Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations

Alexander Hägele[▲], *Elie Bakouch*[◆], *Atli Kosson*[▲], *Loubna Ben Allal*[◆], *Leandro Von Werra*[◆], *Martin Jaggi*[▲]






EPFL MLO | 🤗

Alex Hägele | *@haeggee* | alexander.hagele@epfl.ch

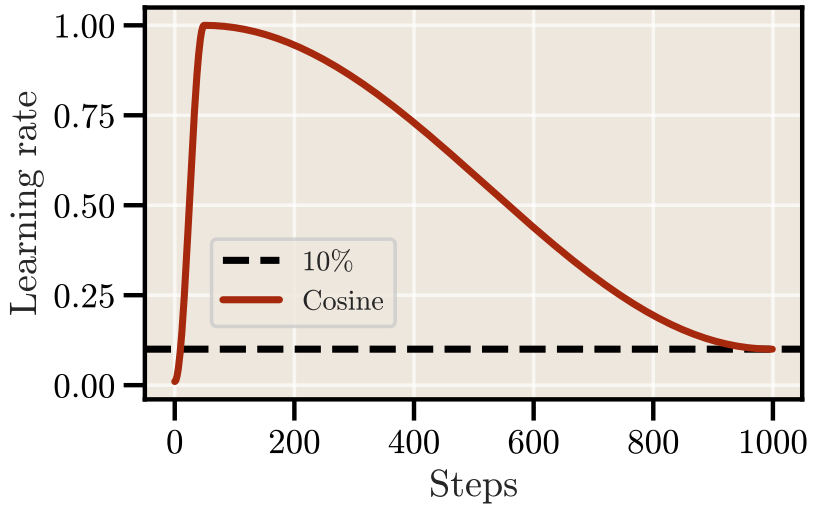
Cosine Schedule: The De-Facto Standard of LLM Training

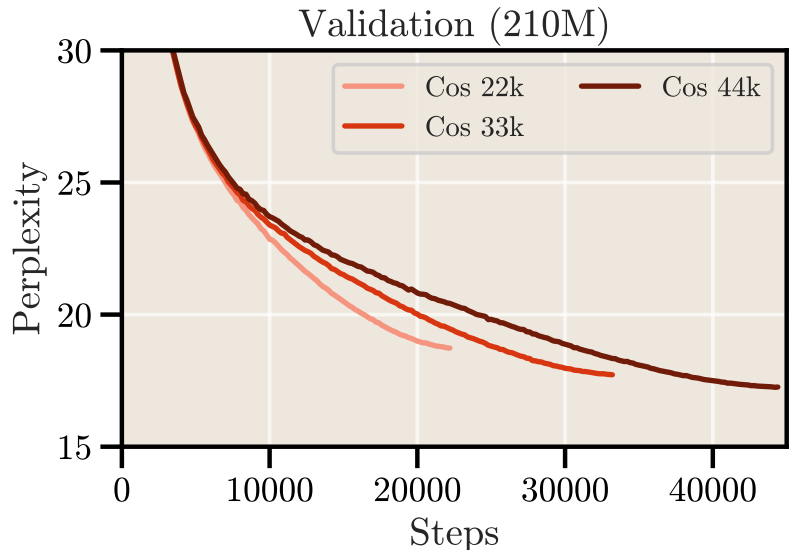
- Tradeoff via slow annealing of learning rate stretched out over training length

- Practitioners know:

Problems

- Needs predefined number of steps
- Optimality only at end; suboptimal before
- Cannot simply continue — learning rate too low

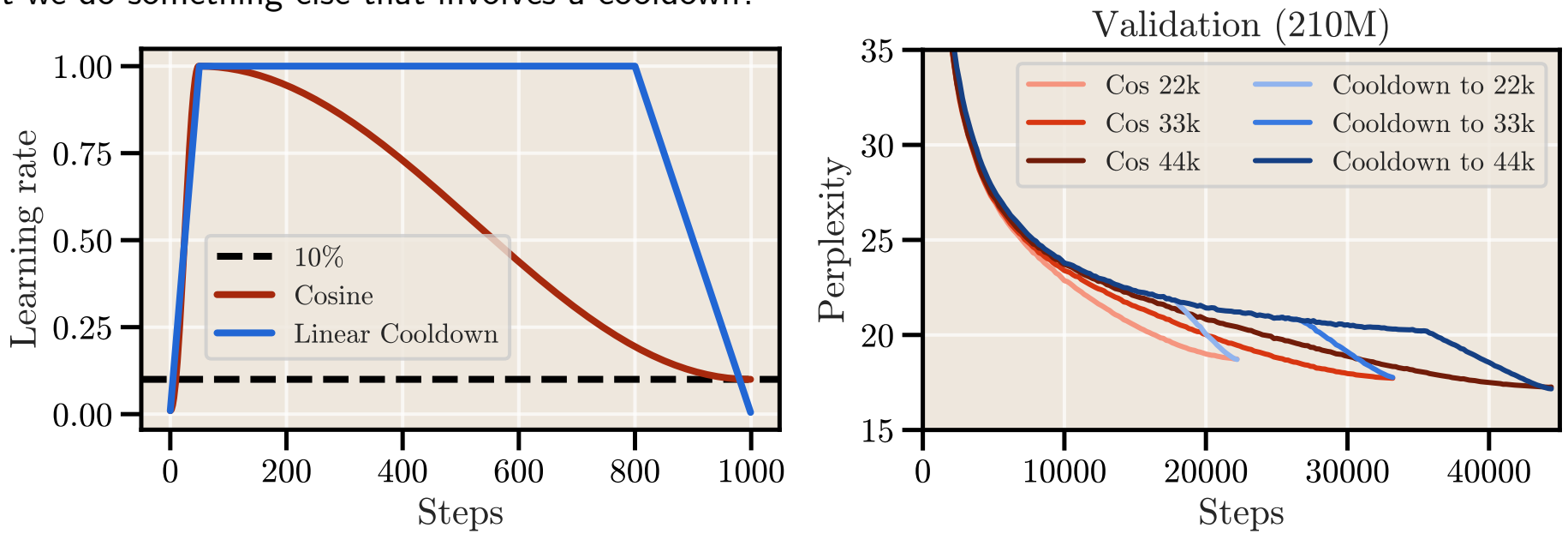




Pretraining on SlimPajama

Do We Really Need Cosine?

- Can't we do something else that involves a cooldown?

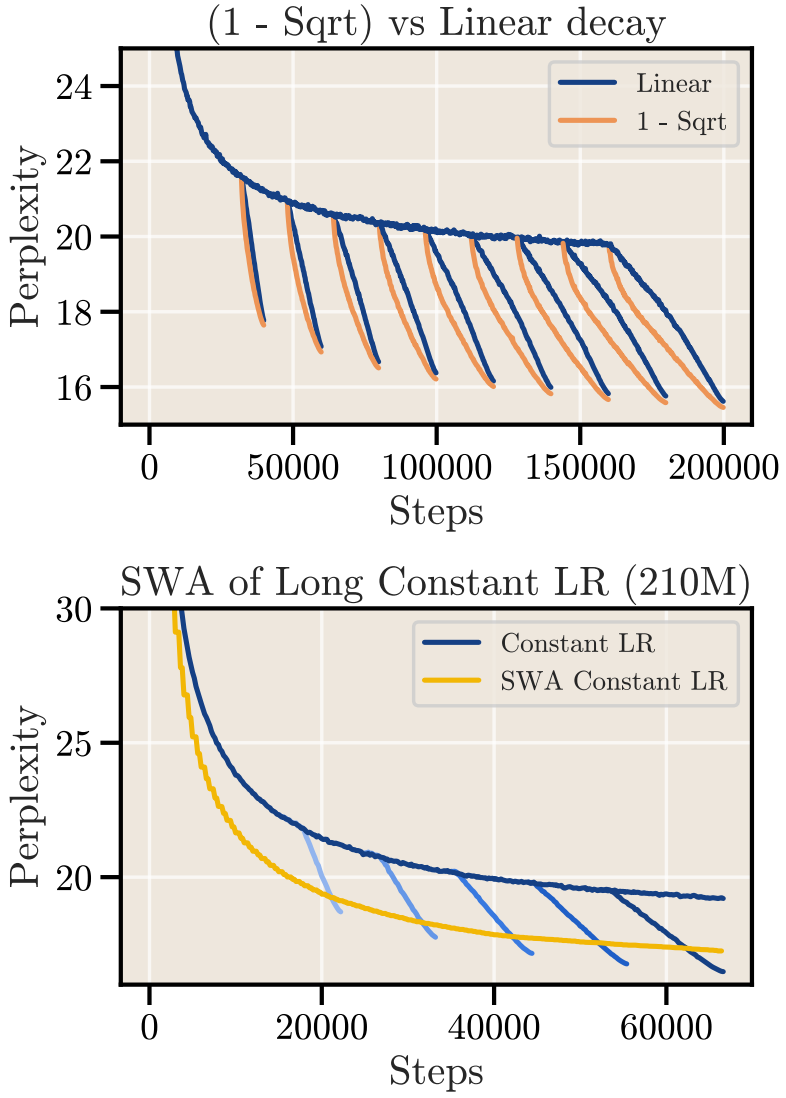


"Trapezoidal Schedule" (Zhai et al., 2021), "WSD" (Hu et al., 2024)

Advantages

- No predefined number of steps: cooldown at any point to see model behavior
- Continual learning via continuation from checkpoint
- Separate pretraining and "fine-tuning" phase: mix in high quality data at the end

Ablations: Cooldown Form, Length, Landscape

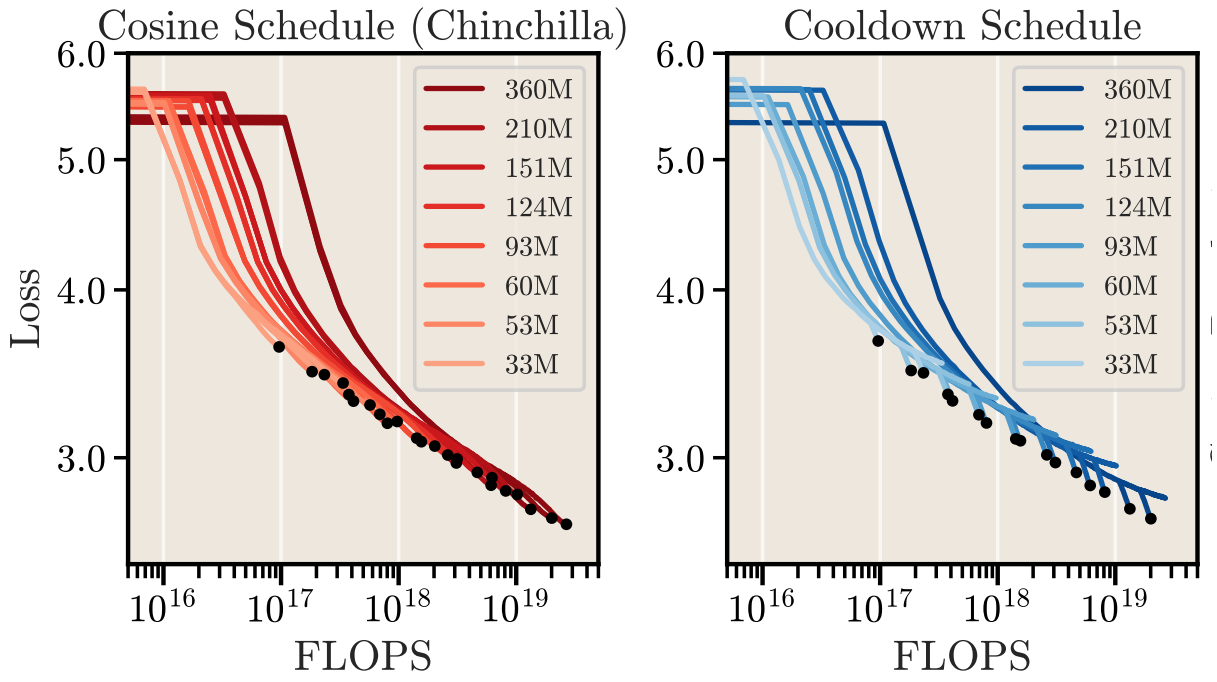


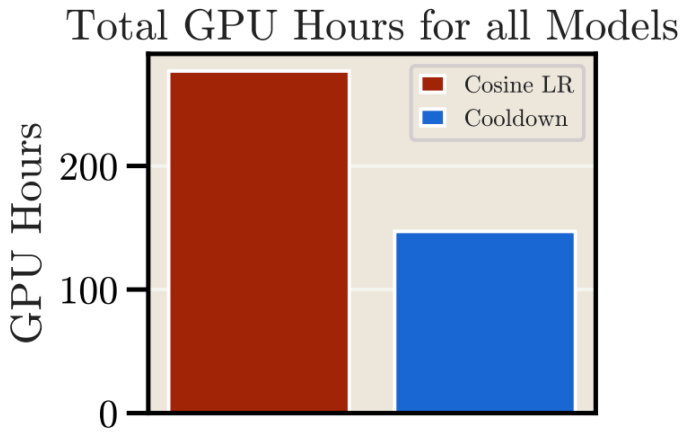
Takeaways

- Longer cooldown helps, sufficient for 10-20% of steps
- A negative square root form outperforms standard linear cooldown!
- Weight averaging serves as a form of simulated LR decay (Sandler et al., 2023) and gives performance boost — at no additional cost

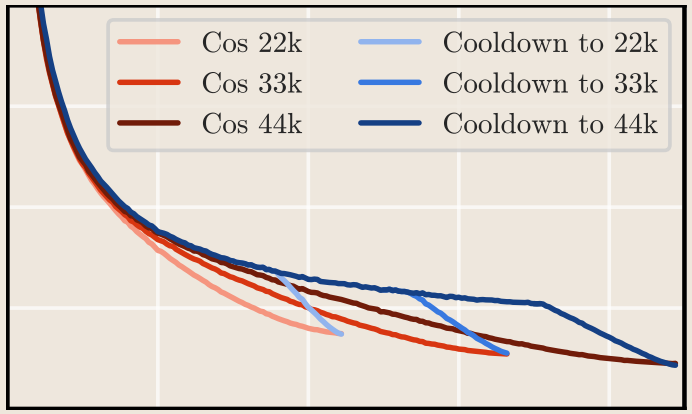
The Implications for Scaling Laws Research

- Instead of training from scratch, reuse checkpoints for ad-hoc cooldowns





Discussion & Messages



- Try it on your model + data and tell us how it works!
- Also: use checkpointing and weight averaging :)
- What is the optimal schedule, really?
  - In practice, let's be smart about training and scaling

Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations

Alexander Hägele[▲], *Elie Bakouch*[◆], *Atli Kosson*[▲], *Loubna Ben Allal*[◆], *Leandro Von Werra*[◆], *Martin Jaggi*[▲]






EPFL MLO | 🤗

Alex Hägele | *@haeggee* | alexander.hagele@epfl.ch

References

- Defazio, A., Yang, X., Mehta, H., Mishchenko, K., Khaled, A., and Cutkosky, A. The Road Less Scheduled. May 2024. URL http://arxiv.org/abs/2405.15682v1
- Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., Zhang, X., Thai, Z. L., Zhang, K., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., Ding, N., Jia, C., Zeng, G., Li, D., Liu, Z., and Sun, M. Minicpm: Unveiling the potential of small language models with scalable training strategies. Apr 2024. URL https://arxiv.org/abs/2404.06395v2
- Sandler, M., Zhmoginov, A., Vladymyrov, M., and Miller, N. Training trajectories, mini-batch losses and the curious role of the learning rate. Jan 2023. URL http://arxiv.org/abs/2301.02312v2
- Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. pp. 12104–12113, 2022. URL http://arxiv.org/abs/2106.04560v2