

LAK-LoRA: Layer-Wise Adaptive Expert Allocation and Principal Component–Based Feature Separation for Medical Multi-Task Learning

Ye Lyu¹, Chuan Li¹, Chengyu Wang², Mingyuan Fan¹, Cen Chen^{1*}

¹East China Normal University, Shanghai, China

²Alibaba Group, Hangzhou, China

{yelyu, lichuan, mingyuan_fmy}@stu.ecnu.edu.cn

chengyu.wcy@alibaba-inc.com, cenchen@dase.ecnu.edu.cn

Abstract

Multi-task learning in the medical domain remains a significant challenge for large language models (LLMs), plagued by conflicting task objectives and substantial computational overhead associated with full fine-tuning. Although parameter-efficient tuning methods help reduce these resource demands, their effectiveness is limited by two critical issues: (i) rigid allocation of expert capacity across layers, and (ii) lack of regularization mechanisms to partition shared components from task-specific signals. To address these limitations, we propose *LAK-LoRA*, a novel architecture that enables adaptive, depth-aware scaling of expert capacity, allowing more efficient resource allocation throughout the model. Furthermore, to resolve conflicting task objectives, *LAK-LoRA* divides experts into two distinct groups dedicated to modeling general and task-specific features, respectively. This separation is implemented via a feature separator module, which disentangles representations into shared and task-specific subspaces before assigning them to the appropriate experts. To foster diversity and minimize representational redundancy, orthogonality constraints are imposed among specialized experts. Extensive experiments on benchmark medical datasets demonstrate that *LAK-LoRA* achieves the best average performance.

1 Introduction

The superior performance of large language models (LLMs) (Chowdhery et al., 2022; Touvron et al., 2023; Qwen et al., 2024; DeepSeek-AI et al., 2024; Team et al., 2025) has opened promising avenues for their application in the medical domain (Liu et al., 2024a; Aljohani et al., 2025). However, the medical domain encompasses a wide range of tasks, and given the enormous number of parameters in LLMs, fine-tuning a separate model for each individual task is computationally prohibitive. As a

* Corresponding author

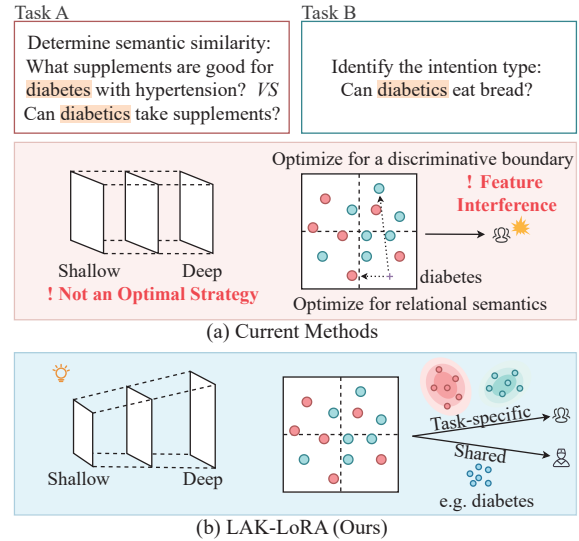


Figure 1: Comparison of methods. (a) Prior methods rely on static expert allocation and entangled representations. (b) LAK-LoRA employs adaptive expert allocation and explicit feature separation.

result, multi-task learning (MTL) has become a critical research direction. Yet, in the medical context, the optimization objectives of different tasks often conflict with each other. For example, medical entity recognition (Monajatipoor et al., 2024) requires fine-grained local representations; intention classification (Shi and Mariano, 2025) emphasizes discriminative boundaries; medical report generation (Jin et al., 2024a) relies on capturing global semantics; and medical dialogue generation (Sun et al., 2025) prioritizes both accuracy and fluency. These divergent requirements hinder the learning of unified feature representations, resulting in significant performance degradation known as the “see-saw phenomenon” (Liu et al., 2025).

To address these challenges, researchers have focused on multi-task learning methods based on parameter-efficient fine-tuning (PEFT) (Mangrulkar et al., 2022). Among them, low-rank adaptation (LoRA) (Hu et al., 2022) has emerged as one

of the most widely adopted approaches. By freezing the pretrained backbone and learning only low-rank update matrices, LoRA drastically reduces the number of trainable parameters. Building upon this foundation, several multi-task extensions of LoRA have been explored, such as ensembling multiple LoRA adapters (Wang et al., 2023), incorporating the mixture-of-experts (MoE) architecture (Liu et al., 2023b), and applying task-specific transformation modules (Yang et al., 2025).

Despite these advancements, existing approaches face two critical limitations, as shown in Figure 1(a). First, prior works (Liu et al., 2023b) commonly adopt a uniform allocation strategy, assigning the same number of experts to each layer. This design overlooks a fundamental characteristic of LLMs: shallow layers predominantly capture general and transferable patterns, whereas deeper layers progressively encode more task-specific semantics. Although recent studies (Gao et al., 2024) indicate that deeper layers benefit from increased expert capacity, they typically rely on manually specified configurations that lack adaptability across different model scales.

Second, existing methods (Liu et al., 2023b; Wang et al., 2023; Yang et al., 2025) treat the input representations at each layer as entangled wholes. This forces experts to simultaneously process both shared, cross-task features and specific, task-dependent information, leading to severe feature interference. For illustration, consider two distinct medical tasks: a semantic similarity task comparing “*What supplements are good for diabetes with hypertension?*” with “*Can diabetics take supplements?*”, and an intention classification task for the query “*Can diabetics eat bread?*”. Although both tasks share the core concept of “diabetes”, their optimization goals differ considerably. Directly processing such entangled representations makes it difficult for experts to effectively separate and model the distinct feature components, resulting in degraded multi-task performance.

To resolve the capacity mismatch and feature entanglement issues, we propose **LAK-LoRA**, a novel MoE-based framework specifically designed for medical multi-task learning. As shown in Figure 1(b), LAK-LoRA integrates two key components: **Layer-Wise Adaptive Expert Allocation** and **Principal Component-Based Feature Separation**. Concretely, the number of activated experts automatically increases with network depth, and experts are divided into one generic expert and

multiple specialized experts. To mitigate cross-task interference, we redesign the intra-layer information flow by introducing a feature separator before routing. Leveraging singular value decomposition (SVD), the separator explicitly projects the representations into two subspaces: a dominant shared subspace and a residual task-specific subspace based on variance structure. The shared subspace is routed to the generic expert for joint cross-task modeling, while the task-specific subspace is dispatched to specialized experts for further refinement. Finally, we impose orthogonality constraints among specialized experts to encourage diverse and complementary adaptations.

Our main contributions are summarized as follows:

- We propose LAK-LoRA, a novel framework that combines layer-wise adaptive expert allocation with principal component-based feature separation for medical multi-task learning.
- We introduce orthogonality constraints on specialized experts to reduce redundancy and promote the learning of diverse, complementary representations.
- Extensive experiments conducted on a comprehensive Chinese medical benchmark comprising 16 diverse NLP tasks demonstrate the superiority of our proposed method.

2 Related Work

2.1 Medical Large Language Models

With the rapid advancement of LLMs, the medical domain has seen the development of sophisticated models such as Med-Gemini (Saab et al., 2024) and Health-LLM (Jin et al., 2024b), which demonstrate strong performance across tasks including medical question answering, report generation, and clinical dialogues. Current efforts encompass a variety of approaches: the Med-PaLM series (Singhal et al., 2023, 2025) focuses on optimizing data for medical examinations and clinical reasoning; DISC-MedLLM (Bao et al., 2023) leverages knowledge graphs to construct dialogue samples; Qibo (Zhang et al., 2024) integrates traditional Chinese medicine knowledge to broaden its scope; Baichuan-M1 (Wang et al., 2025) employs a self-reflection mechanism with a reward model to enhance synthetic data quality and utilizes a multi-stage curriculum learning strategy to incrementally

improve both generic and medical-specific capabilities; and HuatuoGPT-II (Chen et al., 2024) unifies the conventional two-stage process of pre-training and supervised fine-tuning into a single-stage domain adaptation approach.

However, existing medical LLMs predominantly focus on data construction and training strategies, with relatively less emphasis on internal architecture. This gap restricts their potential for multi-task generalization and effective knowledge transfer. In contrast, we pivot towards architectural design by introducing targeted enhancements to the MoE framework, aiming to better address the diverse multi-task demands of the medical domain.

2.2 Multi-Task Learning

MTL is a foundational training paradigm designed to improve generalization by jointly optimizing a single shared model across multiple tasks (Crawshaw, 2020; Zhang and Yang, 2022; Zhang et al., 2023b). With the rise of LLMs, there has been increasing interest in MTL methods based on PEFT, particularly those utilizing LoRA (Hu et al., 2022). For example, MultiLoRA (Wang et al., 2023) integrates multiple independent LoRA modules to reduce inter-task parameter interference, while MOELoRA (Liu et al., 2023b) adopts an MoE framework with a task-driven gating network to dynamically adjust expert fusion weights. MTL-LoRA (Yang et al., 2025) further distinguishes task-specific information through learnable transformation parameters, while shared features are captured within low-dimensional subspaces.

Despite these advances, limitations remain: expert capacity is typically static, and intermediate representations are often treated as indivisible units. Accordingly, we propose layer-wise expert allocation and a mechanism for separating shared and task-specific features from intermediate representations, better capturing multi-task demands.

3 Methodology

3.1 Preliminary: Low-Rank Adaptation

LoRA (Hu et al., 2022) adapts a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$ via a low-rank update $\Delta \mathbf{W} = \mathbf{B}\mathbf{A}$, where $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times k}$ ($r \ll \min(d, k)$). For an input x , the forward pass is:

$$h = \mathbf{W}_0 x + \Delta \mathbf{W} x = \mathbf{W}_0 x + \mathbf{B}\mathbf{A}x. \quad (1)$$

During fine-tuning, $\mathbf{A}$ is initialized using a normal distribution and $\mathbf{B}$ is initialized as a zero matrix,

ensuring $\Delta \mathbf{W} = 0$ initially.

3.2 LAK-LoRA

Overview. Building on the MoE framework, LAK-LoRA enhances LoRA for medical multi-task learning by introducing layer-wise adaptive expert allocation and principal component-based feature separation. As illustrated in Figure 2, each adaptation layer begins with a feature separator module that separates the hidden states into two distinct representations: a shared feature subspace x_s and a task-specific subspace x_t . These disentangled representations are then routed to a layer-wise adaptive MoE module, where expert capacity varies with layer depth. This dual-component design promotes shared feature learning for better generalization while mitigating inter-task negative transfer through dedicated specialization pathways. Additionally, to encourage diversity among specialized experts, we include an orthogonality-promoting term in the training objective.

Mixture-of-Experts Architecture. Each adaptation layer in our framework includes a set of expert networks to learn $\Delta \mathbf{W}$, with each expert implemented as a LoRA module comprising low-rank matrices $\mathbf{A}_i \in \mathbb{R}^{\frac{r}{N_l} \times k}$ and $\mathbf{B}_i \in \mathbb{R}^{d \times \frac{r}{N_l}}$. For an adaptation layer with N_l experts, the total rank r is evenly distributed so the total number of trainable parameters matches that of a standard LoRA module, ensuring a fair comparison in our experiments.

At layer l , the MoE module comprises a generic expert $E_s^{(l)}$ and specialized experts $\{E_{t,i}^{(l)}\}_{i=1}^{N_l-1}$. To prevent interference, we employ deterministic routing: the shared representation x_s is processed solely by the generic expert, yielding the shared update component h_s , while the task-specific representation x_t is processed by specialized experts, yielding the task-specific update component h_t , per Eq. 2 and Eq. 3:

$$h_s = E_s^{(l)}(x_s) = \mathbf{B}_s^{(l)} \mathbf{A}_s^{(l)} x_s. \quad (2)$$

$$\begin{aligned} h_t &= \sum_{i=1}^{N_l-1} g(e_{task}) \cdot E_{t,i}^{(l)}(x_t) \\ &= \sum_{i=1}^{N_l-1} g(e_{task}) \cdot (\mathbf{B}_{t,i}^{(l)} \mathbf{A}_{t,i}^{(l)} x_t). \end{aligned} \quad (3)$$

For dynamic weighting of specialized experts, a task embedding $\mathbf{e}_{task} \in \mathbb{R}^{d_e}$ is trained for each task ID. This embedding is input to the task-driven gate function $g(\cdot)$, which comprises a linear projection

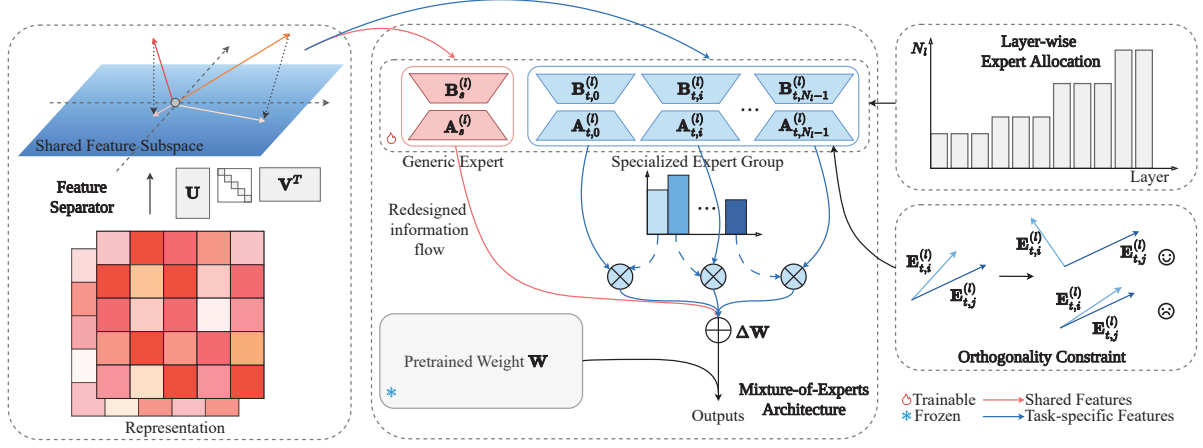


Figure 2: Overview of LAK-LoRA. Each adaptation layer first separates input representations into shared (red) and task-specific (blue) components, processed by their respective dedicated experts. Expert capacity increases with layer depth, and orthogonality constraints are imposed to encourage diversity.

followed by a temperature-scaled softmax:

$$g(\mathbf{e}_{task}) = \text{softmax}(\mathbf{W}_g \mathbf{e}_{task} / \tau), \quad (4)$$

where $\mathbf{W}_g \in \mathbb{R}^{(N_l-1) \times d_e}$ is the learnable gate matrix, and $\tau > 0$ is a temperature hyperparameter controlling the sharpness of the distribution. The final output $\Delta \mathbf{W} \mathbf{x}$ is the sum from both independent paths: $\Delta \mathbf{W} \mathbf{x} = h_s + h_t$.

Layer-Wise Adaptive Expert Allocation. Rather than using a fixed heuristic, the number of experts at layer l , N_l , is treated as a dynamic parameter that adapts to each layer’s role. Specifically, we define a layer-wise adaptation factor α_l that follows a sigmoid growth curve, facilitating a smooth transition from few experts in lower layers to more in higher layers (Liao et al., 2025):

$$\alpha_l = \text{sigmoid} \left(-\epsilon + \frac{2\epsilon \cdot l}{L} \right), \quad (5)$$

where ϵ controls the curve’s steepness and range, and L denotes the total number of layers in the model. We then map α_l to an index in a monotonically increasing sequence of positive integers $\mathbb{V}$, from which N_l is selected: $N_l = \mathbb{V}[\text{round}(\alpha_l \cdot (|\mathbb{V}| - 1))]$.

Principal Component-Based Feature Separation. We disentangle shared features from task-specific components using the statistical structure of the hidden representations.

Theoretical Grounding. Representation learning studies confirm that spectral decomposition isolates essential attributes. Specifically, SVD has been shown to extract high-variance essential knowledge for transfer (Lee and Song, 2022) and decouple safety vectors from general capabilities (Zhou

et al., 2025). Building on this, we posit that shared medical semantics occupy high-variance subspaces, while task-specific features reside in residuals. Assuming shared covariance dominance (analysis in Appendix C.1), extracting the principal subspace effectively recovers shared features.

In practice, we perform inference on n samples from the training set at the target adaptation layer, collecting the d -dimensional hidden states and stacking them column-wise to form the activation matrix $\mathbf{Z} = [z^{(1)} \ z^{(2)} \ \dots \ z^{(n)}] \in \mathbb{R}^{d \times n}$, where the j -th column corresponds to the j -th sample’s activation and the i -th row represents the i -th hidden dimension. To mitigate scale discrepancies, we normalize $\mathbf{Z}$: $\tilde{z}_{ij} = \frac{z_{ij} - \mu_i}{\delta_i}$, where μ_i and δ_i are the mean and standard deviation of dimension i over all samples.

We perform SVD on the normalized matrix: $\tilde{\mathbf{Z}} \approx \hat{\mathbf{U}} \hat{\mathbf{\Lambda}} \hat{\mathbf{V}}^T$, where $\hat{\mathbf{\Lambda}} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ contains the singular values, and $\hat{\mathbf{U}} \in \mathbb{R}^{d \times d}$ the left singular vectors, forming an orthonormal basis. To determine the shared subspace’s dimensionality, we use a singular value entropy criterion based on the magnitude and distribution of squared singular values, with information contribution $p_i = \sigma_i^2 / \sum_j \sigma_j^2$.

We identify the smallest integer r_s such that the ratio of cumulative entropy exceeds a threshold η :

$$\frac{H_{r_s}}{H_n} \geq \eta, \quad \text{where } H_{r_s} = - \sum_{i=1}^{r_s} p_i \log p_i. \quad (6)$$

Using the top- r_s left singular vectors $\hat{\mathbf{U}}_s$, we construct the orthogonal projection matrix onto the shared subspace: $\mathbf{P}_s = \hat{\mathbf{U}}_s \hat{\mathbf{U}}_s^T$. For any input

representation x , we decompose:

$$x_s = \mathbf{P}_s x, \quad (7)$$

$$x_t = (\mathbf{I} - \mathbf{P}_s)x = x - x_s, \quad (8)$$

where x_s is the shared feature component and x_t is the task-specific component.

Objective Function. Our training objective comprises two parts: standard cross-entropy loss $\mathcal{L}_{\text{task}}$ and orthogonality loss $\mathcal{L}_{\text{ortho}}$ applied to the specialized experts. The overall objective $\mathcal{L}_{\text{total}}$ is a weighted sum:

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

where λ is a hyperparameter. $\mathcal{L}_{\text{task}}$ is the mean cross-entropy loss over a training batch of m samples:

$$\mathcal{L}_{\text{task}} = \frac{1}{m} \sum_{j=1}^m \mathcal{L}_{\text{CE}}(y_j, \hat{y}_j), \quad (10)$$

with y_j and $\hat{y}_j$ as the ground-truth label and model’s prediction for sample j , respectively.

$\mathcal{L}_{\text{ortho}}$ mitigates redundancy among specialized experts. Let v_i and v_j be the flattened vector representations of experts $E_{t,i}^{(l)}$ and $E_{t,j}^{(l)}$ ($i \neq j$). The per-layer orthogonality loss is the mean squared cosine similarity over all unique pairs:

$$\mathcal{L}_{\text{ortho}}^{(l)} = \frac{1}{\binom{N_l-1}{2}} \sum_{i < j} \left(\frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \right)^2. \quad (11)$$

The final orthogonality loss is the average across adaptation layers containing multiple specialized experts.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on the Chinese medical PromptCBLUE benchmark (Zhu et al., 2023), which is one of the most widely adopted and standardized benchmarks for the multi-task evaluation of medical LLMs. Detailed statistics are provided in Table 1.

To further validate generalization, we additionally evaluate on several other medical datasets, including CMB-Exam (Wang et al., 2024), CMExam (Liu et al., 2023a), Yidu-N7K (Jiao, 2021), CMedCausal (Li et al., 2023), Chinese medical dialogue data¹, MedQA (Jin et al., 2020) and MedMCQA (Pal et al., 2022). Further details for these datasets are provided in Appendix A.1.

Task Type	Task Name	# Train	# Dev	# Test
Information Extraction	CMeEE-V2	4600	400	400
	CMeIE-V2	4600	400	400
	CHIP-CDEE	2600	400	400
	CHIP-CDN	4600	400	400
	IMCS-V2-NER	4600	400	400
Medical Text Classification	CHIP-CTC	5896	704	704
	KUAKE-QIC	5060	440	440
	IMCS-V2-DAC	4488	512	512
Natural Language Inference	CHIP-STs	4600	400	400
	KUAKE-QQR	4600	400	400
	KUAKE-IR	4600	400	400
	KUAKE-QTR	4600	400	400
Symptom Status	IMCS-V2-SR	4600	400	400
Understanding	CHIP-MDCFNPC	4600	400	400
Medical Content Generation	IMCS-V2-MRG	2549	400	400
	MedDG	4600	400	400

Table 1: Basic statistical information of PromptCBLUE.

Baselines. We conduct experiments on Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct (Qwen et al., 2024; Yang et al., 2024). Our selected baselines span diverse branches of multi-task PEFT: LoRA (Single) (Hu et al., 2022) trains a separate adapter for each task; LoRA (Full) (Hu et al., 2022) employs a single shared LoRA adapter across all tasks; MultiLoRA (Wang et al., 2023) extends LoRA with parallel adapters to capture task diversity; MOELoRA (Liu et al., 2023b) integrates a mixture-of-experts architecture with a task-driven gating function; and MTL-LoRA (Yang et al., 2025) leverages diagonal matrices to isolate task-specific information. We further include domain-specialized medical foundation models, including Qwen-Max (Qwen et al., 2024), MING-MoE (Liao et al., 2024), and Baichuan-M1 (Wang et al., 2025). These models are evaluated using in-context learning to ensure that their outputs follow the required evaluation format. Because they are not adapted using the same training protocol as the PEFT methods, their results are included only as contextual reference points rather than protocol-matched PEFT comparisons. Additionally, we evaluate ChatGLM-6B (GLM et al., 2024) to examine architectural robustness.

Evaluation Metrics. For the main experiments, we adopt task-specific evaluation metrics consistent with Zhu et al. (2023). Specifically, we use micro-F1 for medical information extraction, medical natural language inference, and dialogue-based

¹<https://github.com/Toyhom/Chinese-medical-dialogue-data>

Task	Qwen Max	MING MoE	Baichuan M1	Qwen2.5-7B-Instruct					
				LoRA (S)	LoRA (F)	MultiLoRA	MOELoRA	MTL-LoRA	LAK-LoRA
CMcEE-V2	49.614	29.263	43.155	66.259	65.338	66.435	69.392	68.126	70.196
CMcIE-V2	29.552	4.527	18.004	37.740	41.720	39.350	<u>46.699</u>	42.552	48.012
CHIP-CDN	65.986	35.958	65.240	81.656	80.075	83.508	82.103	<u>84.565</u>	85.421
CHIP-CDEE	31.969	14.494	30.319	60.141	61.924	61.583	65.106	<u>65.284</u>	66.115
IMCS-V2-NER	52.813	43.923	51.228	86.172	85.494	86.919	<u>89.022</u>	87.136	90.894
CHIP-MDCFNPC	66.777	58.193	67.672	76.284	76.398	76.303	<u>78.270</u>	77.740	78.514
IMCS-V2-SR	45.830	19.811	32.734	68.825	<u>71.201</u>	69.764	69.898	62.681	73.005
IMCS-V2-DAC	61.847	16.025	35.612	72.438	67.095	73.071	69.963	<u>73.421</u>	73.955
IMCS-V2-MRG	26.159	28.720	33.262	<u>46.275</u>	44.622	45.608	45.717	46.406	45.846
CHIP-CTC	70.843	12.923	54.818	<u>87.314</u>	84.975	88.071	<u>89.399</u>	86.552	90.029
CHIP-STS	72.978	61.611	69.272	84.000	78.732	81.502	84.751	84.494	<u>84.500</u>
KUAKE-IR	62.315	33.076	58.370	85.721	83.783	85.992	83.268	<u>88.365</u>	89.916
KUAKE-QIC	49.758	43.436	55.848	86.759	88.256	<u>87.724</u>	78.100	<u>86.616</u>	85.917
KUAKE-QQR	57.135	41.634	64.954	<u>76.093</u>	71.628	72.610	74.706	77.703	74.620
KUAKE-QTR	40.482	34.821	41.796	<u>59.679</u>	56.012	59.098	<u>65.125</u>	64.047	66.502
MedDG	10.771	11.030	10.331	13.006	12.212	12.062	11.937	<u>12.214</u>	11.267
Avg. Score	49.677	30.590	45.789	68.023	66.841	68.100	68.966	<u>69.244</u>	70.919
# Trainable (%)	-	-	-	0.1323×16	0.1323	0.4048	0.1347	0.2344	0.1346

Table 2: Performance comparison on the PromptCBLUE benchmark. The “Avg. Score” row shows the average score across all 16 tasks. Bold font indicates the highest score, and underline indicates the second-highest score. “LoRA (S)” and “LoRA (F)” are abbreviations for LoRA (Single) and LoRA (Full), respectively.

symptom understanding; macro-F1 for medical text classification to ensure balanced evaluation across classes; and ROUGE-L (Lin, 2004) for medical content generation. To provide a holistic assessment, we report the average score across all tasks, following Zhu et al. (2023). In the generalization experiments, we evaluate using micro-F1 on CMed-Causal, Yidu-N7K, and CMB-Exam; macro-F1 on CMExam; ROUGE-L on the medical dialogue dataset, and accuracy on MedQA and MedMCQA.

4.2 Main Results

Table 2 compares LAK-LoRA with a range of competitive baselines. The findings are as follows:

- Even when trained on specialized medical data, models such as Qwen-Max, Baichuan-M1, and MING-MoE still lag behind LoRA-based variants and our method by a large margin on PromptCBLUE, underscoring the critical importance of explicit task adaptation for effective medical language understanding.
- LAK-LoRA achieves the best average performance across the benchmark suite, outperforming the strongest baseline, MTL-LoRA, by 1.675 points in average score, attributable to layer-wise adaptive expert capacity scaling and explicit feature separation.
- Conventional multi-task approaches face substantial challenges: LoRA (Full) underper-

forms LoRA (Single), illustrating the severity of conflicting optimization objectives in medical multi-task learning. Furthermore, MOELoRA trails LAK-LoRA by 1.953 points, highlighting the limitations of fixed routing and the absence of feature separation.

Notably, all methods score substantially lower on MedDG, as it is an open-ended dialogue generation task where ROUGE-L scores are limited by diversity and insensitivity to clinical correctness.

Generalization Examination. On Qwen2.5-3B-Instruct (Table 3), LAK-LoRA averages 64.88 on five randomly selected PromptCBLUE tasks, consistently outperforming all baselines and ranking first on four tasks. Additional experiments on five medical datasets show strong results, particularly on CMB-Exam, CMedCausal, and Yidu-N7K, further confirming our method’s generalization across diverse medical tasks. Furthermore, extended evaluations in Appendices B.3, B.4, and B.5 demonstrate that LAK-LoRA consistently outperforms alternative PEFT baselines while maintaining superior performance across knowledge-intensive benchmarks and diverse model architectures like ChatGLM (GLM et al., 2024).

4.3 Ablation Studies

We adopt Qwen2.5-7B-Instruct as the backbone and conduct ablation studies on LAK-LoRA using the PromptCBLUE benchmark, removing one core

Method	CHIP-CDEE	IMCS-V2-SR	IMCS-V2-DAC	IMCS-V2-MRG	KUAKE-QTR	Avg. Score
LoRA (S)	36.135	63.574	50.619	42.802	52.920	49.210
LoRA (F)	60.179	70.289	59.434	45.538	56.028	58.293
MultiLoRA	59.621	68.528	69.218	44.594	57.217	59.836
MOELoRA	61.559	69.871	70.373	44.516	64.380	62.140
MTL-LoRA	<u>63.392</u>	71.599	<u>74.083</u>	44.980	<u>65.073</u>	<u>63.825</u>
LAK-LoRA	64.132	70.178	76.021	45.882	68.173	64.877

Method	CMExam	CMB-Exam	CMedCausal	Med-Dialogue	Yidu-N7K	Avg. Score
LoRA (S)	<u>36.454</u>	<u>22.250</u>	34.831	22.643	64.209	36.077
LoRA (F)	29.634	20.750	58.937	22.972	56.086	37.676
MultiLoRA	33.541	19.750	56.097	23.383	63.658	39.286
MOELoRA	26.517	20.250	68.779	22.267	69.679	41.499
MTL-LoRA	38.589	21.500	66.823	<u>23.189</u>	62.028	<u>42.426</u>
LAK-LoRA	28.027	23.000	74.168	22.882	71.886	43.993

Table 3: Performance comparison of PEFT methods on Qwen2.5-3B-Instruct. The ‘‘Avg. Score’’ row shows the average score. Bold font indicates the highest score, and underline indicates the second-highest score. ‘‘LoRA (S)’’ and ‘‘LoRA (F)’’ are abbreviations for LoRA (Single) and LoRA (Full), respectively.

component at a time while holding all other settings constant. Results are summarized in Table 4. i) Replacing our layer-wise adaptive expert allocation with fixed expert counts (2, 8, and 16) leads to significant performance degradation. Notably, even a large expert pool (e.g., 16) underperforms the adaptive strategy, suggesting that allocating fewer experts in shallower layers is crucial for fostering shared feature learning. ii) Removing the feature separator (denoted as ‘‘w/o feature separation’’) results in a 0.965-point drop, confirming its effectiveness. In this setting, we use a shared adapter for the generic expert and separate adapters for specialized experts without the SVD projection. iii) Omitting the orthogonality constraint causes a 0.136-point decline, validating its role in promoting expert diversity.

Additional ablations, including a per-task ablation study across diverse medical task categories, as well as analyses of the expert growth curve and the number of generic experts, are provided in Appendices B.1 and B.2.

Setting	Avg. Score	Δ
LAK-LoRA	70.020	–
w/o layer-wise expert allocation		
experts = 2	69.126	-0.894
experts = 8	69.196	-0.824
experts = 16	69.627	-0.393
w/o feature separation	69.055	-0.965
w/o orthogonality constraint	69.884	-0.136

Table 4: Results of ablation studies on LAK-LoRA.

4.4 Hyperparameter Analysis

We investigate the impact of hyperparameters τ , η , λ and ϵ on LAK-LoRA’s performance. We use Qwen2.5-7B-Instruct as the backbone and conduct experiments on the PromptCBLUE benchmark. Results are presented in Figure 3. i) τ controls the routing sharpness, and the model’s performance reaches a good level when τ is between 0.1 and 0.3. This indicates that a moderately sharp routing strategy is beneficial. ii) η governs dimension allocation between shared and task-specific subspaces. Performance is highly sensitive to this hyperparameter. When η is too low, insufficient dimensions are allocated to the shared subspace; conversely, the task-specific subspace is overly compressed. The model performs well with η in the range from 0.3 to 0.4. iii) λ scales the orthogonality constraint among experts, with the model achieving optimal performance at $\lambda = 0.01$. However, excessive regularization impairs performance. iv) ϵ modulates the steepness and span of the layer-wise expert growth curve. Performance peaks at $\epsilon = 5$, which yields a balanced expert allocation across layers.

4.5 Overhead Analysis

Training efficiency is a critical evaluation criterion beyond model performance. Here, we compare the trainable parameter ratios of LAK-LoRA against several baselines, as reported in Table 2. LAK-LoRA incurs only 0.1346% trainable parameters, which is significantly lower than 0.4048% for MultiLoRA and 0.2344% for MTL-LoRA, and comparable to standard LoRA and MOELoRA. MultiLoRA suffers from the highest overhead

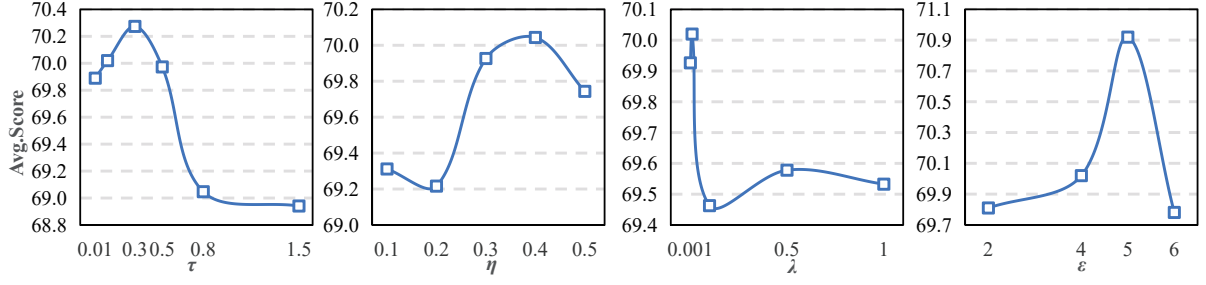


Figure 3: The performance of LAK-LoRA with different hyperparameter configurations.

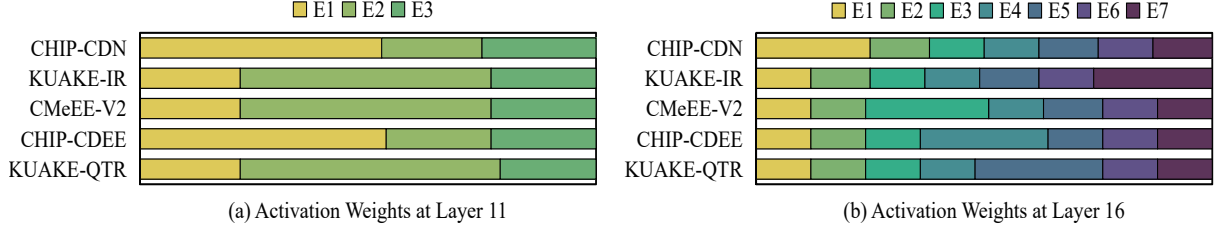


Figure 4: Visualization of specialized expert activation weights for different tasks in LAK-LoRA.

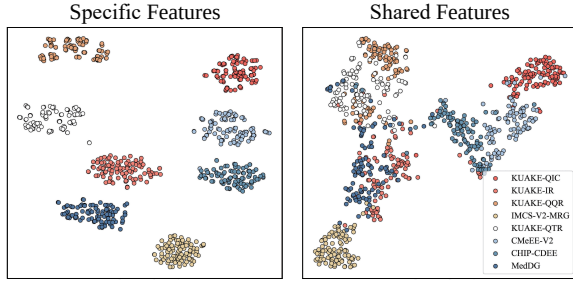


Figure 5: t-SNE visualization of both feature types.

due to maintaining multiple full adapters, while MTL-LoRA introduces multiple diagonal and up-projection matrices, leading to linear parameter growth with the number of tasks. In contrast, MOELoRA and LAK-LoRA increase training load only through expert routing. This highlights that our method achieves substantial performance gains with a negligible increase in parameter budget. Quantitative inference cost analyses and detailed theoretical complexity are presented in Appendices B.6 and C.2.

4.6 Visualization

We randomly select five tasks to visualize their specialized expert activation weights across layers in LAK-LoRA, as shown in Figure 4. This comparative visualization clearly reveals a shift in the division of labor among experts from coarse-grained to fine-grained as network depth and expert capacity increase. In Figure 4(a), with only three specialized experts available, distinct tasks (e.g.,

KUAKE-QTR, KUAKE-IR, and CMeEE-V2) activate similar combinations of experts, indicating that while task-specific differences begin to emerge, processing remains relatively coarse. In contrast, Figure 4(b) shows that tasks activate more distinctive expert ensembles. For example, KUAKE-QTR and KUAKE-IR diverge notably in the activation of expert 5 and expert 7. This demonstrates that, given greater expert capacity, LAK-LoRA enables more fine-grained and task-tailored specialization.

Additionally, we employ t-SNE to visualize the shared and task-specific feature subspaces across tasks in Figure 5. Tasks such as KUAKE-QQR and MedDG are closely intertwined in the shared feature subspace. Conversely, the representations in the task-specific subspace form well-separated clusters with clear boundaries. This confirms that our separator qualitatively isolates task-discriminative signals from the underlying shared features.

5 Conclusion

We propose LAK-LoRA, a novel MoE framework tailored for medical multi-task learning. It introduces layer-wise adaptive expert allocation and principal component-based feature separation, fundamentally redesigning the internal information flow. This enables adaptive allocation of expert capacity across layers, effectively mitigating negative inter-task transfer while promoting acquisition of shared features. Orthogonality constraints further encourage expert diversity among specialized experts. Extensive experiments on a 16-task medical

benchmark demonstrate that LAK-LoRA achieves the best average score among the compared methods, validating its effectiveness. Future work will extend our approach to medical multimodal tasks.

Limitations

Although LAK-LoRA achieves strong performance on various medical tasks, it has not yet been evaluated on larger-scale models due to computational resource constraints. Moreover, our current experiments are confined to text-only settings and do not extend to medical multimodal scenarios. We therefore aim to investigate the applicability of our architecture in multimodal contexts.

Ethics Statements

All datasets used in this work are publicly available and have been anonymized by their creators to exclude personally identifiable patient information. Although our approach targets a range of Chinese clinical language tasks, its misuse as an automated diagnostic tool could lead patients or non-specialist practitioners to over-rely on model outputs, potentially resulting in misdiagnosis or delayed treatment. We emphasize that the system is intended solely as a decision-support aid for qualified healthcare professionals and must not be used as a standalone diagnostic instrument.

Acknowledgments

This work was supported by the Guizhou Provincial Program on Commercialization of Scientific and Technological Achievements (Qiankehezhongyindi [2025] No. 006).

References

Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. 2025. [A comprehensive survey on the trustworthiness of large language models in healthcare](#). *Preprint*, arXiv:2502.15871.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [DISC-MedLLM: Bridging General Large Language Models and Real-World Medical Consultation](#). *Preprint*, arXiv:2308.14346.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. [HuatuoGPT-II, One-stage Training for Medical Adaptation of LLMs](#). *Preprint*, arXiv:2311.09774.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, et al. 2022. [Palm: Scaling language modeling with paths](#). *Preprint*, arXiv:2204.02311.

Michael Crawshaw. 2020. [Multi-Task Learning with Deep Neural Networks: A Survey](#). *Preprint*, arXiv:2009.09796.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. [DeepSeek LLM: Scaling Open-Source Language Models with Longtermism](#). *Preprint*, arXiv:2401.02954.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and V. S. Subrahmanian. 2024. [Higher Layers Need More LoRA Experts](#). *Preprint*, arXiv:2402.08562.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, et al. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.

Zengtao Jiao. 2021. [Sample of Yidu-N7K data set](#).

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Preprint*, arXiv:2009.13081.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024a. [PromptMRG: Diagnosis-driven prompts for medical report generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, and Yongfeng Zhang. 2024b. [Health-LLM: Personalized Retrieval-Augmented Disease Prediction System](#). *Preprint*, arXiv:2402.00746.

Seunghyun Lee and Byung Cheol Song. 2022. [Knowledge transfer via decomposing essential information in convolutional neural networks](#). *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):366–377.

Zihao Li, Mosha Chen, Kangping Yin, Yixuan Tong, Chuanqi Tan, Zhenzhen Lang, and Buzhou Tang. 2023. [Chip2022 shared task overview: Medical causal entity relationship extraction](#). In *Health Information Processing. Evaluation Track Papers*.

- Mengqi Liao, Wei Chen, Junfeng Shen, Shengnan Guo, and Huaiyu Wan. 2025. [HMoRA: Making LLMs more effective with hierarchical mixture of loRA experts](#). In *The Thirteenth International Conference on Learning Representations*.
- Yusheng Liao, Shuyang Jiang, Yu Wang, and Yanfeng Wang. 2024. [Ming-moe: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts](#). *Preprint*, arXiv:2404.09027.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, Barcelona, Spain. Association for Computational Linguistics.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Zhu Lei, and Michael Lingzhi Li. 2023a. [Benchmarking large language models on CMEExam - a comprehensive chinese medical exam dataset](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024a. [A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions](#). *Preprint*, arXiv:2406.03712.
- Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023b. [When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications](#). In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. [DoRA: Weight-decomposed low-rank adaptation](#). In *Forty-first International Conference on Machine Learning*.
- Yuguang Liu, Yiyun Miao, and Luyao Xia. 2025. [Direct routing gradient \(drggrad\): A personalized information surgery for multi-task learning \(mtl\) recommendations](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. [Peft: State-of-the-art parameter-efficient fine-tuning methods](#). <https://github.com/huggingface/peft>.
- Masoud Monajatipoor, Jiabin Yang, Joel Stremmel, Melika Emami, Fazlollah Mohaghegh, Mozhddeh Rouhsedaghat, and Kai-Wei Chang. 2024. [LLMs in biomedicine: A study on clinical named entity recognition](#). *Preprint*, arXiv:2404.07376.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. [Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, et al. 2024. [Qwen2.5: A party of foundation models](#).
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, et al. 2024. [Capabilities of Gemini Models in Medicine](#). *Preprint*, arXiv:2404.18416.
- Wenyu Shi and Vladimir Y. Mariano. 2025. [Enhanced intention recognition model based on text classification](#). *IEEE Access*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, et al. 2023. [Large language models encode clinical knowledge](#). *Nature*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, et al. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*.
- Hongda Sun, Jiaren Peng, Wenzhong Yang, Liang He, Bo Du, and Rui Yan. 2025. [Enhancing medical dialogue generation through knowledge refinement and dynamic prompt adjustment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, et al. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. 2024. [Hydralora: An asymmetric lora architecture for efficient fine-tuning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 9565–9584. Curran Associates, Inc.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.

- Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, et al. 2025. [Baichuan-ml: Pushing the medical capability of large language models](#). *Preprint*, arXiv:2502.12671.
- Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, et al. 2024. [CMB: A comprehensive medical benchmark in Chinese](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. 2023. [Multilora: Democratizing lora for better multi-task learning](#). *Preprint*, arXiv:2311.11501.
- Hermann Weyl. 1912. [Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen \(mit einer Anwendung auf die Theorie der Hohlraumstrahlung\)](#). *Mathematische Annalen*, 71(4):441–479.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.
- Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. 2025. [MTL-LoRA: Low-Rank Adaptation for Multi-Task Learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Yi Yu, Tengyao Wang, and Richard J. Samworth. 2014. [A useful variant of the davis–kahan theorem for statisticians](#). *Preprint*, arXiv:1405.0680.
- Heyi Zhang, Xin Wang, Zhaopeng Meng, Zhe Chen, Pengwei Zhuang, Yongzhe Jia, Dawei Xu, and Wenbin Guo. 2024. [Qibo: A large language model for traditional chinese medicine](#). *Preprint*, arXiv:2403.16056.
- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. [Adaptive budget allocation for parameter-efficient fine-tuning](#). In *The Eleventh International Conference on Learning Representations*.
- Yu Zhang and Qiang Yang. 2022. [A survey on multi-task learning](#). *IEEE Transactions on Knowledge and Data Engineering*.
- Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023b. [A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Guanghao Zhou, Panjia Qiu, Cen Chen, Hongyu Li, Jason Chu, Xin Zhang, and Jun Zhou. 2025. [LSSF: Safety alignment for large language models through low-rank safety subspace fusion](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30621–30638, Vienna, Austria. Association for Computational Linguistics.
- Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. [PromptCBLUE: A Chinese Prompt Tuning Benchmark for the Medical Domain](#). *Preprint*, arXiv:2310.14151.

A Experimental Details

A.1 Datasets and Benchmarks

All datasets and benchmarks used in our studies are publicly released resources that permit academic research. According to their documentation and publication venues, these datasets and benchmarks have undergone prior anonymization to remove or mask personally identifiable information such as patient names.

One of the generalization experiments reported in the main context is conducted using Qwen2.5-3B-Instruct across five medical datasets: CMExam, CMB-Exam, CMedCausal, Med-Dialogue and Yidu-N7K. Due to computational constraints, we apply the following sampling strategies for large-scale datasets: CMExam and Yidu-N7K are randomly subsampled; CMB-Exam is evaluated on a single subcategory selected from its 28 fine-grained categories to maintain manageable scale; Med-Dialogue is uniformly sampled across six clinical departments; and CMedCausal, being of moderate size, is used in its entirety. The data statistics for the aforementioned datasets are summarized in Table 5.

A systematic description of all datasets and benchmarks used in our experiments is provided below:

- **CMExam** (Liu et al., 2023a): It is a dataset derived from the Chinese National Medical Licensing Examination. It comprises over 60,000 multiple-choice questions, each enriched with five types of metadata: disease category, clinical department, medical discipline, competency domain, and question difficulty level. This structured annotation enables fine-grained analysis of model performance across diverse clinical and cognitive dimensions. It is released under the Apache-2.0 License.

- **CMB-Exam** (Wang et al., 2024): It is a hierarchically structured benchmark designed to evaluate large language models’ medical proficiency within the Chinese clinical and educational context. Rooted in Chinese national medical licensing and training frameworks, it encompasses 6 major categories and 28 fine-grained subcategories, spanning basic medicine, clinical specialties, public health, and traditional Chinese medicine (TCM). It is released under the Apache-2.0 License.
- **CMedCausal** (Li et al., 2023): It is a fine-grained causal relation extraction task targeting Chinese medical text, aiming to identify medical concept spans and annotate three types of semantic relations among them: causal, conditional, and hypernym–hyponym relations. It is accessed via the Tianchi competition platform ².
- **Med-Dialogue** ³: It is a large-scale dataset of real-world doctor–patient conversations, spanning six clinical specialties: andrology, internal medicine, obstetrics and gynecology, oncology, pediatrics, and surgery. It contains approximately 800,000 dialogue turns, making it suitable for research in medical dialogue understanding, response generation, and question answering. The dataset is available under the MIT License.
- **Yidu-N7K** (Jiao, 2021): It is a clinical term normalization dataset for Chinese medical text that maps highly heterogeneous and non-standard expressions of diagnoses, procedures, medications, examinations, laboratory tests, and symptoms to a unified set of standardized medical terminologies. It is distributed under the CC BY-NC 4.0 license.
- **MedQA** (Jin et al., 2020): It is a medical question-answering dataset in a multiple-choice format, with questions collected from medical board exams in the United States, Mainland China, and Taiwan. It is released under the MIT License.
- **MedMCQA** (Pal et al., 2022): It is a large-scale multiple-choice question-answering (MCQ) dataset designed to address real-world

medical entrance examination problems. It features a vast collection of high-quality MCQs sourced from the AIIMS and NEET PG entrance exams. The dataset is released under the MIT License.

- **PromptCBLUE** (Zhu et al., 2023): It is a comprehensive benchmark for multi-task learning in the Chinese medical domain, covering five core task categories: medical information extraction, text classification, natural language inference, symptom status understanding, and text generation. Information extraction involves named entity recognition from clinical texts and dialogues (CMEE-V2, IMCS-V2-NER), relation extraction (CMIE-V2), detection of positive and negative clinical findings (CHIP-CDEE), and mapping diagnostic descriptions to standard ICD-10 codes (CHIP-CDN). Text classification includes categorizing clinical trial eligibility criteria (CHIP-CTC) and identifying user or dialogue intentions (KUAKE-QIC, IMCS-V2-DAC). Natural language inference evaluates the model’s ability to assess semantic similarity (CHIP-STs) or relevance (KUAKE-QQR, KUAKE-IR, KUAKE-QTR) between medical text pairs. Symptom status understanding requires joint identification of clinical findings and their status (IMCS-V2-SR, CHIP-MDCFNPC). Finally, text generation tasks directly probe language capabilities, such as summarizing doctor–patient dialogues (IMCS-V2-MRG) or generating physician-like responses (MedDG). It is released under the Apache-2.0 License.

Task Name	# Train	# Dev	# Test
CMEExam	2000	400	400
CMB-Exam	2000	400	400
CMedCausal	2600	400	400
Med-Dialogue	3600	360	360
Yidu-N7K	4000	400	400
MedQA	3000	400	400
MedMCQA	3000	400	400

Table 5: Basic statistical information of seven medical datasets.

A.2 Implementation Details

All experiments are conducted on a server with four H800 GPUs. The rank r is set to 16, while

²<https://tianchi.aliyun.com/competition>

³<https://github.com/Toyhom/Chinese-medical-dialogue-data>

Setting	CHIP-CDEE	CHIP-MDCFNPC	IMCS-V2-MRG	KUAKE-QIC	KUAKE-QQR
LAK-LoRA	65.482	78.714	46.748	86.600	77.266
w/o layer-wise expert allocation					
experts = 2	61.986	76.616	45.977	86.189	74.631
experts = 8	64.319	78.075	45.088	79.430	75.469
experts = 16	66.652	78.119	45.963	87.671	76.744
w/o feature separation	63.661	77.775	44.634	87.382	71.250
w/o orthogonality constraint	62.948	77.310	45.736	83.734	75.304

Table 6: Performance impact of individual components across diverse medical task categories.

the dropout rate is fixed at 0.1 to prevent overfitting. We designate the q_proj , k_proj , v_proj , and o_proj layers as the trainable LoRA target modules, focusing adaptation efforts on these projection layers.

In all experiments based on Qwen2.5-7B-Instruct, we uniformly employ the AdamW optimizer with a weight decay of 0, coupled with a linear learning rate decay strategy. The initial learning rate is fixed at 2×10^{-4} , and gradient accumulation is applied over 8 steps. Training epochs are configured as follows: LoRA (Single) is trained for 5 epochs; LoRA (Full), MultiLoRA, MOELoRA, and LAK-LoRA for 3 epochs each; and MTL-LoRA for 2 epochs. Specific architectural settings include:

- i) In the MultiLoRA experiments, we use 3 LoRA modules.
- ii) For the MOELoRA experiments, the number of experts is fixed at 8, with the task embedding dimension set to 64 for each task.
- iii) In the MTL-LoRA experiments, the number of diagonal matrices is fixed at 16, the number of upper projection matrices is 3, and the temperature coefficient is set to 0.01.
- iv) For the LAK-LoRA experiments, the values of τ , η , and λ are set to 0.1, 0.3, and 0.01, respectively.

To comprehensively evaluate generalization, we conduct two sets of experiments on Qwen2.5-3B-Instruct:

i) First group: Evaluated on 5 datasets randomly sampled from the PromptCBLUE benchmark. In this setting, the number of diagonal matrices in MTL-LoRA is adjusted to 5. LAK-LoRA, MTL-LoRA, MOELoRA, MultiLoRA, and LoRA (Full) are trained for 5 epochs, while LoRA (Single) is trained for 8 epochs.

ii) Second group: Conducted on 5 Chinese medical datasets outside the PromptCBLUE benchmark. All configurations mirror the first group except for

training epochs: MTL-LoRA for 2 epochs, MultiLoRA for 4 epochs, MOELoRA for 3 epochs, LAK-LoRA and LoRA (Full) for 5 epochs each, and LoRA (Single) for 8 epochs.

Notably, in the main experiments with Qwen2.5-7B-Instruct, ϵ in LAK-LoRA is set to 5. Due to computational constraints, ϵ is fixed at 4 in all ablation and comparative experiments. Additionally, the activation matrix $\mathbf{Z}$ is constructed only once at the beginning of training using full-data SVD. For each training sample, we collect the hidden states from the target adaptation layer or module and mean-pool them across tokens to obtain a single representation vector, which forms one column of $\mathbf{Z}$. In the Qwen2.5-7B-Instruct experiments, $\mathbf{Z}$ is constructed using all 71,193 training samples to capture both shared and task-specific feature distributions. This one-time construction avoids repeatedly updating the SVD projections during optimization, thereby limiting the accumulation of routing bias across training rounds. Regarding computational overhead, constructing the SVD projections for Qwen2.5-3B-Instruct using 12,000 samples takes 443.92 seconds and reaches a peak GPU memory usage of 14,353 MiB.

To navigate the hyperparameter space efficiently, we adopt a coordinate-wise grid search strategy. Specifically, we establish a default configuration of $\tau = 0.1$, $\eta = 0.3$, $\lambda = 0.01$, and $\epsilon = 4$. During the search, we isolate and sweep one variable at a time while keeping the remaining parameters fixed at their default values.

For evaluation of text generation tasks, we compute ROUGE-L scores using `rouge_chinese`⁴ (version 1.0.3), a Python implementation of ROUGE tailored for Chinese text.

⁴https://github.com/Isaac-JL-Chen/rouge_chinese

B Extended Experimental Results

B.1 Per-Task Ablation Study

To further investigate the impact of each LAK-LoRA component, we provide a fine-grained ablation study on five representative high-conflict tasks: CHIP-CDEE (Information Extraction), CHIP-MDCFNPC (Symptom Understanding), IMCS-V2-MRG (Medical Content Generation), KUAKE-QIC (Medical Text Classification), and KUAKE-QQR (Natural Language Inference).

As shown in Table 6, key observations include:

i) Impact of adaptive expert allocation: Replacing our depth-adaptive strategy with fixed expert counts (e.g., experts = 2 or 8) leads to a performance drop across most tasks, particularly in CHIP-CDEE and KUAKE-QQR. This supports our hypothesis that allocating fewer experts in shallower layers and more in deeper layers is essential for balancing shared and task-specific learning. ii) Effectiveness of feature separation: The removal of the feature separator results in a notable performance decline across four tasks. Specifically, for the natural language inference task KUAKE-QQR, the score drops from 77.266 to 71.250, the most significant decline among all ablation items. This confirms that our principal component-based separation is critical for mitigating inter-task interference in conflicting task environments. iii) Orthogonality constraint: Omitting the orthogonality-promoting term also leads to degraded results, particularly in KUAKE-QIC and CHIP-CDEE. This validates the role of the constraint in encouraging specialized experts to capture diverse and complementary clinical representations.

B.2 Impact of Expert Growth Curves and the Number of Generic Experts

We further investigate the impact of expert growth curves and the number of generic experts. First, for the growth curve, we compare the sigmoid growth curve with two alternatives: the linear growth curve, where expert capacity increases uniformly with depth, and the exponential growth curve, where expert capacity grows slowly in shallow layers and ramps up sharply in deep ones. Second, we systematically vary the number of generic experts from one to five. As shown in Table 7, the results clearly indicate that the S-shaped growth curve and a single generic expert both yield optimal performance. The former validates that the S-shaped curve aligns well with the hierarchical nature of feature learning

in LLMs, while the latter indicates that one generic expert is sufficient to capture shared feature across tasks.

Setting	Avg. Score	Δ
Expert Growth Curve		
sigmoid growth	70.020	-
linear growth	69.585	-0.435
exponential growth	69.565	-0.455
Number of Generic Experts		
1 generic expert	70.020	-
3 generic experts	69.069	-0.951
5 generic experts	67.848	-2.172

Table 7: Comparative results of expert growth curves and the number of generic experts.

B.3 Robustness Against Alternative PEFT Baselines

To further evaluate the effectiveness of LAK-LoRA, we extend our comparative analysis to include other PEFT methods, specifically AdaLoRA (Zhang et al., 2023a) and DoRA (Liu et al., 2024b). For a fair and consistent comparison, all experiments are conducted using the Qwen2.5-3B-Instruct model under identical experimental settings. As reported in Table 8, LAK-LoRA achieves an average score of 43.993, significantly outperforming DoRA (34.482) and AdaLoRA (16.165). This performance advantage is not only reflected in the overall average but is also consistently observed across most individual tasks, demonstrating the robustness and superiority of our proposed method.

Notably, the performance gaps between LAK-LoRA and baselines are particularly substantial on CMedCausal and Yidu-N7K. This variance primarily stems from LAK-LoRA’s robust ability to maintain structured output formats. Unlike multiple-choice or open-ended tasks, these two tasks impose rigid formatting constraints: Yidu-N7K requires stripping redundant codes to output exact canonical terms, and CMedCausal relies on precise triplet structures where any minor deviation invalidates the answer. Baselines like AdaLoRA and DoRA frequently fail here because cross-task feature interference introduces structural noise. In contrast, LAK-LoRA guarantees high signal purity by effectively decoupling representations through its SVD-based feature separator. Furthermore, its layer-wise adaptive allocation provides expanded expert capacity in deeper layers to handle complex reasoning while strictly adhering to these format constraints.

Method	CMExam	CMB-Exam	CMedCausal	Med-Dialogue	Yidu-N7K	Avg. Score
AdaLoRA	19.705	<u>21.250</u>	15.036	20.509	4.327	16.165
DoRA	37.103	18.500	47.947	22.816	46.043	34.482
LAK-LoRA	<u>28.027</u>	23.000	74.168	22.882	71.886	43.993

Table 8: Performance comparison with other PEFT baselines using the Qwen2.5-3B-Instruct model. Bold and underlined values indicate the best and second-best results, respectively.

Among routing-based PEFT methods, we further compare LAK-LoRA with HydraLoRA (Tian et al., 2024), which adopts an asymmetric LoRA architecture with a shared **A** matrix and multiple routed **B** matrices. As shown in Table 9, LAK-LoRA outperforms HydraLoRA on all five tasks and improves the average score from 62.854 to 64.877. These results suggest that LAK-LoRA’s principal component-based separation more effectively decouples shared medical knowledge from task-specific features on this benchmark.

Task	HydraLoRA	LAK-LoRA
CHIP-CDEE	<u>60.912</u>	64.132
IMCS-V2-SR	<u>69.811</u>	70.178
IMCS-V2-DAC	<u>72.979</u>	76.021
IMCS-V2-MRG	<u>44.708</u>	45.882
KUAKE-QTR	<u>65.857</u>	68.173
Avg. Score	<u>62.854</u>	64.877

Table 9: Performance comparison between HydraLoRA and LAK-LoRA. Bold and underlined values indicate the best and second-best results, respectively.

B.4 Evaluation on Knowledge-Intensive Benchmarks

To evaluate the model’s proficiency in handling specialized domain knowledge, we conduct additional experiments on two representative medical question-answering datasets: MedQA and MedMCQA. Using the Qwen2.5-3B-Instruct backbone, we evaluate LAK-LoRA against the most competitive baselines identified in our primary study. As shown in Table 10, LAK-LoRA demonstrates superior performance in these knowledge-intensive scenarios, achieving the highest average score.

B.5 Cross-Architecture Generalizability

To evaluate the generalizability of our proposed method across different model architectures, we further conduct experiments on ChatGLM-6B (GLM et al., 2024). As shown in Table 11, we provide the performance comparison across five diverse medical tasks. Our proposed LAK-

Method	MedQA	MedMCQA	Avg.Score
LoRA (Full)	77.500	43.750	60.625
MOELoRA	79.250	45.500	<u>62.375</u>
MTL-LoRA	76.500	<u>46.250</u>	61.375
LAK-LoRA	<u>78.500</u>	49.750	64.125

Table 10: Performance comparison on knowledge-intensive medical benchmarks using the Qwen2.5-3B-Instruct backbone. Bold and underlined values indicate the best and second-best results, respectively.

LoRA consistently outperforms all baseline methods, achieving the highest average score of 59.448. Specifically, LAK-LoRA achieves significant performance gains on tasks such as KUAKE-QTR and CHIP-CDEE compared to the standard LoRA (Full).

To provide further architectural validation on diverse contemporary models, we have conducted supplementary experiments on Qwen3-4B-Instruct-2507 (Team, 2025), the latest iteration of the Qwen series. As detailed in Table 12, LAK-LoRA seamlessly transfers to this updated backbone, achieving the highest average score of 65.393 and consistently maintaining its leading performance over strong baselines.

Furthermore, we conduct comprehensive performance tests for Qwen2.5-3B-Instruct across all 16 PromptCBLUE tasks. As presented in Table 13, LAK-LoRA maintains a significant performance lead even at the 3B parameter scale. It achieves the highest overall average score of 64.790, decisively outperforming all existing multi-task PEFT baselines.

B.6 Computational Efficiency Analysis

We further evaluate the practical computational overhead of LAK-LoRA by measuring the average inference time per sample on the MedQA and MedMCQA datasets. As summarized in Table 14, LAK-LoRA exhibits a highly competitive inference profile compared to established LoRA-based baselines. LAK-LoRA requires 1.51 seconds per sample, which represents only a marginal increase

Method	CHIP-CDEE	IMCS-V2-SR	IMCS-V2-DAC	IMCS-V2-MRG	KUAKE-QTR	Avg. Score
LoRA (S)	23.903	55.668	62.314	47.209	39.540	45.727
LoRA (F)	45.011	50.677	51.572	47.310	46.520	48.218
MultiLoRA	56.440	56.054	60.100	47.253	46.714	53.312
MOELoRA	<u>60.136</u>	<u>57.581</u>	68.785	45.226	60.137	58.373
MTL-LoRA	59.941	57.388	<u>69.427</u>	44.133	<u>61.191</u>	<u>58.416</u>
LAK-LoRA	60.438	58.092	69.530	<u>46.869</u>	62.313	59.448

Table 11: Performance comparison on the ChatGLM architecture across five medical tasks. Bold font indicates the highest score, and underline indicates the second-highest score. “LoRA (S)” and “LoRA (F)” are abbreviations for LoRA (Single) and LoRA (Full), respectively.

Method	CHIP-CDEE	IMCS-V2-SR	IMCS-V2-DAC	IMCS-V2-MRG	KUAKE-QTR	Avg. Score
LoRA (S)	57.080	68.586	69.197	<u>47.180</u>	45.662	57.541
LoRA (F)	64.961	69.134	66.259	46.121	56.152	60.525
MultiLoRA	60.523	65.940	70.776	46.917	61.087	61.048
MOELoRA	64.518	<u>72.866</u>	74.051	46.316	57.616	63.073
MTL-LoRA	<u>64.539</u>	71.392	76.218	46.810	<u>64.159</u>	<u>64.624</u>
LAK-LoRA	63.244	74.491	<u>77.264</u>	47.422	64.545	65.393

Table 12: Performance on selected PromptCBLUE tasks using the latest Qwen3-4B-Instruct-2507 architecture. Bold and underlined values indicate the best and second-best results, respectively.

in latency compared to simpler architectures like LoRA (Full).

C Theoretical Analysis

C.1 Theoretical Justification

C.1.1 Mathematical Framework

Let $z \in \mathbb{R}^d$ represent the statistical realization of the hidden representation x at a target adaptation layer. We model z as a composition of a shared component s , a task-specific component t_τ for a specific task $\tau \in \{1, \dots, N\}$, and a noise term ϵ :

$$z = s + t_\tau + \epsilon, \quad (12)$$

where $s \sim \mathcal{N}(0, \Sigma_s)$ represents universal medical features across all N tasks, $t_\tau \sim \mathcal{N}(0, \Sigma_{t,\tau})$ is unique to task τ , and $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is noise.

We assume that s , t_τ , and ϵ are mutually uncorrelated. While normalization layers (e.g., RMSNorm) do not strictly enforce zero-mean across all samples, the empirical mean $\mu = \frac{1}{n} \sum_{i=1}^n z^{(i)}$ is observed to be small in practice. As proved in Eq. 13, mean subtraction introduces only a rank-one correction to the covariance matrix, which does not significantly affect the leading eigenspace. Hence, for simplicity, we proceed with centered representations.

C.1.2 Stability Analysis

Empirical Global Covariance Matrix. During the feature separation phase, we perform inference

on n samples from the training set, and their corresponding d -dimensional hidden states are stacked column-wise to form the activation matrix $\mathbf{Z} \in \mathbb{R}^{d \times n}$. The centered empirical covariance matrix is $\Sigma_c = \frac{1}{n} \sum_{i=1}^n (z^{(i)} - \mu)(z^{(i)} - \mu)^\top$. We define the uncentered second moment $\Sigma_u = \frac{1}{n} \mathbf{Z} \mathbf{Z}^\top$. Then:

$$\begin{aligned} \Sigma_c &= \frac{1}{n} \sum_{i=1}^n (z^{(i)} z^{(i)\top} - z^{(i)} \mu^\top - \mu z^{(i)\top} + \mu \mu^\top) \\ &= \Sigma_u - \mu \mu^\top - \mu \mu^\top + \mu \mu^\top \\ &= \Sigma_u - \mu \mu^\top. \end{aligned} \quad (13)$$

Thus, $\Sigma_u = \Sigma_c + \mu \mu^\top$. Since $\|\mu\|_2^2$ is small in practice, we can approximate $\Sigma_u \approx \Sigma_c$.

Furthermore, given that s, t_τ, ϵ are mutually uncorrelated and zero-mean, all cross-covariance terms vanish (e.g., $E[s t_\tau^\top] = E[s] E[t_\tau^\top] = 0$), the global covariance matrix Σ can be decomposed as:

$$\begin{aligned} \Sigma &= E[zz^\top] \\ &= E[(s + t_\tau + \epsilon)(s + t_\tau + \epsilon)^\top] \\ &= E[ss^\top] + E[t_\tau t_\tau^\top] + E[\epsilon \epsilon^\top] \\ &= \Sigma_s + \Sigma_t + \sigma^2 \mathbf{I}, \end{aligned} \quad (14)$$

where $\Sigma_t = E_\tau[\Sigma_{t,\tau}]$ is the average task-specific covariance. We define the perturbation matrix as $\Delta = \Sigma_t + \sigma^2 \mathbf{I}$.

Spectral Dominance of Shared Covariance. We assume that the shared covariance Σ_s has rank $r_s \ll d$, and task-specific covariances $\Sigma_{t,\tau}$ are heterogeneous and weakly aligned across tasks.

Task	LoRA (F)	MultiLoRA	MOELoRA	MTL-LoRA	LAK-LoRA
CMeEE-V2	54.420	55.302	57.287	<u>60.103</u>	61.246
CMeIE-V2	33.563	32.217	36.979	<u>35.344</u>	38.462
CHIP-CDN	67.498	72.244	73.843	<u>76.879</u>	77.527
CHIP-CDEE	53.021	41.157	53.258	<u>56.352</u>	57.251
IMCS-V2-NER	75.662	77.615	84.394	83.551	84.410
CHIP-MDCFNPC	71.291	70.658	<u>75.032</u>	73.387	75.381
IMCS-V2-SR	68.013	69.185	<u>68.654</u>	67.124	67.468
IMCS-V2-DAC	50.182	54.410	<u>61.073</u>	57.673	65.183
IMCS-V2-MRG	44.133	43.934	44.191	<u>45.363</u>	45.822
CHIP-CTC	76.113	76.062	79.997	<u>83.118</u>	86.100
CHIP-STs	73.576	76.713	79.505	77.748	<u>78.504</u>
KUAKE-IR	75.797	80.231	83.475	<u>83.717</u>	86.723
KUAKE-QIC	74.376	77.535	<u>86.517</u>	87.012	72.189
KUAKE-QQR	66.012	69.871	<u>70.995</u>	69.906	71.122
KUAKE-QTR	48.886	52.661	<u>54.735</u>	53.877	57.710
MedDG	12.837	<u>12.285</u>	11.788	11.928	11.541
Avg. Score	59.086	60.130	63.858	<u>63.943</u>	64.790

Table 13: Full performance comparison on 16 PromptCBLUE tasks using Qwen2.5-3B-Instruct. Bold font indicates the highest score, and underline indicates the second-highest score.

Method	Inference Time (s/sample)
LoRA (Full)	1.37
MultiLoRA	1.43
MOELoRA	1.46
MTL-LoRA	1.44
LAK-LoRA	1.51

Table 14: Efficiency comparison of inference cost. All results are obtained using the Qwen2.5-3B-Instruct backbone and evaluated on the MedQA and MedMCQA datasets.

Eigenvalue Stability. To ensure that the leading eigenvalues of the observed covariance Σ truly represent the shared medical features Σ_s , we apply Weyl’s Inequality (Weyl, 1912). For the perturbed matrix $\Sigma = \Sigma_s + \Delta$, the k -th eigenvalue $\lambda_k(\Sigma)$ is bounded by:

$$\lambda_k(\Sigma_s) + \lambda_{\min}(\Delta) \leq \lambda_k(\Sigma) \leq \lambda_k(\Sigma_s) + \lambda_{\max}(\Delta). \quad (15)$$

Given Σ_t is positive semi-definite, we have $\lambda_{\min}(\Delta) = \lambda_{\min}(\Sigma_t + \sigma^2 \mathbf{I}) \geq \sigma^2$. The maximum eigenvalue can be expressed as:

$$\begin{aligned} \lambda_{\max}(\Delta) &= \lambda_{\max}(\Sigma_t + \sigma^2 \mathbf{I}) \\ &= \|\Sigma_t + \sigma^2 \mathbf{I}\|_2 \\ &\leq \|\Sigma_t\|_2 + \sigma^2. \end{aligned} \quad (16)$$

Substituting these precise bounds into Eq. 15, we obtain the finalized interval for the observed eigenvalues:

$$\lambda_k(\Sigma) \in [\lambda_k(\Sigma_s) + \sigma^2, \lambda_k(\Sigma_s) + \|\Sigma_t\|_2 + \sigma^2]. \quad (17)$$

As task-specific components are heterogeneous and misaligned across tasks, the total energy of task-specific interference $\|\Sigma_t\|_2$ remains relatively small compared to the dominant shared signal $\lambda_{r_s}(\Sigma_s)$. This ensures that the intervals of the top- r_s eigenvalues remain sufficiently separated from the noise floor. Effectively, the observed singular values "overlap" with the true shared singular values plus a bounded perturbation, allowing SVD to consistently identify the shared subspace even in the presence of task-specific noise.

Subspace Stability. While Eq. 17 ensures the stability of eigenvalues, we use the Davis-Kahan Theorem (Yu et al., 2014) to prove that the directions of the eigenvectors (the principal subspace) remain stable under task-specific perturbation.

Let $\mathbf{U}_s \in \mathbb{R}^{d \times r_s}$ be the subspace spanned by the first r_s eigenvectors of the pure shared covariance Σ_s , and let $\hat{\mathbf{U}}_s \in \mathbb{R}^{d \times r_s}$ be the observed principal subspace spanned by the first r_s eigenvectors of Σ . By the definition of eigendecomposition,

$$\begin{aligned} \Sigma \hat{\mathbf{U}}_s &= \Sigma [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_{r_s}] \\ &= [\Sigma \hat{u}_1, \Sigma \hat{u}_2, \dots, \Sigma \hat{u}_{r_s}] \\ &= [\hat{\lambda}_1 \hat{u}_1, \hat{\lambda}_2 \hat{u}_2, \dots, \hat{\lambda}_{r_s} \hat{u}_{r_s}] \\ &= [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_{r_s}] \begin{bmatrix} \hat{\lambda}_1 & 0 & \dots & 0 \\ 0 & \hat{\lambda}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\lambda}_{r_s} \end{bmatrix} \\ &= \hat{\mathbf{U}}_s \hat{\Lambda}_s. \end{aligned} \quad (18)$$

We define the residual matrix R as

$$\mathbf{R} = \Sigma_s \hat{\mathbf{U}}_s - \Sigma \hat{\mathbf{U}}_s = \Sigma_s \hat{\mathbf{U}}_s - \hat{\mathbf{U}}_s \hat{\Lambda}_s. \quad (19)$$

Substituting $\Sigma_s = \Sigma - \Delta$, we get

$$\begin{aligned} R &= (\Sigma - \Delta) \hat{\mathbf{U}}_s - \hat{\mathbf{U}}_s \hat{\Lambda}_s \\ &= (\Sigma \hat{\mathbf{U}}_s - \hat{\mathbf{U}}_s \hat{\Lambda}_s) - \Delta \hat{\mathbf{U}}_s \\ &= -\Delta \hat{\mathbf{U}}_s. \end{aligned} \quad (20)$$

Since $\hat{\mathbf{U}}_s$ is an orthonormal matrix ($\|\hat{\mathbf{U}}_s\|_2 = 1$), the energy of the residual is bounded by the perturbation:

$$\begin{aligned} \|R\|_F &\leq \|\Delta\|_F \|\hat{\mathbf{U}}_s\|_2 \\ &= \|\Delta\|_F \\ &= \|\Sigma_t + \sigma^2 \mathbf{I}\|_F. \end{aligned} \quad (21)$$

The canonical angle $\Theta(\hat{\mathbf{U}}_s, \mathbf{U}_s)$ between the observed and true shared subspaces is then bounded by:

$$\begin{aligned} \|\sin \Theta(\hat{\mathbf{U}}_s, \mathbf{U}_s)\|_F &\leq \frac{\|\mathbf{R}\|_F}{\delta} \\ &\leq \frac{\|\Sigma_t + \sigma^2 \mathbf{I}\|_F}{\lambda_{r_s}(\Sigma_s) - \lambda_{r_s+1}(\Sigma_s)}, \end{aligned} \quad (22)$$

where $\delta = \lambda_{r_s}(\Sigma_s) - \lambda_{r_s+1}(\Sigma_s)$ is the spectral gap. In medical multi-task learning, the persistence of universal clinical concepts ensures a large $\lambda_{r_s}(\Sigma_s)$, while the heterogeneity of tasks keeps the noise floor $\lambda_{r_s+1}(\Sigma_s)$ low, proving that the high-variance subspace identified by SVD is a stable proxy for shared medical features.

C.1.3 SVD Equivalence

The practical implementation of feature separation relies on the Singular Value Decomposition of the activation matrix. Let $\mathbf{Z} \in \mathbb{R}^{d \times n}$ be the centered activation matrix. Its SVD is given by:

$$\mathbf{Z} = \hat{\mathbf{U}} \hat{\Lambda} \hat{\mathbf{V}}^\top, \quad (23)$$

where $\hat{\mathbf{U}} \in \mathbb{R}^{d \times d}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{n \times n}$ are orthogonal matrices containing the left and right singular vectors, respectively, and $\hat{\Lambda} \in \mathbb{R}^{d \times n}$ contains the singular values. The empirical covariance matrix can then be expressed as:

$$\begin{aligned} \Sigma &= \frac{1}{n} \mathbf{Z} \mathbf{Z}^\top \\ &= \frac{1}{n} (\hat{\mathbf{U}} \hat{\Lambda} \hat{\mathbf{V}}^\top) (\hat{\mathbf{V}} \hat{\Lambda}^\top \hat{\mathbf{U}}^\top) \\ &= \hat{\mathbf{U}} \left(\frac{\hat{\Lambda}^2}{n} \right) \hat{\mathbf{U}}^\top. \end{aligned} \quad (24)$$

This demonstrates that the left singular vectors $\hat{\mathbf{U}}$ of the activation matrix $\mathbf{Z}$ are identical to the eigenvectors of the global covariance matrix Σ .

According to our subspace stability analysis, since the leading r_s eigenvectors of Σ are dominated by the shared clinical component s , performing SVD on $\mathbf{Z}$ and selecting the top r_s left singular vectors effectively identifies the shared feature subspace. Consequently, task-specific variations and noise contribute primarily to the lower singular values and their corresponding singular vectors.

Ultimately, this mathematical decoupling assumes the existence of a meaningful and shared subspace, a prerequisite that general-domain multi-task learning struggles to satisfy due to severe semantic heterogeneity. In open-domain scenarios, identical tokens frequently exhibit drastic semantic shifts; for instance, the token ‘‘Apple’’ in financial tasks (the corporation) shares no intrinsic semantic connection with the same token in recipe tasks (the fruit). Applying SVD to such disparate data manifolds mistakenly extracts superficial lexical coincidences as pseudo-shared features, inevitably exacerbating feature interference rather than mitigating it. In contrast, the extreme semantic homogeneity pervasive in medical texts provides the necessary statistical foundation for our principal component-based feature separation to thrive.

C.1.4 Verification of Decoupling

To justify the dual-branch architecture of LAK-LoRA, we prove that the projected shared features and task-specific features are statistically uncorrelated. The shared feature vector x_s and the residual task-specific feature vector x_t are defined in Eq. 7 and Eq. 8. To demonstrate that these two components represent independent information flows within the LAK-LoRA architecture, we derive their cross-covariance:

$$\begin{aligned} \text{Cov}(x_s, x_t) &= E[x_s x_t^\top] \\ &= E[\mathbf{P}_s x ((\mathbf{I} - \mathbf{P}_s) x)^\top] \\ &= E[\mathbf{P}_s x x^\top (\mathbf{I} - \mathbf{P}_s)^\top]. \end{aligned} \quad (25)$$

Since $\mathbf{P}_s = \hat{\mathbf{U}}_s \hat{\mathbf{U}}_s^\top$ is a symmetric projection matrix, $\mathbf{P}_s^\top = \mathbf{P}_s$. Furthermore, since $\hat{\mathbf{U}}_s$ is composed of the leading eigenvectors of the global covariance matrix Σ , the projection matrix $\mathbf{P}_s$ commutes with Σ (i.e., $\mathbf{P}_s \Sigma = \Sigma \mathbf{P}_s$). Using these

properties and the fact that $E[xx^\top] = \Sigma$, we have:

$$\begin{aligned}\text{Cov}(x_s, x_t) &= \mathbf{P}_s \Sigma (\mathbf{I} - \mathbf{P}_s) \\ &= \Sigma \mathbf{P}_s (\mathbf{I} - \mathbf{P}_s) \\ &= \Sigma (\mathbf{P}_s - \mathbf{P}_s^2).\end{aligned}\quad (26)$$

Because any projection matrix is idempotent ($\mathbf{P}_s^2 = \mathbf{P}_s$), the expression simplifies to:

$$\text{Cov}(x_s, x_t) = \Sigma (\mathbf{P}_s - \mathbf{P}_s) = \mathbf{0}. \quad (27)$$

This zero covariance proves that the shared features x_s and task-specific features x_t are statistically decoupled.

C.2 Complexity Analysis

To demonstrate the parameter efficiency of LAK-LoRA, we provide a rigorous breakdown of the trainable components.

Expert Parameterization per Layer. In each adaptation layer l , LAK-LoRA distributes the total rank r among N_l experts. This design ensures a fair comparison with standard LoRA.

- **Standard LoRA Baseline:** A standard LoRA module with rank r consists of two matrices: $\mathbf{A} \in \mathbb{R}^{r \times k}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$. Assuming $d = k$ for projection layers, the total trainable parameters per layer is:

$$P_{\text{LoRA}} = r \cdot k + d \cdot r. \quad (28)$$

- **LAK-LoRA Expert Group:** Each expert (one generic expert $E_s^{(l)}$ and $N_l - 1$ specialized experts $E_{t,i}^{(l)}$) is assigned a sub-rank of r/N_l . For each expert $i \in \{1, \dots, N_l\}$, the parameters are $\mathbf{A}_i^{(l)} \in \mathbb{R}^{\frac{r}{N_l} \times k}$ and $\mathbf{B}_i^{(l)} \in \mathbb{R}^{d \times \frac{r}{N_l}}$. The total expert parameters per layer is:

$$\begin{aligned}P_{\text{LAK-LoRA}}^{\text{experts}} &= \sum_{i=1}^{N_l} \left(\frac{r}{N_l} \cdot k + d \cdot \frac{r}{N_l} \right) \\ &= N_l \cdot \left(\frac{r \cdot k + d \cdot r}{N_l} \right) \\ &= r \cdot k + d \cdot r.\end{aligned}\quad (29)$$

Auxiliary Trainable Components. Beyond the low-rank adapters, LAK-LoRA introduces minimal auxiliary parameters for task-driven routing.

- **Task Embeddings:** For a multi-task setting with T tasks, we train a set of embeddings $\{e_{\text{task}, \tau}\}_{\tau=1}^T$, where each $e_{\text{task}} \in \mathbb{R}^{d_e}$. The total parameters are $T \cdot d_e$.

- **Gate Matrices:** Each layer l contains a learnable gate matrix $\mathbf{W}_g^{(l)} \in \mathbb{R}^{(N_l-1) \times d_e}$.

Across L layers, the total auxiliary parameters are:

$$P_{\text{LAK-LoRA}}^{\text{aux}} = T \cdot d_e + \sum_{l=1}^L (N_l - 1) \cdot d_e \quad (30)$$

D Visual Analysis of Expert Activation Patterns

We visualize the activation patterns when the number of experts is scaled to 16. Compared to lower-capacity settings, the activation map with 16 experts exhibits significantly higher sparsity and task-specific distinctiveness. As shown in Figure 6, expert 5 is predominantly activated by the medical normalization task (CHIP-CDN), whereas expert 6 and expert 13 show clear preference for medical natural language inference (KUAKE-QTR) and entity recognition (CMEE-V2), respectively. This shift from shared to disjoint expert ensembles confirms that LAK-LoRA’s automated allocation strategy successfully enables more fine-grained and task-tailored representations as capacity increases. The heatmap of expert activation across different layers can be found in Figure 7.

E Prompt Templates

To evaluate Qwen-Max (Qwen et al., 2024), MING-MOE (Liao et al., 2024) and Baichuan-M1 (Wang et al., 2025), we employ a few-shot in-context learning approach, designing task-specific prompts for the primary categories within the PromptCBLUE benchmark. These prompts address medical text classification and natural language inference (Table 15), symptom status understanding (Table 16), context generation (Table 17) and information extraction (Tables 18, 19, 20).

F Case Study

Table 21 presents four case studies to intuitively demonstrate the superiority of LAK-LoRA. Case 1 is a fine-grained semantic understanding task where both LoRA and MOELoRA fail to recognize that “dull ache” is a specific manifestation of “pain”. In contrast, LAK-LoRA correctly perceives this hyponymic relationship, identifying “dull ache” as a subconcept of “pain”.

Case 2 highlights a challenging named entity recognition scenario, where the core entity “rash” is modified by the negation “no” and the adverb

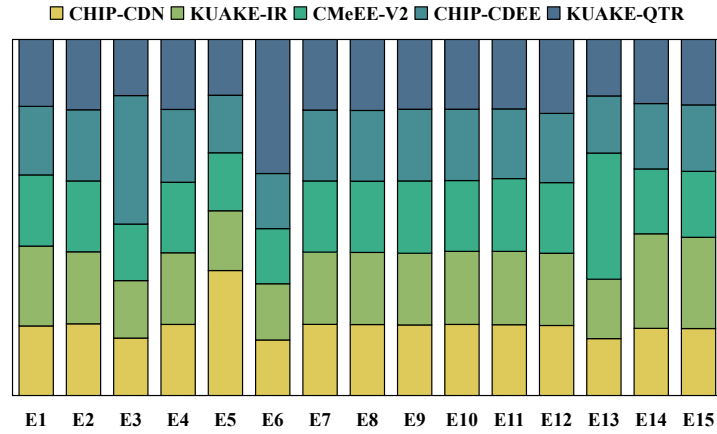


Figure 6: Visualization of specialized expert activation weights for different tasks in LAK-LoRA at Layer 28 ($N_t = 16$).

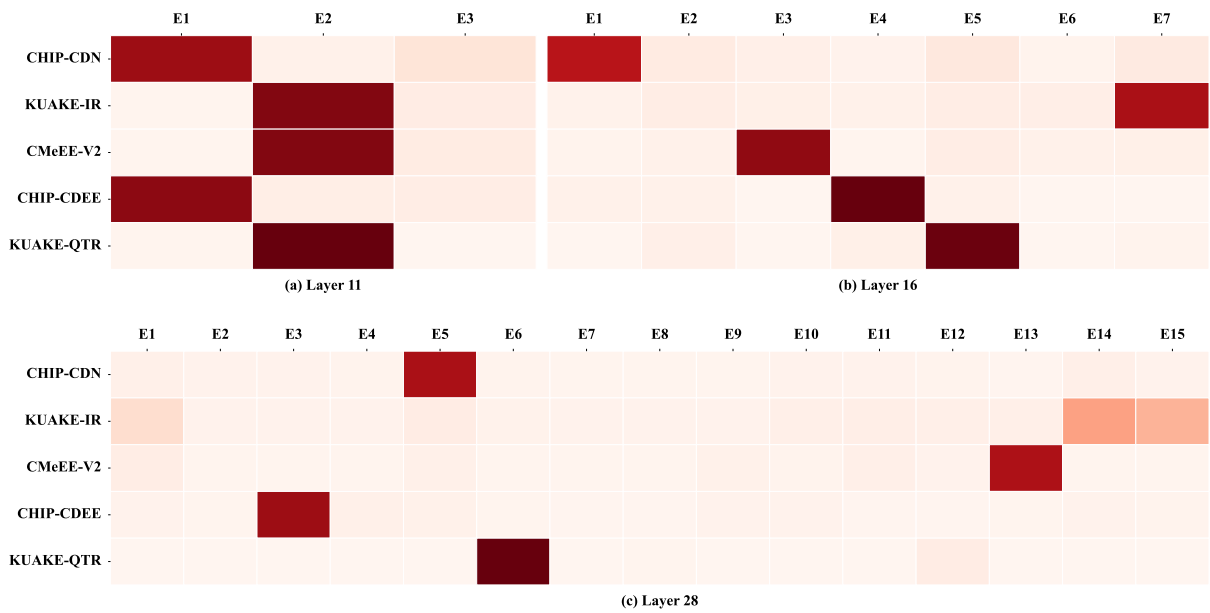


Figure 7: Heatmap of expert activation across different layers.

Please read the “Input” carefully and select the most appropriate answer from the “Options” list. You must and only output the text of the selected option itself.	
— Example Starts —	
# Input: {}	
# Options: {}	
# Output: {}	
— Example Ends —	
— Task Starts —	
# Input: {}	
# Options: {}	
# Output:	

Table 15: Prompt templates for medical text classification and natural language inference tasks.

Based on the context in the “Input”, please select a correct attribute for each mentioned entity from the “Attribute Options”. Strictly follow the example format for the output.	
— Example Starts —	
# Input: {}	
# Attribute Options: {}	
# Output: {}	
— Example Ends —	
— Task Starts —	
# Input: {}	
# Attribute Options: {}	
# Output:	

Table 16: Prompt templates for symptom status understanding tasks.

Please read the input carefully and generate the corresponding text as required. Strictly follow the output format of the examples. — Example Starts — # Input: {} # Output: {} — Example Ends — — Task Starts — # Input: {} # Output:

Table 17: Prompt templates for medical context generation tasks.

Please read the “Input” and select all matching standard terms from the “Candidate Standard Terms” list. Strictly follow the example format for the output, separating multiple standard terms with a comma. — Example Starts — # Input: {} # Candidate Standard Terms: {} # Output: {} — Example Ends — — Task Starts — # Input: {} # Candidate Standard Terms: {} # Output:
--

Table 20: Prompt templates for medical standard term extraction tasks.

From the “Input”, please extract all entities belonging to the “Entity Types” list. Strictly follow the example format for the output. — Example Starts — # Input: {} # Entity Types: {} # Output: {} — Example Ends — — Task Starts — # Input: {} # Entity Types: {} # Output:
--

Table 18: Prompt templates for medical named entity recognition tasks.

From the “Input”, please extract all structured information belonging to the “Relation/Event Types” list. Strictly follow the example format for the output. — Example Starts — # Input: {} # Relation/Event Types: {} # Output: {} — Example Ends — — Task Starts — # Input: {} # Relation/Event Types: {} # Output:
--

Table 19: Prompt templates for medical relation/event extraction tasks.

“consistently”. Both baseline models exhibit boundary detection errors: LoRA incorrectly includes the negation word, while MOELoRA errs more broadly by including both the adverb and the negation. LAK-LoRA, however, accurately defines the entity boundaries, remaining undisturbed by these complex contextual modifiers.

Case 3 tests the model’s ability to perform nuanced user intent classification. The query, “Will getting a tattoo cause AIDS?”, seeks to understand the potential consequences of an action. Both LoRA and MOELoRA fail to capture this specific intent, misclassifying it as a generic “Other Type”. In contrast, LAK-LoRA accurately discerns the user’s underlying concern, correctly identifying the query’s nature as a “Consequence Inquiry”.

Case 4 presents a challenging semantic equivalence problem that requires distinguishing between closely related medical concepts. The baseline models are misled by the topical overlap between “high blood sugar” and “diabetes”, incorrectly concluding that the two queries are equivalent. LAK-LoRA, however, can recognize that “high blood sugar” is a symptom while “diabetes” is a chronic disease, and thus correctly determines that the sentences are not semantically equivalent. This demonstrates its robustness against superficial keyword-based associations.

Case 1	
Input	Determine the semantic relationship between the following two queries: “Pain in the back of both sides of the waist.” and “What is causing the dull ache on both posterior sides of the waist?”
LoRA	No direct semantic correlation. ✗
MOELoRA	No direct semantic correlation. ✗
LAK-LoRA	The latter is a semantic subset of the former. ✓
Case 2	
Input	Identify medical entities in the following text: “Approximately 10%~20% of pediatric patients consistently show no rash symptoms.”
LoRA	Clinical Manifestation: no rash; ✗ Body Part: skin.
MOELoRA	Clinical Manifestation:consistently no rash; ✗ Body Part: skin.
LAK-LoRA	Clinical Manifestation: rash; ✓ Body Part: skin.
Case 3	
Input	Classify the user intent for the query: “Will getting a tattoo cause AIDS?”
LoRA	Other Type. ✗
MOELoRA	Other Type. ✗
LAK-LoRA	Consequence Inquiry. ✓
Case 4	
Input	Determine if the following two sentences are semantically equivalent: “What should I do about high blood sugar?” and “How can I recover from diabetes?”
LoRA	Yes. ✗
MOELoRA	Yes. ✗
LAK-LoRA	No. ✓

Table 21: Case studies on challenging examples.