A soft pool while the anchor settles

> We tried the two changes the anchor-ramp losses suggested, crossed: a pooling temperature° that starts soft and sharpens over the pull warm-up, and a pull on slices 1 to 3 only, with nothing on slice 4. Every arm with the soft start kept the anchor on every seed, but the edit spilled more under each of them than at the fixed temperature. The ramp over slices 1 to 3 at the fixed temperature lost the anchor on one seed, in a new way.

Pooling temperature. How sharply the anchor pull picks its positions....

In the anchor-ramp experiment, a pull that ramps with depth lost the anchor on two or three of eight seeds, where the step in `falloff` lost it on none. The lost runs split off from the kept ones during the pull warm-up, around epochs 16 to 24: on a kept run the example answers rise onto e_1 along with the query answer, and on a lost run they fall back. The losses also followed how little pull slices 1 to 3 had, though that ordering was picked after the counts were in.

The pull is pooled over the positions of each context at a sharp temperature, τ_p = 0.1,[1] so while the query answer is the position closest to e_1, it gets almost all of the pull. A soft pool early on would share the pull with the example answers while they are still deciding.

1 Pooling temperature...

A pull on slices 1 to 3 alone, with nothing on slice 4, would put the pull where it mostly moves the example answers.

The 4 new arms, of 8 runs each, cross the two changes and add a pool over slices 1 to 3 with both temperatures soft. They use the seeds of the split-anti-seeds experiment, so its `falloff` runs serve as the baseline with no new training. Those eight runs are the only ones `falloff` has, so its loss rate is known only roughly (see Scope).

Observations

- Removal and spill (E1)°: the soft start kept the anchor on every seed in all three arms that had it, and the ramp at the fixed temperature lost it on one. Removal on the kept runs was close to that of `falloff` everywhere. But each soft arm had runs over the spill criterion, and the pool over slices 1 to 3 had the most.

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

- When the runs settle (E2): a soft pool holds every position back through the first part of the warm-up, then the example answers come onto e_1 with the query answer on every seed, about four to twelve epochs later than under `falloff`. The one lost run latched onto the query answer for a time: its example answers fell back after rising partway, climbed back only slowly from about epoch 100, and ended below every kept run, with the edit removing almost nothing.
- Where the anchor sits by depth (E3): without a pull on slice 4, the alignment° there falls back from its slice-3 level, and the pooled arm settles where the ramp does. The syntax embeddings stay off e_1 in every arm. The soft pool briefly lifts every position toward e_1 during the rise, the other ops included, and that lift goes with the spill.

  Alignment. How far a state points along e_1…

Scope

This is an exploratory study, with no preregistration and no gate. Two factors are crossed, with `falloff` as the condition where both are at their recipe level:

- The pull shape: the step on slices 2 to 4 (the recipe), or a ramp over slices 1 to 3 at a third, two thirds and the full weight, with nothing on the embedding or slice 4. The ramp adds up to the step over slices 1 to 3 (two each), so the earlier slices keep the pull the step gives them, and the pull on slice 4 goes.
- The position pool temperature τ_p: fixed at 0.1 (the recipe), or soft-to-sharp, starting at 1 and sharpening to 0.1 by epoch 40, the end of the pull warm-up. The schedule is smooth in log τ_p, so the temperature halves at a steady pace and comes to rest at 0.1. At epoch 20, when the anchor-ramp runs split, it is about halfway there in log terms: a query answer well ahead of an example answer takes a few times as much pull as the example answer, against about fifty times at 0.1.

arm	emb	1	2	3	4	τ_p	τ_ℓ
`falloff`	0	0	1	1	1	0.1	–
`step, soft τ`	0	0	1	1	1	1 → 0.1	–
`ramp 1–3`	0	1/3	2/3	1	0	0.1	–
`ramp 1–3, soft τ`	0	1/3	2/3	1	0	1 → 0.1	–
`pool 1–3, soft τ`	0	1	1	1	0	1 → 0.1	4 → 0.4
every arm, anti hold	0.20	0.16	0.12	0.07	0.03		

The arms. *The pull weight on each slice, relative to the anchor weight λ_a on each slice of the step, and the two pooling temperatures. An arrow is a temperature that sharpens from its first value at epoch 0 to its second by the end of the pull warm-up (epoch 40). In `pool 1–3, soft τ` the weights only say which slices are in the pool. The anti hold is the weight the anti term anneals to, the same `falloff` line in every arm, rounded here to two places.*

The fourth arm pools the pull over the slice × position pairs of slices 1 to 3 (the nested pool of the slice-terms experiment), so the pull can choose the slice as well as the position. Its slice pool temperature τ_ℓ starts at 4, near a plain mean over slices, and sharpens to 0.4 over the same window as τ_p.

The end value rests on one seed. In E4 of slice-terms, at seed 701, every τ_ℓ from 0.2 to 1.6 removed about three quarters to nine tenths of the op, with no trend and every run within the criterion, while the joint pool at τ_ℓ = 0.1 lost the anchor on one seed of four. So 0.4 is inside the range that worked there.

A soft pool and a ramp each change more than one thing:

- A softer τ_p spreads the pull of each context over its positions, but the total pull per context stays the same at every temperature (the softmin weights sum to one). In the τ×λ_a sweep, a softer fixed τ_p meant more spill°, so spill is the thing to watch; here it is soft only during the warm-up.
- The ramp pulls two thirds as much in total as the step, all of it moved off slice 4. A difference between the ramp and the step can come from the shape or from the lower total.
- The pool over slices 1 to 3 pulls as much in total as the step, and lets the pull choose where it goes.

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

Every run trains at model seeds 704 to 711 for 400 epochs, paired by seed with its `falloff` twin and with ex-2.2.23's control (same initialization, batches and label draws).

With eight seeds per arm, one or two lost runs could come from the seeds alone, so we weigh the counts across arms together. `falloff` has run only at these eight seeds and lost none, which fits a loss rate anywhere from zero to about 30% (the 95% upper bound for none in eight). Its close relatives in the split-anti-seeds experiment, the uniform anti holds of 0.2 and 0.1, each lost the anchor on one seed of the eight. So one loss in eight is no worse than the recipe as far as eight seeds can tell. Since `falloff` lost no runs here, so the step pair can only show a soft start doing harm, and only the ramp pair can show it helping.

The measurements

The measurements and dose scoring are those of the split-anti-seeds experiment.

- Removal: the net drop in the score on `difference` under the edit, as a share of the way from the clean score to the target null°. The edit projects e_1 out of the state at every position and slice, at a dose of a quarter, a half, three quarters, or all of the e_1 component. Higher is better.

Target null. Where a removal should land…

- Spill°: the largest net drop on another op under the edit. A run is within the criterion when its spill is at most 0.02 at every dose.

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

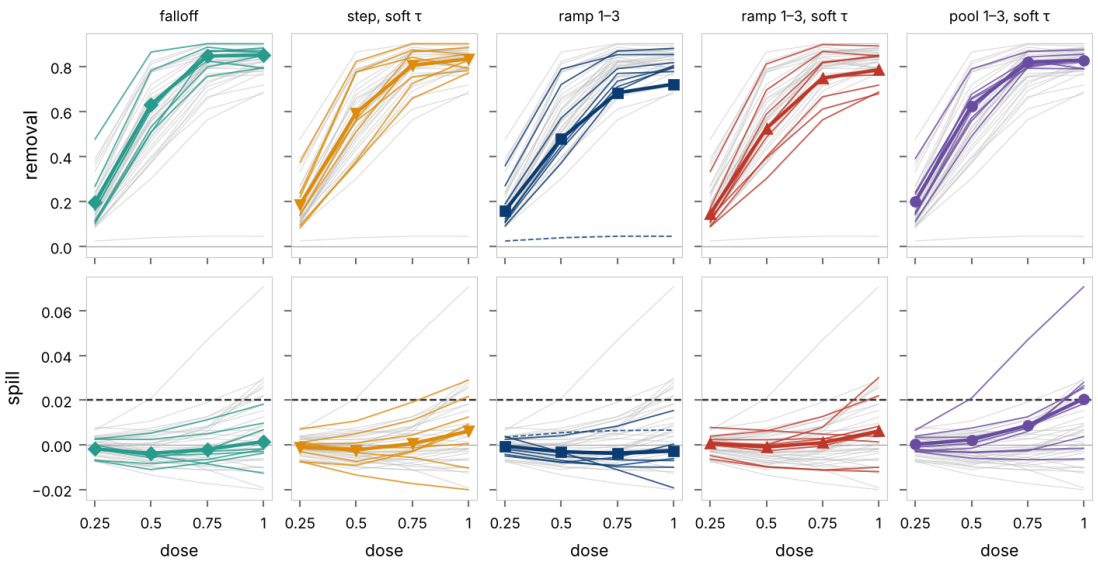
- Both are net of the control at the same seed.
- A run lost the anchor when the full edit removes less than half of the op, as in the split-anti-seeds experiment.
- Task change: the mean score over ops, less that of the control. Zero means no damage.
- The alignment° α with e_1 at the example answers and at the query answer, at every slice, along training and at the end.

Alignment. How far a state points along e_1…

Removal and spill (E1)

Does a soft pool early, or a pull on slices 1 to 3 alone, keep the anchor on every seed, and keep the edit as selective as `falloff`? The main measure is the removal at the full dose, with the spill° at any dose beside it. The soft step against `falloff` isolates the temperature, and the fixed ramp against `falloff` isolates the shape.

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



***Removal and spill by dose**, the edit at every position and slice, one column per arm. One hairline per run, dashed where the run lost the anchor; the seed mean is bold, and the other arms are ghosts behind. The dashed rule is the selectivity criterion.*

Every run but one climbs with the dose to about seven tenths to nine tenths of the op, as `falloff` does. The exception is one run of `ramp 1–3` (the dashed line), which removes almost nothing at any dose.

With the soft start the same ramp kept the anchor on that seed, so in the one pair where the soft start could help, it did. But at a loss rate of one in eight, the soft twin would most often have kept that seed anyway, so this fits the soft start helping without being evidence that it does.

arm	kept the anchor ↑	removal, kept ↑	within criterion ↑	spill, any dose ↓	task change ↑	worst op change ↑
`falloff`	8 of 8	0.85 (0.79 to 0.90)	8 of 8	0.003 (−0.004 to 0.018)	−0.007 (−0.032 to 0.005)	−0.066 (`hue-hsv`, 708)
`step, soft τ`	8 of 8	0.83 (0.77 to 0.90)	6 of 8	0.009 (−0.007 to 0.029)	−0.005 (−0.014 to 0.007)	−0.036 (`hue-hsv`, 705)
`ramp 1–3`	7 of 8	0.82 (0.78 to 0.88)	8 of 8	0.002 (−0.005 to 0.015)	0.001 (−0.008 to 0.012)	−0.016 (`darken`, 709)
`ramp 1–3, soft τ`	8 of 8	0.79 (0.68 to 0.89)	6 of 8	0.008 (−0.006 to 0.030)	−0.001 (−0.022 to 0.007)	−0.046 (`darken`, 708)
`pool 1–3, soft τ`	8 of 8	0.83 (0.79 to 0.88)	4 of 8	0.021 (−0.003 to 0.071)	0.001 (−0.003 to 0.008)	−0.015 (`hue-hsv`, 708)

***Removal, spill and task**, seed mean and range. A run kept the anchor when the full edit removes at least 0.5 of the op, and removal is over those runs, at the full dose. "Spill, any dose" is the largest net spill over the four doses, and "within criterion" counts the runs where it is at most 0.02. The task change is the mean score over ops less that of the control, and the worst op change is the largest loss on any one op at any seed. Arrows mark which direction is better.*

The soft start makes the edit less selective. `falloff` and `ramp 1–3` stayed within the criterion on every seed, and their soft twins went over it on two seeds each, by a little. The pool over slices 1 to 3 went over on half its seeds, one of them (seed 711) by more than three times the criterion. The τ×λ_a sweep found the same for a softer fixed temperature, so this seems to be the price of a soft pool, even one that is soft for the first tenth of training only.

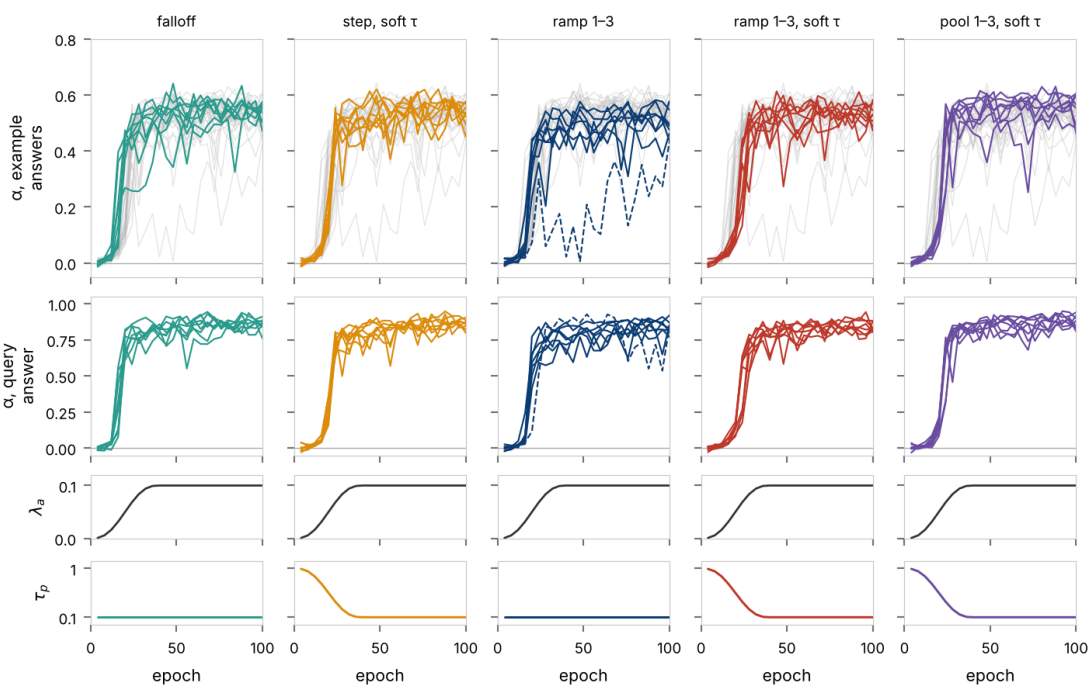
The pool over slices 1 to 3 removes about as much as `falloff`, with a narrow seed range, and more than either ramp at every dose. It pulls as much in total as the step and half again as much as the ramps, and it can put that pull on the slice and position where it moves the concept most, so it acts like the stronger pull of the two shapes on slices 1 to 3. Its spill is the widest of any arm, though.

The ramp also removes a little less. Over the kept runs, both ramps remove a few hundredths less than the step, and `ramp 1–3, soft τ` has three runs near seven tenths. The mean task score is unchanged in every arm, and no new arm loses more on any one op than `falloff` does.

When the runs settle (E2)

Does the soft pool carry the example answers onto e_1 with the query answer through the warm-up, where the lost runs of the anchor-ramp experiment fell back? The main measure is α° at the example answers along training, at slice 3, the deepest slice every arm pulls. The query answer is in the row underneath, and the anchor weight and the temperature in strips below that.

Alignment. How far a state points along e_1...



***Alignment along the first hundred epochs**, at slice 3, on the probe contexts of `difference`, one column per arm: at the example answers (top) and at the query answer (second row). One line per run, dashed where the run lost the anchor; the other arms are ghosts behind the top row. The strips underneath are the anchor weight λ_a and the position pool temperature τ_p (log scale).*

A soft pool slows the start. In the step pair at epoch 16, the query answer under `falloff` is already partway onto e_1 with the example answers following, while under the soft start both are still near zero. By epoch 24 the example answers have come up on every seed of `step, soft τ`, all crossing 0.4 at that epoch.

The ramp starts slowly at slice 3 at either temperature, since it pulls that slice at two thirds of the step weight. With the soft start its runs cross 0.4 within eight epochs of one another; at the fixed temperature one run is left behind.

arm	α query answer, epoch 16	α example answers, epoch 16	first epoch over 0.4	α example answers, epoch 48
`falloff`	0.31 (0.08 to 0.56)	0.19 (0.05 to 0.40)	23 (20 to 40)	0.53 (0.34 to 0.64)
`step, soft τ`	0.06 (0.04 to 0.09)	0.05 (0.03 to 0.07)	24 (24 to 24)	0.54 (0.48 to 0.62)
`ramp 1–3`	0.08 (0.03 to 0.24)	0.06 (0.02 to 0.18)	34 (20 to 100)	0.45 (0.01 to 0.58)
`ramp 1–3, soft τ`	0.05 (0.03 to 0.11)	0.04 (0.03 to 0.07)	26 (24 to 32)	0.51 (0.27 to 0.60)
`pool 1–3, soft τ`	0.04 (0.00 to 0.09)	0.04 (0.01 to 0.07)	25 (24 to 28)	0.55 (0.36 to 0.61)

***How the runs settle.** α at slice 3 on the `difference` probe contexts, seed mean and range over every run: at the query answer and the example answers at epoch 16, the first recorded epoch at which the example answers go over 0.4 (α is recorded every four epochs), and the example answers at epoch 48, just after the warm-up.*

So the soft pool does what it was meant to do, in that no run latched onto the query answer, but it does so by holding every position back. A likely reading is that it spreads the pull so thinly that nothing moves much until the temperature is low enough to choose, and by then the example answers rise in step with the query answer.

On that reading the soft start works mostly as a delay. A warm-up that simply started later might settle the runs the same way, and no arm here tells the two apart. The step pair does show the soft start tightening the spread: under `falloff` one seed (711) crossed 0.4 late, at epoch 40, and under the soft start every seed crossed at epoch 24. Eight seeds per arm fit this reading without proving it.

The lost run of `ramp 1–3` (seed 708) starts as the anchor-ramp losses did and ends differently. Its example answers rose partway with the others, fell back four epochs later, and stayed below 0.4 until about epoch 100, while its query answer stayed on e_1: a partial latch onto the query answer. After that the example answers climbed to about four fifths of the kept level.

arm	seed	removal	spill, any dose	α, example answers	α, query answer
`ramp 1–3`	708	0.05	0.007	0.49	0.93

***The runs that lost the anchor.** Removal at the full dose, the largest net spill at any dose, and α at slice 3 at the end of training on `difference` probe contexts, at the example answers and at the query answer. On the runs that kept the anchor, α at the example answers is 0.62 (0.59 to 0.64) at slice 3.*

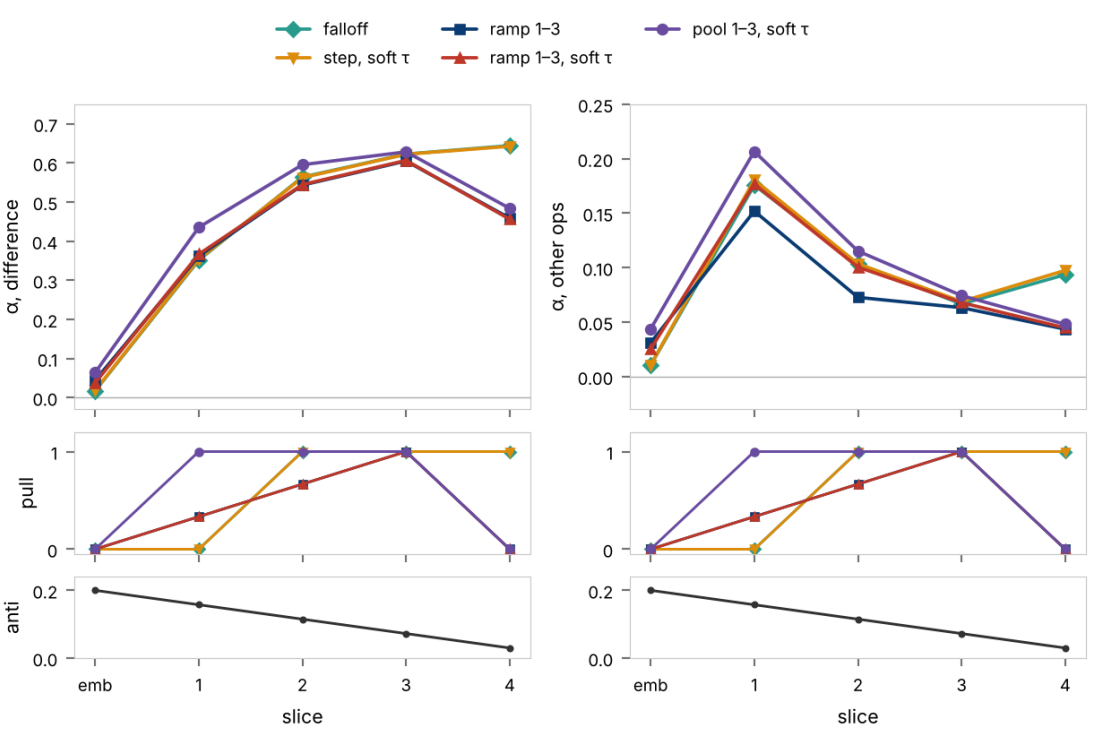
By the end its query answer is near one, as on every run (it is close to the ceiling of α, so it has little room to differ). Its example answers end below every kept run but well above zero, yet the edit removes almost nothing on it, at any slice.

A likely reading is that the model learned to solve `difference` while the anchor was away, along a path that does not go through e_1, and does not use the alignment that came back later. One run can't settle that. But alignment alone would understate this kind of loss: the anchor-ramp losses ended near zero, and this one ends at four fifths of the kept level.

Where the anchor sits by depth (E3)

On the runs that kept the anchor, where does the alignment° sit by slice? With nothing on slice 4, does slice 4 still hold the concept as the residual stream (the running sum each block adds its output to) carries it up? Where does the pool over slices 1 to 3 settle, given that the pool over every block settled on block 4 in the slice-terms experiment? We also check the syntax embeddings, since in the τ×λ_a sweep a soft pool spilled through the `=` embedding coming onto e_1.

Alignment. How far a state points along e_1...



***Alignment at the example answers by slice**, at the end of training, on the probe contexts of `difference` (left) and of the other ops (right), over the runs that kept the anchor: the seed mean in the arm's marker; the two top panels have their own scales. Underneath are the pull weight at each slice, relative to the step (for the pooled arm, the slices in its pool), and the anti hold, which every arm shares.*

Up to slice 3 the arms are nearly alike, and the soft start changes nothing there. At slice 4, where the ramp and the pool put no pull, the example answers fall back to about three quarters of their slice-3 level, and the query answer falls from near one to about four fifths. Under the step they hold their slice-3 level. So the concept rides up past the last pulled slice, though only partly, and the edit still finds it, since it acts on every slice.

The anti term may account for some of the fall at slice 4. It acts on every position, the `difference` answers included, and holds at 0.03 on slice 4 in every arm. Under the step the pull on slice 4 counters it; under the ramp and the pool nothing on slice 4 counters it, so the model gains by moving the concept off e_1 there. No arm here turns the anti term off on slice 4, so this is untested.

The pool over slices 1 to 3 settles much as the ramp does, a little higher at slices 1 and 2. It peaks on slice 3, the last slice in the pool, as the slice-terms pool peaked on its last block. The other ops peak at slice 1 in every arm, as in `falloff`.

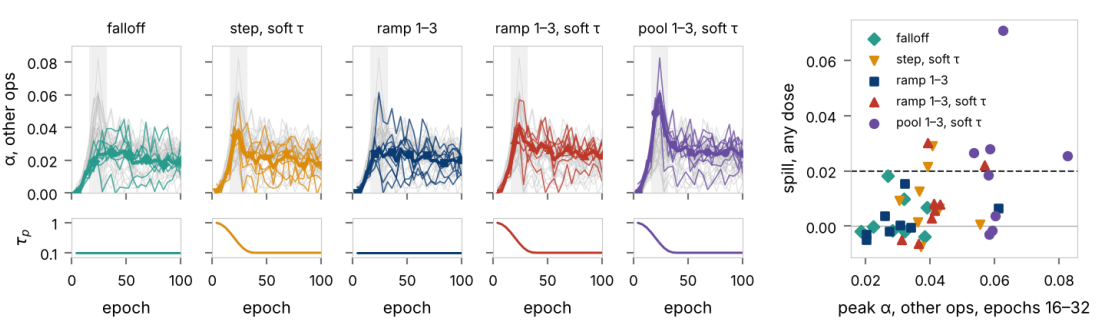
arm	α at 1	α at 2	α at 3	α at 4	query, 3	query, 4	α other ops, 1	α other ops, 2	syntax e_1
`falloff`	0.35	0.56	0.62	0.64	0.98	0.98	0.18	0.10	0.005 (0.001 to 0.009)
`step, soft τ`	0.35	0.56	0.62	0.64	0.97	0.97	0.18	0.10	0.004 (0.002 to 0.005)
`ramp 1–3`	0.36	0.54	0.60	0.46	0.97	0.78	0.15	0.07	0.007 (0.002 to 0.019)
`ramp 1–3, soft τ`	0.37	0.54	0.61	0.46	0.96	0.77	0.18	0.10	0.007 (0.003 to 0.011)
`pool 1–3, soft τ`	0.44	0.60	0.63	0.48	0.97	0.79	0.21	0.12	0.006 (0.001 to 0.011)

***Alignment by slice, and the syntax embeddings.** α at the example answers on `difference` and on the other ops, seed mean over the runs that kept the anchor. "Query" is α at the query answer on `difference`. "Syntax e_1" is the largest e_1 component (in size) among the embeddings of `?`, `=`, `,` and the newline, seed mean and range over every run; an embedding on e_1 would show as a value near one.*

The syntax embeddings stay off e_1 in every arm, within a few hundredths, so the spill° of E1 does not come from the route the τ×λ_a sweep found. The other ops end about as aligned in the step pair, and a little more aligned under the soft ramp than under the fixed one, though not by enough to account for the spill.

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

The trajectories hold a lead. While the pool is soft, its pull reaches every position from the op word on, and during the rise the other ops come partway onto e_1 with it.



***The other ops during the rise.** Left: α on the probe contexts of the other ops along the first hundred epochs, the mean over their positions and slices 1 to 3, one column per arm; one hairline per run, the seed mean bold, and the other arms as ghosts. The shaded band is epochs 16 to 32, and the strip underneath is the position pool temperature τ_p (log scale). Right: the highest α in that band against the spill of E1 at any dose, one point per run; the dashed rule is the selectivity criterion.*

arm	α other ops, epochs 16–32, peak	α other ops, end	spill, any dose
`falloff`	0.030 (0.019 to 0.039)	0.021 (0.013 to 0.027)	0.003 (−0.004 to 0.018)
`step, soft τ`	0.040 (0.030 to 0.055)	0.021 (0.016 to 0.029)	0.009 (−0.007 to 0.029)
`ramp 1–3`	0.031 (0.020 to 0.061)	0.022 (0.015 to 0.038)	0.002 (−0.005 to 0.015)
`ramp 1–3, soft τ`	0.041 (0.031 to 0.057)	0.022 (0.016 to 0.026)	0.008 (−0.006 to 0.030)
`pool 1–3, soft τ`	0.062 (0.054 to 0.082)	0.024 (0.018 to 0.029)	0.021 (−0.003 to 0.071)

***The lift on the other ops, by arm.** α on the probe contexts of the other ops, the mean over their positions and slices 1 to 3: its highest value over epochs 16 to 32, and its value at the end of training, with the spill of E1 beside them; seed mean and range over every run. Over all 40 runs, the correlation of the peak with the spill is r = 0.47.*

Every soft arm lifts the other ops higher during the rise than its fixed twin does, and the pool over slices 1 to 3 lifts them most, in the same order as the spill. By the end of training the lift has gone, and the other ops are about as aligned in every arm. The peak goes with the spill across runs, though loosely, and within an arm it barely separates the runs. So a likely reading is that the model learns something about the other ops while they sit partway on e_1, which the edit still reaches after the alignment has gone. This post-hoc reading would need an arm to test it.

Discussion

The soft start got the example answers onto e1 on every seed it was tried on, apparently by holding the pull back until it can choose. That fits the anchor-ramp reading, that runs are lost in the first tens of epochs when the sharp pool gives the query answer nearly all of the pull. But the evidence for a benefit is one seed, which a one-in-eight loss rate would often keep anyway (E1).

Both soft twins spilled more than their fixed arms, and the pool over slices 1 to 3, which has no fixed twin, spilled most.

The ramp over slices 1 to 3 lost the anchor on one seed of eight at the fixed temperature, where the anchor-ramp ramps from slice 1 lost two or three. That fits the second anchor-ramp reading, that the earlier slices need about as much pull as the step gives them: this ramp gives slices 1 to 3 that much, and drops slice 4 instead.

Its one loss came later and in another way, with the alignment° partly back by the end but no longer doing the work. That is a reason to keep scoring removal, and not alignment alone, when judging a run.

Alignment. How far a state points along e1…

The soft ramp kept the anchor on every seed here without the step between slices 1 and 2, so it would be a candidate if its spill° can be kept down. If the spill comes from the soft pool reaching every position, as E3 suggests, a stronger anti term while the pool is soft would aim at it. Turning the anti term off on slice 4 might also let the ramp keep the concept there. As it stands, `falloff` still keeps the anchor on every seed with the edit inside the criterion, which no new arm matched on both counts.

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

Method

The pooling temperatures[1] are step inputs here, scheduled like the weights (`AnchorSpec.tau_start`, `slice_tau_start` and `tau_anneal_end`); before this experiment they were fixed when the training step was built. The schedule is minimum-jerk in log τ from the start value at epoch 0 to the end value at epoch 40, and holds after. Training otherwise follows the split-anti-seeds experiment unchanged. The trajectory of each run also keeps α° by slice and position for each op separately, and the e_1 component of every syntax embedding, for later use. The 32 runs cost about $11.

Pooling temperature. How sharply the anchor pull picks its positions....

Alignment. How far a state points along e_1...

---

1. Pooling temperature: the pull on a context is a softmin-weighted mix over its positions. A low temperature gives nearly all the weight to the closest position; a high one spreads it evenly. ↩