ORSCALE: ORTHOGONALIZED OPTIMIZATION WITH LAYER-WISE TRUST-RATIO SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon fixes the *direction* of every matrix-valued update at the polar factor of its momentum, while each layer's step *magnitude* is addressed only by a static shape correction. We derive a dynamic per-layer scalar by adapting the LARS/LAMB trust-ratio principle to the orthogonalized setting, where the standard denominator candidates—the raw momentum norm or the polar-factor norm—either live in the wrong unit space or carry no update-scale information. The resulting method, *OrScale*, uses the norm of the parameter-space direction actually applied and anchors each layer's ratio at one via a per-layer calibration, so that the Moonlight recipe (tuned for AdamW, shared with Muon via RMS matching) transfers with *no additional sweep*; a component ablation confirms each design choice is individually load-bearing. Theoretically, OrScale retains a nuclear-norm $O(1/\sqrt{T})$ convergence rate for any clipped multiplier and achieves a strict layer-adaptive descent gain $\kappa_{\text{eff}} > 1$ under two conditions estimable from standard training diagnostics—a bound that predicts the gain should *grow with architectural heterogeneity*. Experiments confirm the prediction: with every hyperparameter inherited verbatim from the Moonlight recipe, OrScale matches or beats Muon+Moonlight across dense 125M–1.1B FineWeb-Edu pre-training, and on a 16B-A3B mixture-of-experts model—where the logged trust ratios separate cleanly by layer class—the gap widens by an order of magnitude to 0.130 nats (3.8% relative) at parity wall-clock cost.

1 INTRODUCTION

Muon has rapidly moved from a compact optimizer idea to a practical ingredient of large-scale training. Its core step takes a matrix-valued momentum, applies a few Newton–Schulz iterations to approximate the polar factor, and subtracts this spectrally flattened direction from the weights; the update can be read as steepest descent under the spectral norm (Bernstein & Newhouse, 2024), and production systems—Moonlight (Liu et al., 2025), Kimi K2.5 (Team et al., 2026a), DeepSeek-V4 (DeepSeek-AI, 2026)—and the modded-nanogpt records (Jordan et al., 2024b;a) have made orthogonalized updates a fixture of current large-scale practice.

Orthogonalization settles the direction of each layer's update but leaves its *magnitude* under-specified: after the polar factor, every matrix layer moves by the same global learning rate, even though attention projections, MLP blocks, routers, and experts see very different gradient statistics and curvature. Moonlight corrects the predictable part with a static per-shape factor $0.2\sqrt{\max(m,n)}$ (Liu et al., 2025), but a static factor cannot track how each layer's update scale evolves *during* training. For SGD and Adam, LARS and LAMB addressed precisely this problem by introducing a *trust ratio*—the weight norm divided by the norm of the direction about to be subtracted—that scales each layer's step so that relative step sizes stay comparable across layers, enabling stable large-batch training (You et al., 2017; 2020). Can the same idea be applied to an orthogonalized optimizer, and if so, what should the ratio look like when the update direction has been spectrally flattened?

With the orthogonalized direction kept untouched, the remaining choice—one positive scalar per layer per step—can be derived rather than searched. Following the LARS/LAMB template, the quantity a layer-wise scalar should regulate is the *relative* step $\|\Delta W_\ell\|_F / \|W_\ell\|_F$; holding it near a common target forces a trust ratio with numerator $\|W_\ell\|_F$ and denominator the norm of whatever the optimizer is about to subtract. After orthogonalization, however, that denominator stops being

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.

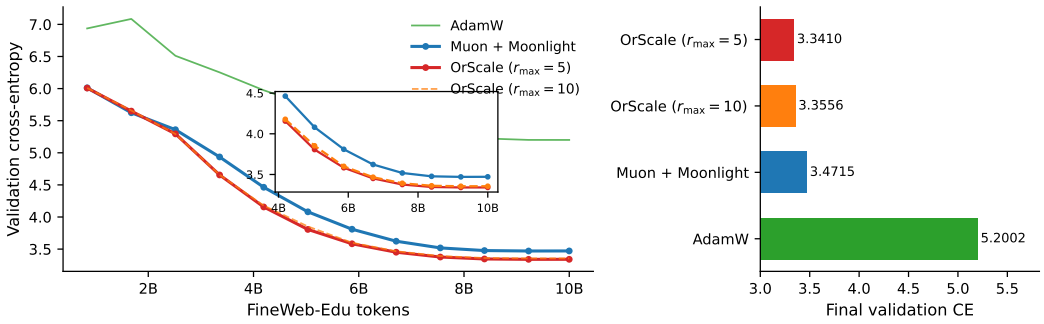


Figure 1: **Moonlight-16B-A3B mixture-of-experts pre-training on 10B FineWeb-Edu tokens**, every hyperparameter inherited verbatim from the Moonlight recipe (LR 4.2×10^{-4} shared by all optimizers; OrScale never swept; Section 5.1). *Left:* validation cross-entropy (inset: final 6B tokens). *Right:* final values. OrScale improves on Muon+Moonlight by 0.130 nats (3.8% relative)—an order of magnitude beyond the $+0.011$-nat dense gap at 1.1B—as the heterogeneity condition of Section 4 predicts (tested in Section 5.3).

a bookkeeping choice: the direction actually applied is the shape-scaled polar factor together with weight decay, not the raw momentum, so the denominator should be the Frobenius norm of that parameter-space direction rather than a gradient-space surrogate. A ratio built this way would still, at initialization, silently shrink an already-tuned step by a layer- and width-dependent factor; a one-time per-layer calibration therefore anchors every ratio at exactly one, making the first update coincide with Muon+Moonlight's. And because a weight-norm numerator can feed its own growth, weight decay is coupled *inside* the scaled update, giving the ratio a finite ceiling. Together these design choices define OrScale (Section 3), with Muon and Muon+Moonlight recovered as the limit $r_{\min}{=}r_{\max}{=}1$; a component ablation shows each choice is individually load-bearing, each deletion producing a distinct, logged failure (Section 5.5).

The calibration anchor doubles as our evaluation protocol: the learning rate and weight decay of the Moonlight recipe—tuned for AdamW, shared with Muon via RMS matching—carry over to OrScale unchanged, so *every OrScale language-model number in this paper, 125M dense through a 16B-A3B mixture-of-experts model, runs at the untouched recipe learning rate; OrScale was never swept at any language-model scale*. The theory then says when the added machinery should pay: the gain κ_{eff} over Muon exceeds 1 whenever the logged trust ratios are heterogeneous across layers and aligned with the layer-wise signal-to-smoothness ratio—two conditions estimable from training diagnostics, both expected to strengthen with architectural heterogeneity. That makes mixture-of-experts models a natural stress test, and the experiment agrees (Figure 1).

Contributions.

- **A trust ratio from first principles.** A forward derivation of Muon's layer-wise trust ratio from the LARS/LAMB template, adapted to the orthogonalized setting; a parameter-light algorithm whose limits recover Muon and Muon+Moonlight; and a component ablation in which each design-choice deletion fails in its own recognizable way (Section 3).
- **Zero-sweep transfer as a protocol.** The calibration anchor lets a Moonlight/AdamW-tuned recipe transfer with no additional sweep—an ability we then lean on at every language-model scale we train (Section 5.1).
- **Theory conditioned on measurable statistics.** A nuclear-norm $O(1/\sqrt{T})$ rate for any clipped multiplier, a strict gain $\kappa_{\text{eff}} > 1$ under two conditions checkable from logged trust ratios and standard per-layer diagnostics rather than oracle smoothness constants alone, and a collapse theorem for unit-inconsistent ratios (Section 4).
- **Validation from 6.5M to 16B parameters.** OrScale ranks first on CIFAR-10/DavidNet, matches or beats Muon+Moonlight over dense 125M–1.1B FineWeb-Edu pre-training, and widens the gap to $+0.130$ nats on 16B-A3B MoE at wall-clock parity (Section 5).

2 BACKGROUND AND POSITIONING

Muon and orthogonalized updates. Muon (Jordan et al., 2024b) updates each matrix-valued parameter $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$ by $W_{\ell,t+1} = W_{\ell,t} - \eta_t Q_{\ell,t}$ (optional weight decay aside), where $Q_{\ell,t} = \mathrm{NS}_k(\widetilde{M}_{\ell,t})$ is a k-step Newton–Schulz approximation to the polar factor $\mathrm{Ortho}(\widetilde{M}_{\ell,t})$ of the Nesterov-lookahead gradient momentum. If $Q_A = \mathrm{Ortho}(A)$ then $\langle A, Q_A \rangle = \|A\|_*$ and $\|Q_A\|_F = \sqrt{\mathrm{rank}\, A}$, making orthogonalization a nuclear-norm/spectral primitive rather than an entrywise adaptive update (Bernstein & Newhouse, 2024). Moonlight adds the static shape factor $0.2\sqrt{\max(m,n)}$, equalizing the per-element RMS of Q across shapes and pinning it to AdamW's typical update RMS so that AdamW-tuned learning rates carry over (Liu et al., 2025); MuonClip (Team et al., 2026b) and AdaMuon (Si et al., 2025) extend Muon along other axes. OrScale keeps the Muon direction untouched and addresses the complementary question of the per-layer scalar magnitude over training time.

Trust-ratio optimizers. LARS (You et al., 2017) scales each layer's SGD step by $\|W_\ell\|_F / \|G_\ell + \lambda W_\ell\|_F$—weight norm over the norm of the exact direction SGD is about to subtract—holding every layer's *relative* step near a common target; LAMB (You et al., 2020) applies the same template to the Adam-normalized direction and enabled large-batch BERT training (Devlin et al., 2019; Goyal et al., 2018). The template silently relies on a unit convention—numerator and denominator are both norms of parameter-space objects—that Muon's pipeline breaks at both ends: the post-orthogonalization norm $\|Q_\ell\|_F$ is a shape constant, and the momentum norm $\|\widetilde{M}_\ell\|_F$ has gradient units. Section 3.1 turns this unit analysis into the denominator that restores the convention (Definition 1); Section 5.5 shows what happens when it is dropped.

Static corrections and width transfer. Moonlight's shape factor and μP-style parameterization (Yang et al., 2022; 2023) handle predictable shape and width effects at initialization; a trust ratio handles training-time drift, and OrScale's LM configuration composes the two—reproducing the Moonlight step exactly at $t{=}0$, adapting thereafter. Preconditioner methods such as Sophia (Liu et al., 2024) and SOAP (Vyas et al., 2025) modify the direction and are orthogonal to the magnitude question here.

Positioning within Muon convergence theory. A rapidly growing line of work analyzes Muon and its relatives: smoothness-based rates for the exact and inexact polar step (Li & Hong, 2025), non-Euclidean trust-region and LMO views (Kovalev, 2025; Pethick et al., 2025), and layer-wise (L^0, L^1) smoothness frameworks with per-layer stepsizes such as Gluon (Riabinin et al., 2025). These analyses obtain sharper constants than ours, but condition their layer-adaptive conclusions on *oracle* smoothness constants unobservable during training. Ours trades sharpness for verifiability: the gain condition of Theorem 3 involves the *logged, clipped trust ratios* $\hat{r}_{\ell,t}$—quantities OrScale computes anyway—and per-layer statistics estimable from training diagnostics, so whether the gain regime holds is checkable on any run; we check it at 16B in Section 5.3, and Section B tabulates the guarantee structures.

3 DERIVING THE ORSCALE TRUST RATIO

Notation. $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$ is a matrix parameter, $p_\ell = \min(m_\ell, n_\ell)$, $\|\cdot\|_F$ and $\|\cdot\|_*$ are the Frobenius and nuclear norms, $\mathrm{Ortho}(\cdot)$ the polar factor, NS_k its k-step Newton–Schulz approximation, $\widetilde{M}_{\ell,t}$ the Nesterov-lookahead momentum, $\mathrm{clip}(\cdot, r_{\min}, r_{\max})$ clamps to an interval, and ε is a small floor. Every Muon-family optimizer in this paper shares the front end

$$M_{\ell,t} = \mu\, M_{\ell,t-1} + G_{\ell,t}, \qquad \widetilde{M}_{\ell,t} = \mu\, M_{\ell,t} + G_{\ell,t}, \qquad Q_{\ell,t} = \mathrm{NS}_k(\widetilde{M}_{\ell,t}), \tag{1}$$

the choice used by every reported large-scale Muon system (Liu et al., 2025; Jordan et al., 2024b); non-matrix parameters (biases, norm gains, embeddings when desired) use an AdamW fallback (Loshchilov & Hutter, 2019). Figure 2 previews the derivation and the update it produces.

3.1 THE DERIVATION

Muon's value lies in its direction: the polar factor equalizes singular values and yields the descent identity $\langle A, Q_A \rangle = \|A\|_*$ on which its analysis rests. We therefore keep the direction and adapt magnitude only: once the direction of the applied step—call it $D_{\ell,t}$—is fixed, the most general layer-wise modification is $W_{\ell,t+1} = W_{\ell,t} - \eta_t\, \rho_{\ell,t}\, D_{\ell,t}$, one positive scalar $\rho_{\ell,t}$ per layer per step. This

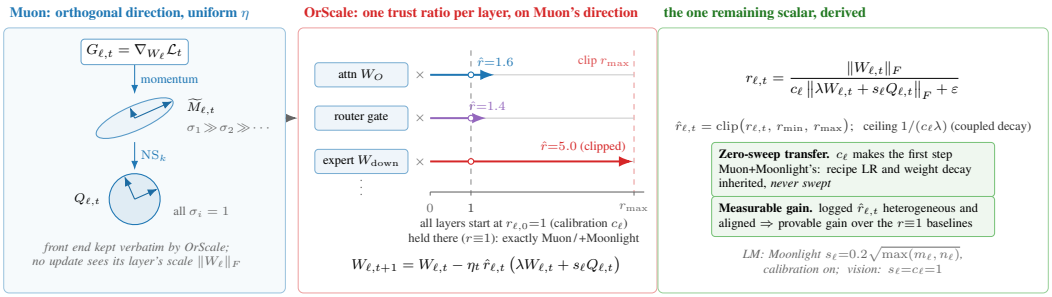


Figure 2: **OrScale at a glance.** *Left:* Muon fixes each update's direction at the polar factor of its momentum, leaving the per-layer magnitude to one layer-blind global η. *Center:* OrScale supplies that magnitude—one clipped trust ratio $\hat{r}_{\ell,t}$ per matrix layer, on the unchanged direction. Arrow lengths are ratios logged in the 16B MoE run at 1.7B tokens, already spread 1.4–5.0 across layer classes with the expert projection at the clip ceiling (Section 5.3; per-class colors and full trajectories in Figure 4); the calibration c_ℓ starts every layer at the r=1 guide (Definition 2), where $r_{\min}$=$r_{\max}$=1 recovers Muon and Muon+Moonlight exactly. *Right:* the ratio—relative-step numerator over applied-direction denominator (Definition 1), bounded through the coupled decay (Lemma 4)—with its gain condition measurable from the logged ratios (Section 4).

scalar should regulate the relative step $\|\Delta W_\ell\|_F / \|W_\ell\|_F$, whose drift across layers is precisely what LARS observed to make layers stall or destabilize (You et al., 2017); holding $\|\eta \rho_\ell D_\ell\|_F / \|W_\ell\|_F$ near a common target gives $\rho_\ell \propto \|W_\ell\|_F / \|D_\ell\|_F$—a trust ratio with numerator $\|W_\ell\|_F$ and denominator the norm of the applied direction. The choice of denominator is where Muon parts ways with SGD and Adam: the ratio should measure what is actually subtracted, and for a Muon step with shape factor s_ℓ and coupled weight decay the direction is $D_{\ell,t} = \lambda W_{\ell,t} + s_\ell Q_{\ell,t}$.

Definition 1 (Real-update-direction trust ratio)**.** Let $D_{\ell,t}$ be the parameter-space direction the optimizer is about to subtract from $W_{\ell,t}$ (before the global learning rate and the trust-ratio scalar). The trust ratio is $r_{\ell,t} = \|W_{\ell,t}\|_F / \big(\|D_{\ell,t}\|_F + \varepsilon\big)$.

For SGD, $D = G + \lambda W$ and Definition 1 *is* LARS. After orthogonalization, however, the choice of denominator acquires content, and the reason is units: the pre-orthogonalization momentum $\widetilde{M}_\ell$ lives in gradient units—its norm set by the loss scale and the noise level—whereas the direction Muon actually subtracts lives in parameter units, so dividing the weight norm by the momentum norm measures no relative step at all, but a dimensionally mixed quantity that can sit orders of magnitude from one. The post-orthogonalization norm fails in the opposite way: $\|Q_\ell\|_F = \sqrt{\mathrm{rank}\, Q_\ell}$ is a shape constant carrying no update-scale information whatsoever (Lemma 1). Definition 1 restores the parameter-space convention on both sides of the ratio.

Two further considerations make the ratio deployable. First, adopting it must not invalidate an already-tuned learning rate—the Moonlight line's main practical asset is that Muon's update RMS is matched to AdamW's, so one recipe serves both (Liu et al., 2025)—yet the raw ratio of Definition 1 would break exactly that: at initialization $\|W_{\ell,0}\|_F / \|s_\ell Q_{\ell,0}\|_F \approx \mathrm{RMS}(W_{\ell,0})/(0.2) \propto 1/\sqrt{\text{fan-in}}$ sits between 0.11 and 0.26 across the layers of our 125M model, shrinking the effective step several-fold, by a different factor for every layer and width. The fix is a one-time anchor:

Definition 2 (Lazy per-layer calibration)**.** At the first step at which W_ℓ receives a non-zero update—write 0 for that step's index in what follows—store the fp32 scalar $c_\ell = \|W_{\ell,0}\|_F / \big(\|\lambda W_{\ell,0} + s_\ell Q_{\ell,0}\|_F + \varepsilon\big)$, and use the calibrated ratio $r_{\ell,t} = \|W_{\ell,t}\|_F / \big(c_\ell \|\lambda W_{\ell,t} + s_\ell Q_{\ell,t}\|_F + \varepsilon\big)$ thereafter.

The constant records each layer's ratio at initialization, so that the trust ratio thereafter measures *growth relative to initialization* rather than absolute scale. By construction $r_{\ell,0} = 1$ (to within the ε floor) for every layer, shape, width, and initialization: the first parameter step coincides with Muon+Moonlight's, the recipe learning rate and weight decay transfer with no additional sweep, and the ratio is width-invariant at initialization in the Moonlight+μP sense (Corollary 1). Afterwards

Algorithm 1 The OrScale update for matrix-valued parameters. The shape factor s_ℓ and the one-time calibration are configuration choices, not separate methods: language-model training uses Moonlight's s_ℓ with calibration on; the vision benchmark uses $s_\ell = 1$ with calibration off.

Require: LR η_t, momentum $\mu \in [0, 1)$, weight decay $\lambda \geq 0$, NS iterations k (default 5), $\varepsilon = 10^{-6}$; shape factor s_ℓ ($0.2\sqrt{\max(m_\ell, n_\ell)}$ for LM training, else 1); calibration on/off (on for LM training); clip bounds $r_{\min}, r_{\max}$ (defaults $0.1, 5.0$ for LM training; $0.5, 1.5$ otherwise).

for each matrix parameter $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$ **do**
  $G_{\ell,t} \leftarrow \nabla_{W_\ell} \mathcal{L}_t$; $M_{\ell,t} \leftarrow \mu M_{\ell,t-1} + G_{\ell,t}$; $\widetilde{M}_{\ell,t} \leftarrow \mu M_{\ell,t} + G_{\ell,t}$
  $Q_{\ell,t} \leftarrow \mathrm{NS}_k(\widetilde{M}_{\ell,t})$ ▷ approximate polar factor
  if calibration on **and** c_ℓ uninitialized **then**
    $c_\ell \leftarrow \|W_{\ell,t}\|_F / (\|\lambda W_{\ell,t} + s_\ell Q_{\ell,t}\|_F + \varepsilon)$ ▷ calibration anchor (Definition 2)
  else if calibration off **then**
    $c_\ell \leftarrow 1$
  end if
  $D_{\ell,t} \leftarrow \lambda W_{\ell,t} + s_\ell Q_{\ell,t}$ ▷ applied direction, WD coupled
  $r_{\ell,t} \leftarrow \|W_{\ell,t}\|_F / (c_\ell \, \|D_{\ell,t}\|_F + \varepsilon)$ ▷ fp32 cast around the division
  $\hat{r}_{\ell,t} \leftarrow \mathrm{clip}(r_{\ell,t}, r_{\min}, r_{\max})$; $W_{\ell,t+1} \leftarrow W_{\ell,t} - \eta_t \, \hat{r}_{\ell,t} \, D_{\ell,t}$
end for
Update non-matrix parameters with the AdamW fallback.

it moves as $r_{\ell,t} \approx \|W_{\ell,t}\|_F / \|W_{\ell,0}\|_F$ in early training (Lemma 5)—the LARS-style weight-norm signal, now referenced to each layer's own initialization.

Second, the adaptation should be bounded. A weight-norm numerator creates a potential feedback loop—growing weights raise r_ℓ, amplifying the step, growing the weights further—and decoupled weight decay (Loshchilov & Hutter, 2019) does not close it, because its pull-back $(1 - \eta\lambda)W$ does not scale with r_ℓ. Coupling the decay into the scaled update—putting λW_ℓ *inside* $D_{\ell,t}$, as Definition 1 already requires—does: as $\|W_\ell\|_F$ grows the denominator becomes $\approx c_\ell \lambda \|W_\ell\|_F$ and the ratio approaches the finite ceiling $1/(c_\ell \lambda)$ (Lemma 4), while a conventional clip $\hat{r}_{\ell,t} = \mathrm{clip}(r_{\ell,t}, r_{\min}, r_{\max})$ guards the transient before the ceiling engages. Section 5.5 deletes each design choice in turn and reports what breaks.

3.2 The resulting algorithm

The update, written out with the calibration c_ℓ of Definition 2, is

$$r_{\ell,t} = \frac{\|W_{\ell,t}\|_F}{c_\ell \, \|\lambda W_{\ell,t} + s_\ell Q_{\ell,t}\|_F + \varepsilon}, \qquad W_{\ell,t+1} = W_{\ell,t} - \eta_t \, \hat{r}_{\ell,t} \, (\lambda W_{\ell,t} + s_\ell Q_{\ell,t}). \tag{2}$$

For language-model training we instantiate OrScale with Moonlight's shape factor $s_\ell = 0.2\sqrt{\max(m_\ell, n_\ell)}$ and enable the calibration, following Moonlight's RMS-matching convention so that the AdamW-tuned recipe—learning rate and weight decay alike—carries over unchanged; the weight-decay coupling inside $D_{\ell,t}$ is part of the core update in both configurations, not an LM-specific choice. On the CIFAR-10 benchmark, where the standard protocol sweeps every optimizer family over its own learning-rate grid, there is no tuned recipe to preserve, so we run Equation (2) with $s_\ell = 1$ and $c_\ell = 1$: the LARS template with the substitution $G \to Q$.

Cost, state, and degenerate cases. With calibration on, OrScale adds one fp32 scalar per matrix layer (none otherwise) and two Frobenius norms per step, both negligible next to Newton–Schulz; measured wall-clock at 16B is at parity with Muon+Moonlight (Table 2). Algorithm 1 contains the baselines as limits: $r_{\min} = r_{\max} = 1$ with $s_\ell = 1$ is Muon; $r_{\min} = r_{\max} = 1$ with the Moonlight s_ℓ is Muon+Moonlight; $\lambda = 0$ removes weight decay while keeping the scaling.

4 Convergence theory

Full proofs are in Section A. We minimize $f(W) = \mathbb{E}_s[\ell(W, s)]$ over h matrix-valued blocks $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$; write $p_\ell = \min(m_\ell, n_\ell)$ and $P = \sum_\ell p_\ell$. Assumptions: **(A1)** layerwise smoothness with constants L_ℓ, $\bar{L}_p = (\sum_\ell L_\ell p_\ell)/P$; **(A2)** bounded variance, $\mathbb{E}\|G_{\ell,t} - \nabla_\ell f\|_F^2 \leq \sigma_\ell^2 / b$; **(A3)** bounded nuclear noise, $\mathbb{E}\|\xi_{\ell,t}\|_* \leq \sqrt{p_\ell}\, \sigma_\ell / \sqrt{b}$ (implied by (A2)); **(A4)** the headline statements set $\mu = 0$,

and Theorem 1 additionally $\lambda = 0$ ($\lambda > 0$ restored in Proposition 1; the $\mu > 0$ extension follows LAMB (You et al., 2020)). Convergence criterion: $\Psi(W) = \frac{1}{\sqrt{P}} \sum_\ell \|\nabla_\ell f(W)\|_*$.

Two properties of polar factors underpin the analysis (Lemma 1): the inner-product identity $\langle A, Q_A \rangle = \|A\|_*$ with its dual bound $|\langle B, Q_A \rangle| \le \|B\|_*$, and the norm identity $\|Q_A\|_F = \sqrt{\operatorname{rank} A}$. Together these yield an expected-descent bound: $\mathbb{E}\langle \nabla_\ell f, Q_{\ell,t} \rangle \ge \|\nabla_\ell f\|_* - \sqrt{p_\ell}\, \sigma_\ell / \sqrt{b}$ (Lemma 2).

Theorem 1 (Rate for any clipped multiplier)**.** *Under (A1)–(A4), for any update $W_{\ell,t+1} = W_{\ell,t} - \eta \hat{r}_{\ell,t} Q_{\ell,t}$ with $\hat{r}_{\ell,t} \in [r_{\min}, r_{\max}]$ arbitrary (possibly sample-dependent), the stepsize $\eta = \sqrt{2(f_0 - f^*)/(r_{\max}^2 T \sum_\ell L_\ell p_\ell)}$ and W_a uniform on $\{W_1, \dots, W_T\}$ give*

$$\big(\mathbb{E}\, \Psi(W_a)\big)^2 \;\le\; \frac{4 r_{\max}^2 \,(f_0 - f^*)\, \bar{L}_p}{r_{\min}^2 \, T} \;+\; \frac{2 r_{\max}^2}{r_{\min}^2} \cdot \frac{\big(\sum_\ell \sqrt{p_\ell}\, \sigma_\ell\big)^2}{P\, b}.$$

Theorem 1 treats $\hat{r}$ as a black box: it guarantees convergence for any clipped multiplier but cannot distinguish OrScale from Muon. The comparison requires a finer quantity. Let $a_\ell = \|\nabla_\ell f\|_*$ and $b_\ell = L_\ell p_\ell$, and define the *descent functional* $\Phi(\hat{r}) = \big(\sum_\ell \hat{r}_\ell a_\ell\big)^2 / \sum_\ell b_\ell \hat{r}_\ell^2$ for per-layer multipliers $\hat{r}$; Muon corresponds to $\hat{r} \equiv 1$. The *effective gain* is $\kappa_{\text{eff}}(\hat{r}) = \Phi(\hat{r})/\Phi(\mathbf{1})$.

Theorem 2 (Layer-adaptive headroom)**.** $\max_{\hat{r}} \Phi(\hat{r}) = \sum_\ell a_\ell^2 / b_\ell$ *at* $\hat{r}_\ell^* \propto a_\ell / b_\ell$*, and* $\kappa_{\text{layer}} := \Phi(\hat{r}^*)/\Phi(\mathbf{1}) \ge 1$ *with equality iff a_ℓ / b_ℓ is constant across layers. Geometrically, with $x_\ell = a_\ell / \sqrt{b_\ell}$ and $y_\ell = \hat{r}_\ell \sqrt{b_\ell}$:* $\kappa_{\text{eff}}(\hat{r}) = \cos^2 \theta_{xy} / \cos^2 \theta_{x,\sqrt{b}}$ *(Lemma 3)—an optimizer beats Muon iff its weighted multipliers align better with the optimum than uniform scaling does.*

A two-layer example makes the scaling concrete: with $L_2 = \beta L_1$ and $a_2 = \alpha a_1$, $\kappa_{\text{layer}} = (1+\beta)(1+\alpha^2/\beta)/(1+\alpha)^2$ evaluates to 1.67, 3.03, 9.10 at $(\alpha, \beta) \in \{(0.1, 1), (1, 10), (0.1, 10)\}$—*the headroom grows with cross-layer heterogeneity*. Whether OrScale captures any of this headroom depends on two statistics of its logged multipliers: **(S1)** heterogeneity, $\nu_{\hat{r}} = \operatorname{Var}_\ell[\hat{r}_\ell \sqrt{b_\ell}] / \mathbb{E}_\ell[\hat{r}_\ell \sqrt{b_\ell}]^2 > 0$; and **(S2)** alignment, $\rho = \operatorname{corr}_\ell(\hat{r}_\ell \sqrt{b_\ell},\, x_\ell) > 0$.

Theorem 3 (OrScale-specific gain)**.** *Under (A1)–(A4), (S1), (S2),* $\kappa_{\text{eff}}^{\text{OrScale}} \ge 1 + \rho^2 \nu_{\hat{r}} \, \dfrac{1 + \nu_{\sqrt{b}}}{1 + \nu_{\sqrt{b}}/\rho^2} - \delta(\lambda)$*, where* $\nu_{\sqrt{b}} = \operatorname{Var}_\ell[\sqrt{b_\ell}] / \mathbb{E}_\ell[\sqrt{b_\ell}]^2$ *and* $\delta(\lambda) = O(\lambda^2 \|W\|_F^2 / \|Q\|_F^2)$.

Theorem 4 (Collapse of unit-inconsistent ratios)**.** *If the clipped ratio saturates ($\hat{r}_{\ell,t} = r_{\max}$ for all ℓ at almost every step—the logged behavior of the raw-momentum denominator variants, Section 5.5), then $\Phi(r_{\max}\mathbf{1}) = \Phi(\mathbf{1})$ and $\kappa_{\text{eff}} = 1$: a uniformly saturated multiplier is Muon with a relabeled learning rate.*

A falsifiable prediction. Theorems 2 and 3 reduce the gain over Muon to two measurable statistics: heterogeneity and alignment of the logged trust ratios. On CIFAR-10/DavidNet diagnostics, $\nu_{\hat{r}} \approx 0.3$, $\rho \approx 0.5$, $\nu_{\sqrt{b}} \approx 0.5$ give $\kappa_{\text{eff}} \in [1.06, 1.15]$, consistent with the observed $+0.35$-point gain (Section 5.6). The same logic predicts *where* the gain should grow: architectures whose layer classes see maximally different update statistics. Mixture-of-experts transformers are the sharpest available case—routers, sparsely-activated experts, and dense attention blocks differ in activation frequency, gradient scale, and curvature—so (S1)–(S2) should strengthen, and with them the gap over Muon+Moonlight; the prediction was recorded before the 16B-A3B run was launched, and Section 5.3 reports the test. Corollary 1 collects the calibrated update's remaining properties (anchor exactness, width-invariance, weight-norm ceiling, early-training LARS signal).

5 Experiments

5.1 Setup and the zero-sweep protocol

Models and data. Language models are trained on FineWeb-Edu (Penedo et al., 2024) in two regimes: *dense*, GPT-2-style pre-norm transformers (Radford et al., 2019) (RMSNorm, RoPE, ReLU^2 MLP, tied embeddings, vocabulary 50,304) at 125M/399M/545M/1.1B following the Moonlight scaling recipe (Liu et al., 2025); and *MoE*, the Moonlight-16B-A3B architecture (DeepSeek-V3-style (DeepSeek-AI, 2024): 27 layers, multi-head latent attention, 64 routed experts with top-6

Table 1: FineWeb-Edu dense pre-training: final validation cross-entropy (lower is better; **bold** best per row). All optimizers share the Moonlight-recipe LR shown per row (125M baselines: best of a four-point grid; OrScale everywhere: recipe LR, unswept). Cells with $\pm\sigma$ are means over three seeds; others are single-seed.

Scale	Tokens	Recipe LR	AdamW	Muon+Moonlight	**OrScale** (ours)
125M	5.24B	1.0×10^{-3}	3.3721	3.2319	$\mathbf{3.2120 \pm 0.0017}$
399M	8.92B	9.503×10^{-4}	2.9966	$\mathbf{2.9198 \pm 0.0016}$	2.9223 ± 0.0022
545M	14.04B	9.143×10^{-4}	2.9235	2.8141 ± 0.0013	$\mathbf{2.8052 \pm 0.0013}$
1.1B	28.54B	8.561×10^{-4}	2.7304	2.6360	**2.6251**

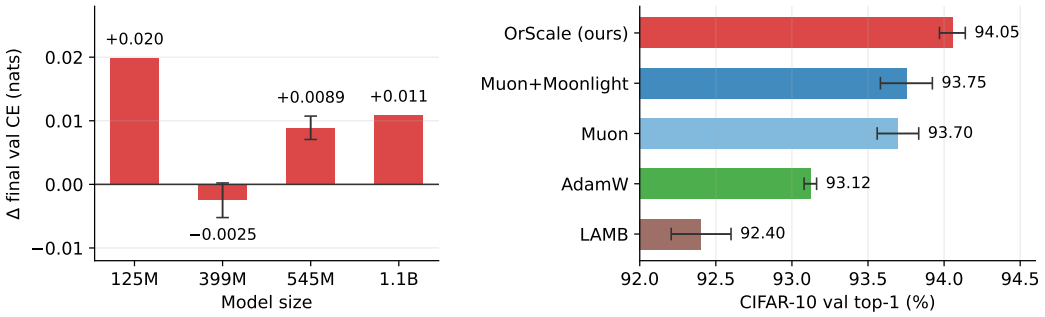


Figure 3: *Left:* per-scale gap (Muon+Moonlight − OrScale) on dense FineWeb-Edu; positive means OrScale better. Values match Table 1: three-seed means at 399M/545M and for OrScale at 125M (error bars where both sides have seed replicates: per-side stds combined in quadrature); single-seed elsewhere; the 399M cell is within noise and reported as a tie. *Right:* CIFAR-10/DavidNet val top-1 at each optimizer family's best learning rate (three seeds, $\pm 1\sigma$, per-family grids); OrScale runs the vision configuration (s_ℓ=1, calibration off; Section 5.6, full sweep in Section G).

routing plus 2 shared experts, aux-loss-free load balancing; 15.96B total, 2.24B active parameters) on 10B tokens at sequence length 8192. Vision cross-check: CIFAR-10 on the 24-epoch DavidNet speedrun (≈6.5M parameters), three seeds per cell. Complete per-scale settings are in Section D, compute in Section H.

The zero-sweep protocol. Our evaluation deliberately stresses the calibration anchor. At 399M, 545M, 1.1B, and 16B-A3B, *every* optimizer runs at the Moonlight-recipe learning rate for that scale (9.503×10^{-4}, 9.143×10^{-4}, 8.561×10^{-4}, 4.2×10^{-4}), with weight decay 0.1 and momentum 0.95: the recipe is tuned for AdamW, shared by Muon+Moonlight via RMS matching, and inherited by OrScale via the calibration anchor. At 125M the baselines were additionally tuned over a four-point LR grid and report their best cell. *OrScale was never swept at any language-model scale in this paper*; its clip bounds stay at the defaults $(0.1, 5.0)$ throughout. Any OrScale win therefore comes with zero tuning advantage—if anything the protocol favors the baselines.

5.2 Dense scaling: 125M–1.1B

Table 1 spans a $48\times$ compute range ($0.046 \rightarrow 2.18$ PFLOP-days). Against Muon+Moonlight, OrScale wins at 125M ($+0.020$), 545M ($+0.0089 \pm 0.0018$ over three seeds—significant), and 1.1B ($+0.011$), and ties at 399M (-0.0025 ± 0.0027—within noise, reported as a tie); Figure 3 (left) plots the gaps at the seed-matched protocol. Against AdamW it wins everywhere by $+0.074$ to $+0.160$ nats. Fitted log–log exponents (-0.054/-0.053/-0.052; Section F) show the advantage on *dense* models is scale-stable rather than growing—the theory attributes the gain to heterogeneity, and a homogeneous stack of identical blocks offers only a fixed amount of it; the prediction is tested where heterogeneity is large, next. A matched 125M LR sweep (shared grid $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$) shows OrScale at or below Muon+Moonlight at every stable cell, with Muon+Moonlight diverging at the top cell OrScale survives (Section F).

Table 2: Moonlight-16B-A3B MoE, 10B FineWeb-Edu tokens, 8×H20, bf16, ZeRO-1, seed 42. All arms use the Moonlight-recipe settings for this scale (LR 4.2×10^{-4}, wd 0.1, μ=0.95, warmup 60, cosine), shared by AdamW and Muon+Moonlight by construction and inherited unchanged by OrScale.

Optimizer	$r_{\min}/r_{\max}$	Final val CE	Δ vs. Muon+ML	Wall-clock
AdamW	—	5.2002	−1.7287	3.93 d
Muon + Moonlight	—	3.4715	—	4.98 d
OrScale ($r_{\max}$=10)	0.1/10	3.3556	+0.1159	4.92 d
OrScale	0.1/5	**3.3410**	+0.1305	4.92 d

Table 3: Calibration ablation (single-seed log values; three-seed means in Table 1). The uncalibrated variant runs Definition 1 with the Moonlight shape factor but no anchor (c_ℓ=1) and must retune its matrix LR upward to train well; calibrated OrScale runs at the untouched recipe LR everywhere.

Scale	Variant	Matrix LR	Final val CE
399M	Muon+Moonlight (reference)	recipe (9.503×10^{-4})	2.9183
399M	OrScale, *uncalibrated*	2.5× recipe (2.376×10^{-3})	2.9939
399M	**OrScale** (calibrated)	recipe, unswept	**2.9247**
125M	OrScale, *uncalibrated*	3× base (3×10^{-3}, tuned)	3.2232
125M	**OrScale** (calibrated)	base (10^{-3}), unswept	**3.2120**
125M	OrScale, calibration in bf16	base (10^{-3})	3.2158

5.3 16B-A3B MIXTURE-OF-EXPERTS: THE HETEROGENEITY STRESS TEST

Figure 1 and Table 2 report the headline test. At 14.5× the parameter count of our largest dense run, with every hyperparameter inherited verbatim, OrScale improves on Muon+Moonlight by 0.130 nats (3.8% relative)—an order of magnitude beyond the dense-regime gaps. OrScale leads at every validation checkpoint from 2.5B tokens onward and passes Muon+Moonlight's *final* loss with a third of the token budget to spare; wall-clock is at parity (425,018 s vs. 430,328 s), replacing the usual <1%-overhead assertion with a measurement. The $r_{\max}$=10 arm doubles the clip ceiling and still beats Muon+Moonlight while losing to $r_{\max}$=5: the win is not an artifact of a tight clip, and the coupled-WD ceiling is doing real work. AdamW at the shared recipe trails by 1.73 nats, consistent with the Moonlight paper's own AdamW-vs-Muon comparison at this architecture; we report it as the recipe-for-recipe reference point (no arm received a per-optimizer retune).

The theory's conditions, measured. The gain condition of Theorem 3 is checkable from logs, and this run checks it: the logged trust ratios separate cleanly and persistently by layer class—routed-expert projections at the ceiling within 1.5B tokens, router gates near 2.2–2.3 and never clipping, attention in between—exactly the heterogeneity (S1) requires, aligned with the layer classes whose update statistics differ most, as (S2) requires. The late-training clip fraction (33% of layers at a bound) is the coupled-WD ceiling engaging for the layer class that wants the largest relative steps, not the uniform saturation pathology of Theorem 4. Per-layer trajectories, calibration-constant statistics, and the weaker dense-125M counterparts of these diagnostics are in Section E (Figure 4).

5.4 CALIBRATION IS LOAD-BEARING

If the calibration constant were cosmetic, removing it would cost a learning-rate sweep and nothing else. Table 3 shows it costs more: the uncalibrated ratio needs a 2.5–3× matrix-LR retune to train well (its uncorrected $r_{\ell,0} \propto 1/\sqrt{\text{fan-in}}$ shrinks early steps several-fold) and *still* finishes 0.069 nats behind the calibrated variant at 399M—behind even the Muon+Moonlight baseline. Calibration converts the trust ratio from a knob that must be re-tuned per model into a drop-in that inherits the recipe, and it is also worth loss. Computing the anchor in bf16 rather than fp32 costs a small but systematic 0.004 nats at 125M; hence the fp32 cast in Algorithm 1.

Table 4: Deleting one design choice at a time; each deletion produces a distinct, logged symptom (accuracy alone can mislead: the top row ties OrScale on CIFAR while its ratio carries no update-scale information). CIFAR: val top-1 at the best LR, three seeds. FW-125M: final CE at each variant's best LR under the instrumented 125M ablation sweep that produced the logged symptoms, single seed; that setup differs from the strict-recipe runs (OrScale there: 3.2120, Table 1), so FW values are comparable within this table only. Symptoms from 125M diagnostics (Section C).

Variant	Omitted	Logged symptom	CIFAR-10 (%)	FW-125M (CE)
$\|Q_\ell\|_F$ denominator	denominator	ratio = shape constant (Lemma 1)	94.06 ± 0.09	3.231
$\|\widetilde{M}_\ell\|_F$ denominator	denominator	$\hat{r}{=}r_{\max}$ on 99.75–100% of steps	93.84 ± 0.15	3.252
$\mathrm{RMS}(\widetilde{M}_\ell)$ + Moonlight	denominator	$\hat{r}{=}r_{\max}$ on 99.5–100% of steps	93.83 ± 0.17	3.230
Uncalibrated ratio	calibration	needs 2.5–3× LR retune (Table 3)	—	3.2232
Decoupled weight decay	WD coupling	10–50× weight-norm growth	93.55 ± 0.12	3.305
OrScale ($s_\ell{=}1$)	—	in-band, layer-structured	94.05 ± 0.08	3.228
OrScale (LM config.)	—	in-band, layer-structured	94.05 ± 0.12	**3.226**

5.5 Design-choice ablation

Table 4 deletes one design choice at a time. The deleted variants coincide with hybrid designs that have been suggested for combining trust ratios with Muon, or that we implemented first ourselves; we report them once, here, as ablations of Section 3.1's design choices rather than as motivation. Dropping the applied-direction denominator either removes adaptivity ($\|Q\|_F$ is a shape constant) or mis-scales the ratio by orders of magnitude so it pins at the clip bound on effectively every step—the empirical premise of Theorem 4, under which $\kappa_{\text{eff}} = 1$ and the "adaptive" method is Muon with a relabeled LR. Dropping the calibration severs learning-rate transfer (Table 3). Decoupling weight decay triggers the runaway feedback loop, with 10–50× weight-norm growth on `mlp` and output projections. Removing the clip entirely is survivable—transient ratios reach 9.6 before the coupled-WD ceiling engages, ending 0.007 nats behind—so the safety-critical boundedness lives in the coupled-WD ceiling, not the clip; at 16B, doubling $r_{\max}$ moves the loss by only 0.015 nats (Table 2).

5.6 Vision cross-check: CIFAR-10 / DavidNet

On the 24-epoch DavidNet speedrun (three seeds, per-family LR grids), OrScale ranks first at $94.05 \pm 0.08\%$ val top-1 (Figure 3, right), improving Muon (93.70 ± 0.14) by $+0.35$ and Muon+Moonlight (93.75 ± 0.17) by $+0.30$ points ($\approx 5\%$ residual-error reduction); AdamW and LAMB trail at 93.12 ± 0.04 and 92.40 ± 0.20. The result cross-validates the derivation outside the LLM regime and supplies the diagnostics behind the numerical κ_{eff} estimate of Section 4; the full sweep, LR-sensitivity analysis, and per-cell table are in Section G.

6 Discussion and conclusion

Muon fixes each update's direction; the remaining per-layer degree of freedom is one scalar per step. We derived that scalar by adapting the LARS/LAMB trust-ratio template to the orthogonalized setting, validating each design choice by a deletion that fails in its own way; the calibration anchor lets a Moonlight-tuned recipe transfer with zero extra sweeps. The theory conditions the gain over Muon on logged statistics rather than oracle constants, predicting growth with architectural heterogeneity—and the sharpest available test agrees: $+0.130$ nats on a 16B-A3B MoE at recipe hyperparameters and parity wall-clock, with the logged ratios exhibiting the layer-class structure the theory requires.

Limitations. The 16B-A3B comparison is single-seed (dense three-seed replication brackets seed noise at ± 0.002–0.003 nats, $\sim 50\times$ below the MoE gap); results are FineWeb-Edu validation cross-entropy without downstream evaluations; claims concern autoregressive next-token pre-training with the Moonlight recipe family, and the dense-regime gains are small (0.009–0.020 nats where significant); what the theory does and does not claim is delimited in Section A.

Ethics statement

This work concerns optimizer design for neural-network pre-training. It involves no human subjects and no new data collection: all experiments use the publicly released FineWeb-Edu corpus and the CIFAR-10 benchmark under their standard terms. The primary ethical consideration is the dual-use nature of training efficiency—OrScale reduces compute, cost, and energy per unit of pre-training progress, and thereby also lowers the barrier to training models for harmful uses; Section I expands on this, and Section H reports the project's full compute footprint. The method introduces no new dataset, surveillance, or model-release pathway beyond the optimizer implementation.

Reproducibility statement

The supplementary material contains the full training and analysis code, one configuration file per reported result row, and the raw training logs for the 16B-A3B runs; every table and figure in the paper regenerates from logged data via the included scripts. Section D specifies architectures, schedules, batch geometry, and every optimizer hyperparameter per scale (none are deferred to external sources); Section H reports hardware and wall-clock; proofs of all statements are in Section A.

AI use statement

Large language models were used as writing and engineering assistants in preparing this manuscript (restructuring and editing text, LaTeX, and log-analysis/plotting scripts). The method, experiments, and theory originate with the authors, who verified all technical claims, numbers, and proofs.

References

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning, 2024. URL https://arxiv.org/abs/2410.21265.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/blob/main/DeepSeek_V4.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. URL https://arxiv.org/abs/1706.02677.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL https://github.com/KellerJordan/modded-nanogpt.

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

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon and further. *arXiv preprint arXiv:2502.02900*, 2025.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training, 2024. URL https://arxiv.org/abs/2305.14342.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. *arXiv preprint arXiv:2502.07529*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon and Scion great again! (Bridging theory and practice of LMO-based optimizers). *arXiv preprint arXiv:2505.13416*, 2025.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer, 2025. URL https://arxiv.org/abs/2507.11005.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin,

Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Chenxiao Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Qizheng Gu, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yang Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Haoyu Lu, Lijun Lu, Yashuo Luo, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Zeyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Lin Sui, Xinjie Sun, Flood Sung, Yunpeng Tai, Heyi Tang, Jiawen Tao, Qifeng Teng, Chaoran Tian, Chensi Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Si Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Haoning Wu, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Jing Xu, Jing Xu, Junjie Yan, Yuzi Yan, Hao Yang, Xiaofei Yang, Yi Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Siyu Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Shaojie Zheng, Longguang Zhong, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2026b. URL https://arxiv.org/abs/2507.20534.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam, 2025. URL https://arxiv.org/abs/2409.11321.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks, 2023. URL https://arxiv.org/abs/2310.02244.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017. URL https://arxiv.org/abs/1708.03888.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2020. URL https://arxiv.org/abs/1904.00962.

A Proofs

A.1 Polar-factor algebra and expected descent

Lemma 1 (Polar-factor algebra). *For $A = U\Sigma V^\top$ of rank $k \le \min(m,n)$ and $Q_A = \mathrm{Ortho}(A) = UV^\top$: (i) $\langle A, Q_A\rangle = \|A\|_*$; (ii) $|\langle B, Q_A\rangle| \le \|B\|_*$ for any B; (iii) $\|Q_A\|_F = \sqrt{k}$ and $\|Q_A\|_{\mathrm{op}} = 1$.*

Proof. (i) $\langle A, Q_A\rangle = \mathrm{tr}(V\Sigma U^\top UV^\top) = \mathrm{tr}(\Sigma) = \|A\|_*$. (ii) Von Neumann's trace inequality gives $|\mathrm{tr}(B^\top Q_A)| \le \sum_i \sigma_i(B)\,\sigma_i(Q_A) \le \sum_{i\le k}\sigma_i(B) \le \|B\|_*$ since all singular values of Q_A are 1 (first k) or 0. (iii) immediate from the singular values of Q_A. □

Lemma 2 (Expected descent). *Let $G_\ell = \nabla_\ell f + \xi_\ell$ with $\mathbb{E}\|\xi_\ell\|_F^2 \le \sigma_\ell^2/b$ and $Q_\ell = \mathrm{Ortho}(G_\ell)$. Then $\mathbb{E}\langle \nabla_\ell f, Q_\ell\rangle \ge \|\nabla_\ell f\|_* - \sqrt{p_\ell}\,\sigma_\ell/\sqrt{b}$.*

Proof. $\langle \nabla_\ell f, Q_\ell\rangle = \langle G_\ell, Q_\ell\rangle - \langle \xi_\ell, Q_\ell\rangle = \|G_\ell\|_* - \langle \xi_\ell, Q_\ell\rangle$ by Lemma 1(i). By Jensen on the convex norm, $\mathbb{E}\|G_\ell\|_* \ge \|\mathbb{E}G_\ell\|_* = \|\nabla_\ell f\|_*$. By Lemma 1(ii), $|\langle \xi_\ell, Q_\ell\rangle| \le \|\xi_\ell\|_* \le \sqrt{p_\ell}\|\xi_\ell\|_F$, so $\mathbb{E}|\langle \xi_\ell, Q_\ell\rangle| \le \sqrt{p_\ell}\sqrt{\mathbb{E}\|\xi_\ell\|_F^2} \le \sqrt{p_\ell}\sigma_\ell/\sqrt{b}$ by Cauchy–Schwarz. □

A.2 Proof of Theorem 1

Step 1 (smoothness). Layerwise smoothness (A1) with $W_{\ell,t+1} - W_{\ell,t} = -\eta\hat{r}_{\ell,t}Q_{\ell,t}$ gives

$$f(W_{t+1}) \le f(W_t) - \eta\sum_\ell \hat{r}_{\ell,t}\,\langle \nabla_\ell f(W_t), Q_{\ell,t}\rangle + \frac{\eta^2}{2}\sum_\ell L_\ell\,\hat{r}_{\ell,t}^2\,\|Q_{\ell,t}\|_F^2.$$

Step 2 (second order). $\|Q_{\ell,t}\|_F^2 \le p_\ell$ (Lemma 1(iii)) and $\hat{r} \le r_{\max}$ bound the last term by $\frac{\eta^2 r_{\max}^2}{2}\sum_\ell L_\ell p_\ell$.

Step 3 (first order, no sign assumption). Write $x_{\ell,t} = \langle \nabla_\ell f(W_t), Q_{\ell,t}\rangle$ and let $x^- = \max(0,-x)$ denote the negative part. The sign of $x_{\ell,t}$ is *not* controlled—when noise dominates, $Q_{\ell,t}$ can be an ascent direction—and $\hat{r}_{\ell,t}$ may depend on the same sample, so $\hat{r}_{\ell,t}x_{\ell,t} \ge r_{\min}x_{\ell,t}$ is false in general. Instead we use the pointwise inequality, valid for any $\hat{r} \in [r_{\min}, r_{\max}]$:

$$\hat{r}_{\ell,t}\,x_{\ell,t} \;\ge\; r_{\min}\,x_{\ell,t} \;-\; (r_{\max}-r_{\min})\,x_{\ell,t}^-,$$

verified by inspection in the two sign cases ($x \ge 0$: $\hat{r}x \ge r_{\min}x$; $x < 0$: $\hat{r}x \ge r_{\max}x = r_{\min}x - (r_{\max}-r_{\min})x^-$).

Step 4 (the negative part is pure noise). By Lemma 1(i)–(ii),

$$x_{\ell,t} = \|G_{\ell,t}\|_* - \langle \xi_{\ell,t}, Q_{\ell,t}\rangle \;\ge\; -\|\xi_{\ell,t}\|_*, \qquad \text{hence} \qquad x_{\ell,t}^- \le \|\xi_{\ell,t}\|_*,$$

and by (A3), $\mathbb{E}[x_{\ell,t}^-] \le \sqrt{p_\ell}\,\sigma_\ell/\sqrt{b}$. Combining Steps 3–4 with Lemma 2,

$$\mathbb{E}\big[\hat{r}_{\ell,t}x_{\ell,t}\big] \;\ge\; r_{\min}\Big(\|\nabla_\ell f\|_* - \frac{\sqrt{p_\ell}\sigma_\ell}{\sqrt{b}}\Big) - (r_{\max}-r_{\min})\frac{\sqrt{p_\ell}\sigma_\ell}{\sqrt{b}} \;=\; r_{\min}\|\nabla_\ell f\|_* - r_{\max}\frac{\sqrt{p_\ell}\sigma_\ell}{\sqrt{b}}.$$

Step 5 (telescope). Substituting into Step 1, taking total expectation, summing over $t = 1,\dots,T$, and rearranging,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_\ell \mathbb{E}\|\nabla_\ell f(W_t)\|_* \;\le\; \frac{f_0 - f^*}{\eta\, r_{\min}\, T} + \frac{r_{\max}}{r_{\min}}\cdot\frac{\sum_\ell \sqrt{p_\ell}\,\sigma_\ell}{\sqrt{b}} + \frac{\eta\, r_{\max}^2}{2r_{\min}}\sum_\ell L_\ell p_\ell.$$

Step 6 (stepsize). With $\eta = \sqrt{2(f_0 - f^*)/(r_{\max}^2 T\sum_\ell L_\ell p_\ell)}$ the first and third terms balance at $\frac{r_{\max}}{r_{\min}}\sqrt{2(f_0-f^*)\sum_\ell L_\ell p_\ell/T}$. Dividing by $\sqrt{P}$ to convert to the Ψ criterion, squaring, and using $(u+v)^2 \le 2u^2 + 2v^2$ yields the claim. □

Remark 1 (Sign-free first-order bound)**.** Because $\hat{r}_{\ell,t}$ may depend on the current sample and the sign of $\langle \nabla_\ell f, Q_{\ell,t} \rangle$ is not controlled, the first-order term is bounded via the pointwise inequality $\hat{r}x \geq r_{\min} x - (r_{\max} - r_{\min})\, x^-$ with $x^- \leq \|\xi_{\ell,t}\|_*$ (Step 3 above). This inflates only the noise term by the factor $r_{\max}/r_{\min}$ visible in Theorem 1, leaving the $O(1/\sqrt{T})$ rate unchanged. The factor is the worst case over arbitrary in-band multipliers—the price of a single statement covering both well-behaved and saturation-prone ratio designs (Section 5.5). No sign assumption on $\langle \nabla_\ell f, Q_{\ell,t} \rangle$ is needed, and downstream statements (Theorems 2 to 4 and Corollary 1) are unaffected, as none consume Theorem 1's constant.

A.3 Coupled weight decay

Proposition 1 (Coupled-WD extension)**.** *Under (A1)–(A3) and the update $W_{\ell,t+1} = W_{\ell,t} - \eta \hat{r}_{\ell,t}(\lambda W_{\ell,t} + s_\ell Q_{\ell,t})$ with $\lambda > 0$, the descent inequality of Section A.2 acquires a first-order term $-\eta \hat{r}_{\ell,t} \lambda \langle \nabla_\ell f, W_{\ell,t} \rangle$ and second-order contributions bounded via $\lambda^2 \|W_{\ell,t}\|_F^2 + 2\lambda s_\ell \langle W_{\ell,t}, Q_{\ell,t} \rangle$. Along OrScale trajectories $\|W_{\ell,t}\|_F$ is bounded by the ceiling mechanism of Lemma 4, so these are $O(\lambda)$ corrections to the constants and the $O(1/\sqrt{T})$ rate is preserved.*

Proof sketch. Substitute the coupled update into layerwise smoothness and split the cross terms; the bookkeeping follows the LAMB weight-decay extension (You et al., 2020, App. B) with the nuclear-norm criterion in place of the ℓ_1 criterion, using Lemma 1(ii) to bound $\langle W, Q \rangle$ by $\|W\|_*$. □

A.4 Layer-adaptive headroom and the geometric identity

Proof of Theorem 2. $\Phi(\hat{r}) = \langle x, y \rangle^2 / \|y\|^2$ with $x_\ell = a_\ell/\sqrt{b_\ell}$, $y_\ell = \hat{r}_\ell \sqrt{b_\ell}$; by Cauchy–Schwarz $\Phi(\hat{r}) \leq \|x\|^2 = \sum_\ell a_\ell^2 / b_\ell$ with equality iff $y \parallel x$, i.e. $\hat{r}_\ell \propto a_\ell / b_\ell$. Muon's value is $\Phi(\mathbf{1}) = (\sum_\ell a_\ell)^2 / \sum_\ell b_\ell$, and $\kappa_{\text{layer}} = \|x\|^2 \|\sqrt{b}\|^2 / \langle x, \sqrt{b} \rangle^2 \geq 1$ with equality iff $x \parallel \sqrt{b}$, i.e. a_ℓ / b_ℓ constant. □

Lemma 3 (Geometric form)**.** $\kappa_{\text{eff}}(\hat{r}) = \cos^2 \theta_{xy} / \cos^2 \theta_{x,\sqrt{b}}$*, directly from* $\Phi(\hat{r}) = \|x\|^2 \cos^2 \theta_{xy}$.

A.5 Proof of Theorem 3

By Lemma 3 it suffices to lower-bound $\cos^2 \theta_{xy} / \cos^2 \theta_{x,\sqrt{b}}$. Decompose

$$y = \alpha \sqrt{b} + \beta \big(x - \pi_{\sqrt{b}}(x) \big) + \rho_\perp,$$

where $\pi_{\sqrt{b}}$ is the orthogonal projection onto $\operatorname{span}(\sqrt{b})$ and $\rho_\perp \perp \operatorname{span}(\sqrt{b}, x)$. The correlation $\rho = \operatorname{corr}_\ell(y_\ell, x_\ell)$ fixes $\beta = \rho \, \|y - \pi_{\sqrt{b}}(y)\| / \|x - \pi_{\sqrt{b}}(x)\| \cdot (1 + o(1))$ after centering against $\sqrt{b}$; (S2) gives $\beta > 0$. (S1) lower-bounds the off-$\sqrt{b}$ mass of y: $\|y - \pi_{\sqrt{b}}(y)\|^2 / \|\pi_{\sqrt{b}}(y)\|^2 \geq \nu_{\hat{r}} / (1 + \nu_{\sqrt{b}})$ by the variance decomposition of $y_\ell = \hat{r}_\ell \sqrt{b_\ell}$ against the direction $\sqrt{b}$ whose own dispersion is $\nu_{\sqrt{b}}$. Substituting the decomposition into $\cos^2 \theta_{xy} = \langle x, y \rangle^2 / (\|x\|^2 \|y\|^2)$, expanding $\langle x, y \rangle = \alpha \langle x, \sqrt{b} \rangle + \beta \|x - \pi_{\sqrt{b}}(x)\|^2$, and discarding the $\rho_\perp$ mass from the numerator while keeping it in $\|y\|^2$ (the worst case, controlled by ρ through $\|\rho_\perp\|^2 \leq (1 - \rho^2) \|y - \pi_{\sqrt{b}}(y)\|^2$) yields

$$\cos^2 \theta_{xy} \;\geq\; \cos^2 \theta_{x,\sqrt{b}} \Big(1 + \rho^2 \nu_{\hat{r}} \, \frac{1 + \nu_{\sqrt{b}}}{1 + \nu_{\sqrt{b}} / \rho^2} \Big) - \delta(\lambda),$$

where $\delta(\lambda) = O(\lambda^2 \|W\|_F^2 / \|Q\|_F^2)$ collects the perturbation of y induced by the λW term inside the denominator (which shifts $\hat{r}_\ell$ from its $\lambda = 0$ expression by a relative $O(\lambda \|W_\ell\|_F / \|s_\ell Q_\ell\|_F)$). Dividing by $\cos^2 \theta_{x,\sqrt{b}}$ gives the claim. □

A.6 Collapse and calibration

Proof of Theorem 4. If $\hat{r}_{\ell,t} = r_{\max}$ for every ℓ, then $y = r_{\max} \sqrt{b}$ is parallel to $\sqrt{b}$, so $\cos^2 \theta_{xy} = \cos^2 \theta_{x,\sqrt{b}}$ and $\kappa_{\text{eff}} = 1$ by Lemma 3: a uniform multiplier cancels in Φ. □

Even a hypothetical re-choice of $r_{\max}$ that brings the raw-momentum ratio in-band does not restore the gain: on FineWeb-Edu 125M diagnostics the cross-layer dispersion of $\|W_\ell\|_F/\|\widetilde{M}_\ell\|_F$ at any fixed step is a factor <2–3 (its variation is dominated by *time*, not layers), giving $\nu_{\hat{r}} \lesssim 0.05$ and $\kappa_{\text{eff}} \lesssim 1.005$ by Theorem 3's formula.

Corollary 1 (Properties of the calibrated ratio)**.** *With $r_{\ell,t}$ as in Equation* (2)*: (i)* anchor*: $r_{\ell,0} = 1$ to within the ε floor; (ii)* width-invariance*: $r_{\ell,0} = 1$ holds for every (m_ℓ, n_ℓ), and the Moonlight invariant $s_\ell \,\mathrm{RMS}(Q_{\ell,t}) = 0.2$ makes the per-element update RMS at the anchor $0.2\,\eta_t$ regardless of width, so the recipe LR transfers in the Moonlight+μP sense (Liu et al., 2025; Yang et al., 2022); (iii)* ceiling*: $r_{\ell,t} \to 1/(c_\ell\lambda)$ as $\|W_{\ell,t}\|_F \to \infty$ (Lemma 4); (iv)* early-training LARS signal*: when $\lambda\|W_{\ell,t}\|_F \ll s_\ell\sqrt{p_\ell}$, $r_{\ell,t} \approx \|W_{\ell,t}\|_F/\|W_{\ell,0}\|_F$ (Lemma 5).*

Lemma 4 (Ceiling)**.** *As $\|W_{\ell,t}\|_F \to \infty$ with $\|s_\ell Q_{\ell,t}\|_F$ bounded, $r_{\ell,t} \to 1/(c_\ell\lambda)$.*

Proof. $\|\lambda W + s_\ell Q\|_F = \lambda\|W\|_F(1 + O(s_\ell\sqrt{p_\ell}/(\lambda\|W\|_F)))$, so the denominator is $c_\ell\lambda\|W\|_F(1 + o(1))$ and the ratio tends to $1/(c_\ell\lambda)$. □

Lemma 5 (Early-training regime)**.** *When $\lambda\|W_{\ell,t}\|_F \ll s_\ell\sqrt{p_\ell}$, the denominator is $\approx c_\ell\, s_\ell\sqrt{p_\ell}$; substituting the leading-order $c_\ell = \|W_{\ell,0}\|_F/(s_\ell\sqrt{p_\ell})$ from Definition 2 gives $r_{\ell,t} \approx \|W_{\ell,t}\|_F/\|W_{\ell,0}\|_F$.*

What the theorems do and do not say. *Do:* the rate is standard for any in-band multiplier (Theorem 1); the gain over Muon is strict under measurable conditions (Theorem 3); saturated unit-inconsistent ratios provably reduce to Muon (Theorem 4); calibration delivers the anchor, transfer, ceiling, and LARS-signal properties (Corollary 1). *Do not:* the Theorem 3 bound is not claimed sharp, and (S1)–(S2) are regime conditions to be checked on diagnostics—as done in Section 5.3—not universal facts.

B Comparison of guarantee structures

Table 5: Guarantee structure on the optimization term, under each method's headline analysis. "Width-invariant" refers to whether the bound's leading factor is independent of d_{model}.

Optimizer	Smoothness factor	Layer-adaptive gain	Width-invariant
SGD	L_∞	none	yes
LARS (You et al., 2017)	L_{avg} (Frobenius)	implicit	yes
LAMB (You et al., 2020)	$\|L\|_1/\sqrt{h}$	implicit	no ($\propto d_{\text{model}}$)
Muon (Jordan et al., 2024b)	$\bar{L}_p$ (nuclear)	none ($\hat{r} \equiv 1$)	yes
Mismatch-denom. variants	$\bar{L}_p$	$= 1$ under saturation (Theorem 4)	yes
OrScale (c_ℓ=1)	$\bar{L}_p$	$\geq 1 + \rho^2\nu_{\hat{r}}(\cdots) > 1$ (Theorem 3)	yes
OrScale (LM configuration)	$\bar{L}_p$	same bound, inherited (Corollary 1)	yes (Corollary 1(ii))

Relative to the Muon convergence literature (Section 2): Gluon-style analyses (Riabinin et al., 2025) obtain layer-wise stepsizes from (L_ℓ^0, L_ℓ^1) smoothness constants, which are sharper but unobservable; Theorem 3 is deliberately conditioned on statistics of the logged $\hat{r}_{\ell,t}$, trading tightness for a condition that any practitioner can check on their own run.

C Design-choice ablation details

This appendix expands the symptoms summarized in Table 4; variant names in the released code are noted for traceability. In the released code, configs, and logs, the LM-configuration arms of the main experiments carry the key `orscale_lm`.

Post-orthogonalization denominator (omitting applied-direction norm) (code: `orscale_muon`/FM1 lineage). With Q_ℓ the polar factor, $\|Q_\ell\|_F = \sqrt{\text{rank}\, Q_\ell} \approx \sqrt{p_\ell}$ (Lemma 1), so $\|W_\ell\|_F/\|Q_\ell\|_F$ depends on weight norm and shape only. The variant can still score on CIFAR—weight-RMS variation supplies a residual per-layer signal—but the ratio carries no information about the applied update, and any adaptive interpretation is unfounded.

Raw-momentum denominator (omitting applied-direction norm) (code: `mutrust`, `muscale`). The numerator has weight units; $\|\widetilde{M}_\ell\|_F$ has gradient units and is orders of magnitude smaller. On FineWeb-Edu 125M with $(r_{\min}, r_{\max}) = (0.5, 1.5)$, the raw cross-layer mean of $\|W\|_F/\|\widetilde{M}\|_F$ spans 24.5–19,128 across the LR grid, and the clipped ratio sits at $r_{\max}$ on 99.75–100% (Frobenius form) and 99.5–100% (RMS/Moonlight form) of diagnostic steps. Theorem 4 then applies: the multiplier is uniform, $\kappa_{\text{eff}} = 1$, and the method is Muon with a relabeled learning rate—which is exactly how it scores.

Uncalibrated ratio (omitting calibration anchor) (code: `orscale_muon_moonlight`). Section 5.4; the uncorrected anchor $r_{\ell,0} \propto 1/\sqrt{\text{fan-in}} \in [0.11, 0.26]$ across our 125M layers.

Decoupled weight decay (omitting WD coupling) (code: FM3 lineage). With the calibrated denominator but a decoupled pull-back $(1 - \eta\lambda)W$, FineWeb-Edu 125M diagnostics show 10–50× growth of $\|W_\ell\|_F$ on `mlp.down`, `mlp.up`, and `attn.out_proj` before the ratio pins at $r_{\max}$. The mechanism is the positive feedback loop of Section 3.1: the pull-back does not scale with r_ℓ, so nothing damps growth until the clip does; coupling the decay into the scaled step creates the finite ceiling of Lemma 4 instead.

Clip removal. With $(r_{\min}, r_{\max}) \to (-\infty, \infty)$ at 125M, the largest transient $\hat{r}$ reaches 9.6 during the first ∼5% of steps before the coupled-WD ceiling engages; final CE 3.2187 vs. 3.2120 with the default clip. At 16B, doubling $r_{\max}$ from 5 to 10 shifts final CE by +0.0146 (Table 2).

D FULL HYPERPARAMETERS

Dense LM (FineWeb-Edu). GPT-2-style decoder-only transformer: pre-norm, RMSNorm, rotary position embeddings, ReLU^2 MLP, no biases, tied input/output embeddings, BPE vocabulary 50,304 (GPT-2 tokenizer). Precision bf16; gradient accumulation to the global batches below; cosine LR schedule with 1,000 warmup steps; matrix parameters use the optimizer under test, non-matrix parameters the AdamW fallback at the same LR with betas $(0.9, 0.95)$ and weight decay 0.1 (matching Moonlight's dense-baseline settings). OrScale (LM configuration): $r_{\min} = 0.1$, $r_{\max} = 5.0$, $\mu = 0.95$, $\lambda = 0.1$, $k = 5$, $\varepsilon = 10^{-6}$, fp32 calibration. Per-scale geometry and shared recipe LRs:

Table 6: Dense-LM scales. "Batch" is examples per optimizer step (global); tokens/step = batch × seq. At 545M+ the sequence length is shortened from Moonlight's 8192 to 4096 with batch examples doubled, keeping tokens/step, steps, and the LR schedule unchanged; every optimizer at a given scale sees the identical setting, so within-scale comparisons are unaffected (absolute losses at 545M+ are not comparable to 8K-context numbers).

Scale	Seq len	Batch (ex.)	Tokens	Steps	Recipe LR	Warmup
125M	1024	256	5.24B	20,000	1.0×10^{-3}	1,000
399M	8192	96	8.92B	11,342	9.503×10^{-4}	1,000
545M	4096	256	14.04B	13,390	9.143×10^{-4}	1,000
1.1B	4096	384	28.54B	18,146	8.561×10^{-4}	1,000

16B-A3B MoE. Moonlight-16B-A3B configuration: 27 layers, hidden size 2048, 16 heads, multi-head latent attention (KV LoRA rank 512, q_{nope} 128, q_{rope} 64, v head dim 128), 64 routed experts with top-6 sigmoid routing and routed scaling factor 2.446, 2 shared experts, MoE intermediate 1408, dense intermediate 10944 (first layer dense), aux-loss-free load balancing via router-bias updates (rate 10^{-3}, aux coefficient 0), vocabulary 163,840, sequence length 8192. Total 15,958.1M parameters, 2,241.7M active per token. Training: 8×H20, bf16, ZeRO-1 optimizer-state sharding, micro-batch 1 with gradient accumulation 256 per rank (= 2048 examples = 16,777,216 tokens/step), 596 steps = 9,999,220,736 FineWeb-Edu tokens (GPT-2 tokenizer), warmup 60 steps, cosine schedule, seed 42, validation every 50 steps over 10 batches. All optimizers: LR 4.2×10^{-4}, weight decay 0.1; Muon-family: $\mu = 0.95$, $k = 5$; AdamW (standalone and fallback groups): betas $(0.9, 0.95)$; OrScale: $r_{\min} = 0.1$, $r_{\max} \in \{5, 10\}$.

CIFAR-10 / DavidNet. 24 epochs, global batch 512, bf16, cosine schedule with 500-step warmup, standard augmentation, label smoothing 0, no gradient clipping; three seeds per cell; per-family

LR grids with weight decay 5×10^{-4}; OrScale clip $(0.5, 1.5)$; AdamW fallback LR 10^{-3}, betas $(0.9, 0.999)$. Val top-1 averages the last three epochs, then seeds.

Uncalibrated variant LRs (Table 3). The uncalibrated variant's matrix LR was tuned to 3×10^{-3} at 125M (grid evidence) and 2.376×10^{-3} at 399M ($2.5\times$ recipe, the rule extrapolated from the 125M ratio); its AdamW group stays at the recipe LR.

E 16B-A3B ADDITIONAL RESULTS

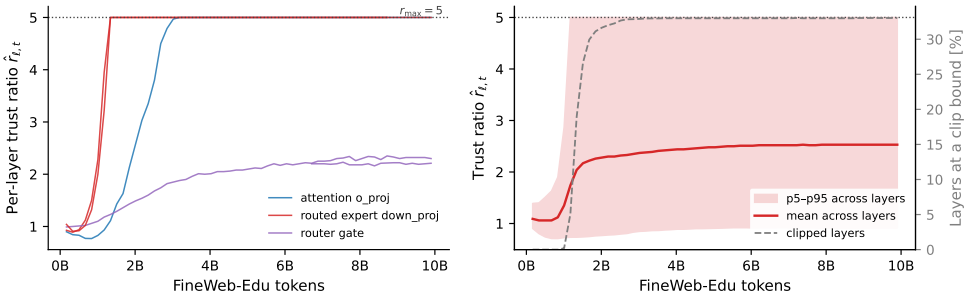


Figure 4: Logged OrScale trust ratios at 16B-A3B ($r_{\max}$=5 arm), behind the heterogeneity check of Section 5.3. *Left:* per-layer trajectories $\hat{r}_{\ell,t}$ by layer class: routed-expert projections rise to the ceiling within 1.5B tokens, attention output projections follow by 3B, router gates settle near 2.2–2.3 and never clip. *Right:* mean and p5–p95 band across all matrix layers; the fraction of layers at a clip bound (gray, right axis) reaches 33% late in training.

Trust-ratio heterogeneity in full. Figure 4 shows the logged ratios of the $r_{\max}$=5 arm, expanding the heterogeneity check of Section 5.3. The calibration constants themselves span an order of magnitude across layers (c_ℓ p5–p95: 0.0047–0.0446), so no static rescaling could reproduce the layer-class structure. The 33% late-training clip fraction is *not* the saturation pathology of Theorem 4—that failure pins *every* layer at $r_{\max}$ from the first steps—but the coupled-WD ceiling engaging for the layer class that wants the largest relative steps while other classes remain in-band; consistently, loosening the ceiling to $r_{\max}$=10 moves the final loss by only $+0.015$ nats (Figure 5). Under the dense 125M diagnostics the same statistics are present but weaker ($\nu_{\text{growth}} \approx 0.4$), matching the smaller dense-regime gains: the theory's heterogeneity dial and the empirical gap move together.

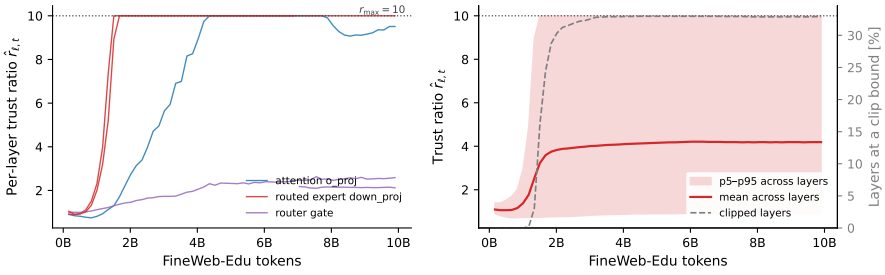


Figure 5: Per-layer trust ratios and layer-aggregate band for the $r_{\max}$=10 arm. The layer-class ordering (experts > attention > router) is identical to Figure 4; expert projections settle near the looser ceiling, and the run finishes 0.0146 nats behind $r_{\max}$=5.

Throughput: AdamW 28.7k tok/s/rank-group, Muon+Moonlight 23.2k, OrScale 23.5k ($r_{\max}$=5) and 23.4k ($r_{\max}$=10); the trust-ratio computation is within measurement noise of Muon+Moonlight, and the AdamW-vs-Muon-family difference is the Newton–Schulz cost, unrelated to OrScale. Figure 6 overlays smoothed training loss for all four arms; validation curves for all arms at all 12 checkpoints are included in the supplementary CSVs.

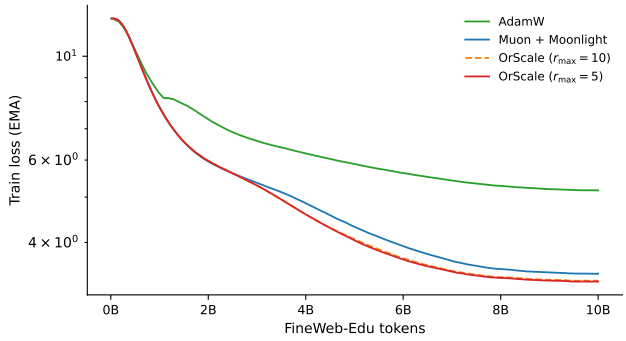


Figure 6: Smoothed training loss (EMA) for all four 16B-A3B arms.

F DENSE-REGIME ADDITIONAL RESULTS

Matched 125M LR sweep. Both optimizers on the same four-point grid, single seed:

Table 7: 125M FineWeb-Edu final validation CE across a shared LR grid. OrScale is at or below Muon+Moonlight at every stable cell; both optima sit at the recipe LR 10^{-3}; Muon+Moonlight diverges at 10^{-2} while OrScale still trains.

Optimizer	3×10^{-4}	10^{-3}	3×10^{-3}	10^{-2}
Muon+Moonlight	3.2841	**3.2319**	3.2372	4.6262 (diverged)
OrScale	3.2588	**3.2120**	3.2196	3.3287

Scaling-law fits. Kaplan-style log–log fits over the four dense scales give $L = 2.826\,C^{-0.054}$ (AdamW), $2.731\,C^{-0.053}$ (Muon+Moonlight), $2.725\,C^{-0.052}$ (OrScale); the AdamW and Muon exponents match Moonlight's published fits to within 0.002, an internal sanity check on the setup. With four points per optimizer we treat the per-scale deltas of Table 1 as primary and the fitted exponents as descriptive; Figure 3 (left, main text) plots the per-scale gap.

G CIFAR-10 DETAILS

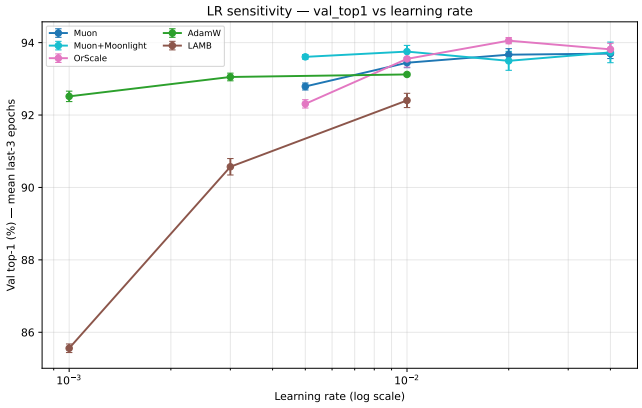


Figure 7: CIFAR-10 / DavidNet validation top-1 vs. learning rate, five main baselines, three seeds per cell.

Two notes on the LR-sensitivity comparison. First, nominal LR grids are not directly comparable across families: Moonlight's shape factor multiplies the effective Muon step by $0.2\sqrt{\max(m, n)} \approx$

Table 8: CIFAR-10 / DavidNet leaderboard at each optimizer's best LR (three seeds, $\pm 1\sigma$).

Rank	Optimizer	LR	Val top-1 (%)
1	**OrScale** (ours)	0.02	$\mathbf{94.05 \pm 0.08}$
2	Muon + Moonlight	0.01	93.75 ± 0.17
3	Muon	0.04	93.70 ± 0.14
4	AdamW	0.01	93.12 ± 0.04
5	LAMB	0.01	92.40 ± 0.20

4.8–13.6 for DavidNet's convolutions, so Muon+Moonlight's nominal grid sits at a Muon-effective position several times higher (empirically, Moonlight at LR 0.005 matches Muon at 0.02: 93.61 vs. 93.67). Second, extending the grid downward, Muon+Moonlight at nominal $\{5 \times 10^{-4}, 10^{-3}\}$ falls to 92.41 ± 0.12 and 93.18 ± 0.09, off its 93.75 plateau—its apparent flatness on the nominal grid partly reflects grid placement. At OrScale's best LR the gain over the best Muon-family baseline is $+0.30$ points with a Welch $t \approx 2.8$ across seeds.

H Compute resources

CIFAR-10/DavidNet: $\sim$2 minutes per run on a single RTX-class GPU; the full sweep $\sim$8 GPU-hours. Dense LM: 8×H20-3E, from $\sim$3 hours (125M) to $\sim$3–5 days (1.1B) per run. 16B-A3B MoE: 3.9–5.0 days per arm on 8×H20 (4 arms; wall-clock in Table 2). Total project compute including preliminary and failed sweeps: $\sim$5,000 GPU-hours (the four 16B-A3B arms alone account for $\sim$3,600).

I Broader impact

OrScale is a pre-training-efficiency modification: its direct effect is reducing compute and energy per unit of pre-training progress. Indirect negative effects mirror those of the systems it accelerates (cheaper pre-training lowers the cost of downstream misuse, bias, and environmental externalities); it introduces no new dataset, surveillance, or model-release pathway beyond the optimizer implementation.