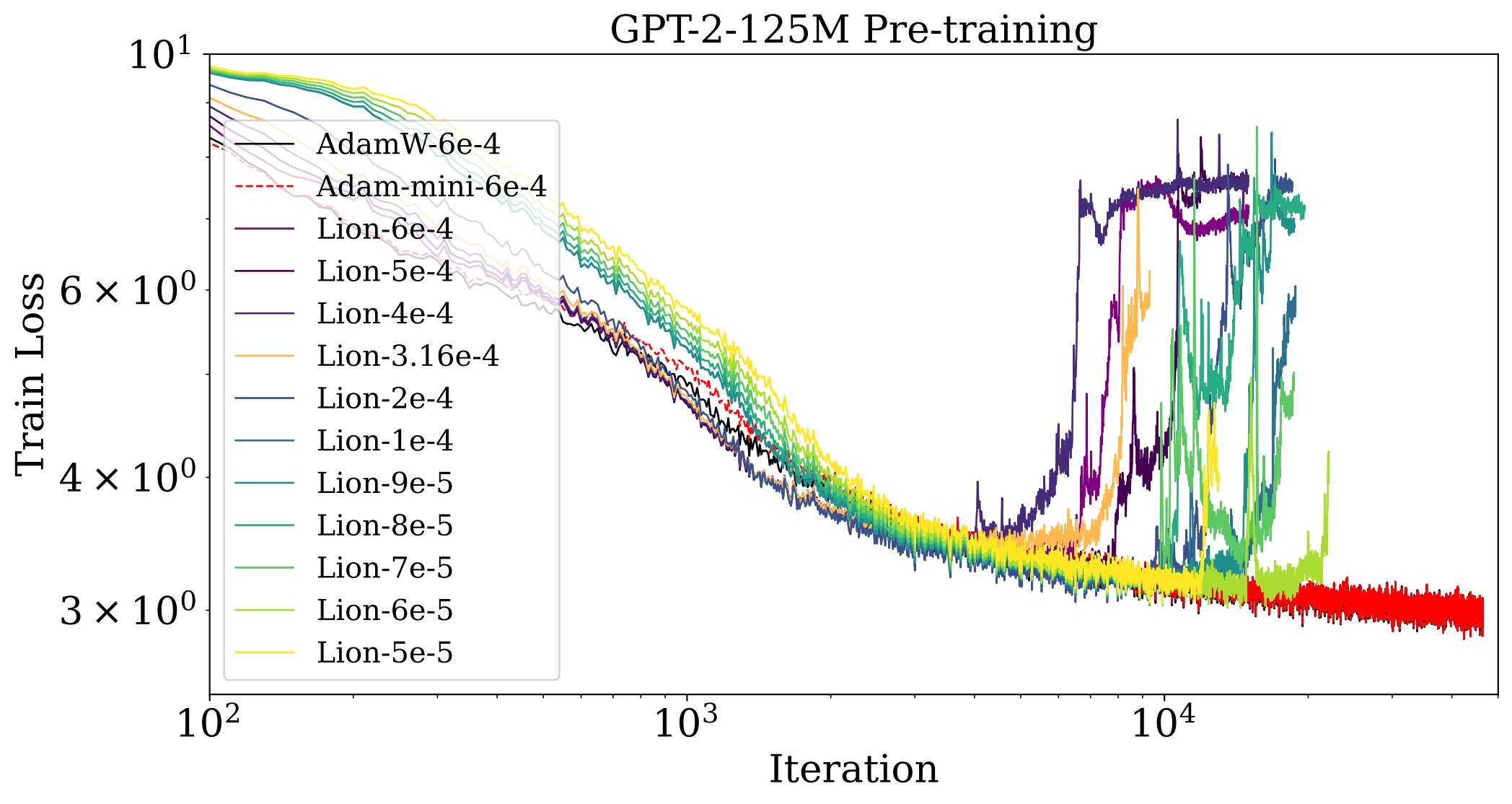
GPT-2-125M Pre-training
Train Loss
Iteration
10^1
6×10^0
4×10^0
3×10^0
10^2
10^3
10^4
AdamW-6e-4
Adam-mini-6e-4
Lion-6e-4
Lion-5e-4
Lion-4e-4
Lion-3.16e-4
Lion-2e-4
Lion-1e-4
Lion-9e-5
Lion-8e-5
Lion-7e-5
Lion-6e-5
Lion-5e-5