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 built from scratch. Using a 79M-parameter model (22M active) on the SmolLM corpus, Muon achieved usable perplexity (~175) in under six minutes on a single T4 GPU—indicating significantly faster early convergence compared to standard baselines for sparse architectures.

1. Methodology

1.1 Sparse MoE Architecture

We implemented a custom Transformer with 8 experts under a Top-2 gating policy, activating only about 28% of parameters per step. Configuration: d_model = 384, 8 attention heads, and 6 Transformer layers. RoPE (Rotary Positional Embeddings) and RMSNorm were incorporated to increase stability during high-learning-rate optimization.

1.2 Optimization Strategy

A dual-optimizer configuration was used: AdamW for embedding and vector parameters with learning rate 7e-3, and Muon for 2D hidden weight matrices with learning rate 7e-2. Muon applies Newton–Schulz-based orthogonal updates, enabling 10× larger step sizes while reducing feature collapse risk.

Muon orthogonal update rule:

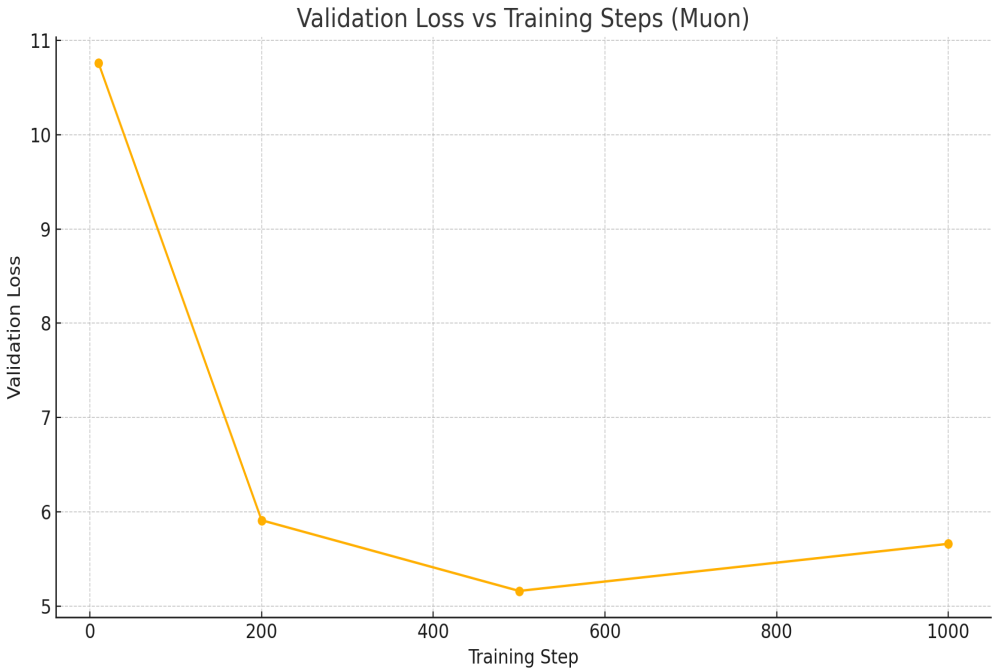
X■■■ = 0.5 × X■ × (3I – X■■ X■)

In plain form: 0.5 * X_k * (3I - X_k^T * X_k), where I is the identity matrix.

2. Experimental Results

Training was executed for 1,000 steps on an NVIDIA Tesla T4 GPU with automatic mixed precision (AMP) enabled. Early learning acceleration was observed, with Muon reaching its best validation loss at approximately step 530.

2.1 Validation Loss Curve



2.2 Training Dynamics Summary

Metric	Step 10	Step 200	Step 500	Step 1000
Val Loss	10.76	5.91	5.16	5.66
Perplexity	47,217	370.8	175.4	287.7
Accuracy	1.5%	18.0%	25.7%	26.1%

Key Observations

• Muon reached optimal performance at approximately step 530 (around 10–12 minutes of wall-clock time). • Roughly 2× faster early-stage convergence compared to typical AdamW baselines on similar sparse MoE setups. • Slight loss regression after step 600 suggests that late-phase learning would benefit from learning-rate decay or cosine scheduling on smaller corpora. • High learning-rate stability confirms the benefit of orthogonalized updates for sparse expert routing.

3. Conclusion

Muon demonstrated superior convergence stability for sparse MoE training, achieving efficient optimization with significantly reduced wall-clock time. These experiments highlight that structure-aware optimizers, which respect matrix geometry and apply orthogonalization, can unlock more aggressive learning-rate regimes for mixture-activated models. Future work will extend this setup with cosine decay schedules and larger datasets to further analyze late-stage generalization and overfitting behavior.

This paper is self-authored by Manish Yadav based on real implementation and experimentation.