ORCA: THE ANNEALED SPECTRAL CONDITIONING OPTIMIZER FOR FASTER, BETTER LLM TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern LLM optimizers such as Muon often produce weight matrices with higher effective rank than Adam, yet further spectral control has delivered only modest gains. We identify a tension behind this result: concentrated spectra can suppress gradient directions in coupled weight matrices and slow optimization, while constraints maintained throughout training can limit task-specific adaptation and raise the attainable loss floor. We introduce ORCA (Orthogonal Regularization, Cooled After), a minimal optimizer intervention that applies strong but temporary soft orthogonality regularization early in training, then removes it. This allows the weights to benefit from a broader spectrum early on and adapt freely afterward. Across LLaMA, Qwen3, and fine-grained mixture-of-experts models ranging from 130M to 8B parameters, ORCA achieves lower final validation loss than Muon. Its loss reduction relative to Muon matches or exceeds Muon's reduction relative to Adam. Ablations support the early-shaping, later-release design. Further, ORCA requires no architectural changes and adds minimal overhead.

1 INTRODUCTION

Recent matrix-based optimizers such as Muon have shown substantial improvements over Adam in LLM training (Liu et al., 2025; Jordan et al., 2024). A striking difference emerges in their weight trajectories: compared with Adam, Muon maintains a markedly flatter singular-value spectrum and a substantially higher effective rank over most of training, where effective rank is measured as $\|\mathbf{W}\|_F^2/\|\mathbf{W}\|_2^2$. The two optimizers start from comparable values, but Adam exhibits a much sharper decline early in training and remains consistently below Muon thereafter. This observation suggests a natural hypothesis: better preserving the effective dimensionality of weight matrices may be an important ingredient behind faster optimization.

This emerging importance of spectral structure has motivated an important and growing line of work that explicitly preserves or promotes favorable spectral structure during training (Qiu et al., 2026a; Shi et al., 2026; Dong & Sawin, 2026). These approaches introduce principled mechanisms for controlling the weight spectrum and have demonstrated promising improvements in LLM optimization. If a higher effective rank is indeed beneficial, however, one might expect stronger spectral control to yield correspondingly larger gains. In practice, the improvements over Muon are often more modest than this intuition would suggest, despite substantially stronger control over the weight spectrum.

To understand this gap, we examine two competing effects of spectral control that arise simultaneously. We begin with a matrix factorization model (Gunasekar et al., 2017) that abstracts the multiplicative coupling common in Transformer architectures. On the benefit side, we illustrate that a well-shaped spectrum allows more gradient directions to contribute to other parameter updates, facilitating faster optimization. On the cost side, however, imposing spectral constraints limits model expressiveness and can raise the attainable loss floor. Correspondingly, controlled interventions during Llama training empirically support both effects. Together, these results point to a seemingly trade-off between faster optimization and model expressiveness.

However, we argue that this trade-off is not fundamental. Spectral regularization may be most valuable early, when it can shape the optimization trajectory, yet become a liability if maintained as the weights adapt to the task. The two effects need not be accepted together: we can use the constraint to shape the spectrum, then release it before it restricts the final solution. This motivates a

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.

minimal method ORCA (Orthogonal Regularization, Cooled After): During an initial phase, ORCA applies a soft orthogonality penalty, $R_{\text{orth}}(\mathbf{W}) = \frac{1}{4}\|\mathbf{W}\mathbf{W}^\top - \mathbf{I}\|_F^2$. Then directly removes it at a single switch point. Training thereafter follows the standard optimizer update, with no gradual decay or further changes. This design lets early regularization influence the optimization path without prescribing the spectrum throughout training. In our experiments, this simple release is followed by faster loss reduction and a lower final validation loss than Muon.

Figure 1.1 illustrates the transition. While the penalty is active, the effective rank of the weight matrix rises sharply above Muon's, although ORCA initially has a higher validation loss. After release, the rank falls from its peak but remains above Muon's, while validation loss decreases rapidly, crosses the Muon curve. The result is consistent with early spectral shaping leaving a useful effect on the subsequent trajectory, even after the explicit penalty is gone. Controlled interventions in Section 5 further investigates into this connection.

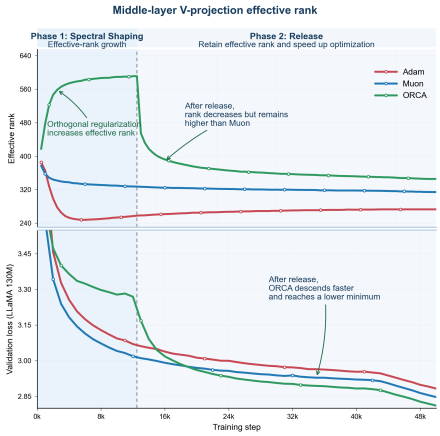


Figure 1.1: LLaMA 130M training dynamics. Effective rank is measured on a middle-layer V projection; the dashed line marks regularization release.

We evaluate ORCA on LLaMA, Qwen3, and fine-grained MoE models ranging from 130M to 8B parameters, including 100B-token training runs. ORCA achieves lower final validation loss than Muon across these settings. In most experiments, the reduction is 0.02–0.04, comparable to or greater than Muon's gain over Adam. To put this loss reduction in perspective, on LLaMA 1.3B, ORCA achieving the 0.024 lower validation loss reaches Muon's 100k-step loss in just 60k steps. Further, downstream evaluations, alternative early spectral interventions, multi-seed runs, and ablations of method parameters and intervention timing further support these gains. Together, these results show that early spectral shaping followed by constraint relaxation can improve training efficiency and final performance across architectures without substantial changes to the optimizer.

To summarize the contributions of the paper: (1) We illustrate a trade-off in spectral control through a simplified model and empirical observations: favorable spectra can accelerate optimization but persistent constraints may harm the attainable solution quality. (2) We introduce ORCA, a minimal shape-then-release optimizer that leverages early orthogonal regularization to improve optimization trajectories while allowing unconstrained adaptation in later training. (3) We demonstrate ORCA's effectiveness across diverse LLM architectures and scales, achieving faster convergence and lower loss than existing optimizers. Through extensive controlled experiments, we further validate the mechanism that early effective-rank shaping drives the observed improvements.

2 Spectral Structure and Optimization

As noted in Section 1, Muon's higher effective rank coincides with faster optimization. This observation suggests that properties of the weight spectrum may be useful for optimization, but whether they provide a general benefit remains unclear. To turn this still-vague possibility into a precise question about optimization dynamics, we ask:

> **Q1.** *Can the spectral structure observed under Muon directly affect optimization speed, and through what mechanism?*

A direct mechanism already appears in 'the multiplicatively coupled weight matrices' of Transformers. Consider for example the value–output pair $\mathbf{W}_V\mathbf{W}_O$ of a single attention head:

$$\text{Attn}(\mathbf{H}) = \text{softmax}\left(\frac{\mathbf{H}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{H}^\top}{\sqrt{d_h}}\right)\mathbf{H}\mathbf{W}_V\mathbf{W}_O. \tag{2.1}$$

The gradient of the value projection takes the form

$$\nabla_{\mathbf{W}_V} L = \mathbf{G}\mathbf{W}_O^\top, \quad \text{with} \quad \mathbf{G} = \nabla_{(\mathbf{W}_V\mathbf{W}_O)} L.$$

Here, $\mathbf{G}$ denotes the gradient with respect to the entire product $\mathbf{W}_V\mathbf{W}_O$. Thus, $\mathbf{W}_O$ acts as a spectral filter on the gradient received by $\mathbf{W}_V$. Writing $\mathbf{W}_O = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, we have $\nabla_{\mathbf{W}_V} L = \sum_i \sigma_i (\mathbf{G}\mathbf{v}_i)\mathbf{u}_i^\top$. This makes the role of the spectrum explicit: each component $\mathbf{G}\mathbf{v}_i$ enters the gradient with weight σ_i. If $\mathbf{W}_O$ is scaled orthogonal, $\mathbf{G}\mathbf{W}_O^\top$ preserves the spectrum of $\mathbf{G}$ up to a common scale. At the opposite extreme, if $\mathbf{W}_O$ has rank r, the update is restricted to contributions from r spectral directions. More generally, as the spectrum becomes more concentrated, directions associated with smaller singular values contribute proportionally less to the update.

In fact, this partner-dependent filtering is ubiquitous in Transformers: analogous coupling appears between the query–key pair $(\mathbf{W}_Q, \mathbf{W}_K)$ and the up- and down-projections of MLP and expert blocks.

To isolate this mechanism from the surrounding Transformer computation, consider the generic factorization $\mathbf{W} = \mathbf{X}\mathbf{Y}$, where $\mathbf{G} = \nabla_{\mathbf{W}} L(\mathbf{W})$. The simple chain rule gives $\nabla_{\mathbf{X}} L = \mathbf{G}\mathbf{Y}^\top$, $\nabla_{\mathbf{Y}} L = \mathbf{X}^\top\mathbf{G}$, so each factor directly filters the gradient received by its partner. Under the continuous-time exact-polar Muon dynamics (Ambrosio et al., 2005; Peyré, 2026),

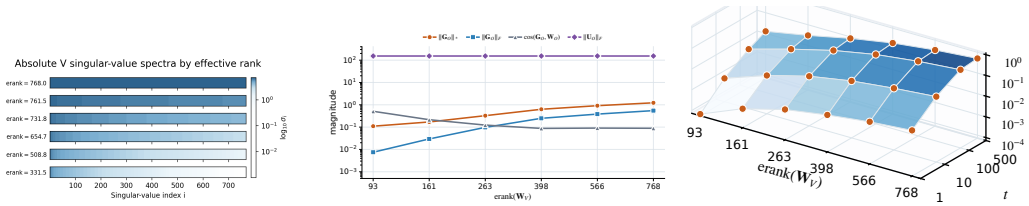
$$\dot{\mathbf{X}} = -\eta_t^X \operatorname{Polar}(\mathbf{G}\mathbf{Y}^\top), \qquad \dot{\mathbf{Y}} = -\eta_t^Y \operatorname{Polar}(\mathbf{X}^\top\mathbf{G}),$$

where $\operatorname{Polar}(\mathbf{M}) = \mathbf{U}\mathbf{V}^\top$ for an SVD $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. The instantaneous loss decrease is therefore

$$-\dot{L} = \eta_t^X \left\|\mathbf{G}\mathbf{Y}^\top\right\|_* + \eta_t^Y \left\|\mathbf{X}^\top\mathbf{G}\right\|_* .$$

Thus, the same filtered gradients above are the quantities that determine Muon's instantaneous loss decrease. The local spectral effect can therefore translate into a local effect on optimization speed.

To empirically demonstrate this effect, we conduct a reciprocal V–O intervention. Let $\mathbf{G}_O = \nabla_{\mathbf{W}_O} L$ denote the gradient with respect to $\mathbf{W}_O$, and let $\mathbf{U}_O = \operatorname{Polar}(\mathbf{G}_O)$ denote the corresponding Muon update direction. At a 10k-step checkpoint, we raise the smaller singular values of $\mathbf{W}_V$ toward its largest singular value while keeping its singular vectors and spectral norm fixed, and update only $\mathbf{W}_O$ with Muon. As shown in Figure 2.1b, with the Muon update scale $\|\mathbf{U}_O\|_F$ held fixed, increasing the effective rank of $\mathbf{W}_V$ increases $\|\mathbf{G}_O\|_*$—the loss decrease under Muon flow—while decreasing the gradient–weight alignment $\cos(\mathbf{G}_O, \mathbf{W}_O)$. These all indicate that the same-sized Muon update has a stronger effect as the effective rank of $\mathbf{W}_V$ increases. Consistently, Figure 2.1c also shows that updating $\mathbf{W}_O$ yields a larger held-out loss decrease at higher effective rank of $\mathbf{W}_V$.



(a) Absolute singular-value spectra. (b) Gradient and update metrics. (c) Held-out loss decrease.

Figure 2.1: Single-layer V-spectrum intervention with O-only Muon updates. $\mathbf{G}_O$ and $\mathbf{U}_O$ denote the gradient and Muon update direction. In (c), rollout steps and loss decrease use logarithmic scales.

These observations suggest a simple principle: spectral concentration can create an optimization bottleneck. For a nonzero matrix $\mathbf{A}$, define its effective rank by $\operatorname{erank}(\mathbf{A}) := \|\mathbf{A}\|_F^2 / \|\mathbf{A}\|_2^2$. To make this intuition concrete, we study a matrix-factorization model. The following bounds illustrate one possible mathematical pathway through which spectral structure may affect optimization.

Theorem 2.1 (Low Effective Rank Limits Descent Speed). *For $k \geq 2$, let $\mathbf{X}$ and $\mathbf{Y}$ be conformable matrices with $\mathbf{W} = \mathbf{X}\mathbf{Y} \in \mathbb{R}^{k\times k}$, and consider the column-wise softmax loss*

$$L(\mathbf{W}) = \sum_{i=1}^{k} \left[\log\left(\sum_{j=1}^{k} e^{\mathbf{W}_{ji}} \right) - \mathbf{W}_{c_i i} \right], \quad c_i \in [k], \quad \textit{with} \quad \mathbf{W} = \mathbf{X}\mathbf{Y} \tag{2.2}$$

let $(\mathbf{X}_t, \mathbf{Y}_t)_{t\in[0,T]}$ be an absolutely continuous solution of the exact-polar Muon flow above, with $\eta_0 \geq 0$ and $0 \leq \eta_t^X \leq \eta_0 / \|\mathbf{Y}_t\|_2$, $\quad 0 \leq \eta_t^Y \leq \eta_0 / \|\mathbf{X}_t\|_2$. Assume that the factor

norms satisfy $\|\mathbf{X}_t\|_2, \|\mathbf{Y}_t\|_2 > 0$ *for all* $t \in [0, T]$. *Writing* $L_t = L(\mathbf{X}_t\mathbf{Y}_t)$, *we have*

$$L_T \geq L_0 \exp\left[-\sqrt{2}\,\eta_0 \int_0^T \left(\sqrt{\operatorname{erank}(\mathbf{X}_s)} + \sqrt{\operatorname{erank}(\mathbf{Y}_s)}\right) ds\right].$$

In this factorization model, the bound relates the rate of loss decrease to the effective ranks of the factors. Naturally, broad spectral support may alleviate this bottleneck, as demonstrated by the intervention experiments in Figure 2.1. We therefore test whether maintaining such support throughout LLaMA pretraining helps by imposing progressively stronger spectral constraints. However, Figure 2.2 shows that persistent orthogonal constraints worsen validation loss. Thus, the path-side advantage identified above does not translate into a benefit when the same structure is enforced persistently. This leads to our second question:

Q2. *If broad spectral structure alleviates an optimization bottleneck, why enforcing it hurts performance?*

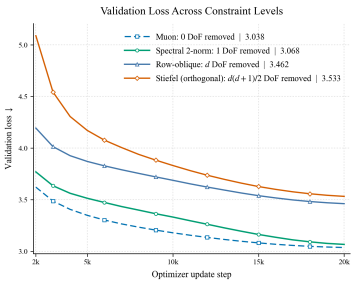


Figure 2.2: Validation loss under persistent spectral constraints.

The key is that a property that is favorable along the optimization trajectory need not be favorable when imposed as a persistent constraint. Spectral control does more than influence how an optimizer moves: when enforced throughout training, it also restricts the spectral structure of the solutions that the optimizer is allowed to reach. For example, an orthogonality constraint $\mathbf{W}\mathbf{W}^\top \approx \mathbf{I}$ forces the singular values of $\mathbf{W}$ to remain nearly equal. If the task prefers a more anisotropic or lower-effective-rank solution, such a constraint can prevent the optimizer from reaching it.

The factorized model above makes this competing effect concrete. For the same column-wise softmax objective in Eq. (2.2), the following theorem shows that sufficiently high effective rank can itself impose a nonzero loss floor.

Theorem 2.2 (High Effective Rank Can Impose a Loss Floor). *Assume* $k \geq 2$, $\mathbf{W} \neq 0$, *and* $\|\mathbf{W}\|_2 \leq B$ *for a fixed scale bound* $B = O(1)$. *Let* $K = |\{c_i : i \in [k]\}|$ *denote the number of distinct target indices. Then*

$$L(\mathbf{W}) \geq C_k(\mathbf{W})\, k \left(1 - \sqrt{\frac{K}{\operatorname{erank}(\mathbf{W})}}\right),$$

where $C_k(\mathbf{W}) > 0$ *is the scale factor determined by* k *and* $\|\mathbf{W}\|_F/\sqrt{k}$. *The rank factor* $1 - \sqrt{K/\operatorname{erank}(\mathbf{W})}$ *increases with effective rank and becomes positive when* $\operatorname{erank}(\mathbf{W}) > K$.

Thus, for this objective, sufficiently high effective rank is not merely unnecessary at the optimum: it can be incompatible with arbitrarily small loss. In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are $k \times k$ scaled orthogonal matrices, then $\operatorname{erank}(\mathbf{X}\mathbf{Y}) = k$, and Theorem 2.2 gives the positive endpoint loss floor $C_k(\mathbf{W})\, k \left(1 - \sqrt{K/k}\right)$ whenever $K < k$.

Together, these results illustrate a tension in the stylized model: broad spectral support can relax a path-side bottleneck, while a persistent spectral constraint can restrict low-loss solutions. We also test whether these intuitions are useful in the empirical training. The same spectral property can therefore be useful as a transient feature of the optimization trajectory while being harmful as a persistent constraint on the final solution.

This distinction helps reconcile the empirical behavior above. The high effective rank naturally maintained by Muon can support optimization without prescribing a fixed spectral geometry for the final weights. By contrast, increasingly restrictive spectral constraints preserve the desired structure by reducing the freedom of the model to adapt its spectrum as training proceeds. The appropriate goal is therefore not to maximize effective rank at every point in training, but to retain enough spectral support to avoid path-side bottlenecks while still allowing the spectrum to adapt to the task.

3 ORCA: A MINIMAL SHAPE-THEN-RELEASE OPTIMIZER

The preceding analysis reveals a tension between the trajectory-side benefit and the endpoint-side cost of spectral control. Broad spectral support can facilitate optimization along the training trajectory, while enforcing the same structure persistently may restrict task-adaptive solutions at the endpoint. This suggests a temporal design principle: spectral control should shape the early optimization trajectory, but be released before it restricts task adaptation.

ORCA (Orthogonal Regularization, Cooled After) is a minimal two-stage realization of this shape-then-release principle. During an initial active stage, ORCA applies spectral shaping to steer the early trajectory away from spectrally concentrated regimes. After this stage, the shaping term is removed completely and training reverts to standard Muon. This separation allows spectral control to influence the optimization path without prescribing the spectral geometry of the final solution.

3.1 ALGORITHM

A minimal realization: soft orthogonal shaping. For this minimal realization, ORCA uses soft orthogonality as its spectral shaping operator. Consider a matrix parameter $\mathbf{W} \in \mathbb{R}^{p\times q}$, and assume $p \le q$ for notational simplicity. We define $\mathcal{R}_{\mathrm{orth}}(\mathbf{W}) = 1/4 \left\|\mathbf{W}\mathbf{W}^\top - I_p\right\|_F^2$, whose gradient is

$$G_{\mathrm{orth}}(\mathbf{W}) := \nabla_{\mathbf{W}}\mathcal{R}_{\mathrm{orth}}(\mathbf{W}) = \left(\mathbf{W}\mathbf{W}^\top - I_p\right) W. \tag{3.1}$$

Writing a thin singular value decomposition as $\mathbf{W} = \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_p)\mathbf{V}^\top$, we have

$$\mathcal{R}_{\mathrm{orth}}(\mathbf{W}) = \frac{1}{4}\sum_{i=1}^{p}\left(\sigma_i^2 - 1\right)^2, \quad \mathbf{G}_{\mathrm{orth}}(\mathbf{W}) = \mathbf{U}\,\mathrm{diag}\left((\sigma_i^2 - 1)\sigma_i\right)_{i=1}^{p}\mathbf{V}^\top. \tag{3.2}$$

Thus, a gradient step on $\mathcal{R}_{\mathrm{orth}}$ decreases singular values above unit scale and increases nonzero singular values below unit scale, tending to reduce spectral concentration and increase the effective-rank measure considered in the preceding analysis. The operator requires neither an explicit SVD nor a projection onto the Stiefel manifold and can be computed using matrix multiplications alone.

ORCA update. Let $\mathbf{U}_t$ denote the standard Muon update direction at step t, η_t the learning rate, λ the decoupled weight-decay coefficient, and γ_t the spectral-shaping strength. For each matrix parameter $\mathbf{W}_t$, the ORCA update can be written as

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t\mathbf{U}_t - \eta_t\lambda\mathbf{W}_t - \eta_t\gamma_t\mathbf{G}_{\mathrm{orth}}(\mathbf{W}_t). \tag{3.3}$$

The weight-decay and spectral-shaping corrections are both evaluated at the pre-update parameter $\mathbf{W}_t$ (Loshchilov & Hutter, 2017), while $\mathbf{U}_t$ is computed in the same way as in standard Muon.

Similar to Muon and other matrix-based optimizers, ORCA is applied only to matrix parameters in the matrix-optimizer group. Embeddings, the language-model head, and one-dimensional parameters continue to follow their original AdamW updates.

Release rule. The defining feature of ORCA is that spectral shaping is transient rather than persistent. Let T denote the total number of training steps and $\tau \in [0, 1]$ the active fraction. The active window has length S, and we use the hard release schedule

$$\gamma_t = \begin{cases}\gamma, & t < S,\\ 0, & t \ge S,\end{cases} \qquad S = \lfloor \tau T \rfloor. \tag{3.4}$$

Equation (3.4) makes the two phases explicit. During the *active stage* ($t < S$), ORCA augments Muon with the spectral-shaping term in Eq. (3.3). During the *released stage* ($t \ge S$), the shaping term vanishes and the optimizer reverts exactly to the corresponding standard Muon training rule.

The hard switch is the simplest implementation of the shape-then-release principle. The early intervention is used to steer the optimization trajectory away from spectrally concentrated regimes, rather than to prescribe a persistent spectral constraint. Once released, the spectrum is free to adapt to the task objective. Empirically, we find that part of the higher-effective-rank state induced during the active stage persists after release, allowing the optimization benefit to continue without maintaining the orthogonal regularizer. Unless otherwise specified, we use $\tau = 0.25$ throughout the paper; Section 5 studies sensitivity to the release point.

Beyond soft orthogonality. Soft orthogonality defines the default ORCA optimizer, but the underlying shape-then-release principle is not tied to this particular spectral operator. In Section 5, we replace soft orthogonal shaping with alternative early spectral interventions, including Pion-start and DiagSVD variants. These methods differ in how the spectral state is shaped during the active stage; once the active window ends, the intervention is removed and all variants continue with the same released Muon dynamics. This provides a controlled way to study whether the benefit of early spectral intervention extends beyond soft orthogonality itself.

Implementation and overhead. In the standard setting, ORCA adds two matrix multiplications per weight matrix during shaping without changing the communication complexity; the remaining operations are fused into the optimizer update. Our Muon baseline performs 24 matrix multiplications across eight Newton–Schulz iterations, so shaping increases the multiplication count by $2/24 = 8.3\%$. This extra work disappears after release. We batch matrices of identical shapes to reduce kernel launches, use dedicated kernels, and form the Gram matrix along the shorter dimension. On the 8B MoE model, average distributed optimizer update time is $227.2\,\mathrm{ms}$ for Muon and $237.4\,\mathrm{ms}$ for ORCA. The $10.2\,\mathrm{ms}$ difference amounts to approximately 0.8% of a roughly $1.3\,\mathrm{s}$ training step. Peak memory is $101.41\,\mathrm{GB}$ and $101.66\,\mathrm{GB}$, respectively. Overall, ORCA incurs negligible additional training overhead in this configuration.

4 Experiment Validations

In this section, we verify the efficacy of ORCA on three specific issues: (1) Llama pretraining with grid learning rate search and larger scale pretraining. (2) Pretraining task on Qwen3 architecture with the additional QK-norm that may violate the matrix factoring reparameterization. This section focuses on the efficacy of ORCA, and for more robustness test of parameters and controlled interventions on gain of ORCA, readers can refer to Section 5.

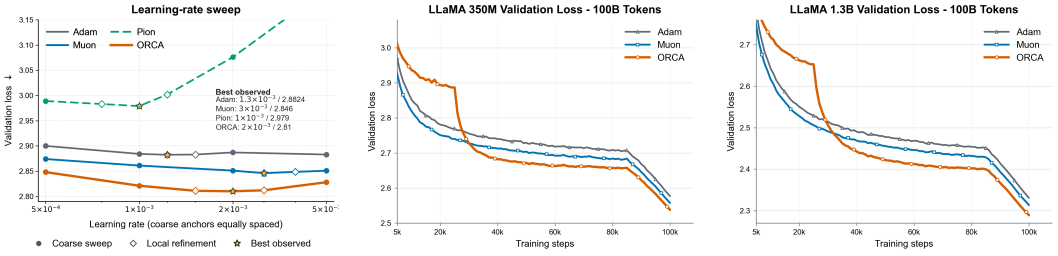
4.1 Experimental Setup and Tuning Protocol

Setups. We pre-train Llama-series models Touvron et al. (2023) ranging from 130M to 1.3B parameters on C4 , using 25B to 100B tokens. We also pre-train Qwen3-series models Yang et al. (2025) ranging from 160M to 0.6B parameters on The Pile (Gao et al., 2020), using 25B to 100B tokens. Finally, we train fine-grained mixture-of-experts models ranging from 0.8B to 8B parameters on OLMo-mix-1124 (OLMo et al., 2024), the data mixture including DCLM-Baseline (Li et al., 2024) using 25B to 100B tokens.

We first run a learning rate grid-search for Adam, Muon, Pion and ORCA. All trials share the same data, batch size, sequence length, and 50k-step training budget. We use $\mathbf{O}_t = 0.2/\mathrm{Rms}(\mathbf{O}_t) \cdot \mathbf{O}_t$ to rescale the Muon and ORCA update where $\mathbf{O}_t$ is the Newton-Schulz orthogonalization output to align the update with the Adam's (Liu et al., 2025). We perform a two-stage coarse-to-fine learning rate grid search. The coarse sweep evaluates $\{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$. Once the coarse sweep locates a candidate learning rate, we evaluate the neighbors on its both sides to refine the best lr within a smaller local range. The refinement grid consists of $\{6, 8\} \times 10^{-4}$, $\{1.3, 1.6\} \times 10^{-3}$, and $\{3, 4\} \times 10^{-3}$, which reduces the local spacing between learning rates candidates to about $1.2\times$–$1.5\times$. All four optimizers use this protocol and the same tuning budget.

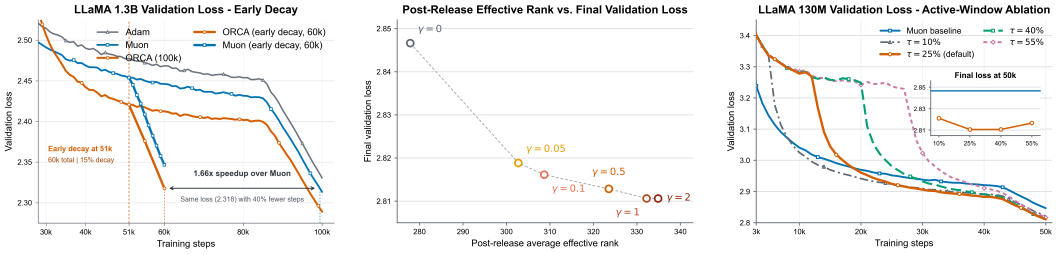
The advantage of ORCA holds over a wide range of learning rates: at every learning rate evaluated for both methods, ORCA reaches a lower final validation loss than Muon, and at all evaluated configurations between 10^{-3} and 5×10^{-3} it is even better than fully tuned Muon. The independent searches select 1.3×10^{-3}, 3×10^{-3}, 10^{-3}, and 2×10^{-3} for Adam, Muon, Pion, and ORCA, with final validation losses of 2.8824, 2.8460, 2.9790, and 2.8106, respectively; ORCA thus lowers the loss by a further 0.0354 relative to the best Muon and stays within 0.002 of its own optimum across 1.6×10^{-3} to 3×10^{-3}, showing that the improvement does not rely on a narrow learning-rate choice.

LLama Pretraining Tasks. We compare ORCA against Adam and Muon on LLaMA 350M and 1.3B, training each model for 100B tokens. As shown in Figure 4.1, Muon outperforms Adam at both scales, while ORCA further lowers the validation loss. Compared with Muon, ORCA reduces the final loss by 0.0194 at 350M and 0.0244 at 1.3B, with consistent gains at both scales. Further, a four-seed robustness result for the 130M, 50k-step setting is reported in Appendix C. We also evaluate the final ORCA and Muon LLaMA 1.3B checkpoints from Figure 4.1 on five downstream



(a) Learning-rate sweep (b) LLaMA 350M (c) LLaMA 1.3B

Figure 4.1: LLaMA pretraining results.



(a) Matched-target token efficiency (b) Shaping strength γ (c) Active-window fraction τ

Figure 4.2: Token efficiency and shaping ablations in LLaMA Pretraining.

tasks (Clark et al., 2018; Zellers et al., 2019; Sakaguchi et al., 2021; Bisk et al., 2020). As shown in Table 4.1, ORCA outperforms Muon on all five tasks. This suggests that the lower pre-training loss of ORCA also leads to better downstream performance.

We also evaluate how the optimization performance translates into token efficiency by measuring the training required to reach a common target loss. We continued ORCA from its original 51k-step checkpoint, starting the final decay early, and reached a validation loss of 2.318 at approximately 60k steps. The standard Muon run, using its original 100k-step settings, reached the same loss at 100k steps (Figure 4.2a). Therefore, ORCA used 40% fewer tokens to reach this target, a 1.66× token-efficiency gain. To save compute, we reused the checkpoint without retuning ORCA for the shorter run: its original 25k-step shaping stage accounts for 41.7% of the 60k-step budget rather than the tuned default of 25%.

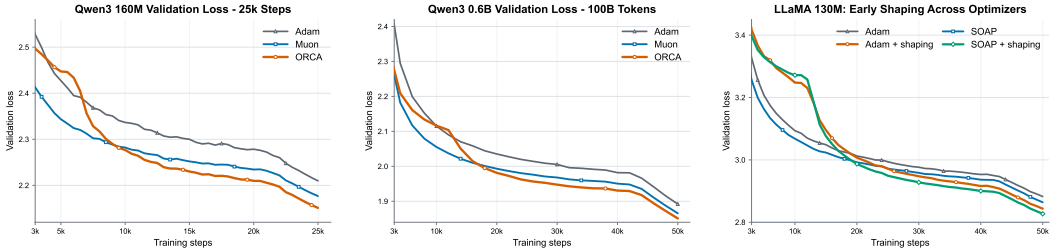
Qwen3 Pretraining Tasks. We compare Adam, Muon, and ORCA on Qwen3-style architecture. Within each model scale, all methods share the same common hyperparameters. The Qwen3 0.6B curves cover 50k optimizer steps (about 105B tokens), and the Qwen3 160M curves cover 25k optimizer steps (about 26B tokens). In both Qwen settings, both sets of loss curves show the same ordering: ORCA may fall temporarily behind Muon early in training, then crosses over and holds the advantage at the end of the observed interval. The validation-loss improvement of ORCA is therefore not confined to LLaMA, but also reproduces on the Qwen3 architecture.

Larger Scale Fine-grained MoE Pretrining. We observed router perturbations under a fixed global release point, so we release the orthogonal constraint layerwise: at 20% of training in the first layer and at 60% in the last layer, with release times linearly interpolated across layers. We use top-4 routing with 64 experts in the MoE 0.8B model (0.24B active parameters) and 64 experts in the MoE 8B model (1B active parameters) (Dai et al., 2024). Due to computational budget constraints, we focus our comparison on Muon and ORCA. ORCA achieves the lowest final loss at both 25B and 50B tokens for the 0.8B model, showing that its gain persists as the token budget increases. The 8B experiment extends this result to a larger model trained on 100B tokens, where ORCA also finishes below Muon (Figure 4.4). Across these MoE experiments, ORCA reduces final validation loss by 0.020–0.025 relative to Muon.

Early orthogonal shaping also benefits other base optimizers. We apply the same soft orthogonal shaping step (3.3) to Adam and SOAP on LLaMA 130M (Vyas et al., 2025). Shaping is active only during the first 25% of training; the base optimizer then continues alone. For both optimizers, shaping temporarily raises validation loss while active, but the shaped runs overtake their baselines after its removal and finish lower. At 50k steps, it reduces Adam's loss from 2.8824 to 2.8435 and SOAP's

Table 4.1: Downstream evaluation on LLaMA 1.3B.

Optimizer	ARC-C	ARC-E	HellaSwag	PIQA	WinoGrande	Avg.
Muon	0.305 (0.013)	0.342 (0.010)	0.549 (0.005)	0.705 (0.011)	0.547 (0.014)	0.489
ORCA	**0.311** (0.014)	**0.357** (0.010)	**0.573** (0.005)	**0.711** (0.011)	**0.553** (0.014)	**0.501**



(a) Qwen3 160M (b) Qwen3 0.6B (c) LLaMA 130M, base optimizers

Figure 4.3: Validation loss across architectures and base optimizers.

from 2.8640 to 2.8271, gains of 0.0389 and 0.0369, respectively. These results show that the benefits of early orthogonal shaping extend beyond Muon to other base optimizers.

5 Where the Gain Comes From: Controlled Interventions

Figure 1.1 offers an initial view of ORCA's mechanism: early shaping raises the effective rank, part of this increase persists after shaping is released, and loss then decreases faster. While Section 4 establishes ORCA's optimization gains, in this section we vary shaping strength and timing, as well as the form of the early spectral intervention, to isolate their contributions to these gains. Additional controls for optimizer state and learning-rate schedules are reported separately in Appendix C.2.

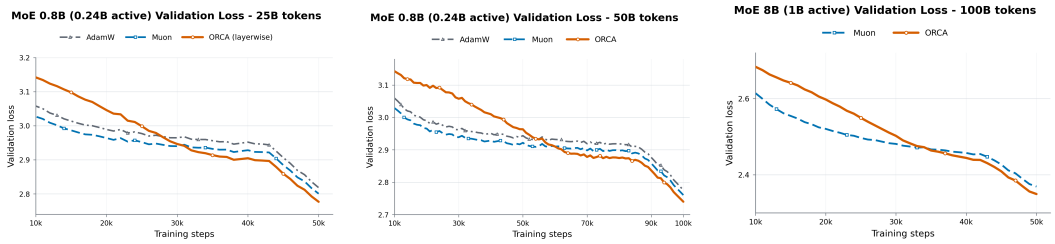
Retained Effective Rank Predicts Post-Release Gain We first test whether the effective-rank difference retained after release tracks ORCA's final gain. Holding all other settings fixed, we sweep the shaping strength over $\gamma \in \{0, 0.05, 0.1, 0.5, 1, 2\}$ and measure both the average post-release effective rank and the final validation loss. The sweep exhibits a clear dose response. As γ increases, the retained effective rank rises from about 278 to 335, while the final validation loss falls from 2.8466 to 2.8106 (Figure 4.2b). The relation flattens at larger γ, indicating diminishing returns. This supports retained effective rank as an important variable behind ORCA's post-release gain.

Release-point sensitivity. We first vary the duration of the initial shaping stage over $\tau \in \{0.10, 0.25, 0.40, 0.55\}$ on LLaMA 130M. Every setting improves over Muon's final loss of 2.8512: the four release points reach 2.8210, 2.8106, 2.8107, and 2.8166, respectively (Figure 4.2c). The broad effective range shows that ORCA does not rely on a precise release point. We use $\tau = 0.25$ by default because it is the shortest tested window in the best-performing range.

Shaping-window ablation. We next compare the default early window $[0, 0.25T)$ with middle $[0.375T, 0.625T)$, late $[0.75T, T)$, full-training, and off-then-on $[0, 0.25T) \cup [0.75T, T)$ schedules. Together with the effective-rank dynamics in Figure 1.1, Figure 5.1 shows two effects qualitatively consistent with the simplified analysis: **(i) shaping is costly while active.** Validation loss rises and remains on a high plateau whenever shaping is imposed, consistent with persistently high effective rank restricting the task-adaptive spectrum. **(ii) The benefit emerges after release and requires time.** Once shaping is removed, the retained high-rank state supports faster Muon descent without the same endpoint restriction. Early shaping leaves the longest released phase and achieves the lowest final loss. Middle shaping also recovers rapidly and finishes slightly below Muon, but remains worse than early; late, full, and off-then-on shaping leave insufficient or no final released phase and fail to recover by the endpoint. These explain ORCA's two-stage design: raise effective rank early, then reserve most of training for unconstrained optimization.



Figure 5.1: Loss curves under different shaping windows on LLaMA 130M.



(a) MoE 0.8B, 25B tokens (b) MoE 0.8B, 50B tokens (c) MoE 8B 1B active, 100B tokens

Figure 4.4: Fine-grained MoE validation loss with layer-wise release.

5.1 Alternative Early Spectral Interventions

ORCA's mechanistic claim extends beyond soft orthogonal regularization: if raising and retaining effective rank early improves optimization after release, other ways of doing so should yield similar gains. We test this with DiagSVD and Pion-start, comparing them with ORCA and standard Muon in 50k-step LLaMA 130M training. Both alternatives replace soft orthogonal shaping during the first 25% of training; afterward, all variants follow the same Muon dynamics.

DiagSVD directly amplifies updates along weaker singular directions. It reweights the Muon direction $\widetilde{\mathbf{u}}$ in the basis $\mathbf{U}$ of the left singular vectors of $\mathbf{W}$:

$$\widehat{\mathbf{u}} = \mathbf{U}\,\mathrm{diag}(s_i)\mathbf{U}^\top\widetilde{\mathbf{u}} + s_r(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\widetilde{\mathbf{u}}, \qquad s_i = 1 + \theta(i/r)^\delta, \tag{5.1}$$

where the singular values are in descending order and $\theta = 128$, $\delta = 3$. DiagSVD assigns larger update coefficients to directions associated with smaller singular values, thereby promoting a higher effective rank of the weight matrix. The exact and lazy variants refresh the singular-vector basis every step and every five steps, respectively. Pion-start instead provides a spectrum-preserving intervention. Given the task gradient $\mathbf{G}$, it constructs the skew-symmetric generators

$$\mathbf{G}_{\text{in}} = \mathbf{W}^\top\mathbf{G} - (\mathbf{W}^\top\mathbf{G})^\top, \qquad \mathbf{G}_{\text{out}} = \mathbf{G}\mathbf{W}^\top - (\mathbf{G}\mathbf{W}^\top)^\top, \tag{5.2}$$

and updates the weight from both sides as

$$\mathbf{W}^+ = E_2(\mathbf{A}_{\text{out}})\mathbf{W}E_2(\mathbf{A}_{\text{in}}), \qquad E_2(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \tfrac{1}{2}\mathbf{A}^2, \tag{5.3}$$

where $\mathbf{A}_{\text{in}}$ and $\mathbf{A}_{\text{out}}$ are obtained from momentum and second-moment estimates of the corresponding generators. The underlying two-sided Lie action preserves the singular spectrum, and hence the effective rank, during the intervention period; E_2 is its second-order approximation. For Pion-start based methods, whose underlying dynamics preserve the singular spectrum, an orthogonalized start improves the final loss from 2.8505 with a random start to 2.8371. This offers an indication that early orthogonality matters. The two DiagSVD variants recover part of ORCA's gain: against Muon's 2.8512, DiagSVD exact and lazy reach 2.8265 and 2.8356, respectively, while ORCA reaches 2.8106. By amplifying early updates along directions of smaller singular values, DiagSVD improves the final loss even after release. These all indicate early spectral shaping can benefit later optimization.

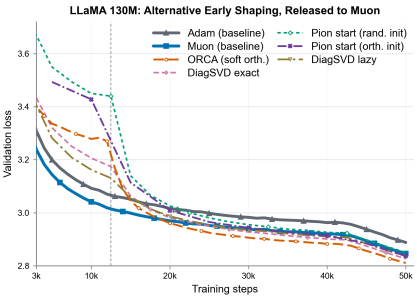


Figure 5.2: Alternative early spectral interventions on LLaMA 130M, released to Muon after 25% of training.

6 Conclusion

ORCA uses temporary early orthogonality regularization to accelerate LLM training, achieving lower validation loss. Its central insight is that optimizers can both choose better update directions and steer early training toward regions where loss is easier to reduce. We hope this perspective inspires future optimizer design.

AI USE STATEMENT

We used generative AI tools to assist with LaTeX formatting, language editing for grammar and clarity, and editing experimental code. We did not use these tools to develop the research ideas or methods, analyze data, formulate scientific claims, or interpret results. The authors reviewed all AI-assisted edits to ensure that the scientific meaning was preserved and take full responsibility for the final content of the manuscript.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

Yihe Dong and Will Sawin. Muon^p: Muon with fractional spectral powers. *arXiv preprint arXiv:2606.13867*, 2026.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

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

Lingkai Kong, Yuqing Wang, and Molei Tao. Momentum stiefel optimizer, with applications to suitably-orthogonal attention, and optimal transport. *arXiv preprint arXiv:2205.14173*, 2022.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

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

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Gabriel Peyré. Muon dynamics as a spectral wasserstein flow. *arXiv preprint arXiv:2604.04891*, 2026.

Zeju Qiu, Simon Buchholz, Tim Xiao, Maximilian Dax, Bernhard Schölkopf, and Weiyang Liu. Reparameterized llm training via orthogonal equivalence transformation. *Advances in Neural Information Processing Systems*, 38:140775–140821, 2026a.

Zeju Qiu, Lixin Liu, Adrian Weller, Han Shi, and Weiyang Liu. Poet-x: Memory-efficient llm training by scaling orthogonal transformation. *arXiv preprint arXiv:2603.05500*, 2026b.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Kexuan Shi, Hanxuan Li, Zeju Qiu, Yandong Wen, Simon Buchholz, and Weiyang Liu. Pion: A spectrum-preserving optimizer via orthogonal equivalence transformation. *arXiv preprint arXiv:2605.12492*, 2026.

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

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Senmiao Wang, Tiantian Fang, Haoran Zhang, Yushun Zhang, Kunxiang Zhao, Alex Schwing, and Ruoyu Sun. Pc layer: Polynomial weight preconditioning for improving llm pre-training. *arXiv preprint arXiv:2606.06470*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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*, pp. 4791–4800, 2019.

Hanqing Zhu, Wenyan Cong, Zhizhou Sha, Sagnik Mukherjee, Xinyuan Song, David González-Martínez, Xiaoxia Wu, Yuandong Tian, Shiwei Liu, David Z Pan, et al. Iso: An rlvr-native optimization stack. *arXiv preprint arXiv:2607.19331*, 2026.

Table B.1: LLaMA configurations used in the standard experiments.

Model	Layers	Hidden	FFN	Q/KV heads	Head dim
130M	12	768	2048	12/12	64
350M	24	1024	2736	16/16	64
1.3B	24	2048	5461	32/32	64

Table B.2: Qwen3 and fine-grained MoE configurations.

Model	Layers	Hidden	Dense FFN	Q/KV heads	Experts	Top-k
Qwen3 160M	20	512	1536	8/4	–	–
Qwen3 0.6B	28	1024	3072	16/8	–	–
MoE 0.8B	16	768	1536	6/3	64	4
MoE 8B	27	2048	4096	16/8	64	4

A Orthogonal and spectral training.

Kong et al. (2022) developed a momentum optimizer that maintains orthogonality constraints. For LLMs, POET and POET-X preserve weight spectra through orthogonal reparameterization, while Pion preserves them through optimizer updates (Qiu et al., 2026a;b; Shi et al., 2026). ISO applies fixed-spectrum optimization to post-training (Zhu et al., 2026). These methods report impressive training stability and downstream or post-training results; Pion and ISO outperform Muon in several downstream or post-training evaluations. However, persistent spectral constraints have not yet been shown to consistently achieve lower LLM pretraining loss than Muon. Similar attention to weight spectra appears at the model level: the PC layer improves convergence by carefully reparameterizing weights to reshape their spectra (Wang et al., 2026). Optimizer-side shaping may be more broadly applicable and less costly because it does not modify the model's forward or backward passes. One key difference in our method is that we remove the orthogonal constraint entirely after early training. We believe this freedom later in training is an important reason for our larger improvement over Muon.

B Experiment Setup

This appendix summarizes the standard pretraining configurations used in the main experiments. We omit the controlled ablations and optimizer-specific interventions discussed in the main text. Unless stated otherwise, all runs use a decoder-only Transformer with RoPE, SwiGLU feed-forward blocks, RMSNorm, and bias-free linear layers. The matrix parameters use Muon or ORCA, while embeddings, the language-model head, and one-dimensional parameters follow the AdamW fallback described in the method section.

B.1 Model and Dataset

We train LLaMA-style models on C4, Qwen3-style models on The Pile, and the fine-grained MoE models on the olmo-mix-1124 mixture (Gao et al., 2020; OLMo et al., 2024). The corresponding tokenizers are the T5 SentencePiece Unigram tokenizer (vocabulary size 32,100), the Qwen3 tokenizer (vocabulary size 151,936), and the OLMo tokenizer (vocabulary size 100,278).

For the LLaMA models, attention is multi-head with equal numbers of query and key/value heads, the RoPE base is 10,000, and the RMSNorm epsilon is 10^{-6}. The model configurations used in the paper are:

Qwen3 uses grouped-query attention with head dimension 128 (Ainslie et al., 2023), a RoPE base of 10^6, RMSNorm epsilon 10^{-6}, and the default QK normalization. The MoE models also use grouped-query attention. Their router uses sigmoid scores, top-4 routing, an auxiliary loss coefficient of 10^{-3}, expert bias, a top-k scaling factor of 2, and one shared expert per layer. The first layer is dense and the remaining layers are MoE layers.

Model	Data	Seq. len.	Global batch	Steps	Tokens	Peak LR	WD
LLaMA 130M	C4	1024	512	50,000	26.2144B	searched	0.01
LLaMA 350M	C4	1024	1024	100,000	104.8576B	1×10^{-3}	0.1
LLaMA 1.3B	C4	1024	1024	100,000	104.8576B	6×10^{-4}	0.1
Qwen3 160M	The Pile	4096	256	25,000	26.2144B	1×10^{-3}	0.01
Qwen3 0.6B	The Pile	4096	512	50,000	104.8576B	6×10^{-4}	0.1
MoE 0.8B	olmo-mix-1124	1024	512	50,000	26.2144B	1×10^{-3}	0.1
MoE 0.8B	olmo-mix-1124	1024	512	100,000	52.4288B	1×10^{-3}	0.1
MoE 8B	olmo-mix-1124	2048	1024	50,000	104.8576B	4×10^{-4}	0.1

Table B.3: Training budgets for the standard experiments. Tokens are computed as sequence length times global batch size times optimizer steps.

For the MoE models, the expert and shared-expert FFN widths are 288 for the 0.8B model and 768 for the 8B model. The MoE models use a RoPE base of 10^6, maximum context length 40960, and RMSNorm epsilon 10^{-6}; their input embeddings and output heads are untied. Qwen3 ties its input and output embeddings. The active parameter counts of the two MoE models are 0.24B and 1B, respectively.

B.2 Training Setup

All standard runs use a 1,000-step linear warmup followed by a WSD schedule. The final 15% of the training steps is a linear decay from the peak learning rate to a minimum learning rate equal to one tenth of the peak. Thus, the middle portion of training uses the peak learning rate. The cosine schedule appears only in the omitted schedule-control ablation.

Table B.3 reports the exact number of processed tokens implied by the global batch size, sequence length, and plotted training steps. The figure titles use rounded budgets (25B, 50B, or 100B tokens), whereas the values in the table are not rounded.

For standard ORCA runs, the shaping strength is $\gamma = 1$ and the active window is the first 25% of training. In the learning-rate sweep, the Muon and ORCA matrix updates are additionally rescaled by $0.2\,\mathbf{O}_t/\,\mathrm{Rms}(\mathbf{O}_t)$, following the matched update normalization used in the main text. For the LLaMA 130M learning-rate sweep, the selected peak learning rates are 1.3×10^{-3} for Adam, 3×10^{-3} for Muon, 1×10^{-3} for Pion, and 2×10^{-3} for ORCA. The remaining models use the peak rates listed in Table B.3.

B.3 Implementation Details

All standard Muon and ORCA runs use eight Newton–Schulz iterations with the Polar Express coefficients and Nesterov momentum of 0.95. The ORCA shaping window is released layerwise for the MoE experiments: the first layer is released at 20% of training, the last layer at 60%, and intermediate layers use linear interpolation. The MoE 8B experiment compares Muon and ORCA under the available budget; the 0.8B experiments additionally include AdamW as shown in the main figure.

C More Experiment Results

C.1 Robustness across random seeds

We repeated the LLaMA 130M experiment for four random seeds under the same 50k-step training protocol. Table C.1 reports the final validation loss for each seed together with the mean and the sample variance across seeds. ORCA achieves the lowest mean loss and the smallest across-seed variance. Relative to Muon, ORCA lowers the mean final loss by 0.0387 and reduces the sample variance by approximately an order of magnitude. The ordering ORCA $<$ Muon $<$ Adam is preserved for every seed, indicating that the observed improvement is not tied to a particular initialization.

Table C.1: Final validation loss across four random seeds for LLaMA 130M at 50k steps. The final two columns report the mean and sample variance across seeds; summary statistics are computed from the displayed values.

Method	Seed 0	Seed 1	Seed 2	Seed 3	Mean	Variance
Adam	2.882	2.880	2.877	2.884	2.8808	8.92×10^{-6}
Muon	2.846	2.851	2.852	2.853	2.8505	9.67×10^{-6}
ORCA	2.811	2.813	2.812	2.811	**2.8118**	$\mathbf{9.17 \times 10^{-7}}$

Variance is the sample variance over the four seeds.

Table C.2: Optimizer-state and learning-rate-schedule controls.

Question	Reference setting	Variant setting	Reference loss ↓	Variant loss ↓	Variant − Reference
Optimizer state	ORCA, state retained	ORCA, state reset	2.8107	2.8106	−0.0001
Schedule: WSD	Muon	ORCA	2.8512	2.8106	−0.0406
Schedule: cosine	Muon	ORCA	2.8595	2.8139	−0.0456

C.2 Optimizer-State and Learning-Rate Controls

We test whether ORCA's gain depends on optimizer state carried across release or on the WSD learning-rate schedule (Hu et al., 2024; Loshchilov & Hutter, 2016). Table C.2 reports the final validation loss at 50k steps.

The gain of ORCA does not come from carrying over the optimizer state. At the end of the active window we keep the model parameters unchanged and reinitialize only the Muon optimizer state, then compare against a matched run that retains that state. The final validation losses at 50k steps are 2.8106 with the state reset and 2.8107 with it retained, a difference of 1×10^{-4}. Resetting the optimizer state does not remove the advantage of ORCA over Muon, so that advantage cannot be attributed to the optimizer buffers inherited at release.

The gain of ORCA does not depend on a particular learning-rate schedule. Under both the WSD and the cosine schedule, ORCA is better than the corresponding Muon run, with a final validation loss lower by 0.0406 and 0.0456 respectively. The advantage of ORCA therefore does not depend on the specific decay shape of WSD.

D Proof of the Path-Dependent Loss Lower Bound

D.1 The Exact-Polar Muon Continuous-Time Flow

For any matrix $\mathbf{A}$, let $\mathbf{Q}(\mathbf{A})$ denote a polar direction satisfying

$$\langle \mathbf{A}, \mathbf{Q}(\mathbf{A}) \rangle = \|\mathbf{A}\|_*, \qquad \|\mathbf{Q}(\mathbf{A})\|_2 \le 1. \tag{D.1}$$

If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is a thin singular value decomposition, one may take $\mathbf{Q}(\mathbf{A}) = \mathbf{U}\mathbf{V}^\top$; in the rank-deficient case, an element satisfying (D.1) can be selected from the subdifferential of the nuclear norm. The argument below uses this defining identity directly for any valid choice of polar direction.

Let $\mathbf{G}_t = \nabla_{\mathbf{W}} L(\mathbf{X}_t \mathbf{Y}_t)$. The exact-polar Muon continuous-time flow with per-factor scheduling is defined by requiring, almost everywhere,

$$\dot{\mathbf{X}}_t = -\eta_t^X \mathbf{Q}(\mathbf{G}_t \mathbf{Y}_t^\top), \qquad \dot{\mathbf{Y}}_t = -\eta_t^Y \mathbf{Q}(\mathbf{X}_t^\top \mathbf{G}_t). \tag{D.2}$$

We assume that $k \ge 2$, $(\mathbf{X}_t, \mathbf{Y}_t)$ is absolutely continuous on $[0, T]$, and that the solution exists with $\eta_0 \ge 0$ and $\|\mathbf{X}_t\|_2, \|\mathbf{Y}_t\|_2 > 0$. The per-factor step-size bounds are the same as those in Theorem 2.1:

$$0 \le \eta_t^X \le \frac{\eta_0}{\|\mathbf{Y}_t\|_2}, \qquad 0 \le \eta_t^Y \le \frac{\eta_0}{\|\mathbf{X}_t\|_2}.$$

The common step-size condition

$$\eta_t^X = \eta_t^Y = \eta_t \quad \text{with} \quad 0 \le \eta_t \le \frac{\eta_0}{\max\{\|\mathbf{X}_t\|_2, \|\mathbf{Y}_t\|_2\}} \tag{D.3}$$

is a special case.

D.2 Two Auxiliary Lemmas

Lemma D.1 (Nuclear norm of a product)**.** *For any conformable matrices* $\mathbf{A}, \mathbf{B}$*,*

$$\|\mathbf{AB}\|_* \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F . \tag{D.4}$$

Proof. By the duality between the nuclear and spectral norms, for any $\|\mathbf{M}\|_2 \le 1$,

$$\langle \mathbf{AB}, \mathbf{M} \rangle = \langle \mathbf{A}, \mathbf{MB}^\top \rangle \le \|\mathbf{A}\|_F \left\|\mathbf{MB}^\top\right\|_F \le \|\mathbf{A}\|_F \|\mathbf{M}\|_2 \|\mathbf{B}\|_F \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F .$$

Taking the supremum over $\mathbf{M}$ gives (D.4). □

Lemma D.2 (Gradient bound for the cross-entropy objective)**.** *In the setting of* (2.2)*, if* $\mathbf{G} = \nabla_{\mathbf{W}} L(\mathbf{W})$ *then*

$$\|\mathbf{G}\|_F \le \sqrt{2}\, L(\mathbf{W}). \tag{D.5}$$

Proof. Let $\mathbf{p}_i$ be the softmax vector of column i, so that the corresponding gradient is $\mathbf{g}_i = \mathbf{p}_i - \mathbf{e}_{c_i}$. Then

$$\begin{aligned} \|\mathbf{g}_i\|_2^2 &= (1 - p_{c_i i})^2 + \sum_{j \ne c_i} p_{ji}^2 \\ &\le (1 - p_{c_i i})^2 + \left(\sum_{j \ne c_i} p_{ji} \right)^2 = 2(1 - p_{c_i i})^2. \end{aligned}$$

Since $p_{c_i i} = e^{-\ell_i}$ and $1 - e^{-x} \le x$ for $x \ge 0$, we get $\|\mathbf{g}_i\|_2 \le \sqrt{2}\, \ell_i$. Therefore

$$\|\mathbf{G}\|_F = \left(\sum_i \|\mathbf{g}_i\|_2^2 \right)^{1/2} \le \sqrt{2} \left(\sum_i \ell_i^2 \right)^{1/2} \le \sqrt{2} \sum_i \ell_i = \sqrt{2} L(\mathbf{W}),$$

which is (D.5). □

D.3 Proof of Theorem 2.1

Since $L(\mathbf{XY})$ is smooth in $(\mathbf{X}, \mathbf{Y})$, the chain rule applies almost everywhere along an absolutely continuous path:

$$\dot{L}_t = \left\langle \mathbf{G}_t \mathbf{Y}_t^\top, \dot{\mathbf{X}}_t \right\rangle + \left\langle \mathbf{X}_t^\top \mathbf{G}_t, \dot{\mathbf{Y}}_t \right\rangle . \tag{D.6}$$

Substituting (D.2) and using the nuclear-norm identity of the polar direction gives

$$-\dot{L}_t = \eta_t^X \left\|\mathbf{G}_t \mathbf{Y}_t^\top\right\|_* + \eta_t^Y \left\|\mathbf{X}_t^\top \mathbf{G}_t\right\|_* \ge 0. \tag{D.7}$$

Applying Lemma D.1 and the step-size condition to the first term,

$$\eta_t^X \left\|\mathbf{G}_t \mathbf{Y}_t^\top\right\|_* \le \frac{\eta_0}{\|\mathbf{Y}_t\|_2} \|\mathbf{G}_t\|_F \|\mathbf{Y}_t\|_F = \eta_0 \|\mathbf{G}_t\|_F \sqrt{\operatorname{erank}(\mathbf{Y}_t)}. \tag{D.8}$$

Similarly,

$$\eta_t^Y \left\|\mathbf{X}_t^\top \mathbf{G}_t\right\|_* \le \eta_0 \|\mathbf{G}_t\|_F \sqrt{\operatorname{erank}(\mathbf{X}_t)}. \tag{D.9}$$

Combining (D.7)–(D.9) and then applying Lemma D.2,

$$-\dot{L}_t \le \sqrt{2}\, \eta_0 L_t \left(\sqrt{\operatorname{erank}(\mathbf{X}_t)} + \sqrt{\operatorname{erank}(\mathbf{Y}_t)} \right). \tag{D.10}$$

For finite matrix entries, (2.2) satisfies $L_t > 0$, so $\log L_t$ is well defined. Dividing (D.10) by L_t and integrating from 0 to T,

$$\log \frac{L_T}{L_0} \ge -\sqrt{2}\, \eta_0 \int_0^T \left(\sqrt{\operatorname{erank}(\mathbf{X}_s)} + \sqrt{\operatorname{erank}(\mathbf{Y}_s)} \right) ds.$$

Exponentiating gives the bound stated in Theorem 2.1.

E Proof of the Endpoint Loss Lower Bound

E.1 A Linear Lower Bound on the Cross-Entropy Loss

For $d \geq 2$ and an index $c \in [d]$, define the unit vector

$$\mathbf{v}_c := \sqrt{\frac{d-1}{d}}\mathbf{e}_c - \frac{1}{\sqrt{d(d-1)}}\sum_{j\neq c}\mathbf{e}_j, \qquad \|\mathbf{v}_c\|_2 = 1. \tag{E.1}$$

Fix a single column $\mathbf{w} \in \mathbb{R}^d$, whose loss is

$$\ell_c(\mathbf{w}) = \log\left(1 + \sum_{j\neq c} e^{w_j - w_c}\right). \tag{E.2}$$

By the AM–GM inequality,

$$\sum_{j\neq c} e^{w_j - w_c} \geq (d-1)\exp\left[-\left(w_c - \frac{1}{d-1}\sum_{j\neq c} w_j\right)\right]. \tag{E.3}$$

A direct computation gives

$$w_c - \frac{1}{d-1}\sum_{j\neq c} w_j = \sqrt{\frac{d}{d-1}}\,\langle \mathbf{v}_c, \mathbf{w}\rangle. \tag{E.4}$$

For any $R > 0$, write

$$\beta_{d,R} := R\sqrt{\frac{d}{d-1}}\,\frac{(d-1)e^{-R\sqrt{d/(d-1)}}}{1 + (d-1)e^{-R\sqrt{d/(d-1)}}} > 0, \tag{E.5}$$

and let $f(t) = \log(1 + (d-1)e^{-t})$ with $t_0 = R\sqrt{d/(d-1)}$. Since f is convex, its tangent at t_0 is a global lower bound. Combining this with (E.3)–(E.4),

$$\begin{aligned} \ell_c(\mathbf{w}) &\geq f\left(\sqrt{\frac{d}{d-1}}\,\langle \mathbf{v}_c, \mathbf{w}\rangle\right) \\ &\geq f(t_0) + \frac{\beta_{d,R}}{R}\left(R - \langle \mathbf{v}_c, \mathbf{w}\rangle\right) \\ &\geq \beta_{d,R} - \frac{\beta_{d,R}}{R}\langle \mathbf{v}_c, \mathbf{w}\rangle, \end{aligned} \tag{E.6}$$

where the last step drops the nonnegative term $f(t_0)$. The bound holds for every $\mathbf{w}$, with R selecting the tangent point.

E.2 Proof of Theorem 2.2

For each target index c, write

$$S_c = \{i : c_i = c\}, \qquad m_c = |S_c|, \tag{E.7}$$

and let $\mathbf{1}_{S_c} \in \mathbb{R}^n$ be the indicator vector of S_c. Summing (E.6) over $i \in S_c$ and applying the Cauchy–Schwarz inequality,

$$\begin{aligned} \sum_{i\in S_c} \ell_c(\mathbf{w}_i) &\geq \beta_{d,R} m_c - \frac{\beta_{d,R}}{R}\left\langle \mathbf{v}_c, \sum_{i\in S_c}\mathbf{w}_i\right\rangle \\ &\geq \beta_{d,R}\left(m_c - \frac{\|\mathbf{W}\|_2}{R}\sqrt{m_c}\right), \end{aligned} \tag{E.8}$$

where we used

$$\left\|\sum_{i\in S_c}\mathbf{w}_i\right\|_2 = \|\mathbf{W}\mathbf{1}_{S_c}\|_2 \leq \|\mathbf{W}\|_2\sqrt{m_c}.$$

Summing over all c,

$$L(\mathbf{W}) \geq \beta_{d,R}\left[n - \frac{\|\mathbf{W}\|_2}{R}\sum_{c=1}^{d}\sqrt{m_c}\right]. \tag{E.9}$$

Now take $R = \|\mathbf{W}\|_F/\sqrt{n}$ and write $C_d(\mathbf{W}) := \beta_{d,R}$. Under the bounded-scale assumption in Theorem 2.2, $R \leq \|\mathbf{W}\|_2 \leq B = O(1)$, so the tangent scale used in the bound remains in the prescribed $O(1)$ regime. Since only K of the m_c are nonzero, the Cauchy–Schwarz inequality gives

$$\sum_{c=1}^{d}\sqrt{m_c} \leq \sqrt{K\sum_{c} m_c} = \sqrt{Kn}, \tag{E.10}$$

while

$$\frac{\|\mathbf{W}\|_2}{R} = \frac{\sqrt{n}\,\|\mathbf{W}\|_2}{\|\mathbf{W}\|_F} = \sqrt{\frac{n}{\operatorname{erank}(\mathbf{W})}}. \tag{E.11}$$

Substituting (E.10)–(E.11) into (E.9) yields

$$L(\mathbf{W}) \geq C_d(\mathbf{W})\, n\left(1 - \sqrt{\frac{K}{\operatorname{erank}(\mathbf{W})}}\right), \tag{E.12}$$

For the square setting of Theorem 2.2, set $d = n = k$; this is exactly the claimed endpoint bound. In the unit-scale setting relevant to the discussion, $\|\mathbf{W}\|_2 = 1$ gives $R = \|\mathbf{W}\|_F/\sqrt{k} = \sqrt{\operatorname{erank}(\mathbf{W})/k} \in [1/\sqrt{k}, 1]$. Since $R \mapsto \beta_{k,R}$ is continuous and strictly positive on this interval, $\underline{C}_k := \min_{R\in[1/\sqrt{k},1]}\beta_{k,R} > 0$ and $C_k(\mathbf{W}) \geq \underline{C}_k$. Consequently, for any $\delta > 0$, the regime $\operatorname{erank}(\mathbf{W}) \geq K + \delta$ has the uniform lower bound

$$L(\mathbf{W}) \geq \underline{C}_k\, k\left(1 - \sqrt{\frac{K}{K+\delta}}\right) > 0.$$

E.3 How the Factor Effective Ranks Control the Product

Lemma E.1 (Two-factor product). *For any nonzero $\mathbf{X} \in \mathbb{R}^{d\times k}$ and $\mathbf{Y} \in \mathbb{R}^{k\times n}$ with $\mathbf{XY} \neq 0$,*

$$\operatorname{erank}(\mathbf{XY}) \geq \operatorname{erank}(\mathbf{X}) + \operatorname{erank}(\mathbf{Y}) - k. \tag{E.13}$$

Proof. Let $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$ and $\mathbf{M} = \mathbf{Y}\mathbf{Y}^\top$, normalized as

$$\widehat{\mathbf{A}} = \frac{\mathbf{A}}{\|\mathbf{X}\|_2^2}, \qquad \widehat{\mathbf{M}} = \frac{\mathbf{M}}{\|\mathbf{Y}\|_2^2}.$$

Then $0 \preceq \widehat{\mathbf{A}}, \widehat{\mathbf{M}} \preceq \mathbf{I}$, and

$$\operatorname{tr}(\widehat{\mathbf{A}}) = \operatorname{erank}(\mathbf{X}), \qquad \operatorname{tr}(\widehat{\mathbf{M}}) = \operatorname{erank}(\mathbf{Y}).$$

Both $\mathbf{I} - \widehat{\mathbf{A}}$ and $\mathbf{I} - \widehat{\mathbf{M}}$ are positive semidefinite, and the trace of a product of two positive semidefinite matrices is nonnegative, so

$$0 \leq \operatorname{tr}\Big[(\mathbf{I} - \widehat{\mathbf{A}})(\mathbf{I} - \widehat{\mathbf{M}})\Big] = k - \operatorname{erank}(\mathbf{X}) - \operatorname{erank}(\mathbf{Y}) + \operatorname{tr}(\widehat{\mathbf{A}}\widehat{\mathbf{M}}). \tag{E.14}$$

On the other hand,

$$\|\mathbf{XY}\|_F^2 = \operatorname{tr}(\mathbf{AM}) = \|\mathbf{X}\|_2^2\,\|\mathbf{Y}\|_2^2\,\operatorname{tr}(\widehat{\mathbf{A}}\widehat{\mathbf{M}}),$$

and $\|\mathbf{XY}\|_2^2 \leq \|\mathbf{X}\|_2^2\,\|\mathbf{Y}\|_2^2$. Therefore

$$\operatorname{erank}(\mathbf{XY}) = \frac{\|\mathbf{XY}\|_F^2}{\|\mathbf{XY}\|_2^2} \geq \operatorname{tr}(\widehat{\mathbf{A}}\widehat{\mathbf{M}}) \geq \operatorname{erank}(\mathbf{X}) + \operatorname{erank}(\mathbf{Y}) - k,$$

where the last step uses (E.14). □