ICML 2026

Stabilizing Native Low-Rank LLM Pretraining

Paul Janson · Édouard Oyallon · Eugene Belilovsky

Concordia University · Mila — Quebec AI Institute · Sorbonne University, CNRS

Low rank is attractive — but not for pretraining

- LLMs keep scaling; **memory and compute** are the bottleneck.
- Factorize weights $W = AB^\top$ with rank $r < \min(m, n)$ → **fewer params & FLOPs**.
- Huge success in **fine-tuning** (LoRA) — but **pretraining from scratch** has **no stable recipe**.
- Prior methods only work by keeping full-rank "guidance" weights.

The question

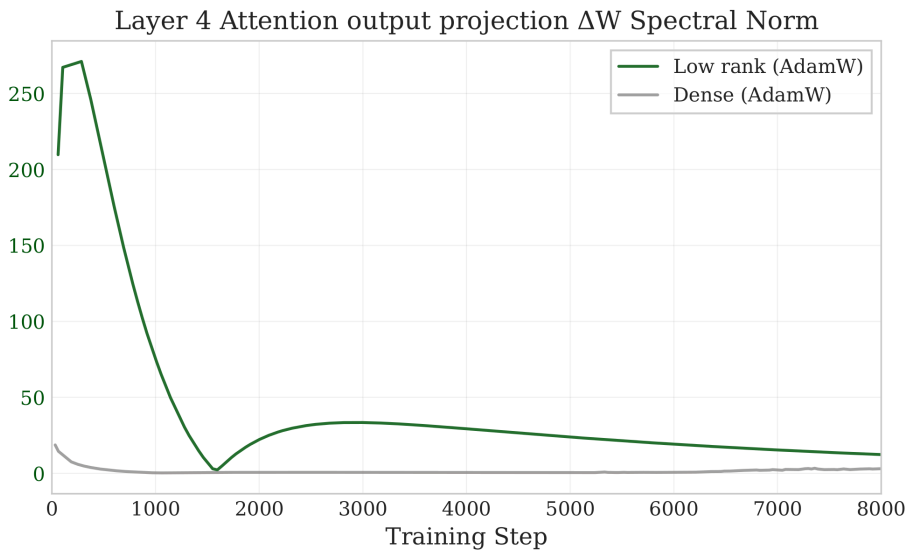
Can we train all non-embedding weights in native low-rank form, from random init — and match the dense model?

Why native low-rank training is unstable

- The factorization is **scale-invariant**:

$$W = (\lambda A)\left(\frac{1}{\lambda}B\right)^{\top}, \quad \forall \lambda > 0$$

- Nothing stops λ from growing → the update's **spectral norm $\|\Delta W\|_2$ explodes**.
- Exploding activations → loss spikes → divergence.
- We pinpoint this as the **dominant cause** of failure — **not** a capacity limit.



Low-rank updates show 10–30× the spectral norm of dense training.

Spectron = orthogonalization + spectral renormalization

Updating factors couples them multiplicatively:

$$\Delta W = \Delta A B^\top + A \Delta B^\top + \Delta A \Delta B^\top$$

Goal: keep $\|\Delta W\|_2 \leq \eta$. Two ingredients:

- **Orthogonalize** each factor update (Newton–Schulz, Muon) → unit singular values.
- **Rescale adaptively** by the factor norms:

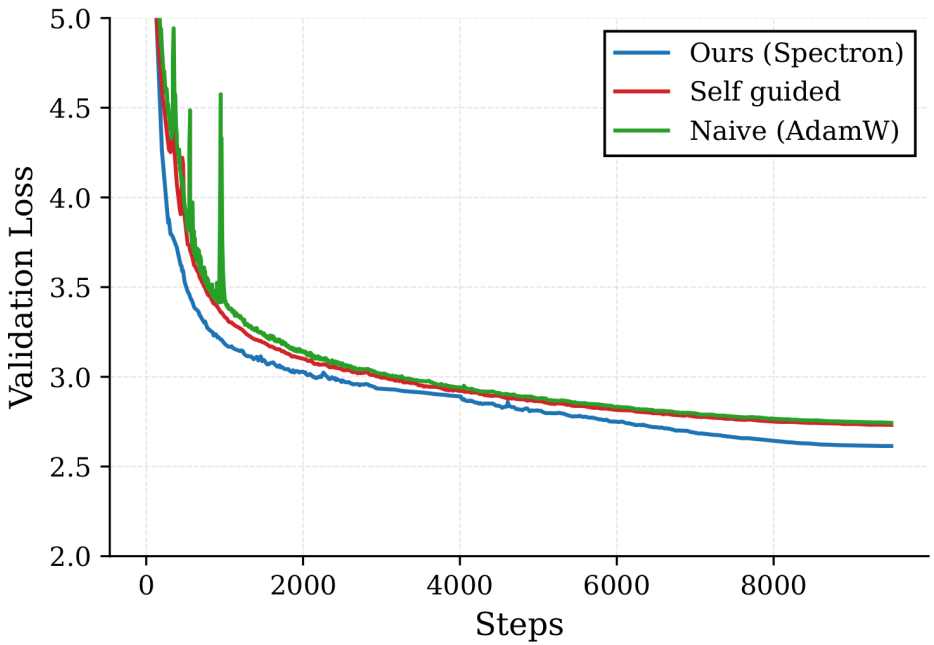
$$\rho = \frac{\eta}{\|A\|_2 + \|B\|_2 + 1}$$

Why it works

Submultiplicativity gives
$\|\Delta W\|_2 \leq \rho(\|A\|_2 + \|B\|_2 + 1) = \eta$
→ composite update is **provably bounded**.

- Spectral norms via **1 power iteration** — only $2mn$ FLOPs.
- **< 1%** overhead, **no auxiliary dense weights**.

Beats prior low-rank training methods

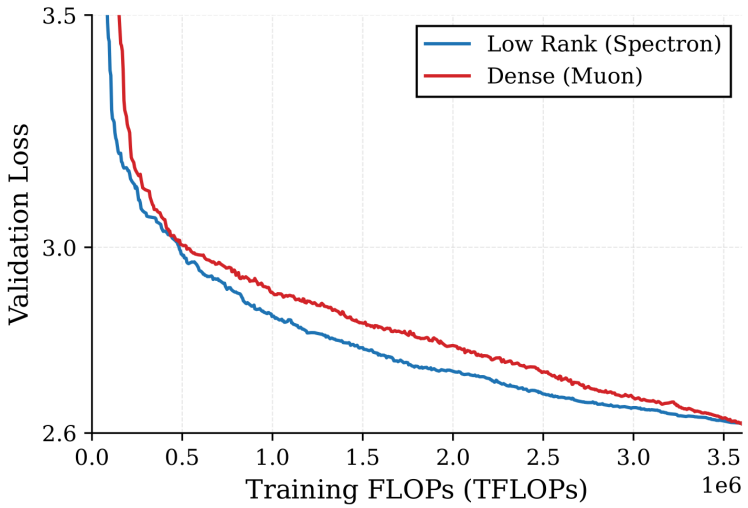


Factorized Transformer-M (297M) on FineWeb.

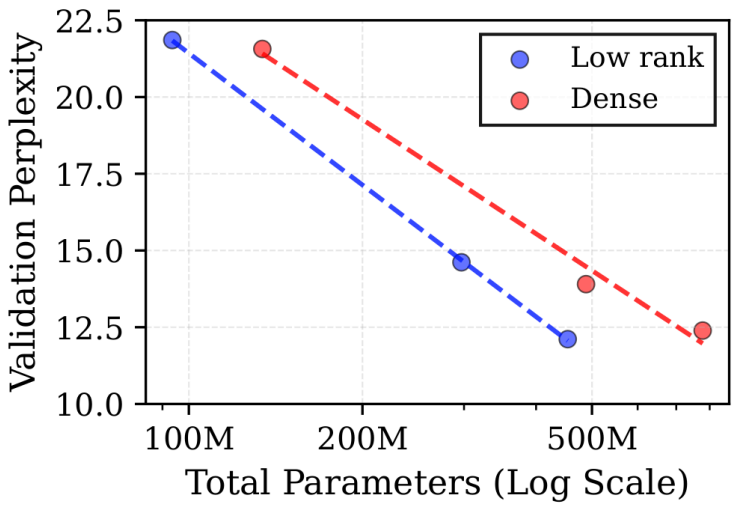
- Naive AdamW **spikes / diverges**.
- Self-guided is stable but needs **+25% FLOPs** of dense guidance.
- Spectron: faster convergence **and** lower final loss at **< 1% overhead**.

Across S / M / L (94–454M):
6–12% lower perplexity vs self-guided,
6–17% vs naive — with downstream gains.

Matches dense quality with far fewer parameters



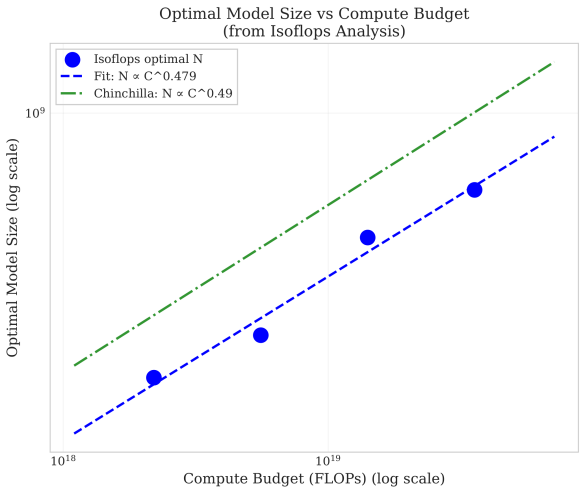
454M low-rank vs 780M dense, equal FLOPs.

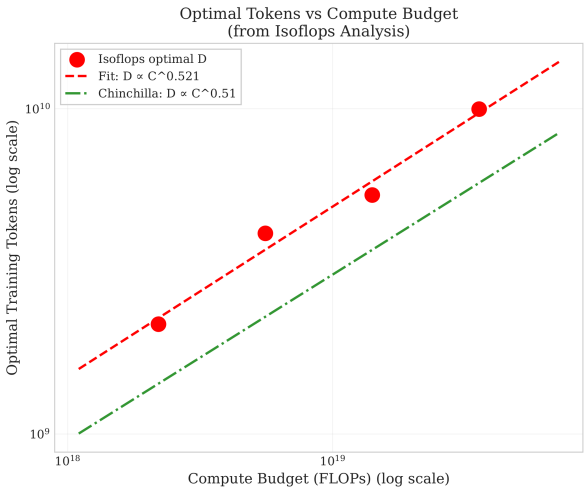


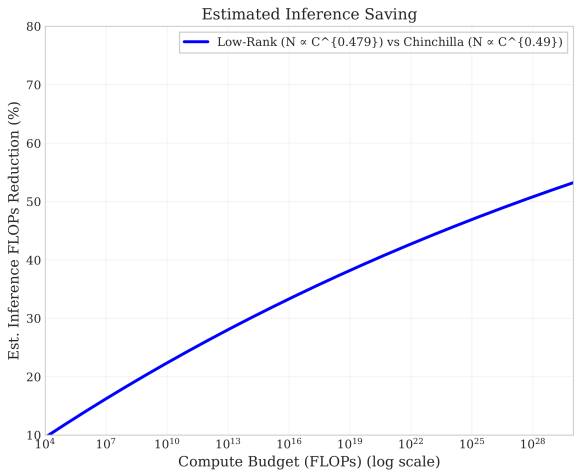
Lower perplexity at every parameter budget.

At equal compute, our 454M model **matches the 780M dense model** — a **42% parameter reduction**, i.e. cheaper inference.

Compute-optimal scaling laws for low-rank transformers







$N_{\text{opt}} \propto C^{0.479}$

(Chinchilla: 0.49)

$D_{\text{opt}} \propto C^{0.521}$

(Chinchilla: 0.51)

up to 50%

inference-cost saving

39 IsoFLOP runs, 47M–1.5B params. Low-rank optima are **smaller and more token-hungry** than dense.

Takeaways

- **Native low-rank pretraining is feasible.** The blocker was spectral instability, not model capacity.
- **Spectron** — orthogonalize + spectrally renormalize factor updates. Provably bounded, < 1% overhead, no auxiliary dense weights.
- Matches dense at equal compute with **42% fewer parameters** follows predictable Chinchilla-style scaling; large inference savings.

Thank you! paul.janson@mila.quebec