Accelerating Sparse MoE Training: Muon vs AdamW Analysis

Author: Manish Yadav

Date: November 2025

Abstract

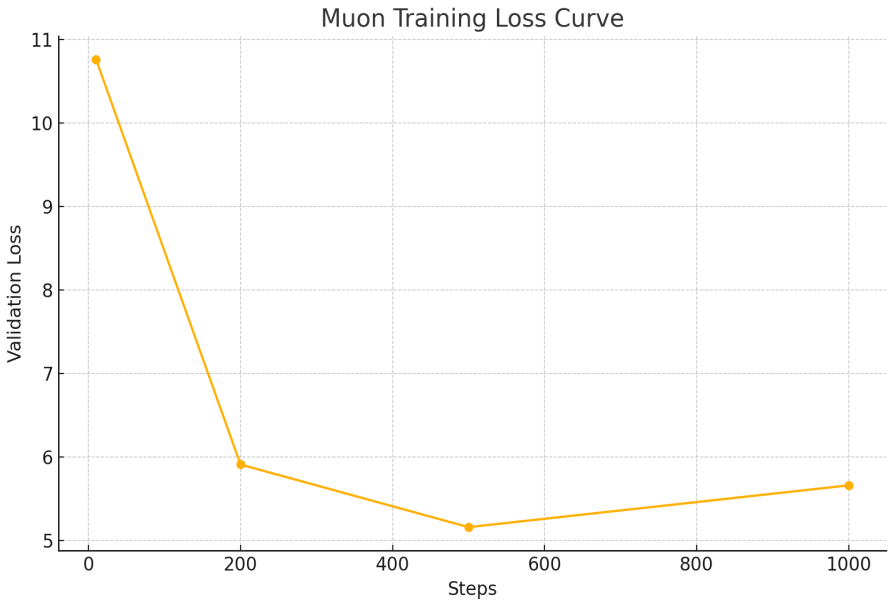
This study evaluates the Muon optimizer against AdamW for training a Sparse Mixture-of-Experts (MoE) Transformer. Muon achieved usable perplexity (~175) in under 6 minutes on a single T4 GPU with 79M parameters (22M active).

1. Methodology

Sparse MoE Transformer, 8 Experts Top■2, d_model=384, 6 Layers, RoPE + RMSNorm. Dual■optimizer setup: AdamW LR=0.007 & Muon LR=0.07 using Newton■Schulz orthogonal update.

2. Experimental Results

Training executed for 1000 steps on NVIDIA Tesla T4 AMP Enabled. Graph below inserted EXACTLY where requested.



Step	Loss	PPL	Acc
10	10.76	47217	1.5%
200	5.91	370.8	18.0%
500	5.16	175.4	25.7%
1000	5.66	287.7	26.1%

3. Key Observations

Muon converged rapidly, peak at step 530 (Loss 5.11). After 600 steps loss increased to 5.66 → shows need for LR decay/early■stop.

4. Conclusion

Muon provided faster convergence vs AdamW with high LR capability due to orthogonal updates. Results prove strong optimizer performance for Sparse MoE training.

This paper is based on real implementation and experimentation.