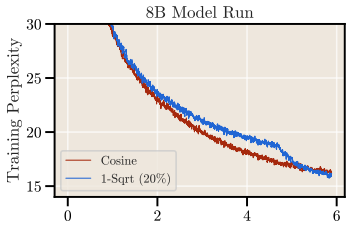

Figure 1: **Results at 8B model scale.** We validate the behavior with a 8B model (architecture of Llama 3.1) for a single short run (6B tokens of FineWeb-Edu), where the cooldown matches the cosine schedule again.

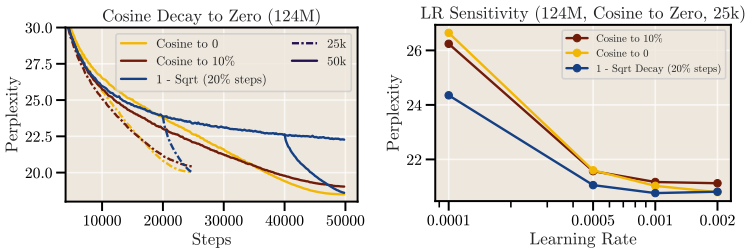

Figure 2: **Comparison of cosine to zero with a 124M model and a LR sweep.** We find that annealing cosine to zero improves the loss, but is still matched by the 1-sqrt cooldown (left, best LR after sweep). It also shows similar LR sensitivity as the standard cosine to 10% (right). Larger LR values led to divergence in training.

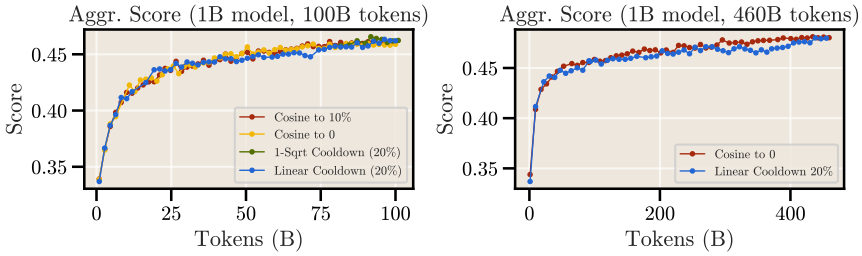

Figure 3: **Aggregate metrics throughout training of a 1B model on 100B and 460B tokens.** We train a 1B model on 100B (left) and 460B tokens (right) of FineWeb, and find that the performance of cosine and cooldown matches. Though cosine to zero improves the loss (Figure 2), it leads to a saturation before the end of training, hurting overall performance.

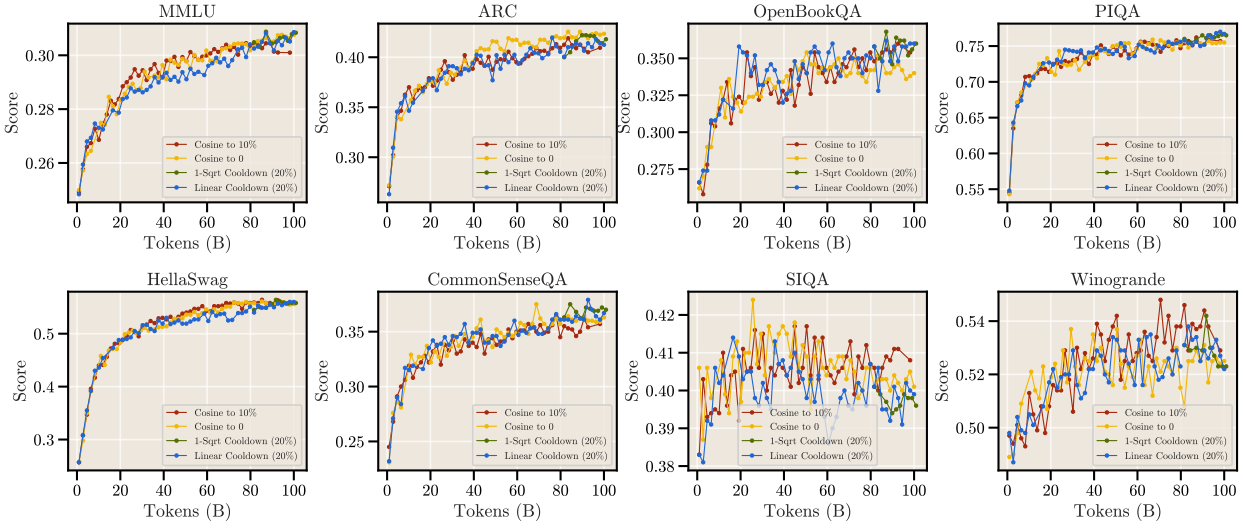

Figure 4: **Detailed Benchmarks throughout training of the 1B model on 100B tokens.** For the cooldown (starting at 80B tokens), we observe a similar uptick in performance for some metrics (e.g., MMLU, HellaSwag) akin to the observed drop in loss. Other metrics do not benefit as clearly from the cooldown (e.g., OpenBookQA).