Anneal confirmation

> The pool over slices 1 to 3, with its slice temperature annealing slowly from 0.1 to 0.025, keeps the anchor on every run at eight fresh seeds. An early call° at epoch 64 finds the anchor kept on every run of a forty-seed count too. Spill° and the task stay within the criteria of the D2.2 design (one run ends at the spill line), and removal° is about that of `falloff`.

Call. The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64....

Spill. The largest drop on any other op, at the dose where it is largest....

Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....

The *soft pool scout*° found a candidate for the recipe of record: the pull pooled over slices 1 to 3, with the slice temperature τ_ℓ held at 0.1.[1] It keeps the anchor and stays within the spill criterion at seeds 704 to 711. But on one or two seeds, spill rises late and ends near the criterion. The *late spill scout*° then found that sharpening τ_ℓ further late in training, to 0.025, lowered that late rise at no measured cost.

soft pool scout. The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.

1 Slice temperature...

late spill scout. The late spill follows the anti anneal, and sharpening τ_ℓ to 0.025 over epochs 64 to 360 lowers it with no measured cost to removal or the task.

Here the two moves become one. τ_ℓ falls from 0.1 at the start to 0.025 at epoch 400, smoothly on a log scale. So it is still close to 0.1 while the anchor settles, and reaches the halfway point (0.05) at epoch 200. We call this recipe `τ_ℓ anneal`.

This is the confirmation run for adopting `τ_ℓ anneal` as the recipe of record. It trains the recipe to the end at eight fresh seeds (712 to 719), with a control at each seed. It also trains it to epoch 64 at forty seeds (704 to 743), to count lost anchors.

Observations

- The anchor at fresh seeds (E1): every run keeps the anchor and passes the task gate. Removal° varies more across seeds than under `τ_ℓ fixed`, and about as much as under `falloff`. One seed is a little below the lowest `falloff` run.

Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....

- Spill at the end and along training (E2)°: no run ends over the criterion. One seed rises late to the line, and the same seed goes over under the edit at block 1 alone. Two seeds have brief bumps over the line in the first hundred epochs, which clear within 64 epochs. The *soft pool scout*° saw none under `τ_ℓ fixed` at the same checkpoints.

Spill. The largest drop on any other op, at the dose where it is largest....

***soft pool scout*.** The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.

- Lost anchors at forty seeds (E3): the call° finds all forty runs kept, each well clear of the band where the call is unsure.

Call. The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64....

Scope and method

This is an exploratory study, with no preregistration and no gate. We set four checks before the runs. They guide the choice of recipe without deciding it, so a recipe that misses one narrowly can still be the better choice:

1. Every run at seeds 712 to 719 keeps the anchor (removal° at least 0.5).
2. No run at those seeds ends over the spill° criterion of 0.02.
3. Removal is no lower than that of `falloff` at seeds 704 to 711. The task passes the gate of the D2.2 design against the control at the same seed: a mean change over ops of at least −0.02, and at least −0.05 on any one op.
4. At most one run of the forty loses the anchor. That is about as few as `step, soft τ` (none of forty) and `falloff` (one) lost in the *loss-rate screen*°, and forty seeds can't tell those two counts apart.

Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....

Spill. The largest drop on any other op, at the dose where it is largest....

***loss-rate screen*.** A call from α at the example answers over epochs 44 to 64 sorts lost anchors from kept ones; at forty seeds `falloff` lost one and `step, soft τ` none.

`τ_ℓ anneal` is `τ_ℓ fixed` from the *soft pool scout*° (in full, `τ_ℓ 0.1, τ_ℓ fixed`) with one change: τ_ℓ moves from 0.1 to 0.025 over epochs 0 to 400, on a minimum-jerk curve in log τ_ℓ.[2] The position temperature τ_p still sharpens from 1 to 0.1 by epoch 40, the anchor weight still warms up over the first 40 epochs, and the anti weight still anneals to the per-slice holds of `falloff` by epoch 360.

***soft pool scout*.** The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.

2 Minimum-jerk...

The forty runs of the count use the same seeds as the `falloff` and `step, soft τ` runs of the *loss-rate screen*. So each run pairs with a run of each of those recipes by initialization, batches, and label draws. Each run stops at epoch 64 for its call°, made as in that screen.

Call. The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64....

The runs at 712 to 719 continue from those stops to epoch 400, keeping a checkpoint every 32 epochs. Each control trains for the full schedule at the same seed, with no pull and no anti term. All runs trained and were measured on L4 GPUs.

The stored recipes ran at seeds 704 to 711, and `τ_ℓ anneal` at 712 to 719. So a difference between them includes the variation between two draws of eight seeds. Under `falloff` and `step, soft τ`, removal varies by about 0.15 across seeds, with one or two low seeds in eight. A difference in the seed mean smaller than a few hundredths is within that variation.

The measurements

Every edit below is the e_1 edit. Each score is the drop under that edit, less the median drop under the same edit along each of the other 63 coordinate axes of the run (the rule of the D2.2 design).

Removal

: The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null. Higher is better, and a run below 0.5 has lost the anchor.

Spill

: The largest drop on any other op, at the dose where it is largest. Lower is better, and a run above 0.02 is over the criterion.

Spill along training

: Spill at the full dose, measured on each checkpoint a run kept (epochs 16, 32, and 64, then every 32 epochs) and at the end.

Task change

: The clean score per op, less that of the control at the seed.

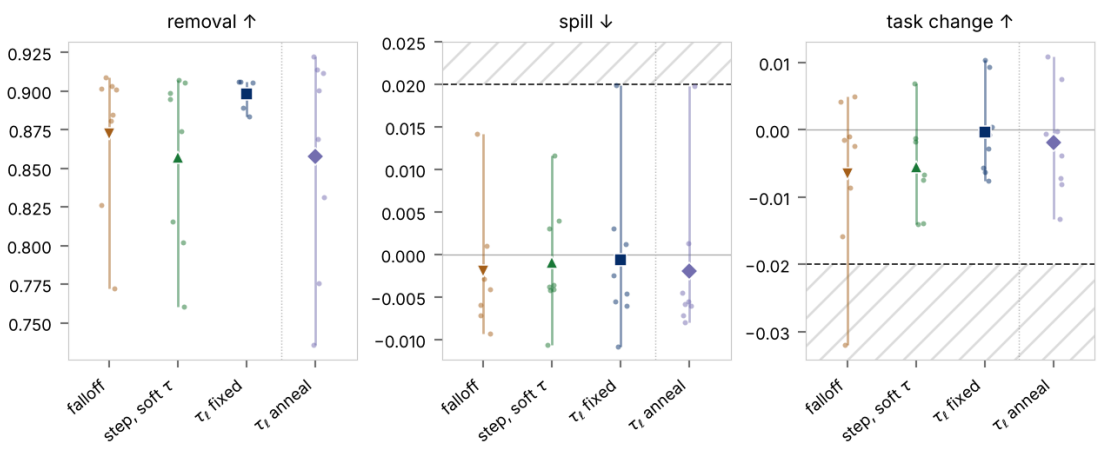
Call

: The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64. A run below 0.15 is called lost, and one from 0.28 kept. A run in between is finished and measured.

The anchor at fresh seeds (E1)

Does every run at the fresh seeds keep the anchor, and are removal° and the task as good as under `falloff`?

Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....



Removal, spill, and task at epoch 400, by recipe. *One column per recipe: the stored runs of `falloff`, `step, soft τ` and `τ_ℓ fixed` at seeds 704 to 711, and `τ_ℓ anneal` at the fresh seeds 712 to 719; the seeds as small dots over a thin bar for their range, the seed mean as the large mark.* ***Left:*** *removal at the full dose.* ***Middle:*** *spill, the largest net drop on another op at any dose; the dashed line is the criterion, with the side over it hatched. Both are net of the median over the other 63 axes of the run.* ***Right:*** *the task change, the mean clean score over ops less that of the control at the seed, with the gate of the D2.2 design dashed.*

recipe	seeds	removal ↑	spill ↓	over criterion ↓	over, block 1 ↓	task change ↑	worst op ↑
`falloff`	704–711	0.87 (0.77 to 0.91)	−0.002 (−0.009 to 0.014)	0 of 8	1 of 8	−0.007 (−0.032 to 0.005)	−0.066
`step, soft τ`	704–711	0.86 (0.76 to 0.91)	−0.001 (−0.011 to 0.012)	0 of 8	1 of 8	−0.005 (−0.014 to 0.007)	−0.036
`τ_ℓ fixed`	704–711	0.90 (0.88 to 0.91)	−0.001 (−0.011 to 0.020)	0 of 8	1 of 8	−0.000 (−0.008 to 0.010)	−0.023
`τ_ℓ anneal`	712–719	0.86 (0.74 to 0.92)	−0.002 (−0.008 to 0.020)	0 of 8	1 of 8	−0.002 (−0.013 to 0.011)	−0.036

Removal, spill and task at epoch 400*, seed mean and range over eight runs, net of the median drop over the other 63 coordinate axes of the run. "Over, block 1" counts the runs over the criterion under the same edit restricted to block 1. The task change is the mean over ops of the clean score less that of the control at the seed; "worst op" is the lowest such change on any op of any run. The three stored recipes are at seeds 704 to 711 and `τ_ℓ anneal` at 712 to 719, so the rows compare different seeds.*

Every run keeps the anchor, and every run passes the task gate. The task stays within about a hundredth of the control on the mean over ops. The worst single op is `darken` at seed 717 (−0.036), against a gate of −0.05. `falloff` has one run that misses the gate on one op.

Removal is close to that of `falloff` on the seed mean, and below that of `τ_ℓ fixed`. Most of the difference comes from two seeds, 719 and 718, which remove about as little as the lowest runs of the step recipes (`falloff` and `step, soft τ`). The lower of the two, seed 719 at 0.74, is the only run below the lowest `falloff` run. It is still well clear of the line where a run counts as lost. The other six spread from 0.83 to 0.92.

So removal under `τ_ℓ anneal` varies across seeds the way it does under the step recipes, while under `τ_ℓ fixed` it was unusually even at its eight seeds. We can't tell from this whether the anneal or the seeds make the difference, since the two recipes never ran at the same seeds.

late spill scout. The late spill follows the anti anneal, and sharpening τ_ℓ to 0.025 over epochs 64 to 360 lowers it with no measured cost to removal or the task.

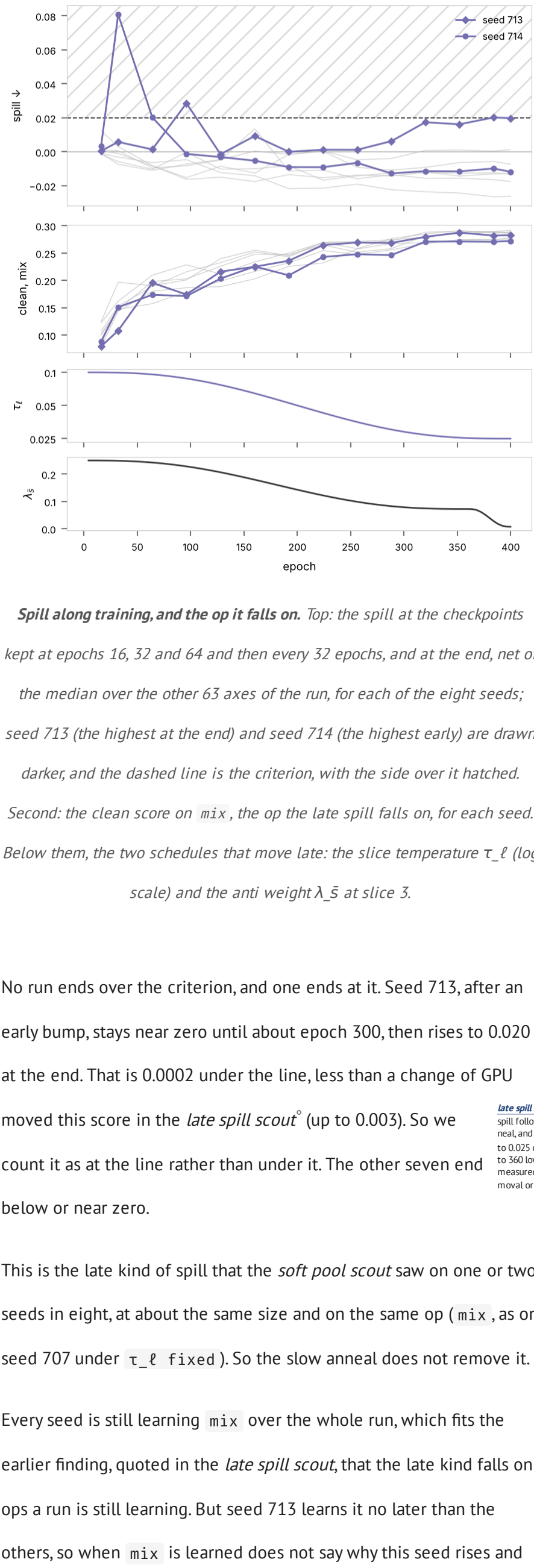
The *late spill scout*° found that sharpening τ_ℓ late kept removal where it was, seed by seed, which points to the seeds. But there the sharpening ran over epochs 64 to 360, carried on from stored epoch-64 states, where here it runs over the whole schedule.

Spill at the end and along training (E2)

Does any run end over the spill° criterion, and does the late rise of the *soft pool scout*° show up along training? The figure shows the spill of each seed along training. Beneath it are the clean score of the op the spill falls on, and the two schedules that move late.

Spill. The largest drop on any other op, at the dose where it is largest....

***soft pool scout*.** The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.



Spill along training, and the op it falls on. *Top: the spill at the checkpoints kept at epochs 16, 32 and 64 and then every 32 epochs, and at the end, net of the median over the other 63 axes of the run, for each of the eight seeds; seed 713 (the highest at the end) and seed 714 (the highest early) are drawn darker, and the dashed line is the criterion, with the side over it hatched. Second: the clean score on* `mix`*, the op the late spill falls on, for each seed. Below them, the two schedules that move late: the slice temperature* `τ_ℓ` *(log scale) and the anti weight* `λ_s̄` *at slice 3.*

No run ends over the criterion, and one ends at it. Seed 713, after an early bump, stays near zero until about epoch 300, then rises to 0.020 at the end. That is 0.0002 under the line, less than a change of GPU moved this score in the *late spill scout*° (up to 0.003). So we count it as at the line rather than under it. The other seven end below or near zero.

***late spill scout*.** The late spill follows the anti anneal, and sharpening τ_ℓ to 0.025 over epochs 64 to 360 lowers it with no measured cost to removal or the task.

This is the late kind of spill that the *soft pool scout* saw on one or two seeds in eight, at about the same size and on the same op (`mix`, as on seed 707 under `τ_ℓ fixed`). So the slow anneal does not remove it.

Every seed is still learning `mix` over the whole run, which fits the earlier finding, quoted in the *late spill scout*, that the late kind falls on ops a run is still learning. But seed 713 learns it no later than the others, so when `mix` is learned does not say why this seed rises and the others don't.

Early on, two seeds have brief bumps over the line. Seed 714 reaches 0.08 at epoch 32, while the position pool sharpens. It is still at the line (0.020) at the next checkpoint, and near zero by the one after. Seed 713 reaches 0.029 at epoch 96, and is near zero again by 128.

These bumps are the early kind of spill. The *soft pool scout* scored `τ_ℓ fixed` at seeds 704 to 711 on the same early checkpoints (16, 32, 64, and 128) and saw no spill above 0.003 there. By epoch 32 the anneal has moved τ_ℓ by about one percent, so the bumps do not come from the anneal.

Two runs of eight against none of eight can't tell whether the two draws of seeds differ here or the bumps are simply uncommon. The stored runs of the step recipes were not scored this densely, so we don't know how common such bumps are under them.

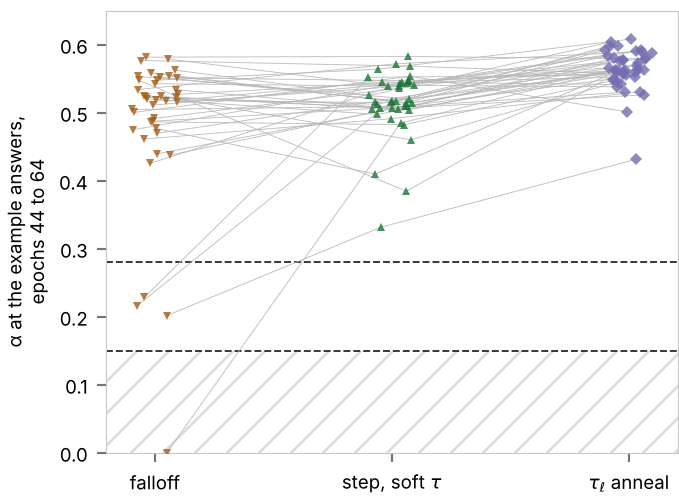
Under the edit at block 1 alone, one run goes over the criterion (713, at 0.028). The run that rises late is also the one that spills at block 1. That fits what *recipe-compare*° found under the pool: the spill enters with the edit at block 1. But one run of every recipe goes over at block 1 (table in E1), the step recipes included, so the block-1 spill on its own does not single out the pool.

***recipe-compare*.** The spill of the pool with a soft slice temperature enters at block 1, the one slice the pool pulls and the step recipes leave alone.

Lost anchors at forty seeds (E3)

How many of the forty runs lose the anchor by the call° at epoch 64, and how does that compare with `falloff` and `step, soft τ` at the same seeds?

Call. The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64....



The early call at forty seeds. *Mean α at the example answers at slice 3 over epochs 44 to 64, for each run, with a grey line joining the runs of a seed.* `falloff` *and* `step, soft τ` *are the calls of the loss-rate screen (and at seeds 704 to 711, the same window mean on their stored full runs). Below 0.15 a run is called lost and from 0.28 kept, with the band between finished and measured.*

The call finds all forty runs kept. The lowest window mean (mean α° over epochs 44 to 64) is seed 724 at 0.43, above the line where a run counts as kept, so no run needed finishing to settle its call.

Alignment. How far a state points along e1....

With none lost of forty, the rate of lost anchors is below 7% (a one-sided 95% bound), the same bound as for `step, soft τ` in the *loss-rate screen*°. The call is made at epoch 64. The eight runs that went on to epoch 400 kept the anchor there too. The other 32 runs were scored only for the call, so the count says nothing about their spill°.

loss-rate screen. A call from α at the example answers over epochs 44 to 64 sorts lost anchors from kept ones; at forty seeds `falloff` lost one and `step, soft τ` none.

Spill. The largest drop on any other op, at the dose where it is largest....

Seed by seed, the window means are above those of `falloff` on 38 of forty seeds and above those of `step, soft τ` on 35. A soft position temperature at the start, which `τ_ℓ anneal` shares with `step, soft τ`, may be what brings the example answers onto e1 early on every seed. But `τ_ℓ anneal` is also above `step, soft τ` on most seeds, so the soft start is not the whole of it. The pool over three slices may add to it, but nothing here separates the two. The soft start was added so the pull would not latch onto a syntax token while it settles, and that is its main job; whether a sharp slice temperature now does that job on its own is not something this run tests.

Discussion

`τ_ℓ anneal` meets three of the four checks outright. Every run at the fresh seeds keeps the anchor (check 1), none ends over the spill° criterion though one ends at it (check 2), and the count of lost anchors is the lowest a count of forty can give (check 4).

Spill. The largest drop on any other op, at the dose where it is largest....

Check 3 has two halves, and the task half passes on every run. On the removal° half, `τ_ℓ anneal` is a little below `falloff` on the seed mean, by less than two draws of eight seeds can differ (Scope and method). So that check is met within the seed spread rather than outright.

Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....

The late rise is still there on one seed in eight, at about the size the *soft pool scout*° saw. So the slow anneal of τ_ℓ does not remove the late kind of spill; at most it keeps it at the line here.

***soft pool scout*.** The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.

The *late spill scout*° found the effect of the late sharpening to be about the size of the difference between two GPUs, so a confirmation at eight seeds can't tell whether it helps. What the run shows is no cost we can separate from the seed draw: removal is lower than under `τ_ℓ fixed` on the mean, and E1 can't say whether the anneal or the seeds did that.

***late spill scout*.** The late spill follows the anti anneal, and sharpening τ_ℓ to 0.025 over epochs 64 to 360 lowers it with no measured cost to removal or the task.

Two things favor the pool recipe over `falloff`. It chooses the slice within a range set by the architecture (slices 1 to $L - 1$), and its early call° is never unsure, where `falloff` stayed in the unsure band on three seeds of forty. Its count of lost anchors, none of forty against one, is a tie at this many seeds.

Call. The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64....

Against the pool recipe, one seed in eight ends at the spill line, while the eight stored runs of `falloff` end at most at 0.014. That is one run, set against the seed draw and the GPU shift, so it is a small mark rather than a difference we can rely on. Under the block-1 edit it goes over on one seed, as one run of every recipe does, and its removal varies across seeds as under the step recipes.

The early bumps on two seeds are the other mark, and the larger one goes well over the line. Under `τ_ℓ fixed` the *soft pool scout* saw none at the same checkpoints, but at epoch 32 the anneal has moved τ_ℓ by about one percent, so the two recipes are nearly the same there and the difference is most likely between the two draws of seeds. Both bumps cleared within 64 epochs, and the step recipes were never scored this early, so we can't say whether they bump too.

References

soft pool scout

The pool over slices 1 to 3 with τ_ℓ held at 0.1 keeps the anchor on all eight seeds with no early spill, and one seed rises late to near the criterion.

late spill scout

The late spill follows the anti anneal, and sharpening τ_ℓ to 0.025 over epochs 64 to 360 lowers it with no measured cost to removal or the task.

loss-rate screen

A call from α at the example answers over epochs 44 to 64 sorts lost anchors from kept ones; at forty seeds `falloff` lost one and `step, soft τ` none.

recipe-compare

The spill of the pool with a soft slice temperature enters at block 1, the one slice the pool pulls and the step recipes leave alone.

---

1. Slice temperature: how sharply the pooled pull picks among slices. The pull is a softmin-weighted mix over slice and position pairs, and τ_ℓ sets how peaked the weights over slices are, as the pooling temperature τ_p does over positions. ↩

2. Minimum-jerk: a smooth S-shaped path that starts and ends at rest, so the change is slow at both ends and fastest in the middle. ↩