Accelerating LLM Pre-Training through Flat-Direction Dynamics Enhancement

Shuchen Zhu[1,*] · Rizhen Hu[1,*] · Mingze Wang[1,†] · Mou Sun[2]
Xue Wang[3] · Kun Yuan[1,‡] · Zaiwen Wen[1]

[1]Peking University · [2]Zhejiang Lab · [3]Meituan

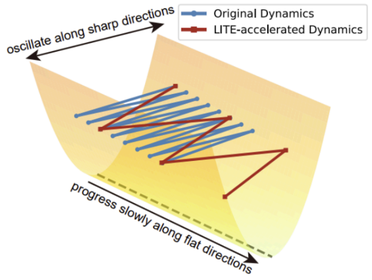
* Equal contribution † Project lead ‡ Corresponding author

Neurips 2026

CODE

Motivation

Progress along sharp directions (Valley) oscillates, whereas progress along flat directions (River) is slow but dominates final loss reduction



Enhance flat-direction dynamics

Euclidean ODE: a continous viewpoint

$$\ddot{w}_t + \alpha_t \dot{w}_t + \beta_t \nabla^2 f(w_t)\dot{w}_t + \gamma_t \nabla f(w_t) = 0$$

- γ_t : gradient driving force — increase it in flat directions.
- $\alpha_t I + \beta_t \nabla^2 f(w_t)$: Hessian damping (momentum mechanism)

Discrete algorithm for adaptive optimizers

$$m_k = (1-\alpha)m_{k-1} + \nabla f(w_k)$$
$$w_{k+1} = w_k - \eta_k F(w_k)^{-1}(m_k + \beta \nabla f(w_k))$$

Semi-implicit Euler discretization. $\beta = 0$ → Heavy Ball; $\beta > 0$ → Nesterov-type momentum. The gradient difference supplies implicit Hessian damping.

LITE Strategy

1 · Momentum $m_k = (1-\alpha)m_{k-1} + \nabla f(w_k)$

2 · Estimate Q_k , the projection to flat directions

$$u_k = (I - Q_k)F_k^{-1}(m_k + \beta_1 \nabla f(w_k)) + \chi Q_k F_k^{-1}(m_k + \beta_2 \nabla f(w_k))$$

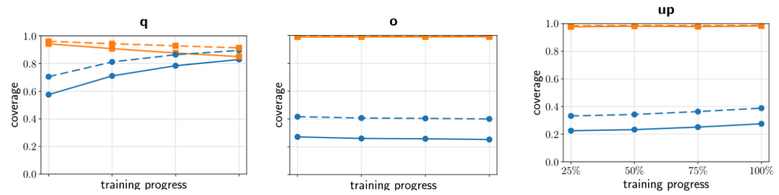
3 · Parameters $w_{k+1} = w_k - \eta_k u_k$

Sharp · $P_k = I - Q_k$: keep (η_k, β_1) for stability. **Flat ·** Q_k : use $(\chi\eta_k, \beta_2)$.
$\chi \geq 1,\ \beta_2 \geq \beta_1 \geq 0$ · $\chi = 1,\ \beta_1 = \beta_2$ recovers the base update. In pratice, we set χ first warm-up, then decay.

Muon-LITE

Observation: Hessian–gradient top-space alignment

Top Hessian eigenspaces are well covered by gradient singular subspaces throughout training (120M LLaMA, layer 2; $k = 200$).



Blue: left · orange: right gradient subspaces · solid: $s = 50$ · dashed: $s = 100$

Flat-direction estimation: Top subspace of momentum

$$\widetilde{M}_k = M_k + G_k/(1-\alpha)$$
$$T_k = \mathrm{msign}(\widetilde{M}_k) + \tfrac{1}{2}\,\mathrm{msign}\Big(\widetilde{M}_k/\tau_k - \mathrm{msign}(\widetilde{M}_k)\Big)$$
$$P_k = T_k^\top T_k, \qquad Q_k = I - P_k$$

Composite msign filtering: adapt τ_{k+1} from $\mathrm{tr}(P_k)$ to maintain $\mathrm{rank}(P_k) \approx d_s$ (the sharp-subspace dim). Two additional msign operations · no extra optimizer state ·

Theory: Two-stage training dynamics (informal)

Stage 1 · River attraction. The trajectory enters an $O(\varepsilon^{1/2})$ neighborhood after $\Theta(\log(\varepsilon^{-1}))$ time.

$$\|w_t - \Phi(w_t)\|_2^2 \lesssim \exp(-\Theta(t)) + \varepsilon$$

Stage 2 · Acceleration on River. The projected trajectory $z_t = \Phi(w_t)$ tracks the decoupled flat-direction dynamics; in the unperturbed case:

$$\tfrac{1}{T}\int_0^T \eta_t \|P_{\mathcal{R}} F^{-1} \nabla f(z_t)\|_F^2 \, dt \leq \tfrac{2(f(z_0) - \inf f)}{\chi \beta_2 T}$$

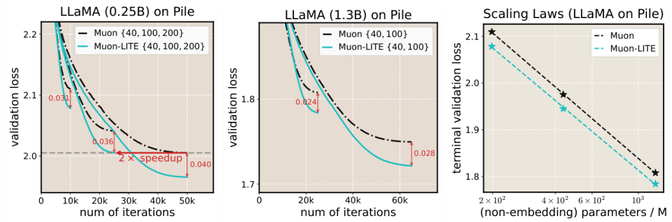
$$\tfrac{1}{T}\int_0^T \eta_t \|s_t\|_F^2 \, dt \leq \tfrac{2(f(z_0) - \inf f)}{\chi \alpha T}$$

χ tightens both time-average bounds; β_2 further tightens the gradient bound — larger flat-direction learning rate and Hessian damping foster optimization.

Experimental Results

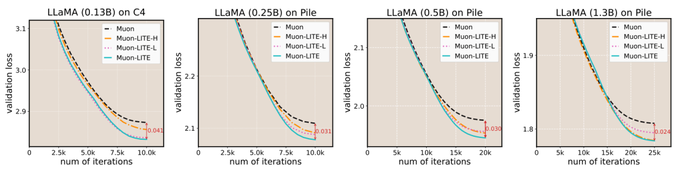
Setup · C4: global batch size 3,600 · sequence length 256
Pile: batch size 1,024 (0.25B / 0.5B), 2,048 (1.3B) · sequence length 1,024

Long-horizon training & scaling laws



40× → 200× token budgets: the loss gap persists; the scaling law shifts downward.

Dense LLaMA · ablations & downstream

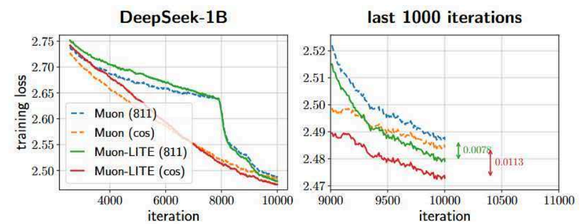


Full Muon-LITE outperforms only amplifying flat-direction LR (L) and only amplifying Hessian damping (H).

Method	ARC_E	ARC_C	PIQA	HellaSwag	BOOLQ	WinoGrande	MMLU	T_mc2	AVG
Muon	52.99	22.10	67.41	35.30	52.45	53.35	23.76	40.14	43.44
Muon-LITE	**54.38**	**23.89**	**67.63**	**36.72**	**60.52**	**54.78**	**25.15**	**40.83**	**45.49**

Limitation

By amplifying update signals in flat directions, LITE also amplifies noise. Its acceleration gains may diminish in practical mixed-precision training.



OUR NEXT WORK

Curvature-Conditioned Multiscale Momentum with Sphere Constraints for LLM Pretraining
Improves the signal-to-noise ratio in flat directions.