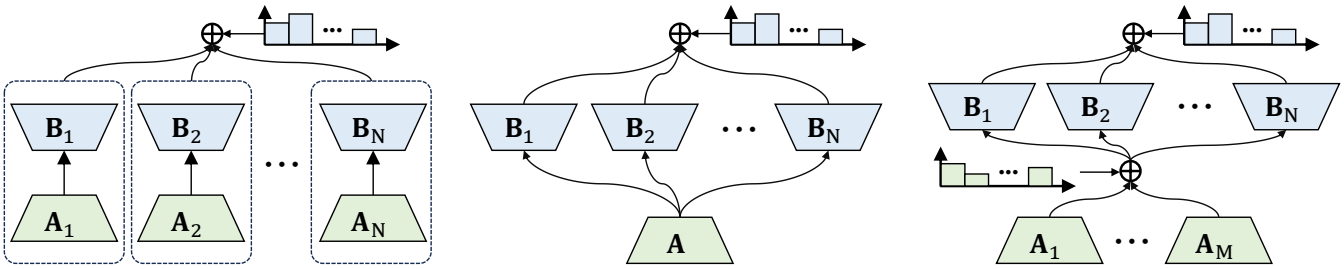
Introduction

Large Language Models (LLMs) have demonstrated remarkable potential in advancing a wide range of downstream tasks, such as commonsense reasoning and question answering. However, adapting these models to specific tasks via Full Fine-Tuning (FFT) is often impractical due to their massive parameter counts, which incur prohibitive computational and memory overhead. To address this challenge, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a critical area of research, focusing on optimizing a minimal number of parameters while maintaining performance comparable to FFT. Among PEFT methods, Low-Rank Adaptation (LoRA) (Hu et al. 2022) has gained significant traction. The basic idea of LoRA is to introduce two trainable low-rank matrices, $\mathbf{A}$ (down-projection) and $\mathbf{B}$ (up-projection). By only updating these matrices, LoRA enables efficient task-specific adaptation while keeping the original pretrained model weights fixed.

To enhance LoRA's performance in multi-task scenarios involving heterogeneous data, recent studies have explored integrating multiple LoRA modules via the Mixture-of-Experts (MoE) mechanism (Jacobs et al. 1991). According to their $\mathbf{A}$-$\mathbf{B}$ interaction paradigms, these methods fall into two categories: one-to-one pairing (Liu et al. 2024a) and one-to-many pairing (Tian et al. 2024). As shown in Figure 1, the one-to-one pairing paradigm utilizes parallel low-rank adapters, with each $\mathbf{A}$ matrix exclusively paired with a fixed $\mathbf{B}$ matrix forming independent experts, where a routing function dynamically adjusts their contributions for parameter adaptation. While effective, this isolated structure inherently limits knowledge transfer among experts, potentially resulting in redundant parameters and suboptimal adaptation performance. In contrast, the one-to-many pairing paradigm employs a shared $\mathbf{A}$ matrix for all $\mathbf{B}$ matrices, with routing conducted only over the $\mathbf{B}$ matrices. This enables cross-expert knowledge sharing while boosting parameter efficiency.

Despite the promising progress made by the one-to-many pairing paradigm, it still suffers from two key limitations. **L1: Constrained Functional Specialization.** While the one-to-many pairing paradigm enhances parameter adaptation by facilitating knowledge sharing, it constrains all experts to operate within a single low-rank subspace defined by the shared $\mathbf{A}$ matrix. This design makes a strong assumption that expert functional specialization can be achieved exclusively through the up-projection matrices ($\mathbf{B}$), ignoring scenarios where optimal adaptation requires specialized transformations in both projection phases. Additionally, relying on a single shared $\mathbf{A}$ may compromise parameter efficiency, as its rigidity could necessitate introducing additional $\mathbf{B}$ matrices to compensate, ultimately degrading performance. **L2: Induced Expert Homogenization.** While one-to-many pairing improves parameter adaptation over one-to-one schemes, it inadvertently induces expert homogenization by forcing all up-projection matrices $\mathbf{B}$ to share the same down-projection matrix $\mathbf{A}$. This violates the core

*Corresponding author.



(a) One-to-One Pairing (b) One-to-Many Pairing (c) Many-to-Many Pairing

Figure 1: Comparison of MoE-enhanced LoRA: (a) One-to-one pairing with independent experts; (b) One-to-many pairing enabling knowledge sharing; (c) Our D^{2}MoRA with asymmetric many-to-many pairing for flexible cross-expert sharing.

MoE principle that experts should develop complementary, specialized functionalities. Our t-SNE (Maaten and Hinton 2008) visualizations of expert diversity in Figures 2(a)–(c) empirically confirm this effect: HydraLoRA (one-to-many pairing) exhibits significant overlap in expert embeddings (marked by distinct colors), whereas MOELoRA (one-to-one pairing) maintains clear separation.[1] This reveals a fundamental trade-off—shared knowledge transfer improves parameter efficiency but risks expert homogenization, while independent experts preserve heterogeneity at the cost of reduced efficiency. Thus, balancing parameter efficiency and expert heterogeneity requires careful consideration.

To address these limitations, we propose a diversity-regulated asymmetric MoE-LoRA decomposition framework (D^{2}MoRA) for efficient multi-task adaptation. This framework enables flexible cross-expert knowledge sharing while preserving expert-specialized parameter adaptation. Unlike conventional approaches that treat each $(\mathbf{A}, \mathbf{B})$ pair as an integral expert unit, we reformulate the expert structure through expert decomposition. Specifically, we introduce two independent sets of base experts: $\mathcal{A} = \{\mathbf{A}_i \in \mathbb{R}^{r\times d_2}\}_{i=1}^{M}$ and $\mathcal{B} = \{\mathbf{B}_j \in \mathbb{R}^{d_1\times r}\}_{j=1}^{N}$ for modeling down-projection and up-projection patterns, respectively. As shown in Figure 1(c), this decomposition facilitates many-to-many $\mathbf{A}$-$\mathbf{B}$ pairing, which not only enables flexible knowledge sharing but also reduces parameter usage compared to the one-to-many pairing paradigm. Crucially, D^{2}MoRA does not require $M = N$, allowing the numbers of down-projection (M) and up-projection (N) experts to scale independently according to their respective complexities. This flexibility supports resource-efficient allocation tailored to the specific demands of each projection type.

Based on the independent base experts $\mathcal{A}$ and $\mathcal{B}$, we decouple the integration of multiple LoRA experts for parameter adaptation into two steps: 1) *Sample-Aware Down-Projection Expert Mixture*: Dynamically weights the down-projection experts ($\mathbf{A}$) based on input features to construct a task-specific low-rank projection. 2) *Low-Rank Embedding-Aware Up-Projection Expert Mixture*: Further specializes the adaptation by weighting up-projection experts ($\mathbf{B}$) conditioned on the projected low-rank embeddings. To prevent expert homogenization caused by knowledge sharing, we introduce dual orthogonality regularization. This regularization is applied separately to the two sets of base experts, $\mathcal{A}$ and $\mathcal{B}$, enhancing their diversity and effectively mitigating parameter redundancy.

Our main contributions can be summarized as follows:

- We propose a novel diversity-regulated, asymmetric LoRA decomposition framework (D^{2}MoRA) for efficient adaptation, enabling flexible cross-expert knowledge sharing. To the best of our knowledge, we are among the first to introduce an asymmetric many-to-many pairing mechanism between down-projection ($\mathbf{A}$) and up-projection ($\mathbf{B}$) matrices.
- We experimentally identify the expert homogenization issue that arises from enabling knowledge sharing across multiple LoRA experts. To address this, we propose dual orthogonality regularization, which effectively enhances the diversity of both down-projection ($\mathbf{A}$) and up-projection ($\mathbf{B}$) experts.
- We conduct extensive experiments on eight public datasets, demonstrating the effectiveness and generalization ability of our framework.

Related Work

Parameter-Efficient Fine-tuning

Parameter-Efficient Fine-Tuning (PEFT) aims to reduce the number of trainable parameters when adapting pre-trained models to downstream tasks. This can be achieved by introducing lightweight trainable components into the model (Houlsby et al. 2019; He et al. 2021; Hu et al. 2022) or by fine-tuning only a small subset of important parameters (He et al. 2023; Lawton et al. 2023; Das et al. 2023). Among existing PEFT methods, LoRA (Hu et al. 2022) has attracted significant attention, spurring extensive research on its enhancement. For example, AdaLoRA (Zhang et al. 2023b) dynamically adjusts the rank during training; DoRA (Liu et al. 2024b) decomposes pre-trained weights into magnitude and direction for separate updates; and MoSLoRA (Wu et al. 2024) constructs a mixture over LoRA subspaces. Other variants, such as KD-LoRA (Azimi et al. 2024) and QLoRA (Dettmers et al. 2023), further explore knowledge distillation and quantization to enhance parameter adaptation.

[1] Detailed explanation is provided in the Methodology section.

MoE-Enhanced Low-Rank Adaptation

As single-LoRA methods underperform in multi-task scenarios, researchers have introduced the MoE mechanism to integrate multiple LoRA modules to improve parameter adaptation. Indeed, heterogeneous data benefit from specialized modeling (Zhang et al. 2023a, 2025). Existing methods can be broadly categorized into two classes: *one-to-one pairing* and *one-to-many pairing*. One-to-one pairing methods adopt multiple parallel LoRA experts, each with a fixed pair of low-rank matrices ($\mathbf{A}$ and $\mathbf{B}$). Various designs have been proposed to improve this architecture. For example, MOELoRA (Liu et al. 2024a) introduces a task-aware expert selection mechanism based on task labels, while LoRAMoE (Dou et al. 2023) employs input-dependent routing and additionally incorporates localized balancing constraint to alleviate world knowledge forgetting. MixLoRA (Li et al. 2024) applies a top-k expert selection strategy along with an auxiliary load balance loss. Octavius (Chen et al. 2023) applies instance-level routing for multimodal learning, and MoCLE (Gou et al. 2023) activates task-customized model parameters based on instruction cluster. While effective, these methods overlook knowledge sharing among LoRA experts. In contrast, the one-to-many pairing approach, as in HydraLoRA (Tian et al. 2024), adopts an asymmetric architecture that shares a down-projection matrix and performs expert routing only on the up-projection matrices to enhance adaptation efficiency.

However, this paradigm still suffers from two key limitations: (1) constrained functional specialization of experts due to the shared $\mathbf{A}$ matrix, and (2) induced expert homogenization. To address these issues, we propose D^{2}MoRA, which decouples the conventional $(\mathbf{A}, \mathbf{B})$ pair-based expert structure into two independent sets of base experts ($\mathcal{A}$ and $\mathcal{B}$). This enables flexible many-to-many interactions between $\mathbf{A}$ and $\mathbf{B}$. Furthermore, we introduce the dual orthogonality regularization to preserve expert diversity and promote efficient parameter adaptation.

Preliminaries

LoRA introduces a trainable low-rank decomposition to the pretrained weight matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ for efficient parameter adaptation as follows,

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{B}\mathbf{A}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_2}$ projects inputs into a low-dimensional space, $\mathbf{B} \in \mathbb{R}^{d_1 \times r}$ projects back to the output dimension, and $r \ll \min(d_1, d_2)$. $\mathbf{W}' \in \mathbb{R}^{d_1 \times d_2}$ is the updated weight matrix used for forward calculation. During fine-tuning, LoRA only updates $\mathbf{A}$ and $\mathbf{B}$ while keeping $\mathbf{W}$ frozen. Conventionally, $\mathbf{A}$ is initialized with Kaiming Uniform (He et al. 2015), while $\mathbf{B}$ is initialized as all zeros, making $\mathbf{W}' = \mathbf{W}$ at the beginning of fine-tuning.

MoE-LoRA (Dou et al. 2023; Liu et al. 2024a) extends the standard LoRA by incorporating the MoE mechanism. Rather than using a single pair of low-rank matrices, this approach employs multiple pairs of matrices, denoted as $\{(\mathbf{A}_1, \mathbf{B}_1), \cdots, (\mathbf{A}_E, \mathbf{B}_E)\}$, where each pair $(\mathbf{A}_i, \mathbf{B}_i)$ constitutes an expert and E represents the total number of LoRA experts. The parameter adaptation is performed through a gated combination as follows:

$$\mathbf{W}' = \mathbf{W} + \sum_{i=1}^{E} \omega_i \mathbf{B}_i \mathbf{A}_i, \tag{2}$$

where $\omega_i = g(\cdot)$ represents the gating weight produced by a learnable router network g (implemented as either a linear projection or feedforward neural network), controlling each expert's contribution.

Methodology

Beyond existing one-to-one and one-to-many LoRA pairing paradigms, we propose D^{2}MoRA, which simultaneously achieves flexible cross-expert knowledge sharing and expert-specialized parameter adaptation. As a key innovation, it decomposes the conventional $(\mathbf{A}, \mathbf{B})$ pair-based expert structure into two sets of base experts $\mathcal{A}$ and $\mathcal{B}$. By this design, it allows flexible many-to-many pairing between $\mathbf{A}$ and $\mathbf{B}$ base experts. D^{2}MoRA consists of two core components: *two-stage hierarchical adaptation* and *dual orthogonality regularization*. The former facilitates dynamic knowledge sharing among LoRA experts, while the latter enhances diversity across them.

Two-stage Hierarchical Adaptation Based on the independent expert sets $\mathcal{A}$ and $\mathcal{B}$, we decouple the expert mixture into two steps. 1) *Sample-Aware Down-Projection Expert Mixture*: A sample-aware router dynamically weights the contributions of the low-rank $\mathbf{A}$ base experts to perform down-projection. 2) *Low-Rank Embedding-Aware Up-Projection Expert Mixture*: A router conditioned on the projected low-rank embedding adaptively weights the contributions of $\mathbf{B}$ base experts to reconstruct the output dimension.

Step 1: Sample-Aware Down-Projection Expert Mixture. Given an input $\mathbf{x} \in \mathbb{R}^{d_2}$, we first compute the importance weights over the down-projection experts using a softmax-based gating function,

$$\boldsymbol{\alpha}(\mathbf{x}) = \mathrm{Softmax}(g_A(\mathbf{x})), \tag{3}$$

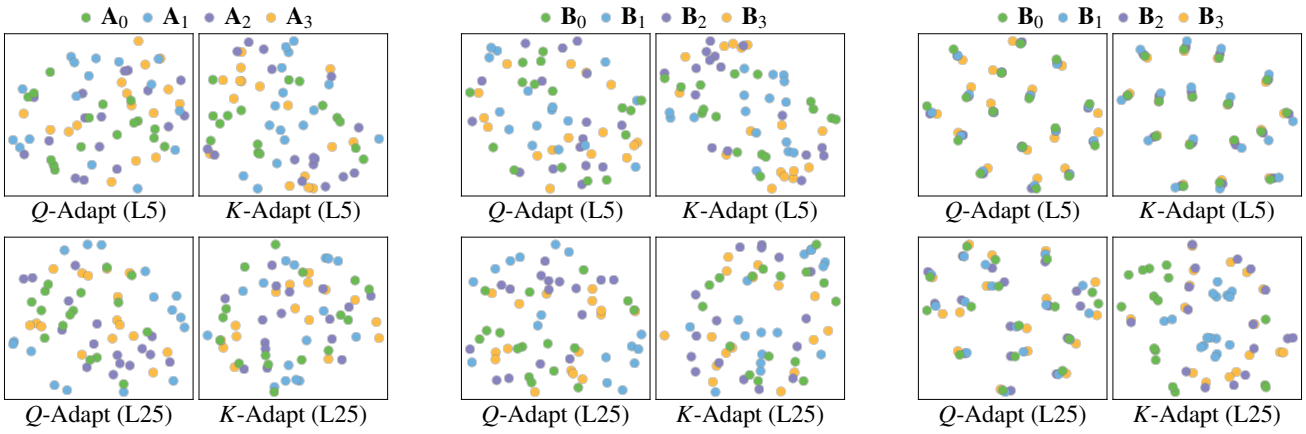
where $g_A(\cdot)$ is a router implemented as a linear transformation of size $d_1 \times M$, $\boldsymbol{\alpha}(\mathbf{x}) \in \mathbb{R}^M$ represents the importance scores for the M down-projection experts. The latent representation is then obtained by aggregating the projections from the set of $\mathbf{A}$ matrices as follows,

$$\mathbf{h}_r = \sum_{i=1}^{M} \alpha_i(\mathbf{x}) \mathbf{A}_i \mathbf{x}, \tag{4}$$

where $\mathbf{h}_r \in \mathbb{R}^r$ represents the projected hidden feature conditioned on the weighted expert composition.

Step 2: Low-Rank Embedding-Aware Up-Projection Expert Mixture. Next, we route the up-projection experts. To better utilize the information encoded by the down-projection experts, we use the intermediate representation $\mathbf{h}_r$ rather than the original input $\mathbf{x}$, to combine up-projection experts as follows,

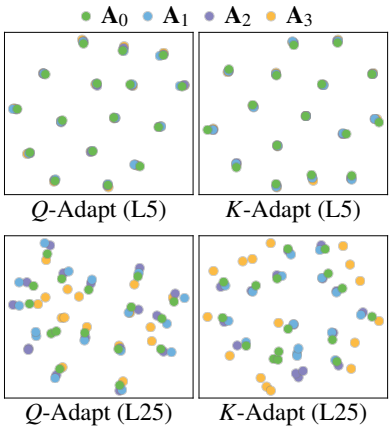
$$\boldsymbol{\beta}(\mathbf{h}_r) = \mathrm{Softmax}(g_B(\mathbf{h}_r)), \tag{5}$$



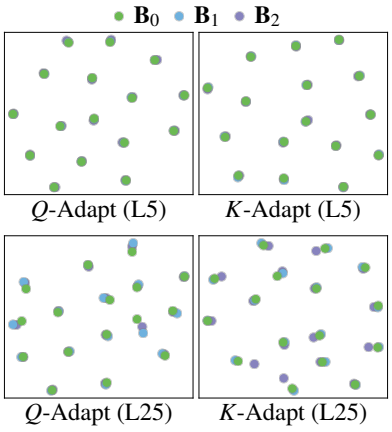
(a) Expert diversity ($\mathbf{A}$) of MOELoRA for $Query/Key$ adaptation

(b) Expert diversity ($\mathbf{B}$) of MOELoRA for $Query/Key$ adaptation

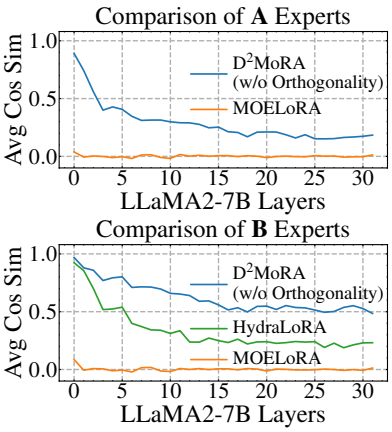
(c) Expert diversity ($\mathbf{B}$) of HydraLoRA for $Query/Key$ adaptation



(d) Expert diversity ($\mathbf{A}$) of D^{2}MoRA (w/o Orthogonality) for $Query/Key$ adaptation



(e) Expert diversity ($\mathbf{B}$) of D^{2}MoRA (w/o Orthogonality) for $Query/Key$ adaptation



(f) Average cosine similarity among experts of the three methods

Figure 2: Expert diversity visualization of MOELoRA$_{\{M=N=4,\, r=16\}}$, HydraLoRA$_{\{M=1,\, N=4,\, r=16\}}$, and our D^{2}MoRA$_{\{M=3,\, N=4,\, r=16\}}$ (w/o Orthogonality) with t-SNE, where experts are learned through parameter adaptation for LLaMA2-7B's self-attention layers using the Commonsense 170K dataset. MOELoRA requires $M=N$ and HydraLoRA is constrained to $M=1$, while D^{2}MoRA has no such constraints. (a-b): visualizations of experts $\mathbf{A}$ and $\mathbf{B}$ in MOELoRA. (c) visualizations of experts $\mathbf{B}$ in HydraLoRA. (d-e): visualizations of experts $\mathbf{A}$ and $\mathbf{B}$ in D^{2}MoRA (w/o Orthogonality). Each expert spans r points (one per rank), sharing the same color. (f): Cosine similarity of experts per layer across the three methods.

where $g_B(\cdot)$ is the router implemented as a linear transformation of size $r \times N$, and $\boldsymbol{\beta}(\mathbf{h}_r) \in \mathbb{R}^N$ represents the importance scores over the N up-projection experts.

The final output of D^{2}MoRA is computed as a weighted sum of the up-projection results as follows,

$$\begin{aligned}\Delta\mathbf{y} &= \sum_{j=1}^{N} \beta_j(\mathbf{h}_r)\mathbf{B}_j\mathbf{h}_r, \\ &= \sum_{j=1}^{N} \beta_j \left(\sum_{i=1}^{M} \alpha_i(\mathbf{x})\mathbf{A}_i\mathbf{x}\right) \mathbf{B}_j \left(\sum_{i=1}^{M} \alpha_i(\mathbf{x})\mathbf{A}_i\mathbf{x}\right),\end{aligned} \tag{6}$$

where $\Delta\mathbf{y}$ is the final output of D^{2}MoRA, which is added to the standard output $\mathbf{y}$ (computed using the frozen pre-trained weight matrix $\mathbf{W}$) for downstream task loss calculation. Let $\mathcal{L}_{\text{task}}$ denote the task-specific loss function used during fine-tuning. Notably, only the two sets of low-rank matrices $\mathcal{A}$ and $\mathcal{B}$, along with the parameters of the routers $g_A(\cdot)$ and $g_B(\cdot)$, are updated during fine-tuning, while $\mathbf{W}$ remains frozen.

Dual Orthogonality Regularization Prior to introducing our orthogonality regularization, we present t-SNE visualizations of expert diversity for three representative methods in Figure 2, including (1) standard MoE-LoRA (without knowledge sharing), (2) HydraLoRA (which enables knowledge sharing via a unified low-rank matrix $\mathbf{A}$), and (3) our D^{2}MoRA without orthogonality constraints. Specifically, we perform parameter adaptation for LLaMA2-7B on its $Query$, Key, and $Value$ projections using the Commonsense 170K dataset (Hu et al. 2023). Notably, due to space

limitations, we only show visualizations of the $Query$ and Key projections at Layer 5 (L5) and Layer 25 (L25).

As shown in Figures 2(a) and (b), the learned $\mathbf{A}$ experts in MOELoRA exhibit high diversity: different-color points remain well separated, indicating that each expert specializes in distinct transformations. In contrast, Figures 2(c)-(e) show that the $\mathbf{B}$ experts in HydraLoRA and our D^{2}MoRA (without orthogonality constraints) tend to collapse into similar functionalities, as points of different colors approximately overlap. This suggests that enabling knowledge sharing across experts may lead to homogeneous behavior. This trend is further supported by Figure 2(f), where MOELoRA yields near-zero average cosine similarity, in contrast to the higher similarity values observed in HydraLoRA and D^{2}MoRA. Such high similarity implies redundant parameterization and weak expert specialization. These findings underscore the necessity of orthogonality regularization to explicitly enforce directional diversity and prevent expert collapse during knowledge sharing.

To address this issue, we introduce orthogonality regularization separately to the base experts $\{\mathbf{A}_i\}_{i=1}^{M}$ and $\{\mathbf{B}_j\}_{j=1}^{N}$ as follows,

$$\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{M-1} \sum_{u=i+1}^{M} \left|\tilde{\mathbf{a}}_i^\top \tilde{\mathbf{a}}_u\right| + \sum_{j=1}^{N-1} \sum_{v=j+1}^{N} \left|\tilde{\mathbf{b}}_j^\top \tilde{\mathbf{b}}_v\right|, \quad (7)$$

where

$$\tilde{\mathbf{a}}_i = \frac{\text{vec}(\mathbf{A}_i)}{\|\text{vec}(\mathbf{A}_i)\|_2 + \epsilon}, \tilde{\mathbf{b}}_j = \frac{\text{vec}(\mathbf{B}_j)}{\|\text{vec}(\mathbf{B}_j)\|_2 + \epsilon}, \quad (8)$$

Here, $\text{vec}(\cdot)$ flattens the matrix into a vector, and ϵ denotes a small constant (*e.g.,* 10^{-6}) to avoid division by zero. This regularization encourages each expert to learn a distinct parameter distribution, thereby enhancing the representational capacity of the mixture model.

Ultimately, the final optimization objective is defined as,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{ortho}}, \quad (9)$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss, and λ controls the strength of the orthogonality regularization.

Experiment

Experiment Setting

Benchmark Following prior work (Wu et al. 2024), we evaluate the model on the Commonsense 170K dataset (Hu et al. 2023), comprising eight commonsense reasoning sub-datasets (each representing a task): **BoolQ** (Clark et al. 2019), **PIQA** (Bisk et al. 2020), **SIQA** (Sap et al. 2019), **HellaSwag** (Zellers et al. 2019), **WinoGrande** (Sakaguchi et al. 2021), **ARC-c** (Clark et al. 2018), **ARC-e** (Clark et al. 2018) and **OBQA** (Mihaylov et al. 2018). These datasets span diverse domains (*e.g.,* Wikipedia, natural science, physical interaction, social interaction, and video captioning) and include various task formats, such as text classification, question answering, sentence completion, and fill-in-the-blank. This diversity enables a comprehensive evaluation of our proposed model in complex multi-task scenarios involving heterogeneous data.

Baseline We compare our approach with five baseline methods, including three single-LoRA methods and two MoE-enhanced multi-LoRA methods:

- **LoRA** (Hu et al. 2022), which employs a pair of trainable low-rank matrices for efficient parameter adaptation;
- **DoRA** (Liu et al. 2024b), which decomposes the pre-trained weights into *magnitude* and *direction* components for fine-tuning;
- **MoSLoRA** (Wu et al. 2024), which decomposes LoRA into multiple subspaces and combines them through a fixed, input-agnostic mixer;
- **MOELoRA** (Liu et al. 2024a), which integrates MoE with LoRA, where each expert is a LoRA module and a task-identity-driven router is used for expert aggregation;
- **HydraLoRA** (Tian et al. 2024), which shares $\mathbf{A}$ across experts while applying MoE only to $\mathbf{B}$ matrices.

Implementation Details We fine-tune the attention projections ($Query/Key/Value$) of LLaMA-7B and LLaMA2-7B for all baselines and our method on the entire Commonsense 170K dataset. All baselines are implemented under the same training and inference settings as our method. We first configure multi-LoRA methods that involve multiple parameters, and then adjust single-LoRA methods accordingly to ensure comparable parameter budgets for fair comparison. Specifically, for multi-LoRA methods (*i.e.,* MOELoRA, HydraLoRA, and D^{2}MoRA), we configure two M, N, r settings per backbone for comprehensive comparison. For MOELoRA, we set $\{M = N = 4, r = 16\}$ and $\{M = N = 8, r = 8\}$; for HydraLoRA, $\{M = 1, N = 6, r = 16\}$ and $\{M = 1, N = 8, r = 12\}$, all applied to both LLaMA-7B and LLaMA2-7B. For D^{2}MoRA, we set $\{M = 3, N = 8, r = 8\}$ and $\{M = 3, N = 4, r = 16\}$ for LLaMA-7B, and $\{M = 3, N = 8, r = 8\}$ and $\{M = 4, N = 3, r = 16\}$ for LLaMA2-7B. For single-LoRA methods, we set $r = 64$ for parity. We set λ to 10^{-4}, and adopt a dropout rate of 0.05, a learning rate of 3×10^{-4}, and a batch size of 4 per A100 GPU (40GB).

Experimental Results

Model Comparison Table 1 shows the performance of different methods in adapting LLaMA-7B and LLaMA2-7B for multi-task learning, including both per-task and average scores. From this table, we make the following observations: 1) MoE-enhanced multi-LoRA methods generally outperform single-LoRA methods, highlighting the benefit of decomposing LoRA into multiple sub-experts for more flexible and efficient parameter adaptation. 2) Among single-LoRA methods, MoSLoRA performs best, further validating the importance of LoRA decomposition even without MoE. 3) Among multi-LoRA methods, D^{2}MoRA attains the highest performance across both LLaMA-7B and LLaMA2-7B, while MOELoRA, restricted to one-to-one $\mathbf{A}$-$\mathbf{B}$ pairings, performs the worst. This underscores the effectiveness of decoupling expert routing and applying dual orthogonality regularization, which together improve the adapter's representational capacity and generalization without increasing the number of trainable parameters.

Model	PEFT Method	Param	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-c	ARC-e	OBQA	Avg.
LLaMA-7B	LoRA$_{\{M=1,\ N=1,\ r=64\}}$	50.3M	68.47	80.09	76.56	78.83	78.69	60.75	76.56	74.60	74.32
	DoRA$_{\{M=1,\ N=1,\ r=64\}}$	51.7M	68.13	79.92	77.64	82.25	80.58	62.80	76.01	76.20	75.44
	MoSLoRA$_{\{M=1,\ N=1,\ r=64\}}$	50.7M	66.82	81.39	**78.40**	81.79	**80.98**	62.63	78.28	77.80	76.01
	MOELoRA$_{\{M=8,\ N=8,\ r=8\}}$	50.4M	69.39	79.90	76.21	81.14	80.76	62.41	78.53	78.70	75.88
	MOELoRA$_{\{M=4,\ N=4,\ r=16\}}$	50.4M	68.47	80.20	77.99	80.81	80.66	63.48	79.00	75.40	75.75
	HydraLoRA$_{\{M=1,\ N=8,\ r=12\}}$	45.6M	68.59	81.56	77.94	83.20	78.61	63.91	78.58	77.40	76.22
	HydraLoRA$_{\{M=1,\ N=6,\ r=16\}}$	46.4M	68.07	81.99	77.64	79.44	79.32	63.82	79.00	**79.20**	76.06
	D^{2}MoRA$_{\{M=3,\ N=8,\ r=8\}}$	**35.8M**	69.48	81.34	78.25	83.89	79.72	64.33	**79.21**	78.40	76.83
	D^{2}MoRA$_{\{M=3,\ N=4,\ r=16\}}$	45.6M	**69.66**	**82.86**	77.22	**85.95**	80.58	**64.68**	**79.21**	77.20	**77.17**
LLaMA2-7B	LoRA$_{\{M=1,\ N=1,\ r=64\}}$	50.3M	70.91	81.34	76.20	81.41	80.19	63.99	77.31	76.80	76.02
	DoRA$_{\{M=1,\ N=1,\ r=64\}}$	51.7M	68.65	81.12	78.45	86.64	81.06	65.02	78.24	79.20	77.30
	MoSLoRA$_{\{M=1,\ N=1,\ r=64\}}$	50.7M	68.64	82.05	77.52	87.66	80.61	67.36	81.62	79.42	78.11
	MOELoRA$_{\{M=8,\ N=8,\ r=8\}}$	50.4M	70.26	82.15	**78.81**	86.23	80.96	65.15	82.81	78.20	78.07
	MOELoRA$_{\{M=4,\ N=4,\ r=16\}}$	50.4M	70.69	81.60	77.43	83.35	**82.06**	66.55	83.54	78.70	77.99
	HydraLoRA$_{\{M=1,\ N=8,\ r=12\}}$	45.6M	69.52	82.81	78.56	87.82	80.58	67.28	81.29	79.80	78.46
	HydraLoRA$_{\{M=1,\ N=6,\ r=16\}}$	46.4M	70.07	82.66	**78.81**	87.53	80.34	66.27	81.82	78.40	78.24
	D^{2}MoRA$_{\{M=3,\ N=8,\ r=8\}}$	**35.8M**	70.40	82.26	78.76	87.72	81.53	**70.65**	**84.01**	78.80	79.27
	D^{2}MoRA$_{\{M=4,\ N=3,\ r=16\}}$	45.6M	**71.31**	**82.86**	78.40	**90.11**	81.68	67.06	83.38	**81.00**	**79.48**

Table 1: Performance (%) comparison of different methods across eight benchmarks for adapting LLaMA-7B and LLaMA2-7B in multi-task learning. Param: number of trainable parameters. The best results are shown in bold.

PEFT Method	Param	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-c	ARC-e	OBQA	Avg.
D^{2}MoRA (w/o Orthogonality)	45.6M	71.25	82.05	78.81	85.79	80.43	67.66	83.50	78.60	78.51
D^{2}MoRA (w/o Orthogonality for $\mathbf{A}$)	45.6M	70.43	83.46	77.12	88.37	81.45	67.83	82.11	80.60	78.92
D^{2}MoRA (w/o Orthogonality for $\mathbf{B}$)	45.6M	72.08	82.75	77.94	89.14	80.51	66.47	81.40	79.80	78.76
D^{2}MoRA	45.6M	71.31	82.86	78.40	90.11	81.68	67.06	83.38	81.00	79.48
HydraLoRA	45.6M	70.07	82.66	78.81	87.53	80.34	66.27	81.82	78.40	78.24
HydraLoRA (w/ Orthogonality)	45.6M	70.70	82.26	78.76	86.01	81.06	68.26	83.17	79.80	78.75

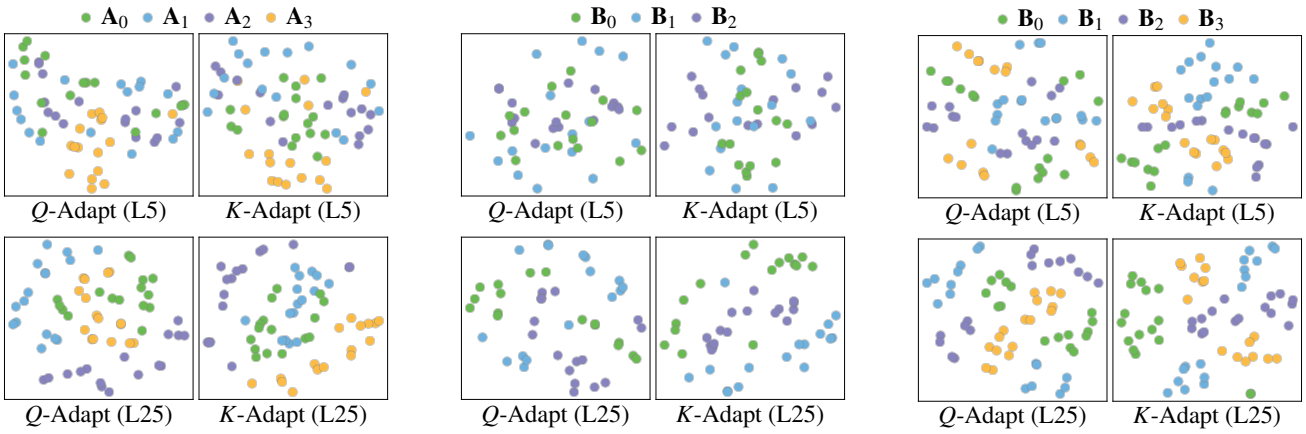
Table 2: Ablation of dual orthogonal regularization across eight benchmarks for adapting LLaMA2-7B in multi-task learning.

Ablation Study In this part, we first evaluate the effect of our dual orthogonality regularization through three variants: D^{2}MoRA (w/o orthogonality), D^{2}MoRA (w/o orthogonality for $\mathbf{A}$), D^{2}MoRA (w/o orthogonality for $\mathbf{B}$), where the corresponding orthogonality regularization terms are removed. To further assess the generalization ability of our regularization, we apply it to HydraLoRA, a model that shares a single $\mathbf{A}$ across experts without explicit diversity constraints, resulting in a new variant: HydraLoRA (w/ Orthogonality).

Table 2 presents the performance of these models in adapting LLaMA2-7B for multi-task learning, where we set $\{M = 4, N = 3, r = 16\}$ for D^{2}MoRA and $\{M = 1, N = 6, r = 16\}$ for HydraLoRA. The results show that removing the orthogonality regularization from either the $\mathbf{A}$ or $\mathbf{B}$ experts in D^{2}MoRA leads to performance drops, while applying dual regularization yields the best results without increasing the number of trainable parameters. This demonstrates that encouraging orthogonality in either expert set promotes expert diversity and expert specialization. Furthermore, HydraLoRA (w/ Orthogonality) outperforms the original HydraLoRA, validating the generalization ability of our orthogonality regularization.

We also visualize the expert distributions of D^{2}MoRA and HydraLoRA (w/ Orthogonality) using t-SNE, where experts are learned from the $Query/Key$ adaptation in LLaMA2-7B, as shown in Figure 3. To ensure comparability with Figure 2, we use the same setup for HydraLoRA: $\{M = 1, N = 4, r = 16\}$. The results show that orthogonality regularization produces well-separated expert representations, in contrast to the overlapping distributions observed in models without regularization (Figure 2). This demonstrates the effectiveness of our dual orthogonality regularization in promoting expert specialization.

Model Generalization To evaluate the generalization ability of our model, we also conduct experiments in a single-task setting, where LLaMA2-7B is fine-tuned individually on BoolQ, SIQA, ARC-e, and PIQA. Compared to the multi-task setup, we reduce the number of trainable parameters accordingly. Specifically, for MOELoRA, we set $\{M = N = 4, r = 8\}$; for HydraLoRA, we set $\{M = 1, N = 4, r = 8\}$; and for D^{2}MoRA, we set $\{M = 2, N = 3, r = 8\}$. For LoRA, DoRA, and MoSLoRA, we set $r = 32$. Other hyperparameters remain the same as in the multi-task setting.



(a) $\mathbf{A}$ of D^{2}MoRA (b) $\mathbf{B}$ of D^{2}MoRA (c) $\mathbf{B}$ of HydraLoRA (w/ orthogonality)

Figure 3: Expert diversity visualization of D^{2}MoRA$_{\{M=4,\,N=3,\,r=16\}}$ and HydraLoRA$_{\{M=1,\,N=4,\,r=16\}}$ (w/ orthogonality) for LLaMA2-7B's $Query/Key$ adaptation. Experts are represented as r points (one per rank) with shared coloring.

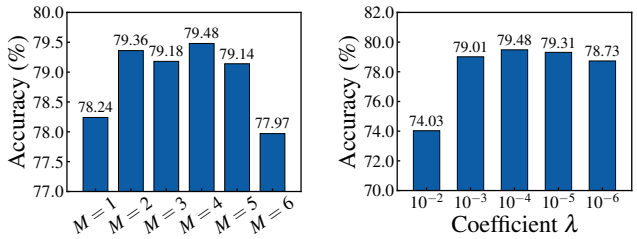
PEFT Method	Param	BoolQ	SIQA	ARC-e	PIQA	Avg.
LoRA$_{\{M=1,\,N=1,\,r=32\}}$	25.2M	71.14	78.26	79.56	78.35	76.83
DoRA$_{\{M=1,\,N=1,\,r=32\}}$	26.5M	72.20	80.12	79.34	79.22	77.72
MoSLoRA$_{\{M=1,\,N=1,\,r=32\}}$	25.3M	72.60	79.62	**81.31**	79.15	78.17
MOELoRA$_{\{M=4,\,N=4,\,r=8\}}$	25.2M	71.09	79.84	80.19	79.64	77.69
HydraLoRA$_{\{M=1,\,N=4,\,r=8\}}$	17.3M	71.87	79.43	80.21	80.65	78.04
D^{2}MoRA$_{\{M=2,\,N=3,\,r=8\}}$	**16.5M**	**72.66**	**80.45**	81.06	**81.07**	**78.81**

Table 3: Performance (%) comparison across four benchmarks for adapting LLaMA2-7B in single-task learning.

Table 3 reports the performance of all methods across four datasets for adapting LLaMA2-7B under single-task learning. As shown, D^{2}MoRA consistently achieves the best performance across all tasks while using the fewest trainable parameters. This demonstrates that our method is not limited to multi-task scenarios and can also deliver performance gains in single-task learning. One possible explanation is that even within a single task, the presence of diverse examples benefits from integrating multiple LoRA modules for more effective semantic understanding.

Sensitivity Analysis We further conduct a sensitivity analysis to examine how the number of $\mathbf{A}/\mathbf{B}$ base experts and the orthogonality regularization coefficient affect the performance of our approach. Figure 4(a) shows D^{2}MoRA's performance for adapting LLaMA2-7B under multitask learning with different $\{M, N\}$ configurations. We fix $M+N=7$ and set the rank $r=16$. We evaluate six settings: $\{M=1, N=6\}, \{M=2, N=5\}, \{M=3, N=4\}, \{M=4, N=3\}, \{M=5, N=2\}$, and $\{M=6, N=1\}$. As can be seen, the most imbalanced configurations (*i.e.,* $\{M=1, N=6\}$ and $\{M=6, N=1\}$) perform the worst, indicating that such extreme allocations constrain the model's representational capacity.

Figure 4(b) plots the model performance across different orthogonality regularization coefficients $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, with $\{M=4, N=3, r=16\}$ fixed for multi-task learning on LLaMA2-7B. The results suggest that a moderate value of λ balances expert diversity and parameter adaptation well. A large λ (*e.g.,* 10^{-2}) leads to suboptimal performance due to overemphasizing expert diversity, whereas a small λ (*e.g.,* 10^{-6}) weakens the effect of orthogonality regularization.



(a) (b)

Figure 4: Sensitivity analysis of (a) expert number $\{M, N\}$ and (b) orthogonality coefficient λ.

Conclusion

In this work, we propose D^{2}MoRA, a diversity-regulated, asymmetric MoE-LoRA decomposition framework for enhancing parameter adaptation in multi-task scenarios. Going beyond existing methods, D^{2}MoRA employs asymmetric expert decomposition to enable flexible knowledge sharing while promoting expert specialization. Through experimentation, we identify and address the expert homogenization phenomenon—caused by cross-expert knowledge transfer—using dual orthogonal regularization. Extensive experiments on eight commonsense reasoning benchmarks demonstrate the superiority of D^{2}MoRA over existing methods in both multi-task and single-task settings. Additionally, quantitative and visualization results confirm the effectiveness of our dual orthogonal regularization.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.: 62376137, No.: 624B2047, No.:62276155, No.:62376140, No.:U23A20315); the Shandong Provincial Natural Science Foundation (No.:ZR2022YQ59); the Special Fund for Taishan Scholar Project of Shandong Province.

References

Azimi, R.; Rishav, R.; Teichmann, M.; and Kahou, S. E. 2024. KD-LoRA: A Hybrid Approach to Efficient Fine-Tuning with LoRA and Knowledge Distillation. *arXiv preprint arXiv:2410.20777*.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.

Chen, Z.; Wang, Z.; Wang, Z.; Liu, H.; Yin, Z.; Liu, S.; Sheng, L.; Ouyang, W.; Qiao, Y.; and Shao, J. 2023. Octavius: Mitigating task interference in mllms via lora-moe. *arXiv preprint arXiv:2311.02684*.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Das, S. S. S.; Zhang, R. H.; Shi, P.; Yin, W.; and Zhang, R. 2023. Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning. *arXiv preprint arXiv:2311.03748*.

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.; Gao, S.; Zhao, J.; Shen, W.; Zhou, Y.; Xi, Z.; Wang, X.; Fan, X.; et al. 2023. LoRAMoE: Alleviate world knowledge forgetting in large language models via MoE-style plugin. *arXiv preprint arXiv:2312.09979*.

Gou, Y.; Liu, Z.; Chen, K.; Hong, L.; Xu, H.; Li, A.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

He, H.; Cai, J.; Zhang, J.; Tao, D.; and Zhuang, B. 2023. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11825–11835.

He, J.; Zhou, C.; Ma, X.; Berg-Kirkpatrick, T.; and Neubig, G. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

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

Hu, Z.; Wang, L.; Lan, Y.; Xu, W.; Lim, E.-P.; Bing, L.; Xu, X.; Poria, S.; and Lee, R. K.-W. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Lawton, N.; Kumar, A.; Thattai, G.; Galstyan, A.; and Steeg, G. V. 2023. Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models. *arXiv preprint arXiv:2305.16597*.

Li, D.; Ma, Y.; Wang, N.; Ye, Z.; Cheng, Z.; Tang, Y.; Zhang, Y.; Duan, L.; Zuo, J.; Yang, C.; et al. 2024. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Xu, D.; Tian, F.; and Zheng, Y. 2024a. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1104–1114.

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

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Tian, C.; Shi, Z.; Guo, Z.; Li, L.; and Xu, C.-Z. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37: 9565–9584.

Wu, T.; Wang, J.; Zhao, Z.; and Wong, N. 2024. Mixture-of-subspaces in low-rank adaptation. *arXiv preprint arXiv:2406.11909*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhang, H.; Liu, M.; Li, Y.; Yan, M.; Gao, Z.; Chang, X.; and Nie, L. 2023a. Attribute-guided collaborative learning for partial person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14144–14160.

Zhang, H.; Liu, M.; Li, Z.; Wen, H.; Guan, W.; Wang, Y.; and Nie, L. 2025. Spatial Understanding from Videos: Structured Prompts Meet Simulation Data. *arXiv preprint arXiv:2506.03642*.

Zhang, Q.; Chen, M.; Bukharin, A.; Karampatziakis, N.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023b. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.