SAME FEATURES, DIFFERENT GEOMETRY: LOCALIZING MUON'S ADVANTAGE TO THE OV CIRCUIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Does the optimizer change *what* a language model learns, or mainly *how* it stores and reaches it? We study Muon and AdamW on a 124M GPT-2-class model under independently tuned recipes (3.292 vs. 3.350, a 0.058-nat gap with seed spreads $\leq$0.002). At matched loss, top-k sparse-autoencoder (SAE) features cross the optimizer boundary as well as they cross a random seed (0.599 vs. a 0.594 AdamW seed ceiling), while Muon is the more seed-canonical substrate (0.618), a gap already set one tenth into training. What changes is geometry: flatter spectra and higher stable rank. To test whether that signature is the cause, we reassign Muon to sub-circuits of a fixed architecture. Muon on the value circuit $W_V W_O$ alone recovers 86% of the loss gap and most of the reproducibility gain, while Muon on every hidden matrix except that circuit recovers none, and neither does Muon on the MLP blocks. Muon on the scoring circuit $W_Q W_K^\top$ raises its stable rank to near-full-Muon levels and recovers neither, and an AdamW control that directly promotes value-circuit rank gains nothing either, its loss degrading as that rank rises. At a fixed learning rate, loss instead falls monotonically with orthogonalization fidelity, reproducibility saturating earlier. Zero-shot HellaSwag tracks the loss split, and both gaps persist at 1.3B. The rank fingerprint that observational analyses highlight is therefore not a sufficient driver: it can be raised without the benefit. Value-pathway sufficiency confirms a published block-level ablation under a tuned baseline; the necessity arm, the rank control, and the fidelity sweep are new. Feature claims are for one SAE family at one or two layers up to 350M parameters, and the reproducibility gap narrows to the seed spread under a GeLU MLP.

1 INTRODUCTION

Muon (Jordan et al., 2024) replaces each momentum update by its nearest orthogonal matrix, yielding spectral-norm steepest descent (Bernstein & Newhouse, 2024a;b). It trains language models faster than AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) and now drives frontier-scale pretraining (Liu et al., 2025; Moonshot AI, 2025). It also leaves a conspicuous weight-space fingerprint: more uniform singular-value spectra and higher stable rank than Adam-trained models, in pretraining (Liu et al., 2025; Wang et al., 2025) and fine-tuning (Qu et al., 2026), resisting the rank-minimization drift of standard training (Yunis et al., 2024; Martin & Mahoney, 2021) and the low-rank-first learning of gradient descent (Dragutinović & Ranganath, 2026). Downstream, Muon yields higher-effective-rank, more robust representations (Ruan et al., 2026). Muon clearly changes weight geometry. The interpretive question is whether that geometry change reaches the *features* a model learns, and *where* the useful part of the optimizer acts inside a transformer.

We answer both with an intervention study in two acts, at a scale where surgical optimizer reassignment is cheap enough to run with seed replication.

Act 1 (the phenomenon). Compared at matched validation loss, Muon and AdamW reach the same top-k sparse-autoencoder (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2024) feature agreement under one operational protocol: SAE features matched by firing-pattern correlation. We claim only protocol-level non-inferiority to the AdamW seed ceiling. What changes is geometry and

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.

seed reproducibility: Muon produces hotter residual streams, broader circuit spectra, and a more seed-canonical solution, with the reproducibility gap set in the first tenth of training.

Act 2 (the intervention). Reassigning Muon to attention sub-circuits (Elhage et al., 2021) shows that Muon on the value circuit $W_V W_O$ is largely sufficient for the measured loss and reproducibility gains, while Muon on the scoring circuit $W_Q W_K^\top$ is not, neither is Muon on the MLP blocks, and Muon on everything except the value circuit recovers none of the loss gain, so the value pathway is necessary as well as sufficient. The sufficiency half replicates, under an independently tuned AdamW baseline, the value-pathway finding of a published block-level ablation (Wang et al., 2025); the necessity arm, the rank control, the fidelity dose-response, and the feature-level results are new. Rank controls show that endpoint stable rank is not sufficient. Varying only the number of Newton–Schulz iterations at a fixed learning rate shows that the benefit is monotone in orthogonalization fidelity, and that reproducibility saturates well before loss does.

The methodological point is that a frozen architecture plus surgical optimizer reassignment can separate a correlational weight signature from an intervention-supported driver. Observationally, Muon raises both scoring and value ranks; interventionally, only the value-side update recovers the benefit in this bench. This makes the negative results informative: if the scoring circuit's rank were the causal knob, muon-QK should help; if endpoint rank were sufficient, rank-promoted AdamW should help; if Muon on any large hidden matrix sufficed, muon-MLP should recover the loss gain. None occurs.

Contributions.

1. A matched-loss, seed-calibrated protocol for comparing sparse features across recipes, with seed pairs as the ceiling and the model seed as the inferential unit (Sections 3, 4).
2. Evidence that optimizer choice preserves top-k SAE firing-pattern content up to the AdamW seed ceiling while changing geometry, dictionary packaging, and seed reproducibility, with the reproducibility gap set by step 500 (Section 4).
3. A causal localization tested in both directions: Muon on everything except W_V, W_O, on MLP blocks, or on W_Q, W_K recovers none of the loss gain, Muon on W_V, W_O most of it; endpoint W_{OV} rank is not sufficient; HellaSwag follows the same split (Section 6).
4. A dose-response test of the update-geometry hypothesis at a fixed learning rate: loss is monotone in Newton–Schulz fidelity, reproducibility saturates earlier, and SOAP partly reproduces the signature (Section 5).
5. The boundaries of the invariance: it holds at three, six, and twelve heads with Muon the more reproducible optimizer throughout and narrows to the seed spread under a GeLU MLP, while a cross-layer control shows that the metric registers content change (Section 6).

2 Related work

Muon and implicit bias. Muon (Jordan et al., 2024) descends from the Shampoo/SOAP preconditioning line (Gupta et al., 2018; Vyas et al., 2024) and the optimizers-as-steepest-descent view (Bernstein & Newhouse, 2024b;a); the polar factor is approximated by Newton–Schulz iteration (Amsel et al., 2025). It runs at production scale (Liu et al., 2025; Moonshot AI, 2025). Theoretically, Muon implicitly constrains spectral norms (Chen et al., 2025) and removes the low-rank-first simplicity bias of gradient descent (Dragutinović & Ranganath, 2026); empirically, it resists rank-minimization drift (Yunis et al., 2024; Martin & Mahoney, 2021) and imprints uniform spectra on weights (Liu et al., 2025; Wang et al., 2025; Qu et al., 2026). Ruan et al. (2026) connect Muon to feature *quality* via robustness and transfer. We work one level down, at the identity of sparse features, and intervene on *which* matrices receive orthogonalized updates and *how faithfully*; SOAP is a further arm.

Optimizer–architecture codesign. Zuo et al. (2026) show that an attention variant can improve over softmax under Muon but not AdamW, with stable-rank analysis tracing the interaction to the scoring circuit. Our intervention result refines that picture on a softmax bench: the scoring circuit shows Muon's largest attention-circuit rank increase, yet orthogonalizing it alone recovers neither loss nor reproducibility; the useful interaction runs through the value/output pathway.

Component ablations of Muon. Wang et al. (2025) run the experiment closest to Act 2, a block-by-block Muon/Adam assignment on a 160M NanoGPT trained on FineWeb, and also find that

the value/output pathway helps, that W_Q, W_K contributes little, and that W_O matters more than W_V. They also report that Muon's spectra are more seed-stable than Adam's and that Muon makes W_Q, W_K more isotropic without that circuit driving the gain; Acts 1 and 2 test these at the feature level and by intervention. The two studies differ on the FFN: Muon on a single FFN matrix closes $\sim$60% of their ungated Adam-to-Muon gap ($\sim$27% gated), which Li et al. (2026) explain through associative memory, whereas our muon-MLP closes none and the complement arm shows the value pathway is necessary. The disagreement is real at the level of loss. Appendix Q lists the protocol differences; the largest are that their Adam baseline trails Muon by 0.36–0.37 nats against our 0.058 under an independently tuned AdamW, so moving any large block onto Muon looks helpful there, and that their hybrids inherit full Muon's learning rate untuned while ours match warmup and sweep the Muon-side rate. Whether the FFN contribution survives a tuned AdamW baseline is what our negative bears on but does not settle.

Feature universality and comparison. Independently trained networks learn overlapping features (Li et al., 2016; Olah et al., 2020), sharpened into representational convergence (Huh et al., 2024; Moschella et al., 2023; Long et al., 2026). Sparse autoencoders (Elhage et al., 2022; Huben et al., 2024; Bricken et al., 2023; Gao et al., 2024; Templeton et al., 2024) make this testable at the level of individual features. SAE features have been matched across architectures and model families (**??**), and top-k SAEs trained with different seeds recover different features (**?**), hence our instrument-noise ceiling. Representation-similarity measures such as SVCCA and CKA (centered kernel alignment) (Raghu et al., 2017; Kornblith et al., 2019) compare subspaces but conflate content with second-order geometry; we use CKA as a deliberately geometry-sensitive complement. Permutation symmetry and the failure of naive linear mode connectivity (Entezari et al., 2022; Ainsworth et al., 2023; Frankle et al., 2020) explain why decoder *directions* cannot be matched across models, motivating behavioral matching. Seed sensitivity is a documented tax (Dodge et al., 2020; D'Amour et al., 2020); our reproducibility result suggests the optimizer modulates it.

3 Setup

Optimizer. For a 2D parameter W with gradient G_t and momentum $M_t = \mu M_{t-1} + G_t$, let $M_t = U_t \Sigma_t V_t^\top$ be the reduced SVD. Muon applies

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

replacing the singular values of the momentum by ones. The direction $U_t V_t^\top = \arg\max_{\|O\|_2 \le 1} \langle M_t, O \rangle$ is steepest ascent of M_t under the spectral norm, so (1) is steepest descent under $\|\cdot\|_2$ (Bernstein & Newhouse, 2024a; Chen et al., 2025). In practice $U_t V_t^\top$ is approximated by a fixed number of Newton–Schulz iterations, avoiding an explicit SVD (Jordan et al., 2024; Amsel et al., 2025); the default is five, and Section 5 varies this count as a fidelity knob (Appendix C). Following Jordan et al. (2024), the bench orthogonalizes the Nesterov look-ahead of M_t rather than M_t itself.

Bench and matched-loss protocol. We use a 124M-parameter GPT-2-class trainer (12 layers, d=768, 6 heads, softmax attention, rotary embeddings, QK-norm, ReLU^2) (Radford et al., 2019; Vaswani et al., 2017; Su et al., 2024) in the tuned optimizer-comparison configuration of Jordan & contributors (2024): AdamW at learning rate 1.8×10^{-3} with 250-step warmup; Muon at 2×10^{-2} with no warmup; both with zero weight decay; embeddings and the language-model head on AdamW. Both recipes were independently tuned, addressing the under-tuned-baseline confound. Three seeds vary initialization only (five for the Act 1 check). Training is on FineWeb (Penedo et al., 2024) with identical data order across seeds, for 5,100 steps at 524,288 tokens/step ($\approx$ 2.67B tokens). Final validation losses: Muon 3.292$\pm$0.001, AdamW 3.350$\pm$0.002. Because Muon reaches any loss earlier, the primary SAE comparison uses *equal-loss* checkpoints (Muon step 4,375, loss 3.347; AdamW step 5,100, loss 3.350); Appendices F and O report the matched-step/token check.

Attention circuits. For residual stream $X \in \mathbb{R}^{T\times d}$, a head computes

$$A = \mathrm{softmax}\Big(\tfrac{1}{\sqrt{d_h}}\, X\, \underbrace{W_Q W_K^\top}_{W_{QK}}\, X^\top\Big), \qquad \mathrm{head}(X) = A\, X\, \underbrace{W_V W_O}_{W_{OV}}, \tag{2}$$

so scores depend on W_Q, W_K only through the scoring circuit W_{QK}, and the readout on W_V, W_O only through the value circuit W_{OV} (Elhage et al., 2021). We summarize each circuit by its stable

rank

$$\mathrm{srank}(M) = \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_i \sigma_i(M)^2}{\sigma_1(M)^2} \in [1, \mathrm{rank}(M)], \tag{3}$$

a smooth, scale-invariant proxy for how many directions carry weight mass (Martin & Mahoney, 2021).

Component-targeted optimizer. Partition the per-layer hidden parameters as $\theta_{\mathrm{QK}}=\{W_Q, W_K\}$, $\theta_{\mathrm{VO}}=\{W_V, W_O\}$, θ_{MLP}, and the rest. Arm **muon-QK** applies Muon to θ_{QK} and AdamW elsewhere; **muon-VO** applies Muon to θ_{VO}; **muon-attn** to both attention circuits; **muon-MLP** to MLP blocks; **muon-V** and **muon-O** to one value-circuit projection each. Every hybrid runs AdamW's 250-step warmup on the shared schedule, so its Muon-side tensors warm up too, unlike pure Muon (the hybrids are far worse without it; Appendix K). Embeddings and head stay on AdamW. The hybrids are deliberately not retuned to their own Pareto front: every reported hybrid loss uses Muon's tuned rate (0.02) on the Muon side and the tuned AdamW recipe on the AdamW side, with no joint retune; separate seed-0 sweeps of the Muon-side rate for muon-VO, muon-QK, and muon-MLP over an $8\times$ range ($32\times$ for muon-QK; Appendix K) check that none of them is reordered at another rate, so hybrid losses are conservative and comparisons are hybrid-versus-hybrid. The first arms test *sufficiency* of Muon on a subset; the complementary arm **muon-xVO**, Muon on every hidden matrix except W_V, W_O with those two on AdamW, is the direct necessity test.

Update-geometry controls. Three further arms vary the update rather than its location: Muon with 1, 2, 3, or 5 Newton–Schulz iterations (everything else fixed); SOAP (Vyas et al., 2024), a matrix-aware preconditioner outside the Muon family, in its tuned configuration from the same optimizer-comparison record; and AdamW with an explicit rank-promoting regularizer on W_{OV} (Section 6).

Variants. We replicate the pure-optimizer comparison at 30M, 350M, and 1.3B parameters (24 layers, d=2048, 16 heads; same token budget and learning rates), with a GeLU MLP, with QK-norm removed, on a second hardware vendor (AMD MI355X, same code, data, and seeds), and at 3 and 12 attention heads with d=768 fixed so that one SAE recipe applies across head counts.

Sparse-autoencoder features. For a layer-ℓ activation $x \in \mathbb{R}^d$, a top-k SAE (Gao et al., 2024) encodes and reconstructs

$$z = \mathrm{TopK}_k\big(W_{\mathrm{enc}}(x - b_{\mathrm{dec}}) + b_{\mathrm{enc}}\big), \qquad \hat{x} = b_{\mathrm{dec}} + W_{\mathrm{dec}} z, \tag{4}$$

keeping the k largest pre-activations. We use 12,288 features, k=32, trained on 50M tokens of layer-6 residual-stream activations, with an identical recipe across models so the model is the only variable. Results are robust to $k \in \{16, 64\}$ (Appendix H).

Cross-model matching and inference. Decoder *directions* are not comparable across independently trained models: even same-recipe seed pairs share no aligned decoder vectors (mean best cosine 0.14), as expected from permutation symmetry (Entezari et al., 2022; Ainsworth et al., 2023). We match *behaviorally*. Let $z_i^A \in \mathbb{R}^T$ be feature i's firing pattern over T=10^6 shared tokens; feature i of A scores

$$m_i^{A\to B} = \max_j \rho\big(z_i^A, z_j^B\big), \tag{5}$$

its best Pearson correlation in B, and a pair scores $s(A,B) = |\mathcal{L}_A|^{-1} \sum_{i\in\mathcal{L}_A} m_i^{A\to B}$ over A's live features, symmetrized over the two directions. Within-optimizer seed pairs bound attainable agreement. CKA is used as a content-blind, geometry-sensitive control (Kornblith et al., 2019). Because pairwise comparisons share models, the model seed, not the feature or the pair, is the inferential unit: intervals are clustered by model, and significance uses an exact permutation over model labels (Appendix N). An instrument-noise control bounds the ceiling: two SAEs trained with different SAE seeds on the *same* model agree at 0.719 (Muon) and 0.769 (AdamW), so much of the distance from perfect matching is SAE-fit noise rather than seed variation.

4 Act 1: The phenomenon

Top-k SAE feature agreement is seed-ceiling limited; geometry is not. Under the fixed SAE/matching protocol, crossing the optimizer boundary costs no more agreement than crossing the random seed: Muon-to-AdamW agreement (0.599) meets AdamW-to-AdamW (0.594), while

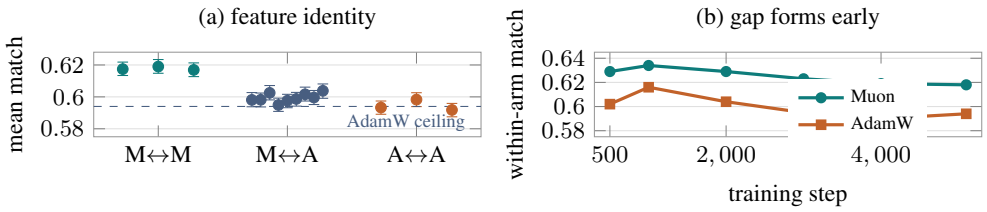

Figure 1: **Act 1: top-k SAE feature agreement is seed-ceiling limited, while reproducibility is higher and set early.** (a) Layer-6 firing-pattern match (Eq. 5) for every model pair at matched loss: cross-optimizer agreement meets the within-AdamW seed ceiling (dashed), while within-Muon is higher. Intervals are clustered by model seed. (b) Within-arm match along training (matched-step checkpoints; final point at matched loss): the Muon advantage is formed by step 500, and a switch run (500 Muon steps, then AdamW) retains it (0.629).

Table 1: Mean feature match (Eq. 5) averaged over seed pairs. Cross-optimizer agreement sits at the within-AdamW ceiling; within-Muon exceeds both. The 350M SAE reads layer 12 of 24, the same relative depth as layer 6 of 12.

Setting	within-Muon	cross-optimizer	within-AdamW
124M, layer 6, matched loss (primary)	0.618	0.599	0.594
124M, layer 6, matched step and tokens	0.619	0.592	0.592
124M, layer 9	0.610	0.586	0.585
30M, layer 3	0.582	0.551	0.544
350M, layer 12	0.630	0.589	0.592

within-Muon (0.618) sits above both (Figure 1a, Table 1; examples in Appendix G). All three within-Muon pairs exceed all twelve other pairs; the exact model-level permutation test over $\binom{6}{3}=20$ relabelings gives $p=0.05$, the smallest value three seeds per arm can attain (Appendix N). Five seeds per arm give 0.619 vs. 0.591. Normalizing by each arm's instrument ceiling widens the gap (0.86 vs. 0.77; Appendix E). The within-Muon $>$ within-AdamW ordering holds at layer 9, at matched tokens rather than matched loss, at 30M, and at 350M, with cross-optimizer agreement at the within-AdamW level in each case (0.589 vs. 0.592 at 350M; Table 1); under a GeLU MLP (three seeds) the ordering is preserved but the gap narrows to the seed spread (0.654 vs. 0.647), a boundary (Table 4).

What changes is the geometry. Muon's weight spectra are wider in stable rank (2.4× on the value circuit, 3.2× on the scoring circuit, 5.3× on MLP-in; Table 3, Appendix F); at matched loss its residual stream is 5–7× hotter, and its dictionary carries more dead features (6–8% vs. 0.2–1.6%) and more crowded ones, at tied reconstruction error (Appendix J). Geometry-sensitive CKA dissociates the two, falling *below* both within-optimizer levels (0.54) while firing-pattern match sees through to shared content.

Muon is more reproducible, and commits early. Muon seed pairs agree more than AdamW's, with non-overlapping intervals. This canonicalization is fully formed by step 500 (Figure 1b). A switch experiment makes the claim causal: 500 Muon steps followed by 4,600 AdamW steps retain reproducibility 0.629, far above AdamW (0.594) and nominally above full Muon (0.618; not under live-set matching, Section 6), at a loss cost (3.390).

The hot residual stream is a different knob. Muon with decoupled weight decay 0.1 eliminates residual-stream inflation (layer-6 norm ∼65 vs. AdamW's ∼700) at a small loss cost (3.297–3.300, 9–14% of the edge) with reproducibility unchanged (0.618), so scale, norm, and reproducibility are separable. Matched-token analysis, activation renormalization, and a relative-noise test leave the split intact (Appendix F).

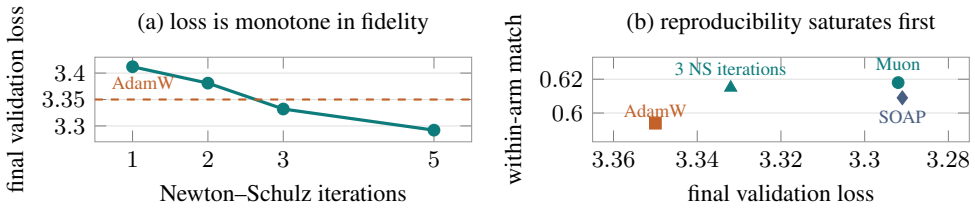

Figure 2: **At a fixed learning rate, loss improves with orthogonalization fidelity, and reproducibility needs less of it than loss.** *(a)* Varying only the Newton–Schulz iteration count (124M, Muon learning rate fixed; one seed at 1 and 2 iterations, three at 3 and 5): one iteration is worse than AdamW, three recover about a third of the edge, five (the default) all of it. *(b)* Loss against within-arm reproducibility (SOAP indicative).

5 How much orthogonalization is needed?

Act 1 attributes the reproducibility gap to Muon's orthogonalized updates. Two controls test that attribution by changing the update itself while leaving the architecture, data order, learning rates, and tensor assignment untouched (Table 2).

Loss is monotone in orthogonalization fidelity; reproducibility saturates earlier. Newton–Schulz approximation quality grows with iteration count. One iteration, a crude non-orthogonal rescaling of the momentum, ends *worse* than AdamW (3.412); two give 3.381; three give 3.332 (three seeds, $\sim$31% of the edge); five, the default, give 3.292 (Figure 2a). Fewer iterations also shrink the update's norm (Appendix C), so at a fixed learning rate this curve varies norm and fidelity together; a norm-matched sweep would separate them. Reproducibility does not track loss along this curve: at three iterations within-arm match is already 0.615 on the plain metric, indistinguishable from full Muon's 0.618 and well above AdamW's 0.594 (Figure 2b). Canonicalization therefore needs less fidelity than the loss edge, a second dissociation, consistent with early commitment.

SOAP partly reproduces the signature. SOAP, which preconditions with second-moment factors rather than orthogonalizing, reaches Muon's loss (3.291 vs. 3.292) and a within-arm match of 0.609, between AdamW and Muon; its features match Muon's (0.602) more closely than AdamW's (0.578). The signature is therefore not unique to Muon's update rule; we treat the SOAP numbers as indicative because the fixed SAE recipe leaves 21–54% of features dead on SOAP models.

Hardware and scale. The loss gap is not a vendor-specific numerical artifact: rerunning the pure arms on AMD MI355X with the same code, data, and seeds gives AdamW 3.349 and Muon 3.292 (three seeds each), matching the NVIDIA values to within 0.001. At 1.3B parameters, with the same tokens and learning rates, the gap persists: Muon 3.034/3.035 against AdamW 3.110/3.108 (Table 2), a 0.075-nat edge (seed spreads $\leq$0.002) against 0.058 at 124M and 0.060 at 350M, and Muon first evaluates below AdamW's final loss at 83% of training (86% at 124M). This is not a scaling law: the run sees $\sim$2 tokens per parameter at untuned 124M rates, and its gap is still narrowing at the end (0.096 at step 4,000; Appendix S).

6 Act 2: The mechanism

Muon on the value circuit is largely sufficient. Optimizer-targeting the two attention circuits separates the effect (Figure 3, Table 3). Orthogonalizing the value circuit alone (muon-VO) reaches loss 3.301, recovering about 86% of the 0.058-nat AdamW-to-Muon gap, with reproducibility 0.632, above full Muon's 0.618 (a pattern muon-attn and the switch arm share, taken up below). Orthogonalizing the scoring circuit alone (muon-QK) reaches 3.360 and 0.588: no better than AdamW on either loss or reproducibility. Five seeds per arm change neither result (Appendix S). The result is not "Muon anywhere in attention helps"; necessity is tested below.

The loss split persists at **350M**: with three seeds per arm, muon-VO reaches 3.140 (full Muon 3.133) while muon-QK stalls at 3.205 (AdamW 3.193). Plain layer-12 reproducibility splits the same way (muon-VO 0.638, muon-QK 0.590; full Muon 0.630, AdamW 0.592), but on the 10,240 most frequent live features muon-VO falls to the AdamW level (0.604 vs. 0.608; full Muon 0.637), so at

Table 2: Update-geometry controls at 124M plus hardware and scale replications of the pure arms. Reproducibility is the within-arm layer-6 match over three seeds; · = not measured.

Arm	Seeds	Loss	Recovers	Reproducibility
AdamW (reference)	3	3.350	0%	0.594
Muon, 1 NS iteration	1	3.412	<0	·
Muon, 2 NS iterations	1	3.381	<0	·
Muon, 3 NS iterations	3	3.332	∼31%	0.615
Muon, 5 NS iterations (default)	3	3.292	100%	0.618
SOAP	3	3.291	∼102%	0.609†
AdamW, AMD MI355X	3	3.349		·
Muon, AMD MI355X	3	3.292		·
AdamW, 1.3B	2	3.109		·
Muon, 1.3B	2	3.034		·

†Indicative: 21–54% dead SAE features on SOAP models.

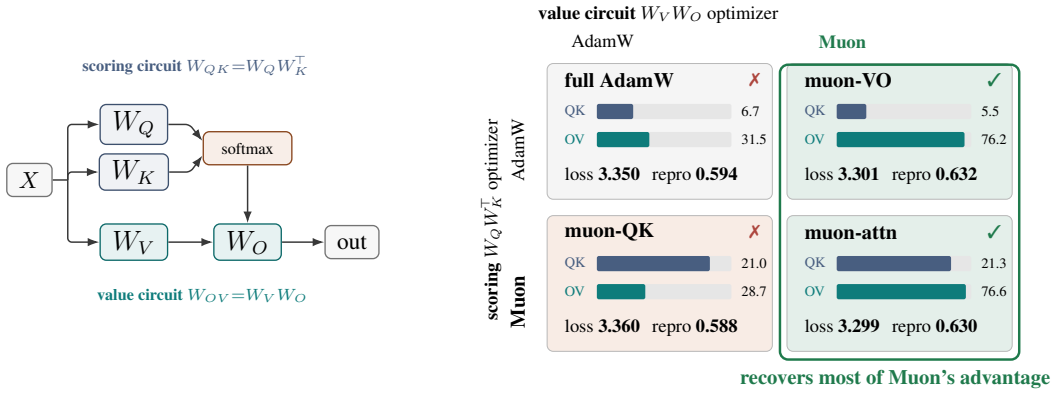

Figure 3: **Muon on the value circuit is largely sufficient; scoring rank is a passenger.** Left: a head factors into scoring and value circuits (Eq. 2). Right: Muon on one circuit at a time (124M, three seeds, matched warmup); cells give scoring/value stable ranks, loss, and reproducibility.

Table 3: Hybrid arms at 124M (three seeds, matched warmup). Only arms that orthogonalize W_V, W_O recover the loss gap. Stable ranks average over heads and layers.

Muon applied to	QK srank	OV srank	Loss	Repro.
nothing (full AdamW)	6.7	31.5	3.350	0.594
W_Q, W_K (muon-QK)	21.0	28.7	3.360	0.588
MLP blocks (muon-MLP)	·	·	3.368	0.603
all hidden except W_V, W_O (muon-xVO)	·	·	3.369	0.606
W_V, W_O (muon-VO)	5.5	76.2	**3.301**	**0.632**
all attention (muon-attn)	21.3	76.6	3.299	0.630
all hidden parameters (full Muon)	21.3	76.6	3.292	0.618

this scale only the loss half is robust to the instrument (Appendix K). It survives removing QK-norm, which might have been thought to pre-empt muon-QK: without it, muon-VO recovers ∼83% of a 0.041-nat edge and muon-QK none (Appendix K). It also survives the per-arm learning-rate sweeps (seed 0): muon-VO stays near Muon and muon-QK stays within the AdamW seed range.

The scoring circuit's high rank is a passenger, and endpoint rank is not a sufficient knob. Scoring rank is Muon's largest attention-circuit rank change and, between the pure optimizers, a clean correlate of loss, yet muon-QK raises $\mathrm{srank}(W_{QK})$ to near full-Muon levels while buying neither loss nor reproducibility, and muon-VO captures most of the loss advantage at an AdamW-

level scoring rank (Figure 3); on this softmax bench the correlation behind scoring-circuit-centered codesign accounts (Zuo et al., 2026) is confounded, because scoring and value ranks rise together under full-attention Muon. The value-side rank is not the knob either: AdamW with an explicit W_{OV} rank-promoting regularizer (a per-head penalty on $\sigma_1(W_V W_O)$, the value circuit of Eq. 2) raises value-circuit stable rank monotonically with the penalty weight, up to full Muon's level on the sweep's own rank instrument, while loss stays at the AdamW level for weak regularization and degrades to 3.549 at Muon-level rank (Appendix K). Elevated OV rank accompanies successful Muon runs but is not a sufficient cause; the fidelity dose-response (Section 5) points to the update itself.

Within the value circuit, the output write dominates. Splitting W_V and W_O one projection at a time, muon-O recovers $\sim$62% of Muon's edge and muon-V $\sim$36%; the two are sub-additive (36%+62% exceeds the joint $\sim$86% of muon-VO), so both projections contribute, with the output write dominant and part of the benefit shared (Appendix L). The passenger pattern recurs: muon-V ends at *higher* OV stable rank (53.1) than muon-O (47.5) yet recovers *less* loss.

Muon on MLP blocks does not recover the advantage. Under the same protocol, Muon on MLP blocks with AdamW elsewhere reaches 3.368 (three seeds), slightly worse than AdamW, with reproducibility 0.603: none of the loss gain and only part of the reproducibility gap. A learning-rate sweep over the same $8\times$ range does not rescue it (3.357–3.418; Appendix K), although its best point is the lowest rate tried; within the swept range this rules out a "Muon on any large hidden matrix helps" reading of Act 2.

The value pathway is necessary as well as sufficient. The complement arm muon-xVO applies Muon to every hidden matrix except W_V, W_O (71M of 85M hidden parameters), under the same matched-warmup protocol. It reaches mean loss 3.369 (three seeds), no better than AdamW (3.350) and indistinguishable from muon-MLP: orthogonalizing everything else recovers none of the loss edge without the value pathway. Its within-arm reproducibility is 0.606, about half of the gap between AdamW (0.594) and full Muon (0.618), whereas muon-VO alone recovers all of it (0.632). With muon-VO this is a double dissociation for the loss advantage; reproducibility depends mostly, but not only, on the same pathway.

Partial Muon looks more reproducible than full Muon; the excess is an instrument effect. On the plain within-arm match, muon-VO (0.632), muon-attn (0.630), and the 500-step switch (0.629) all score above full Muon (0.618), while the arms that keep Muon on the MLP blocks for the whole run score below it (muon-xVO 0.606, muon-MLP 0.603). The match averages over live features, and full-Muon dictionaries have 6–8% dead features against none for muon-VO, so the arms are scored on different live sets. Scoring each pair on its 10,240 most frequent live features (below every arm's live count; re-trained dictionaries reproduce the published means to within 0.001, Appendix F) removes the excess: full Muon 0.627, switch 0.619, muon-VO 0.618, muon-attn 0.616, muon-MLP 0.609, muon-xVO 0.608, AdamW 0.588 (a mid-density band agrees; Table 9). The dead and rare tail, not de-canonicalization by Muon on the MLP, produces the inversion; Muon-family $>$ AdamW and value-pathway arms $>$ MLP-only arms survive every cut.

The split transfers downstream. Zero-shot HellaSwag (Zellers et al., 2019) on the final checkpoints of the Table 3 hybrids and the pure arms' MI355X replicate (acc_norm, 10,042 items) follows the loss ordering: Muon 0.302 and muon-VO 0.299 against AdamW 0.288, muon-QK 0.288, and muon-xVO 0.289, with every Muon-side model above every AdamW-side model (18 models, exact rank test $p{=}2{\times}10^{-5}$; Table 8). The edge is 1.4 points at 124M and 4.0 at 1.3B (0.387 vs. 0.347); within the hybrids the value pathway carries the downstream edge as it carries the loss edge.

The invariance holds across head counts; depth moves the metric. Varying head count at fixed width (Table 4) reproduces both halves of Act 1: cross-optimizer agreement sits at the within-AdamW ceiling at three heads (0.595 vs. 0.594) and at twelve (0.577 vs. 0.573), and Muon stays the more reproducible optimizer (0.625 and 0.611). Retuning both optimizers at three heads (seed-0 sweeps, then three seeds at Muon 0.01 and AdamW $1.2{\times}10^{-3}$) leaves the loss gap at 0.058 nats and the pattern intact (within-Muon 0.632, within-AdamW 0.610, cross-optimizer 0.607). Depth does: layer-6 features of one seed match the layer-3 and layer-9 features of another at 0.590 and 0.570 for Muon and 0.565 and 0.552 for AdamW, against same-layer pairs at 0.618 and 0.590, a drop five to ten times the distance between cross-optimizer agreement and the AdamW ceiling (Appendix I). The boundary we find lies in the MLP: with a GeLU activation the reproducibility gap narrows to the seed spread (0.654 vs. 0.647).

Table 4: Architecture controls at 124M (d=768, layer-6 SAEs): head count (Muon models scored at their first checkpoint below loss 3.35, AdamW models at the final step; six-head learning rates unless retuned) and a GeLU MLP at six heads (three seeds; cross-optimizer agreement not measured). Invariance holds when cross-optimizer agreement reaches the within-AdamW ceiling.

Variant	Muon loss	AdamW loss	within-Muon	within-AdamW	cross-opt.	invariance
3 heads	3.305	3.361	0.625	0.594	0.595	holds
3 heads, retuned	3.301	3.360	0.632	0.610	0.607	holds
6 heads (bench)	3.292	3.350	0.618	0.594	0.599	holds
12 heads	3.300	3.358	0.611	0.573	0.577	holds
6 heads, GeLU	3.331	3.373	0.654	0.647	·	gap = seed spread

7 Discussion

Why the value pathway. A mechanism hypothesis consistent with the passenger result: $W_{OV}{=}W_V W_O$ is each head's linear readout from attended positions into the residual stream. Coordinate-wise AdamW may let this readout drift toward a lower-rank, seed-specific basis (Yunis et al., 2024), whereas Muon spreads updates across the momentum's singular directions (Bernstein & Newhouse, 2024a). This would explain why value-circuit rank moves with successful Muon runs while rank promotion alone fails, why W_O carries more of the edge than W_V, and why reproducibility saturates at lower fidelity than loss: an early, roughly isotropic update fixes the basis, while most of the loss edge needs a nearly orthogonal update. It remains a hypothesis: the necessity arm shows the pathway is required without saying why.

For codesign and for interpretability. The useful Muon–attention interaction in this bench runs through the value/output pathway, which refines scoring-circuit-centered accounts (Zuo et al., 2026); the selective recipe is slower than either pure optimizer (106 vs. 91–92 ms/step; Appendix M), so its value is diagnostic. Stock SAE recipes meet more dead features on Muon models and fail on SOAP at the same budget, so Adam-era budgets need recalibration (Appendix R).

Limits. Most causal runs are 124M with three seeds on one data distribution and one architecture family, not a scaling law, although the pure-optimizer loss and HellaSwag gaps replicate at 1.3B (two seeds, $\sim$2 tokens per parameter, untuned rates). The complement arm shares the hybrid protocol's limits (matched warmup, no joint retune, three seeds). The reproducibility signature holds at three, six, and twelve heads but narrows to the seed spread with GeLU (Table 4). Indicative cells (one seed or a limited instrument): Newton–Schulz at one and two iterations; SOAP (21–54% dead features); muon-V and muon-O (loss only); the 350M hybrid reproducibility scores (cut-dependent); the twelve-head variant (six-head learning rates); QK-norm removal (two seeds). Feature agreement is for one SAE family (top-k) at one or two layers and up to 350M parameters (at 1.3B the fixed recipe leaves 69–76% of features dead for both optimizers), robust to k, with no filtering of generic or positional features.

AI use statement

In this work, we used generative AI tools to help design and prioritize control experiments and to give feedback on the methodology, to recompute summary statistics from training logs and matching outputs, and to interpret and cross-check results against those logs. We have not used generative AI tools to propose or refine the scientific hypotheses, to develop conceptual frameworks, to generate synthetic data, to clean or reformat datasets, to formulate mathematical claims, supply ingredients of their proofs, or write those proofs, or to carry out the qualitative feature analysis of Appendix G, and we did not use them to implement the training and analysis code beyond small, author-reviewed edits; translation is not applicable to this work. Additionally, we used generative AI tools to edit and draft parts of the manuscript, to identify related literature, and to check references. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

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

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

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

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, https://transformer-circuits.pub/2023/monosemantic-features, 2023.

Lizhang Chen, Jonathan Liu, Yifan Lu, et al. Muon optimizes under spectral norm constraints, 2025. arXiv:2506.15054.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning, 2020. arXiv:2011.03395.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. arXiv:2002.06305.

Sara Dragutinović and Rajesh Ranganath. To use or not to use Muon: How simplicity bias in optimizers matters, 2026. arXiv:2603.00742.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, https://transformer-circuits.pub/2021/framework, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition, 2022. arXiv:2209.10652.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, et al. Scaling and evaluating sparse autoencoders, 2024. arXiv:2406.04093.

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

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

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Keller Jordan and contributors. modded-nanogpt: The NanoGPT speedrun and the October 2024 optimizer comparison record. GitHub repository, https://github.com/KellerJordan/modded-nanogpt, 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. Blog post, https://kellerjordan.github.io/posts/muon/, 2024.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws, 2026. arXiv:2602.05725.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations*, 2016.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, et al. Muon is scalable for LLM training, 2025. arXiv:2502.16982.

Junlin Long, Zeyu Zhang, Xu Deng, Yiran Wang, Yue Yang, Luke Borgnolo, Maxwell Twelftree, and Yang Zhao. PlatonicNav: Unveiling semantic correspondence in navigation with Platonic topological maps, 2026. arXiv:2606.01788.

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

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

Moonshot AI. Kimi K2: Open agentic intelligence, 2025. arXiv:2507.20534.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, et al. The FineWeb datasets: Decanting the web for the finest text data at scale, 2024. arXiv:2406.17557.

Xingyu Qu, Peigeng Huang, and Samuel Horvath. Can Muon fine-tune Adam-pretrained models?, 2026. arXiv:2605.10468.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than Adam, 2026. arXiv:2606.09658.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, https://transformer-circuits.pub/2024/scaling-monosemanticity, 2024.

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*, 2017.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, et al. SOAP: Improving and stabilizing Shampoo using Adam, 2024. arXiv:2409.11321.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning, 2025. arXiv:2509.26030.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, et al. Approaching deep learning through the spectral dynamics of weights, 2024. arXiv:2408.11804.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Yifei Zuo, Dhruv Pai, Zhichen Zeng, Alec Dewulf, Shuming Hu, and Zhaoran Wang. Parallax: Parameterized local linear attention for language modeling, 2026. arXiv:2605.29157.

Appendix overview. Appendices A–E give the definitions, derivations, and statistical arguments the main text relies on: the logic of the optimizer-reassignment arms and their parameter accounting (A); the proof that Muon's update is spectral-norm steepest descent and has maximal stable rank (B); what the Newton–Schulz iteration count controls, with explicit per-iteration bounds on the singular values (C); the stable-rank facts used in Act 2 (D); and the invariance and exactness properties of the matching metric and the permutation test (E). The remaining appendices report controls, robustness checks, compute, and seed-level numbers.

A Notation, arms, and the logic of the intervention

Hidden matrices and arms. Let $\mathcal{H}$ be the set of hidden weight matrices of the 124M model: in each of the L=12 layers, $W_Q, W_K, W_V, W_O \in \mathbb{R}^{768\times768}$ and the MLP matrices $W_{\text{in}} \in \mathbb{R}^{3072\times768}$, $W_{\text{out}} \in \mathbb{R}^{768\times3072}$. Embeddings and the language-model head are never in $\mathcal{H}$ and always use AdamW. An *arm* is a subset $S \subseteq \mathcal{H}$: matrices in S are updated by Muon (Eq. 1) and matrices in $\mathcal{H} \setminus S$ by AdamW, with AdamW's 250-step warmup applied to the whole run whenever $S \neq \mathcal{H}$ (Appendix K shows the hybrids are far worse without it). Table 5 lists every arm in the paper with its parameter count.

Table 5: The optimizer-reassignment arms. $|S|$ counts the parameters updated by Muon; the hidden total is 84,934,656. Recovery ρ is defined in Eq. 6; reproducibility r is the within-arm layer-6 match. Losses and r are three-seed means at 124M.

Arm	S (Muon side)	$\lvert S\rvert$	Loss	ρ	r
AdamW	$\emptyset$	0	3.350	0	0.594
muon-QK	$\{W_Q, W_K\}$	14,155,776	3.360	−0.17	0.588
muon-V	$\{W_V\}$	7,077,888	3.330	0.36	·
muon-O	$\{W_O\}$	7,077,888	3.314	0.62	·
muon-VO	$\{W_V, W_O\}$	14,155,776	3.301	0.85	0.632
muon-attn	$\{W_Q, W_K, W_V, W_O\}$	28,311,552	3.299	0.88	0.630
muon-MLP	$\{W_{\text{in}}, W_{\text{out}}\}$	56,623,104	3.368	−0.31	0.603
muon-xVO	$\mathcal{H} \setminus \{W_V, W_O\}$	70,778,880	3.369	−0.33	0.606
Muon	$\mathcal{H}$	84,934,656	3.292	1	0.618

Recovery. Writing $f(S)$ for the final validation loss of arm S, the normalized recovery is

$$\rho(S) = \frac{f(\emptyset) - f(S)}{f(\emptyset) - f(\mathcal{H})}, \tag{6}$$

so $\rho(\emptyset) = 0$ and $\rho(\mathcal{H}) = 1$; the denominator is 0.058 nats at 124M. We call S *sufficient* at tolerance ε if $\rho(S) \geq 1 - \varepsilon$ and *necessary* if $\rho(\mathcal{H} \setminus S) \leq \varepsilon$. With ε=0.15, the arms of Table 5 give: $\{W_V, W_O\}$ is sufficient (ρ=0.86) and necessary ($\rho(\mathcal{H} \setminus \{W_V, W_O\})$= − 0.33); the MLP matrices are neither sufficient (ρ= − 0.31) nor necessary, because their complement $\mathcal{H} \setminus \{W_{\text{in}}, W_{\text{out}}\}$ is exactly the muon-attn arm (ρ=0.88); $\{W_Q, W_K\}$ is not sufficient (ρ= − 0.17), and its necessity was not tested directly (no arm applies Muon to $\{W_V, W_O\}$ and the MLP jointly), although muon-VO's 0.86 leaves at most 0.14 for it. This is the double dissociation stated in Section 6. Recovery is sub-additive

on the value circuit, $\rho(\{W_V\}) + \rho(\{W_O\}) = 0.98 > \rho(\{W_V, W_O\}) = 0.86$, computed from the unrounded seed means, so part of the benefit of orthogonalizing W_V and W_O is shared rather than independent. Three arms ($\rho < 0$) are slightly worse than AdamW: orthogonalized updates on non-value matrices, without the value pathway, cost a little rather than being neutral.

Component-targeted optimizer. Algorithm 1 states the training loop used for every arm; the only difference between arms is the partition of $\mathcal{H}$. Muon uses momentum μ=0.95, learning rate 0.02 (swept in Appendix K), five Newton–Schulz iterations (varied in Section 5), and no weight decay unless stated; the AdamW side uses the bench's tuned hidden-layer settings, learning rate 1.8×10^{-3}, β=(0.9, 0.95), and zero weight decay. The learning-rate schedule (linear warmup, constant, linear cooldown) multiplies both optimizers' rates.

Algorithm 1 One training step of a component-targeted arm S

Input: arm $S \subseteq \mathcal{H}$; batch (x, y) of 512×1024 tokens; step t; schedule $\lambda_t \in [0, 1]$
$g \leftarrow \nabla_\theta \mathcal{L}(x, y; \theta)$ (data-parallel over 8 GPUs, gradients all-reduced)
for each matrix $W \in S$ **do**
  $M_W \leftarrow \mu M_W + g_W$
  $O_W \leftarrow \mathrm{NewtonSchulz}_k(M_W / \|M_W\|_F)$ (Appendix C; each rank owns a slice of S, results all-gathered)
  $W \leftarrow W - \lambda_t \, \eta_{\mathrm{Muon}} \, O_W$
end for
update every $W \in \mathcal{H} \setminus S$, the embeddings, and the head with AdamW at rate $\lambda_t \, \eta_{\mathrm{AdamW}}$

B Muon's update is spectral-norm steepest descent

Let $M \in \mathbb{R}^{m \times n}$ have reduced SVD $M = U \Sigma V^\top$ with $\Sigma = \mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_r > 0)$, $r = \mathrm{rank}(M)$. Write $\|\cdot\|_2$ for the spectral norm (largest singular value) and $\|M\|_* = \sum_i \sigma_i(M)$ for the nuclear norm.

Lemma 1. $\max_{\|O\|_2 \leq 1} \langle M, O \rangle = \|M\|_*$, *and the maximum is attained at* $O^\star = UV^\top$. *If* $r = \min(m, n)$ *the maximizer is unique.*

Proof. By von Neumann's trace inequality, for any O with singular values $\tau_1 \geq \tau_2 \geq \ldots$, $\langle M, O \rangle = \mathrm{tr}(M^\top O) \leq \sum_i \sigma_i(M) \, \tau_i(O) \leq \sum_i \sigma_i(M) \cdot 1 = \|M\|_*$, using $\tau_i(O) \leq \|O\|_2 \leq 1$. For $O^\star = UV^\top$, $\|O^\star\|_2 = 1$ and $\langle M, O^\star \rangle = \mathrm{tr}(V \Sigma U^\top U V^\top) = \mathrm{tr}(\Sigma) = \|M\|_*$, so the bound is attained. Equality in von Neumann's inequality requires O to share singular vectors with M on the support of Σ with $\tau_i = 1$ for $i \leq r$; when $r = \min(m, n)$ this forces $O = UV^\top$. □

Consequently the Muon step $W_{t+1} = W_t - \eta \, U_t V_t^\top$ of Eq. 1 moves along the direction that decreases the first-order loss model $\langle M_t, \Delta \rangle$ fastest among all Δ of unit spectral norm: it is steepest descent in $\|\cdot\|_2$, whose dual norm is $\|\cdot\|_*$ (Bernstein & Newhouse, 2024b;a). The same computation with the entrywise max norm in place of $\|\cdot\|_2$ gives $O^\star = \mathrm{sign}(M)$, the sign-descent direction to which Adam reduces when its second-moment normalization is exact; the difference between the two optimizers studied here is therefore, to first order, the choice of norm under which the momentum is normalized: per entry, or per singular direction.

Corollary 1. *The Muon direction* $O^\star = UV^\top$ *has all* r *nonzero singular values equal to* 1, *so* $\mathrm{srank}(O^\star) = r = \mathrm{rank}(M)$*: every direction present in the momentum is applied with the same magnitude. An entrywise-normalized update carries no such constraint; its stable rank can be as low as* 1.

C What the Newton–Schulz iteration count controls

The bench never computes the SVD. Following Jordan et al. (2024), it first scales the momentum to unit Frobenius norm, $X_0 = M / \|M\|_F$, so that every singular value satisfies $\sigma_i(X_0) \leq 1$, and then applies k steps of the odd matrix polynomial

$$X_{j+1} = a X_j + b \, (X_j X_j^\top) X_j + c \, (X_j X_j^\top)^2 X_j, \qquad (a, b, c) = (3.4445, \, -4.7750, \, 2.0315), \quad (7)$$

in bfloat16, with k=5 by default. Because X_{j+1} is a polynomial in X_j of the form $X_j\, q(X_j^\top X_j)$, it has the same singular vectors as X_j and maps each singular value by the scalar polynomial

$$p(\sigma) = a\sigma + b\sigma^3 + c\sigma^5. \tag{8}$$

So $X_k = U\, p^{\circ k}(\Sigma)\, V^\top$, and the update applied by Muon is $UV^\top$ up to the per-direction factor $p^{\circ k}(\sigma_i)$. The coefficients of Eq. 7 are chosen for a large slope at the origin, $p'(0) = a = 3.44$, at the price of not converging to exactly 1: $p(1) = 0.701$, the positive fixed points of p are 0.868 and 1.264, and $\max_{[0,1]} p = 1.202$. The following facts are obtained by evaluating $p^{\circ k}$ on a grid of 4×10^5 points per interval; since p is a fixed univariate polynomial, they can be checked in a few lines.

Proposition 1. *Let $\sigma \in [\sigma_{\min}, 1]$ be a singular value of X_0. After k iterations of Eq. 7 the value $p^{\circ k}(\sigma)$ lies in the interval given in Table 6. In particular: (i) the band $[0.68, 1.20]$ absorbs every singular value that enters it, and after at most two further iterations the band contracts to $[0.68, 1.13]$; (ii) a singular value below the band grows by a factor $\approx a = 3.44$ per iteration until it enters, so it needs about $\ln(0.68/\sigma)/\ln 3.44$ iterations: 1.5 from σ=0.1, 3.4 from σ=0.01, 5.3 from σ=10^{-3}; (iii) one iteration is a near-linear rescaling of the momentum for all $\sigma \lesssim 0.2$.*

Table 6: Image of $[\sigma_{\min}, 1]$ under k Newton–Schulz iterations (Eq. 8), i.e. the range of singular values of the update after k steps, as a function of the smallest singular value retained. The paper's default is k=5; Section 5 varies $k \in \{1, 2, 3, 5\}$.

k	$\sigma_{\min} = 0.1$	$\sigma_{\min} = 0.01$	$\sigma_{\min} = 10^{-3}$
1	[0.34, 1.20]	[0.034, 1.20]	[0.0034, 1.20]
2	[0.68, 1.20]	[0.12, 1.20]	[0.012, 1.20]
3	[0.68, 1.13]	[0.40, 1.20]	[0.041, 1.20]
4	[0.68, 1.13]	[0.68, 1.20]	[0.14, 1.20]
5	[0.68, 1.13]	[0.68, 1.13]	[0.47, 1.20]

Reading the dose-response through Proposition 1. A 768×768 momentum with a flat spectrum would have $\sigma_i \approx 1/\sqrt{768} = 0.036$ after Frobenius normalization; real momenta have a decaying spectrum in which most directions sit well below that. The iteration count therefore sets how deep into the spectrum the update is equalized: with k=1 the update is essentially $3.4\, M/\|M\|_F$, a rescaled momentum step with none of the equalization (and, at Muon's learning rate, worse than AdamW, Section 5); with k=3 directions with $\sigma \gtrsim 0.02$ are equalized while the tail is not; with k=5 the equalization reaches $\sigma \approx 2 \times 10^{-3}$. That within-arm reproducibility is already at the full-Muon level at k=3 while loss keeps improving to k=5 says that canonicalization needs only the leading directions to be equalized, whereas most of the loss edge (about two thirds of it) needs the tail. We note that the iteration is run in bfloat16 and on the normalized momentum, so "fidelity" here means closeness of $p^{\circ k}(\Sigma)$ to a constant, not to the identity.

D STABLE RANK

Lemma 2. *For $M \neq 0$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $\operatorname{srank}(M) = \sum_i \sigma_i^2/\sigma_1^2$ satisfies: (i) $1 \leq \operatorname{srank}(M) \leq r$, with $\operatorname{srank}(M) = r$ iff all nonzero singular values are equal and $\operatorname{srank}(M) = 1$ iff $r = 1$; (ii) $\operatorname{srank}(cM) = \operatorname{srank}(M)$ for $c \neq 0$ and $\operatorname{srank}(PMQ) = \operatorname{srank}(M)$ for orthogonal P, Q; (iii) $\operatorname{srank}(M)$ is the inverse of the largest weight in the distribution $\sigma_i^2/\sum_j \sigma_j^2$, so it is continuous in M, unlike* rank.

Proof. (i) Each term σ_i^2/σ_1^2 lies in $(0, 1]$ and the first equals 1; the sum is r iff every term is 1, and it is 1 iff there is a single term. (ii) Singular values scale by $|c|$ and are invariant under orthogonal factors. (iii) With $q_i = \sigma_i^2/\sum_j \sigma_j^2$, $\operatorname{srank}(M) = 1/\max_i q_i$, a continuous function of the singular values, which are continuous in M. □

Two consequences are used in Act 2. First, by Corollary 1 each Muon update has maximal stable rank, so the spectral-rank fingerprint of Muon-trained weights is the expected footprint of accumulating equalized updates; the intervention question is whether that footprint is the cause of the benefit.

Second, stable rank is a property of the endpoint W only: two training runs can reach the same $\mathrm{srank}(W_{OV})$ through different update sequences, which is exactly what the rank-promoting AdamW control does (Appendix K), and why matching the endpoint statistic without the orthogonalized updates recovers none of the loss advantage.

E MATCHING METRIC AND INFERENCE: INVARIANCE AND EXACTNESS

Proposition 2 (Invariance of the behavioral score)**.** *Let model A have residual-stream activations $x_t \in \mathbb{R}^d$ at layer ℓ over tokens $t = 1, \ldots, T$ and a top-k SAE with encoder $(W_{\mathrm{enc}}, b_{\mathrm{enc}}, b_{\mathrm{dec}})$ and decoder W_{dec} (Eq. 4). For any invertible $G \in \mathbb{R}^{d\times d}$, the reparametrized model with activations Gx_t and SAE $(W_{\mathrm{enc}}G^{-1}, b_{\mathrm{enc}}, Gb_{\mathrm{dec}}, GW_{\mathrm{dec}})$ produces identical firing patterns z_i for every feature i. Hence the pair score $s(A, B)$ of Eq. 5, and its symmetrization, are invariant to independent reparametrizations of A and of B, and to any permutation of either model's features. The decoder cosine between features of A and B is not invariant: it changes with G.*

Proof. $W_{\mathrm{enc}}G^{-1}(Gx_t - Gb_{\mathrm{dec}}) + b_{\mathrm{enc}} = W_{\mathrm{enc}}(x_t - b_{\mathrm{dec}}) + b_{\mathrm{enc}}$, so the pre-activations, hence the top-k selection and z_i, are unchanged for every t; $s(A, B)$ is a function of the z_i alone, and the maximum over j in Eq. 5 is invariant to relabeling j. The decoder columns become Gw_i, whose cosine with a fixed vector is not preserved unless G is a scaled orthogonal map. □

Independently trained models are related by no canonical linear map between their residual streams, so Proposition 2 is why decoder-direction matching is uninformative across models (mean best cosine 0.14 even between seeds of one recipe) while behavioral matching is well defined. The proposition also makes precise what the matched-loss protocol compares: sets of firing patterns over shared tokens, with no assumption about geometry.

Proposition 3 (Exact permutation test over models)**.** *Let six models be trained, three per optimizer, and let $s(\cdot, \cdot)$ be the symmetrized pair score. Under the null hypothesis that the optimizer label does not affect the joint distribution of the models, the six models are exchangeable, and for any labeling Λ of three models as "Muon" the statistic $T(\Lambda) = \bar{s}_{\text{within Muon}} - \bar{s}_{\text{within AdamW}}$ has the same distribution across the $\binom{6}{3} = 20$ labelings. The exact one-sided p-value is $\#\{\Lambda : T(\Lambda) \geq T_{\mathrm{obs}}\}/20$; its smallest attainable value is $1/20 = 0.05$, achieved iff the observed labeling uniquely maximizes T.*

Proof. Exchangeability under the null means every relabeling is equally likely to be the true one, so T_{obs} is a uniform draw from the multiset $\{T(\Lambda)\}$; the probability of a value at least as large is the stated count over 20. The count is at least 1 because the observed labeling is included. □

Table 7 lists the fifteen layer-6 pair scores behind Figure 1a. Every within-Muon pair (≥ 0.617) exceeds every other pair (≤ 0.604); $T_{\mathrm{obs}} = 0.6178 - 0.5945 = 0.023$ and no other labeling reaches it, giving p=0.05 (Appendix N).

Table 7: Layer-6 symmetrized pair scores at matched loss (Eq. 5) for the three Muon (M0–M2) and three AdamW (A0–A2) seeds, with 95% intervals from the token-level bootstrap used for the error bars of Figure 1a. These fifteen numbers are the input to Proposition 3.

Pair	Score	95% CI	Pair	Score	95% CI
M0–M1	0.617	[0.613, 0.622]	M0–A0	0.598	[0.594, 0.603]
M0–M2	0.619	[0.615, 0.623]	M0–A1	0.598	[0.594, 0.602]
M1–M2	0.617	[0.613, 0.621]	M0–A2	0.603	[0.598, 0.607]
A0–A1	0.593	[0.589, 0.597]	M1–A0	0.595	[0.591, 0.599]
A0–A2	0.598	[0.594, 0.603]	M1–A1	0.598	[0.593, 0.602]
A1–A2	0.592	[0.588, 0.596]	M1–A2	0.599	[0.594, 0.603]
			M2–A0	0.602	[0.597, 0.606]
			M2–A1	0.600	[0.595, 0.604]
			M2–A2	0.604	[0.599, 0.608]

Treating the fifteen pairwise scores as exchangeable units instead is invalid: each model participates in five of the fifteen pairs, so the pairs are dependent, and a permutation over pairs (with $\binom{15}{3} = 455$ arrangements) overstates the evidence by an order of magnitude. The clustered bootstrap of Appendix N resamples models, not pairs, for the same reason.

Ceiling normalization. Let c_A be the agreement between two SAEs trained with different SAE seeds on the same model A (the instrument ceiling; 0.719 for Muon and 0.769 for AdamW at layer 6). Since a pair score cannot exceed what the instrument reproduces on one model, $\tilde{r} = r/c$ expresses within-arm reproducibility as a fraction of the attainable maximum: $0.618/0.719 = 0.86$ for Muon against $0.594/0.769 = 0.77$ for AdamW. The normalization widens the gap because AdamW's dictionaries are the more reproducible instrument yet the less reproducible models.

F ADDITIONAL CONTROLS

Activation-scale control. Muon's residual streams run 5–7× hotter than AdamW's (layer 3: ∼2,800 vs. ∼550; layer 6: ∼4,200 vs. ∼700; layer 9: ∼7,900 vs. ∼1,100). Rescaling activations to mean norm $\sqrt{d}$ before SAE training leaves the packaging separation intact (dead features 4.7% vs. 1.9% at layer 6).

Matched-token control. Muon's matched-loss checkpoints see fewer tokens (first crossing of loss 3.35 at step $4{,}375$ vs. $5{,}100$). Re-running SAE analysis on Muon's final-step checkpoints leaves packaging metrics unchanged to two decimals, and the feature-match ordering is unchanged at matched step and tokens (Table 1: within-Muon 0.619, cross-optimizer 0.592 over nine pairs with range 0.589–0.595, within-AdamW 0.592).

Depth. Dead-feature separation grows with depth: layer 3, 24% vs. 19% ($1.3\times$); layer 6, 6.9% vs. 1.1% ($\sim 6\times$); layer 9, 3.2% vs. 0.12% ($26\times$).

Init-relative spectra. Stable ranks of $W - W_{\text{init}}$ match the raw-weight ratios within a few percent (MLP-in $5.6\times$ vs. $5.3\times$).

Noise robustness. Gaussian noise added to layer-6 activations at $\sigma \in \{0.05, 0.1, 0.2, 0.4\}\times$ each model's stream RMS, evaluated with clean-trained SAEs. Fraction of clean firings preserved at σ=0.4: Muon 0.710–0.718, AdamW 0.684–0.707. Codes are comparably tolerant per unit relative noise (Muon slightly more); absolute headroom therefore scales with stream norm.

Downstream check (HellaSwag). Zero-shot HellaSwag on the final-step checkpoints of the 124M hybrid arms (the runs of Table 3), of the pure arms' MI355X replicate, of the head-count variants, and of the 1.3B pure arms, all trained on the MI355X node (Appendix M), scored by summed (acc) and length-normalized (acc_norm) completion log-likelihood over the $10{,}042$ validation items (Table 8). The standard error of one model's accuracy is about 0.0045, so single-seed differences of one point are not individually significant; the arm-level ordering rests on the seed-consistent separation. Among the six-head models, every Muon-side model (muon-VO and full Muon) scores above every AdamW-side model (AdamW, muon-QK, muon-xVO); with the model as the exchangeable unit, the probability of that complete separation under the null is $1/\binom{18}{8} = 2.3\times10^{-5}$ (eight Muon-side and ten AdamW-side models, muon-QK seeds 0–3; a one-sided Mann–Whitney test with U=0), and the same separation holds for acc. The three- and twelve-head pure arms reproduce the one-point gap. At 1.3B both Muon models score above both AdamW models by at least 0.037 acc_norm, roughly three times the 124M gap, against a per-model standard error near 0.005.

Live-set-matched rescoring. The within-arm match of Eq. 5 averages over each dictionary's live features, and the fixed SAE recipe leaves 6–8% of features dead on full-Muon models, 8% on muon-MLP, 3–16% on muon-xVO, 2–3% on muon-QK, and none on muon-VO. To score every arm on a live set of the same size we re-trained the layer-6 dictionaries of the pure, muon-attn, muon-MLP, and switch arms (same recipe and seed; the re-trained dictionaries reproduce the published within-arm means to within 0.001) and recomputed each pair's score over (i) the N most frequent live features of the scoring side for $N \in \{4{,}096, 8{,}192, 10{,}240\}$, the last below every arm's live count, and (ii) the features whose firing density lies in $[10^{-4}, 10^{-2}]$ (Table 9). Two facts emerge. The $4{,}096$ most frequent features agree equally under every arm (0.53–0.55), so the reproducibility differences live in mid-frequency features. And once live sets are matched, the excess of partial-Muon arms over

Table 8: Zero-shot HellaSwag at 124M and 1.3B on final-step checkpoints (all MI355X-trained). Loss is the mean final validation loss of the scored checkpoints; per-seed acc_norm in the last column.

Arm	Heads	loss	acc	acc_norm	per-seed acc_norm
Muon	6	3.292	0.289	0.302	0.303/0.306/0.299
muon-VO	6	3.301	0.288	0.299	0.300/0.299/0.298/0.300/0.296
muon-QK	6	3.361	0.284	0.288	0.289/0.286/0.289/0.286
muon-xVO	6	3.369	0.283	0.289	0.289/0.290/0.287
AdamW	6	3.349	0.284	0.288	0.287/0.291/0.285
Muon	3	3.305	0.288	0.299	0.300/0.298/0.300
AdamW	3	3.361	0.283	0.290	0.289/0.288/0.293
Muon	12	3.300	0.288	0.300	0.300/0.301/0.300
AdamW	12	3.358	0.283	0.290	0.286/0.291/0.291
Muon, 1.3B	16	3.034	0.330	0.387	0.388/0.386
AdamW, 1.3B	16	3.109	0.310	0.347	0.349/0.345

full Muon disappears while the ordering full Muon and value-pathway arms $>$ MLP-only arms $>$ AdamW and muon-QK remains.

Table 9: Within-arm reproducibility under live-set-matched scoring (layer 6, three seed pairs per arm, symmetrized). "plain" is the published metric, recomputed on these dictionaries (re-trained ones reproduce the published means to within 0.001); top-N scores the N most frequent live features; band restricts to firing density in $[10^{-4}, 10^{-2}]$. The last row is the seed-matched cross-optimizer score.

Arm	live features	plain	top-4,096	top-8,192	top-10,240	band
full Muon	11,414	0.618	0.547	0.610	0.627	0.632
switch (500 Muon steps, then AdamW)	12,280	0.629	0.542	0.599	0.619	0.633
muon-VO	12,288	0.632	0.550	0.592	0.618	0.638
muon-attn	12,282	0.629	0.545	0.589	0.616	0.634
muon-MLP	11,223	0.603	0.543	0.599	0.609	0.615
muon-xVO	11,213	0.606	0.531	0.597	0.608	0.618
AdamW	12,142	0.593	0.541	0.572	0.588	0.598
muon-QK	11,926	0.588	0.540	0.570	0.587	0.593
Muon vs. AdamW (same seed)	11,778	0.592	0.529	0.578	0.595	0.601

G WHAT MATCHED, UNMATCHED, AND DEAD FEATURES LOOK LIKE

For the seed-0 Muon and AdamW models at matched loss (and, as controls, the seed-0/seed-1 pairs within each optimizer) we streamed one million validation tokens through both models, matched layer-6 features as in Section 3, and recorded each feature's firing density, its top-activating contexts, its positional histogram over eight bins of the 1,024-token context, and its *token concentration*: the fraction of its 32 strongest activations that fall on the same token. Table 10 compares the bottom decile of live features by best match ("unmatched", best ≤ 0.27) with the top decile ("matched", best ≥ 0.91).

Matched features are token and lemma detectors. Nearly nine in ten top-decile features fire on one token or subword across contexts, and their partners fire on the same one: a "udd" subword feature (p**udd**les, c**udd**ly, h**udd**ling; match 0.92), "**choice**" after *have no*/*had little* (0.93), "**face**" as a noun (0.93), and "**Human**" in *Human Rights*/*Human Resource* (0.96). These are the features that survive both the seed and the optimizer boundary, and they carry the agreement numbers of Table 1.

Unmatched features are context and format features, and the same kinds fail across seeds as across optimizers. Bottom-decile features have token concentration 0.27–0.30 in every pairing. Reading them, three kinds recur. (i) *Format features*: an all-caps subword-fragment feature (AC-

Table 10: Feature-type statistics for the least- and best-matched deciles of live features (at least 20 firings in the stream). Token concentration is the fraction of a feature's 32 strongest activations on one token; a "token detector" has concentration ≥ 0.5. Positional entropy (normalized) is 0.99 for every group, so no positional features exist at this layer.

Pair	Decile	n	median density	token conc.	token detectors
Muon → AdamW	unmatched	1,140	1.5×10^{-3}	0.27	12%
	matched	1,140	4.4×10^{-4}	0.76	88%
AdamW → Muon	unmatched	1,203	6.8×10^{-4}	0.29	12%
	matched	1,203	5.3×10^{-4}	0.75	87%
Muon → Muon (seeds)	unmatched	1,140	8.4×10^{-4}	0.30	15%
	matched	1,140	4.2×10^{-4}	0.76	87%
AdamW → AdamW (seeds)	unmatched	1,203	7.8×10^{-4}	0.27	10%
	matched	1,203	5.6×10^{-4}	0.74	84%

KNOW**LED**GEMENT, Fo**xt**rot; match 0.05) whose best partner is a different all-caps fragment feature; a number-and-unit feature (**20** Hz, **30**%; 0.06). (ii) *Dense generic features* (density 0.02–0.04) that fire on the first content token after a heading or line break, or on a predicate under negation (*not* **fantastic**, *not* **discouraged**, *not necessarily* **condone**; 0.08); each model has such features but partitions the contexts differently, so no single partner correlates. (iii) A few *coherent lexical features without a counterpart*: a "role noun" feature (**project** manager, **role** model, **postseason** stalwart, **Derby** winner; 0.07) and a "mushroom" feature (0.08), which is also unmatched between two Muon seeds. Rare, document-specific features (a single repeated passage) round out the tail. Nothing in the unmatched decile is specific to the optimizer boundary: the Muon-to-AdamW list reads like the Muon-to-Muon list, which is what the equality of cross-optimizer agreement with the seed ceiling predicts.

Muon's extra dead features are never-recruited directions, not collapsed duplicates. The seed-0 Muon dictionary has 734 features that never fire on the stream or on the SAE's own validation tokens, against 172 for AdamW. Their decoder directions sit far from every live direction (mean nearest-live cosine 0.19 for Muon, 0.17 for AdamW, against 0.29–0.32 for a live feature's nearest live neighbour), and projecting them onto the unembedding gives incoherent token lists (*before, circulated, h, lad*), whereas live features project onto their own tokens. Dead features are therefore leftover initial directions the optimizer never trained, consistent with Muon's tighter packing of the live dictionary (Appendix J) rather than with duplicated or split features. Removing them from the live-set comparison is what the rescoring of Table 9 does.

H SAE ROBUSTNESS

Re-training the SAE at $k \in \{16, 64\}$ preserves the within-Muon > within-AdamW ordering (three seeds, layer 6, each dictionary encoded at its own k during matching): 0.633 vs. 0.596 at k=16; 0.618 vs. 0.594 at k=32; 0.573 vs. 0.556 at k=64. Agreement falls as the code gets denser and the gap narrows at k=64, but Muon remains the more seed-reproducible substrate at every k; the k=16 dictionaries leave 16–44% of features dead against under 1% at k=64.

I WHY THE MATCHING RESULT IS NOT A METRIC ARTIFACT

Three controls make the feature-identity claim falsifiable rather than definitional. (i) The metric registers content change: matching the layer-6 features of one seed against the layer-3 and layer-9 features of another gives 0.590 and 0.570 for Muon and 0.565 and 0.552 for AdamW, against same-layer pairs of 0.618 and 0.590 scored on the same dictionaries, so the metric resolves differences several times larger than the cross-optimizer departure from the AdamW ceiling (≈0.005). Head count at fixed width, by contrast, leaves content near the ceiling (three- vs. twelve-head Muon models 0.608; Table 4). (ii) Geometry-sensitive CKA gives the opposite answer from feature matching: cross-optimizer activations are geometrically farther apart (0.54) even when firing-pattern features match. (iii) Decoder directions themselves do not align across seeds (mean best cosine 0.14). The

claim is that this optimizer change preserves top-k SAE firing-pattern content up to the AdamW seed ceiling at three, six, and twelve heads.

J DICTIONARY DEMOGRAPHICS AT MATCHED LOSS

Table 11 lists the dictionary statistics behind the packaging claims of Section 4: reconstruction ties at matched loss while dead fraction, firing density, and decoder crowding separate with no seed overlap.

Table 11: Dictionary demographics at matched loss (layer 6, ranges over three seeds). Reconstruction ties; every other statistic separates with no seed overlap.

SAE statistic	Muon	AdamW
Reconstruction error (fraction of variance unexplained, FVU)	0.119–0.140	0.121–0.147
Dead features (%)	6.0–8.3	0.2–1.6
Median firing rate ($\times 10^{-4}$)	7.0–7.1	7.9–9.0
Nearest-neighbor cosine	.309–.311	.286–.292
Residual-stream ℓ_2 norm	$\sim$4,200	$\sim$700

K OV-RANK SUFFICIENCY, LEARNING RATES, QK-NORM, AND SCALE

Rank-promoting regularizer. AdamW with the per-head $\sigma_1(W_V W_O)$ penalty at $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 3\times10^{-2}\}$ (seed 0) reaches loss $3.350/3.382/3.476/3.549$ while the value-circuit stable rank rises $50.4 \to 53.4 \to 56.9 \to 64.0$, reaching full Muon's 64.7 on this sweep's own rank instrument. That instrument (mean over layers and heads of the per-head stable rank of $W_V W_O$, computed during training) is not the one used in Figure 3 and Table 3: it places its weakest-penalty run and full Muon closer together (50.4 vs. 64.7) than that instrument places AdamW and full Muon (31.5 vs. 76.6), so the two sets of values are not comparable, and the conclusion rests on the within-sweep trend.

Warmup. Without the matched 250-step warmup the hybrid arms are much worse (muon-attn 3.353–3.362; muon-MLP 3.448–3.477), so all hybrid numbers in the paper use the matched-warmup protocol.

Learning-rate sweep (seed 0). muon-VO loss $3.330/3.309/3.300/3.314$ at Muon learning rate $\{0.005, 0.01, 0.02, 0.04\}$ (always near Muon); muon-QK $3.377/3.358/3.353/3.348/3.347/3.354$ at $\{0.005, 0.01, 0.02, 0.04, 0.08, 0.16\}$ (flat from 0.02 to 0.08 inside the AdamW seed ranges, 3.347–3.351 on the MI355X node and 3.348–3.352 on the B200 node, and worse at 0.16, so the sweep has turned over without leaving the AdamW seed range); muon-MLP $3.357/3.366/3.368/3.418$ at $\{0.005, 0.01, 0.02, 0.04\}$ (never below AdamW, with its best point at the lowest rate tried; the 0.02 entry is the three-seed mean).

QK-norm removal (two seeds). Without QK-norm: AdamW 3.352, full Muon 3.311, muon-VO 3.318 (recovering $\sim$83% of the 0.041-nat edge), muon-QK 3.371 (above AdamW). The value-versus-scoring split is not an artifact of QK-norm.

30M. Pure arms over three seeds (B200 node): Muon 3.690 ($3.691/3.689/3.690$), AdamW 3.748 ($3.749/3.749/3.745$), a 0.058-nat gap.

350M. Pure arms over three seeds (B200 node): Muon 3.133, AdamW 3.193. Hybrids over three seeds: muon-VO $3.139/3.140/3.140$ (seeds 0 and 2 on the MI355X node, seed 1 on the B200 node), muon-QK $3.202/3.207/3.207$ (MI355X node); the original seed-0 hybrids on the B200 node gave 3.139 and 3.200. Layer-12 within-arm match, plain/top-4,096/top-8,192/top-10,240/density band: full Muon $0.630/0.562/0.614/0.637/0.649$, AdamW $0.592/0.584/0.605/0.608/0.606$, and cross-optimizer $0.589/0.541/0.586/0.601/0.607$ (B200 dictionaries, re-trained, reproducing the published plain means; 17–26% of features dead); muon-VO $0.638/0.569/0.586/0.604/0.643$ (under 0.1% dead) and muon-QK $0.590/0.569/0.589/0.598/0.599$ (13–16% dead), with hybrid dictionaries trained and matched on the MI355X node except the muon-VO seed-1 dictionary and its two pairs,

which ran on the B200 node with its rerun. The value-pathway arm keeps its plain and band advantage over AdamW but not its top-N advantage, unlike at 124M (Table 9).

L DECOMPOSING THE VALUE CIRCUIT

Table 12: Decomposing the value circuit (124M, three seeds, matched warmup). Orthogonalizing the output-write W_O alone recovers more than the value-read W_V; the two are sub-additive; higher OV rank does not track higher recovery.

Muon on	Loss	Recovers	OV stable rank
nothing (full AdamW)	3.350	0%	31.5
W_V only (muon-V)	3.330	∼36%	53.1
W_O only (muon-O)	3.314	∼62%	47.5
W_V, W_O (muon-VO)	3.301	∼86%	76.2

M COMPUTE BUDGET, TIMING, AND HARDWARE

Two nodes ran the study with the same code, data files, and seeds. An 8×NVIDIA B200 node trained the pure arms and their SAE study, the matched-token and switch controls, muon-attn, muon-MLP, muon-V, muon-O, the Newton–Schulz and SOAP arms, and the weight-decay, GeLU, QK-norm-removal, 30M, and 350M variants. An 8×AMD MI355X node trained muon-VO and muon-QK with their learning-rate sweeps, muon-xVO, the muon-MLP learning-rate sweep, the head-count variants and the three-head retune, the 350M hybrid seeds, the 1.3B runs, and the cross-vendor replicate of the pure arms, which agrees with the B200 pure arms to 0.001 nats (Table 2); every HellaSwag checkpoint comes from this node, and one of its GPUs ran the 1.3B dictionaries, matching, and checkpoint re-evaluation. Three MI355X runs that died on node faults (a hung all-reduce, a crash at step 0, an allocation timeout) were rerun on four B200 GPUs with the same code and global batch: muon-QK seed 4, three-head Muon seed 2 at the retuned rate, and 350M muon-VO seed 1; the original seed-0 350M hybrids also ran on the B200 node. Smaller controls ran on whichever node was free. Training FLOPs are estimated as $6N_{\text{params}}N_{\text{tokens}}$, not hardware peak. Each 124M run sees $5{,}100 \times 524{,}288 = 2.674 \times 10^9$ tokens, or $\approx 1.99 \times 10^{18}$ FLOPs, in about 8 minutes; the logged budget across 124M runs, 350M replications, and controls is $\approx 2.2 \times 10^{20}$ FLOPs, plus $\approx 8.4 \times 10^{19}$ for the four 1.3B runs. Quiet-node step times at 124M are 92.4 ms (AdamW), 91.0 ms (Muon), and 106.3 ms (muon-VO, two unfused optimizers). Dataset: FineWeb, identical data order across seeds within each comparison.

N CLUSTERED INFERENCE

With three seeds per optimizer, the fifteen pairwise scores of Figure 1a are built from six models, and each model enters five pairs, so pairs are not exchangeable units. We report intervals clustered by model identity: each bootstrap resample draws model seeds with replacement within optimizer, then recomputes pair scores. For significance we permute the optimizer label over models: of the $\binom{6}{3}$=20 ways to label three of the six models "Muon", the observed labeling is the only one for which the within-label-A mean exceeds the within-label-B mean by at least the observed 0.023, giving an exact one-sided p=0.05. This is the smallest value attainable with three seeds per arm; the five-seed replication (0.619 vs. 0.591 over ten pairs each) and the layer, scale, matched-token, and fidelity checks carry the evidential weight, not this single test.

O MATCHED-LOSS, MATCHED-STEP, AND MATCHED-TOKEN COMPARISONS

P ABLATIONS NOT YET CLAIMED

The intervention results establish that Muon on W_V, W_O is largely sufficient in this bench, that Muon on W_Q, W_K and on MLP blocks (protocol-matched) are not, and that Muon on everything except

Table 13: Checkpoints used for SAE analysis. Tokens are step $\times$ 524,288. Matched-step and matched-token coincide because all runs use the same batch configuration; losses are three-seed means.

Comparison	Optimizer	Step	Tokens	Val. loss
Primary matched-loss	Muon	4,375	2.294×10^9	3.347
Primary matched-loss	AdamW	5,100	2.674×10^9	3.350
Matched-step/token	Muon	5,100	2.674×10^9	3.292
Matched-step/token	AdamW	5,100	2.674×10^9	3.350

W_V, W_O (muon-xVO) recovers none of the loss edge, so the pathway is necessary for it. Not yet run or not usable: the joint arm with Muon on W_V, W_O and the MLP blocks (the one missing cell of the QK$\times$VO$\times$MLP factorial); the component arms at 1.3B; SAE analysis at 1.3B, because the fixed recipe does not transfer to that width (32,768 features at layer 12 leave 69–76% of features dead on held-out tokens for both optimizers, FVU 0.27–0.28, and activation normalization does not change this in a 10M-token test), so the scores on the surviving features (within-Muon 0.627, within-AdamW 0.596, cross-optimizer 0.580 over two seeds per arm) reorder on the 4,096 most frequent features (0.580, 0.609, 0.557) and are not claimed; a no-orthogonalization control (one seed at the Muon learning rate diverged to loss 4.75, which we report as uninformative rather than as a data point); downstream evaluation beyond zero-shot HellaSwag; SAE families other than top-k; and matching that filters generic or positional features. The rank-promoting control rules out endpoint W_{OV} stable rank as a sufficient knob, and the Newton–Schulz sweep shows the benefit is monotone in orthogonalization fidelity at a fixed learning rate, but because fewer iterations also shrink the update's norm, matched update-norm controls and optimizers that reproduce Muon-like W_V, W_O update spectra without full Muon would be needed to isolate update geometry as the unique mechanism.

Q Comparison with the associative-memory ablation

Wang et al. (2025) assign Muon to one transformer block at a time with Adam elsewhere ("Independent Blocks"), then combine blocks, on a 160M NanoGPT trained on FineWeb, and replicate at 0.7B. Table 14 places their reported validation losses at 10,000 steps next to our hybrid arms, as fractions of each study's own Adam-to-Muon gap, together with the protocol facts that bear on the one disagreement. Their ungated and gated gaps are $3.924-3.565=0.359$ and $3.884-3.513=0.371$ nats; ours is $3.350-3.292=0.058$.

Two readings are consistent with both tables. The first is that the FFN contribution is real but conditional on how far the Adam baseline sits from Muon: with a 0.36–0.37-nat gap, any large block moved onto Muon closes part of it, and their gated-FFN numbers already halve the FFN share while the value-pathway share moves from 0.45 to 0.53. The second is that the FFN effect depends on the FFN itself (activation, gating) or on the learning-rate assignment for hybrids, which they leave at full Muon's rate and we sweep. Our muon-MLP sweep over an $8\times$ range never drops below the AdamW level, so within the swept range the negative is not a tuning artifact (its best point is the lowest rate tried), but we have not run their protocol and do not claim their result would fail under it.

R Raw cross-seed artifact transfer fails

Evaluating an SAE trained on seed i against layer-6 activations of seed j fails for both optimizers. Cross-seed FVU is 1.83–3.07 for AdamW (mean 2.25) and 2.31–3.19 for Muon (mean 2.67), against own-seed baselines of 0.12–0.15. Feature *content* is shared across seeds, but each model expresses it in a private coordinate frame to which SAE weights are welded. Greater firing-pattern canonicalization does not confer artifact portability. Aligned transfer, fitting a linear map between residual spaces before reuse, is left to future work.

Table 14: Fraction of the Adam-to-Muon loss gap recovered by Muon on one component (Adam elsewhere), and protocol facts. Wang et al. values are computed from Table 1 of their arXiv version (ungated / gated FFN; the ungated losses also appear in Figure 1(c–d) of the ICLR version); ours from Table 3 and Appendix K. Negative fractions are reported as 0.

	Wang et al. (2025)	This paper
Model, data	160M NanoGPT, FineWeb	124M GPT-2-class, FineWeb
Steps	10,000	5,100 (524,288 tokens/step)
Attention / FFN	standard / ungated or gated	QK-norm, rotary / ReLU^2
Adam-to-Muon gap (nats)	0.359 / 0.371	0.058
Adam (β_1, β_2), weight decay	$(0.8, 0.95)$, none	$(0.9, 0.95)$, none
Nesterov momentum (Muon)	disabled	enabled (μ=0.95)
Learning rates, pure arms	grid search	independently tuned recipes
Learning rate, hybrid arms	full Muon's, untuned	Muon side swept 8–32$\times$ (seed 0)
Warmup on the Adam side of hybrids	not reported	matched, 250 steps
Seeds per hybrid arm	not reported	3
Muon on W_Q, W_K	0.09 / 0.09	0
Muon on W_V, W_O	0.45 / 0.53	0.86
Muon on W_V only	0.26 / 0.37	0.36
Muon on W_O only	0.43 / 0.33	0.62
Muon on one FFN matrix	0.58–0.62 / 0.25–0.27	0 (both matrices)
Muon on W_V, W_O + FFN	0.94 / 0.95	not run
Muon on all hidden except W_V, W_O	not run	0 (muon-xVO)
Feature-level measures	none	SAE match, CKA, stable rank

S SEED-LEVEL DETAIL FOR THE FIDELITY, SOAP, AND HEAD-COUNT CONTROLS

Newton–Schulz sweep. Final losses: 1 iteration 3.412 (seed 0); 2 iterations 3.381 (seed 0); 3 iterations 3.329/3.331/3.335; 5 iterations 3.293/3.292/3.292. Within-arm layer-6 match at 3 iterations: 0.617/0.613/0.615.

SOAP. Final losses 3.291/3.291/3.292. Within-arm match 0.617/0.611/0.600; live features per model 5,653–9,647 of 12,288. SOAP-to-Muon pairs average 0.602 and SOAP-to-AdamW pairs 0.578 (nine pairs each).

Cross-vendor. AMD MI355X, AdamW 3.348/3.351/3.347; Muon 3.292/3.292/3.293.

Hybrid seeds 3–4. muon-VO 3.301/3.302 and muon-QK 3.363/3.357 (muon-QK seed 4 trained on the B200 node after an MI355X all-reduce hang); five-seed means 3.301 and 3.360; within-arm layer-6 match over ten pairs 0.632 and 0.587 (top-4,096/top-8,192/top-10,240/band: muon-VO 0.551/0.593/0.619/0.638, muon-QK 0.540/0.573/0.588/0.593).

Necessity arm (muon-xVO). Final losses 3.375/3.359/3.373; within-arm layer-6 match 0.604/0.610/0.603; SAE reconstruction FVU 0.150/0.167/0.129 with 3–16% dead features; layer-6 residual-stream norm ∼2,750–2,880, so the stream inflation of Section 4 follows Muon elsewhere in the network (scoring circuit, MLP blocks, or both) rather than on the value pathway.

1.3B. 24 layers, d=2048, 16 heads (1,310,982,144 parameters with tied embeddings, 1,207,959,552 non-embedding), the same 5,100 steps of 524,288 tokens (≈2 tokens per parameter) and the 124M learning rates, warmup, and schedule, on the MI355X node: Muon seeds 0 and 1 3.034/3.035; AdamW seeds 0 and 1 3.110/3.108. The four runs embed byte-identical training code, and re-evaluating each final checkpoint outside the training loop reproduces its logged loss to within 2.2×10^{-5}. Seed 1 of Muon was run twice (the first attempt hit its allocation limit at step 5,013); the two same-seed trajectories differ by at most 0.010 nats (step 250) and by 0.0014 at step 5,000, the same order as the Muon seed-to-seed spread there (0.0005), so runs on this stack are not bit-reproducible and rerun noise is comparable to seed variation, both far below the 0.075-nat optimizer gap. The Muon-minus-AdamW loss gap is 0.524 at step 500, 0.124 at step 3,000, 0.096 at step 4,000, and 0.075 at the end; both Muon seeds first evaluate below both AdamW final losses at step 4,250.

Each run is $\approx 2.1 \times 10^{19}$ FLOPs and 6.2 (AdamW) to 7.0 (Muon) hours of training on eight GPUs; at this width a Muon step is 10–12% slower than an AdamW step, reversing the 124M ordering.

Head count. Final losses: 3 heads, Muon $3.305/3.305/3.304$, AdamW $3.366/3.358/3.359$; 12 heads, Muon $3.302/3.298/3.300$, AdamW $3.361/3.356/3.356$. Layer-6 within-arm match, each model evaluated at its own head count: Muon-3h $0.623/0.627/0.625$; AdamW-3h $0.592/0.599/0.591$; Muon-12h $0.611/0.611/0.611$; AdamW-12h $0.571/0.576/0.571$. Cross-optimizer, seed-matched: 3 heads $0.597/0.593/0.596$; 12 heads $0.578/0.576/0.578$. Cross-architecture (Muon, 3 vs. 12 heads, seed-matched): $0.608/0.608/0.607$. Retuned three-head rates (seed-0 sweeps; both optima sit at the low edge of their grids): Muon $0.01/0.02/0.04 \rightarrow 3.300/3.305/3.307$, AdamW $1.2/1.8/2.7\times10^{-3} \rightarrow 3.362/3.366/3.378$; at Muon 0.01 and AdamW 1.2×10^{-3}, Muon $3.300/3.302/3.302$ (seed 2 trained on the B200 node after an MI355X crash) and AdamW $3.362/3.359/3.358$; within-arm match Muon $0.631/0.631/0.634$, AdamW $0.611/0.615/0.605$; cross-optimizer $0.607/0.604/0.610$. Matching scores Muon models at step $4{,}500$, their first checkpoint below loss 3.35 (mean losses 3.348 at three heads, 3.343 at twelve, 3.344 retuned), and AdamW models at their final step; the GeLU Muon checkpoints are at step $4{,}875$ (3.343).