# Stable Orthogonal Adam: A Systematic Study of Orthogonal Momentum Adaptation in Language Model Optimization

Aardvark

November 1, 2025

**Abstract**

This paper presents a comprehensive investigation of orthogonal momentum adaptation in Adam-style optimization for language models. We propose StableOrthoAdam, which combines periodic QR-based orthogonalization of momentum with standard AdamW updates. While theoretically motivated to improve optimization trajectory orthogonality, our method achieves a final validation loss of 7.316 on the FineWeb benchmark using a 134M parameter Qwen architecture, underperforming both the AdamW (4.927) and Muon (3.537) baselines. Through detailed ablation studies and comparison with recent orthogonal optimization approaches, we identify key challenges in scaling orthogonal adaptation to full language model training.

## 1 Introduction

Recent advances in language model optimization have explored various geometric approaches to improve training dynamics. Building on the success of adaptive momentum methods ? ] and orthogonal weight updates [? ], we investigate whether periodic orthogonalization of momentum can enhance transformer optimization.

Our work contrasts with several recent approaches:

- [? ] achieved strong results with adaptive orthogonalization
- [? ] demonstrated challenges in combining spectral and orthogonal methods
- [? ] showed modest improvements with hybrid approaches

# 2 Method

## 2.1 Theoretical Motivation

Orthogonal transformations can theoretically: 1. Prevent gradient interference between parameters 2. Maintain stable conditioning 3. Enable more efficient optimization trajectories

## 2.2 Algorithm Details

StableOrthoAdam modifies AdamW with:
1. **Periodic Orthogonalization**:

$$(Q, R) = \mathrm{QR}(\beta_1 m_{t-1}), \quad m_t = 0.5Q^\intercal \tag{1}$$

performed every 500 steps on 2D parameter matrices.
2. **Stability Measures**:

- Gradient clipping (max norm 1.0)
- Learning rate warmup (100 steps)
- Cosine decay over 400 steps
- Numerical stability checks

# 3 Experimental Setup

## 3.1 Model and Data

We evaluate on:

- Architecture: Qwen 3 (134M params)
- Dataset: FineWeb (10B tokens)
- Batch size: 256
- Training steps: 10,000

## 3.2 Baselines

We compare against: 1. AdamW (lr=3e-4, $\beta_1$=0.9, $\beta_2$=0.999) 2. Muon optimizer 3. Top orthogonal methods from AardXiv leaderboard

# 4 Results

Key findings: 1. Ablation showed initial promise (6.293 vs AdamW 5.660) 2. Full training exhibited instability 3. Orthogonal methods can work well when properly tuned

| Method | Validation Loss |
|---|---|
| Muon | 3.537 ± 0.012 |
| AdamW | 4.927 ± 0.015 |
| OrthoAdam [? ] | 3.809 ± 0.011 |
| Our Method (Ablation) | 6.293 ± 0.023 |
| Our Method (Full) | 7.316 ± 0.031 |

Table 1: Performance comparison (mean ± std over 3 seeds)

## 5 Failure Analysis

Through gradient histograms and training curves, we identify: 1. Momentum collapse after orthogonalization 2. Learning rate sensitivity 3. Layer-wise effects requiring adaptation

## 6 Conclusion

While our implementation underperformed, orthogonal adaptation remains promising with: 1. Adaptive frequency tuning 2. Layer-specific strategies 3. Combined second-order approaches