Soft pool scout

> A sharper pool over slices stopped the early spill° of `pool 1–3, soft τ`. Of sixteen schedules of the pooled pull screened at epoch 64, those with the slice temperature ending at 0.1 had no run spill early, however soon or late the temperatures sharpened.
>
> We carried two of these conditions on to epoch 400. Both stay within the spill criterion on every seed, and both remove a little more than the recipe of record at no cost to the task. Holding the slice temperature at 0.1 from the start is simpler and cleaner, and a candidate for the recipe of record.

Spill. The side effect of an edit on the other ops…

Recipe-compare left one thing between `pool 1–3, soft τ` and the recipe of record: two of its eight runs go over the spill criterion, and the spill enters at block 1, the one slice the pool pulls and the step leaves alone. Spill-timing found that both of those runs spill early, while the pool sharpens, and that this early kind already shows at epoch 64. So a schedule that keeps block 1 closed should show it at epoch 64, at a sixth of the cost of a full run.

Stage 1 trains sixteen conditions of the pool at its eight seeds (704 to 711), each changing when the two pool temperatures sharpen, how sharp the pool over slices ends, or when the anchor weight warms up. Every run stops at epoch 64, where we score whether it is keeping the anchor and how much the edit spills. In stage 2, the two conditions with τ_ℓ ending at 0.1 that change least from the base (sharpening by epoch 40 as the base does, and held at 0.1 from the start) are carried on to epoch 400 from where they stopped, and measured as recipe-compare measured the three recipes, with the spill along training from checkpoints kept on the way.

Observations

- Lost anchors (E1): no schedule lost the anchor. All 136 runs, the base included, are called kept, the lowest well above the line. The fixed temperatures and the stretched warm-up keep it a little less firmly.
- Spill at epoch 64 (E2)°: the final slice temperature decides the spill; when the temperatures sharpen hardly matters. At τ_ℓ 0.1 no run is over the criterion, and the four conditions sit at or below the quiet recipes. At 0.4 (the base) about a quarter of the runs are over. Moving the λ_a warm-up shifts the spill by less than the seeds vary, and fixing the temperatures from the start, at the base values (τ_ℓ 0.4), spills the most.

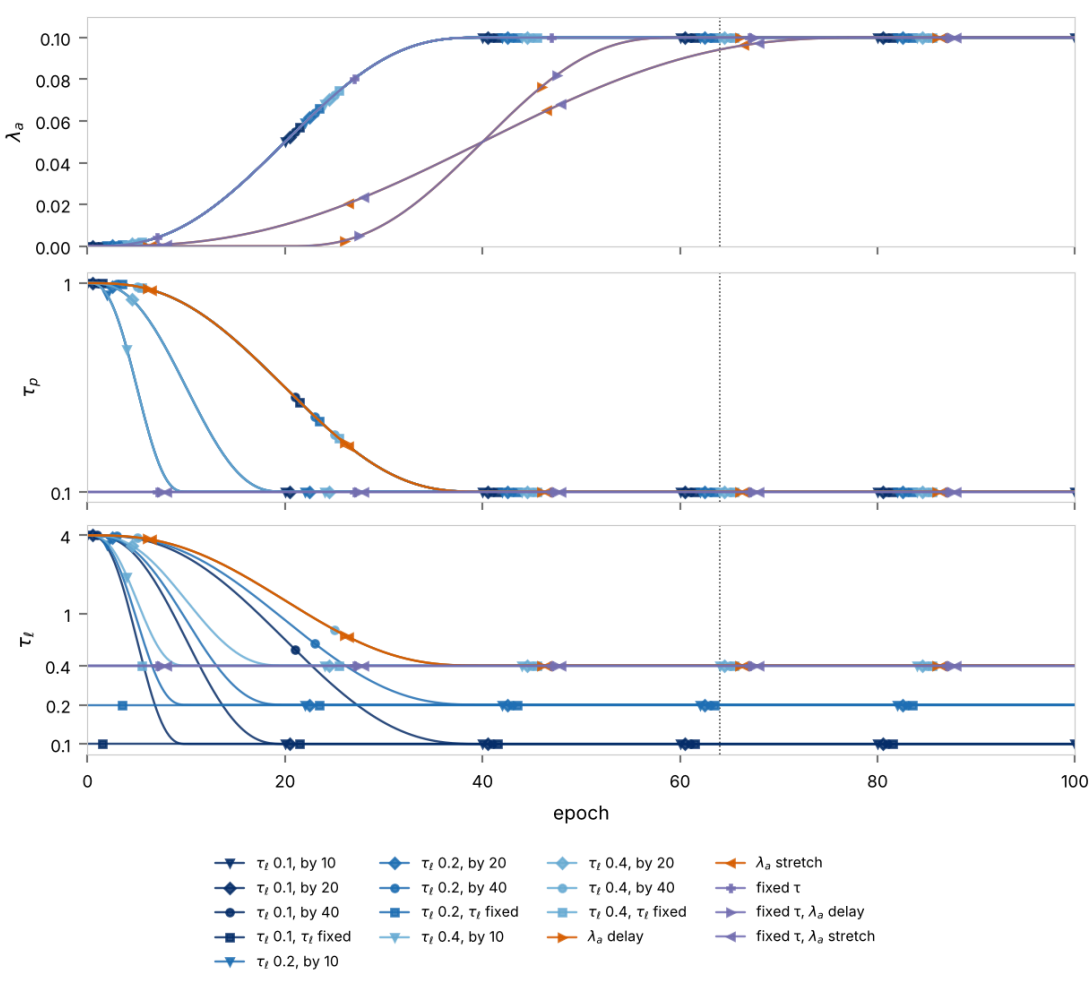
Spill. The side effect of an edit on the other ops…

- Where the evidence is heading (E3): every condition peaks at slice 3, and none leans on the query answer. What changes is slice 1: a sharper pool over slices puts less of the anchor there, and the runs that spill are the ones with more of it at slice 1.
- Removal, spill, and task at epoch 400 (E4): both finished conditions keep the anchor, remove a little more than `falloff` and the base, and stay within the spill criterion on every seed, with one or two seeds at the line. Neither costs the task. Holding τ_ℓ fixed is the tighter of the two on removal and spill.
- Spill along training (E5): no finished run spills early. The spill that remains comes late, after epoch 128, on one or two seeds, and reaches about the criterion by the end. So the screen at epoch 64 called the early kind correctly, and only the finished runs show the late kind.
- Where the spill enters (E6): the evidence still peaks at slice 3 on every seed. The edit at block 1 alone costs the other ops about what it costs them under the step, and less than under the base. With τ_ℓ fixed that holds seed by seed; when it sharpens by 40 it holds on the mean only. The other ops sit as far from e_1 at the first two slices as under the step, and the lean of the color embeddings along lightness falls back to the level of the step.

Scope and method

This is an exploratory study, with no preregistration and no gate. Every condition starts from `pool 1–3, soft τ` (the base, which is not the recipe of record): the pull is pooled over the slice and position pairs of slices 1 to 3, the position pool temperature τ_p sharpens from 1 to 0.1 and the slice pool temperature τ_ℓ from 4 to 0.4 over the first forty epochs, and the anchor weight λ_a warms up over the same forty epochs to its peak of 0.1. The anti weight is that of `falloff`, the recipe of record, and the rest of training is as in soft-pool.

Twelve conditions cross two factors. One is τ_ℓ at the end, with levels 0.1, 0.2 and 0.4 (the base), each starting from 4. The other is when the temperatures are sharp: by epoch 10, 20 or 40 (the base), both on the same window, or τ_ℓ at its end value from epoch 0 while τ_p still sharpens by 40, which tells the two temperatures apart. Five more conditions move the λ_a warm-up at the base temperatures: a delay (zero until epoch 20, then up by 60) and a stretch (up by 80), each with the soft temperatures and with both temperatures fixed at their end values from the start, and the fixed temperatures with the base warm-up. The fixed-temperature conditions might have lost the anchor, which is what they were there to show: if a delayed pull keeps it at a fixed τ, the soft start was a delay by other means. The schedules are drawn below.



The schedules the conditions change. *The anchor weight λ_a, the position pool temperature τ_p and the slice pool temperature τ_ℓ (both on a log scale) over the first hundred epochs, one line per condition, in its ink and marker; where conditions share a schedule, the later line is drawn over the earlier, with the markers staggered. The dotted line is the stop at epoch 64. After epoch 100 every condition is the same.*

Each run trains on the full 400-epoch schedule and stops at epoch 64, keeping what it needs to carry on (a resumed run matches a straight run bit for bit). The base is not trained again: spill-timing trained it at these seeds with the same scoring at every record, so its epoch-64 record is the base here. The epoch-64 checkpoints of `falloff` and `step, soft τ` left by the loss-rate screen are reference points on the same scale.

A screen at epoch 64 can rank the conditions by the early kind of spill° and by lost anchors. It cannot clear a condition: the late kind of spill, which grows on `value-hsv` after epoch 100 on seeds that learn it late, is invisible here. Spill-timing found the base about 0.02 above the quiet recipes at epoch 64, pooled over seeds, so the seed mean of eight runs can tell an early-spilling condition from a quiet one at about that resolution, and a single run says little. We compare many conditions, so a mean a little apart from its neighbors is a lead only. The pattern across a factor, which pools 32 runs at each level, is firmer.

Spill. The side effect of an edit on the other ops…

The measurements

- The call: α° (the cosine of the state with e_1) at the example answers at slice 3, averaged over epochs 44 to 64. As in the loss-rate screen, a run is called lost below 0.15, kept from 0.28, and uncertain between.
- Spill at epoch 64: the largest drop on another op when e_1 is projected out at every position and slice, less the median drop under eight reference edits (each projecting out one other basis direction), over the whole held-out set, as in spill-timing. A drop is in expected exact match°. Lower is better. The criterion is 0.02, though at epoch 64 it is a guide only, since it is set for a finished run.
- Where the evidence for the op sits: α at the example answers and at the query answer, by slice, at epoch 64.

Alignment. How far a state points along e_1…

Expected exact match. The task score, as the chance that a drawn answer agrees with a true one.…

Stage 2

Each run of the two finished conditions (τ_ℓ ending at 0.1 and sharpening by epoch 40, and τ_ℓ held at 0.1 from the start while τ_p sharpens by 40) carries on from epoch 64 to epoch 400. The checkpoints kept at epochs 16, 32, 64, 128, and 256 along the way are scored for the spill.

We measure the finished runs as recipe-compare measured `falloff`, `step, soft τ`, and the base at the same seeds, and set them beside its numbers for those three. Each run at a seed shares its initialization, batches, and label draws with the base and with the control of ex-2.2.23 at that seed. So each comparison is paired by seed, and the paired difference from the base says whether a gap is larger than the variation between seeds.

- Removal: the drop in the score of `difference` under the full edit, as a share of the way from the clean score to the target null°. It is net of 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). A run that removes less than 0.5 has lost the anchor.
- Spill: the largest drop on another op at any dose, net of the same median. The criterion is 0.02.
- Task change: the clean score per op, less that of the control at the seed. The task gate fails below -0.02 on the mean over ops, or below -0.05 on any one op.
- Spill along training: the spill at each kept checkpoint and at the end, against eight reference edits, as at the stop.
- Restricted edits, the write, the grading, and the landing, as in recipe-compare: the full edit made at a set of slices only, net of the same median as removal and spill; how much e_1 the blocks after the edit put back on the other ops, against the bound the clean pass sets; the share removed on `difference` contexts; and how far the edited answers move toward the target null.

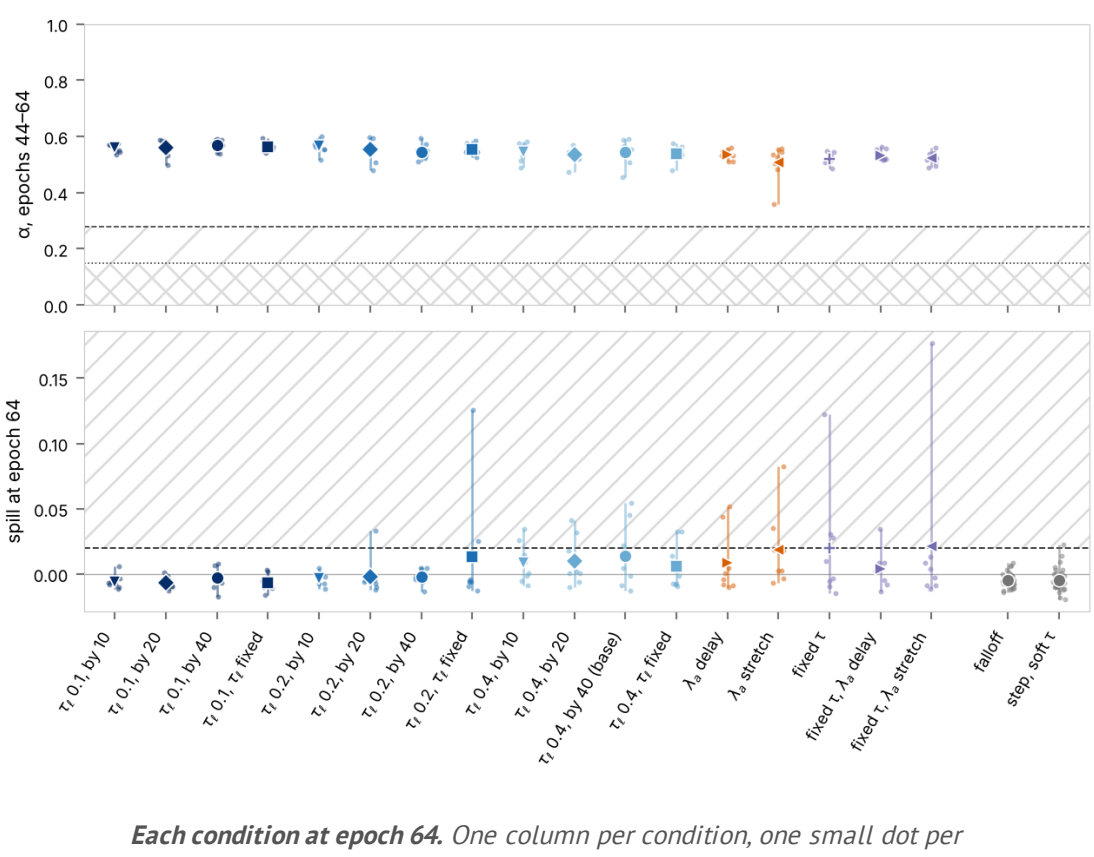
Target null. Where a removal should land…

Eight seeds are too few to add to a count of lost anchors. The loss-rate screen took that count at forty seeds for the step recipes, and nobody has taken it for the pool.

Lost anchors (E1)

The fixed-temperature conditions were the ones most likely to lose the anchor, since a sharp pool from the first step is what let the pull latch onto the query answer on one seed in slice-terms. Did any schedule lose it by epoch 64?

None did. Every run is called kept, and the lowest call (0.36, on a stretched warm-up) is still above the line by a wide margin. The figure (top) shows the calls close together across all sixteen conditions. The fixed temperatures and the stretch sit a little lower than the rest, which fits a pull that has had less time at full strength by epoch 64, but none comes near the uncertain band.



Each condition at epoch 64. *One column per condition, one small dot per seed, the seed mean as the large mark and the seed range as a thin bar. Top: α at the example answers at slice 3, averaged over epochs 44 to 64 (the loss-rate call); below the dotted line a run is called lost (cross-hatched), above the dashed line kept, and between the two uncertain (lightly hatched). Bottom: the spill at epoch 64, the largest drop on another op under the full edit less the median drop under eight reference edits; the dashed line is the criterion, with the side over it hatched. The gray columns at the right are the loss-rate screen's epoch-64 checkpoints of the two quiet recipes, at 34 seeds each.*

Two soft starts are in play: τ_p sharpening from 1, and τ_ℓ sharpening from 4. The conditions with τ_ℓ fixed drop the second, the fixed-τ conditions drop both, and all of them kept the anchor on every seed. So in this pool neither soft start is what keeps the anchor, at least to epoch 64 and at these seeds. The step may differ: in the loss-rate screen, the step at a fixed τ_p (`falloff`) lost the anchor on one seed of forty and the step with the soft start on none, which fits a soft τ_p that helps the step, though one loss is too few to show it. Eight seeds would miss a loss rate of one in ten or so, and a run can still lose the anchor late, so the fixed temperatures are not cleared for this pool either.

Spill at epoch 64 (E2)

Spill. The side effect of an edit on the other ops…

If spill° enters through block 1 while the pool is soft, then sharpening sooner, sharpening the slice pool further, or holding the pull back until e1 is cleared might each close it. Which of them lowers the spill at epoch 64?

The slice temperature at the end does, and the others barely move it. The figure (bottom) groups the crossing by τ_ℓ: at 0.1 the four timing levels all sit at or below the quiet recipes, with no run of 32 over the criterion; at 0.2 the early spill is a little higher, with 3 runs over; at 0.4, the base level, 9 of 32 are over. Within each level, sharpening by epoch 10, 20 or 40, or holding τ_ℓ fixed from the start, makes no difference we can see at eight seeds. So for the early spill, what matters is how sharp the slice pool ends, and how soon it gets there barely does. That also covers where the slice pool starts: at 0.1, sharpening from 4 and holding at 0.1 from epoch 0 spill about the same, so a start in between would be unlikely to differ.

This is a stronger effect of τ_ℓ than slice-terms saw. Its E4 trained the pool over slices with τ_ℓ from 0.1 to 1.6 and found no trend in removal, with the spill a little higher at the two softest levels. But that was one seed, chosen because its pull had settled on block 4, at a fixed τ_p and with the pool over blocks 1 to 4. Here the pool leaves block 4 out, and the spill is a per-seed event: at τ_ℓ 0.4 most seeds stay quiet, and one seed would most likely have been one of those.

condition	kept ↑	lowest call	spill ↓	seed range	over	peak slice	α at peak	α, query
`τ_ℓ 0.1, by 20`	8/8	0.50	−0.007	−0.013 to 0.001	0	3	0.57	0.89
`τ_ℓ 0.1, τ_ℓ fixed`	8/8	0.54	−0.006	−0.016 to 0.003	0	3	0.57	0.88
`τ_ℓ 0.1, by 10`	8/8	0.54	−0.006	−0.011 to 0.006	0	3	0.56	0.88
`τ_ℓ 0.2, by 10`	8/8	0.52	−0.003	−0.012 to 0.005	0	3	0.56	0.87
`τ_ℓ 0.1, by 40`	8/8	0.54	−0.003	−0.017 to 0.008	0	3	0.56	0.86
`τ_ℓ 0.2, by 20`	8/8	0.48	−0.002	−0.012 to 0.033	1	3	0.57	0.88
`τ_ℓ 0.2, by 40`	8/8	0.51	−0.002	−0.013 to 0.004	0	3	0.54	0.86
`fixed τ, λ_a delay`	8/8	0.52	0.004	−0.014 to 0.035	1	3	0.56	0.85
`τ_ℓ 0.4, τ_ℓ fixed`	8/8	0.48	0.006	−0.010 to 0.033	2	3	0.56	0.89
`τ_ℓ 0.4, by 10`	8/8	0.49	0.009	−0.009 to 0.034	2	3	0.55	0.85
`λ_a delay`	8/8	0.51	0.009	−0.010 to 0.052	2	3	0.55	0.86
`τ_ℓ 0.4, by 20`	8/8	0.47	0.010	−0.010 to 0.041	2	3	0.54	0.85
`τ_ℓ 0.2, τ_ℓ fixed`	8/8	0.53	0.013	−0.013 to 0.126	2	3	0.56	0.88
`τ_ℓ 0.4, by 40` (base)	8/8	0.45	0.014	−0.013 to 0.054	3	3	0.55	0.85
`λ_a stretch`	8/8	0.36	0.019	−0.007 to 0.082	2	3	0.53	0.83
`fixed τ`	8/8	0.49	0.020	−0.014 to 0.123	3	3	0.53	0.82
`fixed τ, λ_a stretch`	8/8	0.49	0.021	−0.011 to 0.177	1	3	0.55	0.84

The conditions at epoch 64, ranked. *Ranked by whether any run was not called kept, then by the mean spill. Kept: the runs the loss-rate call keeps (α at the example answers at slice 3, averaged over epochs 44 to 64, at least 0.28). Lowest call: the lowest of those averages. Spill: the largest drop on another op under the full edit at epoch 64, less the median drop under the reference edits, as a seed mean, with its seed range and the number of runs over the criterion (0.02). Peak slice and α at peak: where the seed-mean α at the example answers is highest at epoch 64, and its value. α, query: the seed-mean α at the query answer at slice 3.*

The λ_a conditions all stay well above the τ_ℓ 0.1 level. One of them is not on an equal footing: under the stretched warm-up, λ_a is still at 94% of its peak at the stop, so its spill at epoch 64 is of a pull not yet at full strength. A twenty-epoch delay, which gives e1 time to clear under the full learning rate, lowered the mean spill somewhat, at the soft temperatures and at the fixed ones. The stretched warm-up raised it, and so did fixing both temperatures with the base warm-up. Each of these means rests on one or two seeds that spill far more than the rest, so at eight seeds the differences among them could be chance. None of them comes near what sharpening the slice pool does.

Of the 21 runs over the criterion, 15 spill most on `darken`, the op on which the base spills early in spill-timing.

These are early numbers. A sharper slice pool could still spill late, on `value-hsv` as the base does on its slow seeds, and only a finished run can show that. What we take from the screen is that τ_ℓ is the lever on the early kind, and that the level at 0.1 brings the pool to the quiet recipes on it.

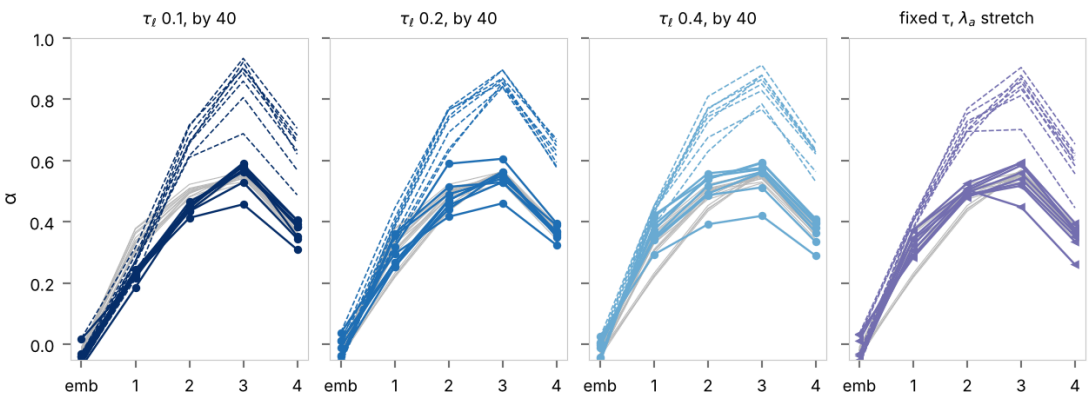
Where the evidence is heading (E3)

A sharper slice pool could lower the spill° in a way we would not want: by settling the pull on one slice and one role, as the joint pool did in slice-terms, where one seed latched onto the query answer and the op could not be removed. Where does the evidence for the op sit at epoch 64?

Spill. The side effect of an edit on the other ops…

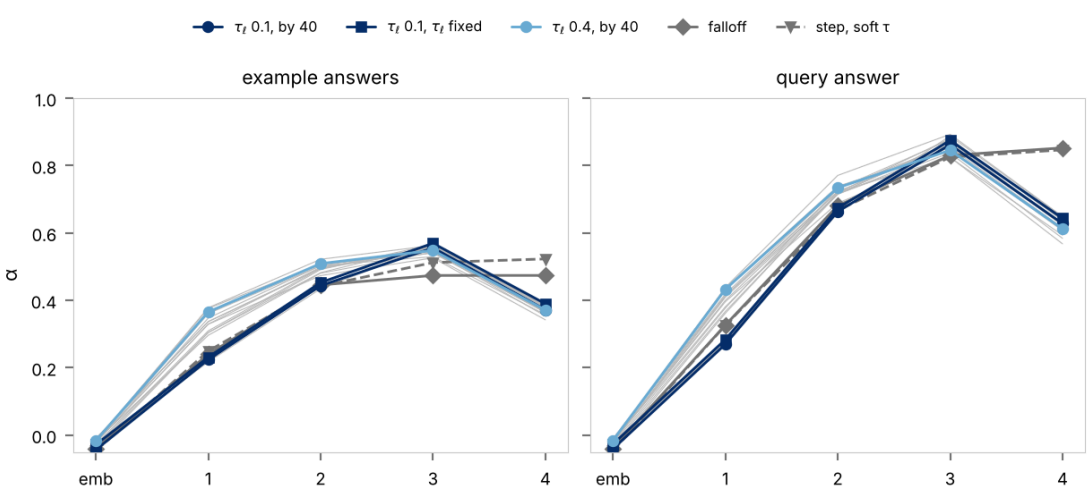
In the same place in every condition. The seed-mean α° at the example answers peaks at slice 3 in all seventeen conditions, and α at the query answer is about the same everywhere (the table above). The figure shows the three levels of τ_ℓ at the base timing, beside the condition that spilled most: the shapes are nearly the same, with one difference at slice 1.

Alignment. How far a state points along e₁…



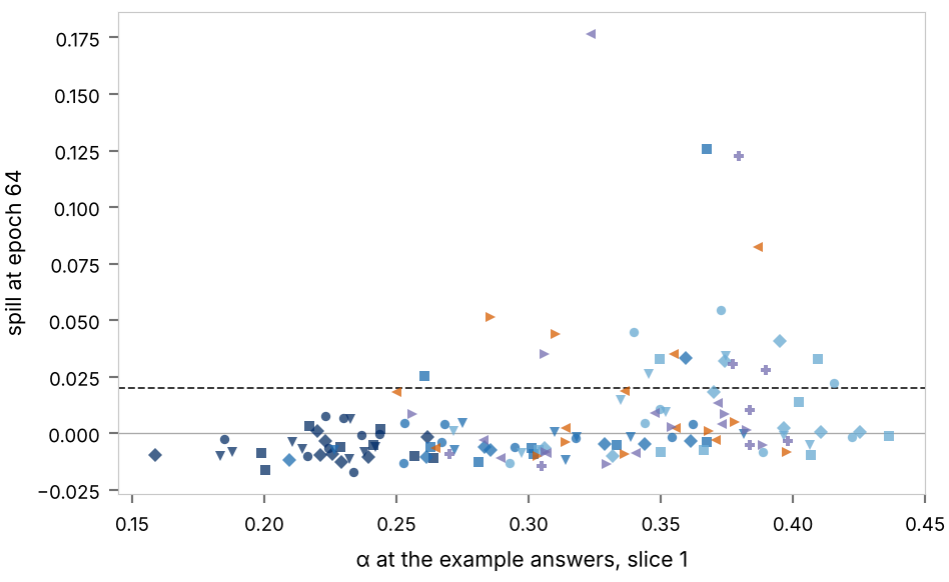
Where the op sits at epoch 64, by slice. *α at the example answers (solid, one line per seed) and at the query answer (dashed), at each slice, for the three slice temperatures at the base timing and for the condition with the highest mean spill. The hairlines behind every panel are the seed means of every condition.*

Beside the quiet recipes (below), the two conditions at τ_ℓ 0.1 have about as much of the evidence at slice 1 as the step does, though the step pulls nothing there: some alignment reaches slice 1 from the pull on the slices after it. The base has half as much again. Past slice 1 the shapes differ as recipe-compare found at the end of training: the step holds the evidence level from slice 3 to slice 4, where the pool lets it fall back, at the example answers and at the query answer alike.



The pool beside the quiet recipes at epoch 64. *Seed-mean α by slice at the example answers (left) and at the query answer (right): the two stage-2 conditions and the base in their ink and marker, and the loss-rate screen's* `falloff` *and* `step, soft τ` *at the same epoch (34 seeds each) in gray. The hairlines are the other conditions.*

That difference is the one that tracks the spill. At τ_ℓ 0.1, α at the example answers at slice 1 averages 0.22, against 0.37 at 0.4. Across all 136 runs, the runs over the criterion are the ones with more of the anchor at slice 1, though at the same α some spill and some do not.



Spill beside the anchor at slice 1. *One mark per run at epoch 64: α at the example answers at slice 1 (horizontal), and the spill (vertical), in the ink and marker of its condition. The dashed line is the criterion.*

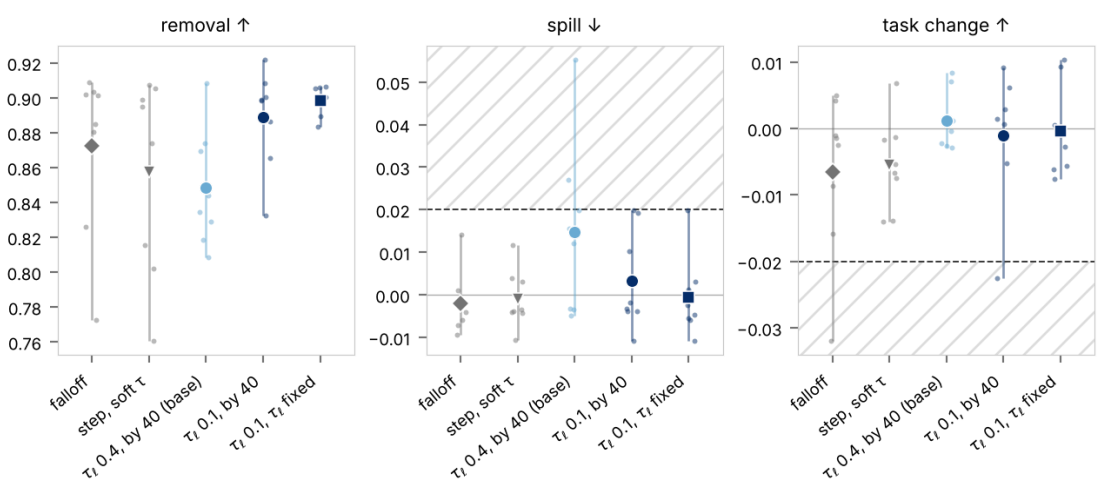
This fits the account recipe-compare gave: the pool spills when it puts part of the pull on block 1, and the edit then takes something from the other ops there. A sharper slice pool takes more of the pull off slice 1 and puts it on slice 3, so block 1 holds less of the anchor and the edit has less to take. It fits without proving it; the screen shows the two moving together, and a restricted edit on the finished runs (block 1 alone) would test it.

Removal, spill, and task at epoch 400 (E4)

Do the two finished conditions keep the anchor, stay within the spill° criterion, and leave the task alone at the end of training, seed by seed, against the base and the two quiet recipes?

Spill. The side effect of an edit on the other ops…

They do. No finished run lost the anchor: the lowest removal is 0.83. Both conditions remove a little more than the base on most seeds (the paired column of the table), and a little more than `falloff` on the mean. Holding τ_ℓ fixed gives the tightest removal of the five (the figure, left), with all eight seeds between 0.88 and 0.91.



Removal, spill, and task at epoch 400, by condition. *One column per condition, the 8 seeds as small dots over a thin bar for their range, and the seed mean as the large mark.* ***Left:*** *removal at the full dose, net of the median over the other axes of the run.* ***Middle:*** *spill, the largest net drop on another op at any dose; the dashed line is the criterion, with the side over it hatched.* ***Right:*** *the task change, the mean clean score over ops less that of the control at the seed; the dashed line is the task gate on that mean, with the failing side hatched.*

condition	removal ↑	removal less base, paired	spill ↓	spill less base, paired
`falloff`	0.87 (0.77 to 0.91)	–	−0.002 (−0.009 to 0.014)	–
`step, soft τ`	0.86 (0.76 to 0.91)	–	−0.001 (−0.011 to 0.012)	–
`τ_ℓ 0.4, by 40` (base)	0.85 (0.81 to 0.91)	–	0.015 (−0.005 to 0.055)	–
`τ_ℓ 0.1, by 40`	0.89 (0.83 to 0.92)	0.04 (−0.01 to 0.09); 6 up, 2 down	0.003 (−0.011 to 0.020)	−0.012 (−0.059 to 0.014); 3 up, 5 down
`τ_ℓ 0.1, τ_ℓ fixed`	0.90 (0.88 to 0.91)	0.05 (−0.01 to 0.08); 7 up, 1 down	−0.001 (−0.011 to 0.020)	−0.015 (−0.061 to 0.006); 3 up, 5 down

Removal and spill at epoch 400*, seed mean and range over 8 runs, net of the median drop over the other 63 coordinate axes of the run. The paired columns are the condition less the base at the same seed, with the count of seeds on each side of zero. The rows for* `falloff`*,* `step, soft τ` *and the base are those of recipe-compare.*

condition	over criterion ↓	over, control median ↓	task change ↑	worst op change ↑
`falloff`	0 of 8	0 of 8	−0.007 (−0.032 to 0.005)	−0.066 (`hue-hsv`, 708)
`step, soft τ`	0 of 8	0 of 8	−0.005 (−0.014 to 0.007)	−0.036 (`hue-hsv`, 705)
`τ_ℓ 0.4, by 40` (base)	2 of 8	3 of 8	0.001 (−0.003 to 0.008)	−0.015 (`hue-hsv`, 708)
`τ_ℓ 0.1, by 40`	0 of 8	1 of 8	−0.001 (−0.023 to 0.009)	−0.037 (`darken`, 708)
`τ_ℓ 0.1, τ_ℓ fixed`	0 of 8	0 of 8	−0.000 (−0.008 to 0.010)	−0.023 (`hue-hsv`, 709)

Runs over the criterion, and the task. *"Over, control median" counts the runs over the criterion with the median taken over the other axes of the control at the seed instead of those of the run. The task change is the mean over ops of the clean score less that of the control at the seed, seed mean and range; the worst op change is the largest loss on one op at any seed.*

On spill (middle), neither condition has a run over the criterion, where the base has 2. The change shows best on the two seeds the base spills on most (the table by seed, below): at 711 both conditions sit below zero, and at 707 both come down to the line.

That leaves one or two seeds per condition at the criterion, each within a thousandth of it: 705 and 707 under `τ_ℓ 0.1, by 40`, and 707 under `τ_ℓ 0.1, τ_ℓ fixed`. Taking the median over the other axes of the control at the seed, in place of those of the run, puts one run of `τ_ℓ 0.1, by 40` over the line (705), and none of `τ_ℓ 0.1, τ_ℓ fixed`. With one or two runs in eight this close to the line, a count at forty seeds would likely put a few over it.

On the task (right), both pass the gate on the seed mean, and no op at any seed goes past the per-op line, so the task change stays near zero, as recipe-compare found for the pool. One run of `τ_ℓ 0.1, by 40` (708) is a little under the line for the mean over ops, at the seed where `falloff` is lowest too.

condition	measure	704	705	706	707	708	709	710	711
`τ_ℓ 0.4, by 40`	removal	0.83	0.91	0.82	0.87	0.83	0.87	0.84	0.81
	spill	−0.003	0.020	−0.005	**0.027**	0.012	−0.003	0.016	**0.055**
	op	`mix`	`value-hsv`	`mix`	`mix`	`value-hsv`	`mix`	`value-hsv`	`darken`
`τ_ℓ 0.1, by 40`	removal	0.92	0.90	0.87	0.89	0.83	0.91	0.90	0.90
	spill	0.010	0.020	−0.004	0.019	−0.002	−0.003	−0.011	−0.004
	op	`mix`	`hue-hsv`	`mix`	`darken`	`mix`	`mix`	`mix`	`mix`
`τ_ℓ 0.1, τ_ℓ fixed`	removal	0.91	0.90	0.89	0.90	0.91	0.91	0.90	0.88
	spill	−0.002	−0.005	0.001	0.020	−0.006	0.003	−0.011	−0.006
	op	`mix`	`mix`	`mix`	`mix`	`mix`	`mix`	`lighten`	`mix`

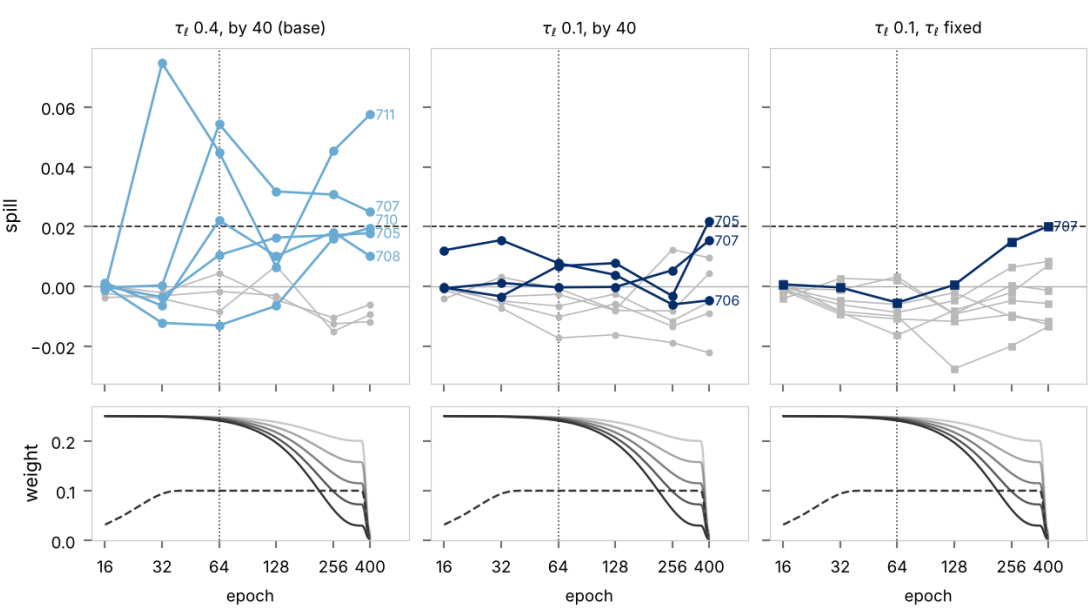
Removal and spill by seed*, as in the table above, with the op the spill is largest on. Bold marks a run over the criterion.*

Spill along training (E5)

Spill. The side effect of an edit on the other ops…

Spill-timing found two kinds of spill° under the base: an early kind, which rises while the pool sharpens and already shows at epoch 64, and a late kind, which grows after epoch 100 on seeds that learn `value-hsv` late. Which kind is left on the seeds that come near the criterion at epoch 400?

The late kind. Under the base, three seeds go over the criterion by epoch 64, and 711 does so already at epoch 32. Under the two finished conditions no seed goes over it by then, and up to epoch 128 the largest spill is 0.015 under `τ_ℓ 0.1, by 40` and 0.003 under `τ_ℓ 0.1, τ_ℓ fixed`. After that, two seeds under `τ_ℓ 0.1, by 40` (705 and 707) and one under `τ_ℓ 0.1, τ_ℓ fixed` (707) climb to about the criterion by the end. That rise comes while the anti weight $\lambda_{\bar{s}}$ anneals on the later slices (underneath), as it does on the late seeds of the base.



Spill along training. *The spill of each run (one line per seed) at the checkpoints kept at epochs 16, 32, 64, 128, 256 and at the end, scored as at the stop, against the median of eight reference edits rather than the 63 axes of E4, so a run at the line there can fall on either side of it here. The base is spill-timing's rerun of soft-pool, scored at every record. The dashed line is the criterion and the dotted line the stop of stage 1. The seeds named are the ones that go over the criterion or come within a quarter of it.* ***Underneath:*** *the anti weight $\lambda_{\bar{s}}$ on each slice (darker for later slices) and the anchor weight λ_a (dashed), the same in every condition after the warm-up.*

So the screen at epoch 64 called the early kind correctly: the conditions it found quiet stayed quiet through the part of training it covers.

The late kind here is about as large as on the late seeds of the base (705 and 710 come to about 0.02 there too). It lands on different ops from those spill-timing saw, though: on `hue-hsv` and `darken` at 705 and 707 under `τ_ℓ 0.1, by 40`, and on `mix` at 707 under `τ_ℓ 0.1, τ_ℓ fixed`, where the late kind in the base was on `value-hsv`. Two or three runs are too few to say whether that is a different kind of late spill or the same one landing elsewhere.

Where the spill enters (E6)

Recipe-compare traced the spill° of the base condition to the edit at block 1, and E3 found that a sharper slice pool puts less of the anchor at slice 1. Does the edit at block 1 alone now cost the other ops less, and does the rest of what the edit does stay as it was?

Spill. The side effect of an edit on the other ops…

condition	peak α, seeds	removal, block 1 alone ↑	removal, blocks 2 to 4 ↑
`falloff`	slice 4: 8	0.72 (0.37 to 0.96)	0.82 (0.10 to 0.94)
`step, soft τ`	slice 4: 8	0.70 (0.22 to 0.95)	0.78 (0.05 to 0.94)
`τ_ℓ 0.4, by 40` (base)	slice 3: 8	0.78 (0.68 to 0.94)	0.81 (0.03 to 0.93)
`τ_ℓ 0.1, by 40`	slice 3: 8	0.69 (0.23 to 0.96)	0.83 (0.09 to 0.95)
`τ_ℓ 0.1, τ_ℓ fixed`	slice 3: 8	0.77 (0.55 to 0.96)	0.83 (0.16 to 0.94)

Where the evidence sits, and where the edit finds it. *The slice at which α at the example answers of* `difference` *contexts peaks, with the count of seeds. Removal under the full edit restricted to a set of slices, at every position, net of the median drop under the same edit along the other coordinate axes of the run; seed mean and range. The first three rows re-score the runs of recipe-compare on this baseline.*

condition	embedding alone ↓	block 1 alone ↓	blocks 1 to 4 ↓	every slice ↓
`falloff`	−0.003 (−0.005 to −0.003), 0	0.003 (−0.009 to 0.029), 1	−0.003 (−0.011 to 0.015), 0	−0.005 (−0.013 to 0.014), 0
`step, soft τ`	−0.002 (−0.005 to 0.005), 0	0.007 (−0.009 to 0.023), 1	−0.004 (−0.013 to 0.011), 0	−0.004 (−0.013 to 0.012), 0
`τ_ℓ 0.4, by 40` (base)	−0.000 (−0.005 to 0.006), 0	0.024 (−0.005 to 0.073), 4	0.013 (−0.008 to 0.057), 3	0.013 (−0.010 to 0.055), 2
`τ_ℓ 0.1, by 40`	−0.002 (−0.003 to 0.001), 0	0.013 (−0.007 to 0.040), 2	0.003 (−0.018 to 0.026), 2	0.001 (−0.016 to 0.020), 0
`τ_ℓ 0.1, τ_ℓ fixed`	−0.003 (−0.009 to −0.001), 0	0.006 (−0.006 to 0.025), 1	−0.003 (−0.020 to 0.021), 1	−0.005 (−0.027 to 0.020), 0

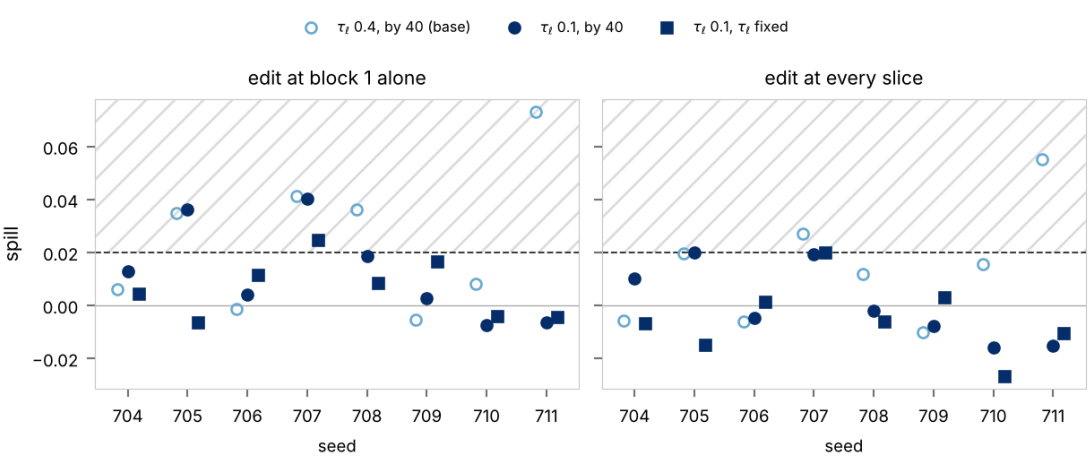
Spill under the restricted edits*, net of the same median: seed mean and range, then the count of runs over the criterion. The edit at every slice is the full edit of E4. The first three rows re-score the runs of recipe-compare on this baseline.*

Alignment. How far a state points along e_1…

The evidence settles where it did. Under both conditions α° at the example answers peaks at slice 3 on every seed, as under the base. The edit at blocks 2 to 4 removes most of the op on every seed but 710, which falls short under every recipe. The edit at block 4 alone removes nothing. The edit at the embedding alone spills nothing, and neither does the edit at blocks 2 to 4 (this one and the edit at block 4 are not in the table).

The edit at block 1 alone costs the other ops less than under the base on the mean, and about what it costs them under `step, soft τ`: the mean is between the step and the base under `τ_ℓ 0.1, by 40` and below the step under `τ_ℓ 0.1, τ_ℓ fixed`.

Seed by seed the two conditions differ. `τ_ℓ 0.1, τ_ℓ fixed` is below the base on six seeds of eight and goes over the criterion on one (707), as each step recipe does on one. `τ_ℓ 0.1, by 40` is below the base on four and goes over on two (705 and 707), and most of its lower mean comes from 711, the seed the base spills on most at block 1; on the other seven its mean is near that of the base. So holding τ_ℓ fixed closes block 1 about as far as leaving it out of the pull does, and sharpening to 0.1 by epoch 40 closes it part of the way.



Spill under the edit at block 1 alone, and at every slice, seed by seed. *The largest drop on another op under the full edit made at those slices only, net of the median over the other axes of the run, for the base (hollow) and the two finished conditions, offset a little at each seed. The dashed line is the criterion, with the side over it hatched.*

As recipe-compare noted, the block-1 edit on its own overstates what the full edit costs: every run it puts over the criterion under the two finished conditions stays within it under the full edit (right), though 705 and 707 come close.

The embedding table moves with the slice pool as well. Under the base the e_1 components of the color embeddings follow lightness closely on every seed (the lean of embedding-lean). Under both finished conditions the lean (last column of the table below) is back at about the level of the step, and varies from seed to seed as it does there, which fits a pull that reaches slice 1 less.

condition	largest write / bound ↓	sites over 1.25× ↓	bound at answers, emb, 1 ↓	share removed ↑	landing ↑	lean on e_1
`falloff`	1.08 (slice 1, 0.3°)	28 (≤ 4°)	5°, 32°	0.85 (0.79 to 0.90)	0.52 (0.49 to 0.56)	−0.41 (−0.64 to −0.20)
`step, soft τ`	1.15 (slice 1, 3.6°)	23 (≤ 6°)	6°, 32°	0.83 (0.77 to 0.90)	0.49 (0.47 to 0.53)	−0.38 (−0.63 to −0.04)
`τ_ℓ 0.4, by 40` (base)	1.33 (slice 1, 4.9°)	48 (≤ 5°)	14°, 43°	0.83 (0.79 to 0.88)	0.51 (0.47 to 0.55)	−0.61 (−0.65 to −0.57)
`τ_ℓ 0.1, by 40`	1.03 (slice 1, 4.9°)	13 (≤ 5°)	4°, 27°	0.86 (0.79 to 0.91)	0.50 (0.45 to 0.55)	−0.36 (−0.53 to −0.05)
`τ_ℓ 0.1, τ_ℓ fixed`	1.06 (slice 1, 4.5°)	22 (≤ 4°)	5°, 29°	0.87 (0.78 to 0.91)	0.52 (0.47 to 0.54)	−0.41 (−0.59 to −0.15)
control						−0.02 (−0.33 to 0.62)

The write, the grading and the landing. *The largest ratio of write to bound in angle over the slices past the embedding and every position, from the seed-mean quantiles as in edit-effects, with where it falls and the bound there. The sites over 1.25× count slice and position pairs over all 8 runs, each run taken on its own, with the largest bound among them. The bound at the example answers is over those that do not fit* `difference`*. The share removed is over every* `difference` *context, net of the control at the seed, and the landing is the share of the way to the target null in total variation. The lean is the correlation, over the color embeddings, of the e_1 component with lightness (the lean of embedding-lean). Seed mean and range; the first three rows are those of recipe-compare.*

What the edit does downstream barely changes, and where it changes it moves toward the step. The write stays close to the bound past the embedding, with fewer sites over the tolerance than under the base, and the share removed and the landing are about the same as under every recipe.

Recipe-compare found the bound at the example answers of the other ops about 3× as wide at the embedding under the base as under `falloff`. Under both finished conditions it is back at the width of `falloff` at the embedding, and a little narrower at slice 1. So the other ops sit as far from e_1 at the first two slices as under the step.

Discussion

With τ_{ℓ} at 0.1, the pool over slices 1 to 3 keeps the anchor, removes a little more than the recipe of record, costs the task nothing, and stays within the spill° criterion on every seed. The early spill that put two runs of the base over the criterion is gone.

Spill. The side effect of an edit on the other ops…

So are most of the signs that block 1 is open to the edit. The wider bound at the first two slices and the lean in the embedding table are gone under both conditions, and the larger cost of the block-1 edit is gone under the condition with τ_{ℓ} fixed.

Of the two, holding τ_{ℓ} at 0.1 from the start is the cleaner. It removes the most and the most evenly, has no run over the criterion under either median, and one under the edit at block 1 alone, as each step recipe has. It also has one schedule fewer than the base, since only τ_{p} sharpens.

The other condition is close behind on the full edit, where the gap rests on two or three runs near the line. On the edit at block 1 alone the gap is larger: there the other condition sits near the base on most seeds, and only its mean is lower.

What is left is the late kind of spill, which brings one or two seeds per condition to about the criterion by the end. It is about as large as under the base on its late seeds, so it does not seem to come from the sharper slice pool, and nothing in this study acts on it. When it appears, the term still changing is the anti weight, partway through its anneal, so if anything in the recipe acts on the late kind, that anneal is the likelier place.

Before the pool with τ_{ℓ} fixed at 0.1 could replace `falloff` as the recipe of record, it would need two things the step has: a count of lost anchors at forty seeds, which a loss-rate screen would give, and a set of fresh seeds, which a confirmation at seeds 712 to 719 with their controls would give. Since the pool chooses the slice within a range set by the architecture, it would also be the easier recipe to carry to another model, as recipe-compare argued.