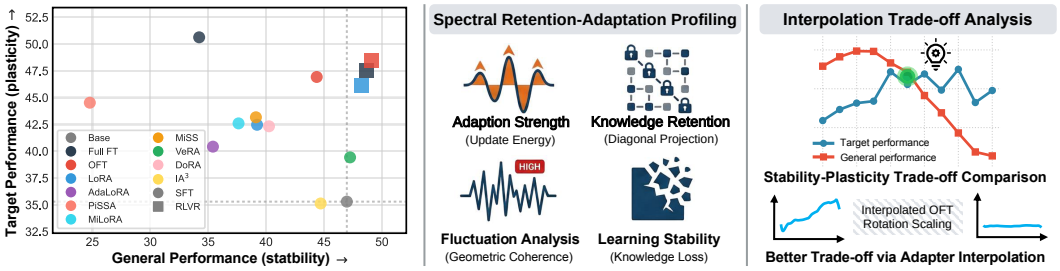
PEFT-ARENA: UNDERSTANDING PARAMETER-EFFICIENT FINETUNING FROM A STABILITY-PLASTICITY PERSPECTIVE

Yangyi Huang[1*] **Ruotian Peng**[2*] **Zeju Qiu**[3] **Jiale Kang**[1] **Yandong Wen**[2]
Bernhard Schölkopf[3] **Weiyang Liu**[1,3]

[1]The Chinese University of Hong Kong [2]Westlake University
[3]Max Planck Institute for Intelligent Systems, Tübingen [*] Equal Contribution.

ABSTRACT

Parameter-efficient finetuning (PEFT) has emerged as a practical solution for adapting large foundation models by updating only a small subset of parameters, and yet existing evaluations largely focus on downstream task performance while overlooking the preservation of pretrained capabilities. In this work, we argue that PEFT should be evaluated through the lens of the stability-plasticity dilemma, which characterizes the fundamental trade-off between efficient task adaptation (plasticity) and resistance to forgetting (stability). To this end, we introduce PEFT-Arena, a unified benchmark that jointly measures downstream performance and general capability retention. Our results show that all PEFT methods exhibit inherent stability-plasticity trade-offs and that different methods produce distinct trade-off patterns, indicating that neither metric alone is sufficient for evaluation. Besides external task-level assessment, we also propose to analyze the spectral geometry of weight updates to uncover the underlying mechanisms that govern the plasticity-stability trade-off. Our results show that PEFT methods achieving better trade-offs exhibit more structured and predictable spectral dynamics, highlighting spectral regularity as an intrinsic factor governing stability and as a guiding principle of the design of future PEFT algorithms. Inspired by the stability-plasticity trade-off, we exploit interpolation between the PEFT-tuned model and the base model. We find that such interpolated models often achieve a better trade-off than either model alone.



(a) Results on PEFT-Arena (b) Internal Spectral Analysis (c) External Performance Analysis

Figure 1: PEFT-Arena is designed to comprehensively evaluate the trade-off between downstream task adaptation and pretrained knowledge retention in LLM post-training with PEFT methods. (a) Stability-plasticity trade-off on the downstream math task; (b) Internal analysis includes studying spectral geometry on weight matrices; (c) External analysis includes target and general performance evaluation and interpolation study.

1 INTRODUCTION

As foundation models continue to scale, finetuning all parameters becomes increasingly infeasible. To address this challenge, parameter-efficient finetuning (PEFT) has been proposed. PEFT aims to adapt large pretrained models to downstream tasks by modifying only a small number of parameters, or by introducing lightweight trainable components, while keeping the majority of the model weights fixed. By drastically reducing the number of trainable parameters, PEFT significantly lowers computational

and memory costs, enabling efficient adaptation across diverse tasks without the need to retrain or store a full model for each task.

Although numerous PEFT methods (Houlsby et al. (2019); Aghajanyan et al. (2020); Hu et al. (2022b); Edalati et al. (2022); Wang et al. (2022); Gheini et al. (2021); Zaken et al. (2022); Guo et al. (2020); Sung et al. (2021); Ansell et al. (2022); Lester et al. (2021); Li & Liang (2021); Vu et al. (2022); He et al. (2021); Mao et al. (2021); Karimi Mahabadi et al. (2021); Liu et al. (2022a); Sung et al. (2022); Chen et al. (2023); Jia et al. (2022); Chen et al. (2022); Zhang et al. (2022); Jie & Deng (2023); Lian et al. (2022); Luo et al. (2023)) have been proposed, their evaluation has largely been limited to downstream task performance. We argue that this metric alone can be highly misleading and insufficient for fairly assessing the effectiveness of PEFT methods. Developing more principled evaluation criteria is therefore crucial for advancing PEFT research and calls for greater attention from the community. To address this limitation, we draw inspiration from the stability-plasticity dilemma (Mermillod et al. (2013)), a well-known challenge in both artificial and biological neural systems. This dilemma characterizes the fundamental trade-off between learning efficiency (*i.e.*, plasticity) and resistance to forgetting (*i.e.*, stability). Guided by this dilemma, we are interested in the following question:

Which PEFT method can achieve the best trade-off between stability and plasticity trade-off?

To address this question, we introduce **PEFT-Arena**, a comprehensive benchmark for evaluating PEFT methods. This new PEFT benchmark focuses on assessing the trade-off between the task adaptation effectiveness (*i.e.*, downstream task performance) and the preservation of pretrained capabilities (*i.e.*, general task performance). By providing a reliable and unified evaluation framework, PEFT-Arena enables effective and fair comparison across different PEFT methods. With PEFT-Arena, we make a few useful observations: (1) neither downstream target performance nor general task performance can serve as a reliable evaluation metric (both metrics are necessary); (2) the stability-plasticity trade-off holds for every PEFT method; (3) different PEFT methods yields different trade-off patterns. More importantly, we find that the stability-plasticity perspective calls into question many recent methodological advances in PEFT, suggesting that future PEFT algorithms should be explicitly designed for improving this trade-off.

While PEFT-Arena offers an effective external empirical measure of stability and plasticity, it remains unclear which internal mechanisms best govern the stability-plasticity trade-off in PEFT. In order to gain deeper insights to designing better PEFT method, we therefore seek to answer the following question:

What internal metrics can characterize a good PEFT method?

To this end, we propose to analyze how the spectral geometry of weight matrices evolves under different PEFT methods. This motivates us to monitor the training dynamics of the singular values and singular vectors of weight matrices. We find that PEFT methods exhibiting a favorable stability-plasticity trade-off tend to be more geometrically principled, in the sense that their weight updates are more spectrally structured and the spectral deviation before and after post-training is more predictable. In general, our empirical study implies that the stability of PEFT largely depends on spectral regularity and consistency, whereas plasticity is more multifaceted and can be achieved through less structured updates.

The stability-plasticity trade-off can be partially characterized by the distance between the finetuned model and the base model. Guided by this insight, we exploit interpolation between these two models, yielding interpolated variants of PEFT methods. We find that these interpolated variants often achieve a substantially better trade-off than either the finetuned or the base model alone. In the light of this observation, we further explore *interpolated* and *geometry-aware* variants of Orthogonal Finetuning (OFT). In particular, we introduce **interpolated Orthogonal Finetuning (iOFT)**, which adjusts OFT updates based on their intrinsic geometric structure. Empirically, iOFT serves as a practical and relatively robust trade-off knob, and can improve the stability-plasticity frontier over the vanilla OFT checkpoint without requiring an exhaustive sweep over a global scaling coefficient. Our contributions are listed below:

- **External understanding of PEFT.** We introduce PEFT-Arena, a unified benchmark that evaluates PEFT via the stability-plasticity trade-off, enabling comparison beyond downstream accuracy.
- **Internal understanding of PEFT.** We characterize PEFT mechanisms by monitoring *spectral characteristics* (singular values/vectors) during post-training, showing that better stability is associated with more structured spectral deviations.

- **Interpolated PEFT improves the stability-plasticity trade-off.** We show that interpolating between the base and finetuned models can often improve the trade-off, and explore iOFT as a geometry-aware OFT variant that can robustly improve the stability-plasticity trade-off frontier.

2 Related Work

PEFT and subspace learning. Parameter-efficient finetuning (PEFT) specializes large pretrained models by optimizing only a small subset of parameters, substantially reducing computational overhead while often achieving performance comparable to full finetuning (Hu et al. (2022a)). Prior PEFT methods can be broadly grouped into three categories. *(1) Additive weight reparameterization.* A representative approach is Low-Rank Adaptation (LoRA) (Hu et al. (2022a)), which freezes pretrained weights and introduces a learnable low-rank update to enable efficient task adaptation with a small number of trainable parameters. Subsequent work extends LoRA along several complementary dimensions. Methods such as AdaLoRA (Zhang et al. (2023)) focus on adaptive rank allocation by dynamically re-allocating rank budgets during training, and DoRA (Liu et al. (2024a)) improves optimization by decoupling update magnitude from direction. Other approaches modify the update structure: MISS (Kang & Yin (2025)) replaces the two-factor decomposition with a single expanded low-rank matrix, and VeRA (Kopiczko et al. (2024)) attains high effective update rank through shared, frozen random projection matrices modulated by lightweight trainable scaling vectors. Another line of work explores initialization strategies informed by the spectral properties of pretrained weights. PiSSA (Meng et al. (2024)) initializes low-rank factors using principal singular components, whereas MiLoRA (Wang et al. (2025)) instead targets minor components by initializing updates on the minor singular subspace while freezing the principal components. *(2) Constrained weight updates.* Beyond low-rank additive updates, Orthogonal Finetuning (OFT) (Qiu et al. (2023); Liu et al. (2024b); Qiu et al. (2025)) constrains the learned transformations to be orthogonal. By preserving the spectral properties of the original weight matrices, such constraints act as an implicit regularizer during adaptation and provide an alternative form of structured parameter efficiency. *(3) Activation- and prompt-based adaptation.* Instead of directly modifying weight matrices, this line of work introduces auxiliary parameters to modulate intermediate representations. For instance, IA^3 learns per-channel scaling vectors that element-wise modulate key activations (Liu et al. (2022a)). Similarly, Prompt Tuning (Liu et al. (2022b)) and Prefix-Tuning (Li & Liang (2021)) prepend trainable continuous embeddings to the input layer or to hidden states across attention blocks, respectively. These methods facilitate task specialization by re-contextualizing frozen representations, effectively inducing localized shifts in the activation space.

Continual learning. Continual learning (CL) aims to incorporate new knowledge while preserving previously acquired capabilities, thereby mitigating catastrophic forgetting Wang et al. (2024a). Prior CL approaches are commonly categorized into data-replay-based methods Aljundi et al. (2019); Silver & Mercer (2002); Tiwari et al. (2022), which require access to historical training data, and data-free methods Wortsman et al. (2022); Lubana et al. (2022); Chen et al. (2024); Panda et al. (2024), which avoid replay by relying instead on architectural constraints or regularization. Since storing or revisiting past data can be impractical due to privacy, storage, or compliance constraints, we focus primarily on data-free CL in this paper. This paradigm operates directly in parameter space, bypassing the need for historical data. Standard approaches employ ℓ_2 regularization Kirkpatrick et al. (2017) or model merging Wortsman et al. (2022); Lubana et al. (2022); Kleiman et al. (2025); Lin et al. (2024) to anchor finetuned parameters to the base model, thereby balancing stability and plasticity. For PEFT, O-LoRA Wang et al. (2023) and InfLoRA Liang & Li (2024) enforce orthogonality between sequential task-specific subspaces to mitigate forgetting. However, these methods often overlook the preservation of pretrained knowledge. KeepLoRA Luo et al. (2026) constrains updates to the non-principal subspace of the original weights, protecting the principal components that encode general capabilities.

Catastrophic forgetting in LLMs. While traditional CL primarily addresses multi-task sequential learning, recent focus has shifted toward catastrophic forgetting during the post-training of LLMs, particularly when finetuning on distributions that diverge significantly from the pre-training data Sanyal et al. (2025); Lin et al. (2025). A prevailing hypothesis identifies the discrepancy between on-policy and off-policy distributions as a primary driver of forgetting. To bridge this gap, several data-centric strategies aim to simulate on-policy learning via reweighting. FLOW Sanyal et al. (2025) prioritizes simpler samples to stabilize updates, while TALR Lin et al. (2025) and EAFT Diao et al. (2026) mitigate forgetting by attenuating learning rates or weights for tokens with high difficulty or low information entropy. Despite the effectiveness of these data-level interventions, a fundamental question remains: how do these stability-promoting effects manifest in the model's

underlying parameter space? Emerging evidence in reinforcement learning suggests that optimization stability is intrinsically linked to parameter-space dynamics. Specifically, the preferential updating of non-principal weight components to constrain parameter shifts (Zhu et al. (2025); Mukherjee et al. (2025)). While preliminary studies have empirically observed that LoRA exhibits superior knowledge retention compared to full finetuning Biderman et al. (2024), it remains poorly understood how such spectral dynamics evolve under different PEFT parameterizations, or how they structurally give rise to the resulting stability-plasticity trade-off. To address this gap, we conduct a systematic evaluation of PEFT methods to identify configurations that yield an optimal trade-off.

Model averaging and interpolation. Task arithmetic (Ilharco et al. (2023)) shows that meaningful task-specific knowledge is encoded in weight differences $W_{\text{ft}} - W_{\text{pre}}$, and that scaling or combining these "task vectors" enables modular control over model behavior. Model soups (Wortsman et al. (2022)) average checkpoints for improved robustness. TIES merging (Yadav et al. (2023)) resolves sign conflicts across task vectors. Fisher merging (Matena & Raffel (2022)) uses second-order information for weighted combinations. Orthogonal merging (Yang et al. (2026)) combines task-specific weights on orthogonal manifolds. Functional dual anchors (Shi et al. (2025)) merge task vectors within input-representation space. These methods primarily combine multiple task-specialized models. We repurpose interpolation differently and view it as a diagnostic and post-hoc improvement tool for single-task PEFT. By sweeping the interpolation coefficient α between the base and finetuned model, we trace explicit stability-plasticity trade-off curves, revealing that final checkpoints consistently overshoot the optimal operating point.

Spectral analysis of neural networks. The singular value spectrum of weight matrices has long been recognized as informative about network behavior. Heavy-tail spectral theory connects weight matrix spectra to generalization (Martin & Mahoney (2021)), and recent work in reinforcement learning has linked stability to preferential updating of non-principal weight components (Zhu et al. (2025); Mukherjee et al. (2025)). Biderman et al. (2024) empirically observes that LoRA retains pretrained knowledge better than full finetuning, but the spectral mechanisms underlying this observation remain unexplored. Qiu et al. (2023) explicitly links the pretrained knowledge retention to spectrum preservation. Our spectral analysis specifically targets PEFT updates, decomposing changes into retention and adaptation components and introducing quantitative metrics that reflect forgetting behavior. This offers a systematic spectral framework that connects internal geometric properties of PEFT updates to external stability-plasticity outcomes.

3 THE PEFT-ARENA BENCHMARK

3.1 EXPERIMENTAL SETUP

In PEFT-Arena, parameter-efficient finetuning methods are evaluated along two axes: (i) target-domain performance (*plasticity*) and (ii) general ability preservation (*stability*). We evaluate two target domains (math and medical reasoning) under both SFT and RLVR. Table 1 summarizes the setup; full details are in Appendix C.

Benchmarks		Models & Methods	
Target (Math)	Math-500 (Lightman et al. (2023)), AMC23, AIME24	Base models	Qwen2.5-7B (Yang et al. (2024)), Llama3.2-3B-Instruct (Dubey et al. (2024))
Target (Med)	MedMCQA (Pal et al. (2022)), MedQA (Jin et al. (2021)), PubMedQA (Jin et al. (2019)), MMLU-Pro (Wang et al. (2024b)), GPQA (Rein et al. (2024)), Lancet/NEJM/MedBullets (Chen et al. (2025)), MedXpertQA (Zuo et al. (2025); Huang et al. (2025))	Additive PEFT	LoRA (Hu et al. (2022a)), AdaLoRA (Zhang et al. (2023)), DoRA (Liu et al. (2024a)), MiSS (Kang & Yin (2025)), VeRA (Kopiczko et al. (2024)), PiSSA (Meng et al. (2024)), MiLoRA (Wang et al. (2025))
		Orth. PEFT	OFT (Qiu et al. (2025))
		Act. PEFT	IA^3 (Liu et al. (2022a))
General	IFEval (Zhou et al. (2023)), NQ (Kwiatkowski et al. (2019)), BBH (Suzgun et al. (2022))	Training	OpenR1-Math (Hugging Face (2025)), MedThink (Gai et al. (2025))
		RL	GRPO (Shao et al. (2024))

Table 1: Experimental setup summary. See Appendix C for full details.

3.2 MAIN RESULTS

Supervised FineTuning (SFT). As shown in Table 2, all methods exhibit a clear stability-plasticity trade-off under SFT. Full FT achieves the largest target gains, but these gains come along with substantial general ability degradation. For example, on Qwen2.5-7B, Full FT increases math target accuracy by 15.33 *percentage points* and medical target accuracy by 7.27, while general performance decreases by 12.74 and 12.56, respectively. On Llama3.2-3B-Instruct (medical setting), the general accuracy decreases by 30.73. These results confirm that single-axis reporting on target tasks is insufficient for evaluating post-training quality.

Method	Hy. Param	Tr. Param (Qwen)	Qwen2.5-7B-base				Tr. Param (Llama)	Llama3.2-3B-Instruct			
			Math Target (%)	Math General (%)	Med Target (%)	Med General (%)		Math Target (%)	Math General (%)	Med Target (%)	Med General (%)
Supervised FineTuning (SFT)											
Base	–	7.61B	35.30 (+0.00)	46.97 (+0.00)	46.36 (+0.00)	46.97 (+0.00)	3.21B	27.63 (+0.00)	53.03 (+0.00)	41.44 (+0.00)	56.76 (+0.00)
Full FT	–	6.53B	50.63 (+15.33)	34.22 (-12.74)	53.63 (+7.27)	34.41 (-12.56)	2.85B	33.90 (+6.27)	39.83 (-13.20)	44.26 (+2.82)	26.03 (-30.73)
OFT	block 16	8.49M	42.33 (+7.03)	42.58 (-4.39)	46.17 (-0.19)	45.09 (-1.88)	7.08M	29.43 (+1.80)	41.08 (-11.95)	39.22 (-2.22)	40.97 (-15.79)
OFT	block 32	17.55M	46.93 (+11.63)	44.37 (-2.60)	48.63 (+2.27)	42.40 (-4.57)	11.55M	30.60 (+2.97)	40.73 (-12.30)	39.50 (-1.94)	40.50 (-16.26)
OFT	block 64	35.68M	46.23 (+10.93)	35.97 (-11.00)	49.47 (+3.11)	39.11 (-7.86)	24.97M	29.30 (+1.67)	39.75 (-13.28)	40.77 (-0.67)	37.70 (-19.06)
OFT	block 128	71.92M	47.77 (+12.47)	36.98 (-9.99)	52.40 (+6.04)	36.88 (-10.08)	47.34M	32.23 (+4.60)	36.26 (-16.76)	42.17 (+0.73)	34.26 (-22.50)
LoRA	r4a8	10.09M	42.33 (+7.03)	41.66 (-5.31)	47.12 (+0.76)	36.42 (-10.55)	6.9M	24.30 (-3.33)	35.79 (-17.23)	36.92 (-4.52)	31.84 (-24.92)
LoRA	r8a16	20.19M	42.47 (+7.17)	39.22 (-7.75)	47.91 (+1.55)	36.06 (-10.91)	12.16M	24.07 (-3.56)	36.57 (-16.46)	38.34 (-3.10)	27.99 (-28.77)
LoRA	r16a32	40.37M	44.87 (+9.57)	34.91 (-12.06)	47.86 (+1.51)	34.86 (-12.11)	24.31M	24.97 (-2.66)	37.55 (-15.48)	39.21 (-2.23)	29.18 (-27.58)
LoRA	r32a64	80.74M	45.37 (+10.07)	38.21 (-8.76)	49.48 (+3.12)	35.50 (-11.47)	48.63M	25.90 (-1.73)	37.20 (-15.83)	39.33 (-2.11)	30.69 (-26.07)
AdaLoRA	r8a16	30.28M	40.43 (+5.13)	35.41 (-11.56)	45.22 (-1.13)	37.34 (-9.63)	18.24M	20.83 (-6.80)	34.53 (-18.49)	37.11 (-4.33)	36.29 (-20.47)
PiSSA	r8a16	20.19M	44.53 (+9.23)	24.78 (-22.19)	26.16 (-20.19)	18.05 (-28.92)	12.16M	0.67 (-26.96)	7.08 (-45.95)	21.17 (-20.27)	9.75 (-47.02)
MiLoRA	r8a16	20.19M	42.60 (+7.30)	37.62 (-9.35)	46.83 (+0.48)	35.88 (-11.09)	12.16M	23.60 (-4.03)	35.59 (-17.44)	37.64 (-3.81)	29.23 (-27.53)
MiSS	r8	11.12M	43.17 (+7.87)	39.12 (-7.85)	48.75 (+2.40)	34.43 (-12.54)	6.19M	23.37 (-4.26)	33.93 (-19.09)	40.16 (-1.28)	31.71 (-25.05)
MiSS	r64	89.00M	46.93 (+11.63)	32.77 (-14.20)	51.90 (+5.54)	32.72 (-14.25)	49.55M	28.63 (+1.00)	34.96 (-18.06)	41.96 (+0.52)	27.07 (-29.69)
VeRA	r256	1.44M	39.43 (+4.13)	47.25 (+0.38)	28.50 (-17.85)	47.01 (+0.04)	0.82M	28.80 (+1.17)	46.79 (-6.23)	40.68 (-0.76)	48.94 (-7.82)
DoRA	r8a16	21.58M	42.33 (+7.03)	40.25 (-6.72)	48.04 (+1.69)	36.06 (-10.91)	12.93M	23.83 (-3.80)	35.65 (-17.37)	38.25 (-3.19)	27.53 (-29.23)
IA^3	–	1.82M	35.13 (-0.17)	44.71 (-2.26)	29.70 (-16.66)	45.72 (-1.25)	0.92M	29.70 (+2.07)	45.72 (-7.30)	39.13 (-2.31)	45.67 (-11.09)
RLVR with Group Relative Policy Optimization (GRPO)											
Full FT	–	6.53B	47.57 (+12.27)	48.68 (+1.71)	46.24 (-0.11)	43.22 (-3.75)	2.85B	29.80 (+2.17)	52.20 (-0.82)	45.88 (+4.44)	51.81 (-4.95)
OFT	block 32	8.49M	48.37 (+13.07)	48.90 (+1.93)	46.79 (+0.44)	47.24 (+0.27)	11.55M	29.97 (+2.34)	50.04 (-2.98)	44.99 (+3.55)	52.31 (-4.45)
LoRA	r8a16	20.19M	46.10 (+10.80)	48.27 (+1.30)	47.08 (+0.73)	42.80 (-4.17)	24.31M	28.83 (+1.20)	52.17 (-0.85)	46.24 (+4.80)	53.54 (-3.23)

Table 2: Main benchmark results. For each domain, average task accuracy is reported in % (higher is better). We also report the absolute change relative to the corresponding base model in parentheses (percentage points, pp).

Across PEFT methods, plasticity and stability vary systematically with parameterization and capacity. Larger trainable capacity often improves target gains (e.g., MiSS $r8$: +7.87 vs. MiSS $r64$: +11.63 on Qwen math target). Some SVD-initialized variants are unstable in this setting: PiSSA $r8$ improves Qwen math target by 9.23 but decreases Qwen general and medical target by 22.19 and 20.19, respectively; on Llama, general drops by 45.95.

Among PEFT baselines, OFT provides the strongest overall frontier under SFT. On Qwen, OFT with block_size $= 32$ improves math target by 11.63 with a 2.60 general decrease, and improves medical target by 2.27 with a 4.57 general decrease. VeRA preserve s general capability best (Qwen math/medical general: +0.38/+0.04), but sacrifices plasticity (e.g., Qwen medical target: -17.85). IA^3 shows a similar pattern of small forgetting but limited target gains.

Reinforcement Learning with Verifiable Rewards (RLVR). Among PEFT baselines, OFT provides the strongest overall frontier under SFT. On Qwen, OFT (block 32) improves math target by 11.63 with a 2.60 general decrease, and improves medical target by 2.27 with a 4.57 general decrease. VeRA preserves general capability best without degradation, but sacrifices plasticity (e.g., Qwen medical target decreased by 17.85). IA^3 similarly shows relatively small forgetting but limited target gains. Furthermore, RLVR post-training imposes minimal requirements on parameter capacity (Schulman & Lab (2025)) for PEFT methods. Both LoRA (rank $= 8$) and OFT (block_size $= 32$) remain competitive with Full FT despite using orders of magnitude fewer trainable parameters, suggesting that the on-policy RL objective can be captured effectively by structured or low-rank updates.

Findings and takeaways. Overall, single-point comparisons at the final checkpoint can obscure the full stability–plasticity frontier, motivating us to move beyond a single operating point and explicitly trace trade-off *curves*. Moreover, the strong stability of OFT suggests that update *geometry* is a key factor; in the next section we connect this behavior to spectrum-level retention/adaptation profiling. Finally, our results also hint at an *overshoot* effect under SFT—the final checkpoint can lie past the best trade-off point—which we later address through interpolation-based and geometry-aware post-hoc adjustment.

Next, we analyze PEFT model geometry to further explain the causes of plasticity-stability trade-off.

4 Spectral Analysis of PEFT Adaptation

Existing work has studied the gap between PEFT and full finetuning by spectral analysis of weight updates. Building on this line, we present a *Spectral Retention-Adaptation Profiling* that connects different PEFT mechanisms to the stability–plasticity trade-off through two complementary dimensions: **Retention** (how much pre-trained structure is preserved) and **Adaptation** (how update energy is distributed), with a specific focus on **spectral smoothness** (local fluctuation across singular-value indices).

Profile I: Diagonal projection on the pre-trained basis (Retention). Let the pre-trained weight be decomposed as $W_{\text{pre}} = U\Sigma V^\top$. We measure how much the fine-tuned weight W_{ft} preserves the pre-trained singular alignment via the diagonal projection

$$P_{\text{diag}}(i) = u_i^\top W_{\text{ft}} v_i. \tag{1}$$

We report $|\Delta P_{\text{diag}}(i)|$ to quantify **component-wise retention**: larger, less consistent deviations indicate stronger interference with pretrained knowledge structure.

Profile II: Update energy spectrum (Adaptation). To capture where the optimization injects energy, we analyze the weight update $\Delta W = W_{\text{ft}} - W_{\text{pre}}$ along the pre-trained input directions v_i:

$$E_\Delta(i) = \|\Delta W v_i\|_2. \tag{2}$$

Compared with $P_{\text{diag}}(i)$, $E_\Delta(i)$ captures a fuller adaptation magnitude because it includes both scaling changes and orthogonal rotations along the i-th latent direction.

Spectral smoothness and fluctuation modes. For both $|\Delta P_{\text{diag}}(i)|$ and $E_\Delta(i)$, we compute a **fluctuation score** as the distance to a moving average over singular-value indices. This indicates whether adjacent components are updated coherently (smooth) or unevenly (spiky). We summarize the observed patterns into four modes, these modes are used to represent tendencies rather than hard categories:

Mode A: Plasticity-dominated The update energy varies smoothly across indices, while $|\Delta P_{\text{diag}}(i)|$ is jagged. This often suggests globally distributed updates that may affect the pre-trained alignment in a more index-local manner; a plausible explanation is that a larger fraction of learning happens off-diagonal as a feature re-allocation rather than by rescaling the same components.

Mode B: Coherent retention, sparse adaptation $|\Delta P_{\text{diag}}(i)|$ is smooth, but $E_\Delta(i)$ has sparse spikes. This pattern is consistent with preserving the global pre-trained scaling structure while injecting strong, orthogonal updates into a few "latent slots" for task-specific knowledge.

Mode C: Global systematic adaptation Both profiles vary smoothly across indices, indicating coherent, systematic transformation (e.g., style/domain shift), where similar features are updated in a similar way, which is often associated with stable optimization behavior.

Mode D: Stochastic degradation Both metrics are highly fluctuating across indices, suggesting that adjacent components are treated inconsistently. This pattern tends to correlate with unstable fitting and a higher likelihood of forgetting, although the severity depends on data/scale/regularization.

4.1 Interpreting PEFT Stability-Plasticity

We apply this framework to compare finetuning mechanisms under *SFT* and *RLVR* (Shao et al. (2024)).

Full finetuning. Under **SFT**, Full FT tends to exhibit **Mode A**-like behavior: broad, high-rank adaptation with more local retention fluctuations. This is consistent with SFT often requiring holistic adjustment across both principal (semantic core) and minor (details) components. On **RLVR**, in contrast, Full FT more often resembles **Mode C**: both retention and adaptation evolve more smoothly. This matches the on-policy nature of RLVR updates, empirically with less destructive interference and better generalization (Zhu et al. (2025)).

Low-rank additive PEFT. LoRA (Hu et al. (2022a)) and its variants (e.g., AdaLoRA (Zhang et al. (2023)), MISS (Kang & Yin (2025))) approximate ΔW with a low-rank additive matrix, which naturally concentrates update energy into a small set of directions. Under **SFT**, this approximation frequently shows a **Mode D** tendency: both profiles become spiky, suggesting a geometric mismatch between the low-rank additive subspace and the pre-trained spectral geometry. Under **RLVR**, however, **LoRA** often shifts toward **Mode B**: retention remains relatively coherent, while adaptation is sparse and targeted. This indicates that although LoRA is less plastic than Full FT in parameter space, it can still achieve task-specific alignment efficiently in RLVR while preserving the pre-trained structure.

Explicit subspace partitioning. PiSSA (Meng et al. (2024)) and MiLoRA (Wang et al. (2025)) target specific spectral regions, but under **SFT** they still lean toward **Mode D**. A likely reason is that SFT couples these regions: changing principal components increases interference, while restricting to minor components limits adaptation.

Fixed update directions. VeRA (Kopiczko et al. (2024)) freezes update directions via random high-rank matrices, usually giving lower plasticity but also lower interference because updates remain closer to orthogonal to dominant pretrained components.

Spectrum Preservation of Orthogonal Finetuning. OFT (Qiu et al. (2023); Liu et al. (2024b); Qiu et al. (2025)) shows **Mode C**-like behavior in **both SFT and RLVR**. This is consistent with

its spectrum-preserving nature: applying an orthonormal transformation primarily rotates weight vectors without strongly distorting their spectral geometry, providing a stability-inducing "geometric barrier." As a result, OFT can achieve strong adaptation in both settings, and remains more robust to SFT overshoot, mitigating forgetting.

Understanding OFT Forgetting through singular vector alignment (SVA). Unlike additive updates, OFT is not naturally characterized by ΔW. We therefore analyze OFT through *Singular Vector Alignment* (SVA). Let $W_{\text{pre}} = U_0 \Sigma_0 V_0^\top$. OFT approximately induces a rotation on the right singular vectors,

$$V^* \approx V_0 R, \quad R^\top R = I, \tag{3}$$

which gives $W^* \approx U_0 \Sigma_0 (V_0 R)^\top$. We measure rotation intensity by cosine similarity between V_0 and V^* at matched indices. As shown in Figure 3, RL training yields nearly uniform rotation across components, while SFT shows spikes in specific layers/components. This explains why OFT mitigates but does not fully remove SFT forgetting.

5 Interpolation Trade-off Curves

Model-merging methods (Yu et al. (2024); Yadav et al. (2023); Matena & Raffel (2022)) expose cross-task trade-offs by combining weights from specialized models. We adopt this idea to analyze the plasticity–stability frontier in SFT: instead of evaluating only the final checkpoint, we evaluate a continuum of interpolated models to trace a Target–General trade-off curve.

Task arithmetic (Ilharco et al. (2023)). Let W_0 be the pre-trained weights and W^* be the aligned weights after SFT on the target domain. The canonical interpolation ("task arithmetic") is

$$W(\alpha) = W_0 + \alpha(W^* - W_0), \qquad \alpha \in [0,1]. \tag{4}$$

Here, $\alpha = 0$ recovers the base model and $\alpha = 1$ recovers the fully adapted model. Sweeping α often yields an explicit trade-off curve and often reveals a "sweet spot" ($\alpha < 1$), where target gains are retained while forgetting is reduced. For additive PEFT methods(LoRA/AdaLoRA/MiSS), interpolation scales the learned update ΔW by α. For OFT, we scale the skew-symmetric generator Q in the Cayley parameterzation (details in Appendix D).

5.1 Results: The Over-Shoot SFT Adaption

Interpolation reveals a "free lunch" region. As shown in Figure 1, we construct interpolation-based trade-off curves by sweeping $\alpha \in \{0.1, 0.2, ..., 0.9\}$ on Qwen2.5-7B for math/medical targets. We include two configurations with comparable parameter budgets (shown as solid vs. dashed curves).

Across both additive (LoRA-family) and orthogonal (OFT) updates, back-interpolating from the final checkpoint ($\alpha = 1$) often markedly improves stability. As shown in Figure 5 (Appendix), many methods achieve near-lossless trade-off gains around $\alpha \approx 0.2$–0.3: target performance stays close to (or slightly above) the final checkpoint while general abilities recover substantially. This indicates that SFT often *overshoots* the best stability–plasticity operating point. A simple one-dimensional interpolation can move the model back to a better operating region.

OFT is Pareto-optimal under interpolation. Overall, OFT yields the most favorable Pareto frontier: for both domains and for both parameter scales, its interpolation curves lie closer to the upper-right corner, achieving stronger target gains at a smaller stability cost than additive PEFT baselines.

Optimization vs. interpolation. We compare the optimization trajectory (training checkpoints) with the interpolation trajectory between W_0 and W^*. They are misaligned and show opposite curvature: optimization is approximately concave with general ability typically drops first, and target performance improves later. In contrast, interpolation is typically convex. As α increases, target performance often improves first. General ability then drops sharply.

This pattern suggests that the overshoot phenomenon is not a simple overfitting issue. Instead, interpolation provides a practical trade-off knob, as well as a diagnostic to reveal stability lost.

5.2 iOFT and Alternative Trade-off Methods

To further analyze overshoot and "free lunch", we go beyond a single global interpolation coefficient and explore geometry-aware adjustment and structured pruning on OFT updates. **iOFT is the main focus of this section**, while the other alternatives as comparisons. Figure 7 (Appendix) and Table 3 summarize these alternatives.

Variant	Target(Math)	General(Math)	Target(Med)	General(Med)
Finetuned	46.93 (+11.63)	44.37 (-2.60)	48.63 (+2.27)	42.40 (-4.57)
Uniform scaling 0.8	46.93 (+11.63)	44.82 (-2.15)	49.17 (+2.81)	45.76 (-1.21)
Uniform scaling 0.4	48.10 (+12.80)	47.53 (+0.56)	48.83 (+2.47)	48.15 (+1.18)
Adaptive scaling (SafeRho)	47.17 (+11.87)	46.69 (-0.28)	50.01 (+3.65)	47.61 (+0.64)
Adaptive scaling (MinRho)	47.83 (+12.53)	46.86 (-0.11)	49.76 (+3.41)	47.79 (+0.82)
BlockDrop 0.2	46.13 (+10.83)	45.70 (-1.64)	48.64 (+2.29)	46.84 (-0.50)
BlockDrop 0.4	44.23 (+8.93)	47.27 (-0.07)	49.86 (+3.50)	47.45 (+0.10)
LayerDrop 18-27	47.03 (+11.73)	46.07 (-1.27)	49.26 (+2.90)	47.35 (+0.01)

Table 3: Ablation of Trade-off Alternatives on OFT Updates.

iOFT: interpolated OFT via layer-wise geometric adjustment. Empirically, we observe that OFT exhibits a pronounced layer profile of rotation strength: later layers rotate much more than early layers, and this gap widens during training. Motivated by this intrinsic update structure, we introduce **iOFT** (interpolated OFT), which rebalances the update by rescaling the generator Q_ℓ per layer so that the effective rotation strengths become more uniform. Concretely, iOFT instantiates two simple reference rules: **SafeRho** matches each layer to the average rotation strength of the first five layers, while **MinRho** matches each layer to the minimum rotation strength across layers. Because iOFT leverages the model's own weight-update geometry, it exhibits more robust trade-off improvements across settings: it can consistently recover general abilities while preserving target gains without requiring an exhaustive sweep over a global scaling factor.

Figure 2 (Appendix) shows that forgetting correlates with *spiky* and *incoherent* spectral deviations in both $|\Delta P_{\text{diag}}(i)|$ and $E_\Delta(i)$. Large layer-wise spikes in rotation strength ρ_ℓ indicate that a few layers (usually later ones) undergo excessive geometric re-allocation, leading to uneven spectral changes and more destructive interference. Flattening ρ_ℓ via layer-wise rescaling of Q_ℓ reduces these spikes, giving a mechanism-level explanation of iOFT's stability gains.

Global scaling requires tuning α across training settings. In contrast, when we apply a uniform scaling to the entire OFT update (analogous to choosing a global interpolation coefficient α), the optimal trade-off is sensitive to the training setting and the overall update magnitude. In practice, selecting the best global scale typically requires enumerating candidate values and evaluating them to find the best α.

Block/layer dropping as coarse alternatives. We also test random block dropping (BlockDrop) and selected layer removal (LayerDrop) on OFT updates. While these methods are generally less effective than iOFT in improving the target–general frontier, they can still mitigate forgetting and often retain a large portion of the target improvements.

Implications. Taken together, these results suggest that overshoot is closely tied to the geometry and distribution of the learned update across layers. Our exploration provides additional motivation and practical references for future work on improving PEFT stability–plasticity trade-offs by exploiting parameter geometry, both within layers (intra-layer) and across layers (inter-layer), rather than relying solely on global scaling and post-hoc α selection.

6 Concluding Remarks

We present **PEFT-Arena**, a benchmark that evaluates PEFT on both **plasticity** (target gains) and **stability** (general ability preservation) across math/medical adaptation settings. Under comparable parameter budgets, **OFT** consistently achieves a stronger Pareto frontier than additive low-rank baselines.

To connect these empirical trade-offs to internal mechanisms, we develop a **spectral retention-adaptation profiling**, which separates **retention** (alignment with the pretrained singular basis) from **adaptation** (distribution of update energy). This analysis suggests two key findings: (i) a **geometric barrier**, where structure-preserving, rotation-like updates maintain more coherent spectral deviations and thus reduce destructive drift; and (ii) a tendency of many constrained additive parameterizations to exhibit **spiky** and **incoherent** spectral changes under SFT, which correlates with increased forgetting.

Finally, we identify a robust **SFT overshoot**: the final checkpoint ($\alpha = 1$) often goes beyond the best stability-plasticity point. To make this explicit, we use **interpolation** between the base and adapted weights to obtain **Interpolation Trade-off Curves**, which frequently uncover a "free-lunch" region that recovers general abilities with little or no target loss. Building on this observation, we propose **iOFT**, a simple empirical, geometry-aware variant that rebalances OFT updates via layer-wise adjustment, providing a practical robust trade-off knob beyond sweeping a global scaling coefficient.

REFERENCES

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. In *ACL*, 2022.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 3563–3599, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.182. URL https://aclanthology.org/2025.naacl-long.182/.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. Parameter-efficient fine-tuning design spaces. In *ICLR*, 2023.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, 2022.

Yupeng Chen, Senmiao Wang, Yushun Zhang, Zhihang Lin, Haozhe Zhang, Weijian Sun, Tian Ding, and Ruoyu Sun. Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning. *arXiv preprint arXiv:2407.20999*, 2024.

Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang, Zhonghao Yan, Yufei Han, Kongming Liang, Weiran Xu, and Zhanyu Ma. Entropy-adaptive fine-tuning: Resolving confident conflicts to mitigate forgetting. *arXiv preprint arXiv:2601.02151*, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J Clark, and Mehdi Rezagholizadeh. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650*, 2022.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. MedThink: a rationale-guided framework for explaining medical visual question answering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the association for computational linguistics: NAACL 2025*, pp. 7438–7450, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.415. URL https://aclanthology.org/2025.findings-naacl.415/.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *EMNLP*, 2021.

Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*, 2020.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International conference on learning representations*, 2022a. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022b.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning in large language models, 2025. URL https://arxiv.org/abs/2504.00869.

Hugging Face. Open R1: a fully open reproduction of DeepSeek-R1, January 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *AAAI*, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421. Number: 6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 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)*, pp. 2567–2577, 2019.

Jiale Kang and Qingyu Yin. MiSS: Revisiting the trade-off in LoRA with an efficient shard-sharing structure, 2025. URL https://arxiv.org/abs/2409.15371. arXiv: 2409.15371 [cs.CL].

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *NeurIPS*, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Anat Kleiman, Gintare Karolina Dziugaite, Jonathan Frankle, Sham Kakade, and Mansheej Paul. Soup to go: mitigating forgetting during continual learning with model averaging. *arXiv preprint arXiv:2501.05559*, 2025.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *The twelfth international conference on learning representations*, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, November 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00276. URL https://direct.mit.edu/tacl/article/43518.

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

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*, 2022.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jiacheng Lin, Zhongruo Wang, Kun Qian, Tian Wang, Arvind Srinivasan, Hansi Zeng, Ruochen Jiao, Xie Zhou, Jiri Gesi, Dakuo Wang, et al. Sft doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in llms. *arXiv preprint arXiv:2509.20758*, 2025.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 580–606, 2024.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *NeurIPS*, 2022a.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024a.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. Parameter-efficient orthogonal finetuning via butterfly factorization. In *ICLR*, 2024b.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022b.

Ekdeep Singh Lubana, Puja Trivedi, Danai Koutra, and Robert Dick. How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation. In *Conference on Lifelong Learning Agents*, pp. 819–837. PMLR, 2022.

Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. Keeplora: Continual learning with residual gradient adaptation. *arXiv preprint arXiv:2601.19659*, 2026.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:2110.07577*, 2021.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, 2022.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. *Advances in Neural Information Processing Systems*, 37:121038–121072, December 2024. doi: 10.52202/079017-3846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in psychology*, 4:504, 2013.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the conference on health, inference, and learning*, volume 174 of *Proceedings of machine learning research*, pp. 248–260. PMLR, April 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms. *arXiv preprint arXiv:2406.16797*, 2024.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2023.

Zeju Qiu, Weiyang Liu, Adrian Weller, and Bernhard Schölkopf. Orthogonal finetuning made scalable. In *EMNLP*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: a graduate-level google-proof Q&a benchmark. In *First conference on language modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. Upweighting easy samples in fine-tuning mitigates forgetting. *arXiv preprint arXiv:2502.02797*, 2025.

John Schulman and Thinking Machines Lab. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. https://thinkingmachines.ai/blog/lora/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Kexuan Shi, Yandong Wen, and Weiyang Liu. Model merging with functional dual anchors. *arXiv preprint arXiv:2510.21223*, 2025.

Daniel L Silver and Robert E Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pp. 90–101. Springer, 2002.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *NeurIPS*, 2021.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *NeurIPS*, 2022.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 99–108, 2022.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *ACL*, 2022.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. Milora: Harnessing minor singular components for parameter-efficient llm finetuning, 2025. URL https://arxiv.org/abs/2406.09044.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024a.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. In *EMNLP*, 2022.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: a more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in neural information processing systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

A Yang, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and others. Qwen2. 5 technical report. *arXiv preprint*, 2024.

Sihan Yang, Kexuan Shi, and Weiyang Liu. Orthogonal model merging. *arXiv preprint arXiv:2602.05943*, 2026.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *ACL*, 2022.

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

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911. arXiv: 2311.07911 [cs.CL].

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. June 2025. URL https://openreview.net/forum?id=IyVcxU0RKI.

A Analysis Figures

This section collects the visualization figures for the spectral and geometric analyses discussed in section 4. Figure 2 presents the dual-view spectral analysis results, including the effective rank of weight updates and the distributions of projected spectrum changes under both the retention and adaptation views, as well as their spectral smoothness. Figure 3 visualizes the singular vector alignment (SVA) analysis of OFT, illustrating how OFT's rotation behavior differs across layers and training paradigms (SFT vs. RLVR).

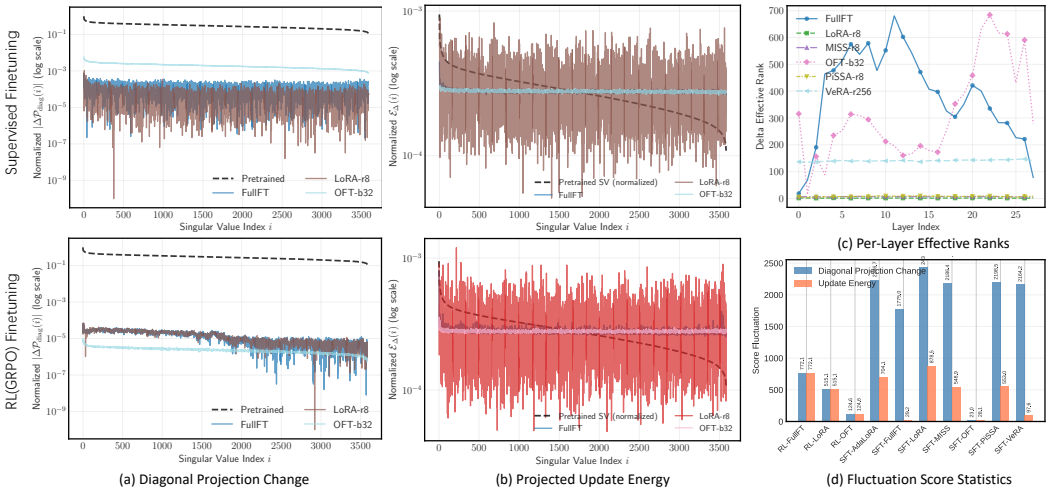


Figure 2: Dual-view spectral analysis. Left: effective rank of weight updates. Mid & right: distributions of projected spectrum changes (retention/adaptation views) and their spectral smoothness.

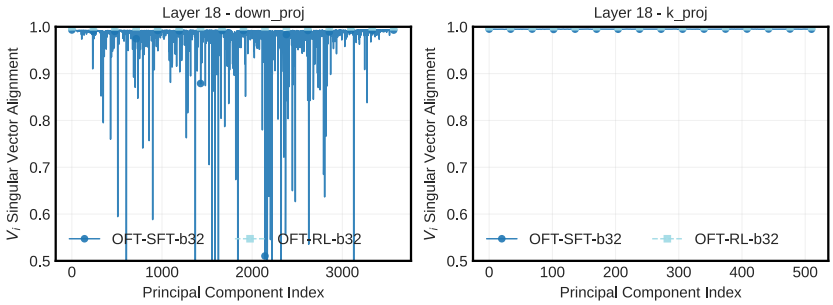


Figure 3: Singular vector alignment (SVA) analysis of OFT updates in different layers.

B Interpolation Trade-off Figures

This section provides the detailed visualization figures for the interpolation trade-off analysis in section 5. Figure 4 shows the interpolation trade-off curves for different PEFT methods, demonstrating that OFT achieves better "sweet spots" under the same parameter scale. Figure 5 presents the target-general performance tradeoff with dual-axes visualization, revealing the "lossless" improvement region at moderate interpolation coefficients. Figure 6 compares the optimization trajectories and interpolation trajectories, further highlighting the overshoot issue of SFT.

C Experimental Setup and Evaluation Details

This section provides the full details of the experimental setup summarized in Table 1.

Target-domain benchmarks. **(1) Math.** We evaluate on Math-500 (Lightman et al. (2023)), AMC23, and AIME24. Each math problem is evaluated by average accuracy@16, with a maximum response length of 8192 tokens and temperature $T = 0.6$. **(2) Medical.** We evaluate on a collection

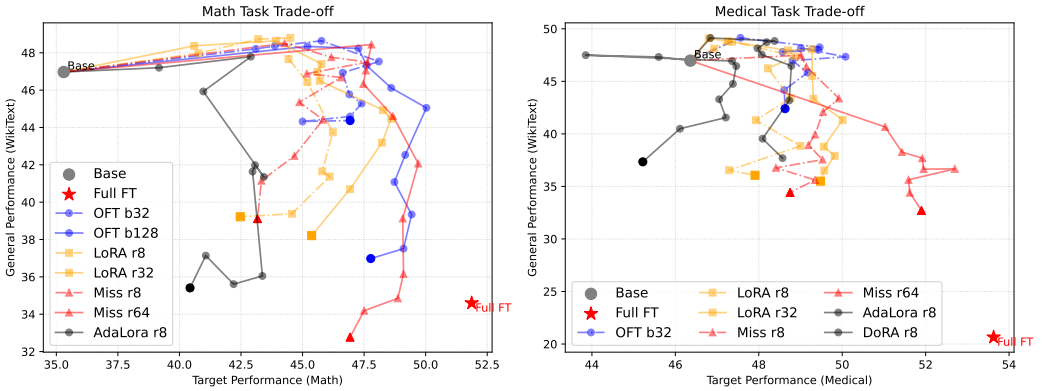


Figure 4: Interpolation PEFT trade-off curves. OFT achieves better "sweet spot" under the same parameter scale.

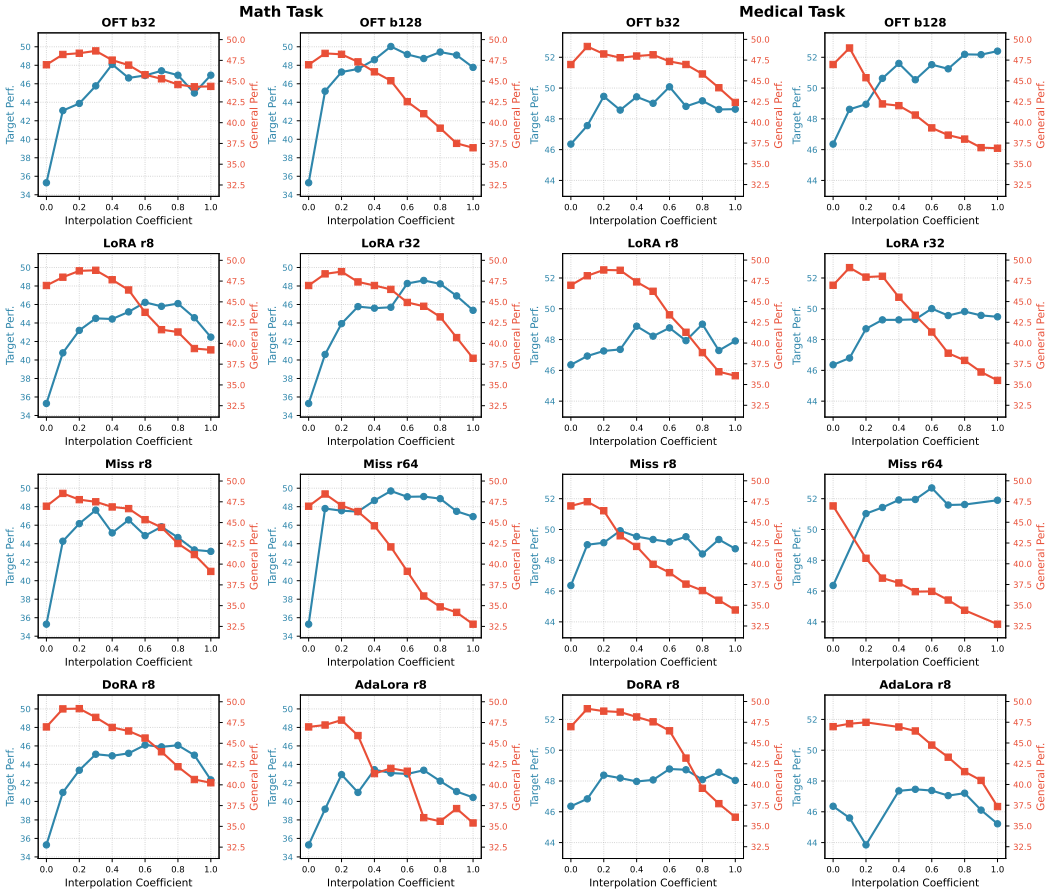


Figure 5: Target-general performance tradeoff with α-interpolation for PEFT methods

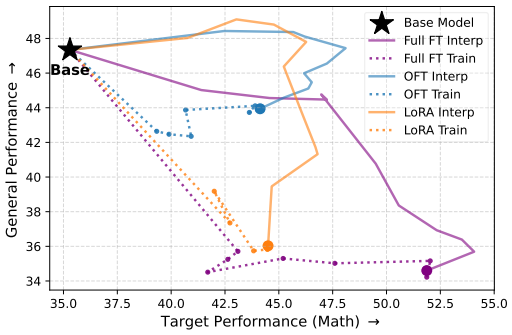


Figure 6: Comparing training trajectories and interpolation trajectories. The misalignment further reveal the overshoot issue of SFT.

of medical reasoning and knowledge benchmarks, including MedMCQA (Pal et al. (2022)), MedQA (USMLE) (Jin et al. (2021)), PubMedQA (Jin et al. (2019)), MMLU-Pro (Wang et al. (2024b)), GPQA (Medical) (Rein et al. (2024)), Lancet, NEJM & MedBullets problems (Chen et al. (2025)), and MedXpertQA (Zuo et al. (2025)) following a dedicate dataset survey (Huang et al. (2025)). We use average accuracy as the primary metric with temperature $T = 0.0$.

General ability preservation. To measure general ability preservation after adaptation, we evaluate on IFEval (Zhou et al. (2023)), NQ (Kwiatkowski et al. (2019)), BBH (Suzgun et al. (2022)), covering instruction following, natural language understanding, general knowledge and general reasoning. We follow the evaluation configuration in OpenCompass[1] with context length 1024, temperature $T = 0.0$, and one sample per query.

Base models and adaptation methods. We use Qwen2.5-7B (Yang et al. (2024)) and Llama3.2-3B-Instruct (Dubey et al. (2024)) as pre-trained backbones to cover different scales and base/instruction-tuned settings. We compare full finetuning (Full FT) against a representative set of PEFT baselines (see section 2 for background). **Additive PEFT (LoRA family).** We include LoRA (Hu et al. (2022a)) and representative variants spanning rank allocation/parameterization/initialization: AdaLoRA (Zhang et al. (2023)), DoRA (Liu et al. (2024a)), MiSS (Kang & Yin (2025)), VeRA (Kopiczko et al. (2024)), PiSSA (Meng et al. (2024)), and MiLoRA (Wang et al. (2025)). **Spectrum Preserving updates (OFT).** We include orthogonal finetuning (OFT) (Qiu et al. (2025)), which constrains updates to structured (approximately) orthogonal transformations. **Activation-based PEFT.** We include IA^3 (Liu et al. (2022a)), a lightweight method that adapts models via learned activation scaling.

Training details. **Compute.** All experiments are conducted on 8× NVIDIA H100 80GB GPUs. **SFT.** We conduct supervised finetuning in both target domains, using 50k filtered samples from OpenR1-Math-330k (Hugging Face (2025)) for math and 23k samples from MedThink (Gai et al. (2025)) for medical. We use an effective batch size of 256, a maximum response length of 8192 tokens, and train for 4 epochs. Full fine-tuning uses a learning rate of 5×10^{-5}, while all PEFT methods use 2×10^{-4}. We adopt a cosine decay learning-rate scheduler. **RL.** We also include RLVR post-training results with GRPO (Shao et al. (2024)) on a representative subset of methods. We use an effective batch size of 256, a mini-batch size of 64, and a group size of 8. The maximum generation length is set to 8192 tokens, consistent with SFT and evaluation. Full fine-tuning uses a learning rate of 10^{-6}, while all PEFT methods use 10^{-5}. We train for 200 steps for all RL experiments.

D Interpolation Implementation Details

Additive-update PEFT. For PEFT methods with an additive parameterization,

$$W^* = W_0 + \Delta W, \qquad W(\alpha) = W_0 + \alpha \Delta W. \tag{5}$$

Interpolation is implemented by scaling the learned update ΔW. For LoRA, $\Delta W = sBA$ (with method-dependent scale s, e.g., $s = \frac{\alpha_{\text{LoRA}}}{r}$). A convenient implementation is to scale both factors,

$$\Delta W(\alpha) = s(\sqrt{\alpha}B)(\sqrt{\alpha}A) = \alpha sBA, \tag{6}$$

which preserves the product structure and avoids excessively scaling a single factor.

[1] https://github.com/open-compass/opencompass

OFT: interpolating rotation by scaling the generator. Using the Cayley form, the rotation is parameterized as

$$R(Q) = (I+Q)(I-Q)^{-1}, \tag{7}$$

where Q is the skew-symmetric generator. We interpolate OFT by scaling its generator,

$$R(\alpha) \triangleq R(\alpha Q) = (I+\alpha Q)(I-\alpha Q)^{-1}, \qquad \alpha \in [0,1]. \tag{8}$$

For small angles, the rotation angle θ satisfies $\|Q\| \approx \tan(\theta/2) \approx \theta/2$, hence θ is approximately linear in $\|Q\|$. Following prior geometric diagnostics, we define the layer-wise rotation strength

$$\rho_\ell \triangleq 1 - \cos(\theta_\ell). \tag{9}$$

Using $\cos\theta \approx 1 - \theta^2/2$ gives $\rho_\ell \approx \theta_\ell^2/2$, and therefore $\rho_\ell \propto \|Q\|^2$. Consequently, to scale the rotation strength by a factor of k in the small-angle regime, we scale the generator by $\sqrt{k}$: $Q' = \sqrt{k}Q$.

E Additional Results

This section provides additional results that complement the analyses in the main text. Figure 8 presents additional visualizations for the spectral analysis in section 4, including distributions of projected update energy and diagonally projected spectrum changes on the pretrained basis. Figure 7 complements the interpolation trade-off discussion in subsection 5.2, visualizing how different scaling/pruning alternatives (uniform scaling, iOFT, BlockDrop, LayerDrop) behave across OFT adapter sizes.

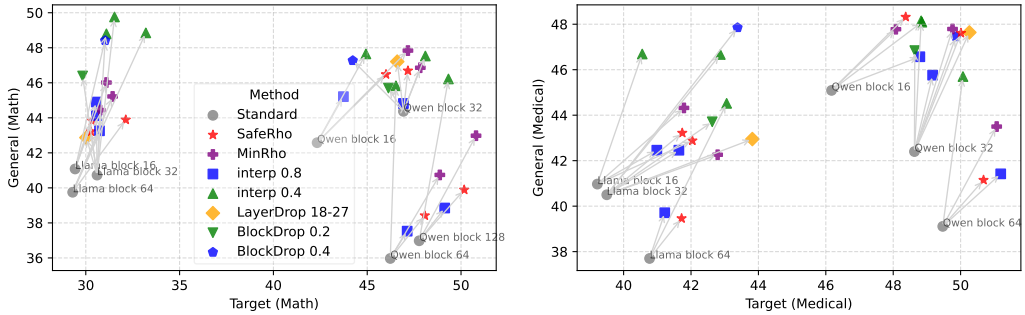


Figure 7: Trade-off alternatives achieve consistent improvement on OFT adapters of different sizes. This figure complements Table 3 by visualizing how different scaling/pruning alternatives behave across adapter sizes.

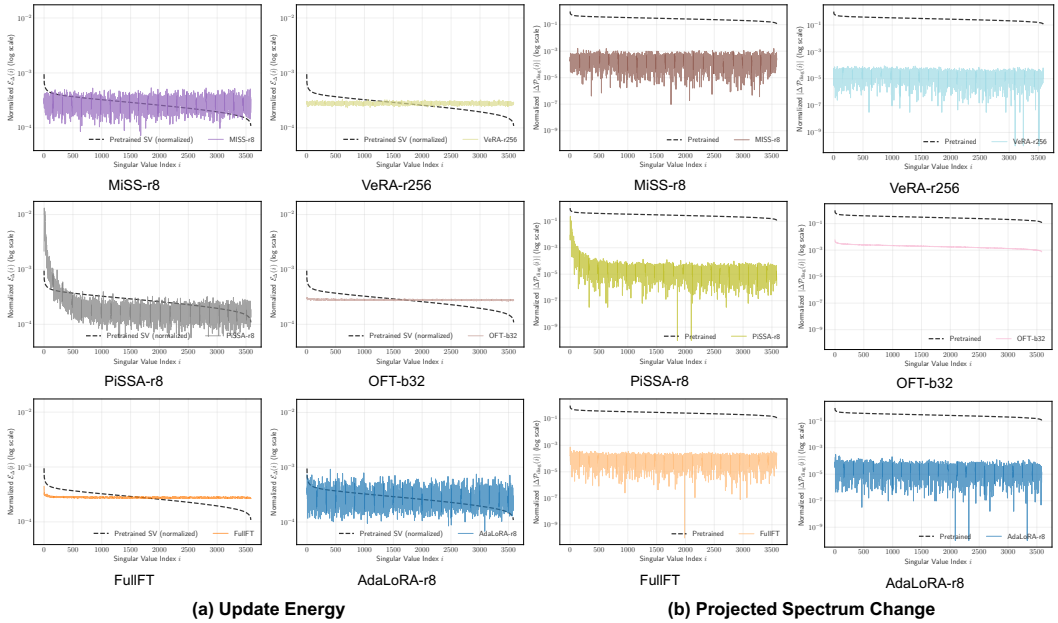


Figure 8: Additional visualization for spectrum analysis in SFT Training. Including distributions of (a) project update energy and (b) diagonally projected spectrum changes on pretrained basis.