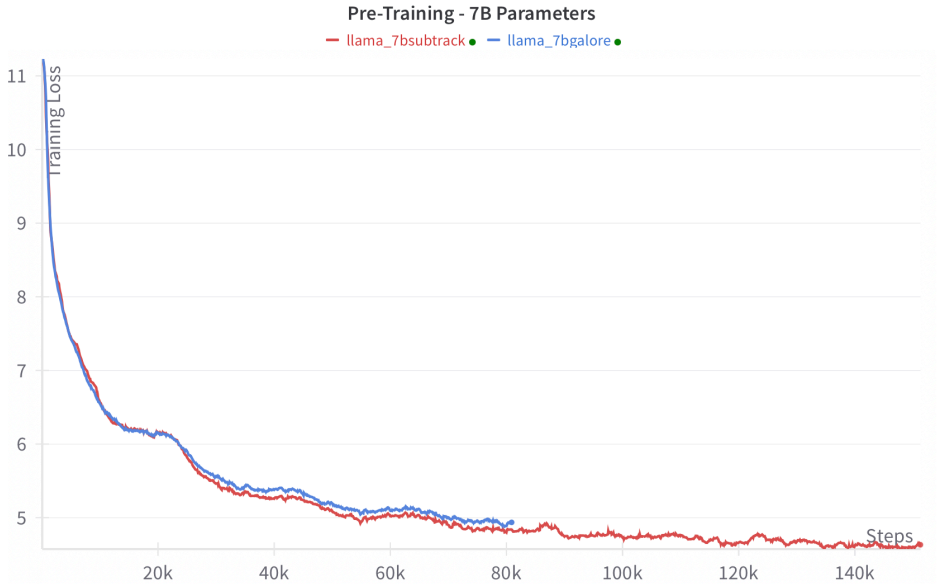
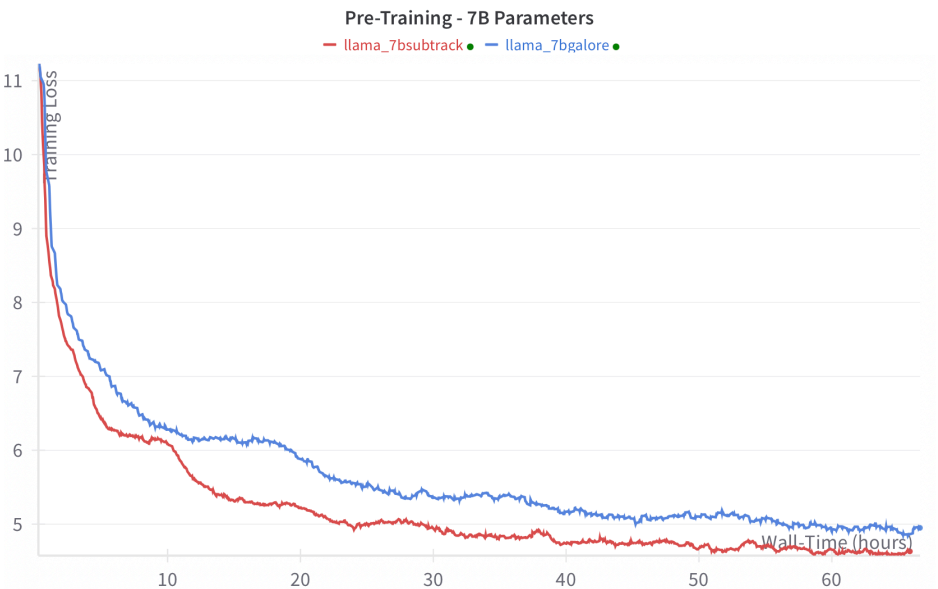


(a) Training loss of SubTrack-Grad and GaLore during pre-training of a LLaMA-based architecture with 7B parameters per iteration. Both methods were started simultaneously. Within the same time frame, GaLore completed roughly half as many iterations as SubTrack-Grad and exhibited a higher training loss.



(b) Training loss of SubTrack-Grad and GaLore during pre-training of a LLaMA-based architecture with 7B parameters, measured with respect to wall-clock time. The figure demonstrates that, under a limited time budget, SubTrack-Grad converges to a lower loss more quickly than GaLore.