Charting Empirical Laws for LLM Fine-Tuning in Scientific Multi-Discipline Learning

Anonymous ACL submission

Abstract

While large language models (LLMs) have achieved strong performance through fine-tuning within individual scientific domains, their learning dynamics in *multi-disciplinary* contexts remains poorly understood, despite the promise of improved generalization and broader applicability through cross-domain knowledge synergy. In this work, we present the first systematic study of multi-disciplinary LLM fine-tuning, constructing a five-discipline corpus and analyzing learning patterns of full fine-tuning, LoRA, LoRA-MoE, and LoRA compositions. Particularly, our study shows that multi-disciplinary learning is substantially more variable than single-discipline training and distills four consistent empirical laws: (1) **Balance-then-Diversity**: low-resource disciplines degrade performance unless mitigated via diversity-aware upsampling; (2) **Merge-then-Align**: restoring instruction-following ability is critical for cross-discipline synergy; (3) **Optimize-then-Scale**: parameter scaling offers limited gains without prior design optimization; and (4) **Share-then-Specialize**: asymmetric LoRA-MoE yields robust gains with minimal trainable parameters via shared low-rank projection. Together, these laws form a practical recipe for principled multi-discipline fine-tuning and provide actionable guidance for developing generalizable scientific LLMs.

1 Introduction

Artificial intelligence (AI) has made significant strides in scientific discovery, supporting tasks such as hypothesis generation, experimental design, and data analysis across disciplines (Xie et al., 2024; AI4Science and Quantum, 2023; Wang et al., 2023; Zhang et al., 2024c). Prior AI4Science works built specialized models to solve specific scientific problems (Lamurias et al., 2023; Pei et al., 2024; Laghuvarapu et al., 2024; Köhler et al., 2023; Hao et al., 2023). While these models show promising performance on specific tasks, they are often constrained by custom architectures and small, specialized labeled datasets, which restrict their generalization capabilities. To address these limitations, large-scale pre-training has emerged as a promising approach for building scientific foundation models (Ji et al., 2021; Chen et al., 2023a, 2024). The scale of pre-training data enables these models to surpass specialized counterparts on downstream tasks. More recently, inspired by the remarkable capabilities of large language models (LLMs) (Zhao et al., 2023), studies have adapted general-purpose LLMs to scientific tasks by fine-tuning them with scientific instructions (Yue et al., 2023; Fang et al., 2023; Han et al., 2023; Zhang et al., 2024b), effectively transferring problem-solving abilities to scientific problems and achieving impressive results.

Despite recent progress in AI4Science, developing models with multi-disciplinary knowledge remains underexplored. Such models can capture cross-domain synergies, support broader generalization, and enable complex scientific reasoning beyond the scope of single-discipline approaches. Recent efforts have begun to address this gap. UniSTD (Tang et al., 2025) develops a spatio-temporal foundation model to enhance performance across four spatio-temporal disciplines. SciReasoner (Wang et al., 2025) unifies representations of proteins, nucleotides, molecules and materials via natural language, enabling reasoning-driven modeling for biology and chemistry. Instead of training from scratch, X-LoRA (Buehler and Buehler, 2024) combines low-rank adaptation adapters (LoRA) fine-tuned on individual disciplines through a mixture-of-experts framework, leveraging single-discipline knowledge for multi-disciplinary tasks. However, these approaches adopt disparate algorithmic design choices without systematically analyzing the differences between multi- and single-discipline learning, limiting their generalizability and making it challenging to extend their designs to other domains.

In this study, we systematically investigate learning dynamics across single- and multi-discipline settings, providing insights into how to develop more generalizable scientific LLMs. We curate a text-based multi-discipline corpus spanning five scientific domains: mathematics, chemistry, biology, medicine, and geography. Leveraging this corpus, we investigate four representative fine-tuning strategies: full fine-tuning (Zhou et al., 2023a), LoRA (Hu et al., 2022), LoRA-MoE (Zadouri et al., 2023), and LoRA composition (Buehler and Buehler, 2024), applied to Qwen2.5 7B Instruct (Yang et al., 2024) under varying data scales. Evaluation on in-domain benchmarks reveals that multi-discipline learning trajectories diverge substantially from those in single-discipline settings. In particular, disciplines with limited data experience greater instability, cross-discipline synergy is inconsistent, and multi-discipline fine-tuning can degrade discipline-specific performance in terms of average accuracy. We further analyze these phenomena to uncover the empirical regularities that govern multi-discipline fine-tuning and distill them into four laws that together form a practical recipe for building robust scientific LLMs:

- **Balance-then-Diversity:** Low-resource disciplines disproportionately hinder multi-discipline learning, leading to reduced accuracy and increased variance. We show that diversity-aware up-sampling, rather than naïve duplication, mitigates this issue by maintaining both balance and diversity in cross-discipline data contributions.

- **Merge-then-Align:** Instruction-following ability is a prerequisite for effective cross-discipline transfer. Multi-discipline fine-tuning alone degrades alignment, while mixing a modest amount of general instruction data restores instruction-following and unlocks synergy across disciplines.

- **Optimize-then-Scale:** Simply enlarging the number of trainable parameters provides negligible gains in multi-discipline settings. Performance improvements instead depend on careful model design and optimization choices, suggesting that scaling should follow—not precede—architectural optimization.

- **Share-then-Specialize:** LoRA-MoE architectures with asymmetric parameter sharing (e.g., shared A matrices) first promote effective cross-discipline knowledge sharing, and then enable stable expert specialization, yielding robust multi-discipline improvements comparable to full fine-tuning while using only a small fraction of the parameters.

2 Related Work

AI4Science applies artificial intelligence to advance scientific research through both specialized task models (Lamurias et al., 2023; Laghuvarapu et al., 2024; Hao et al., 2023) and large-scale discipline-wise scientific foundation models (Ji et al., 2021; Chen et al., 2023a, 2024; Price et al., 2025; Zeni et al., 2025). Large language models (LLMs) have also emerged as scientific foundation models, leveraging their advanced reasoning and language understanding capabilities to support scientific problem solving. Various disciplines have developed their own domain-specific LLMs, including models for mathematics (Cobbe et al., 2021; Toshniwal et al., 2024), biology (Fang et al., 2023), medicine (Li et al., 2023, 2024a), chemistry (Zhang et al., 2024b; Li et al., 2024b), and geoscience (Deng et al., 2024; Ma et al., 2024). However, these models were typically trained and evaluated within isolated domains, neglecting the inherently interdisciplinary nature of many scientific problems. Recent studies (Tang et al., 2025; Xia et al., 2025; Wang et al., 2025; Buehler and Buehler, 2024) attempted multi-disciplinary modeling, yet these approaches adopt diverse strategies and lack systematic investigation into the learning dynamics and challenges inherent to multi-disciplinary fine-tuning. In this work, we aim to bridge this gap by providing a comprehensive analysis of multi-disciplinary fine-tuning paradigms.

LLM Fine-tuning adapts large language models to specific tasks, with instruction tuning (Wei et al., 2021; Ouyang et al., 2022; Zhang et al., 2023c) widely used to align models via curated instruction-output pairs. As model sizes grow, the high cost of full fine-tuning has driven the development of parameter-efficient fine-tuning (PEFT) methods that update only a small subset of parameters. These methods introduces lightweight trainable modules (Houlsby et al., 2019; Liu et al., 2022; Lester et al., 2021; Li and Liang, 2021) or selectively update existing parameters (Zaken et al., 2021). A prominent PEFT approach is low-rank adaptation (LoRA) (Hu et al., 2022) and its variants

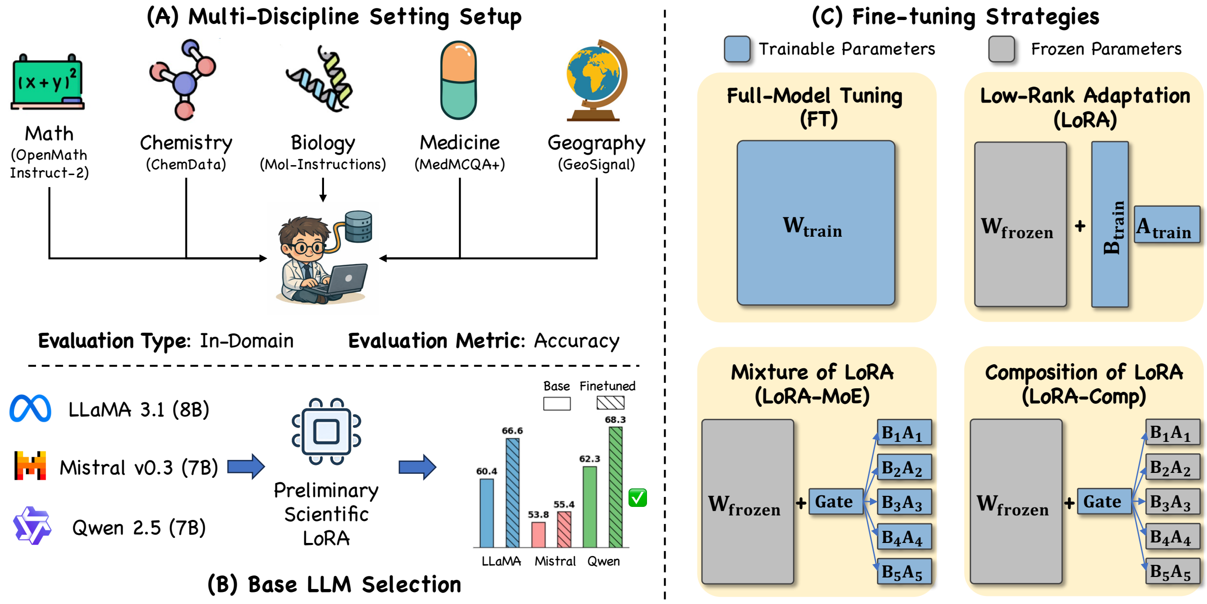


Figure 1: **Overview of the Empirical Study Methodology**. (A) Five scientific disciplines included in our multi-discipline setting. (B) Selection of Qwen 2.5 based on superior pre-na dn post-tuning scientific performance. (C) Fine-tuning strategies investigated by distinguishing trainable and frozen parameters.

(Zhang et al., 2023b; Liu et al., 2024b), which reduce parameter cost by factorizing weight updates. In this work, we analyze both full fine-tuning and PEFT in the context of scientific LLMs.

Multi-task Learning (MTL) improves performance by enabling knowledge sharing across tasks, closely aligning with multi-discipline learning. Traditional MTL shares parameters (Liu et al., 2019; Zhao et al., 2024; Ruder et al., 2019) or balances task gradients (Chen et al., 2018; Liu et al., 2023; Senushkin et al., 2023). In the LLM context, MTL has been adapted using parameter-efficient fine-tuning (PEFT). One line of research composes task-specific LoRA modules (Zhang et al., 2023a; Ostapenko et al., 2024; Prabhakar et al., 2024; Buehler and Buehler, 2024), using arithmetic operations (Zhang et al., 2023a) or dynamic gating (Buehler and Buehler, 2024). Another line integrates LoRA with mixture-of-experts (MoE) architectures, where LoRA-MoE (Zadouri et al., 2023; Luo et al., 2024) achieves competitive performance. Extensions include asymmetric sharing in HydraLoRA (Tian et al., 2024) and hierarchical routing in HMoRA (Liao et al., 2025). We focus on two representative PEFT-based MTL paradigms: LoRA composition and LoRA-MoE.

3 Experiment Setup

Multi-Discipline Setting. To emulate the diversity in quantity and format characteristic of real-world heterogeneous multi-disciplinary data, we construct a multi-discipline training dataset and evaluation benchmarks by aggregating existing open-source datasets from individual disciplines (Figure 1 and Table 1). Specifically, our study encompasses five disciplines: math, chemistry, biology, medicine, and geography. For mathematics, we adopt a 2M-sample subset from OpenMathInstruct-2 (Toshniwal et al., 2024), which is augmented upon MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) to support large-scale supervised fine-tuning for mathematical reasoning. For chemistry, we use ChemData (Zhang et al., 2024b) containing 700K instruction-based QA pairs on molecular properties, reactions, and related tasks. For biology, we select the biomolecular text subset of Mol-Instructions (Fang et al., 2023) for training. For medicine, we compile a large-scale dataset by aggregating samples from MedMCQA (Pal et al., 2022), MedAlpaca (Han et al., 2023), ChatDoctor (Li et al., 2023), MedInstruct-52K (Zhang et al., 2023d), and others (Appendix A). For geography, we adopt GeoSignal (Deng et al., 2024), which includes QA covering geoscientific topics.

Evaluation Setting To assess discipline-specific knowledge acquired through scientific fine-tuning, we adopt in-domain evaluation benchmarks tailored to each discipline. Specifically, mathematics is evaluated using GSM8K test set (Cobbe et al., 2021), while chemistry is assessed with

Discipline	Data Scale		Source (Train / Eval)	Text Statistics	
	Samples	%		Avg. Length	Unique Tokens
Math	2,000,000	60.7	Toshniwal et al. (2024) / Cobbe et al. (2021)	267.74	1,995,829
Chemistry	713,218	21.6	Zhang et al. (2024b) / Zhang et al. (2024b)	57.66	1,232,633
Biology	51,427	1.6	Fang et al. (2023) / Fang et al. (2023)	50.61	22,574
Medicine	490,766	14.9	Pal et al. (2022), etc. / Pal et al. (2022)	139.17	529,564
Geography	39,749	1.2	Deng et al. (2024) / Deng et al. (2024)	114.20	60,209
Total	**3,295,160**	**100.0**	–	**188.17**	**3,840,809**

Table 1: **Multi-discipline corpus statistics.** Average length is measured in words per sample and unique tokens are defined as tokens with fewer than 10 co-occurrences across disciplines.

ChemBench (Zhang et al., 2024b), covering nine representative chemical tasks. For biology, we employ the biomolecular multiple-choice subset of Mol-Instruction test set (Fang et al., 2023). Medicine is evaluated using MedMCQA test set (Pal et al., 2022), and geography is assessed using the multiple-choice subset of GeoBench (Deng et al., 2024) to ensure consistent evaluation formats. We conduct all evaluations using the lm-evaluation-harness framework (Gao et al., 2024), employing accuracy as the primary metric.

Fine-tuning Setting We study the learning patterns for the full fine-tuning and three parameter efficient tuning methods (PEFT) (Figure 1):

- **Full-Model Tuning (FT)**: This standard approach updates all parameters of the language model. Specifically, each forward operation $\mathbf{x}_{\text{out}} = \mathbf{W}\mathbf{x}_{\text{in}}$ will have the pre-trained weight matrix $\mathbf{W}$ fully trainable.

- **Low-Rank Adaptation (LoRA)**: LoRA introduces trainable low-rank matrices $\mathbf{B} \in \mathbb{R}^{r \times d_{\text{out}}}$ and $\mathbf{A} \in \mathbb{R}^{d_{\text{in}} \times r}$ to adapt pre-trained weights $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, where $r \ll \min(d_{\text{in}}, d_{\text{out}})$. During fine-tuning, the model uses $\mathbf{W} + \mathbf{BA}$, updating only $\mathbf{B}$ and $\mathbf{A}$.

- **Mixture of Low-Rank Adaptation (LoRA-MoE):** LoRA-MoE extends LoRA by introducing a set of trainable low-rank matrices $\{\mathbf{B}_i\}_{i=0}^{k}$ and $\{\mathbf{A}_i\}_{i=1}^{k}$, where k denotes the number of experts. The adapted weight is computed as $\mathbf{W} + \sum_{i=1}^{k} \omega_i \cdot \mathbf{B}_i \mathbf{A}_i$, with gating weights ω_i dynamically predicted by a gating network. The gating network is typically implemented as a lightweight MLP, $\omega = f_{\text{MLP}}(\mathbf{x}_{\text{in}})$, and is jointly optimized with the low-rank matrices.

- **Composition of Low-Rank Adaptation (LoRA-Comp):** LoRA-Comp adopts a similar formulation to LoRA-MoE, where the adapted weight is composed of multiple expert LoRA modules. Differently, the expert LoRA modules are pre-trained on individual discipline-specific corpora, and during multi-discipline fine-tuning, only the gating network is optimized to dynamically combine the pre-trained single-discipline LoRA experts.

For full-model tuning, we fine-tune all model parameters using a learning rate of 7×10^{-6}, weight decay of 0.1, and a warm-up ratio of 0.05 for 1 epoch. For PEFT methods, we adopt a learning rate of 1×10^{-4}, weight decay of 0.01, and a warm-up ratio of 0.1, with training conducted for 1 epoch. We set LoRA rank as 16 and number of experts for LoRA-MoE and LoRA-Comp match the number of 5 disciplines. See Appendix A for more details.

Base LLMs We compare three 7B-scale instruction-tuned LLMs—LLaMA 3.1 (Grattafiori et al., 2024), Mistral v0.3 (MistralAI, 2024), and Qwen2.5 (Yang et al., 2024) as base models for preliminary scientific fine-tuning using LoRA (Hu et al., 2022) under identical settings. As shown in Figure 1 (B), scientific fine-tuning performance is highly sensitive to the choice of the pre-trained base model and does not consistently yield gains across tasks. Among the evaluated models, Qwen2.5 7B Instruct achieves the strongest pre-trained performance and the largest improvements after fine-tuning. We therefore adopt Qwen2.5 7B Instruct for all subsequent experiments. More evaluation results can be found in Appendix C.

4 Empirical Laws for Multi-Discipline LLM Fine-Tuning

To compare single- and multi-discipline learning, we analyze their learning patterns across fine-tuning strategies and data scales (Figure 2). In the single-discipline setting, performance scales

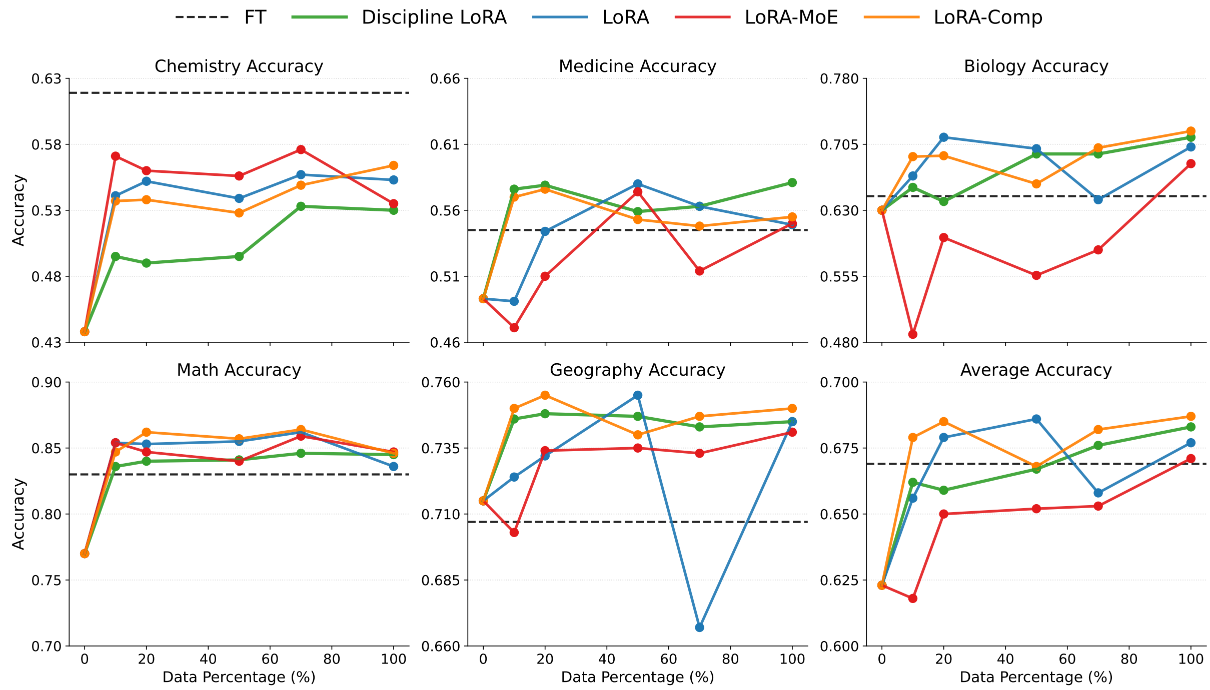


Figure 2: **Learning Dynamics of Fine-tuning Strategies.** Multi-disciplinary fine-tuning is less effective and stable than single-discipline tuning, highlighting trade-offs between generality, performance, and stability across methods.

predictably with data size where modest amounts already yield strong results with continued improvements as data increases (Zhou et al., 2023a; Zhang et al., 2024a). In contrast, multi-discipline learning exhibits markedly less stable scaling behavior.

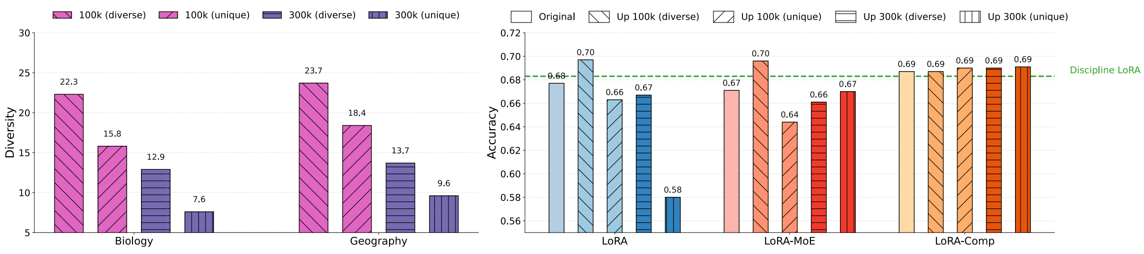
Fine-tuning perspective. Scaling curves exhibit greater variance as data volume changes, a pattern that is particularly pronounced in data-scarce fields such as medicine, biology, and geography. While certain domains (e.g., chemistry) sometimes benefit from cross-disciplinary data, these gains are inconsistent. On average, multi-disciplinary fine-tuning tends to degrade discipline-specific performance, as reflected in lower average accuracy compared to single-discipline fine-tuning. Furthermore, directly applying full-model fine-tuning (FT) on multi-discipline data results in sub-optimal outcomes. In this setting, conflicting learning signals across disciplines and overfitting to dominant data sources could occur, which not only undermines generalization but also leads to performance inferior to PEFT methods, despite FT involving substantially more trainable parameters.

Architectural perspective. Fine-tuning strategies also differ in stability. LoRA-Comp, training a lightweight router on top of discipline-specific LoRA adapters, yields the most stable trajectories and most closely resembles single-discipline behavior. Nonetheless, the router's limited capacity constrains its ability to model cross-disciplinary interactions, producing lower peak performance than LoRA and LoRA-MoE models trained from scratch on aggregated data. Because they are initialized from scratch in the multi-disciplinary regime, LoRA and especially LoRA-MoE exhibit greater performance variability and generally underperform their single-discipline counterparts.

4.1 Balance-then-Diversity Empirical Law

Balance-then-Diversity. Multi-discipline learning first requires balancing data-scarce disciplines to reduce instability, and further gains depend on preserving sample diversity rather than simply increasing data volume.

We begin by examining the impact of data-scarce disciplines on multi-discipline learning. As shown in Table 1, dataset sizes vary significantly across disciplines given its real-world heterogeneity. Figure 2 reveals that geography and biology have high variance in performance, in contrast to the stable trends observed in chemistry and mathematics. We hypothesize that this instability arises from difficulty in learning fine-grained knowledge of low-resource domains. Given the cost of acquiring additional scientific data, we evaluate two straightfor-



(A) Corpus Diversity (B) Average Accuracy of Upsampling Data-Scarce Disciplines

Figure 3: Analysis of how data-scarce discipline affects the multi-disciplinary learning.

ward upsampling methods to expand scarce data: diverse upsampling, which randomly duplicates samples, and unique upsampling, which selectively duplicates samples containing discipline-specific tokens. These strategies are applied to biology and geography, scaled to 100K and 300K samples, while other disciplines remain unchanged.

As shown in Figure 3(B), upsampling data-scarce disciplines improves overall performance, indicating that low-resource disciplines hinder multi-discipline learning. However, gains do not scale consistently with data volume, and diversity-aware upsampling consistently outperforms unique-token-based strategies. This effect stems from reduced sample diversity under large-scale and unique-token-based upsampling, which induces overfitting to repeated prompts. Consistently, the unique n-gram diversity metric (Song et al., 2024) (Figure 3(A)) shows that diversity degradation strongly correlates with performance drops. LoRA-Comp remains relatively stable, likely due to its limited trainable parameters constraining overfitting. Overall, upsampling low-resource disciplines benefits both upsampled and non-upsampled domains (see Appendix B), highlighting a balance-then-diversity principle: effective multi-discipline learning requires first correcting data imbalance and then preserving sample diversity.

4.2 Merge-then-Align Empirical Law

Merge-then-Align. Multi-discipline training should first merge heterogeneous scientific data and then realign the model with general instruction-following data to restore alignment and improve overall multi-discipline synergy.

Multi-discipline fine-tuning requires adapting pre-trained LLMs to heterogeneous scientific corpora, which can degrade their instruction-following capability for effective task generalization. We hypothesize that this degradation contributes to the suboptimal learning patterns observed in multi-discipline settings. To evaluate this, we assess models on the IFEval benchmark (Zhou et al., 2023b), which measures instruction-following ability. As shown in Table 2, all fine-tuned models underperform relative to the base LLM, confirming a decline in alignment quality.

Methods	**IF** (%)	**Science**	**IFEVal**
Qwen 2.5	-	0.623	0.808
Dis. LoRA	-	0.683	-
LoRA	0	0.677	0.725
	70	0.708	0.742
	100	**0.712**	**0.748**
LoRA-MoE	0	0.671	0.763
	70	**0.712**	0.730
	100	0.708	**0.764**
LoRA-Comp	0	0.687	0.769
	70	**0.726**	**0.783**
	100	0.692	0.773

Table 2: Model performance mixing additional IF data.

To address this issue, we augment the multi-discipline training corpus with general-domain instruction-following data from InternLM (Cai et al., 2024), and fine-tune new models on this mixed dataset with varying percentage. This consistently improves instruction-following performance and, importantly, yields stronger results on multi-disciplinary evaluation benchmarks. Overall, these results support a merge-then-align principle: first merging diverse domain data to acquire broad knowledge, followed by explicit alignment to restore and strengthen instruction-following capability for effective multi-discipline generalization.

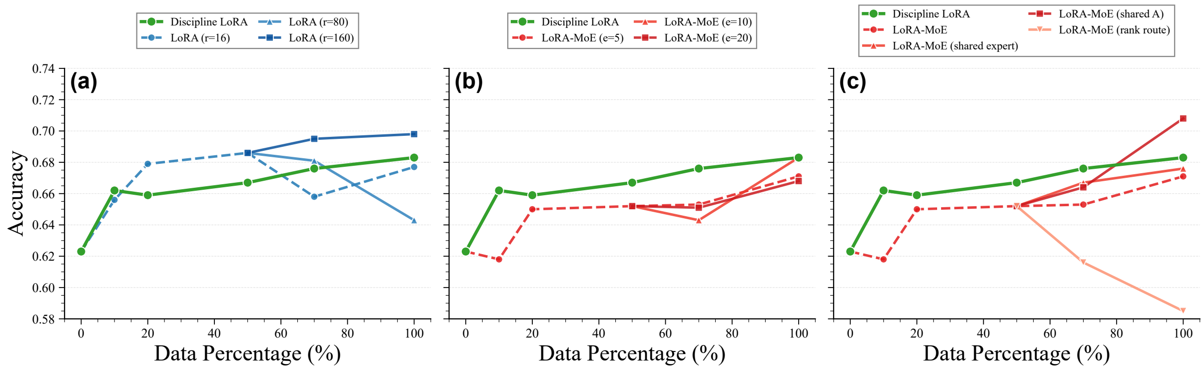


Figure 4: (a,b) Scaling LoRA rank and LoRA-MoE experts; (c) Different MoE designs.

4.3 Optimize-then-Scale Empirical Law

Optimize-then-Scale. The model design choices should be prioritized, as scaling simply yields limited or even negative returns.

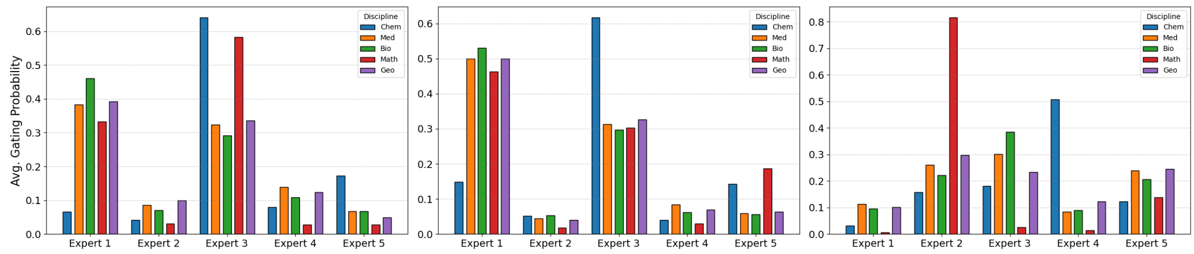
Compared to single-discipline fine-tuning, the multi-discipline setting requires the model to process and integrate a broader and more diverse set of data, potentially increasing optimization difficulty. We hypothesize that part of the observed instability in multi-discipline fine-tuning may stem from insufficient trainable parameter capacity. To examine this hypothesis, we systematically scale the trainable parameters of LoRA and LoRA-MoE by increasing their ranks and numbers of experts, respectively. However, as shown in Figure 4 (a,b), enlarging the trainable parameter size yields only marginal improvements in multi-discipline performance and, in some cases, leads to inverse scaling effects (e.g., LoRA with rank 80). These results are consistent with prior observations (Zhang et al., 2024a) that scaling PEFT capacity alone does not guarantee improved outcomes. Overall, these results indicate that multi-discipline fine-tuning is constrained less by parameter capacity than by optimization and architectural factors, underscoring an optimize-then-scale principle: robust optimization strategies must be established before increasing model or adaptation scale can be effective.

4.4 Share-then-Specialize Empirical Law

Share-then-Specialize. Multi-discipline performance improves robustly when knowledge sharing is enforced latently through asymmetric architectures before enabling selective knowledge specialization.

Despite the success of LoRA-MoE in general MTL (Tian et al., 2024; Lin et al., 2024; Chen et al., 2023b), it surprisingly underperforms in the multi-discipline setting (Figure 2). Given the numerous LoRA-MoE variants proposed to enhance performance, we hypothesize that improved architectural design could better support multi-discipline learning. We therefore investigate two key components: expert design and gating network design. For expert design, inspired by DeepSeek-MoE (Liu et al., 2024a), we introduce an additional shared LoRA expert alongside the five discipline-specific experts to promote shared and specialized knowledge separation. Additionally, following HydraLoRA (Tian et al., 2024), we evaluate an asymmetric design that shares the LoRA A matrix while routing only the B matrices, encouraging cross-discipline commonality. For the gating network, we specifically explore rank-wise routing (Ning et al., 2024), where individual routing weights are assigned to each rank component of the LoRA experts for more nuanced discipline-specific patterns.

Figure 4(c) shows that the shared-expert design yields only marginal gains, whereas sharing the A matrix leads to substantial improvements. Expert activation visualizations in Figure 5 explain this gap. While shared experts in standard MoE (Liu et al., 2024a) are meant to capture common knowledge, in LoRA-MoE they dilute specialization, as indicated by near-uniform activations across disciplines (except chemistry). In contrast, sharing the A matrix aligns LoRA experts in a common latent space, enabling clearer specialization (e.g., expert 2 for mathematics and expert 4 for chemistry). Additionally, rank-wise routing, despite finer-grained control, neither improves performance nor remains stable in data-scarce disciplines (Figure 4(c); Appendix B), likely due to excessive routing flexibility



(a) Original (b) W/ Shared Expert (c) W/ Shared A Matrix

Figure 5: Mean expert routing probabilities computed from intermediate activations at the final feedforward layer.

Method	**In-Domain**							**Out-Domain**	**Param.** (%)	**GPU Hours**
	Chem	Med	Bio	Math	Geo	Avg	$\Delta m(\%)$	MMLU		
QWen 2.5 7B Instruct	0.509	0.493	0.630	0.770	0.715	0.623	-	0.688	0	-
Discipline LoRA	0.530	0.581	0.713	0.845	0.745	0.683	0	-	0.264×5	-
FT (ori.)	0.619	0.545	0.646	0.830	0.707	0.669	-1.135	0.719	100	1920
FT (tuned)	**0.649**	0.588	0.821	0.831	0.743	0.726	+7.376	0.707	100	2080
LoRA-MoE (ori.)	0.535	0.550	0.683	**0.847**	0.741	0.671	-1.780	0.732	1.315	344
LoRA-MoE (tuned)	0.613	**0.603**	**0.823**	0.839	**0.760**	**0.728**	+7.236	**0.739**	0.902	328

Table 3: Results comparison of our recipe-tuned tuned multi-discipline fine-tuning methods.

destabilizing expert selection. Overall, these findings highlight a share-then-specialize law, where early parameter sharing provides a common representation foundation that subsequently promotes clearer expert specialization and improved multidisciplinary performance.

5 Final Multi-Discipline Learning Recipe

Based on our previous analysis, we formulate the following multi-discipline fine-tuning recipe:

- **Data Scarcity Mitigation:** To address performance degradation caused by data-scarce disciplines, we upsample the biology and geography training corpora to 100K samples each, following the diversity-aware upsampling strategy shown to be most effective.

- **Enhancing Instruction Following:** Given its substantial contribution to multi-disciplinary performance, we incorporate the general-domain instruction-following enhancement data from InternLM (Cai et al., 2024) into the multi-discipline training corpus.

- **Fine-Tuning Strategy:** We adopt full-model tuning and LoRA-MoE as the final fine-tuning approaches. For LoRA-MoE, we apply the shared LoRA A matrix design with token-level and layer-wise routing.

All models are fine-tuned using the setup in Section 3, with results summarized in Table 3. In addition to discipline-wise accuracy, we report the delta performance Δm (Agiza et al., 2024), measuring the average per-task performance drop relative to the single-discipline LoRA baseline, along with the percentage of trainable parameters. As shown in Table 3, both full fine-tuning and LoRA-MoE trained with the proposed recipe achieve substantial improvements over the original setting, validating the effectiveness of the combined strategies. Notably, LoRA-MoE largely closes the gap to full fine-tuning, attaining comparable in-domain accuracy while using only a small fraction of trainable parameters. Moreover, LoRA-MoE exhibits stronger out-of-domain generalization on MMLU (Hendrycks et al., 2020), achieving these gains with only 15% of the GPU hours required by full fine-tuning.

6 Conclusion

We present the first systematic analysis of multi-disciplinary LLM fine-tuning, revealing significantly higher training variability than single-discipline settings and distilling four empirical laws which are Balance-then-Diversity, Merge-then-Align, Optimize-then-Scale, and Share-then-Specialize, that govern effective cross-domain learning. Together, these principles provide a practical recipe for building robust and generalizable multi-discipline scientific LLMs.

Limitations

While our study provides comprehensive insights into multi-discipline fine-tuning of LLMs, several limitations should be acknowledged. Firstly, our multi-discipline corpus is constructed by aggregating existing single-discipline datasets. However, these datasets may not fully capture the diversity or specific characteristics of their respective disciplines, particularly in fields with complex or nuanced knowledge structures. As such, the findings and observed learning patterns may vary when applied to other disciplines or to datasets with different properties. Future work should explore broader discipline coverage and incorporate more diverse datasets to enhance representativeness and generalizability. Due to the high computational cost associated with full-model fine-tuning, our analysis is limited to key scenarios, specifically focusing on the multi-discipline setting and its optimized configuration. We do not provide scaling analyses for full fine-tuning across all data volumes, which may limit our understanding of its behavior in lower-resource or discipline-specific settings. Further investigations using more efficient full fine-tuning approximations or resource-scalable evaluations are needed. Lastly, LLM fine-tuning is known to exhibit performance variability due to optimization instability and data randomness. Although we conducted multiple runs for cases showing suspicious outliers to mitigate this concern, inherent variability remains a potential limitation. More robust methods, such as uncertainty estimation or larger-scale repeated trials, could be explored in future studies to further strengthen the reliability of the findings.

References

Ahmed Agiza, Marina Neseem, and Sherief Reda. 2024. Mtlora: Low-rank adaptation approach for efficient multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16196–16205.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20:1–23.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Eric L Buehler and Markus J Buehler. 2024. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2).

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, and 1 others. 2023a. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, and 1 others. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. 2023b. Octavius: Mitigating task interference in mllms via lora-moe. *arXiv preprint arXiv:2311.02684*.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

CogStack. 2023. `OpenGPT`. https://https://github.com/CogStack/OpenGPT.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, and 1 others. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. 2023. Gnot: A general neural operator transformer for operator learning. In *International Conference on Machine Learning*, pages 12556–12569. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Jonas Köhler, Michele Invernizzi, Pim De Haan, and Frank Noé. 2023. Rigid body flows for sampling molecular crystal structures. In *International Conference on Machine Learning*, pages 17301–17326. PMLR.

Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. 2024. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36.

Andre Lamurias, Alessandro Tibo, Katja Hose, Mads Albertsen, and Thomas Dyhre Nielsen. 2023. Metagenomic binning using connectivity-constrained variational autoencoders. In *International Conference on Machine Learning*, pages 18471–18481. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024b. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Mengqi Liao, Wei Chen, Junfeng Shen, Shengnan Guo, and Huaiyu Wan. 2025. Hmora: Making llms more effective with hierarchical mixture of lora experts. In *The Thirteenth International Conference on Learning Representations*.

Tianwei Lin, Jiang Liu, Wenqiao Zhang, Zhaocheng Li, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Hao Jiang, and 1 others. 2024.

Teamlora: Boosting low-rank adaptation with expert collaboration and competition. *arXiv preprint arXiv:2408.09856*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. 2023. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36:57226–57243.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

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

Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. 2024. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *60th Annual Meeting of the Association for Computational Linguistics*.

MistralAI. 2024. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Lin Ning, Harsh Lara, Meiqi Guo, and Abhinav Rastogi. 2024. Mode: Effective multi-task parameter efficient fine-tuning with a mixture of dyadic experts. *arXiv preprint arXiv:2408.01505*.

Oleksiy Ostapenko, Zhan Su, Edoardo Maria Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas Caccia, and Alessandro Sordoni. 2024. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Qizhi Pei, Kaiyuan Gao, Lijun Wu, Jinhua Zhu, Yingce Xia, Shufang Xie, Tao Qin, Kun He, Tie-Yan Liu, and Rui Yan. 2024. Fabind: Fast and accurate protein-ligand binding. *Advances in Neural Information Processing Systems*, 36.

Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. 2024. Lora soups: Merging loras for practical skill composition tasks. *arXiv preprint arXiv:2410.13025*.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, and 1 others. 2025. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4822–4829.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. 2023. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024. Scaling data diversity for fine-tuning language models in human alignment. *arXiv preprint arXiv:2403.11124*.

Chen Tang, Xinzhu Ma, Encheng Su, Xiufeng Song, Xiaohong Liu, Wei-Hong Li, Lei Bai, Wanli Ouyang, and Xiangyu Yue. 2025. Unistd: Towards unified spatio-temporal learning across diverse disciplines. *arXiv preprint arXiv:2503.20748*.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37:9565–9584.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Yizhou Wang, Chen Tang, Han Deng, Jiabei Xiao, Jiaqi Liu, Jianyu Wu, Jun Yao, Pengze Li, Encheng Su, Lintao Wang, and 1 others. 2025. Scireasoner: Laying the scientific reasoning ground across disciplines. *arXiv preprint arXiv:2509.21320*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: system demonstrations*, pages 38–45.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, and 1 others. 2025. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*.

Yutong Xie, Yijun Pan, Hua Xu, and Qiaozhu Mei. 2024. Bridging ai and science: Implications from a large-scale literature analysis of ai4science. *arXiv preprint arXiv:2412.09628*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2023. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, and 1 others. 2025. A generative model for inorganic materials design. *Nature*, pages 1–3.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024a. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, and 1 others. 2024b. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Jinghan Zhang, Junteng Liu, Junxian He, and 1 others. 2023a. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023c. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023d. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024c. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng Zhao, Yftah Ziser, and Shay B Cohen. 2024. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Additional Experiment Setup

Multi-Discipline Setting To expand the medicine discipline dataset and align it with the scale and recency of other disciplines in our multi-discipline setup, we aggregate data from a range of existing sources (Vilares and Gómez-Rodríguez, 2019; Han et al., 2023; Pal et al., 2022; Jin et al., 2021; Ben Abacha and Demner-Fushman, 2019; Mishra et al., 2022; CogStack, 2023; Jin et al., 2019; Li et al., 2023; Ben Abacha et al., 2019; Bodenreider, 2004; Zhang et al., 2023d). The resulting dataset includes:

- 2,657 multiple-choice questions from HeadQA (Vilares and Gómez-Rodríguez, 2019)
- 120,765 multiple-choice questions from MedMCQA (Pal et al., 2022)
- 10,178 multiple-choice questions from MedQA (Jin et al., 2021)
- 119,486 real doctor-patient conversations from ChatDoctor (Li et al., 2023)
- 16,407 QA pairs from MedQuAD (Ben Abacha and Demner-Fushman, 2019)
- 500 biomedical questions from PubMedQA (Jin et al., 2019)
- 689 consumer drug-related questions from MedicationQA (Ben Abacha et al., 2019)
- 79,245 medical QA pairs derived from UMLS (Bodenreider, 2004)
- 24,665 QA pairs from OpenGPT (CogStack, 2023)
- 52,002 machine-generated QA pairs from AlpacaRE (Zhang et al., 2023d)
- A subset of Medical Meadow dataset(Han et al., 2023), including health advice, flashcards, PubMed-based causal data
- Medical-related tasks from the Natural Instructions v2.8 dataset (Mishra et al., 2022), including tasks 179, 181, 1369, 1447, 1449, 1485, 1487, 1495, and 1645

The final aggregation comprises 490,766 samples, ensuring sufficient data diversity and scale to support robust multi-discipline fine-tuning in the medical domain.

Evaluation Setting We evaluate models using the lm-evaluation-harness framework (Gao et al., 2024) with two modes. For math, which involves direct-answer questions, the model generates output until the end-of-sequence token, and the final answer is extracted and compared to the reference. For the remaining disciplines, which use multiple-choice formats, the model selects the option with the highest log-likelihood (Hendrycks et al., 2020), which is then compared to the ground truth.

Finetuning Setting We provide additional details of our fine-tuning setup. For full-model tuning, all parameters are updated using a learning rate of 7×10^{-6}, weight decay of 0.1, and a warm-up ratio of 0.05, trained for 1 epoch. For PEFT methods, we use a learning rate of 1×10^{-4}, weight decay of 0.01, and a warm-up ratio of 0.1, also trained for 1 epoch. Each LoRA module is configured with rank 16 and a scaling factor of 32, applied to all attention and feedforward layers. The gating network in LoRA-MoE and LoRA-Comp is implemented as a single-layer MLP that predicts layer-wise gating weights based on the layer input. Both LoRA-MoE and LoRA-Comp use five experts, aiming for one per discipline, with each expert configured identically to the single LoRA setup. All models are trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with a linear learning rate scheduler, an effective batch size of 128, and are implemented using HuggingFace Transformers (Wolf et al., 2020) and DeepSpeed (Rasley et al., 2020) on 32 NVIDIA A800 GPUs.

B Additional Results

In addition to the visualizations in the main manuscript, we present comprehensive evaluation results in the following tables: Table 4 (fine-tuning methods across all data scales), Table 5 (upsampling strategies), Table 6 (instruction-following enhancement), Table 7 (trainable parameter scaling), and Table 8 (model architectual design).

C Base Model Selection

In Figure 6, we present the effects of preliminary scientific LoRA fine-tuning across different base models(LLaMA 3.1 7B, Mistral v0.3 7B, Qwen 2.5 7B) and scientific disciplines(chemistry, medicine, biology, math, geography). The fine-tuned models achieve higher average performance than their counterparts. Performance differs across base models, with Qwen 2.5 7B exhibiting higher overall performance and improvements across disciplines.

Method	**Data**	**Chem**	**Med**	**Bio**	**Math**	**Geo**	**Avg**
Discipline LoRA	10 %	0.495	**0.576**	0.656	0.836	0.746	0.662
LoRA		0.541	0.491	0.669	**0.854**	0.724	0.656
LoRA-MoE		**0.571**	0.471	0.489	**0.854**	0.703	0.618
LoRA-Comp		0.537	0.570	**0.691**	0.847	**0.750**	**0.679**
Discipline LoRA	20 %	0.490	**0.579**	0.640	0.840	0.748	0.659
LoRA		0.552	0.544	**0.713**	0.853	0.732	0.679
LoRA-MoE		**0.560**	0.510	0.599	0.847	0.734	0.650
LoRA-Comp		0.538	0.576	0.692	**0.862**	**0.755**	**0.685**
Discipline LoRA	50 %	0.495	0.559	0.694	0.841	0.747	0.667
LoRA		0.539	**0.580**	**0.700**	0.855	**0.755**	**0.686**
LoRA-MoE		**0.556**	0.574	0.556	0.840	0.735	0.652
LoRA-Comp		0.528	0.553	0.660	**0.857**	0.740	0.668
Discipline LoRA	70 %	0.533	**0.563**	0.694	0.846	0.743	0.676
LoRA		0.557	**0.563**	0.642	0.862	0.667	0.658
LoRA-MoE		**0.576**	0.514	0.585	0.859	0.733	0.653
LoRA-Comp		0.549	0.548	**0.701**	**0.864**	**0.747**	**0.682**
Discipline LoRA	100 %	0.530	**0.581**	0.713	0.845	0.745	0.683
FT		**0.619**	0.545	0.646	0.830	0.707	0.669
LoRA		0.553	0.549	0.702	0.836	0.745	0.677
LoRA-MoE		0.535	0.550	0.683	**0.847**	0.741	0.671
LoRA-Comp		0.564	0.555	**0.720**	0.846	**0.750**	**0.687**

Table 4: Full evaluation results for all fine-tuning methods under all tested data scales. For each data scale, we highlight the **best value**.

Method	**Upsample**	**Chem**	**Med**	**Bio**	**Math**	**Geo**	**Avg**
LoRA	NA	0.553	0.549	0.702	0.836	0.745	0.677
LoRA-MoE		0.535	0.550	0.683	**0.847**	0.741	0.671
LoRA-Comp		**0.564**	**0.555**	**0.720**	0.846	**0.750**	**0.687**
LoRA (diverse)	100K	**0.570**	**0.586**	**0.765**	0.842	0.720	**0.697**
LoRA (unique)		0.556	0.580	0.588	**0.843**	**0.749**	0.663
LoRA-MoE (diverse)		**0.564**	**0.577**	**0.771**	0.844	**0.725**	**0.696**
LoRA-MoE (unique)		0.559	0.497	0.598	**0.853**	0.712	0.644
LoRA-Comp (diverse)		0.561	0.560	**0.715**	0.845	0.753	0.687
LoRA-Comp (unique)		**0.562**	**0.566**	**0.715**	**0.849**	**0.758**	**0.690**
LoRA (diverse)	300K	**0.541**	**0.547**	**0.688**	0.829	**0.728**	**0.667**
LoRA (unique)		0.511	0.522	0.511	**0.839**	0.515	0.580
LoRA-MoE (diverse)		**0.536**	0.535	0.653	**0.845**	0.734	0.661
LoRA-MoE (unique)		0.535	**0.556**	**0.682**	0.838	**0.739**	**0.670**
LoRA-Comp (diverse)		0.566	0.552	**0.730**	**0.848**	**0.756**	0.690
LoRA-Comp (unique)		**0.567**	**0.556**	0.727	0.846	0.751	**0.691**

Table 5: Full evaluation results for different upsampling strategy. For each upsampling setting and fine-tuning method, we highlight the **best value**.

Method	**Data**	**Chem**	**Med**	**Bio**	**Math**	**Geo**	**Avg**
LoRA	70%	0.557	0.563	0.642	**0.862**	0.667	0.658
LoRA (mix gen)		**0.558**	**0.597**	**0.766**	0.844	**0.761**	**0.705**
LoRA-MoE		**0.576**	0.514	0.585	**0.859**	0.733	0.653
LoRA-MoE (mix gen)		0.572	**0.594**	**0.790**	0.854	**0.753**	**0.713**
LoRA-Comp		0.549	0.548	0.701	0.864	0.747	0.682
LoRA-Comp (mix gen)		**0.551**	**0.554**	**0.720**	**0.870**	**0.753**	**0.690**
LoRA	100%	0.553	0.549	0.702	0.836	0.745	0.677
LoRA (mix gen)		**0.571**	**0.600**	**0.781**	**0.851**	**0.758**	**0.712**
LoRA-MoE		0.535	0.550	0.683	**0.847**	0.741	0.671
LoRA-MoE (mix gen)		**0.542**	**0.600**	**0.805**	0.839	**0.754**	**0.708**
LoRA-Comp		0.564	0.555	0.720	**0.846**	**0.750**	0.687
LoRA-Comp (mix gen)		**0.567**	**0.569**	**0.730**	0.845	**0.750**	**0.692**

Table 6: Full evaluation results for improving instruction following ability. For each data scale and fine-tuning method, we highlight the **best value**.

Method	**Data**	**Chem**	**Med**	**Bio**	**Math**	**Geo**	**Avg**
LoRA (r=16)	70%	0.557	0.563	0.642	**0.862**	0.667	0.658
LoRA (r=80)		0.556	**0.588**	0.671	0.847	0.743	0.681
LoRA (r=160)		**0.585**	0.538	**0.755**	0.848	**0.748**	**0.695**
LoRA-MoE (e=5)		0.576	0.514	**0.585**	**0.859**	0.733	**0.653**
LoRA-MoE (e=10)		**0.594**	**0.566**	0.455	0.841	**0.757**	0.643
LoRA-MoE (e=20)		0.591	0.537	0.550	0.846	0.731	0.651
LoRA (r=16)	100%	0.553	0.549	0.702	0.836	**0.745**	0.677
LoRA (r=80)		**0.562**	0.542	0.547	0.830	0.734	0.643
LoRA (r=160)		0.555	**0.598**	**0.749**	**0.844**	**0.745**	**0.698**
LoRA-MoE (e=5)		0.535	0.550	0.683	**0.847**	0.741	0.671
LoRA-MoE (e=10)		**0.561**	**0.554**	**0.727**	0.821	0.751	**0.683**
LoRA-MoE (e=20)		0.557	0.552	0.644	0.828	**0.757**	0.668

Table 7: Full evaluation results for scaling model trainable parameters. For each data scale and fine-tuning method, we highlight the **best value**.

Method	**Data**	**Chem**	**Med**	**Bio**	**Math**	**Geo**	**Avg**
LoRA-MoE	70 %	0.576	**0.514**	0.585	0.859	0.733	0.653
LoRA-MoE (shared expert)		0.531	0.502	**0.699**	**0.861**	0.741	**0.667**
LoRA-MoE (shared A)		0.578	0.476	0.672	0.848	**0.746**	0.664
LoRA-MoE (rank route)		**0.588**	0.480	0.431	0.848	0.734	0.616
LoRA-MoE	100 %	0.535	0.550	0.683	0.847	0.741	0.671
LoRA-MoE (shared expert)		0.568	0.532	0.677	0.850	0.751	0.676
LoRA-MoE (shared A)		0.567	**0.594**	**0.772**	**0.852**	**0.753**	**0.708**
LoRA-MoE (rank route)		**0.575**	0.496	0.422	0.813	0.620	0.585

Table 8: Full evaluation results for different architecture design. For each data scale, we highlight the **best value**.

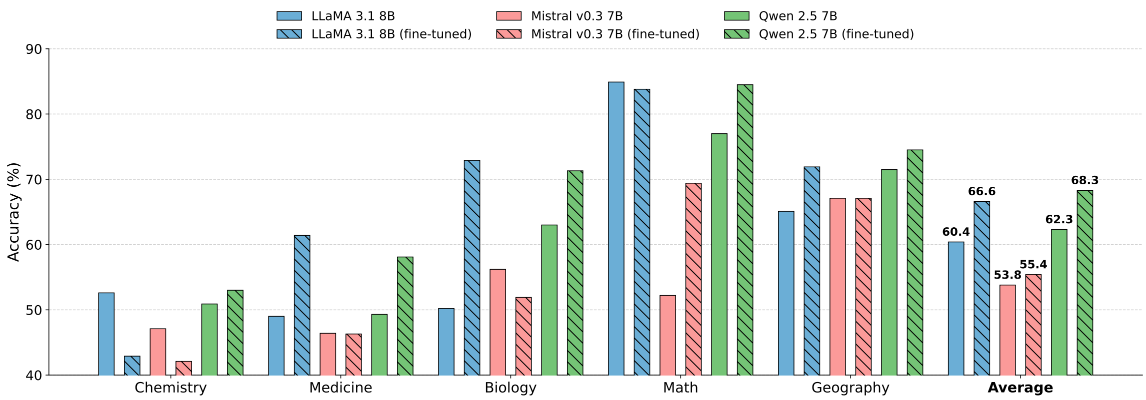


Figure 6: Base models' performance before and after fine-tuning for each discipline by preliminary scientific LoRA.