Figure 1: Pre-training dynamics for Llama 350M (left) and Llama 1.3B (right) on the C4 dataset.

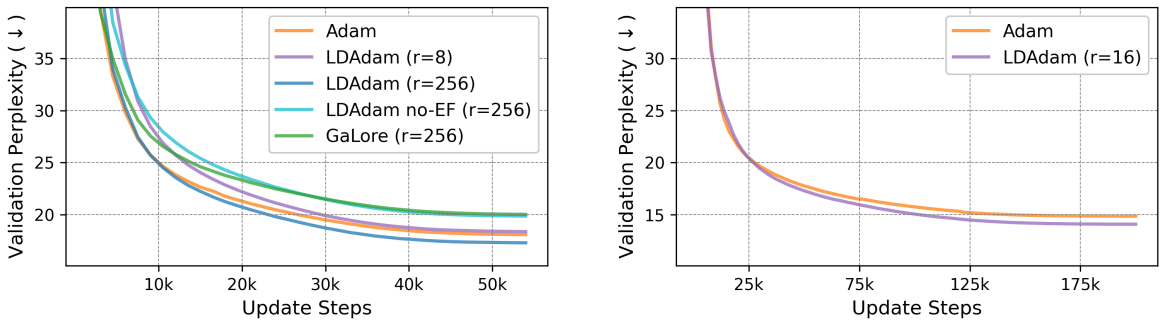

Figure 2: Pre-training dynamics over time for Llama 350M (left) and Llama 1.3B (right) on the C4 dataset.

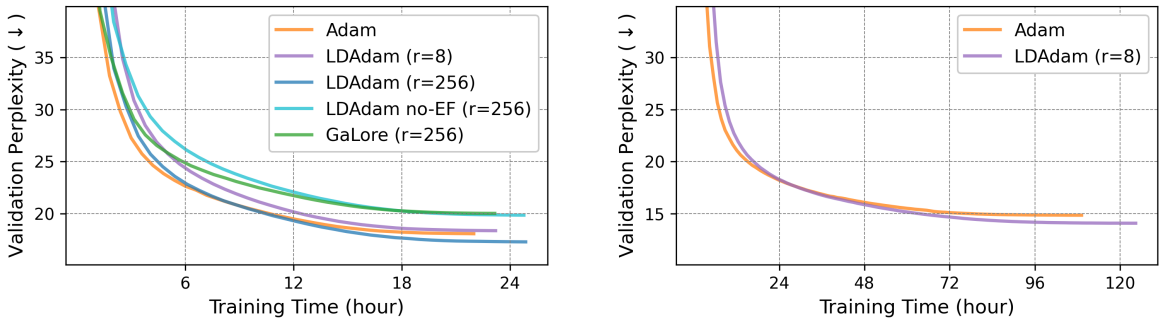

Figure 3: Throughput (token per second) and peak memory (GB) of Adam and LDAdam with respect to rank for pre-training the Llamma 350M model on the C4 dataset, on a single NVIDIA H100 80BG GPU, using micro batch size of 1.

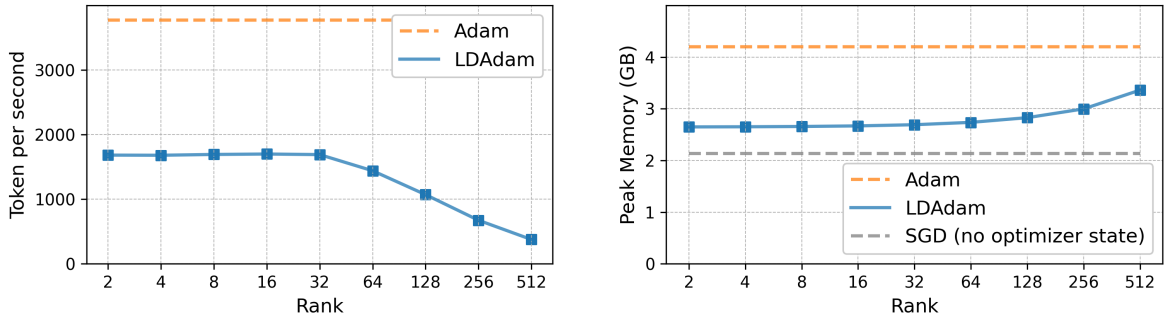

Figure 4: Training dynamics and validation perplexity for various rank when pre-training Llama 350M model. For training dynamics we used a single learning rate of $5e-4$ to allow comparison between runs and provide results for the first 10000 optimization steps. We report the best validation perplexity for learning rates tuned over the set $\{5e-4, 1e-3, 5e-3\}$.

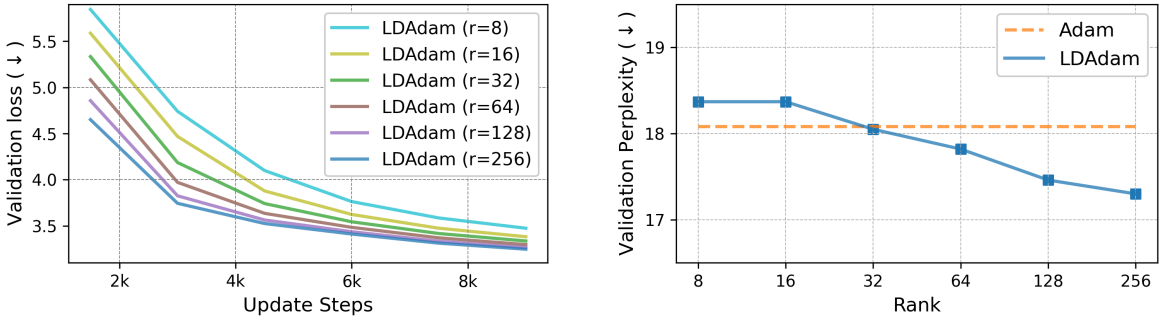

Table 1: Optimizer comparison: parameter count during training for a weight layer of shape $n \times m$ with $n \leq m$ (i.e., left projection), training capabilities, and estimates of optimizer states memory footprint in half precision.

	Adam	LDAdam	GaLore (retaining grad)	GaLore
Token count				
Weights	nm	nm	nm	nm
Gradients	nm	nm	nm	
Optimizer States	$2nm$	$nr + 2rm$	$nr + 2rm$	$nr + 2rm$
Gradient Clipping	✓	✗	✓	✗
Gradient Accumulation	✓	✓	✓	✗
Memory estimates				
RoBERTa-base (`r=8`)	0.46 GB	0.15 GB	0.15 GB	0.15 GB
Llama 350M (`r=256`)	1.37 GB	0.95 GB	0.95 GB	0.95 GB
Llama-2 7B (`r=32`)	25.1 GB	1.22 GB	1.22 GB	1.22 GB
Llama-2 7B (`r=512`)	25.1 GB	4.87 GB	4.87 GB	4.87 GB