Fixed confirmation

> The pool over slices 1 to 3, with its slice temperature held fixed, keeps the anchor on all eight runs at fresh seeds, and on all forty runs of a count that calls° each run at epoch 64. Paired seed by seed with the annealed version of the recipe, it removes about as much, with no late rise in spill°.
>
> But some seeds that did well under one recipe did poorly under the other, and the two runs of a seed parted ways early, while the two recipes were still nearly the same. So the variation between runs seems to come largely from how sensitive a run is to small changes, rather than from either recipe.

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....

The *anneal confirmation*° ran the pool over slices 1 to 3 with its slice temperature τ_ℓ annealing from 0.1 to 0.025, at eight fresh seeds (712 to 719).[1] Every run kept the anchor, but removal° varied across seeds more than under the recipe it grew from: `τ_ℓ fixed` from the *soft pool scout*°, which holds τ_ℓ at 0.1.

***anneal confirmation*.** With τ_ℓ annealing from 0.1 to 0.025, the pool keeps the anchor on all eight runs at seeds 712 to 719 and on all forty in the count, with removal varying across seeds about as under `falloff`.

1 Slice temperature...

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....

***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.

Two seeds removed about as little as the lowest runs of the step recipes. The two recipes had run at different seeds, so either the anneal or the seed draw could account for that.

This report runs `τ_ℓ fixed` at the same fresh seeds, so the two recipes pair seed by seed. It is also the confirmation run for adopting `τ_ℓ fixed`, the simpler of the two, as the recipe of record: it trains to the end at seeds 712 to 719, and to epoch 64 at forty seeds (704 to 743) to count lost anchors. On the stopped runs of both recipes it also scores spill at epochs 16, 32, and 64, which the *anneal confirmation* left unscored on most of its forty runs.

Observations

- The two recipes at the same seeds (E1): every run keeps the anchor. On the seed mean, removal° is a little higher than under `τ_ℓ anneal` and spreads less, but seed by seed the difference runs both ways, by up to about a tenth. Of the two seeds that removed least under the anneal, one removes about as much as the seed mean here and the other is still among the lowest, and a seed that removed well under the anneal is the lowest here. One run misses the task gate, on a seed that learns two ops late.

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 near the spill line, and none rises late, including the seed that rose to the line under the anneal. One run has a brief early bump over the line, which clears by the next checkpoint.

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

- Lost anchors and early spill at forty seeds (E3): the call° finds all forty runs kept. At epochs 16 to 64, five runs of forty spill over the line at some checkpoint under `τ_ℓ fixed` and two under `τ_ℓ anneal`, mostly at different seeds, though the two recipes barely differ before epoch 64.

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, not a preregistered one. It uses the four checks that the *anneal confirmation*° set for its recipe, which guide the choice of recipe without deciding it, so a recipe that misses one narrowly can still be the better choice:

***anneal confirmation*.** With τ_ℓ annealing from 0.1 to 0.025, the pool keeps the anchor on all eight runs at seeds 712 to 719 and on all forty in the count, with removal varying across seeds about as under `falloff`.

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.

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....

`τ_ℓ fixed` (in full, `τ_ℓ 0.1, τ_ℓ fixed`) pulls e_1 at the pool over slices 1 to 3, with τ_ℓ at 0.1 throughout. The position temperature τ_p sharpens from 1 to 0.1 by epoch 40, the anchor weight warms up over the first 40 epochs, and the anti weight anneals to the per-slice holds of `falloff` by epoch 360.

`τ_ℓ anneal` differs in one schedule: τ_ℓ moves from 0.1 to 0.025 over the whole run, slowly at first, so it is still 0.096 at epoch 64. Over the first 64 epochs the two recipes are nearly the same.

The forty runs of the count use the seeds of the *loss-rate screen*°, so each shares its initialization, batches, and label draws with a run of `falloff`, of `step, soft τ`, and of `τ_ℓ anneal`. Each stops at epoch 64 for its call°. At seeds 704 to 711 the stopped runs are those of this recipe from the *soft pool scout*°; the other 32 are new.

***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.

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....

***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 runs at 712 to 719 continue from their stops to epoch 400, keeping a checkpoint every 32 epochs. They share their controls with the *anneal confirmation* (no pull and no anti term, at the same seed). All runs trained and were measured on L4 GPUs, as the *anneal confirmation* was.

Within the fresh seeds, the two recipes pair seed by seed, so a difference between them is not a difference of seed draws. Against the stored recipes at seeds 704 to 711, a difference also includes the variation between two draws of eight seeds. Under `falloff` and `step, soft τ`, removal varies by about 0.15 across seeds.

The measurements

Every edit below is the e_1 edit. Each score is the drop under that edit, less a baseline: the median drop when the same edit is made 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. On the 32 runs stopped at epoch 64 only the first three exist.

Task change

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

Call

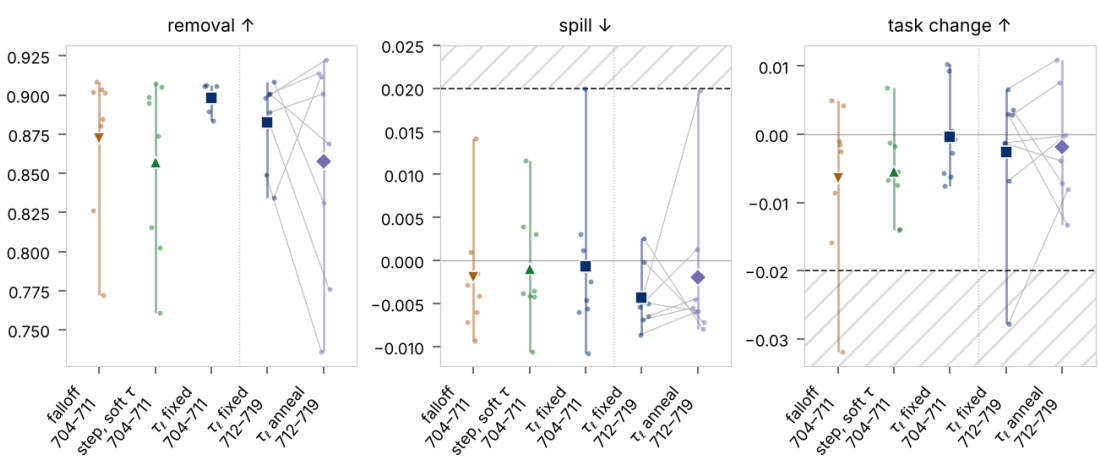
: The early verdict on whether a run kept the anchor, from mean α at the example answers at slice 3 over epochs 44 to 64. A run below 0.15 is called lost, and one at 0.28 or above is called kept. A run in between is finished and measured.

The two recipes at the same seeds (E1)

Does removal° vary across the fresh seeds under `τ_ℓ fixed` as it did under `τ_ℓ anneal`, and does every run keep the anchor and pass the task gate? If the anneal made removal uneven, the two low seeds of the *anneal confirmation*° should remove well here. If those seeds are hard for the pool, they should be low under both.

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....

anneal confirmation. With τ_ℓ annealing from 0.1 to 0.025, the pool keeps the anchor on all eight runs at seeds 712 to 719 and on all forty in the count, with removal varying across seeds about as under `falloff`.



Removal, spill, and task at epoch 400, by recipe. *One column per recipe and seed set: the stored runs of `falloff`, `step, soft τ` and `τ_ℓ fixed` at seeds 704 to 711, and `τ_ℓ fixed` and `τ_ℓ anneal` at the fresh seeds 712 to 719, with a grey line joining the two runs of each fresh seed and a dotted rule between the two seed sets. The seeds are small dots over a thin bar for their range, and the seed mean is 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 fixed`	712–719	0.88 (0.83 to 0.91)	−0.004 (−0.009 to 0.003)	0 of 8	0 of 8	−0.003 (−0.028 to 0.007)	−0.072
`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 last two rows share their seeds, and the controls at those seeds.*

Every run keeps the anchor. Removal under `τ_ℓ fixed` is a little above that of `τ_ℓ anneal` on the seed mean, and above `falloff` too. It spreads more than it did under the same recipe at seeds 704 to 711, though less than under the anneal.

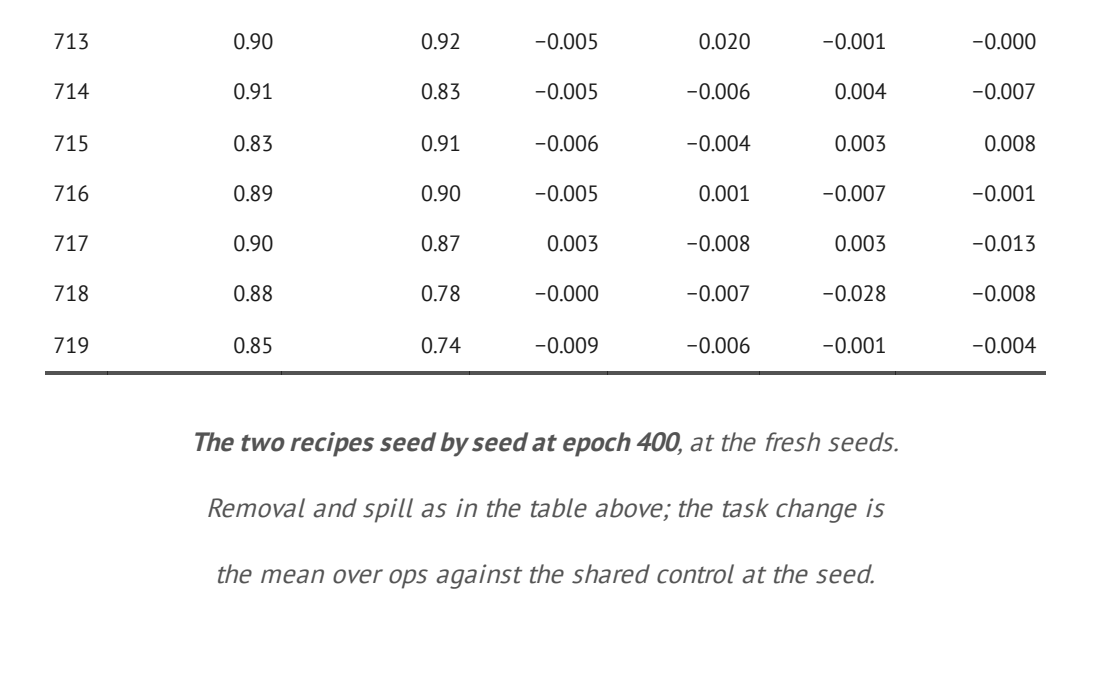
seed	removal, fixed ↑	removal, anneal ↑	spill, fixed ↓	spill, anneal ↓	task, fixed ↑	task, anneal ↑
712	0.90	0.91	−0.007	−0.005	0.007	0.011
713	0.90	0.92	−0.005	0.020	−0.001	−0.000
714	0.91	0.83	−0.005	−0.006	0.004	−0.007
715	0.83	0.91	−0.006	−0.004	0.003	0.008
716	0.89	0.90	−0.005	0.001	−0.007	−0.001
717	0.90	0.87	0.003	−0.008	0.003	−0.013
718	0.88	0.78	−0.000	−0.007	−0.028	−0.008
719	0.85	0.74	−0.009	−0.006	−0.001	−0.004

The two recipes seed by seed at epoch 400*, at the fresh seeds. Removal and spill as in the table above; the task change is the mean over ops against the shared control at the seed.*

Seed by seed, the differences run both ways, by up to 0.11: `τ_ℓ fixed` removes more on four of the eight seeds. The two low seeds of the *anneal confirmation* (719 and 718) remove more here, but seed 715, which removed well under the anneal, is the lowest here, and seed 719 is among the lowest under both.

So neither account from the *anneal confirmation* fits on its own. The low seeds there were not low because of the seed alone, since seed 718 removes about as much as the seed mean here. Nor does the anneal lower removal across the board, since it removes more on some seeds.

By epoch 400 the anneal has taken τ_ℓ down fourfold, so the end gaps could still be a late effect of the anneal that differs by seed. Removal along training shows when the gaps open.



Removal along training, paired by seed. *One panel per fresh seed: removal under the full edit at the checkpoints kept at epochs 16, 32 and 64 and then every 32 epochs, and at the end, for `τ_ℓ fixed` and `τ_ℓ anneal`. Net of the median over the other 63 axes of the run, as a share of the way to the target null of the run at the end. The strips beneath are the slice temperature `τ_ℓ` of each recipe (log scale).*

The runs of a seed part ways early. At epoch 32, where the anneal has moved τ_ℓ by under one percent (0.0994 against 0.1), the two runs of some seeds are up to 0.43 apart in removal, and most pairs come back together by about epoch 160.

The two largest end gaps (seeds 718 and 719, a tenth each) are set by epoch 64 to 96, when τ_ℓ under the anneal is still within about a tenth of 0.1 (0.088 at epoch 96). They change little afterward, while the anneal does most of its work. The smaller lasting gaps, on seeds 714 and 715, wander by a few hundredths along the way.

So the end gaps were set while the two recipes were nearly the same. That fits a picture in which the final removal of a run is sensitive to small changes early on, while the anchor settles. If so, eight seeds of either recipe give a noisy estimate of its mean.

This does not rule out that a slightly sharper τ_ℓ at epochs 64 to 96 matters, but a difference that small would be hard to tell from the same sensitivity. Seed 719, low under both, may also be a hard seed for the pool.

Seven runs pass the task gate, with the mean over ops within about a hundredth of the control. Seed 718 misses it: its mean over ops is −0.028, and two ops, `value-hsv` and `sat-hsv`, end −0.07 and −0.06 below the control.

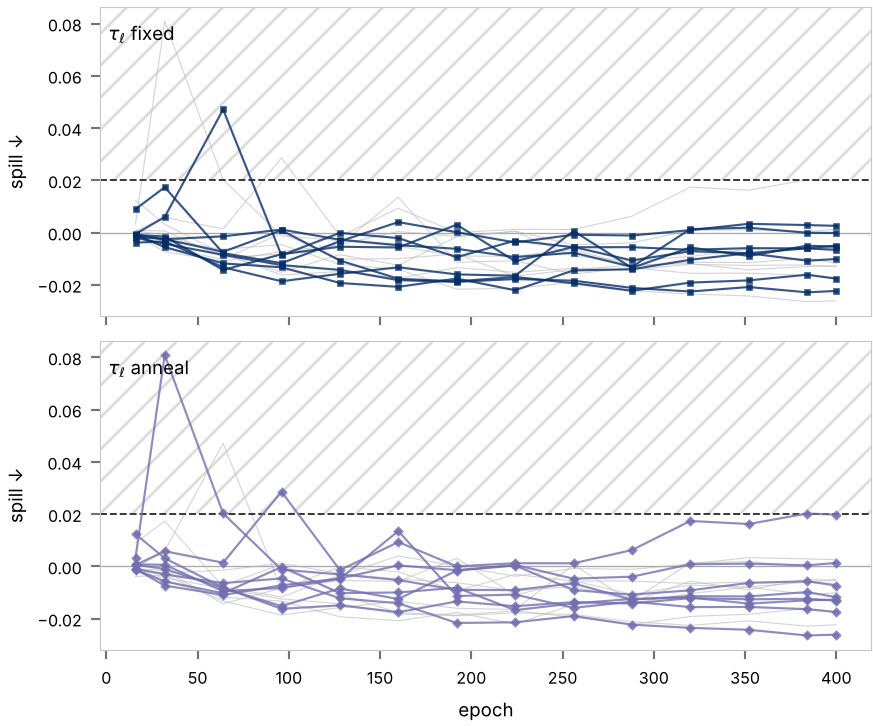
This run learns those two ops late: at epoch 160 its clean score on `value-hsv` is 0.22, where the other seven runs are at 0.33 or more, and it is still rising at the end. Under the anneal, the same seed learned them on time, and another seed (717) was the late one but caught up. So this too looks like a difference between runs more than a property of the recipe. `falloff` also has one run of eight that misses the gate; `step, soft τ`, `τ_ℓ anneal`, and `τ_ℓ fixed` at 704 to 711 have none.

Spill at the end and along training (E2)

Does any run end over the spill° criterion, and do the late rise and the early bumps of the *anneal confirmation*° show up at the same seeds under `τ_ℓ fixed`?

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

anneal confirmation. With τ_ℓ annealing from 0.1 to 0.025, the pool keeps the anchor on all eight runs at seeds 712 to 719 and on all forty in the count, with removal varying across seeds about as under `falloff`.



Spill along training, paired by seed. *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 fresh seeds.* ***Top:*** `τ_ℓ fixed`; ***bottom:*** `τ_ℓ anneal`, *with the other recipe as faint lines behind. The dashed line is the criterion, with the side over it hatched.*

No run ends over the criterion, nor near it: the largest end spill is 0.003, against 0.020 under the anneal. Seed 713, which rose late to the line under the anneal, stays near zero or below to the end here, and no other seed rises late.

The *soft pool scout*° saw the late rise on one seed in eight under `τ_ℓ fixed`, so it comes and goes between runs. With eight seeds in each draw, a late rise on none here against one there is within the variation between two draws.

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.

Under the edit at block 1 alone, no run goes over the criterion. Every other recipe had one run that did (table in E1).

One run has an early bump: seed 716 reaches 0.047 at epoch 64 and is near zero by epoch 96 (−0.009). Under the anneal, the early bumps were on seeds 714 and 713; here those seeds stay below the line throughout. As with removal°, the bumps move between seeds when the recipe changes only a little.

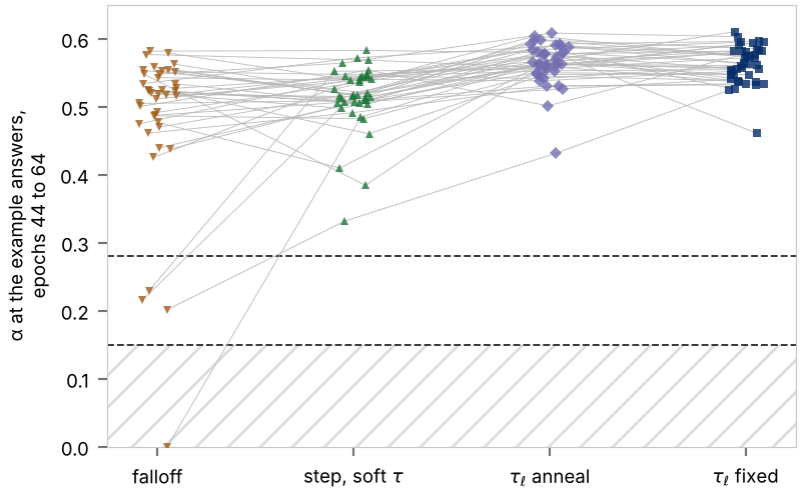
Removal. The drop on `difference` under the full edit at every position and slice, as a share of the way from the clean score to the target null....

Lost anchors and early spill at forty seeds (E3)

How many of the forty runs lose the anchor by the call° at epoch 64, and how often does either recipe spill° early? The second question extends E2 to the forty stopped runs.

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....

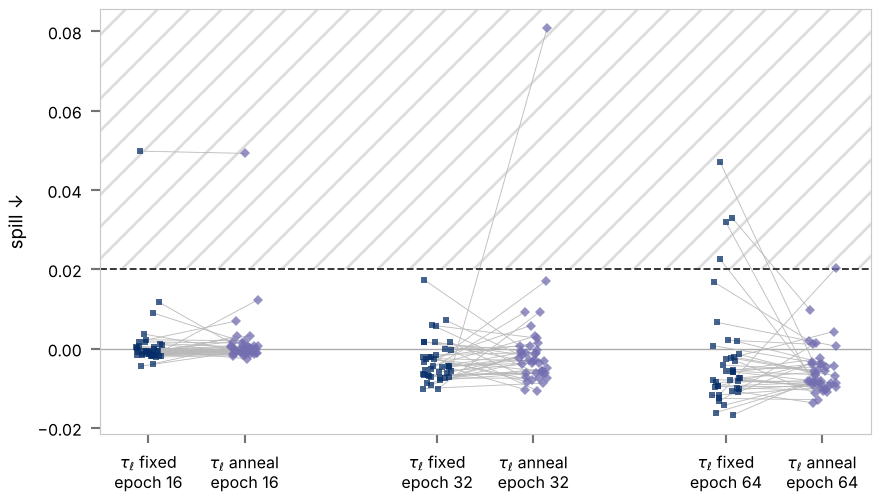


The early call at forty seeds. *Mean α at the example answers at slice 3 over epochs 44 to 64, for each run, with a grey line joining the runs of a seed.* `falloff` *and* `step, soft τ` *are the calls of the loss-rate screen (and at seeds 704 to 711, the same window mean on their stored full runs);* `τ_ℓ anneal` *is the count of the anneal confirmation. Below 0.15 (hatched) 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 717 at 0.46, well above the line where a run counts as kept, so no run needed finishing to settle its call. With none lost of forty, the rate of lost anchors is below 7% (a one-sided 95% bound), the same bound as for `τ_ℓ anneal` and `step, soft τ`.

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

Seed by seed, the window means are above those of `falloff` on 37 of forty seeds and above those of `step, soft τ` on 37, as under the anneal. Against `τ_ℓ anneal` they are above on 20, as expected for two recipes that are nearly the same up to the call.



Early spill at forty seeds. *The spill on the checkpoints kept at epochs 16 and 32 and at the stop at epoch 64, for each of the forty stopped runs of the two recipes, net of the median over the other 63 axes of the run; a grey line joins the two runs of each seed. The dashed line is the criterion, with the side over it hatched.*

recipe	over, epoch 16 ↓	over, epoch 32 ↓	over, epoch 64 ↓	over at any ↓	largest ↓
`τ_ℓ fixed`	1 of 40	0 of 40	4 of 40	5 of 40	0.050
`τ_ℓ anneal`	1 of 40	1 of 40	1 of 40	2 of 40	0.081

Early spill at forty seeds*: the runs over the criterion at each early checkpoint, the runs over it at any of the three, and the largest spill of any run at any of them.*

Early spill is uncommon under both recipes, and brief where we can follow it. `τ_ℓ fixed` goes over the line on five runs of forty at some early checkpoint, mostly at epoch 64, and `τ_ℓ anneal` on two. Only seed 734 spills under both, at epoch 16, where the two recipes barely differ. The runs at 704 to 711 stay at or below 0.002, as the *soft pool scout*° found under its own baseline.

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.

So at the same seeds and nearly the same schedule, the early bumps land on different runs. The runs over the line at epoch 64 under one recipe sit near zero or below under the other at the same seed (paired grey lines in the figure), so these are not runs that sit near the line under both recipes and cross it under one. Five against two is too few to say one recipe spills early more often than the other.

The bump at seed 734 is gone by epoch 64 under both, the one at seed 716 under `τ_ℓ fixed` by epoch 96 (E2), and the one at seed 714 under the anneal by epoch 96. The other runs over the line at epoch 64 stopped there, so we can't say whether their bumps cleared too.

Discussion

`τ_ℓ fixed` meets three of the four checks outright: every run at the fresh seeds keeps the anchor (check 1), none ends near the spill° criterion (check 2), and none of forty loses the anchor by the call° (check 4).

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

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….

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….

Check 3 has two halves. The removal° half passes: removal is above that of `falloff` on the seed mean. The task half misses by one run of eight, on a seed that learns two ops late, as `falloff` also misses by one (E1).

We take the two recipes as removing about the same, since the difference in the seed mean between them is within the differences seen at single seeds.

The pairing with `τ_ℓ anneal` finds that the uneven removal came mostly from neither the anneal nor the seeds: the low runs move between seeds when the recipe changes only a little. So the spread of removal across eight runs is partly how sensitive a run is to small changes, and a seed mean over eight runs carries that noise.

The same holds for the other problems seen at eight seeds: the late rise in spill, the early bumps, and the late learner each appeared under one recipe at a seed and not under the other. A count over more runs, as in E3, is the steadier measure for these.

Between the two, `τ_ℓ fixed` is simpler, and nothing in this confirmation suggests the anneal improves on it. The *late spill scout*° found that the late sharpening lowered the late rise partway on both of its rising runs, and a change of GPU moved one of those runs by more than that. Here the late rise appeared under the anneal and not under the fixed recipe.

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.

We adopt `τ_ℓ fixed` as the recipe of record for D2.2, and the next step is to run the full scoring suite on it.

References

anneal confirmation

With τ_ℓ annealing from 0.1 to 0.025, the pool keeps the anchor on all eight runs at seeds 712 to 719 and on all forty in the count, with removal varying across seeds about as under `falloff`.

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.

---

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. ↩