HiMoLE: Towards OOD-Robust LoRA via Hierarchical Mixture of Experts

Yinuo Jiang[1,2],**Yan Xiaodong**[5], **Keyan Ding**[3], **Deng Zhao**[5],

Lei Liang[5],**Qiang Zhang**[2,4*], **Huajun Chen**[1,2*]

[1]College of Computer Science and Technology, Zhejiang University
[2]ZJU-Ant Group Joint Lab of Knowledge Graph
[3]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[4]ZJU-UIUC Institute, Zhejiang University
[5]Ant Group
{qiang.zhang.cs, huajunsir}@zju.edu.cn

Abstract

Parameter-efficient fine-tuning (PEFT) methods, such as LoRA, have enabled the efficient adaptation of large language models (LLMs) by updating only a small subset of parameters. However, their robustness under out-of-distribution (OOD) conditions remains insufficiently studied. In this paper, we identify the limitations of conventional LoRA in handling distributional shifts and propose **HiMoLE** (**Hi**erarchical **M**ixture of **L**oRA **E**xperts), a new framework designed to improve OOD generalization. HiMoLE integrates hierarchical expert modules and hierarchical routing strategies into the LoRA architecture and introduces a two-phase training procedure enhanced by a diversity-driven loss. This design mitigates negative transfer and promotes effective knowledge adaptation across diverse data distributions. We evaluate HiMoLE on three representative tasks in natural language processing. Experimental results evidence that HiMoLE consistently outperforms existing LoRA-based approaches, significantly reducing performance degradation on OOD data while improving in-distribution performance. Our work bridges the gap between parameter efficiency and distributional robustness, advancing the practical deployment of LLMs in real-world applications.

1 Introduction

Large language models (LLMs) have brought transformative advances across a wide range of domains. However, their unprecedented scale incurs substantial computational and storage costs. To address this issue, parameter-efficient fine-tuning (PEFT) techniques, such as LoRA [1], have emerged as practical solutions. By updating only a small subset of model parameters, PEFT methods reduce storage and computational requirements while achieving performance comparable to full-model fine-tuning. This efficiency makes them especially attractive for real-world deployments.

Despite these advantages, PEFT methods face a critical shortcoming: **limited generalization under out-of-distribution (OOD) conditions**. While deep learning models typically perform well on in-distribution (ID) data, their performance often degrades when faced with data that deviates from the training distribution [2, 3]. This issue persists even in LLMs after full fine-tuning and is particularly pronounced in domains characterized by high heterogeneity [4, 5], such as biomedicine and the

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

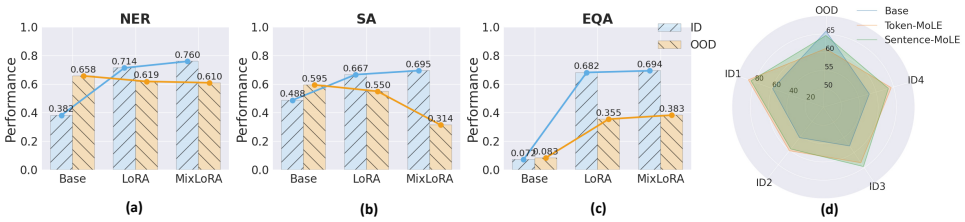


Figure 1: Robustness analysis of parameter-efficient fine-tuning. (a)–(c): In-distribution (ID) and out-of-distribution (OOD) results for the *Base*, *LoRA*, and *MixLoRA* models on three representative tasks: Named Entity Recognition (NER) in biomedicine, Sentiment Analysis (SA) in social science, and Extractive Question Answering (EQA) in general domain. (d): Impact of routing granularity in Mixture-of-LoRA-Experts, where "OOD" refers to an OOD validation dataset, while "ID1" "ID2" "ID3" and "ID4" represent the four ID validation datasets. Token-level routing yields better ID performance but fails to generalize to OOD data. In contrast, sentence-level routing improves OOD robustness at the cost of ID accuracy.

social sciences. Surprisingly, although LoRA and related PEFT methods have gained widespread adoption, their robustness to distributional shifts remains largely unexplored. Our empirical analysis (Fig. 1(a)-(c)) reveals that standard LoRA suffers considerable accuracy drops when applied to tasks requiring adaptation across diverse knowledge domains. These findings suggest intrinsic limitations in LoRA's ability to generalize beyond the training distribution, motivating the need for more robust PEFT strategies.

Mixture-of-parameter-efficient-expert (MoPE) methods attempt to improve generalization by integrating the Mixture-of-Experts (MoE) framework with PEFT, demonstrating effectiveness in multi-task settings [6, 7, 8, 9]. However, their effectiveness under OOD conditions remains underexplored. As shown in Fig. 1(a)(b), both LoRA and MixLoRA suffer notable performance degradation in knowledge-intensive tasks (e.g., Named Entity Recognition (NER) in biomedicine and Sentiment Analysis (SA) in social science) when evaluated on OOD data. As shown in Fig. 1(c), in general-domain tasks (e.g., Extractive Question Answering (EQA)), the gap between ID and OOD performance persists, revealing limited robustness. In some cases, MixLoRA even underperforms standard LoRA, suggesting potential overfitting. We identify a key source of this limitation: token-level routing in MoPE models is prone to expert misallocation under distributional shift. Local token-level features often fail to capture the global semantics necessary for robust generalization, resulting in brittle routing decisions in unseen contexts. These observations motivate the central question of our study: **How can parameter-efficient fine-tuning methods be improved to enhance in-distribution performance while also ensuring robustness to out-of-distribution data?**

To address this issue, we propose **HiMoLE** (**Hi**erarchical **M**ixture **o**f **L**oRA **E**xperts), a novel framework that introduces structural sparsity and hierarchical design into the LoRA architecture. HiMoLE extends conventional MoPE models via the hierarchical architecture which manifests in two dimensions: hierarchical expert design and hierarchical routing strategy. To further improve knowledge utilization and reduce redundancy, we introduce a two-phase training scheme augmented with a diversity-promoting loss. In summary, our work makes the following key contributions:

- **Empirical diagnosis of OOD limitations in LoRA.** We systematically investigate the generalization performance of LoRA under distributional shift, revealing significant weaknesses in its ability to transfer across heterogeneous domains.
- **A novel hierarchical MoPE framework.** We propose HiMoLE, which introduces hierarchical expert architectures and routing strategies into the PEFT paradigm. This structure mitigates negative transfer and promotes positive transfer, offering a new direction for improving PEFT robustness under distributional shift.
- **Theoretical and empirical validation.** We provide theoretical insights into the advantages of hierarchical routing under distributional shift. Experiments across multiple domains show HiMoLE improves OOD generalization while maintaining strong ID performance.

2 Background and Related Work

2.1 OOD Generalization

Out-of-distribution generalization is essential for deploying language models in real-world scenarios, where data distributions are inherently diverse, non-stationary, and unpredictable [10, 11]. This need is especially pronounced in high-stakes domains such as clinical decision support and social science analytics, where knowledge continuously evolves and data often deviate from training distributions. In such contexts, models must demonstrate robustness to emergent semantic patterns, novel entity relationships, and shifting contextual dependencies. Despite the remarkable progress of large language models across a wide range of tasks and benchmarks, recent studies [4, 12, 13] have revealed significant vulnerabilities under distributional shifts. These findings expose the limitations of current fine-tuning strategies and underscore the urgent need for methods explicitly designed to enhance OOD robustness.

2.2 Mixture of Parameter-efficient Experts

Parameter-Efficient Fine-Tuning (PEFT) As the scale of LLMs continues to grow, PEFT has emerged as a practical and cost-effective adaptation strategy. PEFT techniques update only a small subset of model parameters—such as adapter layers or low-rank matrices—while keeping the majority of the pre-trained model frozen [14, 15, 16]. A widely adopted PEFT method is LoRA [1], which inserts trainable low-rank adapters into pre-trained layers and updates them during fine-tuning. LoRA achieves competitive performance with a substantially reduced memory footprint. However, PEFT often struggles to generalize across new distributions. The limited number of trainable parameters can restrict the model's capacity to adapt to distributional shifts or novel task requirements.

Integrating MoE with PEFT (MoPE) To reconcile scalability and efficiency, recent work proposes integrating MoE with PEFT techniques, resulting in the MoPE paradigm. In MoPE, each expert is instantiated using a PEFT configuration (e.g., LoRA), and a routing module dynamically assigns inputs to appropriate experts. This hybrid design seeks to combine the modular adaptability of MoE with the resource efficiency of PEFT. MoPE methods differ in routing granularity. Token-level routing operates at the sub-sentence level. For example, MixLoRA [6] combines multiple LoRA experts with a shared FFN and incorporates an auxiliary load balancing loss to mitigate expert usage imbalance. LoRAMoE [17] employs a router network to reduce knowledge forgetting. HydraLoRA [9] adopts an asymmetric architecture with a shared LoRA A matrix and expert-specific B matrices. In contrast, sentence-level routing mechanisms operate at the input sentence level. MOELoRA [18] performs explicit task-to-expert assignment using task metadata, deterministically routing input sentences based on task identifiers. MOCLE [7] clusters instruction semantics and activates task-specific experts by assigning inputs to their corresponding instruction cluster. Although MoPE frameworks have demonstrated effectiveness in multi-task and in-distribution settings, their robustness under distributional shifts remains largely unexplored. This motivates the need for architectures specifically designed to handle complex domain shifts and enhance OOD generalization.

3 Method

3.1 Preliminaries

Formulation of OOD Generalization. Let x denote the input data and y the output. Out-of-distribution generalization refers to scenarios where the test distribution $P_{\text{test}}(x, y)$ differs from the training distribution $P_{\text{train}}(x, y)$, while preserving core semantic relationships. This work focuses on the joint occurrence of two distributional shifts [19]:

1. Covariate Shift: The input distribution changes ($P_{\text{test}}(x) \neq P_{\text{train}}(x)$), but the conditional distribution remains invariant($P_{\text{test}}(y \mid x) = P_{\text{train}}(y \mid x)$).
2. Concept Shift: The input-conditional distribution changes($P_{\text{test}}(y \mid x) \neq P_{\text{train}}(y \mid x)$), which may arise from label semantics or task definitions evolving across domains.

Robust generalization in this setting requires models to: (1) handle divergent input distributions (covariate shift), and (2) adapt to latent conceptual variations (concept shift).

Identification of OOD Generalization Problem in PEFT Following the protocol established by [4], we consider three key criteria for identifying OOD data: (1) diverse data sources, (2) low SimCSE similarity [20] with the in-distribution dataset, and (3) measurable performance degradation in models. However, in practice, we find that the second criterion—low SimCSE similarity—is not always reliable, especially in knowledge-dense domains. Fine-tuned LLMs can still generalize effectively even when SimCSE scores are low. As such, we primarily rely on criteria (1) and (3) in our OOD dataset selection. Here, performance degradation refers to reduced model performance on OOD data relative to ID data, which can manifest in two ways: (1) the model performs worse than its pre-fine-tuned counterpart, and (2) the performance gain on OOD data is substantially smaller than that on ID data. We fine-tune models using LoRA on each task's ID dataset and evaluate them on both ID and OOD test sets. As shown in Fig. 1(a)(b)(c), LoRA fails to significantly enhance OOD robustness. To further illustrate this, we analyze failure cases in a biomedical NER task. The fine-tuned model demonstrates two main types of errors in OOD data: (1) failure to produce outputs in the correct structured format, and (2) mislabeling of general-domain entities—for instance, tagging symptoms as diseases (see Appendix A for detailed case studies).

Probing OOD Robustness in MoPE Prior research [21, 22] on neural network optimization has shown that sparsely activated architectures often generalize better than dense ones, owing to dynamic parameter specialization and reduced task interference. In LLMs, the MoE framework embodies this principle by leveraging conditional computation through expert routing, enabling scalable and efficient learning. Building on these insights, we investigate the OOD robustness of various MoPE configurations, focusing on token-level and sentence-level routing strategies. As illustrated in Fig. 1(d), our results highlight a key trade-off between routing granularities: **Token-level routing** attains state-of-the-art performance on ID data by exploiting fine-grained contextual cues but exhibits pronounced OOD performance degradation, suggesting overfitting to surface-level patterns. **Sentence-level routing** offers improved OOD robustness by aligning with global semantics but suffers from reduced ID performance due to its limited sensitivity to local details. This finding highlights the necessity of developing a unified approach that can effectively integrate both routing granularities, achieving high ID accuracy while maintaining strong generalization across OOD scenarios.

3.2 Architecture of HiMoLE

In this subsection, we introduce HiMoLE, a Hierarchical Mixture of LoRA Experts model designed to flexibly address the OOD robustness challenges inherent in LoRA-based fine-tuning. An overview of the HiMoLE architecture is shown in Fig. 2(a).

Hierarchical Experts Our parameter-efficient expert architecture is organized into N Knowledge Competition Groups (KCGs), each consisting of M Knowledge Collaboration Experts (KCEs). Since domains with heterogeneous knowledge comprise multiple distinct subdomains, we assign each KCG to a unique subdomain and initialize its parameters by fine-tuning on the corresponding sub-dataset (as detailed in Section 3.3, Training Strategy). All KCEs within a KCG share the same initialization, providing a consistent foundation of subdomain-specific knowledge. Each KCE is implemented as a LoRA module, formally defined as:

$$E = BA, A \in \mathbb{R}^{d_{\text{in}} \times r}, B \in \mathbb{R}^{r \times d_{\text{out}}}, r \ll \min(d_{\text{in}}, d_{\text{out}}). \tag{1}$$

The experts interact through three distinct modes. (1) Intra-group collaboration: KCEs within the same KCG specialize in subdomain-specific patterns while leveraging shared knowledge, enabling efficient adaptation and positive transfer. (2) Cross-group competition: Different KCGs compete to route tokens to the most relevant KCEs, thereby reducing interference across subdomains and mitigating negative transfer. (3) Cross-group collaboration: KCEs across different KCGs may cooperate to improve generalization, promoting knowledge reuse and transferability. This structured interplay between competition and collaboration ensures both specialization and synergistic learning across knowledge boundaries.

Hierarchical Routing Strategy The core of HiMoLE lies in its hierarchical routing strategy, which integrates sentence-level and token-level expert selection to achieve domain-aware and adaptive inference. For a given input sentence, we first compute a sentence-level representation denoted by h_{sen} via applying average pooling over the token-level hidden states h_{token}. This pooled representation

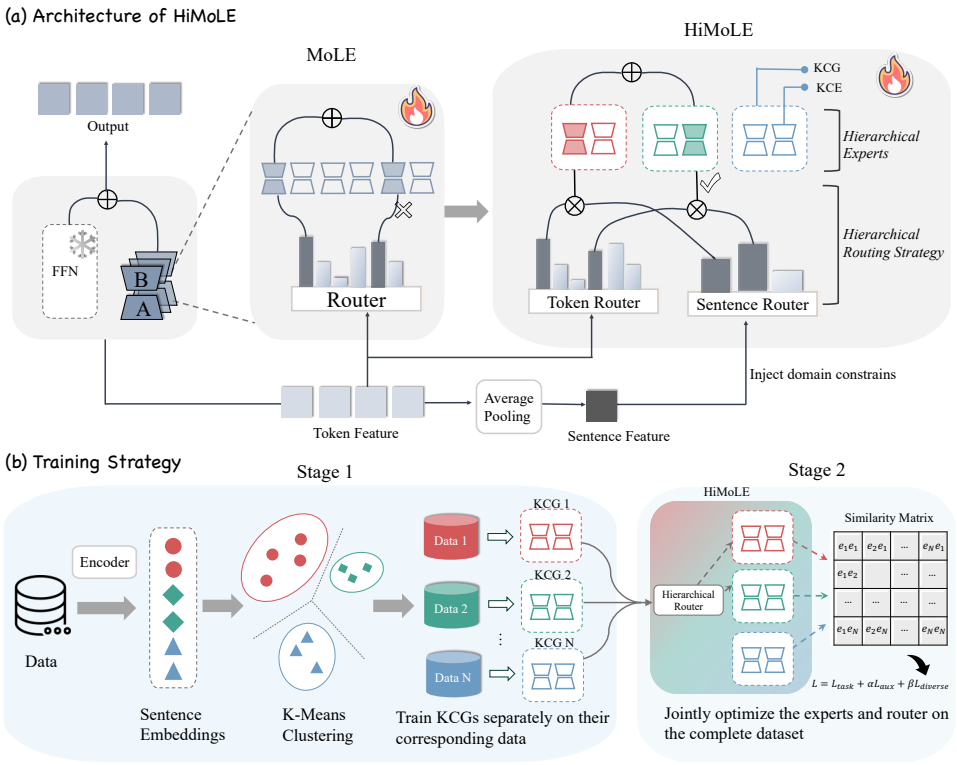


Figure 2: Illustration of the proposed HiMoLE. (a) Architecture of HiMoLE. Unlike MoLE, HiMoLE combines hierarchical experts—organized as Knowledge Competition Groups (KCGs) and their internal Knowledge Collaboration Experts (KCEs)—with a hierarchical routing strategy that performs sentence-level coarse allocation followed by token-level refinement. This architecture is designed to enhance OOD robustness in LoRA-based fine-tuning. (b) Training strategy of HiMoLE, which adopts a two-stage training strategy: first, each KCG is trained independently on a clustered sub-dataset; second, both the expert parameters and hierarchical routing components are jointly optimized.

is processed by a sentence-level router $f_{\text{sen}}(\cdot)$ (implemented as a linear layer parameterized by W_{sen}), to compute the initial allocation scores G_{sen} over KCGs:

$$G_{\text{sen}} = f_{\text{sen}}(h_{\text{sen}}) = W_{\text{sen}} \cdot h_{\text{sen}}. \tag{2}$$

These sentence-level scores serve as a coarse-grained guide, determining which KCGs are most relevant to the input. Subsequently, for each token in the sentence, a token-level router $f_{\text{token}}(\cdot)$ (implemented as a linear layer parameterized by W_{token}) refines this allocation score by integrating the token-specific hidden state h_{token}:

$$G_{\text{token}} = f_{\text{token}}(h_{\text{token}}) = W_{\text{token}} \cdot h_{\text{token}}. \tag{3}$$

The final gating weights matrix G_{hie} are computed using a softmax-normalized fusion of G_{sen} and G_{token}, followed by top-k selection to ensure sparsity:

$$G_{\text{hie}} = \text{KeepTop-}k\left(\text{Softmax}\left(G_{\text{sen}} \odot G_{\text{token}}\right)\right). \tag{4}$$

The forward process of the HiMoLE layer replaced the traditional FFN layer can be represented as:

$$o = W_0 \cdot h_{\text{token}} + \sum_{i=1}^{N \times M} G_{\text{hie}}^{(i)} \cdot E_i \cdot h_{\text{token}}, \tag{5}$$

where W_0 is the parameter matrix of the original FFN layer of the LLM, and o denotes the output. The scalar $G_{\text{hie}}^{(i)}$ modulates the contribution weight of the i-th expert E_i. We provide the definitions of the symbols in Appendix B.

In summary, the hierarchical routing mechanism enables HiMoLE to balance domain specialization and generalization. Sentence-level routing assigns inputs to suitable KCGs based on coarse semantic

cues, while token-level routing fine-tunes expert selection for dynamic, context-aware feature fusion. This design allows for flexible expert collaboration and competition, ultimately enhancing OOD robustness and domain-adaptive inference.

3.3 Training Strategy

As shown in Fig. 2(b), we adopt a two-stage training strategy to construct and optimize the HiMoLE framework. The first stage initializes the Knowledge Competition Groups (KCGs) with specialized domain knowledge, while the second stage jointly optimizes both the expert networks and the hierarchical routing mechanisms.

Stage 1: Initializing Knowledge Competition Groups We begin with the assumption that each task may span multiple subdomains. To capture this diversity, we partition the training dataset into N subsets, each corresponding to a distinct semantic cluster, and train N KCGs in parallel. To perform data clustering, we first use a pre-trained encoder to obtain semantic embeddings for each data instance. We then apply the K-means clustering algorithm to group the data into N clusters. Each cluster is treated as a sub-dataset, and a separate KCG is independently trained on it. This process results in N distinct groups of LoRA-based LLM experts, each specialized in a specific knowledge subdomain.

Stage 2: Co-optimizing the Experts and Routers After initializing the N KCGs, we jointly optimize the expert parameters and the hierarchical routing modules. To encourage diversity among experts and reduce redundancy, we introduce a diversity loss $\mathcal{L}_{\text{diverse}}$. Let e_n denote the output of the n-th KCG, computed as:

$$e_n = \sum_{m=1}^{M} G_{\text{token}}^{(m)} \cdot E_m \cdot h_{\text{token}}, \quad \text{where} \quad E_m \in \text{KCG}_n. \tag{6}$$

We normalize each expert output as: $e_n \leftarrow \frac{e_n}{\max\left(\|e_n\|_2, \epsilon\right)}$, where ϵ is a very small number such as 10^{-8}. We then compute pairwise cosine similarities $S_{nl} = \langle e_n, e_l \rangle$ among all the KCG pairs, and define the diversity loss as the average similarity across all unique expert pairs:

$$\mathcal{L}_{\text{diverse}} = \frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{l=1, l \neq n}^{N} S_{nl}. \tag{7}$$

To control computational cost, $\mathcal{L}_{\text{diverse}}$ is computed every ten layers using a sampled subset of expert outputs. The final training objective combines task loss $\mathcal{L}_{\text{task}}$, auxiliary loss $\mathcal{L}_{\text{aux}}$, which is employed to mitigate the unbalanced load for experts (following [23], see Appendix C for details), and the diversity loss:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{aux}} + \beta \mathcal{L}_{\text{diverse}}. \tag{8}$$

3.4 Theoretical Analysis of Generalizability in Sparse Routing Systems

We analyze how hierarchical expert routing reduces generalization error by mitigating gradient conflicts through structured sparsity in expert selection. Let t index tokens and θ be the parameters of an expert E, which includes the low-rank matrices A and B as defined in Eq. 1. Let $\nabla_\theta \mathcal{L}(\cdot)$ denote the gradient of the loss with respect to θ. We define the expected pairwise gradient similarity as follows:

$$SimGrad := \mathbb{E}_{t \neq t'} \left[\cos\left(g_t, g_{t'}\right) \right], \quad \text{where} \quad g_t := \nabla_\theta \mathcal{L}(h_t). \tag{9}$$

Definition 1 (Gradient Conflict) *Let $h_t \neq h_{t'}$ be inputs from distinct tokens. A* gradient conflict *occurs if* $\cos\left(\nabla_\theta \mathcal{L}(h_t), \nabla_\theta \mathcal{L}(h_{t'})\right) < 0$.

In this context, a higher value of $SimGrad$ indicates better alignment between token gradients and thus fewer gradient conflicts [24].

Theorem 1 (Hierarchical Routing Mitigates Gradient Conflicts) *Let $SimGrad_{hie}$ and $SimGrad_{token}$ denote the expected pairwise gradient similarity under hierarchical and token-only*

routing, respectively. Then hierarchical routing with composition $f_{sen}(h_{sen}) \odot f_{token}(h_{token})$ *yields:*

$$SimGrad_{hie} = SimGrad_{token} + \Delta_\theta, \tag{10}$$

where $\Delta_\theta >= 0$. *Proof. See Appendix D.*

That is, hierarchical routing reduces the prevalence of conflicting gradients by inducing structural sparsity in expert selection. We then connect gradient alignment to generalization through a bound on the generalization error.

Lemma 1 (Generalization Bound via Gradient Variance [25]) *Let* $g_t := \nabla_\theta \mathcal{L}(h_t)$, *and define the gradient variance as* $V(g) := \mathbb{E}\left[||(g_t - \mathbb{E}(g_t)||^2\right]$. *Then, the generalization error of stochastic gradient descent with additive Gaussian noise satisfies:*

$$\mathrm{Gen} \leq \sqrt{\frac{R^2}{b} \sum_{\tau=1}^{T} \frac{\eta_\tau^2}{\sigma_\tau^2} \mathbb{E}\big[V(g)\big]}, \tag{11}$$

where R *represents a constant related to the properties of the loss function and the data distribution, and* b *denotes the number of training samples.* T *is the total number of iterations.* η_τ *is the learning rate at step* τ, *and* σ_τ *is the standard deviation of the Gaussian noise at step* τ.

Theorem 2 (Gradient Conflict Reduction Enhances Model Generalization) *Assume hierarchical routing achieves a lower gradient variance such that* $V(g_{hie}) <= V(g_{token})$, *then under the conditions of Lemma 1, hierarchical routing yields a tighter generalization bound, i.e.,* $\mathrm{Gen}_{hie} \leq \mathrm{Gen}_{token}$.

Proof. By monotonicity of the square root and the inequality on gradient variance, the result follows directly from Lemma 1. See Appendix D for further discussion.

4 Experiments

4.1 OOD benchmark

Name Entity Recognition (NER). To emulate real-world data heterogeneity and enhance the complexity of the NER task, we selected the biomedical domain as our experimental scenario—a knowledge-intensive field characterized by diverse entity types and lexical variations. For the construction of the ID dataset, we rigorously curated English-language resources from the BigBio benchmark [26], with the corpora primarily sourced from PubMed Central (PMC), a premier repository of peer-reviewed biomedical literature. This process resulted in BigBio-NER, the largest dataset for biomedical NER. For the selection of OOD datasets, we adopted the criteria outlined in Section 3.1, choosing the rare disease dataset [27] sourced from the National Organization for Rare Disorders database [28].

Sentiment Analysis (SA). To further enhance the complexity of the SA task and better simulate real-world application scenarios, we frame our SA experiments within social science contexts, where affective expressions exhibit heightened subjectivity and domain-specific connotations. We adopted the sentiment analysis component in SOCIALITEINSTRUCTIONS [29] dataset as our ID dataset. This comprehensive collection of socially-oriented textual interactions contains sentiment labels across various social science scenarios. For the selection of OOD data, we adopted the criteria outlined in Section 3.1 and chose the OPTIMISM [30] dataset.

Extractive Question Answering (EQA). Following previous work [4], we chose SQuAD [31] as the ID dataset, which constructs question-answer pairs based on Wikipedia passages. For the selection of OOD data, we chose NewsQA [32], which writes questions for CNN news articles, each of which requires reasoning to answer.

4.2 Experimental Settings

Base Model and Data Separation For the NER task, we employ OneKE-13B [33] as our base model, which is capable of generalized knowledge extraction across multiple domains and tasks. For

Table 1: Comparative performance of different LoRA methods under out-of-distribution scenarios. Please refer to Appendix E.3 for the metrics details and Appendix E.5 for the complete results. The best results on ID data and the best results on OOD data excluding the base model are highlighted in **boldface** and underlined, respectively.

Task	**NER**						**SA**			**EQA**			
Dataset	ID			OOD			ID	OOD		ID		OOD	
Metric	F1	P	R	F1	P	R	EM	REM	EM	EM	ROUGE-2	EM	ROUGE-2
Base Model	52.5	56.2	51.7	65.8	63.1	76.4	48.8	59.5	56.9	7.2	12.3	8.3	15.8
LoRA	73.9	77.5	73.3	61.9	56.0	77.8	66.7	55.7	55.0	68.2	48.7	35.5	28.2
MixLoRA	76.0	78.2	75.7	61.0	59.1	73.7	69.5	66.5	31.4	69.4	48.6	38.3	27.9
HydraLoRA	77.3	78.4	76.7	62.9	59.4	75.5	70.3	67.2	30.4	68.7	46.4	38.3	26.4
HiMoLE	**77.9**	**78.7**	**77.3**	65.3	64.4	74.3	**73.3**	68.8	32.8	**70.5**	**49.8**	38.9	28.7

the SA and EQA tasks, we utilize Llama-2-7B [34] as our base model. To separate in-distribution data, we use BioBERT [35], BERTweet [36], and DeBERTa-v3-large [37] as encoders to extract sentence-level embeddings for the NER, SA and EQA tasks, respectively. Through sentence feature-based K-means clustering, we obtain subsets of sizes 4, 3 and 3 for each task (see the visualization of the clustering results in Appendix E.1).

Baselines and Settings We compare our HiMoLE method against LoRA and traditional mixture of LoRA experts. In our HiMoLE approach, each Knowledge Component Group consists of 4 Knowledge Component Experts. For the mixture of LoRA experts, we compared state-of-the-art methods, including MixLoRA [6] and HydraLoRA [9]. In all mixture of experts methods, we adopted the tok-k routing and set k in tok-k to 2, while setting the LoRA rank r to 8. We apply MOE to the FFN layer of every Transformer block. To maintain uniformity in parameter sizes across all methods, we set the rank r to 80 in traditional LoRA.

4.3 Primary Results

In Table 1, we provide a comprehensive comparison of HiMoLE with various baselines. There are several observations: (1) Overall, HiMoLE achieves the best performance across all tasks on both ID and OOD datasets compared to other LoRA-based methods. Specifically, compared to the best baseline, HiMoLE achieves improvements of up to 3.0% on ID datasets and 5.0% on OOD datasets. (2) Although the recently proposed HydraLoRA and MixLoRA suggest their ability under multi-task learning scenarios, the lack of adequate sentence-level message integration for routing creates a bottleneck under out-of-distribution data, even resulting in worse performance in knowledge-intensive domains. (3) The generalization improvements from HiMoLE demonstrate domain-dependent variability, showing more pronounced gains in knowledge-intensive domains, with the performance enhancement compared to the best baseline increasing from 0.6% to 5.0% on OOD datasets. (4) All MoPE-based methods still face limitations in enhancing generalization ability when operating under a fixed number of parameters and limited data.

4.4 Effectiveness of HiMoLE on OOD samples

Hierarchical Experts To substantiate the superiority of our hierarchical expert design, we conduct load balancing evaluations on out-of-distribution datasets. Through the visualization of expert selection logits from a randomly sampled OOD instance (Fig. 4, see the complete result in Appendix E.5), we observe that MixLoRA exhibits consistently severe load imbalance across layers compared to HiMoLE, with this disparity intensifying at deeper layers- particularly in its final layer where 78.3% of tokens disproportionately select Expert 4 with probabilities exceeding 0.85. Quantitatively, we adopt the $\text{MaxVio}_{\text{global}}$ metric from [38] (see the definition in Appendix E.3), to evaluate MoPE-layer balancing. As displayed in Fig. 4, HiMoLE consistently improved the load balancing among experts, with particularly notable improvements in knowledge-intensive domains (20.6% reduction in biomedical entity recognition). These results indicates that HiMoLE's hierarchical architecture effectively reduces expert redundancy through structural sparsity constraints, thereby promoting experts specialization and balanced expert utilization.

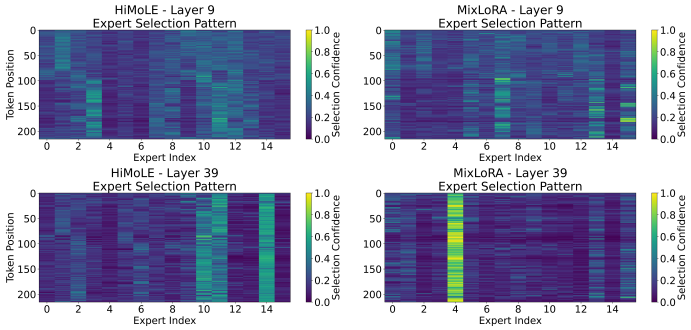


Figure 3: Visualization of expert selection logits produced by HiMoLE and MixLoRA.

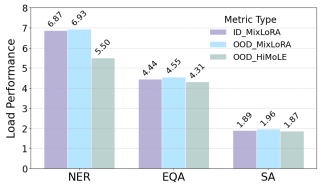


Figure 4: Comparison of experts load balance for different mixture of LoRA experts methods. We utilize MaxVio to evaluate the load performance.

Table 2: Robustness on character-level adversarial attack. Reported results are accuracy scores. The Robustness Ratio is defined as adversarial sample accuracy divided by original accuracy.

Metric	Base	HiMoLE	MixLoRA
original	0.488	**0.733**	0.695
attacked	0.196	**0.330**	0.280
Robustness Ratio	40.2%	**45.1%**	40.3%

Table 3: Ablation study on the two stage training strategy and diverse loss. Reported results are F1 scores. Please refer to Appendix E.5 for the complete results.

Dataset	ID1	ID2	ID3	ID4	OOD
HiMoLE	87.6	64.0	75.3	79.6	65.3
-w/o. Two-stage Training	77.7	32.1	64.1	60.8	53.5
-w/o. Diverse Loss	87.0	63.1	74.7	78.2	65.0

Hierarchical Routing Strategy To further investigate the robustness of the hierarchical routing strategy, we compared its performance against the token routing method (MixLoRA) in the sentiment analysis task. This assessment involved generating adversarial out-of-distribution samples using the TextBugger [39] tool. Specifically, we randomly sampled 500 instances from the in-distribution test set and injected token-level noise via character-level perturbations.

As shown in Table 2, both routing strategies experienced significant performance degradation under adversarial attacks; however, hierarchical router demonstrated not only a superior absolute performance over token-level router, with a 6.0% improvement, but also exhibited a smaller decline in performance, with a 4.8% improvement over the token routing method. This indicates that hierarchical router possesses markedly stronger robustness against adversarial samples when compared to the base model and the token-level MoPE. The results further validate that HiMoLE, through its integration of global features, maintains stable performance against local perturbations by selecting appropriate experts via sentence-level semantic analysis, even in adversarial scenarios.

4.5 Ablations

Two-stage Training Strategy We investigated the influence of different training strategies on model performance by conducting the experiment on the NER task. As evidenced in Table 3, when using single-stage joint training without preliminary KCG initialization, we observed a dramatic performance degradation of up to 31.9% F1 on the ID2 dataset. Furthermore, we examined the bad cases and found that the proportion of incorrect formats had increased significantly (see Appendix A). These observations indicate that (1) unstable optimization trajectories were adopted when learning router-expert interactions from random initialization, and (2) the absence of first-stage specialization prevents KCGs from developing domain-specific inductive biases, resulting in ambiguous routing signals. Thus, proper KCG initialization is critical for our hierarchical routing mechanism to function.

Diversity Loss We investigated the importance of diverse loss on model performance. As evidenced in Table 3, omitting the diversity loss component leads to statistically significant performance degradation across both ID and OOD data. This consistent pattern reveals that diversity regularization can further mitigate expert group redundancy and improve model robustness.

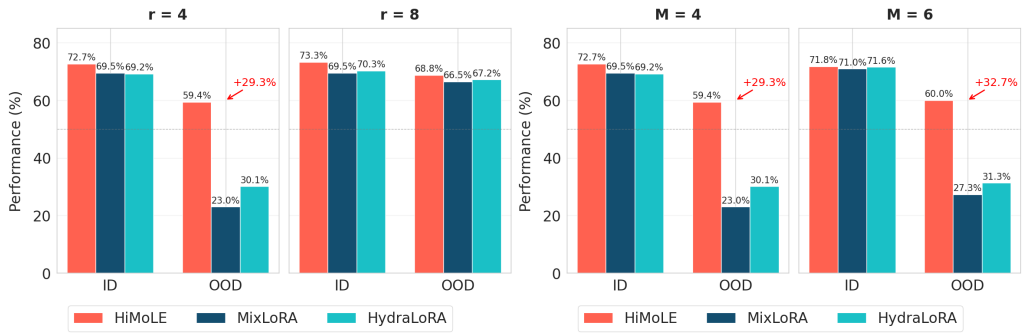


Figure 5: Hyper-parameter Analysis on LoRA rank r and Knowledge Collaboration Experts Numbers M. Performance on OOD dataset are evaluated using REM (Appendix E.3).

4.6 Hyper-parameter Analysis

As illustrated in Figure 5, we further conduct hyper-parameter analysis on the sentiment analysis task examine the impact of the LoRA rank r and the number of knowledge collaboration experts M. In experiments, we fix $r = 4$ when analyzing M, and conversely maintain $M = 4$ when investigating r. The results reveal that HiMoLE demonstrates superior robustness against variations in both parameters, consistently outperforming baseline methods across all configurations. Notably, while competing approaches exhibit significant performance degradation on OOD data with reduced LoRA ranks, HiMoLE achieves a remarkable OOD accuracy improvement from 31.3% to 60.0%, conclusively validating our method's effectiveness.

5 Conclusion

While PEFT methods like LoRA have significantly lowered the barriers for adapting large language models to downstream tasks, our investigation exposes their critical vulnerability to distributional shifts—particularly in knowledge-intensive domains. Our proposed HiMoLE framework alleviates this fundamental problem with by integrating hierarchical experts with a hierarchical routing strategy. This approach leverages sentence-level information to coarsely allocate experts to relevant subdomains and then refines the routing weights using token-level information, enabling efficient acquisition of new knowledge while preserving existing knowledge. By theoretically and empirically validating this approach across three representative NLP tasks, we establish a new paradigm for developing adaptable language models that achieve parameter efficiency with enhanced generalization capacity.

Acknowledgement

This work is funded by Zhejiang Provincial "Jianbing" "Lingyan" Research and Development Program of China (2025C01129), National Natural Science Foundation of China (62302433, 62301480, U23A20496), and Hangzhou West Lake Pearl Project Leading Innovative Youth Team Project (TD2023017). This work was supported by Ant Group and Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph.

References

[1] 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, 2021.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[3] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018.

[4] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations, 2023.

[5] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[6] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts, 2024.

[7] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning, 2024.

[8] Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning, 2023.

[9] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning, 2024.

[10] Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. Out-of-distribution generalization in natural language processing: Past, present, and future. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore, December 2023. Association for Computational Linguistics.

[11] Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in nlp models: A survey, 2022.

[12] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. On the robustness of chatgpt: An adversarial and out-of-distribution perspective, 2023.

[13] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts, 2024.

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

[15] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[16] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning, 2022.

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

[18] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications, 2024.

[19] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning, 2023.

[20] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[21] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[22] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks, 2015.

[23] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models, 2022.

[24] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

[25] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent, 2021.

[26] Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. Bigbio: A framework for data-centric biomedical natural language processing, 2022.

[27] Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. Identifying and extracting rare diseases and their phenotypes with large language models. *Journal of Healthcare Informatics Research*, 8(2):438–461, January 2024.

[28] Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. The raredis corpus: a corpus annotated with rare diseases, their signs and symptoms, 2021.

[29] Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Andrew Schwartz. SOCIALITE-LLAMA: An instruction-tuned model for social scientific tasks. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–468, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[30] Xianzhi Ruan, Steven Wilson, and Rada Mihalcea. Finding optimists and pessimists on Twitter. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 320–325, Berlin, Germany, August 2016. Association for Computational Linguistics.

[31] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[32] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[33] Yujie Luo, Xiangyuan Ru, Kangwei Liu, Lin Yuan, Mengshu Sun, Ningyu Zhang, Lei Liang, Zhiqiang Zhang, Jun Zhou, Lanning Wei, Da Zheng, Haofen Wang, and Huajun Chen. Oneke: A dockerized schema-guided llm agent-based knowledge extraction system, 2025.

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

[35] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[36] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October 2020. Association for Computational Linguistics.

[37] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023.

[38] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts, 2024.

[39] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society, 2019.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and

write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In Section 3.4 and Appendix D

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Appendix E.2

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the datasets are open-source, and the code can be found in the supplementary materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details can be found in Section 4.2 and Appendix E.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: All results from our experiments are presented as either the best or average outcomes.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources can be found in Appendix E.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We make sure the research is conducted in the paper , in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: In Appendix G.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The models and the data are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This study does not involve any research conducted with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve any research conducted with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The LLM is used only for formatting purpose.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Bad case studies

Incorrect Format In the training ID data, the presence of numerous duplicate entity objects in the labels causes LoRA-based fine-tuning methods to inadvertently replicate this pattern. Consequently, when applied to OOD data, the model tends to generate more entities than necessary, often resulting in repetitive output behavior. This ultimately hinders its ability to produce correctly formatted JSON outputs as instructed. As illustrated in Table 4, this issue is particularly pronounced with both simple LoRA and token-level MoLE methods. In contrast, sentence-level MoLE and HiMoLE effectively address this problem. Notably, sentence-level MoLE ensures that all examples produce correctly formatted JSON outputs.

Table 4: Comparison of the Format Robustness

Method	Base	LoRA	Token-MoLE	Sentence-MoLE	HiMoLE	HiMoLE - two stage training
Incorrect Format Ratio(%)	1.19	8.33	7.14	0.00	2.38	44.0

Misclassifications Analyzing the erroneous results revealed that models trained on ID data frequently mislabel certain symptoms as diseases when assessed on OOD data, failing to apply the general knowledge that differentiates the two. HiMoLE significantly mitigates this issue. Figure 6 showcases several comparative results between MixLoRA and HiMoLE.

B Definitions of the Symbols

Table 5: Definitions of the Symbols used in the paper

Symbol	Description
i, m, n, l	the index number
B, A	the low-rank matrices
E	the weight matrix of a LoRA expert
$r, d_{\text{in}}, d_{\text{out}}, d_{\text{emb}}$	the LoRA rank, the dimension of the input and the output of FFN layer, the dimension of the sentence embedding used for dataset clustering
N	the number of the Knowledge Competition Groups
M	the number of the Knowledge Collaboration Experts in a Knowledge Competition Group
$G_{\text{token}}, G_{\text{sen}}, G_{\text{hie}}$	the gating weights matrix derived from the sentence-level router, the token-level router and the hierarchical router, respectively
$W_{\text{token}}, W_{\text{sen}}$	the learnable parameters of the sentence-level router and the token-level router
$f_{\text{token}}, f_{\text{sen}}$	the representation of the sentence-level router and the token-level router
o	the output of a FFN layer
h_{token}	token-level hidden representation
h_{sen}	sentence-level hidden representation
e	the output of a Knowledge Competition Group
S	the cosine similarity between a pair of Knowledge Competition Groups
$\mathcal{L}_{\text{task}}, \mathcal{L}_{\text{aux}}, \mathcal{L}_{\text{diverse}}$	the representations of overall loss, the task loss, the auxiliary loss and the diverse loss
$\nabla_\theta \mathcal{L}(\cdot)$	the gradient of the parameter θ with the loss function $\mathcal{L}$
$SimGrad$	the expected pairwise gradient similarity from all different tokens
$V(g)$	the variance of gradient
$\mathbb{E}$	the notation of the expectation
$\mathbb{R}$	the real number space

Base	MixLoRA	HiMoLE
Alkaptonuria is a rare genetic metabolic disorder characterized by the accumulation of homogentisic acid in the body. Affected individuals lack enough functional levels of an enzyme required to breakdown homogentisic acid. Affected individuals may have dark urine or urine that turns black when exposed to air. However, this change may not occur for several hours after urination and often goes unnoticed. Aside from dark urine that is present from infancy, affected individuals generally do not develop symptoms (asymptomatic) during infancy or childhood and often remain unaware of their condition until adulthood. Affected individuals eventually develop ochronosis, which is the bluish-black discoloration of connective and other tissue within the body. Affected individuals may develop discoloration of the skin overlying cartilage within the body such as over part of the outer ear. In some cases, the whites of the eyes (sclera) may also become discolored. In adulthood, affected individuals also develop progressive arthritis of the spine and large joints. The HGD gene codes for the enzyme required for the breakdown of homogentisic acid. Mutations in the HGD gene cause alkaptonuria. Alkaptonuria affects males and females in equal numbers, although symptoms tend to develop sooner and become more severe in males. More than 1,000 affected individuals have been reported in the medical literature. The exact incidence of alkaptonuria is unknown. In the United States it is estimated to occur in 1 in 250,000-1,000,000 live births. Alkaptonuria has been reported in all ethnic groups. Areas with increased frequencies of the disorder have been identified in Slovakia, the Dominican Republic and Germany.	Alkaptonuria is a rare genetic metabolic disorder characterized by the accumulation of homogentisic acid in the body. Affected individuals lack enough functional levels of an enzyme required to breakdown homogentisic acid. Affected individuals may have dark urine or urine that turns black when exposed to air. However, this change may not occur for several hours after urination and often goes unnoticed. Aside from dark urine that is present from infancy, affected individuals generally do not develop symptoms (asymptomatic) during infancy or childhood and often remain unaware of their condition until adulthood. Affected individuals eventually develop ochronosis, which is the bluish-black discoloration of connective and other tissue within the body. Affected individuals may develop discoloration of the skin overlying cartilage within the body such as over part of the outer ear. In some cases, the whites of the eyes (sclera) may also become discolored. In adulthood, affected individuals also develop progressive arthritis of the spine and large joints. The HGD gene codes for the enzyme required for the breakdown of homogentisic acid. Mutations in the HGD gene cause alkaptonuria. Alkaptonuria affects males and females in equal numbers, although symptoms tend to develop sooner and become more severe in males. More than 1,000 affected individuals have been reported in the medical literature. The exact incidence of alkaptonuria is unknown. In the United States it is estimated to occur in 1 in 250,000-1,000,000 live births. Alkaptonuria has been reported in all ethnic groups. Areas with increased frequencies of the disorder have been identified in Slovakia, the Dominican Republic and Germany.	Alkaptonuria is a rare genetic metabolic disorder characterized by the accumulation of homogentisic acid in the body. Affected individuals lack enough functional levels of an enzyme required to breakdown homogentisic acid. Affected individuals may have dark urine or urine that turns black when exposed to air. However, this change may not occur for several hours after urination and often goes unnoticed. Aside from dark urine that is present from infancy, affected individuals generally do not develop symptoms (asymptomatic) during infancy or childhood and often remain unaware of their condition until adulthood. Affected individuals eventually develop ochronosis, which is the bluish-black discoloration of connective and other tissue within the body. Affected individuals may develop discoloration of the skin overlying cartilage within the body such as over part of the outer ear. In some cases, the whites of the eyes (sclera) may also become discolored. In adulthood, affected individuals also develop progressive arthritis of the spine and large joints. The HGD gene codes for the enzyme required for the breakdown of homogentisic acid. Mutations in the HGD gene cause alkaptonuria. Alkaptonuria affects males and females in equal numbers, although symptoms tend to develop sooner and become more severe in males. More than 1,000 affected individuals have been reported in the medical literature. The exact incidence of alkaptonuria is unknown. In the United States it is estimated to occur in 1 in 250,000-1,000,000 live births. Alkaptonuria has been reported in all ethnic groups. Areas with increased frequencies of the disorder have been identified in Slovakia, the Dominican Republic and Germany.

Base	MixLoRA	HiMoLE
Fitz-Hugh-Curtis syndrome is a rare disorder that occurs almost exclusively in women. It is characterized by inflammation of the membrane lining the stomach (peritoneum) and the tissues surrounding the liver (perihepatitis). The muscle that separates the stomach and the chest (diaphragm), which plays an essential role in breathing, may also be affected. Common symptoms include severe pain in the upper right area (quadrant) of the abdomen, fever, chills, headaches, and a general feeling of poor health (malaise). Fitz-Hugh-Curtis syndrome is a complication of pelvic inflammatory disease (PID), a general term for infection of the upper genital tract in women. Infection is most often caused by Neisseria gonorrhoeae and Chlamydia trachomatis. A diagnosis of Fitz-Hugh-Curtis syndrome is made through the exclusion of other causes of upper right abdominal pain. A diagnosis may be confirmed with a variety of specialized tests including x-ray examination, diagnostic laparoscopy, and certain laboratory exams. X-ray examination may include ultrasound, chest or stomach radiographs, and computed tomography (CT) scanning. X-rays are used to rule out other possible conditions or reveal characteristic inflammation of the perihepatic region. During a laparoscopy, a small, thing tube is inserted in the abdominal cavity through a small incision in the stomach. A laparoscopic exam allows a physician to view the liver and surrounding tissue. Laboratory exams can identify infection with Chlamydia trachomatis or Neisseria gonorrhoeae.	Fitz-Hugh-Curtis syndrome is a rare disorder that occurs almost exclusively in women. It is characterized by inflammation of the membrane lining the stomach (peritoneum) and the tissues surrounding the liver (perihepatitis). The muscle that separates the stomach and the chest (diaphragm), which plays an essential role in breathing, may also be affected. Common symptoms include severe pain in the upper right area (quadrant) of the abdomen, fever, chills, headaches, and a general feeling of poor health (malaise). Fitz-Hugh-Curtis syndrome is a complication of pelvic inflammatory disease (PID), a general term for infection of the upper genital tract in women. Infection is most often caused by Neisseria gonorrhoeae and Chlamydia trachomatis. A diagnosis of Fitz-Hugh-Curtis syndrome is made through the exclusion of other causes of upper right abdominal pain. A diagnosis may be confirmed with a variety of specialized tests including x-ray examination, diagnostic laparoscopy, and certain laboratory exams. X-ray examination may include ultrasound, chest or stomach radiographs, and computed tomography (CT) scanning. X-rays are used to rule out other possible conditions or reveal characteristic inflammation of the perihepatic region. During a laparoscopy, a small, thing tube is inserted in the abdominal cavity through a small incision in the stomach. A laparoscopic exam allows a physician to view the liver and surrounding tissue. Laboratory exams can identify infection with Chlamydia trachomatis or Neisseria gonorrhoeae.	Fitz-Hugh-Curtis syndrome is a rare disorder that occurs almost exclusively in women. It is characterized by inflammation of the membrane lining the stomach (peritoneum) and the tissues surrounding the liver (perihepatitis). The muscle that separates the stomach and the chest (diaphragm), which plays an essential role in breathing, may also be affected. Common symptoms include severe pain in the upper right area (quadrant) of the abdomen, fever, chills, headaches, and a general feeling of poor health (malaise). Fitz-Hugh-Curtis syndrome is a complication of pelvic inflammatory disease (PID), a general term for infection of the upper genital tract in women. Infection is most often caused by Neisseria gonorrhoeae and Chlamydia trachomatis. A diagnosis of Fitz-Hugh-Curtis syndrome is made through the exclusion of other causes of upper right abdominal pain. A diagnosis may be confirmed with a variety of specialized tests including x-ray examination, diagnostic laparoscopy, and certain laboratory exams. X-ray examination may include ultrasound, chest or stomach radiographs, and computed tomography (CT) scanning. X-rays are used to rule out other possible conditions or reveal characteristic inflammation of the perihepatic region. During a laparoscopy, a small, thing tube is inserted in the abdominal cavity through a small incision in the stomach. A laparoscopic exam allows a physician to view the liver and surrounding tissue. Laboratory exams can identify infection with Chlamydia trachomatis or Neisseria gonorrhoeae.

Base	MixLoRA	HiMoLE
Spondyloepiphyseal dysplasia tarda (SEDT; SEDL) is a rare, hereditary skeletal disorder that only affects males. Physical characteristics include moderate short stature (dwarfism), moderate-to-severe spinal deformities, barrel-shaped chest, disproportionately short trunk, and premature osteoarthritis. SEDT does not exhibit any ethnic predisposition. Affected individuals have been described in European, American, Asian, and Australian populations (but not in African-Americans to date). One estimate suggests that the incidence is 2 persons per million.	Spondyloepiphyseal dysplasia tarda (SEDT; SEDL) is a rare, hereditary skeletal disorder that only affects males. Physical characteristics include moderate short stature (dwarfism), moderate-to-severe spinal deformities, barrel-shaped chest, disproportionately short trunk, and premature osteoarthritis. SEDT does not exhibit any ethnic predisposition. Affected individuals have been described in European, American, Asian, and Australian populations (but not in African-Americans to date). One estimate suggests that the incidence is 2 persons per million.	Spondyloepiphyseal dysplasia tarda (SEDT; SEDL) is a rare, hereditary skeletal disorder that only affects males. Physical characteristics include moderate short stature (dwarfism), moderate-to-severe spinal deformities, barrel-shaped chest, disproportionately short trunk, and premature osteoarthritis. SEDT does not exhibit any ethnic predisposition. Affected individuals have been described in European, American, Asian, and Australian populations (but not in African-Americans to date). One estimate suggests that the incidence is 2 persons per million.

Figure 6: Misclassifications Cases. Text highlighted in green represents correct entity annotations, while yellow represents incorrect entity annotations.

C Auxiliary Loss

Given N experts indexed by i =1 to N and a batch B with T tokens. Let $G(\cdot)$ denotes the top-k router, F_i is the fraction of tokens dispatched to expert, and P_i i is the fraction of the router probability allocated for expert i. The final loss is then multiplied by the expert count N to keep the loss constant as the number of experts varies, which can be formulated as following:

$$F_i = \frac{1}{T} \sum_{x \in \mathcal{B}} \mathbb{I}\left(\text{argmax}_k \, R(x)_k = i\right), P_i = \frac{1}{T} \sum_{x \in \mathcal{B}} R(x_i), \tag{12}$$

$$\mathcal{L}_{\text{aux}} = N \cdot \sum_{i=1}^{N} F_i \cdot P_i, \tag{13}$$

D Proof

Here we provide the proof of Theorem 1.

Lemma 2 *Let t represent distinct tokens and s represent sentences. To simplify, let's assume that the sample gradients are a set of independent and identically distributed unit vectors. In this case, $SimGrad$ can be expressed as $||\mathbb{E}(\nabla_{\theta_i} L(h_t))||^2$:*

$$\begin{aligned} SimGrad &= \mathbb{E}(cos(g_t, g_{t'})) = \mathbb{E}(g_t \cdot g_{t'}) \\ &= \mathbb{E}_{g_t}(\mathbb{E}_{g_t'}[g_t \cdot g_{t'} | g_t]) = \mathbb{E}_{g_t}(g_t \cdot \mathbb{E}_{g_{t'}}(g_{t'})) = \mathbb{E}_{g_t}(g_t) \cdot \mathbb{E}_{g_{t'}}(g_{t'}) \\ &= ||\mathbb{E}_{g_t}||^2 = ||\mathbb{E}(\nabla_{\theta_i} L(x_t))||^2 \end{aligned} \tag{14}$$

Proof D.1 *For the i-th expert E_i, let $A_{s,t}$ denotes $\mathbb{I}\left(E_i \in \textit{argmax}_k(f_{\textit{sen}}(h_s) \odot f_{\textit{token}}(h_t))\right)$, $B_{s,t}$ denotes $\mathbb{I}\left(E_i \in \textit{argmax}_k(f_{\textit{token}}(h_t))\right)$. For hierarchical router $G_{\textit{hie}}$, and token router $G_{\textit{token}}$, their gradient operators can be written as follows:*

$$\begin{aligned} \nabla_{\theta_i} \mathcal{L}_{\textit{hie}} &= \mathbb{I}\big(E_i \in \textit{argmax}_k(G_{\textit{hie}})\big) \cdot \nabla \theta_i \mathcal{L}(h_{s,t}) \\ &= \mathbb{I}\big(E_i \in \textit{argmax}_k(f_{\textit{sen}}(h_s) \odot f_{\textit{token}}(h_t))\big) \cdot \nabla \theta_i \mathcal{L}(h_{s,t}) \\ &= A_{s,t} \cdot \nabla \theta_i \mathcal{L}(h_{s,t}) \end{aligned} \tag{15}$$

$$\begin{aligned} \nabla_{\theta_i} \mathcal{L}_{\textit{token}} &= \mathbb{I}\big(E_i \in \textit{argmax}_k(G_{\textit{token}})\big) \cdot \nabla \theta_i \mathcal{L}(x_{s,t}) \\ &= \mathbb{I}\big(E_i \in \textit{argmax}_k(f_{\textit{token}}(x_t))\big) \cdot \nabla \theta_i \mathcal{L}(h_{s,t}) \\ &= B_{s,t} \cdot \nabla \theta_i \mathcal{L}(h_{s,t}) \end{aligned} \tag{16}$$

Combining Lemma 2, Δ_{θ_i} can be written as follows:

$$\begin{aligned} \Delta_{\theta_i} &= SimGrad_{hie} - SimGrad_{token} \\ &= ||\mathbb{E}(\nabla_{\theta_i} L_{\textit{hie}})||^2 - ||\mathbb{E}(\nabla_{\theta_i} L_{\textit{token}})||^2 \\ &= [\mathbb{E}(A_{s,t})^2 - \mathbb{E}(B_{s,t})^2] \cdot \mathbb{E}(\nabla \theta_i L(h_{s,t}))^2 \end{aligned} \tag{17}$$

Since $A_{s,t}, B_{s,t} \in \{0, 1\}$, we have:

$$\mathbb{E}(A_{s,t})^2 = \mathbb{E}(A_{s,t}), \mathbb{E}(\mathbb{B}_{s,t})^2 = \mathbb{E}(B_{s,t})$$

Next, we analyze the relationship between $A_{s,t}$ and $B_{s,t}$. The hierarchical router ensures that the effective selection of token router is not overlooked by positively adjusting the scores. Additionally, by introducing global information through sentence features, it uncovers experts that are neglected under local token features, which guarantee that:

$$B_{s,t} = 1 \Rightarrow A_{s,t} = 1, \textit{ but not vice versa, } A_{s,t} = 1 \not\Rightarrow B_{s,t} = 1$$

As a result,

$$\begin{aligned} \Delta_{\theta_i} &= [\mathbb{E}(A_{s,t})^2 - \mathbb{E}(B_{s,t})^2] \cdot \mathbb{E}(\nabla \theta_i L(h_{s,t}))^2 \\ &= [\mathbb{E}(A_{s,t}) - \mathbb{E}(B_{s,t})] \cdot \mathbb{E}(\nabla \theta_i L(h_{s,t}))^2 >= 0 \end{aligned} \tag{18}$$

Table 6: Description of Datasets used in experiments.

Task	Domain	Train	Test	
			ID	OOD
NER	Biomedical	35132	2060/1267/2764/2746	684
SA	Social Science	11257	2374	1465
EQA	General	87599	10570	3882

Now we turn to the proof of Theorem 2.

Lemma 3 *Let g denotes the gradient, $\mathbb{E}(g)$ denote the average gradient and m denote the number of tokens, then $\mathbb{E}(g)^2$ can be written as follows:*

$$\mathbb{E}(g)^2 = \frac{1}{m^2}\left(\sum_t \|g_t\|^2 + 2\sum_{t\neq t'} g_t \cdot g_{t'}\right) \tag{19}$$

Lemma 4 *Let $V(g)$ represents the gradient variance, then $V(g) \propto -SimGrad$:*

$$\begin{aligned} V(g) &= \mathbb{E}\left[||(g_t - \mathbb{E}(g_t)||^2\right] = \frac{1}{m}\sum_t \|g_t - \mathbb{E}(g)\|^2 \\ &= \frac{1}{m}\left(\|g_t\|^2 - 2g_t \cdot \mathbb{E}(g) + \mathbb{E}(g)^2\right) \\ &= \frac{1}{m}\sum_t \|g_t\|^2 - \mathbb{E}(g)^2 \end{aligned} \tag{20}$$

Combing with Lemma 3, $V(g)$ can be written as:

$$V(g) = \frac{m-1}{m^2}\sum_t \|g_t\|^2 - \frac{2}{m^2}\sum_{t\neq t'} g_t \cdot g_{t'} \propto -SimGrad \tag{21}$$

Proof D.2 *By Theorem 1, $SimGrad_{hie} >= SimGrad_{token}$. By lemma 4, higher $SimGrad$ indicates lower gradient variance, hence $V(g_{hie}) <= V(g_{token})$.*

E Experiments Details

E.1 Datasets

Table 6 summarizes the datasets used in our experiments, including their task names, respective domains, the number of training and test sets. For Biomedical NER, the ID test data was partitioned into 4 sub-datasets using feature-based K-means clustering. All datasets are downloaded from HuggingFace using the DATASETS library in Python. Additionally, we provide a UMAP visualization of the datasets in Fig. 7.

E.2 Hyperparameters and Implementation Details

Table 7: Hyperparameter configurations of LoRA, MixLoRA/HydraLoRA and HiMoLE for fine-tuning LLaMA2-7B and OneKE-13B.

Metric	LoRA	MixLoRA/HydraLoRA	HiMoLE
Cutoff Length	1024	1024	1024
Learning Rate	3e-4	3e-4	3e-4(stage1), 3e-5(stage2)
Optimizer	AdamW	AdamW	AdamW
Batch size	16	16	16
Dropout	0.05	0.05	0.05
Where	Up, Down, Gate	Up, Down, Gate	Up, Down, Gate
LoRA Rank	80	8	8
LoRA Alpha	160	16	16
Top-K	-	2	2

We set a maximum of 10,000 training steps and perform evaluations on the validation sets of all benchmarks every 50 steps. If there is no improvement on the validation set for 10 consecutive evaluations, we will terminate the training early. The best checkpoint, identified by the highest average accuracy across all benchmarks, is then selected for evaluation on the test set.

All experiments are conducted with GPUs having 24GB memory (RTX 4090) for 7B models, GPUs having 40GB memory (RTX A100) for 13B models, and setup with Python 3.8 and Ubuntu 22.04 on x86-64 CPUs.

E.3 Evaluation Metrics

Performance For the metrics in Table 1: In NER, F1 stands for the average F1 score, P stands for the average precision, R stands for the average recall; In SA, EM stands for the exact match, REM stands for the relaxed exact match(we treat the 'positive' label as synonymous with 'optimism' and the 'negative' label as synonymous with 'pessimism'); In QA, EM stands for the exact match, ROUGE-2 is employed to captures phrases that hold vital context.

Load Balance MaxVio is used to quantify the degree of load balance of an MoE layer, defined as $\text{MaxVio} = \frac{\max_i \text{Load}_i - \overline{\text{Load}_i}}{\overline{\text{Load}_i}}$,, where Load_i represents the number of tokens assigned to the i-th expert, and $\overline{\text{Load}_i}$ denotes the expected expert load under perfect load balance.In $\text{MaxVio}_{\text{global}}$, Load_i is calculated on the whole validation set.

E.4 Additional Experiments and Analysis

Impact of Two-stage Training Strategy To disentangle the effects of initialization from our hierarchical architecture, we add baseline comparisons where MixLoRA and HydraLoRA are trained with the same Stage 1 initialization as HiMoLE on NER task. As shown in Table 8, although other methods demonstrate improvements, the two-stage training approach combined with the hierarchical mixture of LoRA experts still delivers the best performance across both in-distribution and out-of-distribution settings, with essential improvement under OOD setting.

Table 8: Impact of Training Strategy across mixture of LoRA experts methods. The reported results(%) are F1.

Training Strategy	HiMoLE		MixLoRA		HydraLoRA	
	ID	OOD	ID	OOD	ID	OOD
two stage	**77.9**	**65.3**	76.4	63.0	77.2	77.3
one stage	61.7	53.5	76.0	61.0	77.3	62.9

Inference Latency HiMoLE adds lightweight gating networks, maintaining nearly the same parameter count as MixLoRA. For inference, the speed is primarily influenced by the base model. Since the parameters of PEFT modules constitute a small fraction of the total model parameters (ranging from 0.65% to 1.05% as shown in Table 9), the inference latency differences across are minimal. Table 9 presents the latency and parameter count during inference using Llama2-7B with different Mixture of LoRA Experts methods, evaluated on the OPTIMISM dataset using a single RTX 4090 GPU. Latency was recorded over 50 random samples. The results show nearly equal latency, but HiMoLE exhibits the highest model performance.

E.5 Complete Results

Table 10 presents the complete results of the comparative performance for the NER task in Table 1. Fig. 8 presents the complete results of the routing logits discussed in Section 4.3. Table 11 displays the complete results of the ablation experiments in Table 3.

F Training Strategy

Algorithm 1: HiMoLE Two-Stage Training

Input: LLM's frozen weights W_0, training data $\mathcal{D}$ (composed of (s, y) pairs), pre-trained encoder Encoder$(\cdot)$, cluster count N
Output: Optimized experts and routers

Stage *1: Knowledge Competition Group Initialization*

foreach *sentence* $s_i \in \mathcal{D}$ **do**
  $\text{emb}_i \leftarrow \text{Encoder}(s_i)$; `//Generate semantic embeddings`
$\{\mathcal{C}_1, ..., \mathcal{C}_N\} \leftarrow K\text{-means}(\{\text{emb}_i\}, N)$; `//Cluster embeddings`
For $k \leftarrow 1$ **to** N **in parallel**
  $\mathcal{D}_k \leftarrow \{s_i, y_i) | \text{emb}_i \in \mathcal{C}_k\}$; `//Build sub-dataset`
  $\theta_{\text{KCG}_k} \leftarrow \arg\min_\theta \sum_{(s,y)\in\mathcal{D}_k} \mathcal{L}_{\text{task}}(f_\theta(s), y)$; `//Train KCGs in parallel`

Stage *2: Joint Optimization of Experts & Routers*

Initialize routing parameters ϕ, load pre-trained $\{\theta_{\text{KCE}_i}\}_{i=1}^{N\times M}$
while *not converged* **do**
  for *batch* $\mathcal{B} \subset \mathcal{D}$ **do**
    foreach $(s, y) \in \mathcal{B}$ **do**
      $g \leftarrow \text{Router}_\phi(s)$; `//Routing weights`
      $\hat{y} \leftarrow W_0 s + \sum_{i=1}^{N\times M} g_i f_{\theta_i}(s)$; `//Weighted combination`
    Compute $\mathcal{L}_{\text{task}} = \ell(\hat{y}, y)$, $\mathcal{L}_{\text{aux}}$, $\mathcal{L}_{\text{diverse}}$; `//Compute loss`
    Update $\theta, \phi \leftarrow \theta, \phi - \eta\nabla(\mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{aux}} + \beta\mathcal{L}_{\text{diverse}})$

G Limitations and Future work

In this section, we discuss the potential limitations of our proposed method HiMoLE. Firstly, has shown effectiveness in addressing simple OOD scenarios, it still struggles to deal with hard OOD samples(e.g., in biomedical NER, it fails to outperform the base model on the OOD dataset). Future

Table 9: Inference Latency vs Performance across mixture of LoRA experts methods. The reported performance is evaluated using REM.

	HiMoLE	**MixLoRA**	**HydraLoRA**
%Param	1.05	1.05	0.65
Latency(s)	127	119	122.0
Performance(%)	**68.8**	66.5	67.2

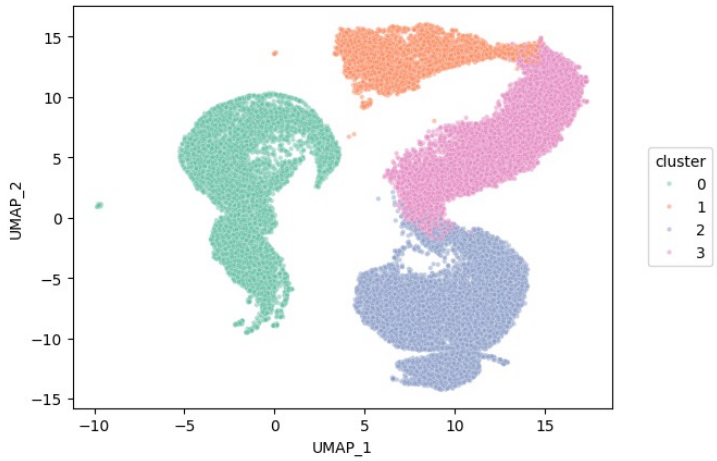
Table 10: Complete NER Results

Dataset	ID1			ID2			ID3			ID4			OOD		
Metric	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Base Model	59.1	60.6	59.0	42.4	52.0	40.4	51.3	53.4	51.8	52.3	57.5	51.4	65.8	63.1	76.4
LoRA	85.9	86.7	86.0	61.4	68.8	60.2	78.2	80.0	78.1	66.2	72.1	65.0	61.9	56.0	77.8
MixLoRA	87.4	88.4	87.3	60.6	64.2	59.9	73.4	75.8	73.0	77.3	79.5	77.0	61.0	59.1	73.7
HydraLoRA	87.8	88.4	87.8	63.2	65.2	62.3	75.1	75.9	74.4	78.2	79.4	77.2	62.9	59.4	75.5
HiMoLE	87.6	88.1	87.8	64.0	67.8	63.5	75.3	75.8	73.4	79.6	79.5	79.8	65.3	64.4	74.3

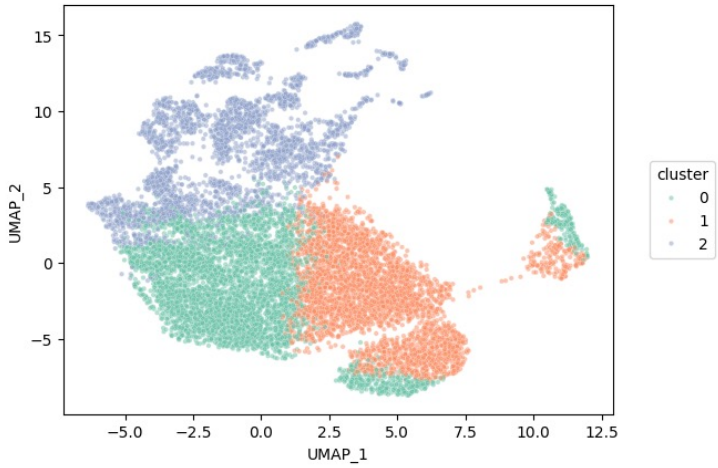
Table 11: Detailed Ablation Results

Dataset	ID1			ID2			ID3			ID4			OOD		
Metric	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
HiMoLE	87.6	88.1	87.8	64.0	67.8	63.5	75.3	75.8	73.4	79.6	79.5	79.8	65.3	64.4	74.3
-two stage training	77.7	80.5	77.0	32.1	45.2	30.5	54.3	62.0	53.2	60.8	66.0	60.1	53.5	57.7	59.0
-diverse loss	87.0	87.5	87.3	62.7	66.2	62.4	74.9	77.2	74.8	78.2	79.8	78.4	65.0	62.8	76.7

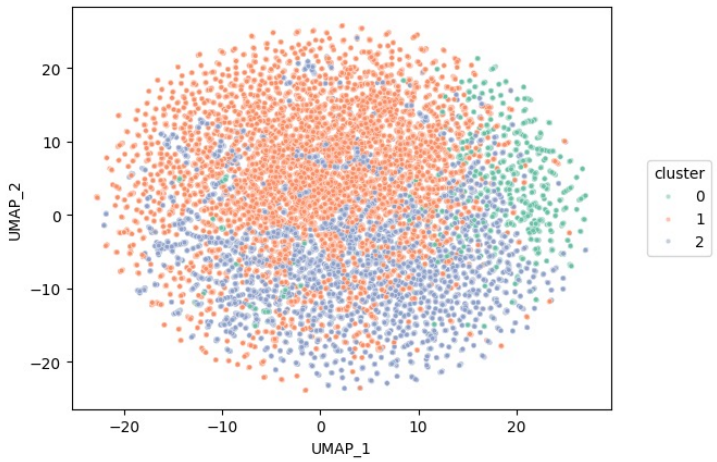
work will incorporate data augmentation to handle with hard OOD samples. Secondly, we constrain the model size to 13B and limit the number of LoRA experts to 16 due to resource and time limitations. As expert numbers scale, the interaction mode among experts would be more intricate and more sophisticated routing topologies like graph can be introduced to adapt to the nuanced patterns that emerge. Subsequent research be will conducted on the larger LLMs and more LoRA experts with more complex interaction mechanisms. This dual-axis expansion (model size + adaptive expert management) could unlock new robustness frontiers without proportional computational overhead.



(a) NER



(b) SA



(c) EQA

Figure 7: UMAP visualization of the datasets. Different colors correspond to different sub-datasets for KCGs initialization.

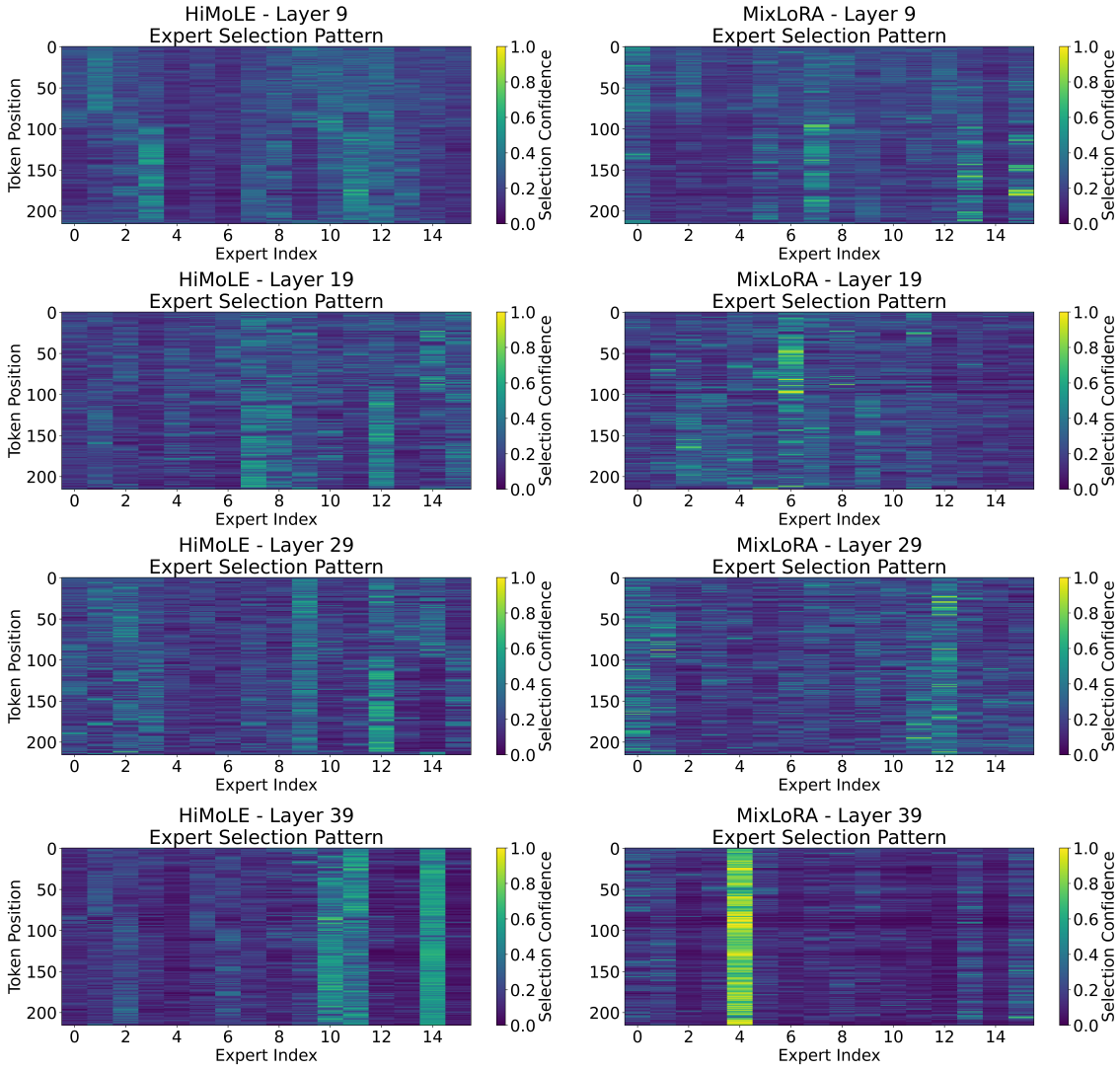


Figure 8: Expert Logits selection pattern.