More concentrated fine-tuning data
⇒ a sharper loss landscape
⇒ more curvature-driven forgetting

Curvature adds to forgetting even without gradient conflict, and its share grows with concentration.

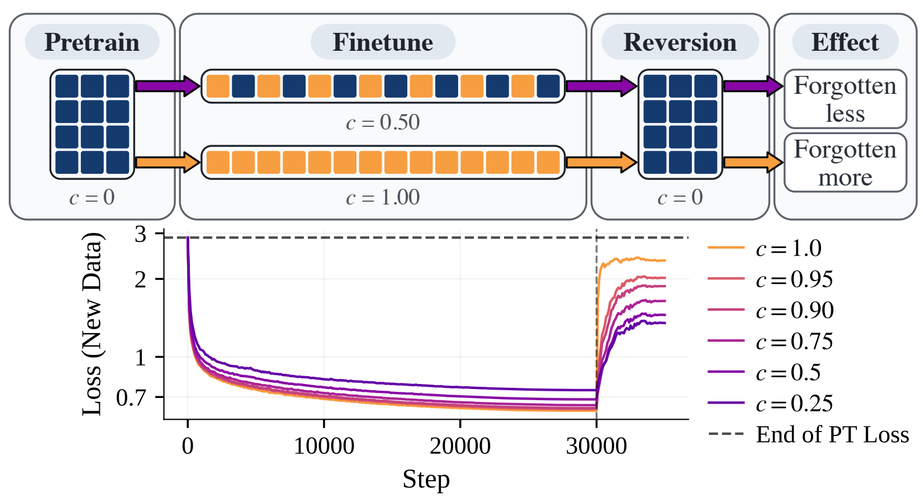
# A loss curvature account of fine-tuning fragility

## The puzzle

Fine-tuning on a narrow distribution is **fragile**: a little further training easily undoes it, which matters for the **durability of safety fine-tuning**. Mixing pre-training data back in is a known fix, but *why* the proportion **c** (**concentration**) of new-task data changes how fast a model forgets was not understood.

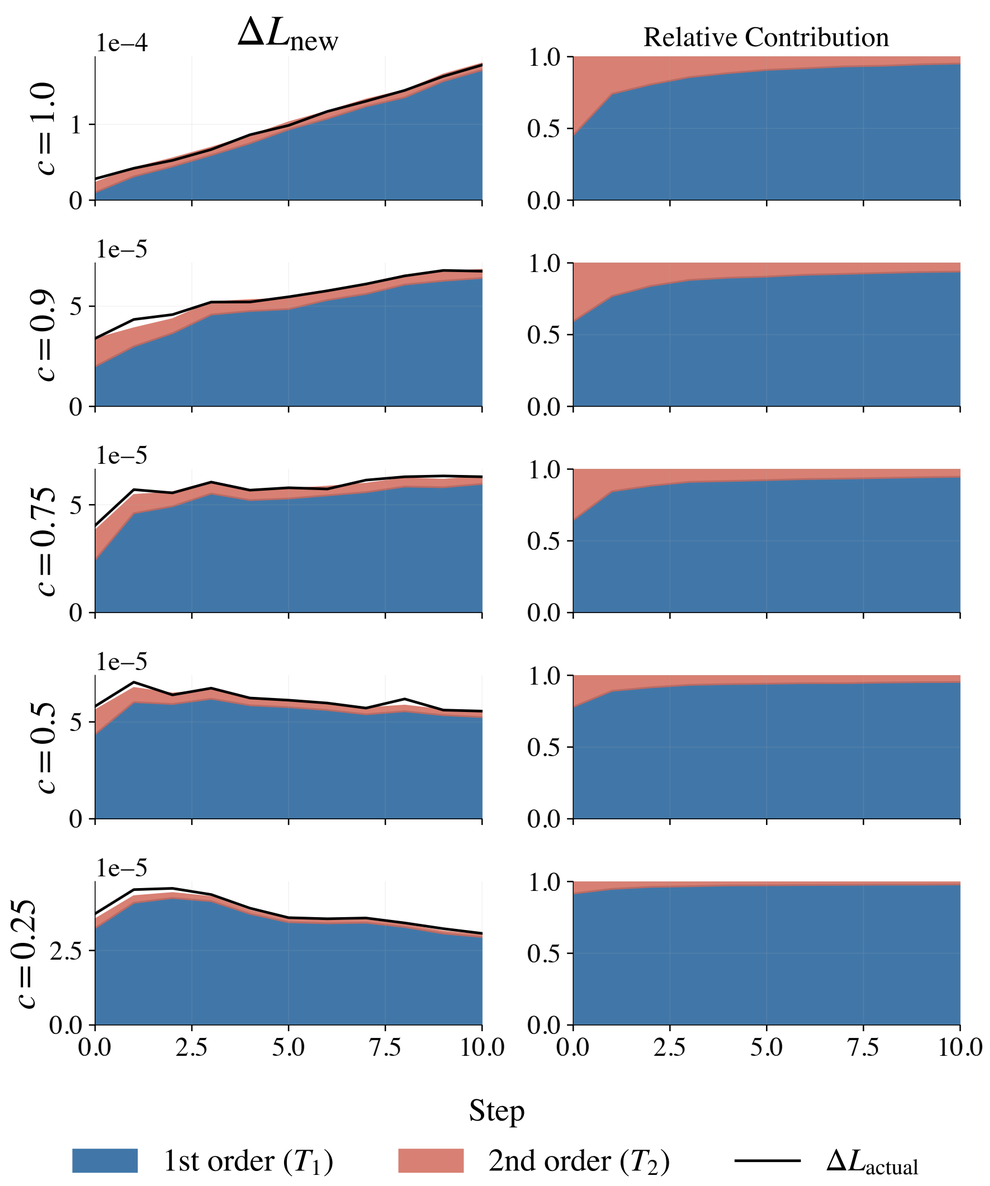
During **reversion** we decompose the per-step fine-tune loss change into first-order ($T_1$, gradient) and second-order ($T_2$, curvature) Taylor terms. Pythia-70M/1B; chemistry, biomedical (+ music in paper).

## The protocol




## Curvature's share grows with c

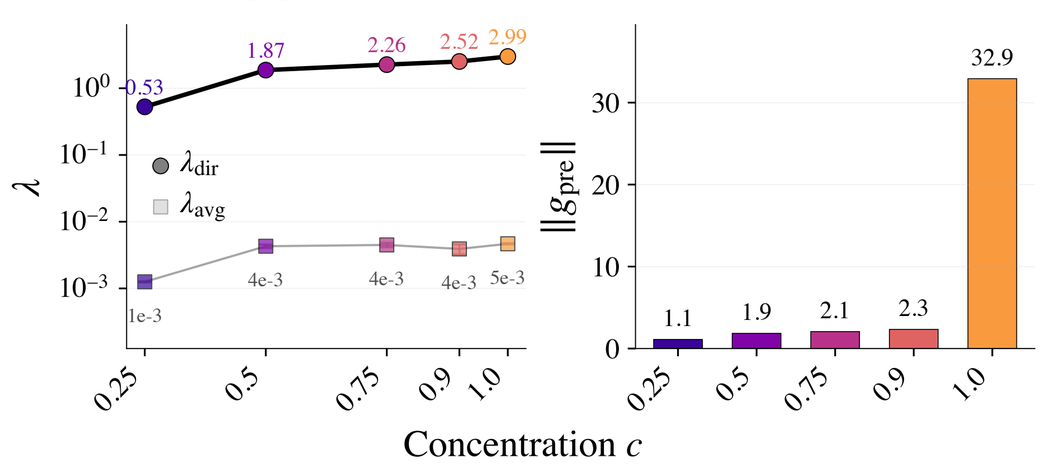
Decompose each reversion step's loss change as $\Delta L_{new} \approx T_1 + T_2$, where $T_1 = g_{new}\Delta\theta$ is the **gradient** term, measuring whether the reversion step $\Delta\theta$ opposes $g_{new}$, and $T_2 = \frac{1}{2}\Delta\theta H \Delta\theta$ is the **curvature** term.




- Left: both $T_1$ and $T_2$ grow as **c** rises, but $T_2$ grows faster.
- Right: $T_2$'s **share** of forgetting is small at **c**=0.25 and grows substantially by **c**=1, though $T_1$ stays larger in absolute terms.

## Sharp along the reversion direction

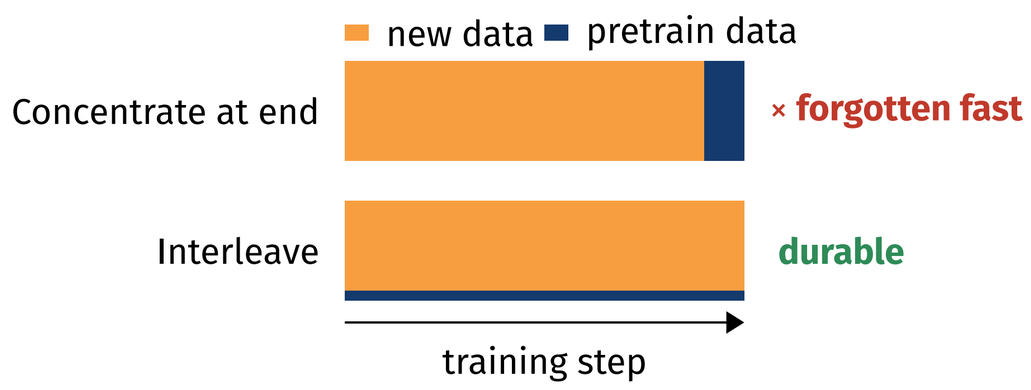
$\lambda_{dir}$ = Hessian sharpness *along* the reversion direction $\Delta\theta$. $\lambda_{avg}$ = average sharpness over all directions. $\|g_{pre}\|$ = pre-training gradient norm at the fine-tuned point.




- $\lambda_{dir} \gg \lambda_{avg}$: the loss is **sharper** along $\Delta\theta$ than along a typical direction, even when $g_{new}$ and $\Delta\theta$ are near-orthogonal.
- $\|g_{pre}\|$ also spikes at **c**=1: high-**c** fine-tunes land in a sharper region, with ~4× less degradation at **c**=0.25 than at **c**=1.

**Takeaway.** The *temporal structure* of training shapes durability: for durable alignment, **interleave safety data throughout training** rather than concentrating it at the end.

### What to do instead





Ivaylo Dimitrov*, Leo Karoubi*, Sunny Howard*, Dmitrii Krasheninnikov
LASR Labs · SUMMIT, Sofia University
Munich Center for Machine Learning


Presented at 3 ICML 2026 Workshops:
- Continual Adaptation at Scale
- High-dimensional Learning Dynamics
- Mechanistic Interpretability

**Read the paper**
openreview.net/forum?id=LhT1YHmyCN