COLLABORATION THROUGH DIVERSITY AND SPECIALIZATION VIA ORTHOGONALITY FOR IMPROVED LORA-BASED MOE

Anonymous authors
Paper under double-blind review

ABSTRACT

Combining the Mixture-of-Experts (MoE) paradigm with Low-Rank Adaptation (LoRA) enables parameter-efficient adaptation of large language models (LLMs) to downstream tasks. However, existing LoRA-based MoE approaches often suffer from expert interference and weak specialization, where experts produce correlated outputs, leading to overlapping representations and reduced effective capacity. A key challenge is to balance collaboration and specialization: experts jointly capture task-specific knowledge while reducing redundancy and conflicts across experts. To address this challenge, we propose a principled framework that characterizes expert interactions in a data-dependent manner. We show that redundancy arises from correlated expert outputs and propose an explicit regularization objective that encourages orthogonality, promoting effective specialization and mitigating interference. Furthermore, built on an asymmetric LoRA framework with shared low-rank projections, we introduce a complementary regularization that encourages diversity in the shared representations. Our experiments across diverse benchmarks demonstrate that our method encourages representational specialization and improves task performance while maintaining the efficiency of parameter-efficient fine-tuning. Our code is available at `https://anonymous.4open.science/r/OD_MoLE-A5F2`.

1 INTRODUCTION

Large Language Models (LLMs) have been powerful in multiple realistic scenarios (Touvron et al., 2023; Achiam et al., 2023; Grattafiori et al., 2024), but high fine-tuning costs make adapting these models to resource-constrained devices for downstream tasks impractical (Hu et al., 2022; Zhao et al., 2024). Low-Rank Adaptation (LoRA) has become a popular technique to solve the challenges when insufficient resources cannot support full fine-tuning (FFT) (Hu et al., 2022). However, the restricted update parameterization can limit adaptation, motivating efforts to improve its capacity and optimization (Wang et al., 2025). Combining LoRA with Mixture-of-Experts (MoE) framework offers a promising direction , which uses multiple low-rank experts provide alternative updates and an input-dependent router determines how to combine them (Zadouri et al., 2024; Liu & Luo, 2024; Tian et al., 2024; Fan et al., 2025).

However, more experts do not necessarily give the router more meaningful choices. Different expert parameters can produce similar responses on the data, allowing different routing weights to yield nearly identical updates. Prior work has identified insufficient expert specialization and expert overlap as challenges in LoRA-based MoE (Guo et al., 2025; Su & Liu, 2026). Routing determines how expert outputs are combined, but does not directly ensure that those outputs are distinguishable (Fan et al., 2025). Similarly, balanced expert utilization does not by itself establish that experts contribute different functions (Guo et al., 2025). Understanding a LoRA mixture therefore requires examining the expert responses that the router actually combines.

Moreover, the LoRA factorization reveals a second source of redundancy. In a LoRA update $\Delta \boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$, $\boldsymbol{A}$ projects inputs into a low-dimensional space and $\boldsymbol{B}$ maps these features back to the output space. Recent work (Tian et al., 2024) leverages this structure to share one projection $\boldsymbol{A}$ matrix across multiple expert-specific reconstruction matrices $\boldsymbol{B}_i$ for each expert i, effectively re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ducing parameter redundancy. But even distinct reconstruction matrices can produce similar outputs when the shared features contain little variation along the directions in which those matrices differ. Preserving these directions is also insufficient to ensure distinguishable mixtures, since different combinations of distinct expert outputs can still produce the same residual. These perspectives shift the focus from expert multiplicity to the geometry of expert contributions and motivates our central question: *How can we encourage distinguishable expert contributions while preserving informative shared features in a parameter-efficient LoRA mixture?*

We address this question through two theoretical results that establish when expert-output decorrelation yields distinguishable mixtures and quantitative routing-recovery bounds, and when a shared projection preserves differences between expert reconstruction matrices in their outputs. Our first result connects expert-output decorrelation to the distinguishability of routing weights. For fixed expert responses, we show that sufficiently decorrelated expert outputs with nonvanishing magnitudes make changes in routing weights observable in the combined residual. Specifically, we derive a lower bound on mixture-output separation in terms of routing-weight separation, which strengthens as correlations decrease while the output-norm bounds remain fixed. This bound establishes a unique oracle mixture for the quadratic residual objective and ensures that small excess residual loss implies small routing-weight error. This result motivates penalizing correlations among expert responses on activations. Our second result explains why output differentiation must be considered alongside the shared representation. We show that controlling the scale and determinant of the shared projection preserves reconstruction differences in mean-square output, provided the input distribution has sufficient variation in the retained directions. Together, these two results motivate reducing output redundancy while preserving informative shared directions. They provide conditional design principles rather than guarantees of training convergence or task-level specialization.

Building on these principles, we propose **OD-MoLE** (**O**rthogonality and **D**iversity for **M**ixture-**o**f-**L**oRA **E**xperts), a simple yet effective framework that promotes both specialization and collaboration. Built on the shared-projection architecture, OD-MoLE combines two complementary regularizers. An output-space orthogonality penalty reduces squared cosine similarities between expert responses on input activations. A shared-projection diversity penalty uses a sampled log determinant to encourage diversity within the shared low-rank representation. The first acts on the responses that the router combines, while the second acts on the features from which those responses are constructed. Our contribution is to connect these regularization targets to distinct properties of LoRA mixtures and evaluate their combined effect, rather than to introduce orthogonality or log-determinant regularization themselves. Experiments compare OD-MoLE with existing methods under both all-expert composition and selective routing, with ablations examining the contribution of each regularizer.

Contributions. This paper makes the following main contributions:

- We establish sufficient conditions under which expert-output decorrelation makes routing weights distinguishable and controls oracle routing-weight error through excess quadratic residual loss (Theorem 1). We also show how scale-controlled shared-projection diversity preserves existing reconstruction differences in mean-square output (Theorem 2).
- Based on this analysis, we propose OD-MoLE, a novel asymmetric LoRA mixture that couples output orthogonality with projection diversity to address output redundancy and shared-feature collapse, respectively.
- Extensive experiments demonstrate that OD-MoLE effectively achieves specialization and improves performance across diverse benchmarks while maintaining parameter efficiency.

2 Related Work

Parameter-Efficient Fine-tuning Parameter-efficient fine-tuning (PEFT) has emerged to improve performance of LLMs under resource-constrained settings. LoRA (Hu et al., 2022) has been a successful method and has various variants, which fundamentally introduce new low-rank trainable parameters within the existing model layer while keeping original parameters frozen. AdaLoRA (Zhang et al., 2023) dynamically allocates rank across layers based on weight importance score, DoRA (Liu et al., 2024) decomposes LoRA updates into magnitude and direction to improve stability during optimization, LoRA+ (Hayou et al., 2024) improves LoRA by setting different learning rates for LoRA adapter matrices, PiSSA (Meng et al., 2024) utilizes SVD to initialize

LoRA weights with principal components of the pre-trained weight for preserving knowledge, and LoRA-Pro (Wang et al., 2025) aligns LoRA's updates with full fine-tuning gradients. Despite these advances, a noticeable performance gap remains between LoRA-based PEFT and full fine-tuning.

Integration of MoE and LoRA The integration of Mixture-of-Experts (MoE) and LoRA is a notable trend in recent advancements aimed at enhancing the performance of LLMs (Wu et al., 2024). MoLoRA (Wu et al., 2024) proposes to integrate the MoE layer with LoRA modules, demonstrating that LoRA experts enhance specialization. AdaMoLE (Liu & Luo, 2024) extends this idea by introducing learnable thresholds for dynamic expert selection. HydraLoRA (Tian et al., 2024) adopts an asymmetric LoRA MoE framework where shared matrix $\boldsymbol{A}$ is combined with multiple expert-specific matrices $\boldsymbol{B}$, effectively reducing the number of expert parameters. GOAT (Fan et al., 2025) introduces a SVD-based MoE framework that adaptively integrates relevant priors and addresses weight misalignment and gradient dynamics through a theoretical scaling strategy. Guo et al. (2025) introduce complementary orthogonality and variance losses to encourage distinct expert token assignments and more discriminative routing decisions. FlyLoRA (Zou et al., 2025) is an implicit LoRA MoE that enables rank-wise experts in matrix $\boldsymbol{B}$ and a fixed sparse random projection for matrix $\boldsymbol{A}$ as an implicit router to reduce parameter interference.

3 Preliminaries

In this section, we introduce the notations and briefly review the Mixture-of-Experts (MoE) formulation and its low-rank adaptation parameterization.

Problem Definition. In standard MoE architecture, the output $f(\boldsymbol{x}) \in \mathbb{R}^d$ for an input $\boldsymbol{x} \in \mathbb{R}^k$ is formulated as the weighted combination of N specialized networks $\{f_i\}_{i=1}^N$:

$$f(\boldsymbol{x}) = \sum\nolimits_{i=1}^{N} g_i(\boldsymbol{x})\, f_i(\boldsymbol{x}), \tag{1}$$

where $f_i(\boldsymbol{x})$ is expert i output, and $g_i(\boldsymbol{x})$ is a nonnegative gating coefficient for expert i. We assume that the router produces a probability vector $\boldsymbol{g}(\boldsymbol{x}) \in \Delta_N = \{\boldsymbol{g} \in \mathbb{R}^N : g_i \geq 0, \mathbf{1}^\top \boldsymbol{g} = 1\}$. A common parameterization uses learnable scoring functions $\{\phi_i\}_{i=1}^N$, often parameterized as:

$$g_i(\boldsymbol{x}) = \exp(\phi_i(\boldsymbol{x})) / \sum\nolimits_{j=1}^{N} \exp(\phi_j(\boldsymbol{x})) \tag{2}$$

LoRA MoE. In LoRA MoE framework, we leverage the efficiency of LoRA to parameterize individual experts within the mixture. Given a frozen pretrained model weight $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$, each expert i is defined by a modular additive update. Decompose update $\Delta\boldsymbol{W}_i$ into two low-rank matrices $\boldsymbol{B}_i \in \mathbb{R}^{d\times r}$, $\boldsymbol{A}_i \in \mathbb{R}^{r\times k}$, where $r \ll \min(d, k)$. The effective weight for expert i becomes

$$\boldsymbol{W}_i = \boldsymbol{W}_0 + \Delta\boldsymbol{W}_i = \boldsymbol{W}_0 + \frac{\gamma}{r}\boldsymbol{B}_i\boldsymbol{A}_i, \tag{3}$$

where scaling factor γ scales update and we define $c = \gamma/r$ as the residual scaling factor. By applying this to input $\boldsymbol{x}$, the expert-specific transformation $f_i(\boldsymbol{x})$ becomes $f_i(\boldsymbol{x}) = \boldsymbol{W}_i\boldsymbol{x} = \boldsymbol{W}_0\boldsymbol{x} + \gamma/r\boldsymbol{B}_i\boldsymbol{A}_i\boldsymbol{x}$. Here, $\boldsymbol{W}_0$ remains frozen during training, while only $\{\boldsymbol{A}_i, \boldsymbol{B}_i\}$ are updated.

Substituting this into the global MoE aggregation formula, and noting that the frozen term $\boldsymbol{W}_0\boldsymbol{x}$ is independent of expert index i, we can simplify by factoring out the shared linear transformation:

$$f_{\mathrm{MoE}}(\boldsymbol{x}) = \boldsymbol{W}_0\boldsymbol{x} + c\sum\nolimits_{i=1}^{N} g_i(\boldsymbol{x})\,(\boldsymbol{B}_i\boldsymbol{A}_i\boldsymbol{x})\,. \tag{4}$$

This parameterization effectively reduces per-expert training overhead to $\{\boldsymbol{A}_i, \boldsymbol{B}_i\}$, significantly enhancing parameter efficiency of MoE layer while maintaining the expressive capacity afforded by gating-dependent low-rank residuals. However, the integration of LoRA into mixture paradigms introduces distinct challenges. Specifically, we need to evaluate whether this architecture successfully partitions latent space among experts or if it inadvertently promotes redundant, correlated behaviors.

4 Theoretical Analysis and Design Principles

We now analyze the shared-projection LoRA MoE through (1) when different routing weights produce distinguishable mixtures, and (2) when the shared projection preserve differences between experts. The insights from this analysis motivate the design of our proposed method in Section 5.

For a model input $\boldsymbol{x}$, let $\boldsymbol{h} = \boldsymbol{h}_m(\boldsymbol{x})$ denote the input activation to module m. Suppressing the module index, we define the unscaled expert residual output of expert i as $\boldsymbol{z}_i(\boldsymbol{h}) = \boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$, and collects all expert residuals as columns of $\boldsymbol{Z}(\boldsymbol{h}) = [\boldsymbol{z}_1(\boldsymbol{h}), \ldots, \boldsymbol{z}_N(\boldsymbol{h})]$. The adapted layer output is $\boldsymbol{W}_0\boldsymbol{h} + c\boldsymbol{Z}(\boldsymbol{h})\boldsymbol{g}(\boldsymbol{h})$. For fixed $\boldsymbol{h}$ and fixed expert parameters, consider

$$Q_{\boldsymbol{h}}(\boldsymbol{g}) = \frac{1}{2}\|c\boldsymbol{Z}(\boldsymbol{h})\boldsymbol{g} - \boldsymbol{t}_{\boldsymbol{h}}\|_2^2, \quad \boldsymbol{g} \in \Delta_N \tag{5}$$

where $\boldsymbol{t}_{\boldsymbol{h}} \in \mathbb{R}^d$ is a fixed target residual. For squared-error prediction with target $\boldsymbol{y}$, $\boldsymbol{t}_{\boldsymbol{h}} = \boldsymbol{y} - \boldsymbol{W}_0\boldsymbol{h}$ makes $Q_{\boldsymbol{h}}$ the prediction loss. For hidden-layer adaptation, it is an analytical surrogate.

Expert-output decorrelation and router recovery. We first examine whether different routing vectors produce distinguishable mixtures at a fixed input activation $\boldsymbol{h}$. The Hessian of $Q_{\boldsymbol{h}}$ (Eq. 5) with respect to routing weights is $c^2\boldsymbol{Z}(\boldsymbol{h})^\top\boldsymbol{Z}(\boldsymbol{h})$, Thus, expert-output alignment determines the curvature of the routing objective. To quantify directional alignment, assume $N \geq 2$ and $\boldsymbol{z}_i(\boldsymbol{h}) \neq 0$ for every expert, and define average squared cosine similarity

$$\omega(\boldsymbol{h}) = \frac{1}{N(N-1)}\sum_{i\neq j}\left(\frac{\boldsymbol{z}_i(\boldsymbol{h})^\top\boldsymbol{z}_j(\boldsymbol{h})}{\|\boldsymbol{z}_i(\boldsymbol{h})\|_2\|\boldsymbol{z}_j(\boldsymbol{h})\|_2}\right)^2, \quad \delta(\boldsymbol{h}) = \sqrt{N(N-1)\omega(\boldsymbol{h})} \tag{6}$$

Here quantity $\omega(\boldsymbol{h})$ measures average squared similarity between expert residuals, treating positive and negative alignment symmetrically. The quantity $\delta(\boldsymbol{h})$ measures how far the normalized output Gram matrix is from the identity in Frobenius norm, giving the following sufficient condition.

Theorem 1 (Output decorrelation and routing recovery)**.** *Fix input activation $\boldsymbol{h}$, the expert parameters, the target residual $\boldsymbol{t}_{\boldsymbol{h}}$, and the scaling factor $c > 0$. Suppose that $0 < \underline{z} \leq \|\boldsymbol{z}_i(\boldsymbol{h})\|_2 \leq \overline{z}$ for every expert i and that $\delta(\boldsymbol{h}) < 1$. Then, for any $\boldsymbol{g}, \widehat{\boldsymbol{g}} \in \Delta_N$,*

$$\|c\boldsymbol{Z}(\boldsymbol{h})(\boldsymbol{g} - \widehat{\boldsymbol{g}})\|_2^2 \geq c^2\underline{z}^2(1 - \delta(\boldsymbol{h}))\|\boldsymbol{g} - \widehat{\boldsymbol{g}}\|_2^2. \tag{7}$$

Moreover, $Q_{\boldsymbol{h}}$ has a unique minimizer $\boldsymbol{g}^*(\boldsymbol{h})$ on Δ_N, and each $\boldsymbol{g} \in \Delta_N$ satisfies

$$\|\boldsymbol{g} - \boldsymbol{g}^*(\boldsymbol{h})\|_2^2 \leq \frac{2[Q_{\boldsymbol{h}}(\boldsymbol{g}) - Q_{\boldsymbol{h}}(\boldsymbol{g}^*(\boldsymbol{h}))]}{c^2\underline{z}^2(1 - \delta(\boldsymbol{h}))}. \tag{8}$$

The norm lower bound controls output scale, even orthogonal outputs provide weak separation if their magnitudes approach zero. The condition $\delta(\boldsymbol{h}) < 1$ requires $\omega(\boldsymbol{h}) < 1/[N(N-1)]$ and is sufficient rather than necessary for distinguishable mixtures, as opposite expert outputs can still uniquely determine routing weights (see Appendix, Example 1). Under the theorem's assumptions, small excess residual loss implies proximity to the best mixture weights for $Q_{\boldsymbol{h}}$, motivating penalizing expert-output alignment directly on input activations, since even orthogonal residuals can yield identical predictive distributions under different routing weights (Appendix, Example 2). The oracle need not represent a semantic task assignment, and the result does not guarantee that joint training achieves a small loss gap.

Shared-projection diversity and expert distinguishability. We next examine whether differences between expert matrices remain visible after applying the shared projection. Even substantially different matrices $\boldsymbol{B}_i$ and $\boldsymbol{B}_j$ can produce identical outputs if their differences act only on latent directions that the input distribution does not activate. Assume $\mathbb{E}\|\boldsymbol{h}\|_2^2 < \infty$, and define $\boldsymbol{\Sigma}_{\boldsymbol{h}} = \mathbb{E}\left[\boldsymbol{h}\boldsymbol{h}^\top\right]$ and $\boldsymbol{K}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{h}}\boldsymbol{A}^\top$. The smallest eigenvalue of $\boldsymbol{K}_{\boldsymbol{A}}$ measures the least expected squared activation along a unit latent direction. In particular,

$$\mathbb{E}_h\|\boldsymbol{z}_i(\boldsymbol{h}) - \boldsymbol{z}_j(\boldsymbol{h})\|_2^2 \geq \lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})\|\boldsymbol{B}_i - \boldsymbol{B}_j\|_{\mathrm{F}}^2. \tag{9}$$

The following result bounds this eigenvalue using the shared projection and the input distribution.

Theorem 2 (Shared diversity preserves expert differences)**.** *Suppose $r > 1$, $tr(\boldsymbol{A}\boldsymbol{A}^\top) = r$, and $\log\det(\boldsymbol{A}\boldsymbol{A}^\top) \geq -\eta$, where $\eta \geq 0$. Assume $\boldsymbol{u}^\top\boldsymbol{\Sigma}_{\boldsymbol{h}}\boldsymbol{u} \geq \nu\|\boldsymbol{u}\|_2^2$ for every $\boldsymbol{u} \in row(\boldsymbol{A})$, with $\nu > 0$. Then*

$$\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}}) \geq \nu e^{-\eta}\left(\frac{r-1}{r}\right)^{r-1}. \tag{10}$$

The trace fixes the total squared singular-value scale, and the determinant prevents individual directions from collapsing at that scale. The input assumption ensures nonzero second-moment energy in

the directions retained by $\boldsymbol{A}$. Together with Eq. 9, Theorem 2 shows that these controls preserve existing expert differences in mean square. But they do not create differences between identical experts or imply the pointwise routing condition. For example, one expert can equal the average of two others while all pairs retain positive mean-square output differences, allowing distinct routing weights to produce identical mixtures (Appendix, Example 3). Thus, the two results motivate complementary regularizers as output decorrelation discourages redundant mixtures and shared-projection diversity helps preserve the directions through which experts differ. Proofs are provided in Appendix A, and the connection to the sampled log-determinant penalty are detailed in Appendix A.6.

5 OD-MoLE: Orthogonality and Diversity for LoRA MoE

Building on Section 4, we introduce OD-MoLE, which combines two regularizers in a shared-projection LoRA mixture. Output-space orthogonality discourages redundant expert responses, while shared-projection diversity encourages the projection to retain distinct latent directions. These objectives are motivated by the routing-recovery and expert-difference preservation results, respectively. We first describe the architecture and then define the two regularizers.

Asymmetric LoRA MoE Formulation. We adopt an asymmetric LoRA MoE architecture motivated by the distinct functional roles of LoRA components to navigate the trade-off between expert specialization and collaborative diversity. In LoRA, $\boldsymbol{A}$ projects inputs into a low-rank feature space and $\boldsymbol{B}$ reconstructs this representation into the output space. Prior work has shown their asymmetric roles that optimizing $\boldsymbol{B}$ is particularly important for capturing task-specific variations (Zhu et al., 2024). This observation motivates an MoE with a shared $\boldsymbol{A}$ and multiple expert-specific $\boldsymbol{B}_i$ matrices. For an input activation $\boldsymbol{h} \in \mathbb{R}^k$, the effective weight update is

$$\Delta \boldsymbol{W} = c \sum_{i=1}^{N} g_i(\boldsymbol{h}) \boldsymbol{B}_i \boldsymbol{A}, \tag{11}$$

where $\boldsymbol{A} \in \mathbb{R}^{r \times k}$, $\boldsymbol{B}_i \in \mathbb{R}^{d \times r}$, and $c = \gamma / r$. Sharing $\boldsymbol{A}$ reduces parameter redundancy, while the expert-specific matrices $\boldsymbol{B}_i$ provide different mappings from shared features to output residuals. This structure also separates the two regularization targets: output orthogonality acts on $\boldsymbol{B}_i \boldsymbol{A} \boldsymbol{h}$, while projection diversity acts on $\boldsymbol{A}$. We next describe how these regularizers address output redundancy and shared-feature collapse.

5.1 Specialization via Output Orthogonality

Theorem 1 connects expert-output correlations to the distinguishability of routing weights. For fixed expert outputs, the Hessian of the quadratic residual objective is $c^2 \boldsymbol{Z}(\boldsymbol{h})^\top \boldsymbol{Z}(\boldsymbol{h})$. Under the theorem's output-norm assumptions, sufficiently small correlations provide a positive curvature bound and control routing-weight error through excess residual loss. This motivates regularizing expert responses on input activations. Specifically, recalling $\boldsymbol{z}_i(\boldsymbol{h}) = \boldsymbol{B}_i \boldsymbol{A} \boldsymbol{h}$ for expert i, we penalize the average squared cosine similarity among experts as follows:

$$\mathcal{L}_{\text{orth}} = \mathbb{E}_h \left[\frac{1}{N(N-1)} \sum_{i \neq j} \left(\frac{\boldsymbol{z}_i(\boldsymbol{h})^\top \boldsymbol{z}_j(\boldsymbol{h})}{\|\boldsymbol{z}_i(\boldsymbol{h})\|_2 \|\boldsymbol{z}_j(\boldsymbol{h})\|_2} \right)^2 \right]. \tag{12}$$

This expression assumes nonzero expert outputs. The normalization separates directional alignment from output magnitude, and squaring penalizes both positive and negative alignment. Because the penalty is evaluated on activations, it targets similarities between the responses actually combined by the router. Minimizing this penalty encourages output decorrelation. It does not enforce the output-norm lower bound or guarantee that every input satisfies the sufficient correlation condition in Theorem 1. The theorem therefore motivates the regularization target without guaranteeing task-specific specialization during training.

5.2 Collaboration via Shared-Projection Diversity

Output decorrelation addresses similarities between expert responses, but those responses also depend on the shared features $\boldsymbol{A}\boldsymbol{h}$. Differences between reconstruction matrices can remain unexpressed if the shared features contain little energy in the corresponding latent directions. Theorem 2

shows that controlling the projection's scale and determinant preserves these differences in mean-square output, provided the input distribution has sufficient variation in the retained directions.

Motivated by this result, we encourage projection diversity through a sampled log-determinant penalty. In LoRA module m, for $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ with $r \leq k$, sample a set $\mathcal{S}$ of r columns uniformly. Let $\boldsymbol{A}_{\mathcal{S}} \in \mathbb{R}^{r\times r}$ denote the submatrix formed by these columns. We define

$$\mathcal{L}_{\text{div}} = -\log\det\left(\boldsymbol{A}_{\mathcal{S}}^{\top}\boldsymbol{A}_{\mathcal{S}} + \epsilon\boldsymbol{I}\right), \qquad \epsilon > 0 \tag{13}$$

where $\boldsymbol{I}$ is $r \times r$ identity matrix, serving as stabilization term keeping the logarithm well-defined even when the sampled columns are linearly dependent. The determinant measures the squared volume spanned by the sampled columns before stabilization. At a controlled scale, larger volume discourages nearly dependent directions. The scale assumption is essential to this interpretation. Rescaling $\boldsymbol{A} \mapsto s\boldsymbol{A}$ and $\boldsymbol{B}_i \mapsto \boldsymbol{B}_i/s$ for any scalar $s > 1$ preserves expert outputs but can decrease the diversity penalty. Consequently, the penalty alone does not enforce the assumptions of Theorem 2 or guarantee good conditioning. The theorem provides conditional motivation for preserving shared directions through which existing expert differences can affect the output.

We note that this objective is closely related to determinantal point processes (DPPs) (Derezinski & Mahoney, 2021), which favor diverse and non-redundant subsets by maximizing the determinant of a kernel matrix. In our setting, the determinant of $\boldsymbol{A}_{\mathcal{S}}^{\top}\boldsymbol{A}_{\mathcal{S}}$ equals the squared volume spanned by the sampled columns. We use this quantity as a regularization objective to encourage diversity in the shared projection. We also note that our approach is aligned with prior research (Huang et al., 2025; Albert et al., 2025; Shen, 2025; Xu et al., 2026).

5.3 Overall Training Objective

We optimize OD-MoLE by augmenting the task loss with two complementary regularization terms that promote expert specialization and collaborative expressiveness. Let $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\}$ denote the training dataset. We use m to index all LoRA MoE modules $\mathcal{M}$ across layers and submodules (e.g., query, key, value, and output projections), and let $\boldsymbol{h}_m(\boldsymbol{x})$ denote the input hidden representation to module m induced by $\boldsymbol{x}$. We define the output-space orthogonality loss:

$$\mathcal{L}_{\text{orth}}(\boldsymbol{x}) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{1}{N(N-1)}\sum_{i\neq j}\left(\frac{\langle \boldsymbol{B}_{m,i}\boldsymbol{A}_m\boldsymbol{h}_m(\boldsymbol{x}), \boldsymbol{B}_{m,j}\boldsymbol{A}_m\boldsymbol{h}_m(\boldsymbol{x})\rangle}{\|\boldsymbol{B}_{m,i}\boldsymbol{A}_m\boldsymbol{h}_m(\boldsymbol{x})\|_2\,\|\boldsymbol{B}_{m,j}\boldsymbol{A}_m\boldsymbol{h}_m(\boldsymbol{x})\|_2}\right)^2, \tag{14}$$

which penalizes squared cosine similarity between expert outputs. OD-MoLE training objective is:

$$\begin{aligned}\mathcal{L}_{\text{OD-MoLE}} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\Big[\ell(f_{\boldsymbol{W}_0+\Delta\boldsymbol{W}_{\text{MoE}}}(\boldsymbol{x}), \boldsymbol{y}) + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}(\boldsymbol{x})\Big] \\ - \frac{\lambda_{\text{div}}}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\mathbb{E}_{\mathcal{S}_m}\left[\log\det\left(\boldsymbol{A}_{m,\mathcal{S}_m}^{\top}\boldsymbol{A}_{m,\mathcal{S}_m} + \epsilon\boldsymbol{I}\right)\right].\end{aligned} \tag{15}$$

Here, $\ell(\cdot)$ denotes the task loss, while the two regularizers encourage less redundant expert outputs and more diverse shared projection directions. The coefficients λ_{orth} and λ_{div} control their contributions to training. Their effects on task performance are evaluated empirically in Tab. 3.

Efficient approximation. Direct computation of the regularization terms over all tokens and modules can be expensive in Eq. 15. In practice, we adopt stochastic approximations: the orthogonality loss is estimated using a randomly sampled subset of token-level activations, while the diversity regularizer is computed on a randomly selected LoRA module m^* and a subset $\mathcal{S}$ of its columns then compute $\mathcal{L}_{\text{div}}$ via Eq. 13 at each training step. These sampling procedures reduce regularization overhead. The completed algorithm is in Algorithm 1.

6 Experiments

We first describe the experimental setup, followed by the main results, then present ablation studies and additional analyses to provide deeper insights into our method. Additional experimental results are shown in Appendix B.

Baselines. We evaluate OD-MoLE against both single-LoRA fine-tuning and existing LoRA MoE approaches. For single-LoRA baselines, we consider LoRA (Hu et al., 2022) with ranks

Algorithm 1 OD-MoLE

Require: Dataset $\mathcal{D}$, frozen $\boldsymbol{W}_0$, LoRA modules $\mathcal{M}$, N experts, top-K routing, $\lambda_{\text{orth}}, \lambda_{\text{div}}$

1: Initialize $\{\boldsymbol{A}_m, \boldsymbol{B}_{m,i}\}$, routers ϕ_m for each module $m \in \mathcal{M}$
2: **for** each minibatch $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}$ **do**
3: For each module m, $\boldsymbol{g}_m \leftarrow \text{TopKSoftmax}(\phi_m(\boldsymbol{h}_m), K)$
4: Compute $\boldsymbol{z}_{m,i} = \boldsymbol{B}_{m,i}\boldsymbol{A}_m\boldsymbol{h}_m$ for each expert i, and routed residual $c\sum_i g_{m,i}\boldsymbol{z}_{m,i}$
5: Randomly select a module m^* and columns $\mathcal{S}_{m^*}$ of $\boldsymbol{A}_{m^*}$ and compute $\mathcal{L}_{\text{div}}$ via Eq. 13
6: Compute output-space orthogonality loss $\mathcal{L}_{\text{orth}}$ via Eq. 14
7: Compute $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}$
8: Update $\{\boldsymbol{A}_m, \boldsymbol{B}_{m,i}, \phi_m\}$ by backpropagation while keeping $\boldsymbol{W}_0$ frozen
9: **end for**

Table 1: Comparison of PEFT methods on commonsense reasoning tasks using Llama-2-7B. LoRA MoE methods are grouped into (i) all-expert composition and (ii) selective routing, including both adaptive routing (AdaMoLE) and top-k selection. Task scores are mean accuracies (%) across three seeds $\{0, 1, 2\}$. Avg. reports the mean $\pm$ sample standard deviation of the per-seed average accuracy across eight tasks. **Bold** and underlined indicate the best and second-best results within each group.

Method	ComQA	CosQA	SIQA	PIQA	ScQA	BoolQ	ARC-C	OBQA	Avg
Single-LoRA baselines									
LoRA (r:8)	73.31	81.71	78.33	82.97	88.15	76.80	58.45	83.60	77.92 ± 1.96
LoRA (r:16)	73.23	81.68	77.97	83.11	88.85	77.40	59.25	84.80	78.29 ± 1.97
LoRA (r:32)	73.06	82.78	78.04	83.23	88.46	75.20	58.14	85.00	77.99 ± 1.91
PiSSA	72.60	81.15	77.45	81.85	87.18	74.80	57.97	81.40	76.80 ± 1.32
Compose all experts									
HydraLoRA	75.76	82.78	78.25	82.92	88.62	80.40	58.70	86.60	79.25 ± 1.31
MoLoRA	74.86	**84.19**	**78.66**	83.35	**89.32**	78.60	59.98	87.20	79.52 ± 1.35
GOAT	73.96	82.74	77.83	**84.01**	87.37	**82.20**	59.53	86.20	79.23 ± 1.56
OD-MoLE	**76.41**	83.45	78.20	83.03	89.18	81.21	**61.26**	**87.60**	**80.04** ± 1.19
Route to selective experts									
HydraLoRA	78.21	84.05	**78.86**	82.56	90.11	82.20	61.69	**89.40**	80.89 ± 1.43
MoLoRA	77.31	84.02	77.99	81.18	90.16	81.60	59.64	87.20	79.89 ± 1.38
AdaMoLE	74.46	83.08	78.13	**84.12**	89.03	80.80	60.87	86.40	79.61 ± 1.22
GOAT	76.63	83.75	78.36	81.47	**90.32**	82.80	61.88	88.40	80.45 ± 1.67
OD-MoLE	**78.79**	**84.25**	78.51	83.03	89.74	**83.60**	**62.37**	88.80	**81.14** ± 1.07

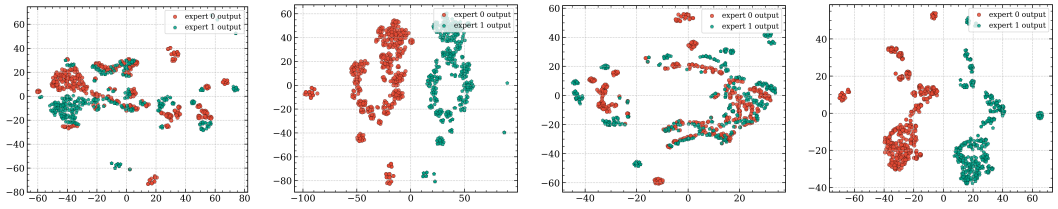
$r \in \{8, 16, 32\}$ and PiSSA (Meng et al., 2024). For LoRA MoE methods, we compare with HydraLoRA (Tian et al., 2024), GOAT (Fan et al., 2025), AdaMoLE (Liu & Luo, 2024), and MoLoRA (Zadouri et al., 2024).

Model and Datasets. We conduct experiments on two pre-trained models, Llama-2-7B and Qwen2.5-7B (Yang et al., 2024). We consider a diverse set of commonsense reasoning benchmarks, including CommonsenseQA (ComQA) (Talmor et al., 2019), CosmosQA (CosQA) (Huang et al., 2019), SocialIQA (SIQA) (Sap et al., 2019), PhysicalIQA (PIQA) (Bisk et al., 2020), ScienceQA (ScQA) (Lu et al., 2022), BoolQ (Clark et al., 2019), ARC-Challenge (ARC-C) (Clark et al., 2018), and OpenBookQA (OBQA) (Mihaylov et al., 2018).

Experimental Setup. For all LoRA MoE methods, we assign LoRA experts to four attention projection submodules (W_q, W_k, W_v, W_o) in every Transformer layer and fix the total effective expert rank at 32, excluding gating-related parameters. We consider two configurations: (i) *expert composition*, with 2 experts per submodule, all active; and (ii) *selective routing*, with 8 experts per submodule and top-2 gating. Single-LoRA baselines use ranks $r \in \{8, 16, 32\}$. Following (Liu & Luo, 2024), we truncate input sequences to a maximum length of 256 tokens. We use a total batch size of 16 and a learning rate of 10^{-4}, with $\lambda_{\text{orth}} = \lambda_{\text{div}} = 10^{-3}$. The number of training epochs for each dataset is selected based on validation performance and is reported in Appendix Tab. 5. To promote balanced expert utilization, we include a load-balancing loss (Fedus et al., 2022) with coefficient

Table 2: Comparison of OD-MoLE and HydraLoRA on commonsense reasoning tasks using Qwen2.5-7B in all-expert composition and selective top-k routing. Task scores are mean accuracies (%) across seeds $\{0, 1, 2\}$. Avg. reports mean $\pm$ sample standard deviation of per-seed average accuracy across eight tasks. **Bold** indicates the highest mean average within each routing group.

Method	ComQA	CosQA	SIQA	PIQA	ScQA	BoolQ	ARC-C	OBQA	Avg.
Compose all experts									
HydraLoRA	85.42	84.35	80.14	87.65	92.49	83.20	83.36	92.80	86.18±1.25
OD-MoLE	85.50	84.85	79.63	87.65	92.58	83.80	83.96	94.80	**86.60**±0.92
Route to selective experts									
HydraLoRA	85.01	84.57	81.24	87.25	93.32	84.10	84.73	93.20	86.68±1.37
OD-MoLE	85.09	84.93	81.35	87.84	93.11	84.50	84.73	94.80	**87.04**±0.95



HydraLoRA (q-proj) OD-MoLE (q-proj) HydraLoRA (k-proj) OD-MoLE (k-proj)

Figure 1: t-SNE visualization of expert output representations $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$ for q- and k-projection modules at layer 0. From left to right: HydraLoRA (q), OD-MoLE (q), HydraLoRA (k), and OD-MoLE (k). OD-MoLE exhibits clearer separation between expert outputs, improving specialization.

10^{-3}. All experiments use Hugging Face Transformers (Wolf et al., 2019) and PEFT on a single NVIDIA A100 GPU.

6.1 Overall Performance

Tab. 1 reports the performance of OD-MoLE and baseline methods on commonsense reasoning benchmarks using Llama-2-7B. We group LoRA MoE methods into (i) expert composition and (ii) selective routing to enable a systematic comparison under different routing strategies. Under all expert compositions, OD-MoLE outperforms baselines on average performance, attaining the best results on ComQA, ARC-C, and OBQA, and remains competitive on other tasks. Under selective routing, OD-MoLE further improves the average performance, outperforming HydraLoRA, GOAT, and AdaMoLE. It achieves the best results on ComQA, CosQA, BoolQ, and ARC-C, and ranks among top performers on most other tasks. Tab. 2 compares OD-MoLE with HydraLoRA on Qwen2.5-7B. OD-MoLE outperforms HydraLoRA under both expert composition and top-k routing, achieving higher average accuracy in both settings. The improvements are consistent across most tasks, with notable gains on CosQA, BoolQ, and OBQA. These results demonstrate that enforcing output-level orthogonality and diversity leads to more effective and robust expert representations.

6.2 Ablation Study

Expert Output Representation Analysis. To better understand how different methods affect expert specialization, we visualize expert output representations $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$ using t-SNE (Van der Maaten & Hinton, 2008). Fig. 1 compares HydraLoRA and OD-MoLE on the q- and k-projection modules at layer 0 under expert composition, using Llama-2-7B on ARC-C. We observe that HydraLoRA produces heavily overlapping clusters between expert outputs, indicating that different experts tend to learn similar transformations and exhibit limited specialization. In contrast, OD-MoLE yields clearly separated clusters for different experts across both q- and k-projection modules. It suggests that OD-MoLE encourages experts to capture more diverse and distinct output representations.

Impact of Each Regularizer. Tab. 3 evaluates the contributions of output orthogonality and shared-projection diversity. The results identify output orthogonality as the larger individual contributor.

Table 3: Regularizer ablation on Qwen2.5-7B in all-expert composition. Orth. and Div. denote output orthogonality and shared-projection diversity. Task scores are mean accuracies (%) across seeds $\{0, 1, 2\}$. Avg. is mean $\pm$ standard deviation of per-seed average accuracy across eight tasks.

Orth.	Div.	ComQA	CosQA	SIQA	PIQA	ScQA	BoolQ	ARC-C	OBQA	Avg.
×	×	85.42	84.35	**80.14**	87.65	92.49	83.20	83.36	92.80	86.18 ± 1.25
×	✓	**85.76**	84.45	79.93	87.17	91.27	82.97	83.81	94.40	86.22 ± 1.23
✓	×	84.85	84.43	79.50	**87.97**	92.38	82.83	**84.83**	94.60	86.42 ± 1.14
✓	✓	85.50	**84.85**	79.63	87.65	**92.58**	**83.80**	83.96	**94.80**	**86.60** ± 0.92

The diversity regularizer is comparatively weak in isolation but provides a complementary improvement when combined with output orthogonality.

Effect of the Number of Experts. We vary $N \in \{8, 16, 32\}$ while fixing the total expert rank $Nr_i = 32$ and routing ratio $K/N = 1/4$, keeping the active expert rank at $Kr_i = 8$, where r_i is rank of each expert i. Using Qwen2.5-7B, OD-MoLE achieves average accuracies of 87.04%, 86.23%, and 86.53%, exceeding HydraLoRA by 0.36, 0.18, and 0.54 percentage points, respectively. Increasing expert count while reducing per-expert rank therefore does not consistently improve performance. Full results are reported in Appendix Table 7.

Table 4: Comparison of OD-MoLE and HydraLoRA in peak memory usage, training time per epoch, and performance on CommonsenseQA using Llama-2-7B and Qwen2.5-7B. Both methods are evaluated under identical settings on a single NVIDIA A100 GPU.

	Llama-2-7B			Qwen2.5-7B		
Method	Peak Mem	Time/Epoch	Accuracy(%)	Peak Mem	Time/Epoch	Accuracy(%)
Compose all experts						
HydraLoRA	37.30GB	34.0min	75.76	33.07GB	37.0min	85.42
OD-MoLE	37.18GB	40.0min	**76.41**	34.77GB	42.5min	**85.50**
Route to selective experts						
HydraLoRA	39.10GB	48.5min	78.21	35.35GB	52.5min	85.01
OD-MoLE	39.33GB	56.5min	**78.79**	36.85GB	59.0min	**85.09**

Computation Analysis. Tab. 4 compares the computational cost and performance of different LoRA MoE methods under identical experimental settings. We evaluate both peak memory usage and per-epoch training time on a single NVIDIA A100 GPU. OD-MoLE achieves consistently better performance than HydraLoRA across both routing strategies, with only a modest increase in computational cost. In *compose-all-experts* setting, OD-MoLE improves accuracy while incurring a negligible increase in memory and a small increase in training time. In *selective routing* setting, OD-MoLE again outperforms HydraLoRA, but with a more noticeable increase in training time. We attribute this overhead to the additional computations required for enforcing output diversity and orthogonality among dynamically selected experts. Despite this increase, the memory usage remains comparable, indicating that our method does not introduce significant additional parameters. These results highlight a favorable trade-off: OD-MoLE consistently improves model performance while maintaining similar memory consumption and only moderately increasing training time.

7 Conclusion

In this work, we have presented OD-MoLE, which combines expert-output orthogonality and shared-projection diversity in an asymmetric LoRA mixture. Our analysis establishes sufficient conditions under which output decorrelation makes routing weights distinguishable and shared-projection diversity preserves existing expert differences in mean-square output. These results motivate the two regularization targets while making their assumptions explicit. Experiments on Llama-2-7B and Qwen2.5-7B show OD-MoLE performs well under both all-expert composition and selective routing. Complementarily, the results support the usefulness of the combined regularization, establishing task-specific expert functions requires further evaluation beyond output separation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the manuscript and generating latex code for the experimental tables.

We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, generating research ideas, and LLM-generated synthetic data sets are not applicable to this work.

Additionally, we used generative AI tools for assisting in the writing of proofs, for implementing methods, and for organizing the ReadMe instructions of the code.

We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and does not involve human-subject experiments, participant recruitment, private user data, personally identifiable information, or newly collected human annotations. The datasets and pretrained models used in the experiments are used in accordance with their publicly available licenses and terms. The method does not provide a formal privacy guarantee against information leakage from shared model updates. Such privacy protection is outside the scope of this work and may be incorporated through complementary privacy-preserving mechanisms when required.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs for the main theoretical results in the appendix. The paper and appendix describe the dataset construction, expert-mixture generation procedure, model and training configurations, baselines, hyperparameters, routing and extraction procedures, ablations, evaluation protocol, and computational setup. We provide the anonymous link for code.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Paul Albert, Frederic Z Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Randlora: Full-rank parameter-efficient fine-tuning of large models. *arXiv preprint arXiv:2502.00987*, 2025.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Michał Derezinski and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021.

Chenghao Fan, Zhenyi Lu, Sichen Liu, Chengfeng Gu, Xiaoye Qu, Wei Wei, and Yu Cheng. Make loRA great again: Boosting loRA with adaptive singular values and mixture-of-experts optimization alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=SUxq4HeIAd.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better moe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iydmH9boLb.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=NEv8YqBROO.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243/.

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

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

Zefang Liu and Jiahua Luo. AdamoLE: Fine-tuning large language models with adaptive mixture of low-rank adaptation experts. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ndY9qFf9Sa.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=HjwK-Tc_Bc.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6ZBHIEtdP4.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Yixin Shen. Kron-lora: Hybrid kronecker-lora adapters for scalable, sustainable fine-tuning. *arXiv preprint arXiv:2508.01961*, 2025.

Ye Su and Yong Liu. Variational inference, entropy, and orthogonality: A unified theory of mixture-of-experts. *arXiv preprint arXiv:2601.03577*, 2026.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng zhong Xu. HydraloRA: An asymmetric loRA architecture for efficient fine-tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qEpi8uWX3N.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-pro: Are low-rank adapters properly optimized? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWvKBCYh4S.

Yixing Xu, Guanchen Li, Chao Li, Xuanwu Yin, Dong Li, Spandan Tiwari, Ashish Sirasao, and Emad Barsoum. Low kruskal-rank adaptation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UcnpfSsXtu.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EvDeiLv7qc.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hYHsrKDiX7.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=txRZBD8tBV.

Heming Zou, Yunliang Zang, Wutong Xu, Yao Zhu, and Xiangyang Ji. FlyloRA: Boosting task decoupling and parameter efficiency via implicit rank-wise mixture-of-experts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=nGQLYn13Xf.

APPENDIX

The appendix is organized as follows:

- Appendix A provides proofs and supporting theoretical analysis.
  - Proof of the routing-recovery theorem (Section A.1).
  - Task-weighted extension and gradient outer products (Section A.2).
  - Population guarantees for output decorrelation (Section A.3).
  - Proof of the expert-output difference bound (Section A.4).
  - Proof of the shared-diversity theorem (Section A.5).
  - Sampled determinants and the implemented regularizer (Section A.6).
- Appendix B provides experimental details and additional analyses.
  - Datasets (Section B.1).
  - Baselines (Section B.2).
  - Dataset-specific training schedules (Section B.3).
  - Sensitivity to regularization coefficients (Section B.4).
  - Expert-output similarity analysis (Section B.5).
  - Effect of the number of experts (Section B.6).
  - Multi-task learning with task-assigned experts (Section B.7).
  - Comparison with alternative specialization objectives (Section B.8).
- Appendix C compares the parameter counts of OD-MoLE and the baselines.

A PROOFS AND DETAILS OF THE THEORETICAL ANALYSIS

In this section, we follow the notation of Sections 3 and 4. Recall, the unscaled expert residual is $\boldsymbol{z}_i(\boldsymbol{h}) = \boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$, and the full output is $\boldsymbol{W}_0\boldsymbol{h} + c\sum_{i=1}^{N} g_i(\boldsymbol{h})\boldsymbol{z}_i(\boldsymbol{h})$, where $c = \gamma/r > 0$. The routing analysis fixes $\boldsymbol{h}$ and expert parameters, whereas the diversity analysis takes expectations over a fixed activation distribution with finite second moments. We first prove routing recovery and clarify its scope, then establish the shared-projection bound and its connection to the regularizer.

A.1 PROOF OF THEOREM 1

Proof. Fix an input activation $\boldsymbol{h}$, the expert parameters, and the target $\boldsymbol{t}_{\boldsymbol{h}}$. Assume $c > 0$, $0 < \underline{z} \le \|\boldsymbol{z}_i(\boldsymbol{h})\|_2 \le \overline{z} < \infty$ for every expert i, and $\delta(\boldsymbol{h}) < 1$.

Define

$$\boldsymbol{D} = \mathrm{diag}(\|\boldsymbol{z}_1\|_2, \ldots, \|\boldsymbol{z}_N\|_2), \qquad \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{D}^{-1}, \qquad \boldsymbol{R} = \boldsymbol{V}^\top\boldsymbol{V}. \tag{16}$$

The positive norm lower bound makes $\boldsymbol{D}$ invertible. The columns of $\boldsymbol{V}$ are normalized expert outputs, so $\boldsymbol{R}$ has unit diagonal and pairwise cosine similarities off the diagonal. Hence,

$$\|\boldsymbol{R} - \boldsymbol{I}\|_{\mathrm{F}}^2 = N(N-1)\omega(\boldsymbol{h}) = \delta(\boldsymbol{h})^2. \tag{17}$$

Since the operator norm is bounded by the Frobenius norm, we have $\|\boldsymbol{R} - \boldsymbol{I}\|_{\mathrm{op}} \le \delta(\boldsymbol{h})$. Every eigenvalue of $\boldsymbol{R}$ therefore lies between $1 - \delta(\boldsymbol{h})$ and $1 + \delta(\boldsymbol{h})$, giving

$$(1 - \delta(\boldsymbol{h}))\boldsymbol{I} \preceq \boldsymbol{R} \preceq (1 + \delta(\boldsymbol{h}))\boldsymbol{I}. \tag{18}$$

Define the routing Gram matrix $\boldsymbol{G} = c^2\boldsymbol{Z}^\top\boldsymbol{Z}$. The $\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$ and the output-norm bounds imply

$$c^2\underline{z}^2(1 - \delta(\boldsymbol{h}))\boldsymbol{I} \le \boldsymbol{G} \le c^2\overline{z}^2(1 + \delta(\boldsymbol{h}))\boldsymbol{I}, \qquad \boldsymbol{G} = c^2\boldsymbol{Z}^\top\boldsymbol{Z}. \tag{19}$$

For any $\boldsymbol{g}, \widehat{\boldsymbol{g}} \in \Delta_N$, we have

$$\|c\boldsymbol{Z}(\boldsymbol{g} - \widehat{\boldsymbol{g}})\|_2^2 = (\boldsymbol{g} - \widehat{\boldsymbol{g}})^\top\boldsymbol{G}(\boldsymbol{g} - \widehat{\boldsymbol{g}}) \ge c^2\underline{z}^2(1 - \delta(\boldsymbol{h}))\|\boldsymbol{g} - \widehat{\boldsymbol{g}}\|_2^2. \tag{20}$$

This proves Eq. 7.

Since Δ_N is compact and $Q_{\boldsymbol{h}}$ is continuous, a minimizer $\boldsymbol{g}^*$ exists. Its Hessian with respect to routing weights is $\boldsymbol{G} \succ 0$ by $\delta(\boldsymbol{h}) < 1$ and $c > 0$, so this minimizer $\boldsymbol{g}^*$ is unique. At this constrained minimizer,

$$\nabla Q_{\boldsymbol{h}}(\boldsymbol{g}^*)^\top (\boldsymbol{g} - \boldsymbol{g}^*) \geq 0, \qquad \boldsymbol{g} \in \Delta_N. \tag{21}$$

Since the simplex is convex, moving from $\boldsymbol{g}^*$ toward any feasible $\boldsymbol{g}$ remains feasible. Therefore, the optimality of $\boldsymbol{g}^*$ requires the directional derivative along $\boldsymbol{g} - \boldsymbol{g}^*$ to be nonnegative.

For $\boldsymbol{v} = \boldsymbol{g} - \boldsymbol{g}^*$, the exact quadratic expansion gives

$$Q_{\boldsymbol{h}}(\boldsymbol{g}) - Q_{\boldsymbol{h}}(\boldsymbol{g}^*) = \nabla Q_{\boldsymbol{h}}(\boldsymbol{g}^*)^\top \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^\top \boldsymbol{G}\boldsymbol{v} \geq \frac{c^2 \underline{z}^2 (1 - \delta(\boldsymbol{h}))}{2} \|\boldsymbol{g} - \boldsymbol{g}^*\|_2^2. \tag{22}$$

Rearranging the above proves Eq. 8. Eq. 22 also proves uniqueness, since the distance between two minimizers must be zero. The proof also applies when some optimal routing weights are zero. □

The output orthogonality regularizer of OD-MoLE targets $\omega(\boldsymbol{h})$ on input activations. The theorem explains when the norm assumptions hold and $\delta(\boldsymbol{h}) < 1$, differences in routing weights cannot diminish in the mixture output. Consequently, a small excess residual loss implies proximity to the oracle mixture weights for the fixed target $\boldsymbol{t}_{\boldsymbol{h}}$.

Role and scope of the assumptions. The positive norm lower bound matters because directional separation alone does not control the magnitude of output changes. For example, outputs of the form $\boldsymbol{z}_i = \epsilon \boldsymbol{e}_i$ are mutually orthogonal, but satisfy

$$\|c\boldsymbol{Z}(\boldsymbol{g} - \widehat{\boldsymbol{g}})\|_2 = c\varepsilon \|\boldsymbol{g} - \widehat{\boldsymbol{g}}\|_2. \tag{23}$$

As ϵ approaches zero, distinct routing vectors become difficult to distinguish from their outputs even though their normalized directions remain perfectly separated.

The condition $\delta(\boldsymbol{h}) < 1$, equivalently $\omega(\boldsymbol{h}) < 1/[N(N-1)]$, ensures positive definiteness of the normalized Gram matrix. It is a sufficient rather than necessary condition for routing identifiability. In particular, it implies that $\boldsymbol{Z}$ has full column rank and hence $N \leq d$.

Theorem 1 imposes these conditions at the activation being analyzed. If $\boldsymbol{A}\boldsymbol{h} = 0$, every expert residual vanishes. More generally, for any scalar $s > 0$, because $\boldsymbol{z}_i(s\boldsymbol{h}) = s\boldsymbol{z}_i(\boldsymbol{h})$, a uniform positive lower bound cannot hold on an input set containing arbitrary small scalings of a nonzero activation.

Example 1 (routing recovery without decorrelation). The following example shows that routing weights can be recovered uniquely and stably even when expert outputs point in opposite directions, so the correlation condition is sufficient but not necessary for recovery. Fix a nonzero vector $\boldsymbol{v} \in \mathbb{R}^d$ and let $\boldsymbol{z}_1 = \boldsymbol{v}$ and $\boldsymbol{z}_2 = -\boldsymbol{v}$. Their squared cosine similarity is one, so $\omega(\boldsymbol{h}) = 1$ and $\delta(\boldsymbol{h}) = \sqrt{2} > 1$. Nevertheless, for $\boldsymbol{g} = (a, 1-a)^\top$ with $a \in [0, 1]$,

$$c\boldsymbol{Z}\boldsymbol{g} = c(2a - 1)\boldsymbol{v}. \tag{24}$$

This output uniquely determines a. More precisely, for $\widehat{\boldsymbol{g}} = (b, 1-b)^\top$,

$$\|c\boldsymbol{Z}(\boldsymbol{g} - \widehat{\boldsymbol{g}})\|_2^2 = 2c^2 \|\boldsymbol{v}\|_2^2 \|\boldsymbol{g} - \widehat{\boldsymbol{g}}\|_2^2. \tag{25}$$

Thus, squared-cosine regularization can penalize configurations that already distinguish routing weights. The theorem assumes nonzero expert outputs so that cosine similarities are well-defined, but this assumption is not necessary for routing weights to be uniquely recoverable.

The relevant routing directions. Example 1 above is possible because routing differences satisfy $\mathbf{1}^\top(\boldsymbol{g} - \widehat{\boldsymbol{g}}) = 0$. For $N \geq 2$, define

$$m_\Delta(\boldsymbol{h}) = \min_{\substack{\mathbf{1}^\top \boldsymbol{v} = 0 \\ \|\boldsymbol{v}\|_2 = 1}} \|\boldsymbol{Z}(\boldsymbol{h})\boldsymbol{v}\|_2^2. \tag{26}$$

Different routing vectors in Δ_N produce different mixture outputs if and only if $m_\Delta(\boldsymbol{h}) > 0$. Equivalently, the expert outputs are affinely independent, so each mixture has a unique representation by routing weights. If the minimum is zero, there exists a nonzero vector $\boldsymbol{v}$ satisfying $\mathbf{1}^\top \boldsymbol{v} = 0$ and

$\boldsymbol{Z}(\boldsymbol{h})\boldsymbol{v} = \mathbf{0}$. For sufficiently small $s > 0$, both routing vectors $\boldsymbol{g}_0 = \frac{1}{N}\mathbf{1}$ and $\boldsymbol{g}_0 + s\boldsymbol{v}$ belong to Δ_N and produce the same mixture output. Conversely, a positive minimum gives

$$\|\boldsymbol{Z}(\boldsymbol{h})(\boldsymbol{g} - \widehat{\boldsymbol{g}})\|_2^2 \geq m_\Delta(\boldsymbol{h})\|\boldsymbol{g} - \widehat{\boldsymbol{g}}\|_2^2, \tag{27}$$

so distinct routing vectors produce distinct outputs. We use the same reasoning as in the previous proof, where the routing weights must remain in the simplex. Then we have

$$\|\boldsymbol{g} - \boldsymbol{g}^*\|_2^2 \leq \frac{2[Q_{\boldsymbol{h}}(\boldsymbol{g}) - Q_{\boldsymbol{h}}(\boldsymbol{g}^*)]}{c^2 m_\Delta(\boldsymbol{h})}. \tag{28}$$

Theorem 1 provides a convenient lower bound $m_\Delta(\boldsymbol{h}) \geq \underline{z}^2(1 - \delta(\boldsymbol{h}))$. Unique routing recovery can hold with more experts than the theorem's sufficient condition permits. There are two different requirements. First, affine independence only requires that different routing weights summing to one give different outputs. In $\mathbb{R}^d$, up to $d + 1$ expert outputs can satisfy this. Second, full column rank requires all N output vectors to be linearly independent, a stronger condition that permits at most d vectors in $\mathbb{R}^d$. Thus, $\delta(\boldsymbol{h}) < 1$ is sufficient for routing recovery but excludes some configurations where recovery is still possible.

A.2 Extension to task-weighted routing objectives

Theorem 1 measures residual differences in Euclidean distance. We now consider a weighted quadratic objective to account for different sensitivities along output directions. Fix the input $\boldsymbol{h}$, expert parameters, target $\boldsymbol{t}_{\boldsymbol{h}}$, and a symmetric matrix $\boldsymbol{C}_h \succ 0$. Define

$$Q_{\boldsymbol{h}}^C(\boldsymbol{g}) = \frac{1}{2}(c\boldsymbol{Z}\boldsymbol{g} - \boldsymbol{t}_{\boldsymbol{h}})^\top \boldsymbol{C}_h(c\boldsymbol{Z}\boldsymbol{g} - \boldsymbol{t}_{\boldsymbol{h}}), \qquad \boldsymbol{g} \in \Delta_N, \tag{29}$$

where $c > 0$ and $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{h})$. The Euclidean objective corresponds to $\boldsymbol{C}_h = \boldsymbol{I}$.

Assume the hypotheses of Theorem 1. Suppose further that

$$c_-\|\boldsymbol{u}\|_2^2 \leq \boldsymbol{u}^\top \boldsymbol{C}_h \boldsymbol{u} \leq c_+\|\boldsymbol{u}\|_2^2, \qquad 0 < c_- \leq c_+ < \infty, \tag{30}$$

for every $\boldsymbol{u}$ in the span of the expert outputs. Writing $\boldsymbol{G} = c^2\boldsymbol{Z}^\top\boldsymbol{Z}$, the Hessian of the weighted objective satisfies

$$c_-\boldsymbol{G} \preceq \boldsymbol{G}_C := c^2\boldsymbol{Z}^\top\boldsymbol{C}_h\boldsymbol{Z} \preceq c_+\boldsymbol{G}. \tag{31}$$

Thus, with $\mu_h = c^2\underline{z}^2(1 - \delta(\boldsymbol{h})) > 0$, the weighted objective has curvature at least $c_-\mu_h$. Applying the same constrained-optimality argument gives a unique minimizer $\boldsymbol{g}_C^*$ on Δ_N and

$$\|\boldsymbol{g} - \boldsymbol{g}_C^*\|_2^2 \leq \frac{2[Q_{\boldsymbol{h}}^C(\boldsymbol{g}) - Q_{\boldsymbol{h}}^C(\boldsymbol{g}_C^*)]}{c_-\mu_h}. \tag{32}$$

The lower comparison bound ensures that the weighted objective remains sensitive to changes in the expert mixture. The upper bound controls smoothness but is not needed for recovery. If $\boldsymbol{C}_h$ represents a local approximation to task curvature, this result applies to the quadratic surrogate. Extending it to the task loss itself requires additional control of curvature and approximation error.

Example 2 (separated residuals with identical task predictions). The following example highlights a key limitation of residual-space separation showing that even orthogonal expert residuals can produce identical predictive distributions, leaving the routing weights unidentifiable from the task loss. Consider two-class logit residuals with $c = 1$, zero base logits, and

$$\boldsymbol{z}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{z}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}. \tag{33}$$

Both have unit norm and are orthogonal, so $\omega(\boldsymbol{h}) = \delta(\boldsymbol{h}) = 0$. For $\boldsymbol{g} = (a, 1 - a)^\top$, however,

$$\boldsymbol{Z}\boldsymbol{g} = \begin{pmatrix} a \\ a - 1 \end{pmatrix}, \qquad \mathrm{softmax}(\boldsymbol{Z}\boldsymbol{g}) = \begin{pmatrix} \frac{e}{e+1} \\ \frac{1}{e+1} \end{pmatrix}. \tag{34}$$

Different values of a produce different logits, but their probabilities remain unchanged because softmax is invariant to adding the same constant to every logit. And cross-entropy loss for any fixed label is constant in a and cannot distinguish the routing weights. There is no contradiction with Theorem 1, which concerns the Euclidean residual objective. The cross-entropy Hessian has zero curvature along this routing-induced direction and therefore fails the positive-curvature assumption needed to transfer the recovery argument to the task loss.

A.3 Population guarantees for output decorrelation

The routing theorem applies at a fixed input, whereas OD-MoLE averages its output penalty over sampled activations. To connect these settings, fix expert parameters and consider an activation distribution under which all expert outputs are nonzero with probability one, so the cosine similarities are well-defined. Then $\omega(\boldsymbol{h})$ is defined, and Markov's inequality gives, for $0 < \zeta < 1$,

$$\Pr_{\boldsymbol{h}}[\delta(\boldsymbol{h}) > \zeta] = \Pr_{\boldsymbol{h}}[\omega(\boldsymbol{h}) > \zeta^2/(N(N-1))] \leq \frac{N(N-1)\mathbb{E}_{\boldsymbol{h}}\left[\omega(\boldsymbol{h})\right]}{\zeta^2}. \tag{35}$$

Therefore, a small expected penalty limits the probability of violating the correlation threshold. On inputs satisfying both $\delta(\boldsymbol{h}) \leq \zeta$ and $\min_i \|\boldsymbol{z}_i(\boldsymbol{h})\|_2 \geq \underline{z} > 0$, the recovery constant is at least $c^2\underline{z}^2(1-\zeta)$. The expected penalty alone does not enforce the output-norm condition.

A.4 Proof of Eq. 9

Proof. We next examine how the shared representation affects differences between expert outputs. Let $\boldsymbol{D}_{ij} = \boldsymbol{B}_i - \boldsymbol{B}_j$. By linearity of expectation and the trace identity, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{h}}\|\boldsymbol{z}_i(\boldsymbol{h}) - \boldsymbol{z}_j(\boldsymbol{h})\|_2^2 &= \mathbb{E}_{\boldsymbol{h}}\|\boldsymbol{D}_{ij}\boldsymbol{A}\boldsymbol{h}\|_2^2 = \operatorname{tr}(\boldsymbol{D}_{ij}\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{D}_{ij}^\top) \\ &\geq \lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})\operatorname{tr}(\boldsymbol{D}_{ij}\boldsymbol{D}_{ij}^\top) = \lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})\|\boldsymbol{B}_i - \boldsymbol{B}_j\|_{\mathrm{F}}^2. \end{aligned} \tag{36}$$

The inequality follows from $\boldsymbol{K}_{\boldsymbol{A}} \succeq \lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})\boldsymbol{I}$. □

Thus, a positive lower bound on $\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})$ ensures that every nonzero reconstruction difference remains observable in mean-square output. The next theorem provides such a bound. For scaled expert residuals $c\boldsymbol{z}_i(\boldsymbol{h})$, the corresponding lower bound includes an additional factor c^2.

A.5 Proof of Theorem 2

Proof. Let $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{A}^\top$ with eigenvalues $0 < \lambda_1 \leq \cdots \leq \lambda_r$. The trace and log-determinant assumptions imply $\sum_j \lambda_j = r$ and $\prod_j \lambda_j \geq e^{-\eta}$. By the arithmetic–geometric mean inequality,

$$e^{-\eta} \leq \lambda_1 \prod_{j=2}^{r} \lambda_j \leq \lambda_1 \left(\frac{r-\lambda_1}{r-1}\right)^{r-1} \leq \lambda_1 \left(\frac{r}{r-1}\right)^{r-1}. \tag{37}$$

Therefore,

$$\lambda_{\min}(\boldsymbol{P}) \geq e^{-\eta}\left(\frac{r-1}{r}\right)^{r-1}. \tag{38}$$

For any $\boldsymbol{w} \in \mathbb{R}^r$, $\boldsymbol{A}^\top\boldsymbol{w}$ belongs to the row space of $\boldsymbol{A}$. Applying the activation assumption gives

$$\boldsymbol{w}^\top\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{w} = (\boldsymbol{A}^\top\boldsymbol{w})^\top\boldsymbol{\Sigma}_{\boldsymbol{h}}(\boldsymbol{A}^\top\boldsymbol{w}) \geq \nu\|\boldsymbol{A}^\top\boldsymbol{w}\|_2^2 = \nu\boldsymbol{w}^\top\boldsymbol{P}\boldsymbol{w}. \tag{39}$$

Hence $\boldsymbol{K}_{\boldsymbol{A}} \succeq \nu\boldsymbol{P}$, which proves Eq. 10. □

This highlights how projection diversity plays a supporting role by maintaining the output behavior of the original reconstructions whenever inputs activate those primary directions.

Role of the assumptions. The trace controls the sum of squared singular values of $\boldsymbol{A}$, while the determinant controls their product. Eigenvalues $(\varepsilon, \varepsilon^{-1})$ have determinant one, whereas $(\varepsilon, 2-\varepsilon)$ have trace two, but both have a vanishing minimum as $\varepsilon \to 0$. The log-determinant assumption also requires full row rank, hence $r \leq k$. The input assumption is needed only on the row space of $\boldsymbol{A}$. It ensures nonzero second-moment energy in the retained directions. Conditioning $\boldsymbol{A}$ cannot supply signal absent from the data. When $\boldsymbol{A}$ changes during training, the condition must be checked for its resulting row space. At $\eta = 0$, trace normalization forces $\boldsymbol{P} = \boldsymbol{I}$, giving the sharper bound $\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}}) \geq \nu$. The same conclusion holds for $r = 1$ under trace normalization.

Example 3 (pairwise expert differences with an ambiguous mixture). We provide a simple example that demonstrates while two routing assignments can yield identical mixture outputs, pairwise expert separation alone is insufficient to guarantee unique routing recovery. Suppose the assumptions of Theorem 2 hold. Choose $\boldsymbol{B}_1 \neq \boldsymbol{B}_2$ and set

$$\boldsymbol{B}_3 = \frac{1}{2}(\boldsymbol{B}_1 + \boldsymbol{B}_2). \tag{40}$$

All three matrices are pairwise different. Eq. 9 and the positive lower bound on $\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})$ therefore imply positive mean-square output differences for every pair. Yet linearity gives, for every activation,

$$\boldsymbol{z}_3(\boldsymbol{h}) = \frac{1}{2}\boldsymbol{z}_1(\boldsymbol{h}) + \frac{1}{2}\boldsymbol{z}_2(\boldsymbol{h}). \tag{41}$$

Consequently, the distinct routing vectors $\boldsymbol{g} = (\frac{1}{2}, \frac{1}{2}, 0)^\top$ and $\widehat{\boldsymbol{g}} = (0, 0, 1)^\top$ satisfy $c\boldsymbol{Z}(\boldsymbol{h})\boldsymbol{g} = c\boldsymbol{Z}(\boldsymbol{h})\widehat{\boldsymbol{g}}$. This illustrates why preserving pairwise expert differences does not ensure routing recovery. The shared-projection result preserves existing differences in mean square, while routing recovery requires separation of all nonzero zero-sum combinations of expert outputs.

A.6 Connection to the sampled diversity regularizer

Theorem 2 uses the determinant of the full projection Gram matrix, whereas the regularizer uses sampled columns.We now relate this condition to the penalty evaluated on sampled columns of $\boldsymbol{A}$, keeping the projection fixed during the sampling analysis.

Sampled determinants. Assume $r \leq k$, and let $\mathcal{S}$ be sampled uniformly from all sets of exactly r columns. Then $\boldsymbol{A}_{\mathcal{S}} \in \mathbb{R}^{r\times r}$, and the Cauchy–Binet formula gives

$$\det(\boldsymbol{A}\boldsymbol{A}^\top) = \sum_{|\mathcal{S}|=r} \det(\boldsymbol{A}_{\mathcal{S}})^2 = \sum_{|\mathcal{S}|=r} \det(\boldsymbol{A}_{\mathcal{S}}^\top \boldsymbol{A}_{\mathcal{S}}). \tag{42}$$

Under uniform sampling of r columns, this equals $\binom{k}{r}\mathbb{E}_{\mathcal{S}}\left[\det(\boldsymbol{A}_{\mathcal{S}}^\top \boldsymbol{A}_{\mathcal{S}})\right]$. Taking logarithms and applying Jensen's inequality gives

$$\log\det(\boldsymbol{A}\boldsymbol{A}^\top) \geq \log\binom{k}{r} + \mathbb{E}_{\mathcal{S}}\left[\log\det(\boldsymbol{A}_{\mathcal{S}}^\top \boldsymbol{A}_{\mathcal{S}})\right]. \tag{43}$$

We note that if any sampled matrix has a determinant of zero, the expected log determinant is $-\infty$ under uniform sampling, and the inequality has no finite lower bound.

Finite-ϵ regularization. For the implemented penalty, define its expectation over column sampling as

$$D_\epsilon = -\mathbb{E}_{\mathcal{S}}\left[\log\det(\boldsymbol{A}_{\mathcal{S}}^\top \boldsymbol{A}_{\mathcal{S}} + \epsilon\boldsymbol{I})\right], \quad \epsilon > 0 \tag{44}$$

Since $\boldsymbol{A}\boldsymbol{A}^\top \succeq \boldsymbol{A}_{\mathcal{S}}\boldsymbol{A}_{\mathcal{S}}^\top$ and $\boldsymbol{A}_{\mathcal{S}}$ is square, we have

$$\log\det(\boldsymbol{A}\boldsymbol{A}^\top + \epsilon\boldsymbol{I}) \geq -D_\epsilon. \tag{45}$$

If $\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^\top) \leq T_A$ and $r > 1$, we have

$$\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top) \geq \left[e^{-D_\epsilon}\left(\frac{r-1}{T_A + (r-1)\epsilon}\right)^{r-1} - \epsilon\right]_+, \tag{46}$$

where $[x]_+ = \max(x, 0)$. Under the activation assumption of Theorem 2, $\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}}) \geq \nu\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)$. Thus, multiplying $\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^\top)$ by ν gives a lower bound on $\lambda_{\min}(\boldsymbol{K}_{\boldsymbol{A}})$.

Scale control. For $s > 0$, the transformation $\boldsymbol{A} \mapsto s\boldsymbol{A}$ and $\boldsymbol{B}_i \mapsto \boldsymbol{B}_i/s$ keep every expert output unchanged. If a sampled submatrix has positive singular values $\sigma_1, \ldots, \sigma_q$, with $q > 0$, its negative log determinant becomes

$$-\sum_{j=1}^{q}\log(s^2\sigma_j^2 + \epsilon) - (r-q)\log\epsilon \longrightarrow -\infty \quad \text{as } s \to \infty. \tag{47}$$

Thus, the diversity penalty can decrease through rescaling without changing the expert output. A scale constraint or suitable norm control is therefore needed to connect the penalty to preservation of latent directions. Theorem 2 provides this connection under its stated scale and input assumptions, minimizing the log-determinant penalty alone does not enforce them.

B EXPERIMENT DETAILS

B.1 DATASETS

We summarize the details of the used datasets as follows:

- CommonsenseQA (ComQA) (Talmor et al., 2019): it challenges the model's grasp of everyday knowledge.
- CosmosQA (CosQA) (Huang et al., 2019): it tests inferential reasoning through narrative comprehension.
- SocialIQA (SIQA) (Sap et al., 2019): it examines the model's understanding of social interactions by asking questions about motivations, reactions, or outcomes in social contexts.
- PhysicalIQA (PIQA) (Bisk et al., 2020): it examines the model's understanding of physical phenomena to select the most plausible solution to a given problem.
- ScienceQA (ScQA) (Lu et al., 2022): it examines the model's understanding of scientific principles.
- BoolQ (Clark et al., 2019): it's a QA task where each example consists of a short passage and a yes/no question about the passage, which is to assess the model's ability to evaluate the veracity of statements.
- ARC-Challenge (ARC-C) (Clark et al., 2018): it's a more difficult subset of ARC, containing questions that require advanced reasoning and knowledge.
- OpenBookQA (OBQA) (Mihaylov et al., 2018): it's a question-answering task requiring reasoning and knowledge from a small "open book" of science facts.
- COPA (Roemmele et al., 2011): it's a causal reasoning task where a system is given a premise sentence and must determine either the cause or the effect of the premise from two possible choices.

B.2 BASELINES

- LoRA (Hu et al., 2022): it utilizes trainable low-rank matrices for parameter-efficient fine-tuning. We change the rank from 8 to 32.
- PiSSA (Meng et al., 2024): it initializes LoRA matrices with the principal components of the pre-trained weights.
- HydraLoRA (Tian et al., 2024): it proposes an asymmetric LoRA framework that improves parameter efficiency and performance by addressing training inefficiencies.
- MoLoRA (Zadouri et al., 2024): it combines MoE architecture with lightweight experts, enabling extremely parameter-efficient fine-tuning by updating less than 1% of model parameters.
- AdaMoLE (Liu & Luo, 2024): it uses adaptive mechanisms to optimize the selection of experts.
- GOAT (Fan et al., 2025): it adaptively integrates pre-trained knowledge using an SVD-structured MoE framework, addressing the limitations of non-informative or static priors.

B.3 DATASET-SPECIFIC TRAINING SCHEDULES

Different commonsense reasoning datasets exhibit varying levels of complexity and convergence behavior. To ensure stable and fair evaluation, we adopt dataset-specific training schedules rather than using a fixed number of epochs across all tasks. Tab. 5 summarizes the number of training epochs used for each dataset. Importantly, we select the training epochs based on validation performance to avoid underfitting or overfitting across tasks. This design ensures that all compared methods are evaluated under comparable convergence conditions, allowing for a fair assessment of model capability rather than training duration. We note that using a uniform training schedule across datasets leads to suboptimal performance for harder tasks and premature convergence for easier ones, highlighting the importance of task-aware training budgets in multi-dataset evaluation.

Table 5: Dataset-specific training schedules used in our experiments.

Dataset	#Epochs
ComQA	2
CosQA	2
SIQA	4
PIQA	7
ScQA	10
BoolQ	8
OBQA	9
ARC-C	16

Table 6: Sensitivity of OD-MoLE to λ_{orth} and λ_{div} coefficients on COPA using Qwen2.5-7B. Results are reported as accuracy under different coefficient combinations in the expert composition setting (2 experts per submodule).

$\lambda_{\text{orth}} \backslash \lambda_{\text{div}}$	10^{-5}	10^{-4}	10^{-3}
10^{-5}	0.95	0.94	0.94
10^{-4}	0.95	0.95	0.95
10^{-3}	0.96	0.96	0.96

B.4 Sensitivity to Regularization Coefficients

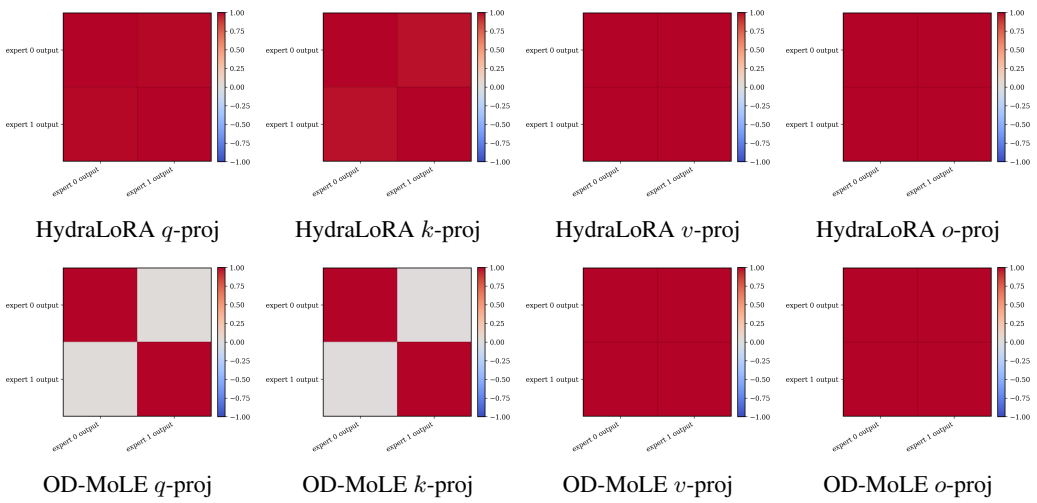
We analyze the sensitivity of OD-MoLE to the orthogonality and diversity coefficients, λ_{orth} and λ_{div}. We vary one coefficient while fixing the other to 10^{-3} and report the results using Qwen2.5-7B on the COPA dataset in Tab. 6. We use two experts per submodule and compose their outputs during training. As the coefficients increase from 10^{-5} to 10^{-3}, performance improves, suggesting that moderate regularization effectively encourages expert diversity and reduces interference. Overall, the results show that OD-MoLE is robust to a reasonable range of coefficient values, with 10^{-3} providing a good balance between promoting diversity and preserving model capacity.

B.5 Expert Output Similarity Analysis

To quantitatively assess expert specialization, we analyze the pairwise cosine similarity between expert output representations $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$. Figs 2 and 3 present the similarity heatmaps across four attention projection modules (q, k, v, and o) at both early (layer 0) and deeper (layer 31) layers of Llama-2-7B. We use two experts per submodule and compose their outputs during training. We observe that HydraLoRA consistently exhibits high off-diagonal similarity values across all projection modules, indicating that different experts produce highly correlated outputs and thus learn redundant transformations. In contrast, OD-MoLE substantially reduces off-diagonal cosine similarities, particularly in the q- and k-projection modules, where the off-diagonal entries approach zero. This indicates that expert outputs become nearly orthogonal, demonstrating that our method effectively encourages diverse and complementary representations. The improvement is consistent across both shallow and deep layers, showing that OD-MoLE promotes specialization throughout the network. Interestingly, both methods exhibit relatively high similarity in the v- and o-projection modules. Overall, these results provide strong quantitative evidence that OD-MoLE reduces output-level interference and improves expert specialization by enforcing diversity in the representation space, which aligns with the qualitative observations from t-SNE visualizations and the improvements in downstream task performance.

B.6 Effect of the number of experts

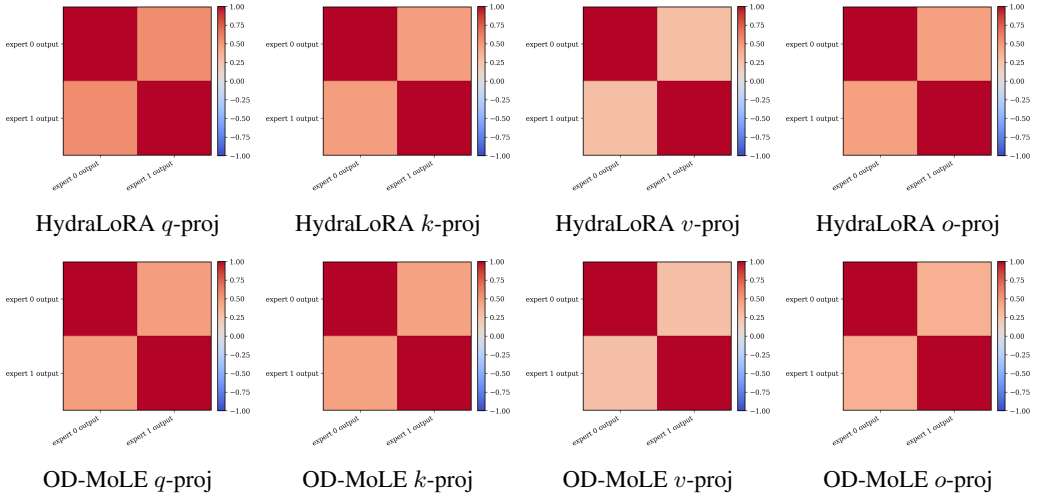
We vary $N \in \{8, 16, 32\}$ while fixing the total expert rank $Nr_i = 32$ and routing ratio $K/N = 1/4$, keeping the active expert rank at $Kr_i = 8$, where r_i is rank of each expert i. Tab. 7 shows that OD-MoLE outperforms HydraLoRA in all three configurations on Qwen2.5-7B and achieves its highest



HydraLoRA q-proj HydraLoRA k-proj HydraLoRA v-proj HydraLoRA o-proj

OD-MoLE q-proj OD-MoLE k-proj OD-MoLE v-proj OD-MoLE o-proj

Figure 2: Heatmap of expert output representations $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$ from four attention projection modules at layer 0. The first row shows HydraLoRA, while the second row shows OD-MoLE. Compared with HydraLoRA, OD-MoLE exhibits clearer separation between expert output distributions, indicating improved expert specialization and reduced output-level interference.



HydraLoRA q-proj HydraLoRA k-proj HydraLoRA v-proj HydraLoRA o-proj

OD-MoLE q-proj OD-MoLE k-proj OD-MoLE v-proj OD-MoLE o-proj

Figure 3: Heatmap of expert output representations $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$ from four attention projection modules at layer 31. The first row shows HydraLoRA, while the second row shows OD-MoLE. Compared with HydraLoRA, OD-MoLE exhibits clearer separation between expert output distributions, indicating improved expert specialization and reduced output-level interference.

accuracy at $N = 8$. Increasing expert count while reducing per-expert rank therefore does not consistently improve performance.

B.7 Multi-Task Learning with Task-Assigned Experts

Experimental setup. We evaluate OD-MoLE and HydraLoRA in a multi-task setting using Qwen2.5-7B and all eight commonsense reasoning benchmarks. We use 1,000 training examples per task, eight experts with rank four per expert, and five training epochs. Each task i is assigned to one expert-specific reconstruction matrix $\boldsymbol{B}_i$, and all tokens from that task are routed to the assigned expert, bypassing the learned router. The task loss for task i therefore updates $\boldsymbol{B}_i$ and the shared projection $\boldsymbol{A}$. This restriction concerns task-loss gradients and regularization may additionally couple the experts.

Table 7: Effect of expert count on Qwen2.5-7B across eight commonsense reasoning tasks. The total expert rank $Nr_i = 32$ and routing ratio $K/N = 1/4$ are fixed, giving an active expert rank of $Kr_i = 8$, where r_i is the rank of each expert. Bold indicates the higher accuracy within each setting.

Experts N	Top-K	Rank r_i	Active rank Kr_i	HydraLoRA	OD-MoLE	Gain
8	2	4	8	86.68	**87.04**	+0.36
16	4	2	8	86.05	**86.23**	+0.18
32	8	1	8	85.99	**86.53**	+0.54

Table 8: Multi-task results on Qwen2.5-7B with one expert assigned to each task. Difference denotes OD-MoLE minus HydraLoRA in percentage points. **Bold** indicates the higher accuracy in each column.

Method	ComQA	CosQA	SIQA	PIQA	ScQA	BoolQ	ARC-C	OBQA	Avg.
HydraLoRA	**84.68**	79.77	**74.77**	**85.50**	84.84	85.29	85.24	91.60	83.96
OD-MoLE	84.11	**80.74**	74.72	84.90	**85.40**	**85.57**	**85.75**	**93.20**	**84.30**
Difference	-0.57	$+0.97$	-0.05	-0.60	$+0.56$	$+0.28$	$+0.51$	$+1.60$	$+0.34$

Results and discussion. Tab. 8 shows that OD-MoLE achieves an average accuracy of 84.30, exceeding HydraLoRA by 0.34 percentage points and improving five of eight tasks. These results suggest that OD-MoLE's benefits extend to explicit task-to-expert assignment, although improvements are not uniform across tasks.

Task assignment does not eliminate cross-task interaction, since all tasks update the shared projection $\boldsymbol{A}$. The experiment therefore evaluates task-specific reconstruction mappings within a jointly trained shared representation. It requires task identity at inference and does not measure learned-router quality.

B.8 Comparison with Alternative Specialization Objectives

We compare OD-MoLE with the specialization objectives proposed by Guo et al. (2025). We implement their objectives within the same LoRA MoE framework and evaluate both methods using Qwen2.5-7B, the same datasets and LoRA rank, and two experts per submodule with both experts active. This comparison evaluates the objectives within our experimental setting.

Table 9: Comparison with the objectives of Guo et al. (2025) implemented in the same LoRA MoE framework. Both methods use Qwen2.5-7B with all-expert composition. **Bold** indicates the higher score in each column.

Method	ComQA	CosQA	SIQA	PIQA	ScQA	BoolQ	ARC-C	OBQA	Avg.
Guo et al. (adapted)	83.46	76.25	76.41	**87.87**	78.13	80.90	82.76	**95.20**	82.62
OD-MoLE	**85.50**	**84.85**	**79.63**	87.65	**92.58**	**83.80**	**83.96**	94.80	**86.60**

Tab. 9 shows that OD-MoLE outperforms Guo et al. (2025) on average. The adapted objectives achieve higher scores on PIQA and OBQA. Guo et al. (2025) regularize selected dense-expert representations and routing-score dispersion in a general MoE, while OD-MoLE is designed in the asymmetric LoRA framework. We regularize data-dependent residuals $\boldsymbol{B}_i\boldsymbol{A}\boldsymbol{h}$ to reduce redundancy among expert-specific mappings while separately preserving the capacity of shared low-rank projection $\boldsymbol{A}$. These results support the effectiveness of OD-MoLE's regularization design in the evaluated LoRA MoE setting.

Table 10: Comparison of trainable and activated parameters across vanilla LoRA with varying rank and different adaptation methods.

Method	**Trainable parameters**		**Activated parameters**	
	Count	Proportion	Count	Proportion
LoRA ($r = 8$)	$8drL$	0.12%	—	—
LoRA ($r = 16$)	$8drL$	0.25%	—	—
LoRA ($r = 32$)	$8drL$	0.50%	—	—
MoLoRA	$8drNL + 4dNL$	0.56%	$8drkL + 4dNL$	0.19%
GOAT	$8drNL + 4dNL$	0.56%	$8drkL + 4dNL$	0.19%
AdaMoLE	$8drNL + 4dNL$	0.56%	$8drNaL + 4dNL$	0.19%
HydraLoRA	$4(N+1)drL + 4dNL$	0.34%	$4(N+1)drL + 4dNL$	0.34%
OD-MoLE	$4(N+1)drL + 4dNL$	0.34%	$4(N+1)drL + 4dNL$	0.34%

C Parameter Analysis

We provide a parameter analysis for each baseline and our proposed method based on the Llama-2-7B model in Tab. 10. We assume d represents the model dimension, r denotes the rank of LoRA, N indicates the number of experts, L indicates the number of layers, top-k is top-2, and a is the selected number of experts using threshold in AdaMoLE. LoRA is applied to the attention projection matrices $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v, \boldsymbol{W}_o$, and only the corresponding LoRA parameters are trainable. The proportion is measured relative to the total number of whole model parameters.