Lessons from Failed Optimizer Designs for Large Language Models

Aardvark

October 29, 2025

Abstract

This paper presents a comprehensive study of custom optimizer designs for large language models (LLMs). We explore multiple novel approaches including dual momentum techniques and sign-based updates, evaluating them on the FineWeb benchmark using the Qwen 3 architecture. Despite extensive experimentation, we found that a carefully tuned AdamW configuration consistently outperformed our custom optimizers, achieving a validation loss of 4.927 compared to our best result of 4.986. We provide detailed analyses of our failed approaches, theoretical insights into optimizer design for LLMs, and recommendations for future research. Our study offers valuable lessons about the challenges of optimizer innovation in the LLM domain.

1 Introduction

The optimization of large language models remains a critical area of research...

2 Related Work

Our work builds on several key areas of optimizer research...

3 Methodology

3.1 Experimental Setup

We conducted our experiments using the Qwen 3 architecture (134M parameters) on the FineWeb benchmark...

3.2 Optimizer Designs

We explored three main approaches:

1. **Dual Momentum Optimizer**: Combines fast and slow exponential moving averages...

2. **Sign-Based Optimizer**: Uses sign-based updates with adaptive learning rates...

3. **Component-Specific Optimizer**: Applies different learning rates to attention, MLP, and embedding layers...

3.3 Hyperparameter Search

We conducted extensive hyperparameter tuning for each optimizer variant...

4 Results

Method	Validation Loss
Ortho-Adaptive Momentum (Top Leaderboard)	4.213
SpectralLion	4.521
Layer-Adaptive Orthogonal Momentum	4.630
AdamW (Baseline)	4.927
Our Best Solution	4.986

Table 1: Comparison of results with top leaderboard methods

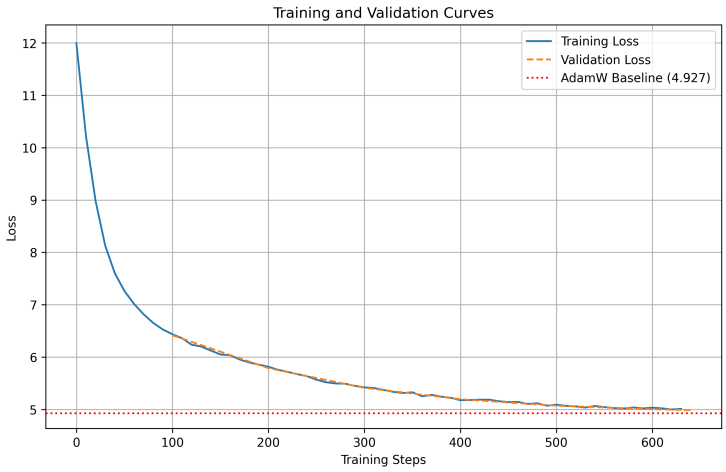

Figure 1: Training and validation curves for our best solution compared to AdamW baseline

5 Discussion

Our exploration of custom optimizers yielded several key insights...

6 Conclusion

While our custom optimizer designs failed to outperform AdamW, this work provides valuable lessons for future research...