NanoGPT - 125M speedrun

Training time on 8xH100, minutes

- llm.c baseline
- Introduced the Muon optimizer
- Architectural modernizations & tuned learning rate
- Muon improvements
- Pad embeddings & architectural improvements
- Upgraded PyTorch from 2.4.1 to 2.5.0
- Distributed the overhead of Muon
- Shortcuts & tweaks
- Untied embed and lm_head
- U-net & 2x lr
- Bfloat16 activations
- Attention window warmup
- FlexAttention

Date (2024–2025)