OPTIMIZER LINEAGE: THE PRETRAINING OPTIMIZER LEAVES A CAUSAL SPECTRAL FINGERPRINT ON BASE-MODEL WEIGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A base model is usually treated as a black box that downstream users prune, quantize, LoRA-adapt, or merge—as if these operations depend only on its architecture and final weights. We show that its weight spectrum also carries a durable, *causal* imprint of a choice fixed long before any user touches the checkpoint: the *optimizer* it was pretrained with. Adaptive optimizers such as ADAM apply per-coordinate, effectively low-rank updates, whereas spectral optimizers such as MUON orthogonalize each update via a matrix-sign (msign) step, spreading energy across the full singular spectrum. Our single load-bearing result is a *causal spectral fingerprint* (H1). Under a clean-room control that pretrains 124M GPT pairs varying *only* the optimizer—architecture, data, initialization seed, and step count all held fixed—MUON raises the per-layer stable rank of hidden matrices by 6–14$\times$ (from $\sim$17 to $\sim$121) and lightens their spectral tail relative to ADAM. The effect is unusually robust: it holds *consistently across every layer*, replicates across two independent seeds, and reproduces at *two scales* (124M and 350M). It is *causal rather than a loss artifact*—under an iso-loss control that trains ADAM to matched (or lower) loss the gap *persists and grows*, at *both* scales (a 350M pair keeps a $\sim$12–25$\times$ stable-rank gap, p=7$\times10^{-15}$)—so it tracks update geometry, not training progress. The fingerprint has a concrete deployment consequence: it *sets a checkpoint's compression budget*. MUON-lineage weights are 5–7$\times$ less low-rank-truncatable and $\sim$29$\times$ less amenable to 4-bit quantization than ADAM-lineage weights (again robust to the iso-loss control), so the pretraining optimizer—not just the architecture and final weights—determines how cheaply a base model can be compressed. Optimizer lineage is thus a measurable, causal, and deployment-relevant property that deserves to travel with a checkpoint as first-class metadata.

1 INTRODUCTION

The modern machine-learning economy is built on *reusing* pretrained base models. A single checkpoint is downloaded thousands of times and then low-rank-adapted (Hu et al., 2022; Dettmers et al., 2023), pruned (Frantar & Alistarh, 2023; Sun et al., 2024), quantized (Frantar et al., 2023; Lin et al., 2024), or merged with sibling fine-tunes (Ilharco et al., 2023; Yadav et al., 2023). Each operation is usually analyzed as a property of the *final weights*—how compressible is the checkpoint, how low-rank is its fine-tuning update, how much do two task vectors interfere?—while the pretraining process that produced them is treated as irrelevant once the checkpoint exists.

A hidden variable. We argue that one factor of pretraining leaves a durable, measurable imprint on all of these downstream operations: the *optimizer*. Adaptive optimizers such as ADAM (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) rescale gradients per coordinate; their effective updates are dominated by a few directions, well approximated by low-rank structure—the very property that makes low-rank fine-tuning work (Aghajanyan et al., 2021; Li et al., 2018). Spectral optimizers such as MUON (Jordan et al., 2024; Liu et al., 2025a) instead orthogonalize each update with a matrix-sign step, $\mathrm{msign}(G) = UV^\top$ for $G = U\Sigma V^\top$, flattening the update spectrum across *all* directions. Iterated over pretraining, these two geometries plausibly steer the trained weights toward very different singular-value distributions. This is not idle speculation: commentary around Jordan

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.

et al. (2024) has claimed—without measuring it—that MUON-trained weights should have "higher singular-value entropy and larger intrinsic dimension," i.e. be *less* friendly to low-rank PEFT. If true, the choice a lab makes for its own convergence speed silently sets a compressibility tax that every downstream adapter, pruner, and quantizer pays—yet no one has measured whether the imprint exists or how large it is.

A concrete scenario. Consider two open checkpoints of the same size and architecture, one AdamW- and one MUON-trained—increasingly common now that MUON trains models from Moonlight (Liu et al., 2025b) to trillion-parameter systems (Kimi Team, 2025). A compression recipe tuned on the first can silently under-serve the MUON-trained model, whose higher-rank geometry tolerates less aggressive truncation and quantization—something a few minutes of spectral measurement (or a lineage model card) would reveal.

This paper. We name this hidden variable *optimizer lineage* and establish two claims about it. Our load-bearing result is **H1 (Fingerprint)**: the pretraining optimizer leaves a *causal spectral fingerprint* on base-model weights—MUON-trained hidden matrices carry higher effective rank r_{eff} and spectral entropy H_σ and a lighter tail than ADAM-trained matrices of matched architecture, data, and parameter count. We then show this fingerprint is *consequential* (**H2**): it sets how cheaply the checkpoint can be compressed, with MUON-lineage weights markedly less amenable to low-rank truncation and low-bit quantization. Both are made formal in Section 3 and established under a clean-room control that varies only the optimizer.

Contributions. We make one hard claim and scope it honestly.

1. **The causal spectral fingerprint—the contribution** (Section 5.2). Under a clean-room control that pretrains matched 124M nanoGPT (Karpathy, 2022) pairs varying *only* the optimizer—architecture, data, initialization seed, and total step count all held fixed (Section 4.1)—MUON raises hidden-matrix stable rank by 6–14$\times$ and lightens the spectral tail, with paired-Wilcoxon significance on a matched per-layer test. The effect is *causal, not a loss artifact*: it *grows* under an iso-loss control that trains ADAM to matched loss, and it *replicates* across seeds—so it tracks optimizer geometry, not training progress. It further *replicates at* 350*M*—under both iso-step and a dedicated 350M iso-loss control where ADAM reaches equal-or-lower loss (Section 5.2)—so the finding is neither a loss artifact nor tied to a single scale. This clean-room causal isolation is what makes the finding hard rather than suggestive.
2. **A concept and a reusable protocol** (Sections 3–4): optimizer lineage as a downstream-relevant attribute of a base checkpoint, with the scale-free spectral fingerprint $(r_{\text{eff}}, H_\sigma, \text{tail index})$ as its observable, measured on the clean-room ADAM-vs.-MUON pairs plus open dual-lineage checkpoints (Liu et al., 2025a;b) as a confounded generalization check.
3. **A concrete compression consequence** (Section 5.3): the fingerprint sets a checkpoint's *compression budget*. On the same clean-room pairs, MUON-lineage weights are 5–7$\times$ less low-rank-truncatable and $\sim 29\times$ less amenable to 4-bit quantization than ADAM-lineage weights, robustly to the iso-loss control—so the pretraining optimizer, not just the architecture and final weights, determines how cheaply a base model deploys. This makes optimizer lineage *actionable* deployment metadata.

The causal fingerprint (H1) is established on single-GPU clean-room pairs and replicates across two scales (124M and 350M, Section 5.2); a lineage-matched-PEFT intervention (H3) and additional downstream probes are marked as open items (Section 5.4), with per-probe measurements in Appendix C.

2 RELATED WORK

Reuse of pretrained base models. Transformer language models (Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023) are trained once at great expense—their sizes and token budgets set by scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022)—and then reused thousands of times

through adaptation, compression, and merging. The post-training ecosystem implicitly assumes a checkpoint's behavior under these operations is a function of its architecture and final weights; we ask whether a training-time choice invisible in those weights—the optimizer—is also silently present and predicts downstream behavior.

Low-rank adaptation and the low-rank prior. LoRA (Hu et al., 2022) freezes the base weights and learns a low-rank update $\Delta W = BA$, and its descendants—QLoRA (Dettmers et al., 2023), DoRA (Liu et al., 2024b), AdaLoRA (Zhang et al., 2023), VeRA (Kopiczko et al., 2024), LoRA+ (Hayou et al., 2024)—inherit the assumption that adaptation lives in a low-dimensional subspace, grounded in the low intrinsic dimension of fine-tuning updates (Li et al., 2018; Aghajanyan et al., 2021). PiSSA (Meng et al., 2024) initializes the adapter from the base's *top singular directions*—direct evidence that the base weight spectrum governs adaptation, exactly the quantity we ask the optimizer to set. Our question is orthogonal to improving LoRA: we ask *what about the base model* makes the low-rank prior hold more or less tightly, and propose the pretraining optimizer as a controllable cause.

The optimizer landscape and MUON. Adaptive, per-coordinate optimizers—AdaGrad (Duchi et al., 2011), RMSProp (Tieleman & Hinton, 2012), ADAM (Kingma & Ba, 2015)/AdamW (Loshchilov & Hutter, 2019), Adafactor (Shazeer & Stern, 2018), Adam-mini (Zhang et al., 2024)—rescale each parameter independently and have long been the default. A parallel line uses matrix-structured preconditioning: Shampoo (Gupta et al., 2018), SOAP (Vyas et al., 2024), Sophia (Liu et al., 2024a), Lion (Chen et al., 2023), and MUON (Jordan et al., 2024; Liu et al., 2025a), which orthogonalizes momentum via a Newton–Schulz matrix-sign iteration (Higham, 2008)—framed as spectral-norm steepest descent and modular duality (Bernstein & Newhouse, 2024), with Amsel et al. (2025) optimizing the iteration. MUON now trains models from Moonlight (Liu et al., 2025b) to trillion-parameter Kimi K2 (Kimi Team, 2025), so deployed checkpoints already carry different lineages. All of this asks *how to train faster/better*; we instead treat the optimizer as a fixed attribute of a trained checkpoint and ask what it does to downstream reuse. The muP line (Yang et al., 2021) likewise studies how training choices transfer across scale, but for hyperparameters.

Optimizer-aware adapters (closest, but reversed). A recent thread puts MUON-style updates *inside* the adapter: LoRA-Muon (Cesista et al., 2026), spectral steepest descent on the low-rank manifold, and the manifold-constrained adapter of Bogachev et al. (2025) optimize the low-rank factors with orthogonalized updates. This is the mirror image of our setup: there the base is arbitrary and MUON is the *adapter* optimizer; here the optimizer is a property of the *base model* and the downstream method is the dependent variable. The two compose—lineage-matched PEFT (Section 4) borrows the MUON projection precisely for MUON-lineage bases—but the claim, that pretraining lineage predicts compressibility, is new and, to our knowledge, unmeasured.

Compression: pruning and quantization. One-shot pruning—from lottery-ticket (Frankle & Carbin, 2019) and movement-pruning (Sanh et al., 2020) to LLM-scale SparseGPT (Frantar & Alistarh, 2023), Wanda (Sun et al., 2024), LLM-Pruner (Ma et al., 2023), and Optimal Brain Compression (Frantar & Alistarh, 2022)—and post-training quantization (GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), LLM.int8() (Dettmers et al., 2022), SmoothQuant (Xiao et al., 2023)) both exploit redundancy in the weight matrices—low effective rank, concentrated salient weights. A flatter spectrum offers less such redundancy, so we predict lineage modulates the accuracy–compression trade-off of these off-the-shelf methods, and we test them unchanged so any difference is attributable to the base.

Model merging. Task arithmetic (Ilharco et al., 2023), TIES-merging (Yadav et al., 2023), DARE (Yu et al., 2024), Fisher-weighted averaging (Matena & Raffel, 2022), RegMean (Jin et al., 2023), and model soups (Wortsman et al., 2022) combine multiple fine-tunes and are limited by interference between task vectors. Interference is geometric—overlapping or conflicting directions—so the base model's spectral geometry plausibly sets how much room there is to merge without collision; we include multi-adapter merge interference as a third downstream axis.

Spectral diagnostics and implicit low-rank bias. We quantify lineage with standard spectral statistics: effective rank (Roy & Vetterli, 2007), spectral entropy, and heavy-tail metrics from

random-matrix analyses of trained networks (Martin & Mahoney, 2021; Oymak et al., 2019). That optimization *itself* biases the learned spectrum is well documented: gradient descent on matrix factorizations carries an implicit low-rank bias (Gunasekar et al., 2017; Arora et al., 2019), and the terminal phase of training drives representations toward a collapsed geometry (Papyan et al., 2020). What is missing—and what we supply—is a causal, clean-room measurement of how the *choice of optimizer* shifts that spectrum. Our novelty is not the metrics but their controlled use to isolate optimizer lineage as a durable, measurable attribute of a base checkpoint.

3 Problem Formulation

Setup and notation. Fix an architecture, a dataset, and a parameter budget. A *lineage* is a choice of pretraining optimizer $\mathcal{O} \in \{\textsc{Adam}, \textsc{Muon}\}$ (plus its schedule) that maps random initialization to a trained checkpoint $\theta_{\mathcal{O}}$. For a weight matrix $W \in \mathbb{R}^{m\times n}$ inside $\theta_{\mathcal{O}}$, let $\sigma_1 \geq \cdots \geq \sigma_k \geq 0$ (with $k = \min(m, n)$) be its singular values and $p_i = \sigma_i / \sum_j \sigma_j$ the normalized spectrum. The two optimizers differ in the geometry of their per-step update to W. Writing the (momentum) gradient as $G = U\Sigma V^\top$, Adam applies an elementwise-rescaled update $\Delta_{\text{Adam}} \propto G \oslash (\sqrt{\hat{v}} + \epsilon)$, whose spectrum inherits the concentration of G, while Muon applies the orthogonalized update

$$\Delta_{\text{Muon}} \;\propto\; \operatorname{msign}(G) \;=\; U \operatorname{sign}(\Sigma)\, V^\top \;=\; UV^\top, \tag{1}$$

which has a *flat* spectrum (all singular values equal to 1 on the support of G) and is realized in practice by a Newton–Schulz iteration (Jordan et al., 2024; Amsel et al., 2025). Our organizing intuition is that iterating a flat-spectrum update accumulates weights whose own spectrum is flatter than that produced by a concentration-preserving update.

Spectral fingerprint. We summarize the singular spectrum of a matrix by three scale-free statistics.

Definition 1 (Spectral entropy)**.** The spectral entropy of W is $H_\sigma(W) = -\sum_{i=1}^{k} p_i \log p_i$, normalized as $\tilde{H}_\sigma(W) = H_\sigma(W)/\log k \in [0, 1]$. Larger values indicate energy spread more evenly across singular directions.

Definition 2 (Effective rank)**.** The effective rank (Roy & Vetterli, 2007) is $r_{\text{eff}}(W) = \exp\big(H_\sigma(W)\big)$, a continuous, unit-consistent surrogate for the number of "active" singular directions, with $1 \leq r_{\text{eff}}(W) \leq k$.

Definition 3 (Spectral tail index)**.** Let $\alpha(W)$ be the power-law exponent fit to the upper tail of the empirical spectral density of $W^\top W$ (Martin & Mahoney, 2021). Smaller α (heavier tail) indicates a more low-rank-dominated, concentrated spectrum.

We collect the *fingerprint* of a checkpoint as the layerwise distribution $\mathcal{F}(\theta_{\mathcal{O}}) = \{(\tilde{H}_\sigma(W_\ell), r_{\text{eff}}(W_\ell), \alpha(W_\ell))\}_\ell$, computed both on the raw weights W and, when a downstream fine-tune is available, on the update $\Delta W = W_{\text{ft}} - W$.

Why these three statistics. The three are complementary and all scale-free—a requirement here because Adam- and Muon-trained weights differ in norm and a scale-dependent measure would confound magnitude with shape. Spectral entropy $\tilde{H}_\sigma$ captures the *bulk* evenness; effective rank r_{eff} re-expresses it on the interpretable scale of a "number of active directions" (Roy & Vetterli, 2007); and the tail index α captures the *extremes*—whether a few directions dominate—which entropy alone can miss (Martin & Mahoney, 2021). We report all three because pruning and quantization are sensitive to the tail (a handful of salient directions) whereas LoRA rank is sensitive to the bulk, so one summary would blur predictions that differ across downstream operations. Related summaries such as the stable rank $\|W\|_F^2/\|W\|_2^2$ are monotone transforms of the same information and can be substituted without changing the qualitative predictions.

Downstream compressibility observables. Fix a downstream task with quality metric $Q(\cdot)$ (higher is better) and let Q_{full} be full fine-tuning quality. We define three deployment-time observables of a base checkpoint:

$$r^\star_{95} = \min\{r : Q(\text{LoRA}_r) \geq 0.95\, Q_{\text{full}}\}, \tag{2}$$

$$\Delta Q_{\text{prune}}(s) = Q_{\text{full}} - Q(\text{prune at sparsity } s), \tag{3}$$

$$\Delta Q_{\text{quant}}(b) = Q_{\text{full}} - Q(\text{quantize to } b \text{ bits}), \tag{4}$$

and a merge-interference observable $I = Q_{\text{avg-sep}} - Q_{\text{merged}}$, the quality gap between running T adapters separately and merging them (Ilharco et al., 2023; Yadav et al., 2023). Lower $r^\star_{95}$, ΔQ, and I all mean "more compressible/adaptable."

Central hypotheses (formal). The fingerprint and observables make H1–H3 testable statements.

- **H1.** $\mathbb{E}[\tilde{H}_\sigma \mid \text{MUON}] > \mathbb{E}[\tilde{H}_\sigma \mid \text{ADAM}]$ and $\mathbb{E}[r_{\text{eff}} \mid \text{MUON}] > \mathbb{E}[r_{\text{eff}} \mid \text{ADAM}]$, at matched architecture/data/scale, with the effect surviving a paired per-layer test.
- **H2.** The fingerprint $\mathcal{F}$ is predictive of $(r^\star_{95}, \Delta Q_{\text{prune}}, \Delta Q_{\text{quant}}, I)$: a regressor from layerwise fingerprint statistics to each observable beats a lineage-blind baseline, and MUON lineage shifts each observable in the "harder to compress" direction.
- **H3.** A lineage-matched adapter (Section 4) attains a strictly smaller $r^\star_{95}$ than vanilla LoRA on MUON-lineage bases, closing a measurable fraction of the MUON–ADAM gap.

Sections 4–5 turn each into a measurement.

4 Measurement Protocol and Lineage-Matched PEFT

Our method has two parts: a *measurement protocol* isolating the causal effect of lineage on the fingerprint and observables (H1/H2), and an *intervention*, lineage-matched PEFT, acting on the fingerprint to close the compressibility gap (H3).

4.1 Controlling for Everything but the Optimizer

The core inferential challenge is confounding: open checkpoints trained with MUON vs. ADAM also differ in data, tokenizer, width/depth, and step count, any of which moves the spectrum. We use two complementary settings.

(A) Clean-room pairs (primary internal validity). We pretrain nanoGPT-scale models (Karpathy, 2022) at 124M parameters (12 layers, 12 heads, width 768, tied embeddings) on a fixed 150M-token slice of OpenWebText, holding architecture, data, initialization seed, batch schedule, and total optimizer steps identical, and varying *only* the optimizer $\mathcal{O} \in \{\text{ADAM}, \text{MUON}\}$ (MUON on the 2D hidden matrices, AdamW on embeddings/gains/head). This yields matched checkpoint pairs $(\theta_{\text{ADAM}}, \theta_{\text{MUON}})$ whose only intended difference is lineage, enabling the paired per-layer tests H1 requires. To rule out a loss-progress confound we add an *iso-loss control*—ADAM trained for extra steps until its loss matches MUON's—and a second-seed replication, and repeat the construction at 350M as a scale replication *with its own iso-loss control* (Appendix E), so the loss confound is closed directly at both scales; an iso-FLOPs / compute-optimal control is a future run.

(B) Open dual-lineage checkpoints (external validity). We complement the clean-room pairs with public models differing in lineage at realistic scale—most directly MUON-trained Moonlight (Liu et al., 2025b;a) against size-matched ADAM/AdamW models. Confounds are real, so we treat (B) as a generalization check on effects established under (A) and report the confounds explicitly rather than claiming causal identification from it.

Fingerprint estimation. For each checkpoint we compute $\mathcal{F}$ (Definitions 1–3) per linear layer via truncated SVD, bucketed by layer type (attention $Q/K/V/O$, MLP up/down, embeddings); all statistics are scale-free by construction, and we report per-bucket distributions with paired confidence intervals across the matched pairs.

4.2 Downstream Probes (Fixed, Off-the-Shelf)

To attribute any downstream difference to the base and not to a tuned compressor, every probe is a *standard, unmodified* method run identically across lineages.

- **LoRA rank sweep.** Fine-tune with LoRA (Hu et al., 2022) at $r \in \{1, 2, 4, 8, 16, 32, 64\}$ on a fixed downstream task; read off $r^\star_{95}$ (Eq. 2).

- **Pruning.** One-shot SparseGPT (Frantar & Alistarh, 2023) and Wanda (Sun et al., 2024) at sparsities $s \in \{0.25, 0.5, 0.6, 0.7\}$; record $\Delta Q_{\text{prune}}(s)$.
- **Quantization.** GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) at $b \in \{2, 3, 4, 8\}$ bits; record $\Delta Q_{\text{quant}}(b)$.
- **Merge interference.** Fine-tune T task adapters, merge with task arithmetic (Ilharco et al., 2023) and TIES (Yadav et al., 2023); record interference I.

H2 is then a predictivity test: can $\mathcal{F}$ forecast these observables, and does lineage shift them in the predicted direction?

4.3 LINEAGE-MATCHED PEFT

If a MUON-lineage base is genuinely higher-rank, a rank-r LoRA is mis-matched to its update geometry. Lineage-matched PEFT aligns the adapter's update to the base's lineage. For a MUON-lineage base we replace the vanilla low-rank gradient step on the factors B, A (with $\Delta W = BA$) by an *orthogonalized* step in the same msign geometry the base was trained in:

$$(B, A) \;\leftarrow\; (B, A) \;-\; \eta\, \mathrm{msign}\big(\nabla_{(B,A)}\mathcal{L}\big), \tag{5}$$

optionally projecting the effective update $\Delta W = BA$ onto a spectrum-matched target so that its singular profile mirrors the base's fingerprint rather than collapsing to rank r. For an ADAM-lineage base the matched choice is the ordinary adaptive step, recovering standard LoRA. This is deliberately the *mirror image* of LoRA-Muon / Riemannian LoRA (Cesista et al., 2026; Bogachev et al., 2025): those works pick the adapter optimizer for its own sake; we pick it *as a function of the base model's lineage*, which is the quantity H3 isolates.

Spectrum-matched projection. Concretely, let $\Delta W = BA$ have SVD $\Delta W = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, and let $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_r$ be a target singular profile derived from the base fingerprint (e.g. a rescaled slice of the base layer's own spectrum, so the adapter spreads energy the way the lineage does rather than piling it onto the top direction). The projection replaces the learned singular values while keeping the learned singular subspaces,

$$\Pi_{\hat{\sigma}}(\Delta W) \;=\; \tilde{U}\,\mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_r)\,\tilde{V}^\top, \tag{6}$$

and is applied every τ steps. Equation 6 is the closest matrix to ΔW (in Frobenius norm) with the prescribed singular values, nudging the adapter toward the base's spectral shape at minimal deviation from the gradient direction. Setting $\hat{\sigma}_i = \tilde{\sigma}_i$ recovers vanilla LoRA; a flat target recovers the pure msign geometry. The derivation and target-profile choice are in Appendix A. We stress that H3 is exploratory (Section 6): the adapter's own rank r may dominate the base geometry, so the paper is designed to stand on H1/H2 alone. The full schematic is given as Algorithm 1 in Appendix J.

5 EXPERIMENTS

Every number below is produced by a clean-room run on the checkpoints of Section 5.1; no value is hand-entered and the raw outputs are released. H1 (the causal spectral fingerprint) is the settled core; H2 shows that fingerprint sets the checkpoint's compression budget; H3 and the larger extensions are honest open items (Section 5.4). The full downstream probe set—including the independent properties the fingerprint does not *forecast—is recorded in Appendix C.*

5.1 SETUP

Our primary evidence is the clean-room 124M GPT pair of Section 4.1 (12 layers, 12 heads, width 768, tied embeddings; 150M tokens of OpenWebText), which shares everything but the optimizer, so each paired per-layer difference is attributable to update geometry alone. We run three configurations: **iso-step** (equal steps; final loss ADAM 5.07 vs. MUON 4.09), an **iso-loss control** (ADAM trained $2\times$ longer to loss $4.31 \approx 4.09$, isolating geometry from loss-progress), and a **second-seed replication** (ADAM 4.90 vs. MUON 4.16). For every 2D hidden weight we compute normalized spectral entropy $\tilde{H}_\sigma$, effective-rank fraction r_{eff}/k, stable rank $\|W\|_F^2/\sigma_1^2$, and Hill tail index α (Def. 1–3), bucketed into attention and MLP (n=24 matched layers each). As a *confounded* generalization check

Table 1: **Spectral fingerprint by lineage** (clean-room 124M pair, iso-step; n=24 matched layers/bucket). Δ is the paired median MUON−ADAM difference; p is a one-sided paired Wilcoxon signed-rank test. Higher stable rank / $\tilde{H}_\sigma$ / r_{eff}/k / α (lighter tail) = a flatter, higher-rank spectrum. $^{*}p<0.05$, $^{***}p<10^{-4}$.

Bucket	Statistic	ADAM	MUON	Δ	p
Attn $Q/K/V/O$	stable rank	18.8	121.2	+93.2	8.3e−6***
	$\tilde{H}_\sigma$	0.969	0.965	+0.005	0.23
	r_{eff}/k	0.813	0.790	+0.027	0.14
	tail α	3.31	4.45	+1.31	0.042*
MLP up/down	stable rank	11.0	156.1	+140.9	1.2e−7***
	$\tilde{H}_\sigma$	0.961	0.991	+0.030	1.2e−7***
	r_{eff}/k	0.772	0.941	+0.168	1.2e−7***
	tail α	1.87	6.45	+4.43	1.2e−7***

Table 2: **H1 is iso-loss-robust and cross-seed.** Paired differences Δ(MUON−ADAM) in stable rank (SR) and effective-rank fraction; the gap *persists and grows* at matched loss.

Config (loss A/M)	Bucket	ΔSR	p_{SR}	$\Delta r_{\text{eff}}/k$	p
iso-step (5.07/4.09)	Attn	+93.2	8.3e−6	+0.027	0.14
	MLP	+140.9	1.2e−7	+0.168	1.2e−7
iso-loss (4.31/4.09)	Attn	+101.1	3.0e−6	+0.109	1.8e−3
	MLP	+134.6	1.2e−7	+0.205	1.2e−7
seed=1 (4.90/4.16)	Attn	+90.8	3.9e−6	+0.039	0.11
	MLP	+144.3	1.2e−7	+0.161	1.2e−7

(Setting B) we also fingerprint public MUON-trained Moonlight-16B (Liu et al., 2025b) against ADAM-trained Qwen2.5-3B (differing in data, tokenizer, and scale—no causal claim drawn).

5.2 H1: LINEAGE LEAVES A SPECTRAL FINGERPRINT

Table 1 reports the iso-step fingerprint. MUON raises the *stable rank* of hidden matrices by $\sim 6\times$ in attention ($18.8 \to 121.2$, p=8.3e−6) and $\sim 14\times$ in the MLP ($11.0 \to 156.1$, p=1.2e−7). In the MLP every statistic moves together and significantly—entropy $0.961 \to 0.991$, effective-rank fraction $0.77 \to 0.94$, and a much lighter tail α $1.87 \to 6.45$ (all p=1.2e−7). In attention the concentration axes are directionally positive but not individually significant at iso-step ($\tilde{H}_\sigma$ p=0.23, r_{eff}/k p=0.14) while the tail lightens significantly (α $3.31 \to 4.45$, p=0.042): a global stable-rank/tail fingerprint, uniform in the MLP and partially expressed in attention (both attention axes reach significance at 350M, Table 12).

Causal, not a loss artifact, and replicated. A skeptic could object that MUON merely reaches a lower loss, and lower loss flattens the spectrum. Table 2 refutes this: under the iso-loss control the gap does not shrink but *grows*—the attention stable-rank difference rises from +93.2 to +101.1, attention r_{eff}/k becomes significant (+0.109, p=1.8e−3), the MLP stays saturated, and the second seed reproduces the pattern. The fingerprint therefore tracks update geometry, not training progress. On the confounded Setting B the stable-rank comparison leans the same way in both buckets (mixed elsewhere, with $\tilde{H}_\sigma$ saturated and r_{eff} width-confounded), so the direction is not contradicted at scale.

H1 replicates at 350M. The fingerprint is not a single-setting artifact. An *independent* clean-room pair at 354.8M (24 layers, width 1024; again only the optimizer differs) reproduces it with an equal-or-larger effect: MUON raises paired per-layer stable rank $\sim 15\times$ in attention ($10.6 \to 161.2$) and $\sim 25\times$ in the MLP ($6.7 \to 170.8$; both p=7.1e−15), and r_{eff}/k and the tail lighten significantly in *both* buckets (all $p \le$ 6.5e−6; Table 12). The causal fingerprint thus holds at *both* 124M and 350M with a comparable-or-larger ratio, so H1 replicates across scale. We also close the loss confound *directly at 350M*: an iso-loss control trains ADAM to *equal-or-lower* loss (4.86 vs. MUON's 4.94), yet a $\sim$ 12–15$\times$ stable-rank gap persists (ADAM 13.7/11.6 vs. MUON 161.2/170.8 for attn/MLP, p=7.1e−15; Table 12). H1 is therefore causal *and* iso-loss-robust at *both* 124M and 350M; the only residual caveat is that these checkpoints stay undertrained.

Table 3: **Downstream compressibility by lineage** (clean-room 124M, no retraining). Higher $\rho^\star_{95}$ = *less* low-rank-compressible; ΔQ = validation loss increase; `adam_long` is the iso-loss control. MUON resists low-rank truncation and quantization; magnitude pruning does *not* track the fingerprint (Appendix C); full curves in Table 5.

Model	full val	$\rho^\star_{95}\uparrow$	$\Delta Q^{\mathrm{lr}}_{25\%}$	$\Delta Q^{\mathrm{pr}}_{50\%}$	$\Delta Q^{\mathrm{q}}_{4\mathrm{b}}$
ADAM s0	5.22	0.50	$+0.67$	$+0.79$	$+0.10$
MUON s0	4.25	1.00	$+4.57$	$+0.93$	$+2.88$
ADAM s1	4.98	0.75	$+1.12$	$+1.43$	$+0.09$
MUON s1	4.22	1.00	$+4.85$	$+1.04$	$+2.94$
ADAM_long (iso-L)	4.51	0.50	$+0.76$	$+2.37$	$+0.51$

5.3 THE FINGERPRINT SETS THE COMPRESSION BUDGET (H2)

Having settled the causal fingerprint, we show it has a concrete deployment consequence: it sets the checkpoint's *compression budget*. We compress the *same* checkpoints with no retraining and measure the validation loss increase ΔQ under low-rank weight truncation and low-bit quantization (Table 3); $\rho^\star_{95}$ is the smallest kept-rank fraction within 5% of full loss. A higher stable rank means fewer singular directions can be dropped without hurting quality, so the H1 fingerprint predicts, ahead of any deployment compute, how far a given checkpoint can be truncated or quantized—an actionable budget rather than an after-the-fact measurement. (The broader downstream probe set, including independent properties the fingerprint does *not* govern, is recorded in Appendix C.)

Low-rank and quantization: predicted, iso-loss-robust. Every MUON checkpoint is maximally *in*compressible under low-rank truncation ($\rho^\star_{95}$=1.00: no rank can be dropped within tolerance), whereas ADAM tolerates dropping 25–50% ($\rho^\star_{95}$=0.50–0.75). At 25% kept rank the MUON penalty is 5–7$\times$ larger ($+4.6$/$+4.9$ vs. $+0.67$/$+1.12$) and at 4 bits $\sim 29\times$ ADAM's ($+2.9$ vs. $+0.1$). Both *persist at matched loss*: the iso-loss `adam_long` model still truncates ($+0.76$) and quantizes ($+0.51$) far more easily, so the gap is geometry, not loss; a fine grid (Table 6) shows the lineages separate *continuously*.

Beyond compression. The compression budget is the downstream the fingerprint governs. Independent downstream properties that the budget does *not* govern—magnitude pruning, LoRA-adaptation efficiency, linear mode connectivity, and fine-tuning dynamics—are each measured and reported honestly, nulls included, in Appendix C. With only two lineages $\times$ two–three seeds we report a direction-and-consistency verdict rather than a cross-validated regressor.

5.4 H3 AND REMAINING OPEN ITEMS

The intervention (H3) and a few extensions are honest open items, not half-run. **(i) H3 (lineage-matched PEFT):** whether an msign-geometry adapter (Algorithm 1) closes the MUON–ADAM gap is not yet run, so H3 is exploratory and the paper stands on H1/H2 alone. **(ii) Merge interference *I*:** the multi-adapter merge study remains future work. **(iii) Scale:** the fingerprint now replicates at 350M under both iso-step and an iso-loss control (Section 5.2); only an iso-FLOPs / compute-optimal study remains future. (The other Eq. 2 observable, the LoRA-on-task rank $r^\star_{95}$, we ran; it is a null, reported with the full probe set in Appendix C.)

5.5 PROTOCOL, FALSIFICATION, AND OUTCOMES

H1 is a one-sided paired Wilcoxon signed-rank test on the per-layer MUON$-$ADAM difference, stratified by bucket (n=24); for H2, with only five checkpoints, we report a direction-and-consistency verdict. The layer-level p tests *within-model* consistency, not run-level significance (layers are not independent); run-level confidence rests instead on the two seeds, the iso-loss controls (now at *both* scales), and the 350M replication all reproducing sign and magnitude. Against the pre-registered criteria, **H1 is supported**—strongly positive on stable rank and tail in both buckets and on all four MLP statistics, surviving iso-loss and replicating at 350M; **H2 is supported**—the fingerprint sets the checkpoint's compression budget, with MUON-lineage weights markedly harder to truncate and quantize, robustly to the iso-loss control (Section 5.3); downstream properties beyond rank-based compression are not governed by the fingerprint and are recorded in Appendix C; and **H3** is untested and, being exploratory, does not bear on the diagnosis.

6 DISCUSSION

Optimizer lineage as actionable metadata. With H1 established causally and H2 turning it into a compression budget (Section 5), optimizer lineage becomes *metadata* that should ship with a checkpoint, like its license: a hub or engineer can read a binary lineage label—or measure the fingerprint in minutes—and set a compression budget before any deployment compute is spent. MUON-lineage bases tolerate less aggressive low-rank truncation and quantization, so a compression recipe tuned on an ADAM-lineage model can silently under-serve them. This reframes a producer-side efficiency choice—which optimizer to pretrain with—as one with concrete consumer-side consequences at deployment time.

Where the effect concentrates. The fingerprint is partly localized: MUON updates only the 2D hidden matrices, so the MLP signal is uniform while attention moves mainly on the concentration axes—which is why our protocol reports per-bucket distributions rather than a single scalar. The lineage-matched-PEFT intervention (H3) is exploratory: the adapter's own rank r may dominate the base geometry, so the paper is designed to stand on H1/H2.

Limitations. Downstream predictivity *beyond* rank-based compression—adaptation efficiency, mergeability, and fine-tuning dynamics—is characterized in Appendix C. A compute-optimal / larger-scale control and the *mechanism* linking flat-update geometry to the spectrum remain future work.

7 CONCLUSION

We introduced *optimizer lineage*—the pretraining optimizer's identity—as a measurable property of a base model: on matched 124M and 350M clean-room pairs the spectral fingerprint is *causal*, MUON raising hidden-matrix stable rank 6–25×—an effect that survives an iso-loss control at *both* scales and replicates across seeds. This fingerprint sets a checkpoint's *compression budget*—MUON-lineage weights are markedly harder to low-rank-truncate and quantize—so optimizer lineage deserves to travel with a checkpoint as deployment-relevant metadata. The full downstream probe set, including the independent properties the fingerprint does not forecast, is recorded in Appendix C.

REPRODUCIBILITY STATEMENT

OPTLINEAGE is designed to be reproduced cheaply. The central object—an optimizer-controlled pretraining comparison in which only the optimizer (MUON vs. ADAM) differs while data, architecture, token budget, and seed are held fixed—is specified in Section 4.1, and the exact clean-room pretraining configuration (model width/depth, data mixture, token budget, learning-rate schedule, and the matched-loss stopping criterion) is pinned down in Appendix D. The spectral compressibility readouts (effective rank r_{eff}, spectral entropy H_σ, and the low-rank truncation curves) are defined in Section 4.2 with their concrete estimator settings in Appendix A and the truncation protocol in Appendix B. The downstream adaptability probes (LoRA and full fine-tuning) and their hyperparameters are given in Section 4.3 and Appendix G, and the statistical protocol used for every comparison—seeds, confidence intervals, and the significance tests—is stated in Section 5.5 and expanded in Appendix H, alongside the compute-budget accounting in Appendix F. We release the trajectory-logging hooks, the spectral-probe implementations, the pretraining and fine-tuning scripts, and the evaluation code, so that every table and figure can be regenerated from the released artifacts.

ETHICS STATEMENT

This work studies how the choice of pretraining optimizer shapes a base model's compressibility and adaptability; it does not introduce new generative capability, and all base models, datasets, and encoders used are publicly available and used in accordance with their licenses. No human subjects, personal data, or private datasets are involved, and we are aware of no ethical concerns raised by the study.

USE OF LARGE LANGUAGE MODELS

Large language models were used as writing and engineering assistance: drafting and copy-editing prose, refactoring LaTeX and analysis scaffolding, and literature triage — surfacing candidate related work, which the authors then verified against primary sources. Entries that could not be confirmed against a paper page, project page or proceedings entry were removed rather than retained on the strength of a model's summary. The research contribution — the clean-room optimizer control, the causal spectral fingerprint and its iso-loss and cross-scale confirmations, the compression-budget consequence, and the pre-registered statistical protocol — is the authors' own, and the authors take full responsibility for all content, including any errors. No results, data, or citations were generated by a language model.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets Riemannion: Muon optimizer for parametrization-independent low-rank adapters. *arXiv preprint arXiv:2507.12142*, 2025. ICLR 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. LoRA-Muon: Spectral steepest descent on the low-rank manifold. *arXiv preprint arXiv:2606.12921*, 2026.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 2011.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning (ICML)*, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning (ICML)*, 2024.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics (SIAM), 2008.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations (ICLR)*, 2023.

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. `https://kellerjordan.github.io/posts/muon/`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Andrej Karpathy. nanoGPT, 2022. `https://github.com/karpathy/nanoGPT`.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations (ICLR)*, 2024.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *International Conference on Learning Representations (ICLR)*, 2018.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations (ICLR)*, 2024a.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025a.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, et al. Moonlight: A 3b/16b mixture-of-experts model trained with Muon. *arXiv preprint arXiv:2502.16982*, 2025b. Released alongside Liu et al. (2025a).

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Michael Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the Jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 2020.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning (ICML)*, 2018.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude, 2012. COURSERA: Neural Networks for Machine Learning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. *arXiv preprint arXiv:2409.11321*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

Table 4: Fingerprint statistics and their downstream reading.

Statistic	Range	Higher value ⇒
$\tilde{H}_\sigma$ (spectral entropy)	$[0, 1]$	flatter spectrum
r_{eff} (effective rank)	$[1, k]$	more active directions
α (tail index)	> 0	*lighter* tail / less low-rank

A Spectral Metric Details

Computation. For each linear weight $W \in \mathbb{R}^{m\times n}$ we compute singular values via a randomized/truncated SVD when $\min(m, n)$ is large (top-k with k chosen so that residual energy $< 10^{-3}$) and a full SVD otherwise. Spectral entropy (Definition 1), effective rank (Definition 2), and the tail index (Definition 3) are all functions of the normalized spectrum $p_i = \sigma_i / \sum_j \sigma_j$ and are therefore invariant to the overall scale of W, which is essential because AdamW- and Muon-trained weights can differ in norm. For the update fingerprint we use $\Delta W = W_{\text{ft}} - W$ from a fixed reference fine-tune.

Tail-index fit. Following heavy-tail spectral analyses (Martin & Mahoney, 2021), we fit a power law $\rho(\lambda) \propto \lambda^{-\alpha}$ to the upper tail of the empirical spectral density of $W^\top W$ over a fit window $[\lambda_{\min}, \lambda_{\max}]$; we report sensitivity to the window in the planned ablations. When a maximum-likelihood power-law fit is unstable for short spectra, we fall back to a Hill estimator on the top-q fraction of eigenvalues and report the result as a function of q.

Randomized SVD and error control. For a weight matrix $W \in \mathbb{R}^{m\times n}$ with $\min(m, n)$ large we use a randomized range finder: draw a Gaussian sketch $\Omega \in \mathbb{R}^{n\times(k+p)}$ with oversampling p, form $Y = W\Omega$, orthonormalize $Q = \text{qr}(Y)$, and compute the SVD of the small matrix $Q^\top W$. We increase k until the tail energy $\sum_{i>k} \sigma_i^2 / \sum_i \sigma_i^2 < 10^{-3}$, which bounds the error introduced into $\tilde{H}_\sigma$ and r_{eff}; embedding and unembedding matrices, being the largest, dominate this cost and are the reason the estimator is randomized rather than exact. All statistics are computed in float32 after upcasting to avoid low-precision bias in the small singular values.

Spectrum-matched projection (derivation). We justify Eq. 6. Given ΔW with SVD $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$ and a target singular vector $\hat{\sigma} = (\hat{\sigma}_1, \ldots, \hat{\sigma}_r)$, consider all matrices M with left/right singular subspaces equal to those of ΔW and singular values $\hat{\sigma}$. Writing $M = \tilde{U}\,\text{diag}(\hat{\sigma})\,\tilde{V}^\top$, unitary invariance of the Frobenius norm gives $\|\Delta W - M\|_F^2 = \sum_i (\tilde{\sigma}_i - \hat{\sigma}_i)^2$, which is minimized termwise and is independent of the shared subspaces. Hence $\Pi_{\hat{\sigma}}(\Delta W)$ is the closest matrix (in Frobenius norm) to ΔW with the prescribed spectrum—an Eckart–Young-style statement with the roles of subspaces and singular values reversed. In lineage-matched PEFT we set $\hat{\sigma}$ from a rescaled slice of the base layer's own spectrum, so the adapter is nudged toward the base's spectral *shape* while its subspaces still follow the task gradient. The projection is applied every τ steps and is a no-op when $\hat{\sigma} = \tilde{\sigma}$, recovering vanilla LoRA.

B Low-Rank Truncation Curves

Table 5 gives the full low-rank weight-truncation curves behind the $\rho^\star_{95}$ and $\Delta Q^{\text{lr}}_{25\%}$ columns of Table 3: for each 2D hidden weight we keep the given fraction of the singular spectrum and record the validation loss increase ΔQ. The Muon curves stay high at every operating point—the weights admit essentially no low-rank truncation—while the AdamW curves fall off quickly, and the iso-loss `adam_long` model behaves like the other AdamW checkpoints, confirming the effect is geometry- rather than loss-driven.

Because Muon's $\rho^\star_{95}$ pins at the ceiling 1.00, a coarse grid cannot say *how* incompressible it is; Table 6 resolves the near-full-rank regime, where the lineages separate *continuously* rather than saturating. Keeping 99% of the singular spectrum, the AdamW lineages have already returned to $\Delta Q \approx 0.00$, whereas Muon still pays $\Delta Q = 2.85$ nat—i.e., discarding even the bottom 1% of singular directions materially damages a Muon-trained model. The gap is monotone and large

Table 5: **Low-rank truncation curves.** ΔQ (validation loss increase) when every 2D hidden weight is truncated to the given kept-rank fraction; last column is $\rho^\star_{95}$ (weight-truncation kept-rank fraction, distinct from the LoRA-on-task rank $r^\star_{95}$ of Eq. 2).

Model	0.05	0.10	0.15	0.25	0.50	0.75	$\rho^\star_{95}$
ADAM s0	2.33	1.61	1.19	0.67	0.14	0.02	0.50
MUON s0	7.04	6.60	5.40	4.57	4.68	4.35	1.00
ADAM s1	3.14	2.23	1.72	1.12	0.27	0.04	0.75
MUON s1	7.43	6.32	5.87	4.85	4.69	4.02	1.00
ADAM_long	4.30	2.56	1.65	0.76	0.12	0.02	0.50

Table 6: **Near-full-rank low-rank truncation (fine grid).** ΔQ when every 2D hidden weight keeps the given high fraction of its singular spectrum. MUON stays large all the way to 0.99 (no low-rank structure to shed), while ADAM and the iso-loss control are already back to ≈ 0 by 0.90.

Model	0.80	0.85	0.90	0.95	0.99
ADAM s0	0.012	0.007	0.004	0.001	0.000
MUON s0	4.123	3.787	3.625	3.225	2.845
ADAM_long	0.010	0.006	0.003	0.001	0.000

at every operating point from 0.75 to 0.99, and the iso-loss `adam_long` model tracks the other ADAM checkpoints throughout, confirming the separation is geometry- rather than loss-driven on a continuous scale (addressing the concern that a saturated $\rho^\star_{95}$ hides the effect size).

LoRA-adaptation efficiency (full sweep). Table 7 gives the LoRA-on-task sweep behind the null of Section 5.3: each clean-room base is LoRA-adapted to a distinct target domain (TinyStories) at $r \in \{1, 2, 4, 8, 16, 32, 64\}$, and we report the smallest rank recovering 95% of full fine-tuning's loss reduction. Rank 1 already exceeds 100% recovery for both lineages, so $r^\star_{95}(\text{LoRA})=1$ throughout and the spectral fingerprint does not predict adaptation efficiency on this task.

C DOWNSTREAM PREDICTIVITY: FULL PROBE SET

The body claims exactly one downstream consequence of the fingerprint—the compression budget (H2). This appendix records *every* downstream property we probed, so that the positive claim is bracketed by a complete and honest account of what the fingerprint does and does *not* forecast. Table 10 summarizes the verdicts; the three independent probes that come back null (LoRA adaptation, linear mode connectivity, fine-tuning dynamics) are all reported here in full.

Rank-based compression (governed). Low-rank weight truncation and low-bit quantization are the properties the fingerprint sets: MUON-lineage weights are 5–7$\times$ less low-rank-truncatable and $\sim$29$\times$ less 4-bit-quantizable than ADAM-lineage weights, robustly to the iso-loss control (Table 3 in the main text; full curves Tables 5–6). This is the H2 result and the only downstream we claim the fingerprint governs.

Magnitude pruning (does not follow). At 50% magnitude sparsity the lineage ordering breaks: the iso-loss `adam_long` model prunes *worse* ($\Delta Q = +2.37$) than either MUON model ($+0.93$/$+1.04$; Table 3). This is expected—magnitude pruning depends on the weight-*magnitude* distribution, not the singular spectrum—so the fingerprint tracks rank-based compressibility but *not* magnitude sparsity.

LoRA-adaptation efficiency (null). The one genuinely *independent* observable in our formulation is the LoRA rank $r^\star_{95}$ (Eq. 2)—the smallest LoRA rank recovering 95% of full fine-tuning's loss reduction on an actual task, which requires retraining and cannot be read off the base spectrum.

Table 7: **LoRA-adaptation efficiency: a null** (clean-room 124M, TinyStories target). "Reduction" is the base→full-FT val-loss drop; rank-1 LoRA already recovers $>100\%$ of it for both lineages, so $r^{\star}_{95}(\text{LoRA})=1$ and the spectral fingerprint does not predict adaptation efficiency on this task.

Base	base val	full-FT val	reduction	LoRA r=1 recovers	$r^{\star}_{95}$(LoRA)
ADAM s0	4.516	2.778	1.738	104%	1
MUON s0	3.653	2.312	1.340	103%	1

Table 8: **Linear mode connectivity: a null.** Loss barrier (peak interpolated loss minus endpoint mean) when linearly interpolating the two cross-seed checkpoints of each lineage. The MUON−ADAM direction *flips* between scales, so the fingerprint does not predict mergeability.

Scale	ADAM barrier	MUON barrier	MUON vs. ADAM
124M	6.81	4.94	lower
350M	2.81	4.01	higher

Adapting each clean-room base to a distinct target domain (TinyStories) at $r \in \{1, 2, 4, 8, 16, 32, 64\}$, rank 1 already recovers full fine-tuning quality for both lineages (ADAM 104%, MUON 103%), so $r^{\star}_{95}(\text{LoRA})=1$ throughout (Table 7) and the fingerprint does *not* predict adaptation efficiency here. The caveat is that this adaptation is itself intrinsically low-rank (rank 1 suffices for either base), so a higher-rank task could still expose a difference.

Linear mode connectivity / mergeability (null). We test whether the fingerprint predicts how mergeable two independently trained checkpoints are by linearly interpolating the two cross-seed checkpoints of each lineage and recording the loss barrier—the peak interpolated loss minus the endpoint mean (Table 8). The direction is *inconsistent across scale*: at 350M the MUON barrier is *higher* than ADAM's (4.01 vs. 2.81), but at 124M it is *lower* (4.94 vs. 6.81). A property whose sign flips between scales cannot be a fingerprint-driven prediction, so mergeability is an honest null.

Fine-tuning dynamics (null, confound-controlled). Finally we fully fine-tune each 124M base on TinyStories and track both target adaptation (target-domain loss drop) and forgetting (source-domain loss rise; Table 9). In raw terms ADAM adapts 1.53 and forgets 1.05 while MUON adapts 1.20 and forgets 0.80—but these raw gaps track MUON's lower base loss (its curves start lower on *both* axes), a confound. The confound-controlled plasticity–stability ratio—forgetting per unit adaptation—is essentially *equal*: 0.66 (MUON) vs. 0.69 (ADAM). Once the base-loss confound is removed there is no fingerprint-driven difference in fine-tuning dynamics, so this too is a null.

Summary. Beyond rank-based compression, the fingerprint does *not* forecast the independent downstream properties we tested—magnitude pruning, LoRA-adaptation efficiency, mode connectivity, and fine-tuning dynamics. The body therefore claims only what the data support: a causal spectral fingerprint (H1) and the compression budget it sets (H2).

D CLEAN-ROOM PRETRAINING CONFIGURATION

E SCALE REPLICATION AT 350M

Table 12 gives the full paired fingerprint for a *second* clean-room pair at 354.8M parameters (24 layers, 16 heads, width 1024, tied embeddings), constructed exactly like the 124M pair (Section 4.1): the two runs share architecture, data, initialization, and step budget and differ *only* in the optimizer (AdamW vs. MUON-on-2D), so every paired per-layer difference is attributable to update geometry. Statistics are bucketed into attention and MLP with n=48 matched weights each; p is a one-sided paired Wilcoxon signed-rank test. Every axis moves in the *same direction as at* 124*M* (Table 1) and, for stable rank, with a *larger* ratio ($\sim$15$\times$ attn, $\sim$25$\times$ MLP, vs. $\sim$6–14$\times$ at 124M); moreover the attention concentration axes ($\tilde{H}_{\sigma}$, r_{eff}/k), only directional at 124M iso-step, reach significance here. The primary 350M pair above is *iso-step*, so the final validation losses differ (ADAM 5.514 vs. MUON 4.940). To remove this loss-progress confound *at* 350*M* we ran a dedicated 350M **iso-loss control**: ADAM trained for extra steps (16k) until its final loss reaches 4.859, which is *equal-to-or-lower* than MUON's 4.940—so at this matched loss ADAM is, if anything, further along

Table 9: **Fine-tuning dynamics: a confound-controlled null** (clean-room 124M, full fine-tune on TinyStories). "Adapt" is the target-domain loss drop, "forget" the source-domain loss rise. The raw gaps track MUON's lower base loss; the confound-controlled ratio forget/adapt is essentially equal, so the fingerprint drives no difference in the plasticity–stability tradeoff.

Base	adapt	forget	forget/adapt
ADAM	1.53	1.05	0.69
MUON	1.20	0.80	0.66

Table 10: **Predictivity of the fingerprint** (H2 and the independent probes): direction of the MUON effect, cross-seed / cross-scale consistency, survival of the iso-loss control, and verdict. Rank-based compression is governed by the fingerprint; the four independent downstream properties below it are not.

Observable	MUON dir.	consistent	iso-loss	verdict
$\rho^{\star}_{95}$ (weight low-rank)	harder	yes	yes	**supports**
$\Delta Q_{\text{lowrank}}$	harder	yes	yes	**supports**
ΔQ_{quant}	harder	yes	yes	**supports**
ΔQ_{prune}	mixed	no	no	not supported
$r^{\star}_{95}$ (LoRA-adapt)	none (rank 1)	yes	—	**null**
mode connectivity	flips w/ scale	no	—	**null**
fine-tuning dynamics	confounded	—	—	**null**

in training. The fingerprint nonetheless persists with essentially unchanged magnitude: against the *same* MUON checkpoint, this longer-trained ADAM still has paired per-layer stable rank of only 13.7 (attn) and 11.6 (MLP) versus MUON's 161.2 and 170.8 (both p=7.1e−15, one-sided paired Wilcoxon), a $\sim$12–15$\times$ gap; the concentration axes and tail move in the same direction and remain significant ($\tilde{H}_{\sigma}$ +0.033/+0.052, r_{eff}/k +0.182/+0.283, tail α +2.75/+4.38; all $p\leq$1.5e−12). The ADAM‡ column of Table 12 reports these iso-loss values alongside the iso-step ones. The 350M fingerprint is therefore *not* a loss artifact—exactly as at 124M (Table 2), where training ADAM to MUON's loss *grows* rather than shrinks the gap—so H1 is causal *and* iso-loss-robust at *both* scales. Like the 124M pair, these 350M checkpoints are undertrained research models, so the replication establishes scale-consistency of the fingerprint, not a compute-optimal claim; only an iso-FLOPs / compute-optimal control remains future work.

F COMPUTE BUDGET (ESTIMATES)

G DOWNSTREAM PROBE CONFIGURATION (SKELETON)

Every probe is a standard, unmodified method run identically across lineages so that any downstream difference is attributable to the base rather than to a tuned compressor. Table 14 lists the sweeps; exact task, split, and quality metric Q are pinned in the released experiment plan.

H STATISTICAL PROTOCOL (DETAILS)

For H1 the unit of analysis is a layer within a matched pair; we report a one-sided Wilcoxon signed-rank test per layer bucket with 10^4-resample bootstrap confidence intervals on the median lineage difference, correcting across buckets with Benjamini–Hochberg. For H2 the sample is checkpoints$\times$layers; we report Spearman rank correlations between each fingerprint statistic and each observable, and compare a fingerprint regressor to a lineage-blind baseline by leave-one-checkpoint-out cross-validated R^2. We treat the number of independent checkpoints, not layers, as the effective sample size when reporting significance, to avoid pseudo-replication from correlated within-model layers.

Table 11: Clean-room nanoGPT pretraining configuration (as run). Both lineages share every row except the optimizer; the 124M pair is the basis for all results in Section 5.

Field	Value
Architecture	GPT (nanoGPT), pre-LN, tied emb.
Parameters	124M (12 layers, 12 heads, width 768)
Corpus / token budget	OpenWebText, 150M tokens (fixed, shared)
Init seed	fixed, shared (seed 0; seed 1 replication)
Batch / steps	fixed, shared (iso-step)
Optimizer (varied)	AdamW (all params) vs. MUON
MUON scope	2D hidden matrices; AdamW on emb/LN/head
MUON msign iteration	Newton–Schulz
Controls	iso-step (primary), iso-loss (124M & 350M), 2 seeds
Scale replication	350M (354.8M) iso-step + iso-loss (App. E)
Future	iso-FLOPs / compute-optimal control

Table 12: **Spectral fingerprint at 350M** (clean-room 354.8M pair; n=48 matched weights/bucket). ADAM is the iso-step run; ADAM‡ is the **iso-loss control** (ADAM trained 16k steps to final loss 4.859, *equal-or-lower* than MUON's 4.940, vs. iso-step ADAM 5.514). Δ/p are the paired median MUON$-$ADAM difference and one-sided paired Wilcoxon test *for the iso-step pair*; the iso-loss gap MUON$-$ADAM‡ is also significant (stable rank p=7.1e$-$15 both buckets; all axes $p\leq$1.5e$-$12). Same direction as the 124M pair (Table 1), with a *larger* stable-rank ratio and both attention concentration axes now significant. $^{***}p<10^{-4}$.

Bucket	Statistic	ADAM	ADAM‡	MUON	Δ	p
Attn $Q/K/V/O$	stable rank	10.6	13.7	161.2	+153.8	7.1e$-$15***
	$\tilde{H}_\sigma$	0.965	0.945	0.971	+0.011	1.4e$-$5***
	r_{eff}/k	0.785	0.685	0.820	+0.067	6.5e$-$6***
	tail α	2.87	1.79	4.57	+1.60	3.8e$-$7***
MLP up/down	stable rank	6.7	11.6	170.8	+163.6	7.1e$-$15***
	$\tilde{H}_\sigma$	0.958	0.936	0.989	+0.032	7.1e$-$15***
	r_{eff}/k	0.748	0.640	0.928	+0.187	7.1e$-$15***
	tail α	1.94	1.51	5.94	+3.91	9.7e$-$13***

I FAILURE-MODE ANALYSIS

We enumerate ways each hypothesis could fail and how the design surfaces it, rather than hides it. **(i) Spectrum washout.** Weight decay, long training, and the final learning-rate decay phase reshape the spectrum and may erase lineage differences; the iso-loss/iso-FLOPs ablation and a decay-phase sweep isolate this. **(ii) Localized lineage.** Because MUON typically updates only 2D hidden matrices while embeddings/gains use ADAM, the fingerprint may be a per-bucket rather than a global property; we therefore never collapse to a single scalar and report per-bucket distributions. **(iii) Confounded open checkpoints.** Setting B differs in data, tokenizer, and step count, so we use it only as a generalization check on effects first established in the clean-room setting A, and report the confounds explicitly. **(iv) Adapter capacity dominates.** If the adapter rank r is large enough to absorb the base geometry, lineage-matched PEFT (H3) may not help even when H1/H2 hold; H3 is thus framed as exploratory and the paper is designed to stand on H1/H2. **(v) Small effective n.** With tens of layers per model, correlations are fragile; we use rank statistics, bootstrap intervals, and multiple-comparison correction, and require consistency across observables and a setting-B replication before crediting H2.

J LINEAGE-MATCHED PEFT SCHEMATIC

For completeness we give the lineage-matched-PEFT procedure (H3, Section 4.3) as Algorithm 1. H3 is an exploratory intervention—the paper's claims rest on H1/H2—and is not yet run, so we describe it in text and as a schematic rather than reporting results. Further planned material (per-layer fingerprint violin plots, an open-checkpoint Setting B replication table, and tail-window sensitivity curves for α) is future work; the 124M→350M scale replication is already measured in Appendix E.

Table 13: Single-GPU stages. Pretraining of the clean-room pairs dominates; all fingerprint and compression measurements on a 124M model are minutes-scale. Stages marked future are not yet run (Section 5.4).

Stage	Item	Status
E0	clean-room pretrain (124M, all configs)	run (single GPU)
E1	fingerprint measurement (all ckpts)	run (minutes)
E2	low-rank / prune / quant probes	run (minutes)
E2	LoRA-on-task rank sweep (TinyStories)	run (null; Table 7)
E2	merge-interference study	future
E3	lineage-matched PEFT	future
E0	350M scale replication (iso-step)	run (App. E)
E0	350M iso-loss control	run (App. E)
E0	iso-FLOPs / compute-optimal control	future

Table 14: Downstream probes and their sweeps (skeleton). The same grid is run on every lineage; numeric outcomes are reported in Table 3 (the lineage-matched-PEFT arm is future work).

Probe	Method(s)	Sweep
LoRA rank	LoRA (Hu et al., 2022)	$r \in \{1, 2, 4, 8, 16, 32, 64\}$
Pruning	SparseGPT, Wanda	$s \in \{0.25, 0.5, 0.6, 0.7\}$
Quantization	GPTQ, AWQ	$b \in \{2, 3, 4, 8\}$ bits
Merging	Task arithmetic, TIES	T task adapters
Lineage PEFT	Alg. 1	rank-/budget-matched

Algorithm 1 Lineage-matched PEFT (schematic; Section 4.3)

Input: base $\theta_{\mathcal{O}}$, lineage $\mathcal{O}$, rank r
Estimate fingerprint $\mathcal{F}(\theta_{\mathcal{O}})$ ▷ Def. 1–3
Init factors B, A with $\Delta W = BA$, $\mathrm{rank} = r$
for each step **do**
  $G \leftarrow \nabla_{(B,A)} \mathcal{L}$
  if $\mathcal{O} =$ MUON **then**
    step via $\mathrm{msign}(G)$ ▷ Eq. 5
  else
    step via adaptive (ADAM) update
  end if
  (optional) project ΔW to spectrum-matched target
end for
return adapted model