Branch: main
Date: December 7-8, 2024
Hardware: 8× NVIDIA H100 80GB HBM3 (via DistributedDataParallel)
Objective: Rigorously test whether Moderate LMD (β=0.98→0.90) significantly improves training performance compared to baseline, and conduct ablation study to verify that benefits arise from layer-wise scheduling rather than global momentum values.

Experimental Design

Based on preliminary sweep results showing Moderate configuration as the most promising LMD variant, we conducted a focused three-way comparison: Baseline vs. Moderate LMD vs. Ablation controls (Fixed High/Low). This phase used 8 GPUs to increase throughput and enable more rigorous statistical testing.

Primary Comparison:

- **Baseline:** Uniform momentum β=0.95 across all 12 layers
- **Moderate LMD:** Linear momentum β=0.98 (layer 0) → β=0.90 (layer 11)

Ablation Study:

- **Fixed High:** Uniform momentum β=0.98 (tests if 0.98 globally is better than 0.95)
- **Fixed Low:** Uniform momentum β=0.90 (tests if 0.90 globally is better than 0.95)

Training details:

- Model: NanoGPT
- Optimizer: Muon (MomentUm Orthogonalized by Newton-Schulz) for transformer blocks, AdamW for output head
- Dataset: FineWeb-10B
- Iterations: 5,100 steps
- Global batch size: 512 (64 per GPU × 8 GPUs)
- Learning rate: 0.0036 (base), 0.00036 (Muon blocks)
- Validation frequency: Every 125 steps

Sample sizes achieved:

- Baseline: n=2
- Moderate: n=4
- Fixed High: n=1
- Fixed Low: n=1

Note: Originally planned for n=10 per primary condition, but completed n=2 baseline and n=4 moderate due to compute budget constraints (~$80 GPU cost). Ablations received n=1 each as they served validation rather than primary hypothesis testing.

Log Files

Configuration	Seeds	Log Files
Baseline	92605, 923	baseline_seed92605_c2411477.txt
Moderate	23513, 38685, 62289, 32437	moderate_seed23513_7b39b493.txt moderate_seed38685_6c71abb2.txt moderate_seed62289_fcec204b.txt
Fixed High	67440	fixed_high_seed67440_cce03dba.txt
Fixed Low	2075	fixed_low_seed2075_4cecc575.txt

Code:

- train_gpt_final.py - Main training script with layer-wise momentum implementation
- train_gpt.py - Original training script (baseline)
- main.ipynb - Analysis notebook for processing results and generating figures

Results Summary

Primary Comparison: Baseline vs. Moderate

Final Validation Loss:

- Baseline: 3.2930 ± 0.0022 (n=2), range [3.2914, 3.2945]
- Moderate: 3.3041 ± 0.0029 (n=4), range [3.3007, 3.3077]
- **Difference:** +0.0111 (+0.34% worse for Moderate)
- **Statistical test:** t=-4.72, p=0.009 → **Statistically significant (p<0.05)**
- **Conclusion:** Moderate LMD significantly degrades final validation loss

Training Time:

- Baseline: 817.42 ± 10.53 s (n=2), range [810.0, 824.9]
- Moderate: 808.59 ± 61.93 s (n=4), range [723.4, 870.1]
- **Difference:** -8.83 s (-1.09% nominally faster for Moderate)
- **Statistical test:** t=0.19, p=0.859 → **Not statistically significant**
- **Conclusion:** No reliable speed difference between configurations

Ablation Study: Fixed High vs. Fixed Low vs. Baseline

Final Validation Loss:

- Baseline: 3.2930 ± 0.0022 (n=2)
- Fixed High (β=0.98): 3.3027 (n=1)
- Fixed Low (β=0.90): 3.2994 (n=1)

Training Time:

- Baseline: 817.42 ± 10.53 s (n=2)
- Fixed High: 824.76 s (n=1)
- Fixed Low: 827.23 s (n=1)