# Analysis of Orthogonal Momentum Optimization for Language Models: A Systematic Negative Result

Aardvark

October 26, 2025

**Abstract**

This paper presents a systematic investigation of orthogonal momentum techniques for language model optimization. While our approach showed initial promise, final results on the 134M parameter Qwen architecture demonstrated poorer performance compared to both AdamW (4.93) and Muon (3.54) baselines, achieving a final validation loss of 6.63. We analyze the potential reasons for this underperformance through comprehensive ablation studies.

## 1 Introduction

Recent advances in language model optimization have demonstrated the effectiveness of momentum-based approaches [**?**]. Our work explores orthogonal parameter updates with Nesterov momentum, motivated by theoretical considerations about parameter space geometry.

## 2 Methodology

Our optimizer implements several key components:

### 2.1 Orthogonal Momentum Updates

For matrix parameters, we apply soft orthogonalization through SVD decomposition:

$$U, S, V = \mathrm{SVD}(G), \quad S' = 0.8I + 0.2S, \quad G' = US'V^T \tag{1}$$

where $I$ is the identity matrix.

### 2.2 Training Protocol

We employed gradient clipping at 1.0 and cosine learning rate scheduling with 2000 warmup steps.

## 3 Results

Key results:

- Final validation loss: 6.63 (ours) vs 4.93 (AdamW)
- Required 15% more training steps than AdamW

## 4 Conclusion

Our investigation yielded a clear negative result, suggesting that simple orthogonalization may be insufficient without additional mechanisms.