

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

Lintao Wang^{1,2,†,‡}, Zhuqiang Lu^{1,†}, Yilin Zhu^{1,†}, Kun Hu³,
Zhenfei Yin⁴, Shixiang Tang^{2,5}, Zhiyong Wang¹, Wanli Ouyang^{2,5}, Xinzhu Ma^{2,6,*},§

¹The University of Sydney, ²Shanghai AI Laboratory, ³Edith Cowan University,
⁴University of Oxford, ⁵The Chinese University of Hong Kong, ⁶Beihang University

[†]Equal contribution. ^{*}Corresponding author.

Abstract

While large language models (LLMs) perform strongly after adaptation within individual scientific domains, their learning dynamics in multi-disciplinary contexts remain poorly understood, despite the promise of improved generalization and broader applicability through cross-domain knowledge synergy. We present a systematic study of multi-discipline LLM fine-tuning across five scientific disciplines and compare the learning dynamics of full fine-tuning, LoRA, LoRA-MoE, and LoRA composition. In our naturally imbalanced setting, we show that multi-disciplinary learning is substantially more variable than single-discipline training and distill four empirical patterns: (1) Balance-then-Diversity: increasing scarce-domain exposure helps, but broad source-pool diversity must be preserved; (2) Merge-then-Align: general instruction data restores instruction-following behavior needed for effective joint adaptation; (3) Optimize-then-Scale: increasing trainable capacity offers limited gains without prior design optimization; and (4) Share-then-Specialize: asymmetric LoRA-MoE improves multi-discipline performance through a shared low-rank projection. Together, these patterns 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., 2024d). Prior AI4Science works built specialized models to solve specific scientific problems (Lamurias et al., 2023; Pei et al., 2023; Laghuvarapu et al., 2023; 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., 2023, 2024a). 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., 2026), studies have adapted general-purpose LLMs to scientific tasks by fine-tuning them with scientific instructions (Yue et al., 2024; Fang et al., 2024; 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 may support cross-discipline transfer and broader generalization beyond isolated 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; Tang et al., 2026) unifies representations of proteins, nucleotides, molecules and materials via natural language, enabling reasoning-driven scientific modeling. 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 ex-

[‡]wanglintao@pjlab.org.cn

[§]xinzhuma@buaa.edu.cn

tend 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., 2024), 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, occasional transfer gains are inconsistent, and joint fine-tuning can reduce average discipline-specific accuracy. Our evaluation measures scientific question answering and problem solving; it does not establish compositional interdisciplinary reasoning in the strict sense of combining knowledge from multiple disciplines within the same problem. We further analyze these phenomena to uncover the empirical regularities that govern multi-discipline fine-tuning and distill them into four patterns 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 instability. Diversity-preserving uniform upsampling mitigates this issue more effectively than upsampling on discipline-specific-token examples.
- **Merge-then-Align:** Instruction-following ability is important for effective joint adaptation. Multi-discipline fine-tuning alone degrades alignment, while mixing general instruction data restores instruction-following and improves aggregate performance.
- **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., 2023; Hao et al., 2023) and large-scale discipline-wise scientific foundation models (Ji et al., 2021; Chen et al., 2023, 2024a; 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., 2025), biology (Fang et al., 2024), medicine (Li et al., 2023b,a), chemistry (Zhang et al., 2024b; Li et al., 2024), 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., 2022; Ouyang et al., 2022; Zhang et al., 2026) 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 introduce 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.,

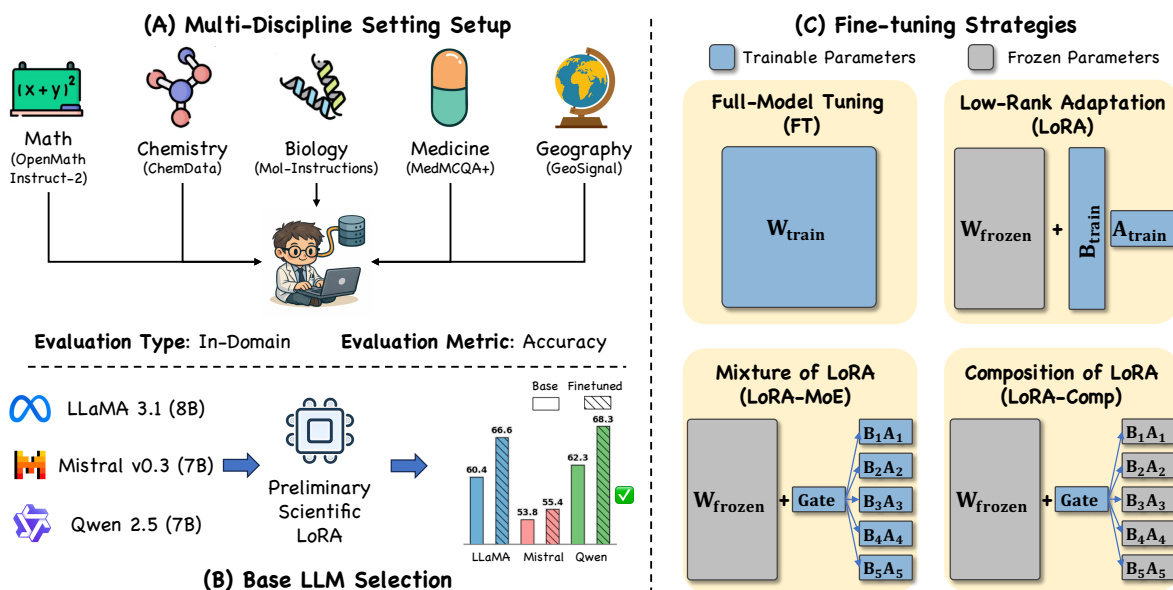


Figure 1: **Overview of the Empirical Study Methodology.** (A) Five scientific disciplines included in our multi-discipline setting. (B) Selection of Qwen 2.5 based on superior pre- and post-tuning scientific performance. (C) Fine-tuning strategies investigated by distinguishing trainable and frozen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

2022). A prominent PEFT approach is low-rank adaptation (LoRA) (Hu et al., 2022) and its variants (Zhang et al., 2023b; Liu et al., 2024), 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; Yu et al., 2020; Liu et al., 2021, 2023; Senushkin et al., 2023). Data-mixture methods instead optimize the sampling weights assigned to domains (Xie 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., 2025; 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., 2024; 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). Our comparison is organized around parameter-sharing structure rather than an exhaustive PEFT or MTL

leaderboard: full fine-tuning permits unrestricted adaptation, LoRA uses one jointly trained shared adapter, LoRA-MoE uses jointly trained routed experts, and LoRA-Comp composes independently trained discipline adapters through a learned router. Gradient balancing and adaptive mixture weighting are complementary to this architectural axis and can be combined with these methods.

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., 2025), which is augmented upon MATH (Hendrycks et al., 2021b) 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., 2024) for training. For medicine, we com-

Discipline	Data Scale		Source (Train / Eval)	Text Statistics	
	Samples	%		Avg. Length	Unique Tokens
Math	2,000,000	60.7	Toshniwal et al. (2025) / 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. (2024) / Fang et al. (2024)	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.

pile a large-scale dataset by aggregating samples from MedMCQA (Pal et al., 2022), MedAlpaca (Han et al., 2023), ChatDoctor (Li et al., 2023b), MedInstruct-52K (Zhang et al., 2023c), and others (Appendix A). For geography, we adopt GeoSignal (Deng et al., 2024), which includes QA covering geoscientific topics. The final training mixture is also heterogeneous in response format: a corpus audit identifies 54.08% chain-of-thought, 7.49% conversational, 5.87% multiple-choice, and 32.56% short or open QA examples, with median raw input and target lengths of 69 and 217 Qwen2.5 tokens, respectively (Appendix A.1).

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 the GSM8K test set (Cobbe et al., 2021), while chemistry is assessed with ChemBench (Zhang et al., 2024b), covering nine representative chemical tasks. For biology, we employ the biomolecular multiple-choice subset of the Mol-Instructions test set (Fang et al., 2024). Medicine is evaluated using the MedMCQA test set (Pal et al., 2022), and geography is assessed using the multiple-choice subset of GeoBench (Deng et al., 2024). We conduct all evaluations using the lm-evaluation-harness framework (Gao et al., 2024) and report accuracy. GSM8K uses free generation followed by answer extraction; the other four benchmarks use likelihood-based multiple-choice or classification evaluation. Thus, the demonstrated scope is scientific question answering and problem solving rather than long-form scientific explanation, hypothesis generation, or experimental design.

Fine-Tuning Setting. We study full fine-tuning and three parameter-efficient fine-tuning (PEFT) methods (Figure 1). Together they span unrestricted adaptation, one shared adapter, jointly trained routed experts, and composition of inde-

pendently trained discipline adapters:

- **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 $\mathbf{W} + \mathbf{B}\mathbf{A}$, 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 the LoRA rank to 16, and the number of experts for LoRA-MoE and LoRA-Comp matches the five disciplines. See Appendix A for details.

Base LLMs. We use three 7B-scale instruction-tuned LLMs—LLaMA 3.1 (Grattafiori et al., 2024), Mistral v0.3 (Mistral AI, 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 the main controlled experiments. Additional base-model results and transfer of the final recipe to Llama 3.2 3B Instruct are reported in Appendix B.

4 Empirical Pattern for Multi-Discipline LLM Fine-Tuning

To compare single- and multi-discipline learning, we analyze their scaling patterns across fine-tuning strategies and data scales (Figure 2). Single-discipline learning scales predictably with increasing data (Zhou et al., 2023a; Zhang et al., 2024a), whereas multi-discipline learning under our naturally heterogeneous corpus shows substantially less stable behavior. To rule out optimization noise, we additionally perform three independent runs for the main methods and report the mean and variance in Appendix F.

Fine-Tuning Perspective. Scaling curves exhibit greater instability 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-discipline data, these gains are inconsistent. On average, multi-discipline fine-tuning tends to degrade discipline-specific performance relative to single-discipline fine-tuning. Directly applying full-model fine-tuning (FT) also produces 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

Balance-then-Diversity. In naturally imbalanced scientific corpora, increasing scarce-domain exposure can reduce instability, while further gains depend on preserving broad diversity coverage rather than concentrating duplication.

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 performance instability, in contrast to the more stable trends observed in chemistry and mathematics. We hypothesize that limited exposure to these disciplines contributes to this behavior. Given the cost of acquiring additional scientific data, we evaluate two upsampling methods. *Diversity-preserving uniform upsampling* samples with replacement from the complete biology or geography pool. The comparison, *discipline-token-concentrated upsampling*, restricts duplicated examples to those containing discipline-specific tokens. The unique n-gram score is computed only as a post-hoc diversity diagnostic; it is not an n-gram or semantic-clustering sampler. Full procedures are provided in Appendix A.5. 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), increasing scarce-domain exposure improves overall performance in our setting. However, gains do not scale consistently with target size, and diversity-preserving uniform upsampling consistently outperforms the discipline-token-concentrated strategy. The con-

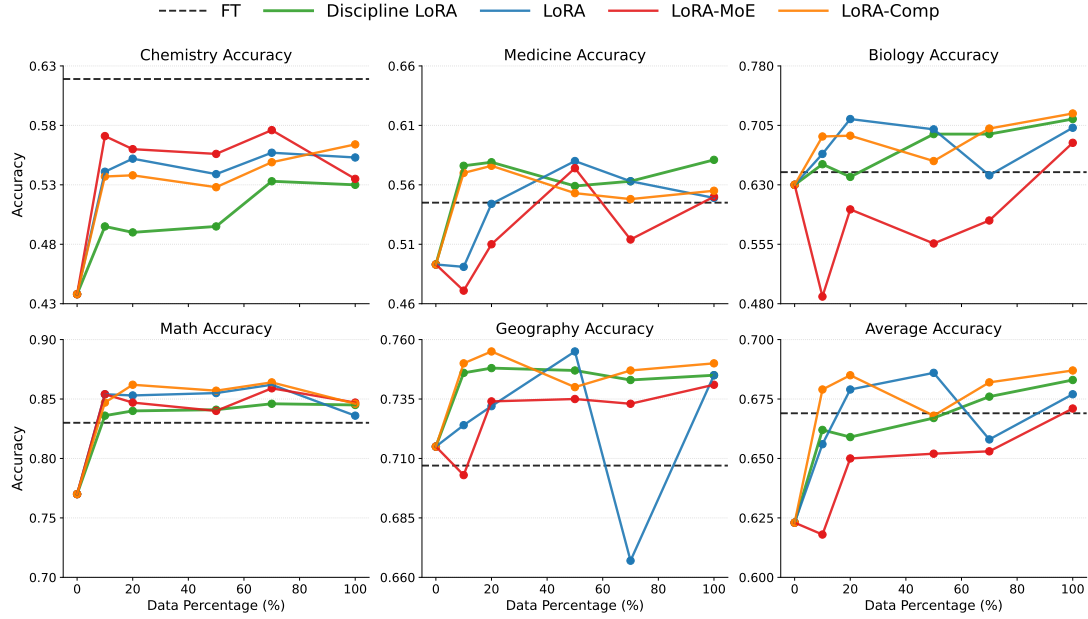


Figure 2: **Learning Dynamics of Fine-tuning Strategies.** Under the naturally heterogeneous corpus studied here, multi-discipline fine-tuning is less effective and stable than single-discipline tuning, highlighting trade-offs between generality, performance, and stability across methods.

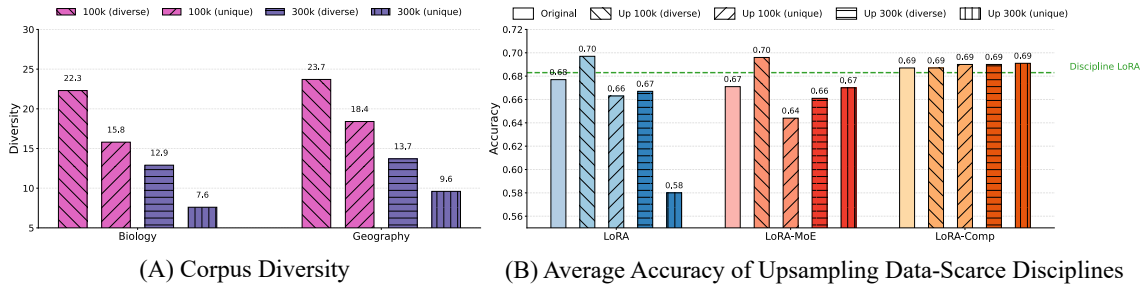


Figure 3: Analysis of how data-scarce discipline affects the multi-disciplinary learning.

centrated strategy reduces diversity coverage and repeatedly exposes a narrower set of prompts. Consistently, the post-hoc unique n-gram diversity metric (Song et al., 2024) (Figure 3(A)) correlates with the observed performance drops. LoRA-Comp remains relatively stable, likely because its limited trainable capacity constrains overfitting. Overall, upsampling biology and geography benefits both upsampled and non-upsampled domains (Appendix C). This supports a balance-then-diversity pattern for imbalanced scientific corpora for effective multi-discipline learning.

4.2 Merge-then-Align

Merge-then-Align. Mixing general instruction and multi-discipline data restores alignment and improves aggregate performance.

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.

To address this issue, we jointly mix general-domain instruction-following data from InternLM (Cai et al., 2024) with the multi-discipline corpus at the tested ratios and fine-tune new models. The mixture improves instruction-following performance and yields stronger aggregate results on

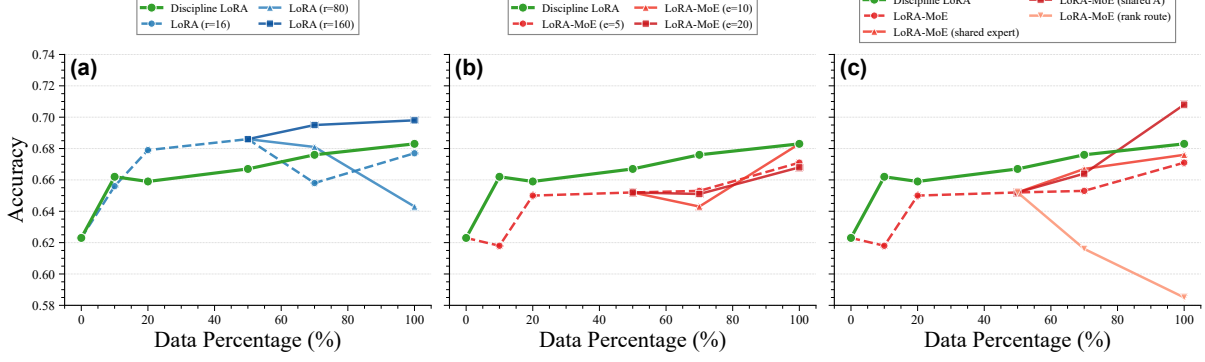


Figure 4: (a, b) Scaling LoRA rank and LoRA-MoE experts; (c) Different MoE designs.

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.

the scientific 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.

4.3 Optimize-then-Scale

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

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., 2025; Chen et al., 2024b), 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 (DeepSeek-AI et al., 2024), we introduce an additional shared LoRA expert alongside the five discipline-specific experts to promote shared and specialized knowledge separation. Additionally,

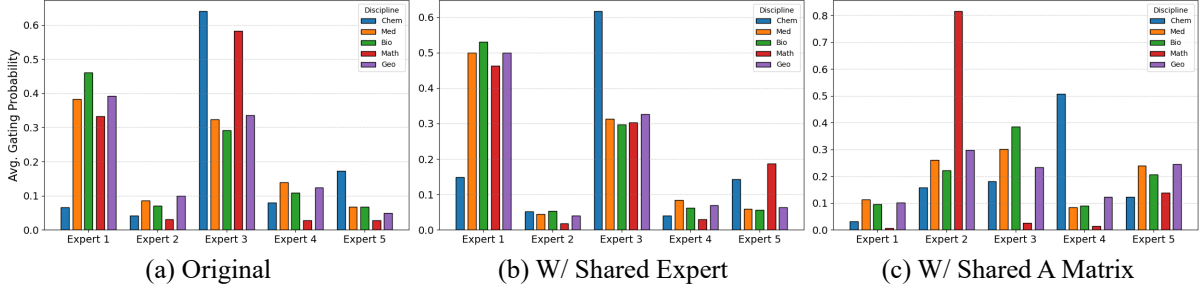


Figure 5: Mean expert routing probabilities computed from intermediate activations at the final feedforward layer.

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	<u>0.845</u>	<u>0.745</u>	0.683	0	-	0.264×5	-
FT (ori.)	<u>0.619</u>	0.545	0.646	0.830	0.707	0.669	-1.135	0.719	100	1920
FT (tuned)	0.649	<u>0.588</u>	<u>0.821</u>	0.831	0.743	<u>0.726</u>	+7.376	0.707	100	2080
LoRA-MoE (ori.)	0.535	0.550	0.683	0.847	0.741	0.671	-1.780	<u>0.732</u>	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: Comparison of multi-discipline fine-tuning methods before and after applying the final recipe.

								Qwen2.5-7B setting	Avg
Model	Chem	Med	Bio	Math	Geo	Avg	MMLU		
Llama 3.2 3B Instruct	0.468	0.455	0.462	0.758	0.594	0.547	0.615	Original LoRA-MoE	0.671
LoRA-MoE (tuned)	0.507	0.568	0.843	0.767	0.701	0.677	0.627	Upsampling only	0.696
								Instruction data only	0.708
								Shared-A only	0.708
								Complete recipe	0.728

Table 4: Additional validation. *Left*: transfer of the final recipe to another model family and parameter scale. *Right*: aggregate compatibility of the three executable recipe components on Qwen2.5-7B.

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., 2025), 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 (DeepSeek-AI et al., 2024) 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 C), likely due to excessive routing flexibility destabilizing expert selection. Overall, these findings support a share-then-specialize pattern in our setting: sharing the input projection constrains experts to a common low-rank representation before the routed output projections specialize to disciplines.

5 Final Multi-Discipline Learning Recipe

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

- **Data Scarcity Mitigation:** We upsample the biology and geography training corpora to 100K examples each using diversity-preserving uniform upsampling.
- **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., 2021a), achieving these gains with only 15% of the GPU hours required by full fine-tuning.

Cross-backbone transfer. On Llama 3.2 3B Instruct, the final shared-A LoRA-MoE recipe improves the five-discipline average from 0.547 to 0.677 and MMLU from 0.615 to 0.627, with gains in every discipline (Table 4). This result supports transfer across model families and to a different parameter scale. Configuration details are in Appendix B.

Component compatibility and deployment. The complete Qwen2.5-7B recipe reaches 0.728, exceeding the strongest individual intervention (0.708, Table 4), which supports aggregate compatibility. Inference profiling also reveals a deployment trade-off: evaluating all five experts raises mean latency from 1.771 to 4.620 seconds per example and reduces throughput from 72.32 to 27.71 tokens/s, with similar peak memory (14.72 versus 15.00 GiB). Appendix E reports the complete protocol and distributional statistics.

6 Conclusion

We systematically study joint LLM fine-tuning on a naturally imbalanced, format-heterogeneous corpus spanning five scientific disciplines. Across controlled experiments, robustness depends less on scaling data or trainable capacity than on coordinating exposure, alignment, and parameter sharing. Diversity-preserving upsampling stabilizes scarce disciplines without concentrating duplication, general instruction data restores instruction-following during joint adaptation, and shared-A LoRA-MoE combines a common latent projection with routed expert specialization. The resulting Balance-then-Diversity, Merge-then-Align, Optimize-then-Scale, and Share-then-Specialize patterns form a practical decision sequence: balance the data mixture, preserve alignment, optimize the adaptation design, and then scale capacity. On Qwen2.5-7B, the complete recipe reaches 0.728 average accuracy while training 0.902% of the parameters. It also improves every evaluated discipline on Llama 3.2 3B Instruct, providing evidence across architectures and parameter scales. Together, the results support efficient joint adaptation under natural heterogeneity, while compositional interdisciplinary reasoning remains an open evaluation problem.

Limitations

Three limitations motivate clear extensions of this study. (1) Controlled heterogeneity. Our conclusions characterize joint adaptation on naturally imbalanced corpora, where discipline co-varies with data volume, response format, and sequence length. Matched-format, equal-token, and balanced unique-data controls could disentangle these factors. Crossed mixtures could reveal when coverage transfers beyond upsampled disciplines. (2) Broader generalization. Our evidence centers on scientific QA and problem solving with Qwen2.5-7B and transfer to Llama-3.2-3B. Larger backbones, additional disciplines, and expert-grounded long-form or compositional benchmarks would test how broadly the four patterns transfer and connect discipline-wise gains to richer scientific reasoning. (3) Recipe optimization and deployment. The current design evaluates controlled interventions and the complete recipe, but not all pairwise combinations. Factorial studies, broader PEFT and multi-task baselines, additional seeds, and sparse routing could map interactions and preserve accuracy with lower inference cost.

Acknowledgments

Lintao Wang gratefully acknowledges the support of Shanghai Artificial Intelligence Laboratory, where this work was conducted during his internship.

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):511.
- Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. [Bridging the gap between consumers’ medication questions and trusted answers](#). In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, volume 264, pages 25–29. IOS Press.
- 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):026119.
- 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, Zhao Ye Fei, Yang Gao, Jiaye Ge, Chena 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, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. 2023. [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 F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. 2024a. [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, Lyu Sheng, Wanli Ouyang, and Jing Shao. 2024b. [Octavius: Mitigating task interference in MLLMs via LoRA-MoE](#). In *International Conference on Learning Representations*, pages 24652–24673.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. [GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803. PMLR.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- CogStack. 2023. OpenGPT. <https://github.com/CogStack/OpenGPT>.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. [DeepSeek-V3 technical report](#). *arXiv preprint arXiv:2412.19437*.
- Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 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. ACM.
- Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huanjun Chen. 2024. [Mol-Instructions: A large-scale biomolecular instruction dataset for large language models](#). In *International Conference on Learning Representations*, pages 48221–48251.
- 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, Amy Yang, Angela Fan, Anirudh

- Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 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, Alexei Figuerola, Alexander Löser, Daniel Truhn, and Ken K. Bressen. 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12556–12569. PMLR.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc.
- 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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- 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 Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.
- Jonas Köhler, Michele Invernizzi, Pim De Haan, and Frank Noé. 2023. [Rigid body flows for sampling molecular crystal structures](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17301–17326. PMLR.
- Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. 2023. [CoDrug: Conformal drug property prediction with density estimation under covariate shift](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 37728–37747. Curran Associates, Inc.
- Andre Lamurias, Alessandro Tibo, Katja Hose, Mads Albertsen, and Thomas Dyhre Nielsen. 2023. [Metagenomic binning using connectivity-constrained variational autoencoders](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18471–18481. PMLR.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. [LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.
- Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. [Towards 3D molecule-text interpretation in language models](#). In *International Conference on Learning Representations*, pages 17352–17371.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. [ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI \(LLaMA\) using medical domain knowledge](#). *Cureus*, 15(6):e40895.
- 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*, pages 91309–91330.
- Tianwei Lin, Jiang Liu, Wenqiao Zhang, Yang Dai, Haoyuan Li, Zhelun Yu, Wanggui He, Juncheng Li, Jiannan Guo, Hao Jiang, Siliang Tang, and Yueting Zhuang. 2025. **TeamLoRA: Boosting low-rank adaptation with expert collaboration and competition**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13622–13637. Association for Computational Linguistics.
- Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. 2023. **FAMO: Fast adaptive multitask optimization**. In *Advances in Neural Information Processing Systems*, volume 36, pages 57226–57243. Curran Associates, Inc.
- Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. 2021. **Conflict-averse gradient descent for multi-task learning**. In *Advances in Neural Information Processing Systems*, volume 34, pages 18878–18890. Curran Associates, Inc.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. **Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning**. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. **DoRA: Weight-decomposed low-rank adaptation**. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR.
- 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. 2019. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.
- 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 Khoshabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. **Cross-task generalization via natural language crowdsourcing instructions**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487. Association for Computational Linguistics.
- Mistral AI. 2024. **Mistral-7B-Instruct-v0.3**. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>.
- Lin Ning, Harsh Lara, Meiqi Guo, and Abhinav Rastogi. 2025. **MoDE: Effective multi-task parameter efficient fine-tuning with a mixture of dyadic experts**. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8248–8261. Association for Computational Linguistics.
- Oleksiy Ostapenko, Zhan Su, Edoardo Ponti, Laurent Charlin, Nicolas Le Roux, Lucas Caccia, and Alessandro Sordani. 2024. **Towards modular LLMs by building and reusing a library of LoRAs**. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38885–38904. PMLR.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. **Training language models to follow instructions with human feedback**. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. **MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering**. In *Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, 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. 2023. **FABind: Fast and accurate protein-ligand binding**. In *Advances in Neural Information Processing Systems*, volume 36, pages 55963–55980. Curran Associates, Inc.
- Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. 2025. **LoRA soups: Merging LoRAs for practical skill composition tasks**. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 644–655. Association for Computational Linguistics.
- Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. 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. ACM.
- Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. [Latent multi-task architecture learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):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](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14358–14369. ELRA and ICCL.
- 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](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29213–29224.
- Chen Tang, Yizhou Wang, Jianyu Wu, Lintao Wang, Shixiang Tang, Pengze Li, Encheng Su, Jun Yao, Jiabei Xiao, Yuqi Shi, Jielan Li, Hongxia Hao, Zhangyang Gao, Fang Wu, Ben Fei, Xiangyu Yue, Pan Tan, Bozitao Zhong, Jinouwen Zhang, and 10 others. 2026. [Accurate, interdisciplinary and transparent structure-property understanding with deep native structural reasoning](#). *arXiv preprint arXiv:2607.07708*.
- Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. 2024. [HydraLoRA: An asymmetric LoRA architecture for efficient fine-tuning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 9565–9584. Curran Associates, Inc.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisanin, Alexan Ayrapetyan, and Igor Gitman. 2025. [OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data](#). In *International Conference on Learning Representations*, pages 19243–19275.
- 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. Association for Computational Linguistics.
- Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, and 11 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, Guohang Zhuang, Yuchen Ren, Ben Fei, Ming Hu, Xin Chen, Dongzhan Zhou, Junjun He, Xiangyu Yue, Zhenfei Yin, and 13 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. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 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. Association for Computational Linguistics.
- Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, Zekun Guo, Yeqi Bai, Pan Deng, Yaosen Min, Ziheng Lu, Hongxia Hao, Han Yang, Jielan Li, Chang Liu, and 27 others. 2025. [Nature language model: Deciphering the language of nature for scientific discovery](#). *arXiv preprint arXiv:2502.07527*.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. [DoReMi: Optimizing data mixtures speeds up language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.
- 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, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. [Gradient surgery for multi-task learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2024. [MAMmoTH: Building math generalist models through hybrid instruction tuning](#). In *International Conference on Learning Representations*, pages 40320–40341.
- Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2024. [Pushing mixture of experts to the limit: Extremely parameter efficient MoE for instruction tuning](#). In *International Conference on Learning Representations*, pages 25401–25420.
- Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. [BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9. Association for Computational Linguistics.
- Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, and 7 others. 2025. [A generative model for inorganic materials design](#). *Nature*, 639(8055):624–632.
- Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024a. [When scaling meets LLM finetuning: The effect of data, model and finetuning method](#). In *International Conference on Learning Representations*, pages 44694–44713.
- Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. 2024b. [ChemLLM: A chemical large language model](#). *arXiv preprint arXiv:2402.06852*.
- Jingfan Zhang, Yi Zhao, Dan Chen, Xing Tian, Huanran Zheng, and Wei Zhu. 2024c. [MiLoRA: Efficient mixture of low-rank adaptation for large language models fine-tuning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17071–17084.
- Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023a. [Composing parameter-efficient modules with arithmetic operation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 12589–12610.
- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. [AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning](#). In *International Conference on Learning Representations*.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2026. [Instruction tuning for large language models: A survey](#). *ACM Computing Surveys*, 58(7):169:1–169:36.
- Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023c. [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. 2024d. [A comprehensive survey of scientific large language models and their applications in scientific discovery](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, and 4 others. 2026. [A survey of large language models](#). *Frontiers of Computer Science*, 20(12):2012627.
- 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. Association for Computational Linguistics.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. [LIMA: Less is more for alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 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 Experimental Details

This appendix provides additional details about the experimental setup used in our multi-disciplinary scientific fine-tuning study. We first report a corpus-format audit and describe the construction of the medical-domain training data, then summarize the evaluation protocol, fine-tuning configuration, up-sampling procedures, and major controlled variables. We also clarify several implementation choices that are important for interpreting the results.

A.1 Corpus Format Audit

We audit the final training mixture with deterministic text-based indicators for explicit rationale or final-answer markers, dialogue turns, and enumerated answer options. Each example is assigned one mutually exclusive category; examples without one of the first three signatures are assigned to short/open QA. This lightweight taxonomy is descriptive rather than a semantic classifier, and examples near a category boundary may exhibit more than one underlying behavior.

Response-format category	Share (%)
Chain-of-thought	54.08
Conversational	7.49
Multiple-choice	5.87
Short/open QA	32.56

Table 5: Deterministic response-format audit of the training mixture.

Using the Qwen2.5 tokenizer on the raw input and target fields, the median lengths are 69 and 217 tokens, respectively. Together, the format shares and length statistics document substantial heterogeneity beyond discipline.

A.2 Medical Data Construction in the Multi-Discipline Setting

To expand the medical-domain data and align it with the scale and recency of the other disciplines in our multi-discipline setup, we aggregate data from a range of existing medical and biomedical 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., 2023b; Ben Abacha et al., 2019; Bodenreider, 2004; Zhang et al., 2023c). The resulting medical dataset consists of the following sources:

- 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., 2023b);
- 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 MedInstruct-52K, released with AlpaCare (Zhang et al., 2023c);
- a subset of the Medical Meadow training data released with MedAlpaca (Han et al., 2023), including health advice, flashcards, and PubMed-based causal reasoning data;
- medical-related tasks from Natural Instructions v2.8 (Mishra et al., 2022), including tasks 179, 181, 1369, 1447, 1449, 1485, 1487, 1495, and 1645.

After aggregation, the final medical-domain dataset contains 490,766 samples, matching the corpus statistics reported in Table 1.

A.3 Evaluation Protocol

All evaluations use lm-evaluation-harness (Gao et al., 2024). For mathematics, the model generates until the end-of-sequence token, after which the final answer is extracted and compared with the reference. The other benchmarks use likelihood-based multiple-choice or classification evaluation, selecting the candidate with the highest log-likelihood. Models are fine-tuned on the sources in Table 1 and evaluated on the benchmark test sets described in Section 3.

A.4 Fine-Tuning Configuration

We provide additional details about the fine-tuning setup. For full-model fine-tuning, all model parameters are updated with a learning rate of 7×10^{-6} , a weight decay of 0.1, and a warm-up ratio of 0.05. Models are trained for 1 epoch.

For PEFT methods, we use a learning rate of 1×10^{-4} , a weight decay of 0.01, and a warm-up ratio of 0.1. These models are also trained for 1 epoch. Unless otherwise specified, the default LoRA module uses rank 16 and a scaling factor of 32, and is applied to all attention and feedforward layers. The parameter-scaling experiments separately evaluate ranks 80 and 160.

For the default LoRA-MoE and LoRA-Comp configurations, the gating network is implemented as a single-layer MLP that predicts gating weights at each adapted layer from the layer input. Both use five routed experts, corresponding to the five scientific disciplines, and each expert is configured identically to the default single-LoRA setup. The expert-scaling experiment separately evaluates 10 and 20 experts, while the shared-expert ablation adds one shared expert alongside the five discipline-specific experts.

All models are optimized using AdamW (Loshchilov and Hutter, 2019) with a linear learning rate scheduler. We use an effective batch size of 128. All experiments are implemented with HuggingFace Transformers (Wolf et al., 2020) and DeepSpeed (Rasley et al., 2020), and are trained on 32 NVIDIA A800 GPUs.

A.5 Upsampling Procedures

For a discipline dataset D and target size $T > |D|$, both procedures retain every original example and append sampled duplicates until the dataset contains T examples. They differ only in the candidate pool used for those additional draws:

1. **Diversity-preserving uniform upsampling.** Initialize $D' \leftarrow D$. Until $|D'| = T$, draw an example uniformly with replacement from the complete source pool D and append it to D' .
2. **Discipline-token-concentrated upsampling.** Construct $C \subseteq D$ from examples whose input or target contains at least one token in the fixed discipline-specific token list. Membership is binary: matching additional tokens does not increase sampling weight. Initialize $D' \leftarrow D$, then draw uniformly with replacement from C until $|D'| = T$.

We apply each procedure independently to biology and geography with $T \in \{100K, 300K\}$ and leave the other discipline datasets unchanged. The unique n-gram statistic in Figure 3 is computed after constructing D' and serves only as a post-hoc diversity diagnostic; neither procedure optimizes an n-gram score or uses semantic clustering. Diversity-preserving uniform upsampling is used in the final recipe at 100K examples per upsampled discipline.

A.6 Overview of Experimental Variables

To improve reproducibility and clarify the experimental protocol, we summarize the major controlled variables in Table 6. Our study considers multiple base models, data scales, fine-tuning strategies, upsampling settings, instruction-data mixtures, and LoRA-MoE architectural variants.

Unless otherwise specified, the main controlled experiments are conducted using Qwen2.5-7B as the backbone. Mistral v0.3 7B and LLaMA 3.1 8B are included in preliminary backbone selection, while the complete recipe is additionally evaluated on Llama 3.2 3B Instruct.

A.7 Additional Implementation Clarifications

We further clarify several implementation details that are important for interpreting the experimental results.

First, Discipline LoRA and LoRA use different training paradigms. Discipline LoRA trains one LoRA adapter independently for each scientific discipline, whereas LoRA trains a single adapter jointly on data from all disciplines. Therefore, Discipline LoRA reflects discipline-specific specialization, while LoRA evaluates whether a single shared adapter can capture transferable knowledge across disciplines.

Second, full fine-tuning is reported as a fixed 100%-data baseline, while the controlled data- and capacity-scaling analyses focus on LoRA-based methods. Accordingly, we do not infer the scaling behavior of full fine-tuning from these experiments.

Finally, Merge-then-Align names a conceptual design principle rather than a sequential training schedule. Operationally, scientific and general instruction-following examples are mixed and optimized jointly in a single training stage. This design tests whether general instruction data preserves response formatting, option following, and direct-answer behavior during multi-disciplinary scientific fine-tuning.

Variable	Values	Section
Base model	Qwen2.5-7B, Mistral-v0.3-7B, LLaMA-3.1-8B; Llama-3.2-3B transfer	Section 3
Data amount	0%, 10%, 20%, 50%, 70%, 100%	Section 4
Fine-tuning method	Full FT, LoRA, LoRA-MoE, LoRA composition	Section 4
Upsampling method	Discipline-token-concentrated, diversity-preserving uniform	Section 4.1
Upsampling amount	100K, 300K	Section 4.1
Instruction data amount	0%, 70%, 100%	Section 4.2
LoRA rank	16, 80, 160	Section 4.3
LoRA-MoE expert number	5, 10, 20	Section 4.3
LoRA-MoE design	Shared expert, shared-A, rank-wise routing	Section 4.4

Table 6: Summary of experimental variables used in our multi-disciplinary scientific fine-tuning study. The table consolidates the main backbone choices, data scaling settings, fine-tuning methods, data balancing strategies, instruction-data mixture ratios, and LoRA-MoE architectural variants.

B Base Model Selection and Cross-Backbone Transfer

Before conducting the main controlled experiments, we compare different base models under preliminary scientific LoRA fine-tuning. As shown in Figure 6, we evaluate LLaMA 3.1 8B, Mistral v0.3 7B, and Qwen2.5-7B across five scientific disciplines: chemistry, medicine, biology, mathematics, and geography.

Overall, preliminary scientific LoRA fine-tuning improves the average performance of all base models, with Qwen2.5-7B achieving the strongest overall performance and consistent gains across disciplines. We therefore use Qwen2.5-7B for the main controlled experiments and report the final-recipe transfer test separately below.

For direct cross-backbone validation, we apply the complete recipe to Llama 3.2 3B Instruct: diversity-preserving upsampling to 100K biology and geography examples, the tested general-instruction mixture, and a five-expert shared-A LoRA-MoE with rank 16 and token-level, layer-wise routing. We retain the five Qwen2.5 benchmark protocols and additionally evaluate MMLU. The gains across all five disciplines and on MMLU in Table 4 provide evidence of transfer across architectures and parameter scales.

C Additional Quantitative Results

This section provides the full quantitative results corresponding to the analyses in the main paper. Table 7 reports the results of different fine-tuning methods across all tested data scales. Table 8 compares different upsampling strategies. Table 9 reports per-discipline mixed-data results at two scientific-data scales; the instruction-data amount sweep itself is reported in Table 2. Table 10 studies the effect of increasing trainable parameters. Table 11 compares different LoRA-MoE architectural variants.

Averaged across LoRA, LoRA-MoE, and LoRA-Comp, diversity-preserving upsampling yields higher overall accuracy at both targets: 0.693 versus 0.666 at 100K and 0.673 versus 0.647 at 300K. This aggregate advantage, together with the stronger gains at the selected 100K recipe setting, supports Balance-then-Diversity while allowing method-specific variation.

At full scientific-data scale, the rows without and with instruction data correspond to the IF= 0% and IF= 100% conditions in Table 2. The 70% block provides the same per-discipline comparison under reduced scientific-data exposure.

Capacity scaling is non-monotonic and method-dependent. Rank 160 improves LoRA at both scales, whereas rank 80 degrades at full data and additional LoRA-MoE experts yield inconsistent changes. Capacity can help in selected settings,

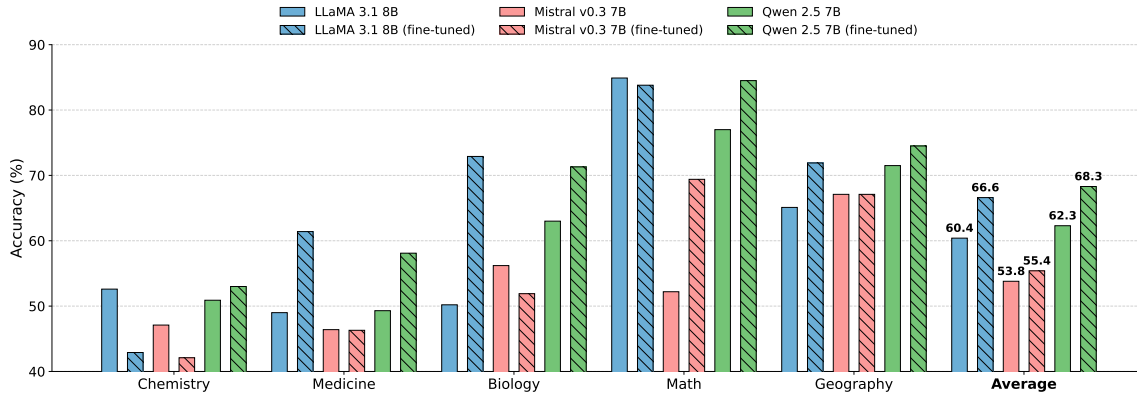


Figure 6: Base-model performance before and after preliminary scientific LoRA fine-tuning across five scientific disciplines. Qwen2.5-7B achieves the strongest overall performance and is therefore selected as the main backbone for the controlled experiments.

but it is not a reliable substitute for architectural optimization.

D Recipe Component Compatibility

The executable recipe contains three interventions: diversity-preserving upsampling, general instruction-data mixing, and the shared-A LoRA-MoE architecture; Optimize-then-Scale is a design-order observation rather than a fourth binary switch. The right panel of Table 4 shows that the complete recipe reaches 0.728, exceeding the strongest individual intervention by 0.020 and supporting aggregate compatibility. The gain is diminishing rather than additive, and without the three pairwise configurations we make no factorial, additive, or conditional-effect claim.

E Inference Efficiency

We profile dense Qwen2.5-7B-Instruct and the final shared-A LoRA-MoE on one NVIDIA H200 GPU. Both use BF16 without quantization, batch size 1, 512 input tokens, and 128 generated tokens. Each measurement follows three warm-up runs and covers 20 timed runs using PyTorch 2.3.0+cu121, Transformers 4.45.2, and FlashAttention 2.5.8. Throughput is measured over generated tokens, and peak allocated device memory is recorded for the same runs.

Metric	Dense	Share-A
Mean latency (s/example)	1.771	4.620
P50 latency (s/example)	1.772	4.625
P95 latency (s/example)	1.834	4.669
Throughput (tokens/s)	72.32	27.71
Peak memory (GiB)	14.72	15.00

Table 12: Batch-one inference profile on a single H200 GPU.

The shared-A implementation evaluates all five experts for every token at 140 routed adapter hooks. Its mean latency is therefore 160.9% higher than the dense baseline even though peak memory changes little. This separates parameter-efficient training from inference efficiency. Prompt-level top- k expert selection, as explored by MiLoRA (Zhang et al., 2024c), could amortize routing and avoid evaluating every expert at every generated token; adapting such sparse routing while preserving the observed accuracy is left to future work.

F Multi-Seed Stability Analysis

We assess run-to-run variability for a representative original-setting configuration using three independent runs. We compare Discipline LoRA, jointly trained LoRA, and LoRA-MoE and report benchmark-wise means and variances. This is a configuration-level robustness check rather than seed replication of every data-scaling point.

We report results on five discipline-specific benchmarks: ChemBench, MedMCQA, Mol-Instructions, GSM8K, and GeoBench. We also report the overall average across disciplines.

Table 13 shows small overall variance for Discipline LoRA and LoRA, while LoRA-MoE varies more, particularly on Mol-Instructions and in the overall average. The reruns therefore support the stability of the two non-MoE baselines and independently reproduce the broader instability pattern for LoRA-MoE.

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 7: Full evaluation results for different fine-tuning methods across all tested data scales. For each data scale, the best value in each discipline and the best average score are highlighted in **bold**.

G Qualitative Error Analysis for Merge-then-Align

In Section 4.2, we show that mixing general instruction-following data with multi-disciplinary scientific data improves overall performance. To better understand this effect, we provide a qualitative comparison between two models: one fine-tuned only on scientific data and another fine-tuned with both scientific data and general instruction-following data.

The examples in Table 14 illustrate response-behavior differences associated with mixed-data training, especially when the model must parse answer choices, follow the requested format, and provide a direct answer. They complement the quantitative IFEval results but are not a controlled attribution experiment.

In Example 1, the scientific-data-only model fails to recognize that the candidate options are already included in the prompt and incorrectly asks the user to provide answer choices. This behavior is consistent with weakened instruction following because the model misreads the input format. In contrast, the mixed-data model correctly identifies

the available options and directly selects the answer.

In Example 2, the scientific-data-only model follows the requested option format but selects the wrong answer, whereas the mixed-data model correctly selects lead. This example complements the first by illustrating improved answer selection rather than a pure formatting correction.

Overall, these examples illustrate the Merge-then-Align pattern: mixed-data training is associated with more reliable adherence to the requested response behavior and correct option selection, consistent with the IFEval and scientific-benchmark gains.

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 (diversity-preserving)	100K	0.570	0.586	0.765	0.842	0.720	0.697
LoRA (token-concentrated)		0.556	0.580	0.588	0.843	0.749	0.663
LoRA-MoE (diversity-preserving)		0.564	0.577	0.771	0.844	0.725	0.696
LoRA-MoE (token-concentrated)		0.559	0.497	0.598	0.853	0.712	0.644
LoRA-Comp (diversity-preserving)		0.561	0.560	0.715	0.845	0.753	0.687
LoRA-Comp (token-concentrated)		0.562	0.566	0.715	0.849	0.758	0.690
LoRA (diversity-preserving)	300K	0.541	0.547	0.688	0.829	0.728	0.667
LoRA (token-concentrated)		0.511	0.522	0.511	0.839	0.515	0.580
LoRA-MoE (diversity-preserving)		0.536	0.535	0.653	0.845	0.734	0.661
LoRA-MoE (token-concentrated)		0.535	0.556	0.682	0.838	0.739	0.670
LoRA-Comp (diversity-preserving)		0.566	0.552	0.730	0.848	0.756	0.690
LoRA-Comp (token-concentrated)		0.567	0.556	0.727	0.846	0.751	0.691

Table 8: Full evaluation results for diversity-preserving uniform and discipline-token-concentrated upsampling. For each upsampling amount and fine-tuning method, the best value is highlighted in **bold**.

Method	Science Data	Chem	Med	Bio	Math	Geo	Avg
LoRA	70%	0.557	0.563	0.642	0.862	0.667	0.658
LoRA (+ instruction data)		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 (+ instruction data)		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 (+ instruction data)		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 (+ instruction data)		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 (+ instruction data)		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 (+ instruction data)		0.567	0.569	0.730	0.845	0.750	0.692

Table 9: Full per-discipline results for mixing general instruction-following data at two scientific-data scales. The block percentage denotes scientific-data scale; the mixed rows add the general-instruction subset used for that setting. For each scale and method, the better result is highlighted in **bold**.

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 10: Full evaluation results for scaling trainable parameters. We vary the LoRA rank and the number of LoRA-MoE experts. For each data scale and method group, the best value is highlighted in **bold**.

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-wise)		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-wise)		0.575	0.496	0.422	0.813	0.620	0.585

Table 11: Full evaluation results for different LoRA-MoE architectural variants. For each data scale, the best value is highlighted in **bold**.

Discipline / Benchmark	Method	Mean	Variance
Chemistry / ChemBench	Discipline LoRA	0.5000	0.0000127
	LoRA	0.5527	0.0000683
	LoRA-MoE	0.5630	0.0000320
Medicine / MedMCQA	Discipline LoRA	0.5670	0.0003207
	LoRA	0.5120	0.0005580
	LoRA-MoE	0.5283	0.0016970
Biology / Mol-Instructions	Discipline LoRA	0.6640	0.0000647
	LoRA	0.6597	0.0000440
	LoRA-MoE	0.6083	0.0077200
Mathematics / GSM8K	Discipline LoRA	0.8327	0.0000483
	LoRA	0.8543	0.0000016
	LoRA-MoE	0.8533	0.0000168
Geography / GeoBench	Discipline LoRA	0.7437	0.0000029
	LoRA	0.7250	0.0000487
	LoRA-MoE	0.7310	0.0004250
Overall Average	Discipline LoRA	0.6617	0.0000373
	LoRA	0.6607	0.0000350
	LoRA-MoE	0.6570	0.0008090

Table 13: Multi-seed stability analysis over three independent runs. We report the mean and variance of Discipline LoRA, LoRA, and LoRA-MoE on five discipline-specific benchmarks and their overall average.

Example 1: Lake chemical deposition order

Instruction: The half order of lake chemical deposition in arid climate region is: A) Sulfate-chloride-carbonate B) Chloride-sulfate-carbonate C) Carbonate-sulfate-chloride

Scientific-data-only model: The answer to the question is not provided in the prompt. Please provide the options for the answer so I can assist you further.

Mixed-data model: (C) Carbonate-sulfate-chloride

Example 2: Non-essential metal mineral

Instruction: Which of the following is a non-essential metal mineral? What of the following is the right choice? (A) Sodium (B) Manganese (C) Iron (D) Lead

Scientific-data-only model: The correct answer is (A) Sodium.

Mixed-data model: (D) Lead

Table 14: Qualitative comparison between a model fine-tuned only on scientific data and a model fine-tuned with mixed scientific and general instruction-following data. In these examples, the scientific-data-only model fails to follow the requested format or selects an incorrect choice, whereas the mixed-data model directly provides the correct option.