SPECTRAL MOMENTUM STEERING MITIGATES OPTIMIZER MISMATCH OF MUON FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has emerged as a compelling alternative to Adam for training matrix-valued parameters and has been adopted in the pretraining of frontier language models. Its use in post-training, however, is hindered by *optimizer mismatch*: fine-tuning an Adam-pretrained checkpoint with Muon can yield worse performance than fine-tuning with Adam itself. Because most publicly available checkpoints are pretrained with Adam, resolving this mismatch would enable broader application of Muon to post-training. Analyzing how pretrained singular directions align with layer representations, we find that pretrained weights read from input directions associated with spectral outliers beyond both edges of the Marchenko–Pastur (MP) distribution, yet write predominantly into the subspace spanned by the principal left singular vectors. Building on this, we propose Spectral Momentum Steering (SMS), which performs steepest descent under a plasticity-weighted anisotropic norm constraint that limits movement along the principal output directions encoding foundational representation structure. Across Adam-pretrained model families, SMS consistently improves pretrained capability preservation while matching or exceeding the downstream adaptation performance of matched Adam or standard Muon fine-tuning, substantially alleviating optimizer mismatch. [Code]

1 INTRODUCTION

Muon (MomentUm Orthogonalized by Newton-Schulz) (Jordan et al., 2024) has recently emerged as a compelling alternative to Adam (Kingma & Ba, 2015; Loshchilov & Hutter, 2017) for training the matrix-valued parameters of neural networks. Unlike Adam, which applies element-wise preconditioning, Muon orthogonalizes the momentum matrix, which can be interpreted as steepest descent under the spectral norm (Chen et al., 2026; Fan et al., 2025). This matrix-level preconditioning equalizes the update magnitude across the singular directions of the momentum, amplifying informative but weak directions that would otherwise be overshadowed by a few dominant ones, and thereby yields faster convergence and improved training stability (Shah et al., 2025; Liu et al., 2025). These advantages have contributed to its adoption in frontier large language model (LLM) pretraining, including the Kimi series (Team-Kimi et al., 2026a;b), GLM-5 (Zeng et al., 2026), DeepSeek-V4 (DeepSeek-AI et al., 2026), and Qwen3.8 (Qiu et al., 2026). However, existing studies of Muon have focused almost exclusively on pretraining, and extending its benefits to post-training remains largely unexplored.

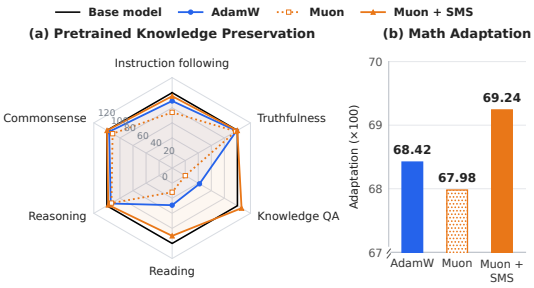

Figure 1: Optimizer mismatch of Muon fine-tuning and its mitigation by spectral momentum steering.

A key obstacle to using Muon for post-training is *optimizer mismatch*: models pretrained with Adam often exhibit lower downstream performance when fine-tuned with Muon than when fine-tuned with Adam, and vice versa. Our results in Fig. 1 illustrate this effect: relative to Adam, Muon causes substantially greater degradation of pretrained capabilities (Fig. 1(a)) as well as lower downstream adaptation performance (Fig. 1(b)). Since most publicly available foundation models are pretrained with Adam, resolving this mismatch is critical for extending Muon's benefits beyond pretraining to the post-training of existing foundation models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Pioneering work (Qu et al., 2026) attributes optimizer mismatch to the distinct implicit biases of the two optimizers. Adam applies element-wise preconditioning, associated with max-norm geometry, whereas Muon applies matrix-level preconditioning associated with spectral-norm geometry. These differences can lead the two optimizers to weight configurations with different structural properties, even when they achieve comparable training losses. When an Adam-pretrained model is fine-tuned with Muon, the distinct implicit biases of the two optimizers can induce updates that are potentially misaligned with the geometry established during Adam pretraining. This mismatch can perturb pretrained weight structure and consequently degrade foundational capabilities (Fig. 1(a)). Although Qu et al. (2026) show that regularization enforced by low-rank adaptation (LoRA) (Hu et al., 2022) partially alleviates the mismatch, we find that Muon still forgets substantially more than Adam under LoRA, and that existing forgetting-aware LoRA methods (Wang et al., 2024; Yang et al., 2024; Tang et al., 2026; Xiong & Xie, 2026) do not close this gap in our experiments. These observations suggest that optimizer mismatch is closely tied to update geometry, motivating an optimizer-level solution.

In this work, we propose *Spectral Momentum Steering* (SMS), an optimizer-level solution to mitigating optimizer mismatch of Muon fine-tuning. To identify an appropriate update geometry for preserving pretrained structure, we analyze how pretrained weights interact with layer representations. Measuring the alignment of the singular vectors with the layer input and output activations reveals an asymmetric spectral alignment pattern. On the input side, the weight *reads* from high-variance input directions aligned with spectral outliers beyond both edges of the Marchenko–Pastur (MP) distribution. On the output side, it *writes* predominantly into the subspace spanned by the principal left singular vectors. We hypothesize that this principal output subspace encodes a substantial portion of the model's foundational knowledge in a geometry shaped by Adam pretraining, and that Muon's update, which approximately equalizes magnitude across the singular directions of the momentum, can disrupt this pretrained structure, causing *catastrophic forgetting* of pretrained knowledge.

Following up on these observations and this hypothesis, we redesign the Muon update to steer movement away from the principal activation subspace established by Adam. SMS assigns each output singular direction of the pretrained weight a plasticity coefficient that decreases with its singular value, and performs steepest descent under an *anisotropic* spectral norm that penalizes movement along directions associated with the principal activation subspace. SMS preserves Muon's efficient Newton–Schulz orthogonalization (Kovarik, 1970) and requires only a one-time SVD of the pretrained weights prior to fine-tuning. Unlike CorDA (Yang et al., 2024) and LoRA-Null (Tang et al., 2026), SMS does not require calibration data serving as a proxy for the pretraining data distribution, which can be difficult to obtain in practice.

Across benchmarks on Adam-pretrained model families, SMS-enhanced Muon consistently outperforms standard Muon and even exceeds the performance of matched Adam-pretraining/Adam-fine-tuning configurations. We attribute these gains to preserving the foundation structure established by the pretraining optimizer. By carefully steering Muon's momentum away from this structure, SMS better preserves the pretrained geometry and, consequently, the foundational knowledge encoded within it. Through careful ablations of the preservation–adaptation trade-off across learning rates and adapter ranks, we further reveal that preserving this foundational structure and the pretrained knowledge it supports is critical for effective downstream adaptation in the post-training pipeline.

Our main contributions are summarized as follows:

- We analyze the spectral structure of pretrained weights and its alignment with layer representations, showing that pretrained weights *read* from input directions associated with spectral outliers beyond the MP bulk yet *write* predominantly into the principal output subspace.
- We propose SMS, an extension of Muon derived as steepest descent under a plasticity-weighted anisotropic spectral norm. SMS steers Muon's momentum away from the principal activation subspace to reduce disruption of the pretrained representation structure.
- We empirically validate that preserving the foundational representation structure during fine-tuning not only improves retention of broad non-target capabilities, but can also benefit downstream adaptation, challenging the conventional stability–plasticity dilemma.
- Across two AdamW-pretrained model families, SMS consistently improves foundation capability preservation while matching or surpassing the downstream performance of AdamW fine-tuning, standard Muon, and the evaluated forgetting-aware fine-tuning baselines, substantially alleviating optimizer mismatch.

2 PRELIMINARIES

Muon Optimizer. Muon (Jordan et al., 2024; Liu et al., 2025) is an optimizer designed for matrix-valued parameters. In particular, Muon can be interpreted as *normalized momentum descent* (NMD) with respect to the spectral norm (Fan et al., 2025; Chen et al., 2026). Consider a weight matrix $W_t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ at iteration t, and let $G_t := \nabla_W \mathcal{L}(W_t)$ denote the corresponding gradient. Given the momentum $M_t = \beta M_{t-1} + G_t$, Muon updates the parameter as

$$W_{t+1} = W_t - \eta_t \Delta_t, \qquad \Delta_t \in \arg\max_{\|\Delta\|_{\text{op}} \le 1} \langle M_t, \Delta \rangle_F , \tag{1}$$

where η_t is learning rate, $\|\cdot\|_{\text{op}}$ denotes the spectral (operator) norm, and $\langle A, B\rangle_F := \text{tr}(A^\top B)$ is the Frobenius inner product. Thus, Δ_t is a unit-spectral-norm direction that maximizes alignment with the momentum M_t, corresponding to the steepest descent direction under the spectral-norm geometry. Let the singular value decomposition (SVD) of the momentum be $M_t = U_t \Sigma_t V_t^\top$, where $\Sigma_t = \text{diag}(\sigma_1, \ldots, \sigma_r)$ and $r = \text{rank}(M_t)$. Then, a solution to Eq. 1 is given by

$$\Delta_t = U_t V_t^\top = \text{Polar}(M_t), \tag{2}$$

i.e., the polar factor of M_t, yielding a unit-spectral-norm steepest-descent direction. We provide a detailed derivation of this result in Appendix C.1. In practice, Muon approximates the polar factor through Newton–Schulz iterations (Kovarik, 1970), $\Delta_t \approx \text{NS}(M_t)$, providing an efficient approximation to the momentum orthogonalization step (Bernstein & Newhouse, 2024). Intuitively, this momentum orthogonalization produces updates whose nonzero singular values are approximately equal, thereby equalizing the update magnitude across the singular directions of the momentum. This singular-value equalization mitigates directional imbalance in parameter updates and can promote better-conditioned weight matrices during optimization, contributing to fast and stable convergence.

Optimizer Mismatch. Liu et al. (2025) first investigated the interchangeability of Adam and Muon across the pretraining and fine-tuning stages, identifying an *optimizer mismatch* phenomenon: models pretrained with Adam exhibit worse fine-tuning performance with Muon than with Adam, and vice versa. Qu et al. (2026) further studied this phenomenon and showed that low-rank adaptation (LoRA) (Hu et al., 2022) can mitigate the mismatch. Specifically, Qu et al. (2026) hypothesize that this incompatibility arises from the distinct implicit biases induced by the two optimizers: Adam employs element-wise preconditioning and can be associated with max-norm geometry, whereas Muon applies matrix-level preconditioning and can be associated with spectral-norm geometry. To illustrate these distinct geometric preferences, consider a learning problem with objective $\mathcal{L}(W)$ and define the set of optimal solutions as

$$\mathcal{S} := \arg\min_W \mathcal{L}(W). \tag{3}$$

In the simplified underdetermined linear-regression setting studied by Qu et al. (2026), SignGD (Bernstein et al., 2018) is used as a proxy for Adam, while Muon is idealized by exact orthogonalization. In this setting, their implicit biases admit the following minimum-norm characterizations:

$$W_A^\star \in \arg\min_{W \in \mathcal{S}} \|W\|_{\max}, \qquad W_\mu^\star \in \arg\min_{W \in \mathcal{S}} \|W\|_{\text{op}}. \tag{4}$$

That is, among the solutions minimizing the same training objective, the Adam proxy, SignGD, favors a minimum max-norm solution, whereas idealized Muon favors a minimum spectral-norm solution. Although this exact characterization is established only in the simplified linear-regression setting, it motivates a geometric interpretation of the distinct implicit biases observed more generally for Adam and Muon: even when they achieve comparable training objective values, they may converge to weight configurations with substantially different geometric structures. Consequently, fine-tuning with a mismatched optimizer may induce parameter updates that are small under the fine-tuning optimizer's native norm but comparatively large under the norm associated with the pretraining optimizer. For example, Muon fine-tuning of an Adam-pretrained model can induce a larger max-norm displacement than the corresponding Adam fine-tuning. Qu et al. (2026) hypothesized that such incompatible updates disrupt the pretrained weight structure and thereby exacerbate *catastrophic forgetting*.

Language Model Forgetting. A growing body of work studies the degradation of pretrained capabilities during language model adaptation. Guo et al. (2025) empirically show that post-training

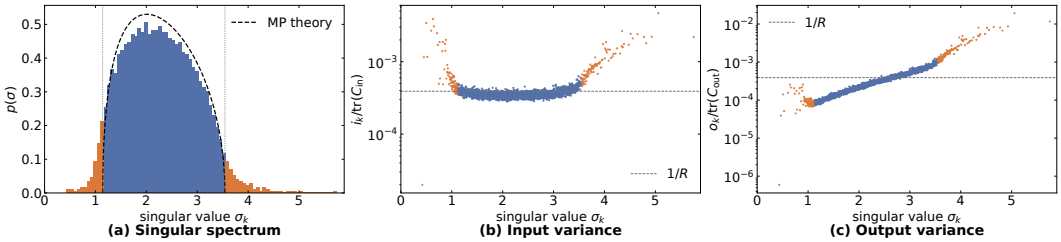

Figure 2: Spectral structure of a pretrained weight matrix. (a) Empirical singular-value distribution and reference Marchenko–Pastur (MP) density; orange denotes spectral outliers outside the MP bulk. (b) Normalized input variance along right singular directions. (c) Normalized output variance along left singular directions.

can impair several foundational capabilities inherited from pretraining. Harmon et al. (2026) further characterize forgetting across domains and find factual knowledge retrieval to be particularly susceptible to degradation. Recent evidence suggests that on-policy adaptation may better preserve pretrained capabilities than off-policy adaptation. For instance, Shenfeld et al. (2026b) observe substantially less forgetting under reinforcement learning than supervised fine-tuning (SFT), while Shenfeld et al. (2026a) shows that on-policy self-distillation (SDFT) can further improve capability retention. A complementary line of work mitigates forgetting in SFT through low-rank adaptation (LoRA). MiLoRA (Wang et al., 2024) performs an SVD of pretrained weight matrices and initializes LoRA adapters along minor singular components, biasing adaptation toward directions that are less prominent in the pretrained weights. CorDA (Yang et al., 2024) and LoRA-Null (Tang et al., 2026) instead identify initialization subspaces using activation covariance statistics estimated from calibration data intended to approximate the pretraining distribution. OPLoRA (Xiong & Xie, 2026) uses hard projection to preserve singular components of pretrained weight matrices during training.

3 Methods

3.1 Observations: Geometry of Pretrained Weights and Layer Representations

Consider a randomly initialized weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with $R = \min(d_{\text{out}}, d_{\text{in}})$ and $n = \max(d_{\text{out}}, d_{\text{in}})$, whose entries are independently and identically distributed with zero mean and finite variance v. In the large-dimensional limit, with R/n approaching a fixed aspect ratio, the empirical singular-value distribution converges to the Marchenko–Pastur (MP) law (Marčenko & Pastur, 1967). Writing $\alpha = R/n$ and $\tau = \sqrt{nv}$, the singular value density is

$$\rho_{\text{MP}}(\sigma) = \frac{\sqrt{(\sigma_+^2 - \sigma^2)(\sigma^2 - \sigma_-^2)}}{\pi \alpha \tau^2 \sigma} \mathbf{1}_{[\sigma_-, \sigma_+]}(\sigma), \qquad \sigma_\pm = \tau(1 \pm \sqrt{\alpha}). \tag{5}$$

Training, however, can introduce correlations and structure into the weights, leading to deviations from this MP bulk. In particular, singular values lying outside the MP spectral boundaries have been associated with learned structure (Martin et al., 2020; Thamm et al., 2022). Fig. 2(a) shows that the central singular-value bulk of a rectangular `Layer-16 mlp.up_proj` weight matrix from Qwen3-4B-Base closely follows the predicted MP density (dashed curve), while distinct outlier populations emerge below $\sigma_- = 1.14$ and above $\sigma_+ = 3.54$, highlighted in orange.

To explore how pretrained weights interact with layer representations, we consider the alignment between the left (output) and right (input) singular vectors of the pretrained weight matrix and the corresponding layer output and input representations, respectively. Let the thin singular value decomposition (SVD) of the pretrained linear layer weight matrix W_0 be $W_0 = U_R \Sigma_R V_R^\top$, where $U_R = [u_1, \ldots, u_R] \in \mathbb{R}^{d_{\text{out}} \times R}, \Sigma_R = \text{diag}(\sigma_1, \ldots, \sigma_R) \in \mathbb{R}^{R \times R}, V_R = [v_1, \ldots, v_R] \in \mathbb{R}^{d_{\text{in}} \times R}$. Let $h \in \mathbb{R}^{d_{\text{in}}}$ denote the token representations entering the linear layer with its mean $\bar{h} = \mathbb{E}[h]$. Define the centered input covariance as

$$C_{\text{in}} = \mathbb{E}\big[(h - \bar{h})(h - \bar{h})^\top\big]. \tag{6}$$

Because the right singular vectors v_k lie in the input space of W_0, we quantify their alignment with the input representations by measuring the variance captured along each input singular direction:

$$i_k = v_k^\top C_{\text{in}} v_k, \qquad \widehat{i}_k = \frac{i_k}{\text{tr}(C_{\text{in}})}. \tag{7}$$

Similarly, we quantify the alignment of each left singular vector u_k with the output representations:

$$o_k = u_k^\top C_{\text{out}} u_k, \qquad \widehat{o}_k = \frac{o_k}{\text{tr}(C_{\text{out}})}, \tag{8}$$

where $C_{\text{out}} = W_0 C_{\text{in}} W_0^\top$. Fig. 2(b) and (c) report these quantities for the same rectangular weight matrix from Qwen3-4B-Base, using token representations obtained from a proxy pretraining corpus constructed following Marek et al. (2026). The dashed lines indicate the equal-share variance reference, corresponding here to uniform variance across the R singular directions, as expected under an isotropic random-matrix baseline. Fig. 2(b) shows that most right (input) singular directions associated with both small and large spectral outliers capture above-average input variance, suggesting that pretrained weights align with high-variance input directions on both sides of the MP bulk. Consequently, the pretrained weights preferentially *read* from input directions associated with both large and small spectral outliers. Meanwhile, as shown in Fig. 2(c), the variance along the left singular directions increases strongly with the corresponding singular values. In particular, most of the output variance is concentrated along left singular directions associated with large singular values. Thus, pretrained weights primarily *write* into the principal left-singular subspace.

Given that layer output representations encode information and factual associations acquired during training (Hewitt & Manning, 2019; Meng et al., 2022; Huben et al., 2024), we posit that the dominant components of these representations capture a substantial portion of the pretrained capabilities within a geometric structure shaped by the implicit bias of the pretraining optimizer. We therefore hypothesize that updates induced by a mismatched optimization bias can disproportionately disrupt this structure, leading to degradation of pretrained capabilities. Based on this hypothesis, we modify Muon to steer its momentum away from the dominant output representation subspace, thereby reducing interference with the representation geometry established during pretraining.

3.2 Spectral Momentum Steering of Muon Optimizer

Coordinate Plasticity. For each output singular direction u_i, we define $p_i \in (0, 1]$, a plasticity coefficient that controls movement along that direction. Directions with larger pretrained singular values are assigned smaller plasticity coefficients. First, we define the normalized log-spectral score

$$q_i = \frac{\log(\sigma_i + \epsilon) - \min_j \log(\sigma_j + \epsilon)}{\max_j \log(\sigma_j + \epsilon) - \min_j \log(\sigma_j + \epsilon)}, \qquad \forall i \in \{1, 2, \ldots, R\} \tag{9}$$

where $\epsilon > 0$ is a small constant added for numerical stability, $q_i = 1$ corresponds to the largest singular value, and $q_i = 0$ to the smallest. We then set the plasticity coefficient for direction u_i as

$$p_i = \exp(-\gamma q_i), \qquad \forall i \in \{1, 2, \ldots, R\} \tag{10}$$

where $\gamma > 0$ is a hyperparameter controlling the strength of direction-wise spectral regularization. Hence, large-singular-value directions are assigned smaller plasticity coefficients, while small-singular-value directions remain more plastic. To apply this direction-wise plasticity to a weight update expressed in the original output coordinates, we define the output-space plasticity operator

$$\mathcal{P} = U_R P_\sigma U_R^\top \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}, \quad \text{where} \quad P_\sigma = \text{diag}(p_1, \ldots, p_R) \in \mathbb{R}^{R \times R}. \tag{11}$$

Spectral Momentum Steering. Our goal is to choose an update direction Δ_t that maximally decreases the objective while respecting the output-space plasticity induced by $\mathcal{P}$. To this end, we impose two complementary constraints. First, we restrict the update to the pretrained output subspace $\text{range}(U_R)$, preventing movement into output directions unused by the pretrained weight. Second, within this subspace, we impose an *anisotropic* spectral-norm constraint that penalizes movement more strongly along the principal output subspace encoding the pretraining geometric structure:

$$(I - U_R U_R^\top)\Delta_t = 0, \qquad \|\mathcal{P}^\dagger \Delta_t\|_{\text{op}} \le 1, \tag{12}$$

where $\mathcal{P}^\dagger = U_R P_\sigma^{-1} U_R^\top$ is the Moore–Penrose pseudoinverse of $\mathcal{P}$. The second constraint makes movement along the protected directions with low plasticity more expensive. In particular, since $\mathcal{P} u_i = p_i u_i$, we have $\mathcal{P}^\dagger u_i = u_i / p_i$. Hence, when p_i is small, an update component along u_i is strongly amplified by the constraint norm, restricting its allowable magnitude of physical update. Together, these constraints are equivalently parameterized as

$$\Delta_t = \mathcal{P} Z_t, \qquad \|Z_t\|_{\text{op}} \le 1. \tag{13}$$

Since $\mathcal{P}$ is symmetric, i.e., $\mathcal{P}^\top = \mathcal{P}$, the objective in Eq. 1 can be rewritten as

$$\langle M_t, \Delta_t \rangle_F = \langle M_t, \mathcal{P} Z_t \rangle_F = \mathrm{tr}(M_t^\top \mathcal{P} Z_t) = \mathrm{tr}((\mathcal{P} M_t)^\top Z_t) = \langle \mathcal{P} M_t, Z_t \rangle_F. \tag{14}$$

Therefore, an optimal direction in the normalized coordinates is obtained from

$$Z_t^\star \in \arg \max_{\|Z\|_{\mathrm{op}} \le 1} \langle \mathcal{P} M_t, Z \rangle_F. \tag{15}$$

Following the same spectral-norm steepest-descent argument as in Eq. 1, a solution to Eq. 15 is

$$Z_t^\star = \widetilde{U} \widetilde{V}^\top = \mathrm{polar}(\mathcal{P} M_t), \tag{16}$$

where $\widetilde{U} \widetilde{\Sigma} \widetilde{V}^\top$ is an SVD of $\mathcal{P} M_t$. Consequently, the optimal physical update direction is

$$\Delta^\star = \mathcal{P} Z^\star = \mathcal{P}\, \mathrm{polar}(\mathcal{P} M_t). \tag{17}$$

As in Muon (Jordan et al., 2024), we approximate the polar factor of momentum using the Newton–Schulz iteration (Kovarik, 1970), yielding

$$\mathrm{NS}(\mathcal{P} M_t) \approx \mathrm{polar}(\mathcal{P} M_t). \tag{18}$$

With the scaling factor c (Jordan et al., 2024; Liu et al., 2025), resulting weight update rule becomes

$$W_{t+1} = W_t - \eta_t \cdot c \cdot \mathcal{P}\, \mathrm{NS}(\mathcal{P} M_t). \tag{19}$$

Intuition. Since $\mathcal{P}$ acts only from the left, SMS constrains where the update *writes* while leaving where it *reads* unconstrained, thereby letting the update read freely from the full input representation. This matches the pretrained geometry: output variance concentrates in the principal left-singular subspace, whereas informative input variance spans both ends of the MP spectrum. The two appearances of $\mathcal{P}$ play distinct roles: inner $\mathcal{P}$ suppresses protected directions before the orthogonal update direction selection, while the outer $\mathcal{P}$ limits their actual movement. Thus, high-singular-value directions neither dominate direction selection nor undergo large updates, preserving the representation geometry established during pretraining even under a mismatched fine-tuning optimizer.

4 Experiments

Models and Benchmarking. We evaluate supervised fine-tuning (SFT) on AdamW-pretrained foundation models, namely Qwen3-4B-Base (Bai et al., 2023; Yang et al., 2025) and Llama3.1-8B-Base (Grattafiori et al., 2024), using MetaMathQA (Yu et al., 2023) as the downstream math adaptation dataset. We evaluate downstream adaptation on the test sets of GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) benchmarks. We further evaluate preservation of pretrained capabilities across six categories: (1) Instruction Following, (2) Truthfulness, (3) Knowledge Retrieval, (4) Reading Comprehension, (5) Reasoning, and (6) Commonsense. For each base model (e.g., Qwen3-4B-Base), we report the raw benchmark performance as a fraction, and for each finetuned model, we quantify pretrained capability preservation as the ratio of its category score to that of the corresponding base model, expressed as a percentage. Thus, 100% denotes matching the base-model score, while values above 100% indicate improvement above the base model. We further report Avg1, defined as the arithmetic mean of the preservation percentages across the six categories. For downstream adaptation, we report the raw GSM8K and MATH accuracies, also as percentages, together with their arithmetic mean, denoted by Avg2. To jointly assess pretrained capability preservation and downstream adaptation, we report GM, defined as the geometric mean of Avg1 and Avg2. Further benchmark details are provided in Appendix. A.1.

Baselines and Implementation. We compare SMS-enhanced Muon against standard Muon (Liu et al., 2025) and AdamW (Loshchilov & Hutter, 2017). We further include several forgetting-aware fine-tuning baselines, namely MiLoRA (Wang et al., 2024), CorDA (Yang et al., 2024), LoRA-Null (Tang et al., 2026), and OPLoRA (Xiong & Xie, 2026), each optimized with the standard Muon optimizer. For a fair comparison with these LoRA-based methods, the main comparison uses Low-Rank Adaptation (LoRA) with rank 128 for all methods. Approaches that prescribe their own adapter initialization (MiLoRA, CorDA, LoRA-Null) follow their respective prescribed initialization strategies, while all remaining methods use the standard LoRA initialization, with zero-initialized LoRA up-projection and Kaiming-initialized down-projection (He et al., 2015).

Table 1: Foundation-preservation and adaptation performance. Preservation values for finetuned models are percentages relative to the base model; base-model rows report raw scores as fractions. GM is the geometric mean of Avg1 and Avg2. Bold indicates the best finetuned result. Results are averaged over three seeds.

(a) Qwen3-4B-Base finetuned on math (MetaMathQA)

	Foundation-Preservation							Adaptation			
Methods	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	GSM8K	MATH	Avg2	GM
Qwen3-4B-Base	0.42 ± 0.00	0.53 ± 0.00	0.23 ± 0.00	0.18 ± 0.01	0.51 ± 0.00	0.67 ± 0.00	0.42	-	-	-	-
AdamW (LoRA)	89.04 ± 3.62	98.59 ± 0.04	41.86 ± 3.63	49.24 ± 5.47	94.07 ± 0.76	96.53 ± 0.89	78.22	85.67 ± 0.21	51.17 ± 0.21	68.42	73.16
Muon (LoRA)	73.98 ± 5.26	97.41 ± 0.99	20.12 ± 0.20	32.04 ± 2.00	92.94 ± 0.22	91.26 ± 3.03	67.96	85.97 ± 0.32	49.98 ± 0.03	67.98	67.97
Muon (OPLoRA)	82.79 ± 5.56	97.32 ± 0.27	23.07 ± 1.66	36.24 ± 0.97	94.62 ± 1.29	94.25 ± 0.13	71.38	86.13 ± 0.11	50.04 ± 0.40	68.08	69.71
Muon (LoRA-Null)	52.49 ± 0.30	96.61 ± 1.46	7.76 ± 4.13	13.75 ± 4.75	81.36 ± 3.31	79.56 ± 0.82	55.26	82.83 ± 0.27	43.80 ± 0.23	63.31	59.14
Muon (MiLoRA)	57.87 ± 0.30	97.45 ± 0.88	5.57 ± 0.80	11.04 ± 2.66	83.51 ± 1.01	78.37 ± 3.47	55.63	84.76 ± 0.09	46.29 ± 0.04	65.53	60.38
Muon (CorDA)	48.00 ± 2.79	88.78 ± 1.59	7.02 ± 7.85	1.98 ± 0.18	67.34 ± 0.70	53.78 ± 1.47	44.48	77.29 ± 0.16	34.44 ± 0.82	55.87	49.85
Muon+SMS (LoRA)	**95.45** ± 6.68	**99.81** ± 0.40	**106.36** ± 5.35	**89.95** ± 0.18	**97.84** ± 1.04	**99.79** ± 0.04	**98.20**	**86.73** ± 0.02	**51.74** ± 0.93	**69.24**	**82.46**

(b) Llama3.1-8B-Base finetuned on math (MetaMathQA)

	Foundation-Preservation							Adaptation			
Methods	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	GSM8K	MATH	Avg2	GM
Llama3.1-8B-Base	0.12 ± 0.00	0.44 ± 0.00	0.32 ± 0.00	0.17 ± 0.00	0.42 ± 0.00	0.60 ± 0.00	0.34	-	-	-	-
AdamW (LoRA)	**202.06** ± 9.31	104.46 ± 2.75	16.15 ± 8.34	11.22 ± 11.88	101.77 ± 5.10	97.58 ± 0.77	88.87	78.52 ± 0.86	28.54 ± 0.93	53.53	68.97
Muon (LoRA)	189.28 ± 11.62	**111.27** ± 0.50	13.55 ± 0.04	11.12 ± 0.96	97.93 ± 5.86	96.04 ± 2.08	86.53	76.88 ± 0.97	27.66 ± 0.13	52.27	67.25
Muon (OPLoRA)	179.94 ± 4.71	110.83 ± 0.03	8.77 ± 2.86	7.73 ± 2.88	97.08 ± 2.24	94.05 ± 1.19	83.07	77.79 ± 0.48	27.78 ± 0.10	52.78	66.22
Muon (LoRA-Null)	170.36 ± 19.38	105.04 ± 2.48	0.86 ± 0.15	2.58 ± 0.64	90.80 ± 0.58	83.42 ± 0.78	75.51	73.54 ± 0.91	23.32 ± 0.16	48.43	60.47
Muon (MiLoRA)	158.47 ± 14.03	105.94 ± 3.00	2.61 ± 2.29	2.88 ± 5.41	89.70 ± 2.27	74.56 ± 3.66	72.36	75.51 ± 1.34	25.74 ± 0.10	50.63	60.53
Muon (CorDA)	156.22 ± 6.45	106.67 ± 4.78	8.46 ± 0.24	12.43 ± 6.26	88.24 ± 4.42	81.55 ± 1.02	75.59	72.02 ± 0.75	22.14 ± 0.03	47.08	59.66
Muon+SMS (LoRA)	184.42 ± 15.76	107.41 ± 0.53	**61.42** ± 1.25	**49.42** ± 1.89	**104.02** ± 2.92	**101.13** ± 0.63	**101.30**	**78.85** ± 0.48	**29.62** ± 0.24	**54.23**	**74.12**

Liu et al. (2025) introduce a Muon update scaling factor $c = 0.2\sqrt{\max(d_{\text{in}}, d_{\text{out}})}$ designed to calibrate the RMS magnitude of Muon updates to a scale comparable to that of AdamW, enabling the use of a common learning-rate scale across the two optimizers. We therefore adopt the same scaling factor in our implementation. For a fair comparison, we first tune the AdamW learning rate via a learning-rate sweep and select the value that yields the best downstream adaptation performance (Avg2). We then use the selected learning rate for standard Muon and SMS-enhanced Muon.

4.1 Comparison with Forgetting-aware fine-tuning Methods

We evaluate the downstream adaptation and foundation-preservation performance of SMS-enhanced Muon against the baselines in Tab. 1a and Tab. 1b. On Qwen3-4B-Base, fine-tuning leads to clear degradation of pretrained capabilities across multiple domains. In particular, Instruction Following, Knowledge Retrieval, and Reading Comprehension are disproportionately affected. We further observe that standard Muon fine-tuning (Muon (LoRA)) causes greater degradation of foundational capabilities than AdamW fine-tuning (AdamW (LoRA)) and also achieves slightly lower downstream adaptation performance. This behavior is consistent with the optimizer mismatch phenomenon reported by Liu et al. (2025) and further suggests that LoRA alone does not fully resolve the mismatch. Existing forgetting-aware fine-tuning methods do not effectively mitigate the optimizer mismatch arising in Muon fine-tuning, often achieving preservation performance comparable to, or even worse than, standard Muon. None of the evaluated initialization- or projection-based baselines closes the gap between standard Muon and AdamW. On the other hand, SMS-enhanced Muon effectively preserves pretrained capabilities across all evaluated domains, retaining 98% of the base model's performance on average in Avg1. Muon+SMS also achieves a higher downstream adaptation score (Avg2) than standard AdamW or Muon, showing that its preservation gains do not come at the expense of downstream adaptation. We further investigate this behavior in Sec. 4.2 and Sec. 4.3.

Evaluation on Llama-3.1-8B-Base shows consistent gains from SMS-enhanced Muon, which achieves the highest overall preservation (Avg1) and downstream adaptation (Avg2). One noteworthy observation is that when the base model has a low score on a particular pretrained capability, and fine-tuning improves that capability rather than degrading it, SMS can partially limit the improvement. For example, on Instruction Following, the Llama base model is relatively weak, and both AdamW and standard Muon substantially improve its score, whereas the gain under SMS is somewhat smaller. This behavior is consistent with the additional regularization imposed by SMS.

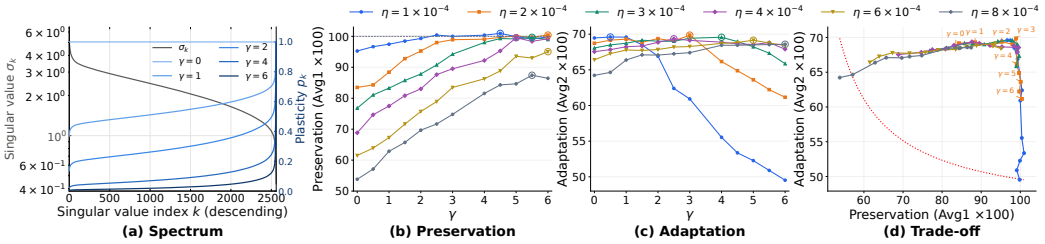

Figure 3: (a) Singular value spectrum and corresponding plasticity coefficients under different γ. (b) Foundation-capability preservation (Avg1) and (c) downstream adaptation (Avg2) as functions of γ at different learning rates η; circles mark the best-performing γ for each learning rate. (d) Preservation–adaptation tradeoff across γ and learning rates; the red dotted curve illustrates the conventional stability–plasticity tradeoff.

4.2 Ablation Studies

γ controls the spectral profile of regularization. In SMS, the regularization strength is governed by the hyperparameter γ, which controls the spectral characteristics of momentum steering. Fig. 3(a) shows the singular value spectrum of `Layer-16 mlp.up_proj` together with the corresponding plasticity coefficients p_i under different choices of γ. As γ increases, the distribution of direction-wise plasticity coefficients becomes more skewed, such that output directions associated with small singular values remain highly plastic, while output directions associated with large singular values, which we hypothesize capture the foundational representation geometry, become increasingly constrained. Through this mechanism, the activation induced by the weight update is steered toward output subspaces that are less utilized by the foundational representation structure, thereby reducing interference with the representation geometry established by pretraining optimizer.

Learning Rates. Fig. 3(b) and (c) illustrate the preservation (Avg1) and adaptation (Avg2) performance of SMS as a function of γ under different learning rates η. The highest achieved performance for each learning rate is highlighted with a colored circle. As shown in Fig. 3(b), higher learning rates lead to greater degradation of foundational capabilities, consistent with larger updates more strongly perturbing the dominant pretrained representation structure. Across all learning rates, increasing γ generally improves preservation by imposing stronger regularization on this dominant pretrained structure. Interestingly, as shown in Fig. 3(c), moderate increases in γ can also improve downstream adaptation performance, with the optimal value depending on the learning rate. For most learning rates, adaptation improves up to an intermediate value of γ, after which stronger regularization degrades performance. In general, we observe that the optimal γ tends to increase with the learning rate, indicating that larger learning rates benefit from stronger spectral regularization. Fig. 3(d) jointly illustrates this effect. The curves trace the preservation (Avg1) and adaptation (Avg2) performance of SMS-enhanced Muon across different values of γ for each learning rate. As the regularization imposed by spectral momentum steering becomes stronger, adaptation performance can improve together with preservation. However, once preservation approaches saturation, further increasing γ can reduce adaptation performance due to excessive regularization.

Rank of LoRA. Fig. 4(a) and (b) show the preservation and adaptation performance of AdamW, Muon, and SMS-enhanced Muon across different LoRA ranks, extending up to full-weight fine-tuning (FFT). For SMS, we sweep γ at each rank and annotate the value yielding the best downstream performance. We observe that preservation performance degrades with increasing LoRA rank for AdamW and standard Muon, consistent with the larger adaptation subspace providing greater capacity to perturb the pretrained representation structure. For downstream adaptation, increasing the rank is beneficial up to a moderate level for AdamW and standard Muon, after which performance declines. In general, standard Muon underperforms its AdamW counterpart across LoRA ranks, consistent with the optimizer mismatch phenomenon. Introducing SMS substantially improves preservation while also improving adaptation performance relative to standard Muon and the AdamW baseline across all LoRA ranks and in the full-weight fine-tuning setting.

Efficiency. Storing the singular basis U_R and applying SMS introduce additional memory and computational overhead. Fig. 4(d) compares the peak GPU-allocated memory, training time, preservation and adaptation performance of several memory-efficient SMS variants against AdamW and

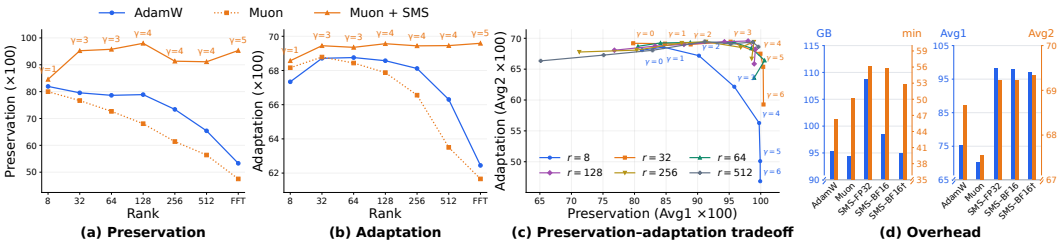

Figure 4: (a) Preservation (Avg1) and (b) adaptation (Avg2) across LoRA ranks and full-weight fine-tuning (FFT); labels denote the selected γ for SMS. (c) Preservation–adaptation tradeoff across LoRA ranks and γ. (d) Peak memory, training time, Avg1, and Avg2 for memory-efficient SMS variants and baselines.

standard Muon. SMS-FP32 and SMS-BF16 (our default) store the singular basis U_R in FP32 and BF16, respectively. SMS-BF16† further skips momentum steering for the MLP up-projection, whose large output dimension induces a large U_R matrix. Memory-efficient variants, such as SMS-BF16 and SMS-BF16†, reduce additional memory and training-time overhead to near-standard Muon levels, while maintaining comparable preservation and adaptation performance to SMS-FP32.

4.3 Preserving the Pretrained Foundation Enables Better Adaptation

The stability–plasticity trade-off is a long-standing challenge in neural network adaptation, referring to the tension between preserving previously acquired knowledge and effectively learning new capabilities (Turrigiano & Nelson, 2004; Carpenter & Grossberg, 1988). The conventional stability–plasticity view suggests that greater plasticity during adaptation often comes at the cost of eroding previously acquired knowledge, thereby reducing stability. Conversely, stronger efforts to preserve existing knowledge may restrict the model's ability to acquire new knowledge, leading to lower plasticity. This trade-off is illustrated by the red curve in Fig. 3(d).

Our findings suggest that this trade-off does not necessarily hold when fine-tuning highly overparameterized pretrained foundation models. In particular, appropriately enforced regularization can improve stability without sacrificing downstream adaptation and, in many cases, can even enhance it. Fig. 4(c) jointly illustrates the preservation–adaptation trade-off of SMS-enhanced Muon across different LoRA ranks under varying values of γ. As the regularization imposed by SMS becomes stronger with increasing γ, adaptation performance can improve together with preservation up to a moderate level of regularization, exhibiting a trend consistent with that observed in Fig. 3(d). Here, the learning rate (update magnitude) and LoRA rank (dimensionality of update) govern the overall strength of the adaptation signal. In contrast, γ controls the spectral profile of the regularization applied to this signal. Across both Fig. 3(d) and Fig. 4(c), we observe that spectral regularization can improve downstream adaptation together with preservation of pretrained capabilities.

One possible explanation is that downstream adaptation relies on capabilities inherited from pretraining. If these foundational capabilities are substantially degraded, the newly acquired task knowledge may become less usable in downstream inference. For example, even when fine-tuning successfully acquires mathematical knowledge, degradation in instruction-following or reading-comprehension ability may prevent the model from properly interpreting mathematical context or user instructions, ultimately limiting downstream performance. SMS substantially reduces fine-tuning-induced degradation of foundational capabilities by steering updates away from the representation geometry shaped by the implicit bias of the pretraining optimizer. By avoiding interference with this optimizer-specific structure, SMS preserves foundational capabilities while enabling stronger downstream adaptation, thereby substantially alleviating optimizer mismatch.

5 Conclusion

We introduce Spectral Momentum Steering (SMS), an extension of Muon that mitigates optimizer mismatch by preserving the pretrained representation geometry during finetuning. Across AdamW-pretrained model families, SMS consistently mitigates optimizer mismatch. Ablations across learning rates, adapter ranks, and full-weight finetuning further validates its effectiveness, suggesting that maintaining pretrained structure can benefit effective downstream adaptation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing, code development and debugging, preparation of plotting and analysis scripts, literature exploration, and checking the clarity of mathematical derivations and technical explanations. Generative AI was also used during research discussions to help explore alternative formulations and experimental designs. We did not use generative AI to generate experimental results or to make final scientific conclusions on behalf of the authors. All AI-assisted code was reviewed and tested by the authors, mathematical arguments were independently checked, and references and technical claims were verified against the corresponding source materials where applicable. All reported experimental results were obtained from our implemented training and evaluation pipelines rather than generated by AI tools. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies optimization methods for fine-tuning existing language models using publicly available model checkpoints and research benchmarks. The study consists of computational experiments and does not involve recruiting human participants or collecting new personal data. We cite the original sources of the models and datasets used in our experiments.

SMS aims to preserve pretrained capabilities during downstream adaptation. However, preserving these capabilities can also preserve biases or undesirable behaviors inherited from the base model. The preservation and adaptation metrics evaluated in this work do not establish guarantees of fairness, privacy, or safe behavior. Applications of the method should therefore assess these properties separately in their intended deployment settings.

REPRODUCIBILITY STATEMENT

Section 3 describes the spectral analysis, plasticity coefficients, and update rule underlying Spectral Momentum Steering. Appendix A.2 specifies the training data preparation, adapter parameterization, hyperparameters, and learning-rate selection procedure. Appendix A.1 describes the evaluation benchmarks, evaluation protocols, and aggregation of preservation and adaptation metrics. Additional ablations and optimizer comparisons are provided in Appendix B, including the Moonlight experiment in Appendix B.4. The derivation of Muon's spectral-norm formulation is given in Appendix C.1. We also include anatomized SMS implementation to support inspection and reproduction of the proposed method.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1533–1544. ACL, 2013. doi: 10.18653/V1/D13-1160. URL https://doi.org/10.18653/v1/d13-1160.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:7763588.

Gail A. Carpenter and Stephen Grossberg. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21:77–88, 1988. URL https://api.semanticscholar.org/CorpusID:14625094.

Lizhang Chen, Jonathan Li, and qiang liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Blz4hjxLwU.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mo Bavarian, Clemens Winter, Phil Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021. URL https://api.semanticscholar.org/CorpusID:235755472.

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. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Anyi Xu, Bang Lin, Bing Xue, Bing-Li Wang, Bin Xu, Bo Wu, Bowei Zhang, Chao Lin, Chengyao Dong, Chen Ling, Chengda Lu, Chen Zhao, Chengqi Deng, Cheng Hou, Chen Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, De-Bin Yang, Deli Chen, Dong-Hui Li, Dong-Li Ji, Erhang Li, Fangchen Wei, Fangyun Lin, Fang Yuan, Fei Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guo Cao, Guo-Hui Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Hao Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Hao Yuan, Haowei Zhang, Haowen Luo, Hao Chen, Haozhe Ji, Heng Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J. Yang, JQ Zhu, Jianming Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, JingChang Chen, Jing Zhou, Jing Xiang, Jingyang Yuan, Jing Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Junliang Ran, Jun Jiang, Junjie Qiu, Junlong Li, Junming Zheng, Jun-Mei Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lin Luo, Lin Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, Mingxi Di, My Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mi Zhou, Min Han, Ning Wang, Pan Huang, Pan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qi Jiang, Rui Tian, Rui qin Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Run Da Chen, Runqiu Yin, Runxin Xu, Ruo-Han Shen, Ruoyu Zhang, Ruyi Chen, Sh. Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shan-Shan Chen, Shaofei Cai, Shao Nie, Shao-Ping Wu, Shaoyuan Chen, Shengding Hu, Sheng Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tianhong Pei, Tian Ye, Tianle Lin, Tian Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tu-Tu Wang, W. Y. Zhang, Wl Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjin Yao, Wenjun Gao, Wenkai Yang, Wen-Wen Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiang Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xu Fu, Yc Yan,

YQ Wang, Yw Ma, Yanfen Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yingxia Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yi-Yu Xiong, Yi Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yong Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yukun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yue Xu, Yuhan Wu, Yu Meng, Yu Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yu Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, Zhanghua Wu, Ze Wang, Ze Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhe Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixiang Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zi-Long Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zong xiang Yao. Deepseek-v4: Towards highly efficient million-token context intelligence. 2026. URL https://api.semanticscholar.org/CorpusID:289623513.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In *Neural Information Processing Systems*, 2025. URL https://api.semanticscholar.org/CorpusID:290969503.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qCIg2WGudx.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551, 2017. URL https://api.semanticscholar.org/CorpusID:26501419.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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 V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. URL https://api.semanticscholar.org/CorpusID:86611921.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Jingyuan Liu, Jianling Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Meng Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *ArXiv*, abs/2502.16982, 2025. URL https://api.semanticscholar.org/CorpusID:276575212.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Martin Marek, Dongkyu Cho, Shikai Qiu, Rumi Chunara, Pavel Izmailov, and Andrew Gordon Wilson. Forgetting in language models: Capacity, optimization, and self-generated replay, 2026. URL https://arxiv.org/abs/2605.26097.

Charles H. Martin, Tongsu Peng, and Michael W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12, 2020. URL https://api.semanticscholar.org/CorpusID:211132500.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems 35*, 2022. URL https://api.semanticscholar.org/CorpusID:255825985.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Xingyu Qu, Huang Peigeng, and Samuel Horváth. Can muon fine-tune adam-pretrained models? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=NKKwTEYdAm.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

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. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. 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. 4463–4473, 2019.

Essential AI Darsh J Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of muon for pretraining. *ArXiv*, abs/2505.02222, 2025. URL https://api.semanticscholar.org/CorpusID:278327183.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=7HNRYT4V44.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14670–14728, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Pengwei Tang, Xiaolin Hu, Yong Liu, Lizhong Ding, Dongjie Zhang, Xing Wu, and Debing Zhang. Put the space of lora initialization to the extreme to preserve pre-trained knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33232–33240, 2026.

Team-Kimi, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026a.

Team-Kimi, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026b.

Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix analysis of deep neural network weight matrices. *Physical review. E*, 106 5-1:054124, 2022. URL https://api.semanticscholar.org/CorpusID:247763175.

Gina G. Turrigiano and Sacha B. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5:97–107, 2004. URL https://api.semanticscholar.org/CorpusID:14535839.

Hanqing Wang, Zeguan Xiao, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. Milora: Harnessing minor singular components for parameter-efficient llm finetuning. In *North American Chapter of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:270440848.

Yifeng Xiong and Xiaohui Xie. Oplora: Orthogonal projection lora prevents catastrophic forgetting during parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34088–34096, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. Corda: Context-oriented decomposition adaptation of large language models. *ArXiv*, abs/2406.05223, 2024. URL https://api.semanticscholar.org/CorpusID:270370784.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12834–12859, 2024.

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

A Experiment Details

A.1 Benchmark

Tab. 2 summarizes the evaluation suite. We organize the benchmarks into two families. *Foundation-preservation* benchmarks measure how much of the capabilities inherited from pretraining survives fine-tuning, and are grouped into six categories so that no single capability dominates the aggregate. *Downstream-adaptation* benchmarks measure how well the model acquires the target mathematical skill.

Foundation-preservation benchmarks. **Instruction Following** is measured with IFEval (Zhou et al., 2023), which scores verifiable constraints (length, format, keyword) attached to each prompt. **Truthfulness** is measured with TruthfulQA-MC2 (Lin et al., 2022), a multiple-choice benchmark of questions that invite imitative falsehoods. **Knowledge Retrieval** is measured with three open-domain question-answering datasets, TriviaQA (Joshi et al., 2017), NQ-Open (Kwiatkowski et al., 2019), and WebQuestions (Berant et al., 2013), which probe factual knowledge stored in the parameters without any retrieval context. This category is of particular interest, since factual recall has been reported to be the capability most susceptible to degradation during post-training (Harmon et al., 2026). **Reading Comprehension** is measured with SQuAD-v2 (Rajpurkar et al., 2016), which additionally requires abstaining on unanswerable questions and therefore penalizes models that have drifted toward unconditional answer generation. **Reasoning** is measured with GPQA-Diamond (Rein et al., 2023) for graduate-level scientific reasoning, MMLU (Hendrycks et al., 2021a) for broad multi-domain knowledge-intensive reasoning, and MuSR (Sprague et al., 2024) for multi-step soft reasoning over long narratives. **Commonsense** is measured with CommonsenseQA (Talmor et al., 2019) and Social-IQA (Sap et al., 2019), covering physical-conceptual and social commonsense, respectively. Within a category, the category score is the arithmetic mean of its constituent benchmark scores.

Downstream-adaptation benchmarks. We fine-tune on MetaMathQA (Yu et al., 2023) and evaluate mathematical problem solving on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b). We use held-out subsets of the GSM8K and MATH test splits for hyperparameter selection and report final performance on the remaining examples. Both benchmarks are scored by exact match of the extracted final answer.

Evaluation protocol. All foundation-preservation benchmarks are evaluated with the `lm-evaluation-harness` using a vLLM backend in `bfloat16` with a maximum context length of 4,096 tokens. We do not override the per-task few-shot configuration, so every preservation benchmark is evaluated zero-shot, which isolates the model's intrinsic capability from its ability to exploit in-context demonstrations. The two adaptation benchmarks are evaluated with greedy decoding (temperature $= 0$, top-$p = 1$) under the same instruction template used during fine-tuning, with a generation budget of 1,024 tokens for GSM8K and 2,048 tokens for MATH. All methods, including the unmodified base models, are evaluated with an identical protocol, and adapters are merged into the base weights before evaluation so that no method receives an inference-time advantage.

Aggregation. Let s_c^{base} and s_c^{ft} denote the score of the pretrained base model and of a fine-tuned model on foundation capability category c. We report the *preservation ratio*

$$\text{Pres}_c = \frac{s_c^{\text{ft}}}{s_c^{\text{base}}}, \tag{20}$$

which normalizes away the differing difficulty and scale of the six categories and allows direct comparison between the two model families. Avg1 is the arithmetic mean of Pres_c over the six categories, Avg2 is the arithmetic mean of the raw GSM8K and MATH accuracies, and

$$\text{GM} = \sqrt{\text{Avg1} \cdot \text{Avg2}} \tag{21}$$

is their geometric mean. We use a geometric rather than arithmetic mean for the overall score because it penalizes methods that succeed on only one of the two axes, which is precisely the failure mode of interest: a method that adapts well while substantially degrading pretrained capabilities, or that preserves the pretrained model by failing to adapt, should not score well.

Table 2: Evaluation suite. All foundation-preservation benchmarks are evaluated zero-shot with the `lm-evaluation-harness`; the two adaptation benchmarks are evaluated with greedy decoding under the fine-tuning instruction template. EM denotes exact match.

Category	Benchmark	#Test	Metric
Foundation preservation			
Instruction Following	IFEval (Zhou et al., 2023)	541	Strict/loose acc. (avg.)
Truthfulness	TruthfulQA-MC2 (Lin et al., 2022)	817	Accuracy
Knowledge Retrieval	TriviaQA (Joshi et al., 2017)	17,944	EM
	NQ-Open (Kwiatkowski et al., 2019)	3,610	EM
	WebQuestions (Berant et al., 2013)	2,032	EM
Reading Comprehension	SQuAD-v2 (Rajpurkar et al., 2016)	11,873	EM
Reasoning	GPQA-Diamond (Rein et al., 2023)	198	Accuracy
	MMLU (Hendrycks et al., 2021a)	14,042	Accuracy
	MuSR (Sprague et al., 2024)	756	Normalized acc.
Commonsense	CommonsenseQA (Talmor et al., 2019)	1,221	Accuracy
	Social-IQA (Sap et al., 2019)	1,954	Accuracy
Downstream adaptation			
Math	GSM8K (Cobbe et al., 2021)	1,319	EM (final answer)
	MATH (Hendrycks et al., 2021b)	5,000	EM (final answer)

A.2 Training Details and Hyperparameters

Adapter parameterization. For each adapted linear layer with frozen pretrained weights $W_0 = U_R \Sigma_R V_R^\top$, we use LoRA parametrization of

$$W_{\text{eff}} = W_0 + \frac{\alpha}{r} BA, \quad B \in \mathbb{R}^{d_{\text{out}} \times r}, A \in \mathbb{R}^{r \times d_{\text{in}}} \tag{22}$$

where we consistently use $\alpha = r$. To adapt the SMS mechanism in LoRA adapter, we optimize the up-projection (output projection) matrix B using SMS-enhanced Muon, while optimizing the down-projection (input projection) matrix A using standard Muon. With the output-space plasticity operator $\mathcal{P} = U_R P_\sigma U_R^\top$ where $P_\sigma = \text{diag}(p_1, \ldots, p_R)$, The weight update LoRA adapter is therefore

$$B_{t+1} = B_t - \eta\, c\, \mathcal{P}\, \text{NS}(\mathcal{P} M_t^B) \tag{23}$$

$$A_{t+1} = A_t - \eta\, c\, \text{NS}(M_t^A), \tag{24}$$

where M_t^B and M_t^A are the momentum of B and A matrix. Following standard LoRA initialization, B is zero-initialized and A is initialized with Kaiming initialization. The singular vectors U_R and the plasticity coefficients p_i are obtained from a single exact thin SVD of W_0 computed once before training in FP32 and cached in specific precision (e.g., FP32, BF16); they are frozen thereafter and incur no per-step cost beyond two diagonal scalings.

Setup. We fine-tune on the first 100,000 examples of MetaMathQA for one epoch, using the Alpaca-style instruction template with the `query` and `response` fields and a maximum sequence length of 512 tokens. All methods share the same data order, schedule, and adapter budget; only the optimizer and, for SMS, the plasticity schedule differ. All experiments are performed using a single NVIDIA GB200 node. Tab. 3 lists the full configuration.

Learning-rate selection. For the Qwen3-4B-Base and Llama-3.1-8B-Base experiments in Sec. 4, the learning rate is selected once per model family using AdamW and then shared by every method. Concretely, we sweep the AdamW learning rate over $\{1, 5\} \times 10^{-5}$ and $\{1, 2, 3, 4, 5\} \times 10^{-4}$, select the value that maximizes downstream adaptation performance, and reuse that value for standard Muon, for the forgetting-aware LoRA baselines, and for SMS-enhanced Muon. This yields $\eta = 3 \times 10^{-4}$ for Qwen3-4B-Base and $\eta = 4 \times 10^{-4}$ for Llama-3.1-8B-Base.

Sharing a single η across optimizers is part of the comparison protocol rather than a convenience. Liu et al. (2025) observe that the RMS magnitude of a raw orthogonalized Muon update depends on the

Table 3: Training hyperparameters. The upper block is shared by every method; the lower blocks give the optimizer-specific settings. $R = \min(d_{\text{out}}, d_{\text{in}})$ denotes the full thin-SVD rank of the pretrained weight, which is distinct from the LoRA rank r.

Hyperparameter	**Value**
Shared	
fine-tuning dataset	MetaMathQA (first 100,000 examples)
Epochs	1
Max sequence length	512
Per-device batch size	16
Gradient accumulation steps	8
Effective batch size	128
LR schedule	Cosine decay
Warmup ratio	0.03
Weight decay	0
Precision	`bfloat16` (TF32 matmuls; FP32 adapters)
Gradient checkpointing	Enabled
Learning rate η	3×10^{-4} (Qwen3-4B-Base) 4×10^{-4} (Llama-3.1-8B-Base)
LoRA rank r / scale α	128 / 128
LoRA dropout	0
Target modules	`q,k,v,o,gate,up,down_proj`
AdamW	
Learning rate	η
(β_1, β_2)	$(0.9, 0.999)$
ϵ	10^{-8}
Muon	
Learning rate	η
Momentum μ	0.95
Newton–Schulz steps	5
Nesterov momentum	Disabled
Update scale c	$0.2\sqrt{\max(d_{\text{in}}, d_{\text{out}})}$
SMS-enhanced Muon	
Learning rate (A and B)	η
Optimizer for A / B	Muon / SMS-Muon
Plasticity sharpness γ	4.0 (Qwen3-4B-Base) 1.5 (Llama-3.1-8B-Base)
Spectral basis rank	Full thin SVD (R)
Momentum, NS steps, Nesterov	Same as Muon (Liu et al., 2025)

shape of the parameter matrix, and introduce the update scale $c = 0.2\sqrt{\max(d_{\text{in}}, d_{\text{out}})}$ precisely so that the RMS of a Muon update is calibrated to that of an AdamW update on the same parameter. With this calibration a single learning rate denotes a comparable effective step size under both optimizers, so transferring an AdamW-tuned η to Muon is the intended usage rather than an approximation, and we adopt exactly this scaling factor (Tab. 3) in every Muon-based run, including SMS-enhanced Muon. This is what makes the shared learning rate the fair choice: because update magnitudes are matched, the differences we report are attributable to the geometry of the update, rather than to one optimizer effectively taking larger steps than another. Tuning η separately per optimizer would reintroduce exactly this confound, since a method could then appear to preserve foundation capability merely by having been assigned a smaller effective step size, which Tab. 4 identifies as the single most influential factor for preservation.

Tab. 4 reports the complete sweep, and two trends are visible on both model families. First, foundation preservation degrades monotonically with the learning rate, and the categories identified in Sec. 4 as most fragile, Knowledge Retrieval and Reading Comprehension, degrade fastest: on Qwen3-4B-Base they fall from 1.069 and 0.971 at $\eta = 1 \times 10^{-5}$ to 0.444 and 0.531 at the selected $\eta = 3 \times 10^{-4}$, while Truthfulness, Reasoning, and Commonsense remain above 0.93 throughout. Second, adaptation

Table 4: AdamW learning-rate sweep used to select the shared learning rate η for each model family. Foundation-preservation columns are ratios relative to the corresponding base model; adaptation columns are raw accuracies. The selected row, maximizing Avg2, is shown in bold.

(a) Qwen3-4B-Base. Selected $\eta = 3 \times 10^{-4}$.

	Foundation-Preservation							Adaptation		
η	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	GSM8K	MATH	Avg2
1×10^{-5}	0.940	0.988	1.069	0.971	0.981	1.003	0.992	0.466	0.464	0.465
5×10^{-5}	0.874	0.986	0.726	0.706	0.955	0.997	0.874	0.800	0.521	0.660
1×10^{-4}	0.878	0.985	0.549	0.628	0.935	0.990	0.827	0.843	0.512	0.678
2×10^{-4}	0.850	0.982	0.450	0.509	0.964	0.980	0.789	0.851	0.519	0.685
$\mathbf{3 \times 10^{-4}}$	**0.865**	**0.986**	**0.444**	**0.531**	**0.935**	**0.972**	**0.789**	**0.858**	**0.513**	**0.686**
4×10^{-4}	0.813	0.971	0.331	0.458	0.903	0.958	0.739	0.856	0.504	0.680
5×10^{-4}	0.778	0.969	0.288	0.401	0.924	0.951	0.718	0.854	0.508	0.681
1×10^{-3}	0.507	0.903	0.002	0.234	0.570	0.462	0.446	0.000	0.000	0.000

(b) Llama3.1-8B-Base. Selected $\eta = 4 \times 10^{-4}$.

	Foundation-Preservation							Adaptation		
η	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	GSM8K	MATH	Avg2
1×10^{-5}	1.391	1.070	0.922	1.267	1.021	1.022	1.115	0.710	0.254	0.482
5×10^{-5}	2.092	1.072	0.506	0.805	1.059	1.006	1.090	0.763	0.286	0.525
1×10^{-4}	1.843	1.084	0.369	0.455	1.062	1.006	0.970	0.769	0.297	0.533
2×10^{-4}	2.175	1.068	0.297	0.221	1.032	1.006	0.967	0.772	0.293	0.532
3×10^{-4}	1.972	1.049	0.240	0.173	0.985	0.964	0.897	0.782	0.290	0.536
$\mathbf{4 \times 10^{-4}}$	**2.001**	**1.043**	**0.173**	**0.102**	**1.020**	**0.989**	**0.888**	**0.786**	**0.291**	**0.539**
5×10^{-4}	1.686	1.016	0.103	0.064	0.831	0.741	0.740	0.767	0.275	0.521

saturates well before preservation stabilizes. On Qwen3-4B-Base, Avg2 is already within one point of its maximum at $\eta = 2 \times 10^{-4}$ and varies by less than 0.01 across the entire range $[1 \times 10^{-4}, 5 \times 10^{-4}]$, whereas Avg1 falls from 0.827 to 0.718 over the same range; Llama-3.1-8B-Base behaves identically, with Avg2 spanning 0.532–0.537 while Avg1 falls from 0.970 to 0.740. The learning rate that is optimal for the target task therefore sacrifices a large fraction of the pretrained capability for an adaptation gain that is close to negligible. This gap between the adaptation-optimal and preservation-optimal operating points is exactly what SMS closes, and it is why we hold η fixed at the selected value and vary only γ in Sec. 4.2. We extend the Qwen3-4B-Base sweep to $\eta = 1 \times 10^{-3}$ to confirm that the trend terminates in collapse: at that learning rate the model produces no correct answer on either math benchmark and retains less than half of its foundation capability.

On Llama-3.1-8B-Base the Instruction Following ratio exceeds one at every learning rate, peaking above 2.0. This reflects the weak instruction-following ability of the Llama base checkpoint, for which supervised fine-tuning on MetaMathQA is a net improvement rather than a degradation, as discussed in Sec. 4. Preservation ratios above one should therefore be read as fine-tuning having supplied a capability the base model largely lacked, not as forgetting having been avoided.

Baselines. All LoRA-based forgetting-aware fine-tuning baselines use the same rank, scale, target modules, schedule, and learning rate as above, and are optimized with standard Muon. MiLoRA, CorDA, OPLoRA, and LoRA-Null follow their prescribed implementation and recommended best-operating hyperparameter setup as specified in their original documents. We specify the method-specific hyperparameters for each baseline as follows:

- **MiLoRA**: No additional method-specific hyperparameters; we use the method-specific LoRA initialization.
- **CorDA / LoRA-Null**: We use 256 NQ-Open (Kwiatkowski et al., 2019) samples to construct the calibration set and apply the corresponding method-specific LoRA initialization.
- **OPLoRA**: We set $k = 128$ for the projection matrices P_L and P_R and use standard LoRA initialization.

Table 5: Decomposition of SMS into subspace restriction and spectral steering on Qwen3-4B-Base. $\gamma = 0$ retains only the restriction of the update to the pretrained column space $\mathrm{range}(U_R)$; $\gamma = \gamma^\star$ additionally applies spectral steering, with $\gamma^\star$ tuned per learning rate. Δ_{res} and Δ_{steer} report the preservation gain attributable to each mechanism. All values are percentages.

	Avg1 (foundation preservation)					**Avg2 (downstream adaptation)**			
η	Muon	SMS (γ=0)	SMS ($\gamma^\star$)	Δ_{res}	Δ_{steer}	Muon	SMS (γ=0)	SMS ($\gamma^\star$)	$\gamma^\star$
1×10^{-4}	86.20	95.26	**96.65**	+9.06	+1.39	67.61	69.47	**69.60**	0.5
2×10^{-4}	77.10	83.50	**98.92**	+6.40	+15.42	67.87	68.73	**69.87**	3.0
3×10^{-4}	68.05	76.81	**98.00**	+8.76	+21.18	67.87	68.07	**69.57**	4.0
4×10^{-4}	63.21	68.83	**89.52**	+5.62	+20.69	67.07	67.55	**69.13**	3.0
6×10^{-4}	56.72	61.50	**88.82**	+4.78	+27.33	64.38	66.41	**68.84**	4.5
8×10^{-4}	49.90	53.82	**86.45**	+3.92	+32.63	60.78	64.23	**68.59**	6.0
Mean	66.86	73.29	**93.06**	+6.42	+19.77	65.93	67.41	**69.27**	–

B Additional Experiments

B.1 Effect of Spectral Momentum Steering

SMS modifies the Muon update in two distinct ways, and it is natural to ask how much each of them contributes.

(i) Null-space restriction. The plasticity operator $\mathcal{P} = U_R P_\sigma U_R^\top$ is built from the left singular basis of the pretrained weight, so every feasible update lies in $\mathrm{range}(U_R)$. For a layer with $d_{\mathrm{out}} > d_{\mathrm{in}}$, most prominently the MLP gate and up projections, $\mathrm{range}(U_R)$ is a proper subspace of $\mathbb{R}^{d_{\mathrm{out}}}$, and its orthogonal complement is the left null space of W_0, i.e. the set of output directions into which the pretrained layer never writes. SMS cannot move along these directions at all, whereas standard Muon explores the entire output space.

(ii) Spectral steering. For $\gamma > 0$ the plasticity coefficients decay with the pretrained singular values, so movement is additionally graded *within* $\mathrm{range}(U_R)$, in proportion to how much of the layer's output energy each direction carries.

Setting $\gamma = 0$ separates the two. From Eq. 10 we then have $p_i = 1$ for every i, so $P_\sigma = I$ and $\mathcal{P} = U_R U_R^\top$ reduces to an orthogonal projector onto the pretrained column space: mechanism (i) is active while mechanism (ii) is switched off, and the update rule of Eq. 19 becomes a projected Muon step with no direction-dependent weighting. Tab. 5 compares standard Muon, this $\gamma = 0$ control, and SMS at its best γ, sweeping the learning rate over $\{1, 2, 3, 4, 6, 8\} \times 10^{-4}$ on Qwen3-4B-Base.

Restricting the update away from the left null space already helps. At every learning rate the $\gamma = 0$ control improves preservation over standard Muon, by between 3.92 and 9.06 points, with the largest gain at the smallest learning rate. This is worth emphasizing because $\gamma = 0$ applies no direction-dependent regularization whatsoever: the sole difference from standard Muon is that the update is confined to the subspace the pretrained layer actually writes into. Movement along the remaining output directions therefore appears to be a pure liability. It perturbs the layer's output geometry, and hence the representation structure that downstream layers consume, without contributing to the target task, which is consistent with the read–write asymmetry identified in Sec. 3. Notably the $\gamma = 0$ control also improves adaptation at every learning rate, so this restriction is not a stability-for-plasticity trade: the null-space component of a Muon update is simply not useful.

Staying inside the pretrained output subspace modifies features the network already knows how to process. Moving into the null-space creates entirely new activation directions that were never produced during pretraining, and these new directions can affect computation of downstream network on many general inputs, perturbing pretrained capabilities.

Spectral steering contributes the larger share. From $\eta = 2 \times 10^{-4}$ upward, turning on spectral steering adds a further 15.42 to 32.63 preservation points over the $\gamma = 0$ control, between $2.4\times$ and $8.3\times$ the gain from restriction alone, and across the full sweep it contributes +19.77 points on

average against $+6.42$ for restriction. Restriction decides *which* output directions may be written at all, but among the admissible directions it still treats a dominant singular direction and a negligible one identically; steering is what grades them. Only with both mechanisms active does preservation approach the base model, reaching 98.92 and 98.00 at $\eta = 2 \times 10^{-4}$ and 3×10^{-4}. The one exception is the smallest learning rate, $\eta = 1 \times 10^{-4}$, where restriction alone already reaches 95.26 preservation and 69.47 adaptation, and steering adds only 1.39 and 0.13 points respectively; we return to this regime below.

The balance shifts from restriction to steering as the step size grows. Standard Muon degrades sharply with the learning rate, losing 36.30 preservation points between $\eta = 1 \times 10^{-4}$ and $\eta = 8 \times 10^{-4}$, whereas SMS stays within 86.45–98.92 across the entire sweep. The effect is large enough that SMS at $\eta = 8 \times 10^{-4}$ matches the preservation of standard Muon at $\eta = 1 \times 10^{-4}$ (86.45 versus 86.20) while adapting better (68.59 versus 67.61), at an eight times larger step size. The two mechanisms scale in opposite directions, however. The restriction-only gain broadly *shrinks* with the learning rate ($9.06 \to 3.92$), whereas the steering gain *grows* almost monotonically ($1.39 \to 32.63$), so the ratio between them rises from $0.15\times$ at $\eta = 1 \times 10^{-4}$ to $8.3\times$ at $\eta = 8 \times 10^{-4}$. Projecting away the null space bounds the damage a single update can do, but that bound is fixed, whereas steering rescales with the step size and is therefore what carries the method in the aggressive-update regime.

Correspondingly, the best γ increases with the learning rate (0.5, 3.0, 4.0, 3.0, 4.5, 6.0), matching the trend reported in Sec. 4.2, and small learning rates call for markedly weaker steering: at $\eta = 1 \times 10^{-4}$ the optimum is $\gamma^\star = 0.5$, an almost flat plasticity profile. When the step is small, removing its null-space component already recovers most of the preservation, leaving little for steering to protect, while a sharp plasticity profile would only throttle adaptation along the directions the task requires, the excessive-regularization regime identified in Sec. 4.2. As the step grows, the update increasingly perturbs the dominant pretrained directions, and correspondingly stronger steering is needed to redirect it.

Adaptation is never traded away. SMS attains the best Avg2 at every learning rate. Its margin over standard Muon stays at roughly two points up to $\eta = 4 \times 10^{-4}$ (1.70 to 2.06) and then widens sharply, to 4.46 and 7.81 points at $\eta = 6 \times 10^{-4}$ and 8×10^{-4}. More striking is the stability: SMS's adaptation score varies by only 1.28 points across the entire sweep (69.87 down to 68.59), whereas Muon's falls by 7.09 points (67.87 down to 60.78). Because SMS regularizes the update in proportion to how much pretrained output energy each direction carries, an overly large global step size is absorbed by the plasticity profile rather than being allowed to overwrite the foundation structure. In practice this makes downstream performance substantially less sensitive to learning-rate choice, which reinforces the observation in Sec. 4.3 that preserving the pretrained foundation supports, rather than competes with, downstream adaptation.

The AdamW-selected learning rate transfers to Muon. The learning rate used in the main experiments, $\eta = 3 \times 10^{-4}$, was selected by the AdamW sweep of Appendix A.2 and reused for Muon without retuning. The present sweep shows that it is also where standard Muon adapts well: Muon's Avg2 is nearly flat between $\eta = 1 \times 10^{-4}$ and 3×10^{-4} (67.61 to 67.87) and peaks at 3×10^{-4} before degrading. This agreement is consistent with the role of the update scaling factor $c = 0.2\sqrt{\max(d_{\text{in}}, d_{\text{out}})}$, which matches the RMS magnitude of Muon updates to that of AdamW so that the two optimizers can share a learning rate.

B.2 The Benefit of SMS Is Not Explained by a Smaller Update

Because the plasticity operator is applied twice in Eq. 19, SMS takes a physically smaller step than standard Muon at the same learning rate. A natural concern is therefore that the gains reported in Sec. 4 merely reflect a reduced update magnitude rather than the geometry of the update. This section controls for that possibility.

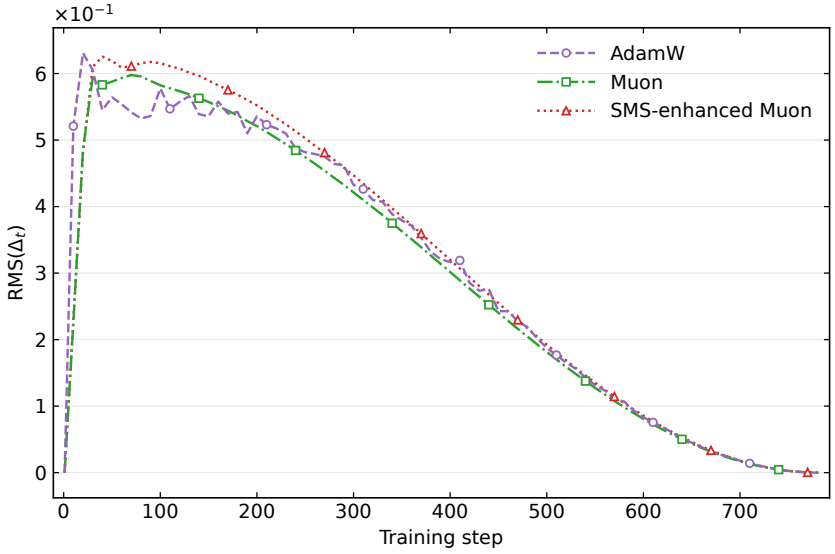

Figure 5: Update magnitude over training on Qwen3-4B-Base. Each curve plots the Frobenius norm of the weight update $\|\Delta W_t\|_F$ of Eq. 25 at every optimizer step, for AdamW, standard Muon, and SMS-enhanced Muon. The learning rate of each method is tuned individually so that the three curves coincide, equalizing how far each optimizer moves the weights per step. The step-averaged values are reported in the first column of Tab. 6.

Table 6: Comparison at matched update magnitude on Qwen3-4B-Base. $\|\Delta W\|_F$ is the step-averaged Frobenius norm of the weight update (Eq. 25); the learning rate of each method is tuned so that all three operate at the same update magnitude, isolating the effect of the update *direction* from that of its *scale*. Foundation-preservation columns are percentages relative to the base-model score; adaptation columns are raw accuracies. Bold indicates the best result.

		Foundation-Preservation							**Adaptation**		
Method	$\|\Delta W\|_F$	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	GSM8K	MATH	Avg2
AdamW (2×10^{-4})	0.2978	91.36	99.07	54.01	49.93	95.40	97.24	81.17	86.20	**51.80**	69.00
Muon (2×10^{-4})	0.2906	91.60	98.73	38.30	48.74	94.85	95.17	77.90	86.05	50.64	68.35
Muon+SMS (3×10^{-4})	0.3064	**97.94**	**99.89**	**106.36**	**81.01**	**97.49**	**98.92**	**96.93**	**87.11**	51.28	**69.20**

For a parameter set $\{W_\ell\}$ we measure the magnitude of a single optimizer step by the Frobenius norm of the weight update,

$$\|\Delta W_t\|_F = \left(\sum_\ell \left\| W_\ell^t - W_\ell^{t-1} \right\|_F^2 \right)^{1/2}, \tag{25}$$

and summarize a run by the mean of $\|\Delta W_t\|_F$ over all training steps. We then tune the learning rate of each method individually so that the three runs trace essentially the same update-magnitude curve throughout training, as shown in Fig. 5. This equalizes how far each optimizer moves the weights per step and leaves only the *direction* of the update as the difference between methods.

Tab. 6 reports foundation preservation and downstream adaptation under this matched-magnitude protocol, together with the mean update norm actually attained; the three runs agree to within 5.3% of their common mean, so the comparison is genuinely at equal update energy. Two conclusions follow. First, standard Muon preserves less than AdamW (77.90 versus 81.17 Avg1) and also adapts slightly worse (68.35 versus 69.00 Avg2), confirming that optimizer mismatch is not an artifact of Muon taking larger steps: it persists once step size is held fixed. Second, SMS-enhanced Muon improves on both baselines on every foundation-preservation category and on Avg1, Avg2 and GSM8K, the single exception being MATH, where AdamW is 0.52 points higher. Since SMS is not moving the weights any less than the baselines here, its advantage must come from *where* it moves them. This is direct evidence that the mechanism is geometry-aware regularization rather than an implicit reduction in effective step size.

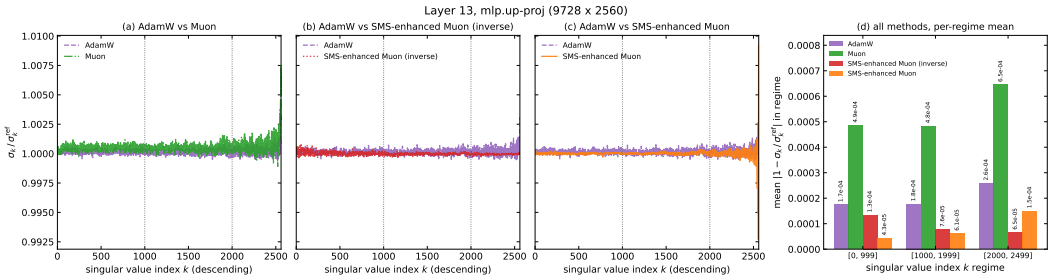

Figure 6: Perturbation of the pretrained singular-value spectrum induced by fine-tuning, measured on `Layer-13 mlp.up_proj` of Qwen3-4B-Base at the matched update magnitudes of Tab. 6. Bars report the relative deviation $|1 - \sigma_k/\sigma_k^{\text{ref}}|$ of Eq. 26, averaged over the principal ($k \in [0, 999]$), intermediate ($k \in [1000, 1999]$) and minor ($k \in [2000, 2499]$) regimes of the rank-2560 spectrum. (a) AdamW versus standard Muon: the mismatched optimizer displaces the pretrained spectrum further in every regime. (b) SMS-enhanced Muon (inverse), which reverses the plasticity profile via Eq. 27 and concentrates perturbation on the principal components. (c) SMS-enhanced Muon, which leaves the principal components largely intact and confines perturbation to the minor components.

B.3 Where fine-tuning Perturbs the Pretrained Spectrum

The matched-magnitude comparison shows that the update direction is what matters. We now examine that direction directly, by measuring how fine-tuning deforms the singular-value spectrum of the pretrained weights.

Let σ_k^{ref} and σ_k denote the k-th singular value of a given layer before and after fine-tuning, with indices sorted in descending order and aligned between the two models. We quantify the perturbation at index k by the relative deviation

$$\delta_k = \left| 1 - \frac{\sigma_k}{\sigma_k^{\text{ref}}} \right|, \tag{26}$$

so that $\delta_k = 0$ means the pretrained singular value is left untouched. To summarize where along the spectrum a method concentrates its perturbation, we average δ_k over three contiguous index regimes of the rank-2560 spectrum, corresponding to its principal ($k \in [0, 999]$), intermediate ($k \in [1000, 1999]$) and minor ($k \in [2000, 2499]$) components. All runs compared here use the matched update magnitudes of Sec. B.2, so differences in δ_k cannot be attributed to differences in step size.

Fig. 6(a) contrasts AdamW with standard Muon. Muon perturbs the pretrained spectrum more than AdamW in every regime, which is the spectral signature of optimizer mismatch: AdamW's implicit bias is the one under which the checkpoint was formed, so its updates sit more comfortably within the existing weight structure, whereas Muon's spectral-norm geometry displaces it further. Fig. 6(c) adds SMS-enhanced Muon. Its perturbation is markedly smaller on the principal components and is instead concentrated on the minor ones, exactly the redistribution that spectral momentum steering is designed to produce: plasticity decays with the pretrained singular value, so the update is pushed away from high-energy directions and into the tail of the spectrum.

To confirm that this redistribution is the operative mechanism rather than an incidental correlate, we invert it. Replacing the plasticity coefficient of Eq. 10 by

$$p_i = \exp(-\gamma(1 - q_i)) \tag{27}$$

assigns *higher* plasticity to principal directions and lower plasticity to minor ones, reversing the spectral profile while leaving every other aspect of the method unchanged. We denote this variant SMS-enhanced Muon (inverse). Fig. 6(b) confirms that it does what is intended: the perturbation now concentrates on the principal components. Its performance collapses accordingly, to 78.40 Avg1 and 68.28 Avg2, against 96.93 and 69.20 for SMS under the same update magnitude. Both numbers land essentially on standard Muon's (77.90 and 68.35), so reversing the spectral ordering does not merely reduce the benefit of SMS but removes it entirely. The advantage of SMS can therefore be attributed to spectral regularization that limits perturbations to the principal components of the pretrained activation structure.

Table 7: Optimizer comparison across AdamW- and Muon-pretrained foundation models. Pretraining and fine-tuning columns indicate the respective optimizers. The learning rate is selected using the pretraining optimizer and shared by both fine-tuning optimizers. Avg1 is the mean foundation-preservation percentage relative to each model's own pretrained baseline; Avg2 is the mean GSM8K and MATH accuracy in percent. Bold indicates the higher score within each model.

Model	Pretraining	Learning rate	fine-tuning	Avg1 ↑	Avg2 ↑
Qwen3-4B-Base	AdamW	3×10^{-4}	AdamW	**78.22**	**68.42**
			Muon	67.96	67.98
Llama-3.1-8B-Base	AdamW	4×10^{-4}	AdamW	**88.87**	**53.53**
			Muon	86.53	52.27
Moonlight-16B-A3B	Muon	9×10^{-5}	AdamW	100.97	65.04
			Muon	**106.15**	**65.38**

B.4 Optimizer Mismatch of Muon-pretrained Foundation Model

We further examine optimizer mismatch on Moonlight-16B-A3B, a foundation model pretrained with Muon (Liu et al., 2025), and compare its behavior with the AdamW-pretrained Qwen3-4B-Base and Llama-3.1-8B-Base models. We use a consistent learning-rate selection rule: first select the learning rate using the optimizer employed during pretraining, then evaluate the alternative fine-tuning optimizer at that same rate. For Qwen3-4B-Base and Llama-3.1-8B-Base, AdamW fine-tuning determines $\eta = 3 \times 10^{-4}$ and $\eta = 4 \times 10^{-4}$, respectively, which are also used for Muon fine-tuning. For Moonlight-16B-A3B, we instead select $\eta = 9 \times 10^{-5}$ using Muon fine-tuning and evaluate AdamW at this learning rate.

Table 7 shows that AdamW fine-tuning achieves higher foundation preservation and downstream adaptation on both AdamW-pretrained models. Its Avg1 advantage over Muon is 10.26 percentage points on Qwen3-4B-Base and 2.34 points on Llama-3.1-8B-Base, with corresponding Avg2 advantages of 0.44 and 1.26 points. The ranking reverses on the Muon-pretrained Moonlight model: Muon achieves Avg1 of 106.15, compared with 100.97 for AdamW, and a slightly higher Avg2 of 65.38, compared with 65.04. Thus, Muon improves Avg1 by 5.18 points and Avg2 by 0.34 points over AdamW in this setting. Since Avg1 is normalized to each model's pretrained baseline, values above 100 indicate an average improvement over that baseline. This reversal is consistent with the optimizer mismatch reported in prior work (Liu et al., 2025; Qu et al., 2026): under the evaluated settings, the optimizer used during pretraining also yields higher preservation and adaptation scores during fine-tuning.

B.5 Evaluation on Various Downstream Adaptation Domains

Setup. To examine whether the benefits of SMS extend beyond mathematical reasoning, we repeat the Qwen3-4B-Base comparison on code generation. We fine-tune on the first 100,000 examples of CodeFeedback-Filtered-Instruction (Zheng et al., 2024), using its query and answer fields as instruction and response, and keep every other setting of the main experiments (App. A.2): one epoch, a maximum sequence length of 512 tokens, an effective batch size of 128, a cosine schedule with 3% warmup, no weight decay, and LoRA with $r = \alpha = 128$ on all seven linear projections. AdamW, standard Muon and SMS-enhanced Muon share the learning rate of the main Qwen3-4B-Base experiments, $\eta = 3 \times 10^{-4}$, which is selected by a validation set learning rate sweep, and SMS uses $\gamma = 1$, the value that maximizes Avg2 in a sweep over $\gamma \in [0, 4]$. Downstream adaptation is measured by pass@1 on HumanEval (Chen et al., 2021) (164 problems, zero-shot) and MBPP (Austin et al., 2021) (500 problems, three-shot), both with greedy decoding and a 512-token generation budget through `lm-evaluation-harness`. For HumanEval, we normalize the indentation of the first generated line before the completion is appended to the prompt and executed, since a misindented first line otherwise raises an `IndentationError` and scores the sample as failed regardless of its correctness. Foundation preservation is evaluated on the same six-category suite as in the main experiments (App. A.1).

Results. Tab. 8 reports the results. Compared with the math adaptation setting, we observe milder forgetting across the preservation benchmarks, suggesting a smaller distribution shift between the

Table 8: Code adaptation on Qwen3-4B-Base fine-tuned on CodeFeedback. The base-model row reports raw benchmark scores as fractions; all remaining rows report percentages, foundation-preservation columns relative to the base-model score and adaptation columns as raw pass@1. Bold indicates the best result.

	Foundation-Preservation							Adaptation			
Methods	Instruction	Truthfulness	Knowledge	Reading	Reasoning	Commonsense	Avg1	MBPP	HumanEval	Avg2	GM
Qwen3-4B-Base	0.42	0.53	0.23	0.18	0.51	0.67	0.42	-	-	-	-
Adam (LoRA)	71.09	102.00	50.94	73.16	98.37	100.65	82.70	**69.20**	68.29	68.75	75.40
Muon (LoRA)	**72.42**	101.62	42.46	76.36	98.31	100.80	81.99	68.00	62.81	65.40	73.22
Muon+SMS (LoRA)	70.05	**102.07**	**73.59**	**90.83**	**99.02**	**100.84**	**89.40**	69.00	**71.95**	**70.48**	**79.38**

pretraining and downstream data. The optimizer mismatch nonetheless persists, as Muon adapts noticeably worse than AdamW. SMS-enhanced Muon achieves the strongest preservation by a clear margin, with the largest gains in Knowledge Retrieval and Reading Comprehension, and it not only closes Muon's adaptation gap but surpasses AdamW, attaining the best HumanEval score as well as the best Avg2 and GM overall. Due to the milder forgetting compared to the math adaptation task, a smaller γ value is enough to compensate for the forgetting.

C Additional Proofs

C.1 Muon as Spectral-Norm Normalized Momentum Descent

For completeness, recall that the nuclear norm of a matrix $A \in \mathbb{R}^{m\times n}$ is

$$\|A\|_* := \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i(A). \tag{28}$$

The nuclear and spectral norms are dual:

$$\|A\|_* = \max_{\|B\|_{\mathrm{op}}\le 1} \langle A, B\rangle. \tag{29}$$

Proof. Let

$$M_t = U_t \Sigma_t V_t^\top \tag{30}$$

be the compact singular value decomposition of M_t, with $r_t = \operatorname{rank}(M_t) > 0$. We first derive an upper bound on the objective in equation 1. For any feasible matrix Δ satisfying $\|\Delta\|_{\mathrm{op}} \le 1$, von Neumann's trace inequality gives

$$\langle M_t, \Delta\rangle \le \sum_{i=1}^{\min(d_{\mathrm{out}}, d_{\mathrm{in}})} \sigma_i(M_t)\sigma_i(\Delta) \tag{31}$$

$$= \sum_{i=1}^{r_t} \sigma_{t,i}\sigma_i(\Delta). \tag{32}$$

Since

$$\sigma_i(\Delta) \le \sigma_1(\Delta) = \|\Delta\|_{\mathrm{op}} \le 1, \tag{33}$$

we obtain

$$\langle M_t, \Delta\rangle \le \sum_{i=1}^{r_t} \sigma_{t,i} = \|M_t\|_*. \tag{34}$$

Equivalently, this bound follows directly from the duality between the nuclear and spectral norms:

$$\langle M_t, \Delta\rangle \le \|M_t\|_* \|\Delta\|_{\mathrm{op}} \le \|M_t\|_*. \tag{35}$$

It remains to construct a feasible matrix that attains this upper bound. Define

$$Q_t := U_t V_t^\top. \tag{36}$$

Because U_t and V_t have orthonormal columns, the nonzero singular values of Q_t are all equal to one. Therefore,

$$\|Q_t\|_{\mathrm{op}} = 1, \tag{37}$$

and Q_t is feasible for equation 1. Moreover,

$$\begin{aligned}
\langle M_t, Q_t \rangle &= \mathrm{tr}\left(M_t^\top Q_t\right) && (38)\\
&= \mathrm{tr}\left(V_t \Sigma_t U_t^\top U_t V_t^\top\right) && (39)\\
&= \mathrm{tr}\left(V_t \Sigma_t V_t^\top\right) && (40)\\
&= \mathrm{tr}\left(\Sigma_t V_t^\top V_t\right) && (41)\\
&= \mathrm{tr}(\Sigma_t) && (42)\\
&= \sum_{i=1}^{r_t} \sigma_{t,i} = \|M_t\|_*. && (43)
\end{aligned}$$

Thus, $Q_t = U_t V_t^\top$ attains the upper bound in equation 34, proving that

$$U_t V_t^\top \in \underset{\|\Delta\|_{\mathrm{op}} \le 1}{\arg\max} \langle M_t, \Delta \rangle. \tag{44}$$

Substituting this direction into the NMD update gives

$$W_{t+1} = W_t - \eta_t U_t V_t^\top, \tag{45}$$

which is the exact Muon update.

When M_t is rank deficient, the maximizer of equation 1 need not be unique because the objective is insensitive to components acting entirely within the null singular subspaces. The compact polar factor $U_t V_t^\top$ provides a canonical maximizer. When $M_t = 0$, every matrix satisfying $\|\Delta\|_{\mathrm{op}} \le 1$ is optimal; by convention, we set $\Delta_t = 0$. □