ICML
International Conference
On Machine Learning

Carnegie Mellon University

Sharpness-Aware Pretraining Mitigates Catastrophic Forgetting

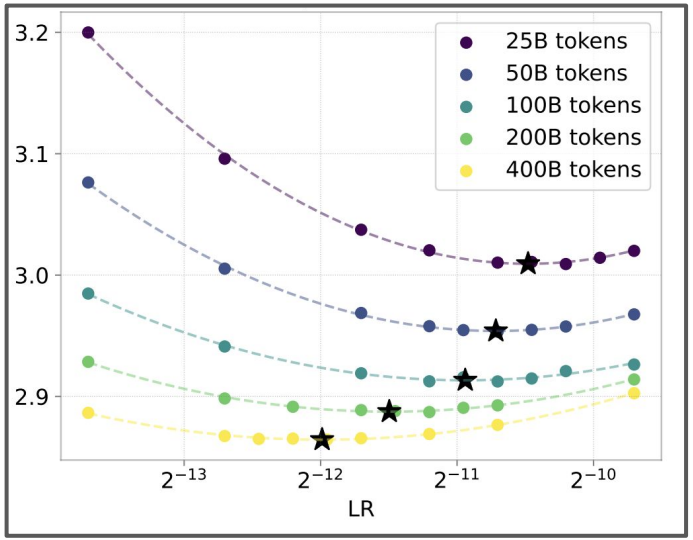
ICML 2026

Ishaan Watts*, Catherine Li*, Sachin Goyal,
Jacob Mitchell Springer†, Aditi Raghunathan†

*Equal contribution †Equal advising

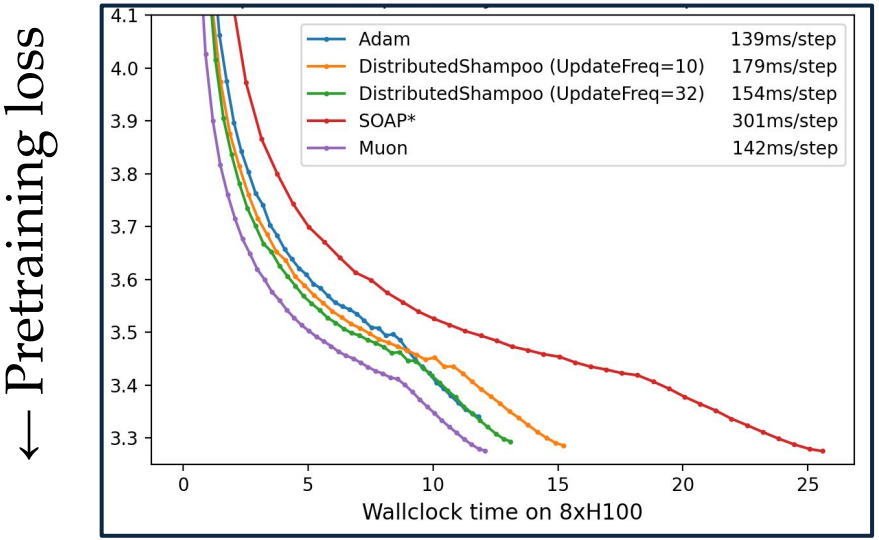
Pretraining choices are tuned to optimize the base model.

Learning Rate



Bjorck & Benhaim. ***Scaling Optimal LR across Token Horizons****, 2025.*

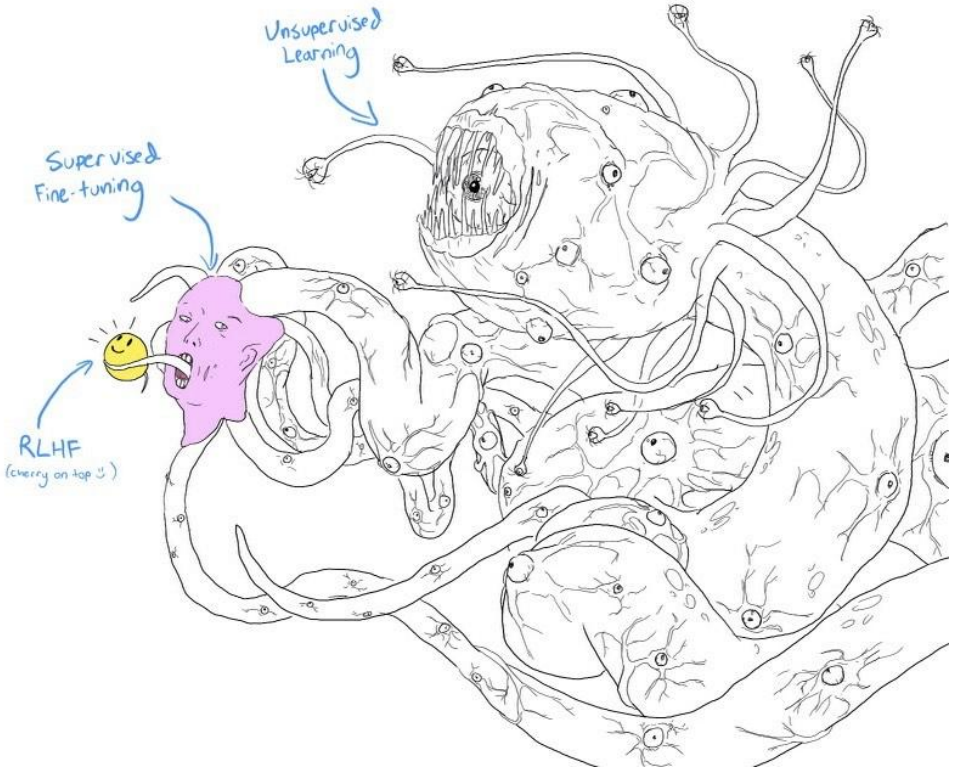
Optimizer



Jordan et al. ***Muon: An optimizer for hidden layers in neural networks****, 2024.*

Multi-stage LLM-training pipeline.

We want the base model to be good after *further modification*!

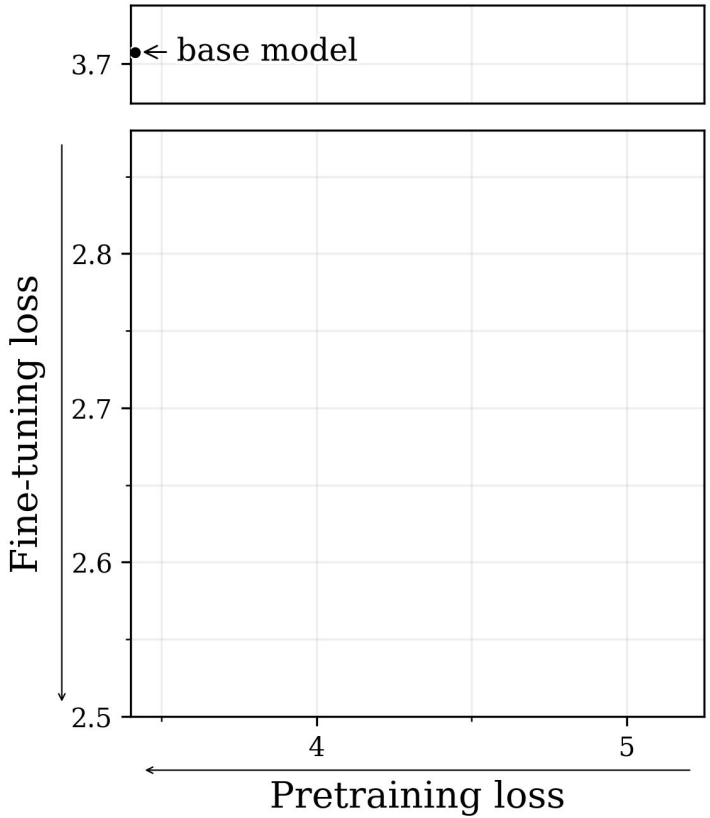


Shoggoth with Smiley Face. Courtesy of twitter.com/anthrupad

Inherent tradeoff in learning-forgetting.

Pretraining loss alone doesn't capture base model *adaptivity*.

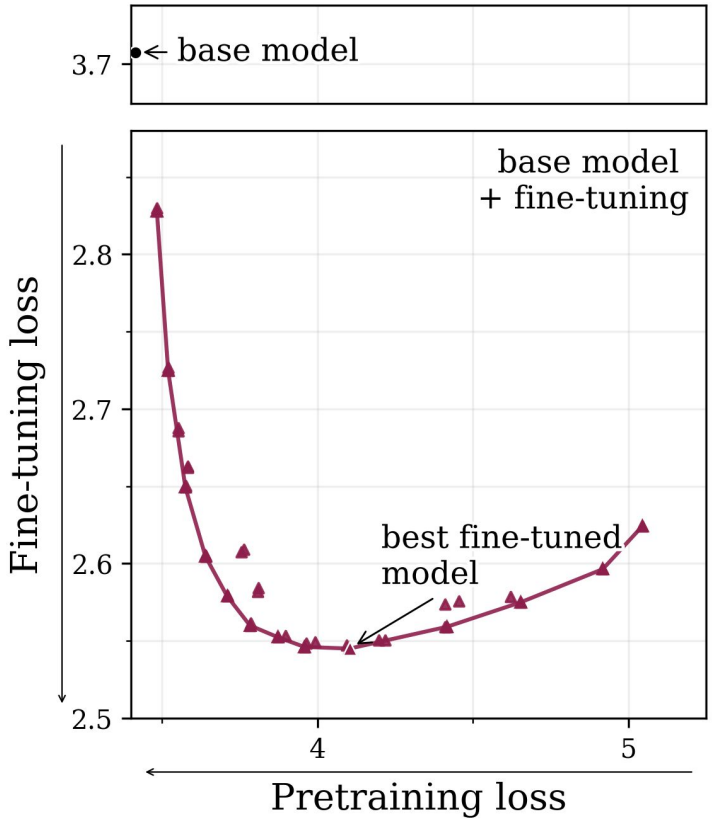
Instead, we study the *learning-forgetting tradeoff*.



Inherent tradeoff in learning-forgetting.

Pretraining loss alone doesn't capture base model *adaptivity*.

Instead, we study the *learning-forgetting tradeoff*.



Reducing the *sharpness* of pretraining loss *improves* the learning-forgetting tradeoff.

We evaluate three changes to the optimizer to *minimize sharpness*.

We evaluate three changes to the optimizer to *minimize sharpness*.

Explicit → *Sharpness-Aware Minimization (SAM)* optimizer.

Sharpness-Aware Minimization:

$$L^{\text{SAM}}(w) = \min_{w} \max_{||\varepsilon|| \leq \rho} L(w + \varepsilon)$$

Foret et. al. Sharpness-aware minimization for efficiently improving generalization 2020

We evaluate three changes to the optimizer to *minimize sharpness*.

Explicit → *Sharpness-Aware Minimization (SAM)* optimizer.

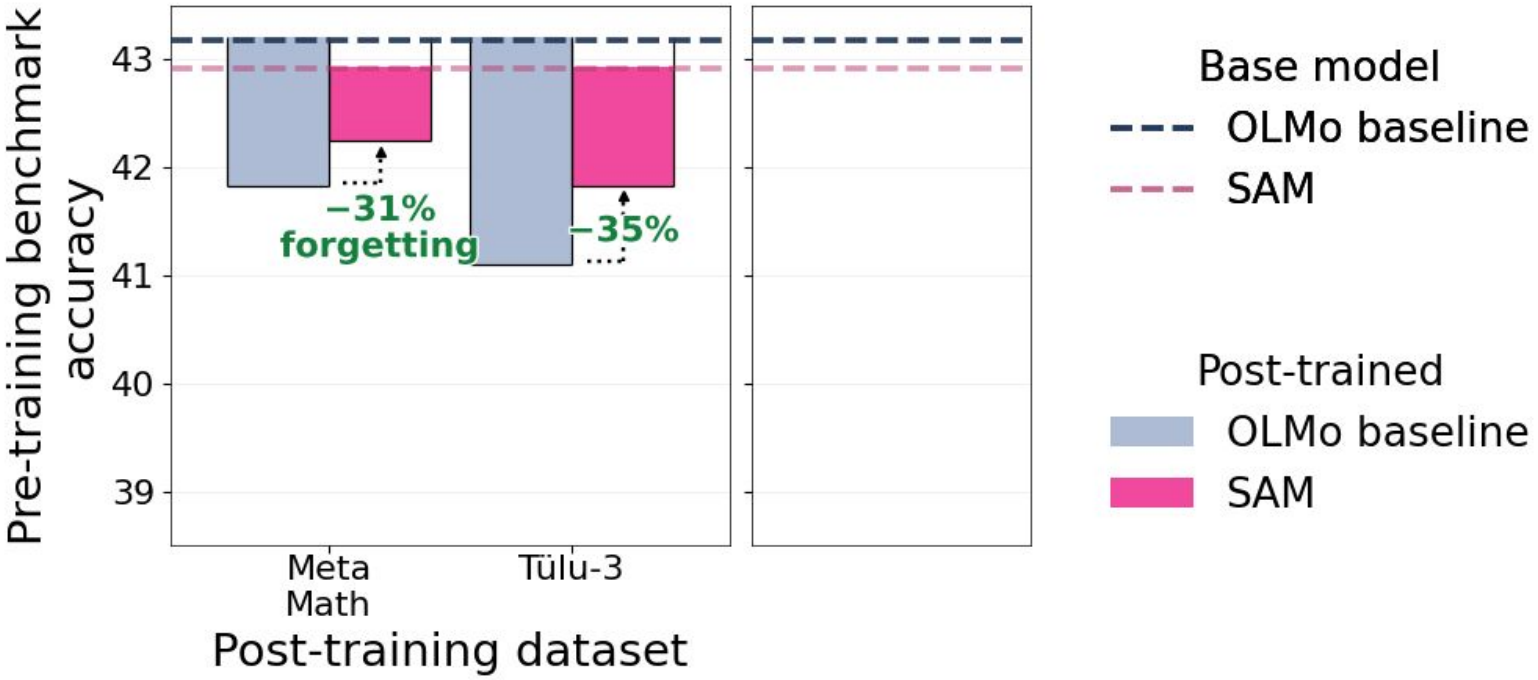
Implicit → *Peak learning rates* and *learning rate decay time.*

Sharpness-Aware Minimization:

$$L^{\text{SAM}}(w) = \min_{w} \max_{||\varepsilon|| \leq \rho} L(w + \varepsilon)$$

Foret et. al. Sharpness-aware minimization for efficiently improving generalization 2020

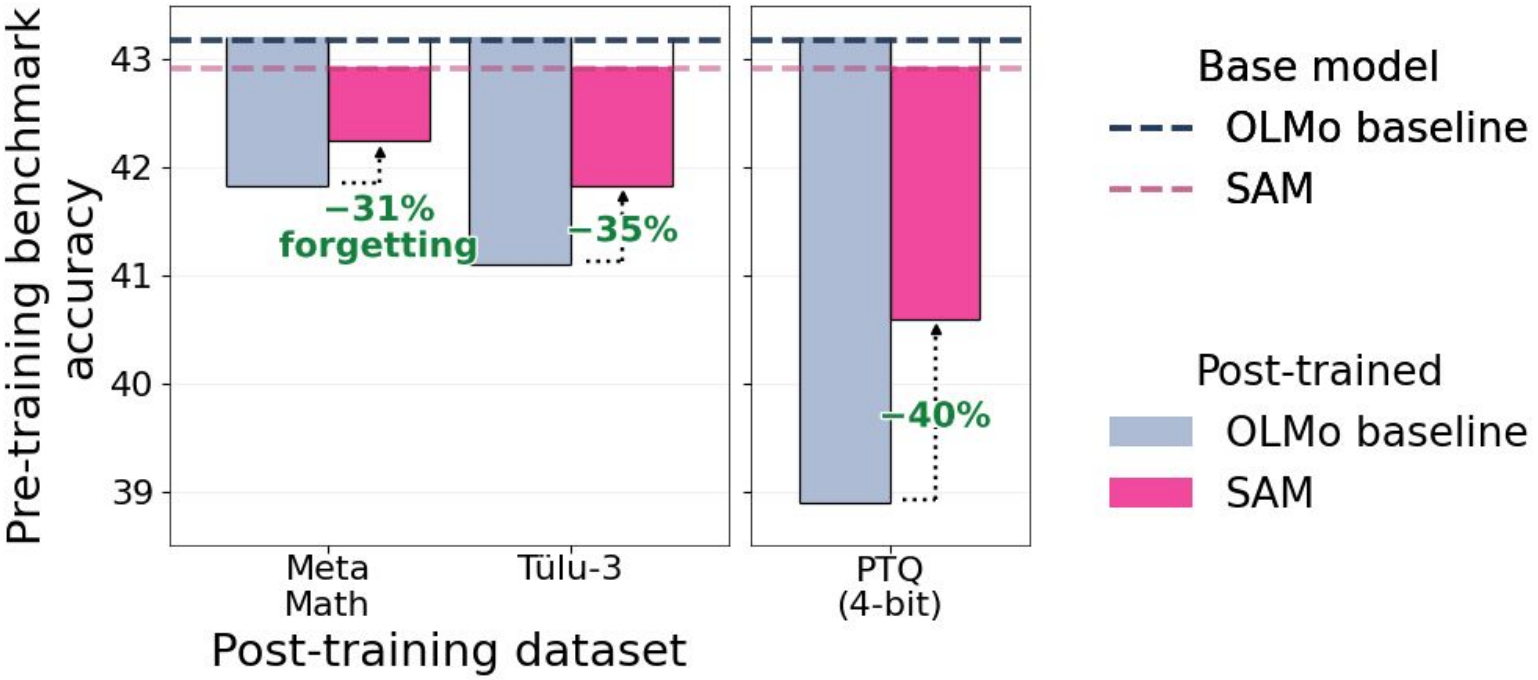
Main results from OLMo-2-1B.



SAM yields **35% less forgetting after post-training on Tülu-3** and **Meta-Math[1].**

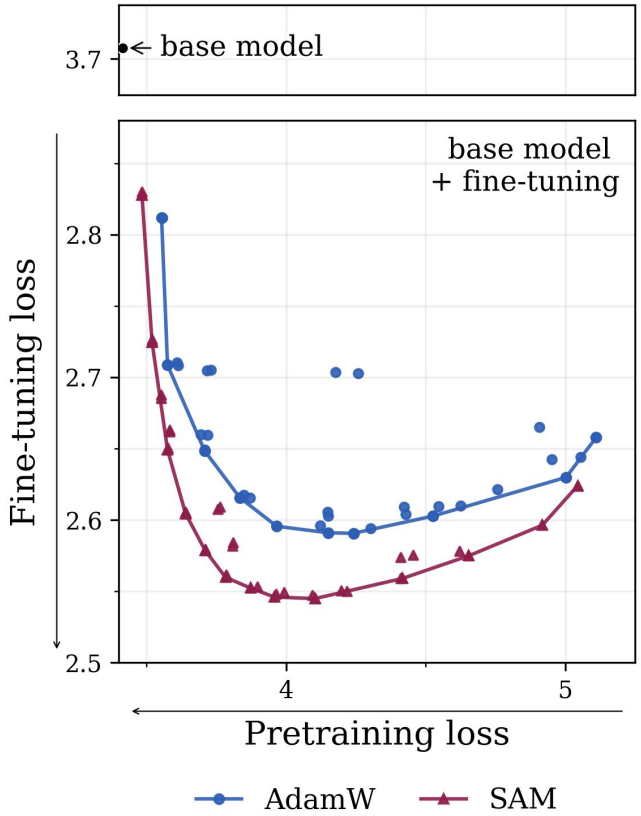
[1] More datasets in poster!

Main results from OLMo-2-1B.



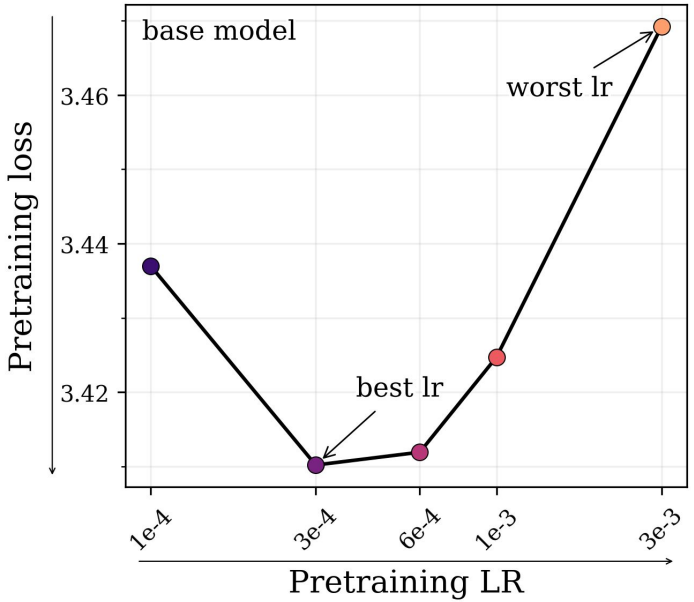
SAM yields **40% less forgetting after 4-bit quantization.**

Main LF-tradeoff curves from SAM.

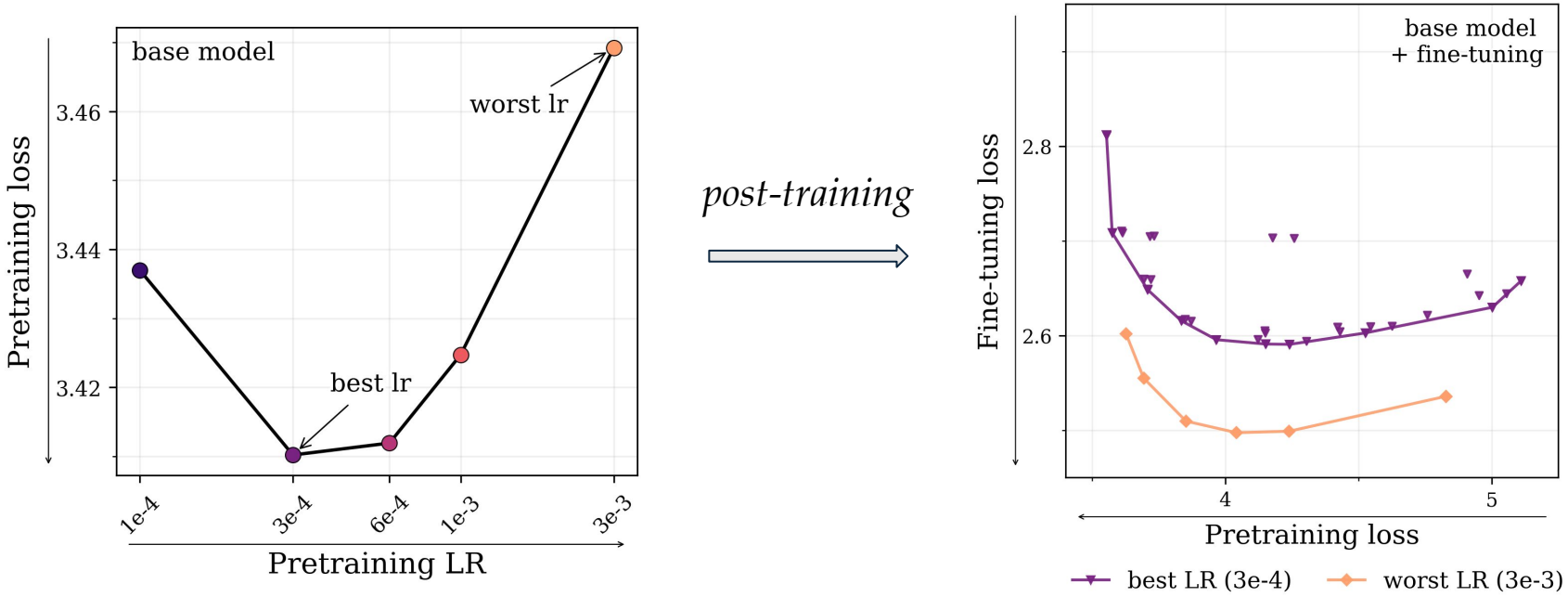


SAM achieves a *superior* learning-forgetting tradeoff.

Main LF-tradeoff curves from peak learning rate.



Main LF-tradeoff curves from peak learning rate.



Worst pretrain LR: *Worse pretrain loss* but *better learning-forgetting tradeoff*!

Come visit our poster!
[7th July 2:00-3:45pm Hall A]

- *Intuition* behind minimizing sharpness to mitigate forgetting.
- *Empirical analysis* of sharpness-minimization.
- *Scaling behaviours* of SAM and more…

Thank you





Coauthors (left to right): Ishaan Watts, Catherine Li, Sachin Goyal, Jacob Mitchell Springer, Aditi Raghunathan