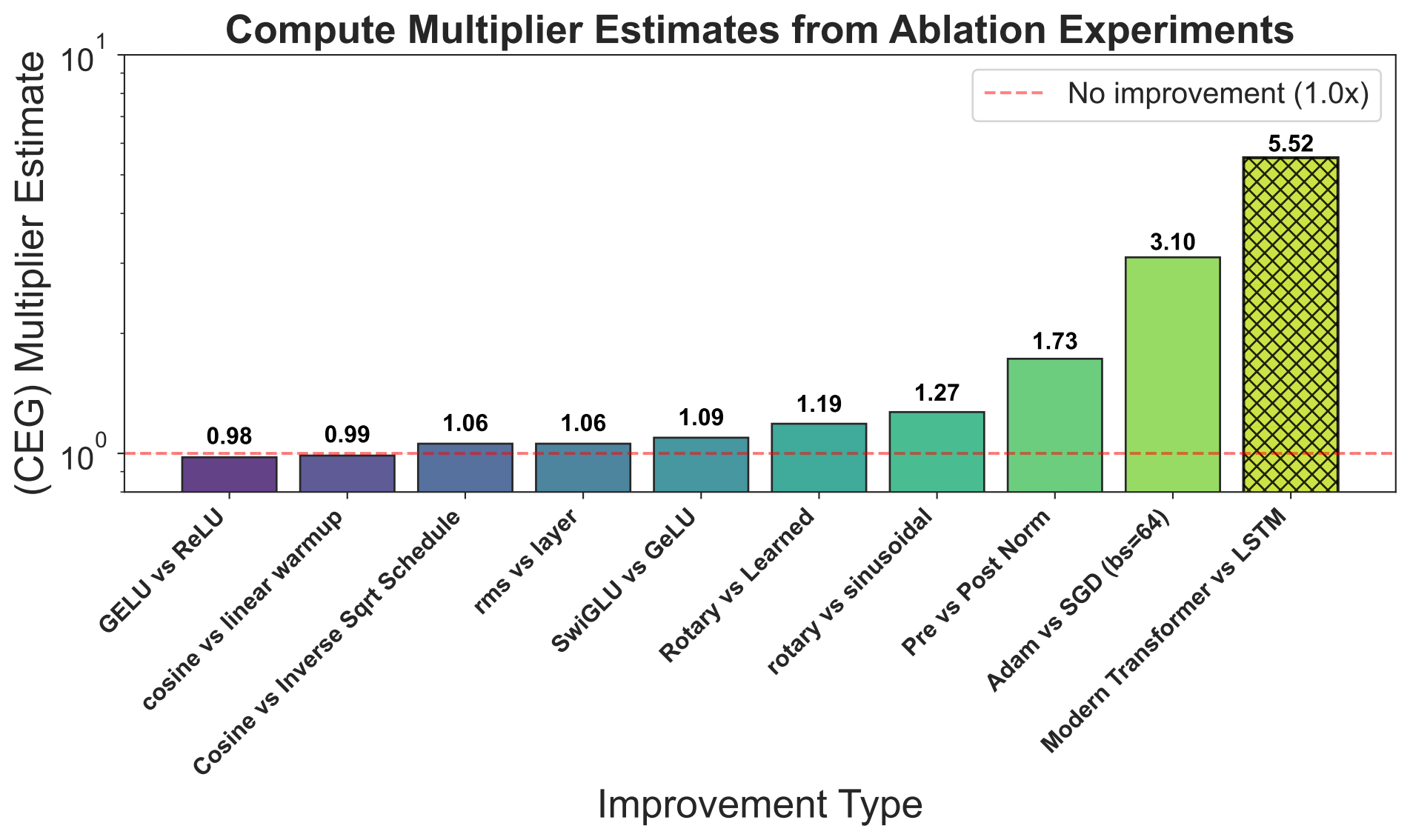
Compute Multiplier Estimates from Ablation Experiments
(CEG) Multiplier Estimate
10^1
10^0
No improvement (1.0x)
0.98
0.99
1.06
1.06
1.09
1.19
1.27
1.73
3.10
5.52
GELU vs ReLU
cosine vs linear warmup
Cosine vs Inverse Sqrt Schedule
rms vs layer
SwiGLU vs GeLU
Rotary vs Learned
rotary vs sinusoidal
Pre vs Post Norm
Adam vs SGD (bs=64)
Modern Transformer vs LSTM
Improvement Type