

The split anti weight at fresh seeds

We trained the candidate recipes from [the slice-terms experiment](#) at eight fresh seeds. Holding the anti-subspace weight high on the first slices only kept the op removable and the edit selective, and a weight that falls off linearly with depth did as well. Holding it at one level on every slice, high or gentler, lost the anchor altogether on one seed in each arm.

We chose `falloff` as the recipe of record. By the frozen gate on damage to the task only `split-anti` qualified, and `falloff` missed that gate on one op at one seed.

Findings

- [The split keeps both removal and selectivity \(H1\)](#) – **pass**. `split-anti` removed more than `deep`, `hold 0.2` on 6 of 8 seeds and stayed within the criterion on 7, with the whole gain on the four seeds where the uniform hold anchored less well or not at all.
- [A linear falloff does as well as the split \(H2\)](#) – **pass**. `falloff` removed about as much as `split-anti` and stayed within the criterion on every seed.
- [Which slices the edit needs \(E1\)](#) – the edit on slices 2 to 4 alone falls short on the same seed in every arm, which fits the seed rather than the terms deciding where it misses. The edit on slice 1 alone does depend on the arm: it removes less under the high uniform hold.
- [Decision](#) – `falloff` is the recipe of record. Only `split-anti` cleared the frozen task gate; `falloff` missed it on one op at one seed, and we kept it for its spill and its shape.

How to read this report

This report was preregistered: the predictions and the criteria for the decision were frozen at commit `ff44a17`, before any run of this experiment. Each section opens with what we expect, and the results replaced the placeholders in place. The subsection on runs that lost the anchor, in H1, was not planned, and the decision departs from the frozen task gate.

Why this experiment

The pull indirectly asks the states at the example answers of `difference` contexts to align with e_1 , and the anti-subspace term asks every other state to stay off it. When the anti weight anneals low, the color embedding table and slice 1 drift onto e_1 on the other ops, and the edit that projects e_1 out takes some of those ops with it: the spill. [The late-pull experiment](#) found that a higher hold everywhere keeps the spill within the criterion, at some cost in removal. [The slice-terms experiment](#) traced that cost to the hold countering the pull at the pulled slices, and held the weight high on the embedding and slice 1 only. On its four seeds this gave the removal back and kept the spill within the criterion on three.

Those four seeds are the ones the arms were chosen on, so the result may be partly fitted to them. Before the split becomes the recipe a deliverable rests on, it should hold at fresh seeds, beside the arms it was chosen over.

The split also sets a boundary by hand: high on the first two slices, low after. If a weight that falls off smoothly with depth does as well, the recipe has one less choice in it.

Parameters

Every arm pulls `difference` toward e_1 at slices 2 to 4, pooled over the positions of each labeled context, so that most of the pull lands on the example answers. The anti-subspace term acts on every slice (the `deep` pull of the late-pull experiment). The anchor weight λ_a (0.1), the pooling temperature τ_p (0.1), the label, the corpus and the schedules are the recipe values, at 400 epochs. The anti weight $\lambda_{\bar{s}}$ starts at 0.25 on every slice and anneals to its *hold* by epoch 360, as in the slice-terms experiment. The arms differ only in that hold:

arm	emb	1	2	3	4
deep, hold 0.2	0.2	0.2	0.2	0.2	0.2
deep, hold 0.1	0.1	0.1	0.1	0.1	0.1
split-anti	0.2	0.2	0.03	0.03	0.03
falloff	0.2	0.16	0.12	0.07	0.03

The hold of the anti weight by slice, for each arm. Slice 0 is the embedding. The `falloff` holds are a straight line from 0.2 to 0.03, rounded here to two places.

`falloff` runs in a straight line between the two holds of `split-anti`, so it matches the split at the embedding and slice 4 and differs between them. `deep, hold 0.1` is the gentler uniform hold, a candidate for the recipe in its own right.

Each arm trains at model seeds 704 to 711, 32 runs in all. Ex-2.2.23 trained 400-epoch controls (no pull, no anti term) at these seeds, so every run has a control with the same initialization, batches and label draws. No arm has been trained at these seeds before.

`deep` at the recipe hold (0.03 everywhere) is left out, since it spilled over the criterion on every seed so far.

Measurements

The measurements and dose scoring are those of the slice-terms 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, on a set of slices, at a dose of a quarter, a half, three quarters, or all of the e_1 component. Higher is better.
- 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.
- Both are net of the control at the same seed.
- Landing: the share of the distance to the target null that the full edit covers, as total variation on `difference` contexts. Higher is better.
- Task change: the mean score over ops, less that of the control. Zero means no damage; below zero is a loss.
- The alignment α with e_1 at the example answers, at every slice, at the end of training, and the slope of the color embedding table (the e_1 component of the color embeddings against lightness).

The split keeps both removal and selectivity (H1)

✓ Pass

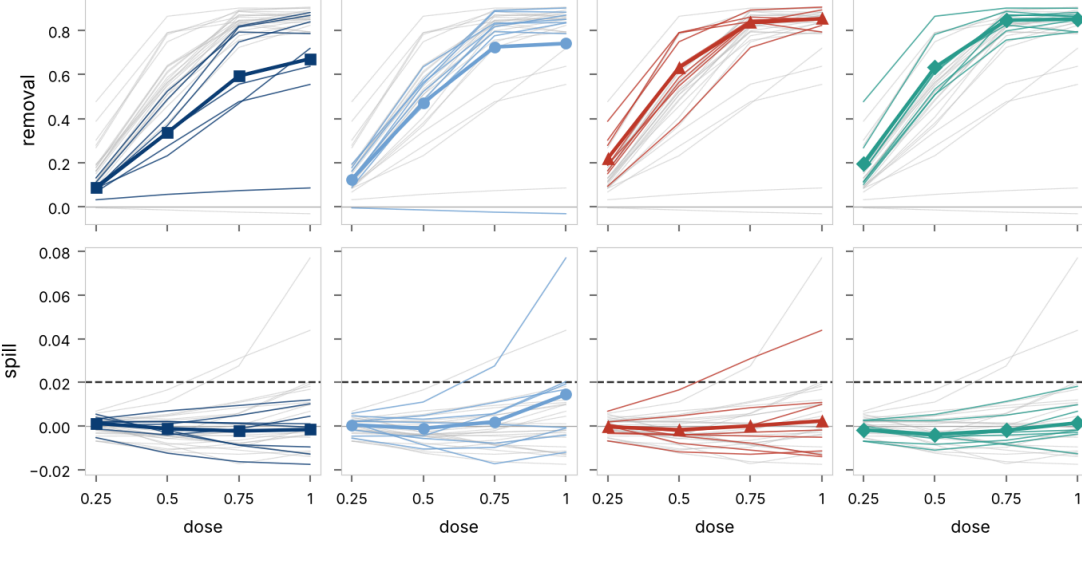
What we expect. On the slice-terms seeds, `split-anti` removed more than `deep, hold 0.2` at every seed and stayed within the criterion on three of four. We expect the same at the fresh seeds. The main number is the removal at the full dose, paired by seed against `deep, hold 0.2`.

A pass needs both removal and spill to hold:

- A pass would be `split-anti` removing more than `deep, hold 0.2` on at least 6 of 8 seeds, by at least 0.05 on average, with at least 7 of 8 runs within the criterion.
- A partial pass would be the removal gain holding with 5 or 6 runs within the criterion.
- A miss would be no removal gain (fewer than 5 seeds ahead, or an average gain under 0.02), or 4 or fewer runs within the criterion.

If the removal gain lands between 0.02 and 0.05 on average, or the seeds split evenly on its direction, the result would be outside the plan, and the verdict would be Unresolved.

What we saw. `split-anti` removed more than `deep, hold 0.2` on 6 of 8 seeds, by 0.18 on average, and stayed within the criterion on 7. By the frozen rule that is a pass.



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

The gain is uneven across seeds. On four of them the two arms remove about the same share of the op, with the split a little behind on two. On the other four the high hold everywhere removed much less, and on one of those it removed almost nothing.

seed	split-anti	deep, hold 0.2	gain
704	0.86	0.87	-0.01
705	0.79	0.79	+0.01
706	0.82	0.88	-0.06
707	0.86	0.84	+0.02
708	0.91	0.09	+0.82
709	0.85	0.56	+0.29
710	0.89	0.72	+0.17
711	0.85	0.64	+0.21
mean	0.85	0.67	+0.18

***Removal at the full dose, paired by seed**, and the gain of `split-anti` over `deep, hold 0.2`.*

So the high hold at the pulled slices does cost removal, as the slice-terms experiment found, but here only on some seeds. With the weight split by slice, no seed shows that cost.

The gain seems to come from where the anti weight sits, and not only from how much there is. `deep, hold 0.1` sums to about the same weight over the five slices as `split-anti` (0.50 against 0.49), and it removed less, lost the anchor on one seed, and spilled over the criterion on two.

The one `split-anti` run over the spill criterion (seed 704) goes over at the two highest doses only. `falloff` and `deep, hold 0.2` were within the criterion on every seed, and `deep, hold 0.1` on 6.

arm	removal ↑	within criterion ↑	spill, any dose ↓	raw spill ↓	landing ↑	task change ↑
deep, hold 0.2	0.67 (0.09 to 0.88)	8 of 8	0.004 (-0.005 to 0.012)	0.010 (0.000 to 0.017)	0.43 (0.09 to 0.55)	-0.015 (-0.054 to 0.011)
deep, hold 0.1	0.74 (-0.03 to 0.90)	6 of 8	0.017 (-0.002 to 0.077)	0.038 (0.012 to 0.102)	0.45 (0.01 to 0.54)	-0.011 (-0.036 to 0.008)
split-anti	0.85 (0.79 to 0.91)	7 of 8	0.007 (-0.007 to 0.044)	0.019 (-0.002 to 0.047)	0.52 (0.47 to 0.56)	-0.004 (-0.015 to 0.006)
falloff	0.85 (0.79 to 0.90)	8 of 8	0.003 (-0.004 to 0.018)	0.017 (0.003 to 0.038)	0.52 (0.49 to 0.56)	-0.007 (-0.032 to 0.005)

***Removal, spill, landing and task**, seed mean and range over 8 runs. Removal is 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 raw spill is the largest drop on another op at the full dose without subtracting the control. The landing is the share of the distance to the target null that the full edit covers, and the task change is the mean score over ops less that of the control. Arrows mark which direction is better.*

The edit lands a little closer to the target null in the two arms with a split or falloff hold than in the uniform ones. The task change is small in every arm, and smallest in `split-anti`.

Pass

`split-anti` removed more than `deep, hold 0.2` on 6 of 8 seeds, by 0.18 on average, with 7 of 8 runs within the criterion.

Runs that lost the anchor (post hoc)

Each uniform hold lost the anchor on one seed. On those runs the example answers stay off e1 at every slice, and the query answer aligns instead, as in seed 700 of `pool-slices` in the slice-terms experiment. The prediction of the query answer is made at the query = , before that position, so an edit there cannot reach it.

arm	seed	removal	α , example answers	α , query answer
deep, hold 0.2	708	0.09	0.01	0.98
deep, hold 0.1	706	-0.03	0.04	1.00

***The runs that lost the anchor.** Removal at the full dose, and α at slice 4 at the end of training on `difference` probe contexts, at the example answers and at the query answer.*

Neither the split nor the falloff lost the anchor on any seed. One reading is that a high hold at the pulled slices makes the example answers costly to align, so the pull settles on the position that aligns most readily. Eight seeds per arm give two such runs in all, so this is a pattern to watch rather than an established cause. A different pooling temperature τ_p might also keep the pull on the example answers; we have not tried one here.

A linear falloff does as well as the split (H2)

✓ Pass

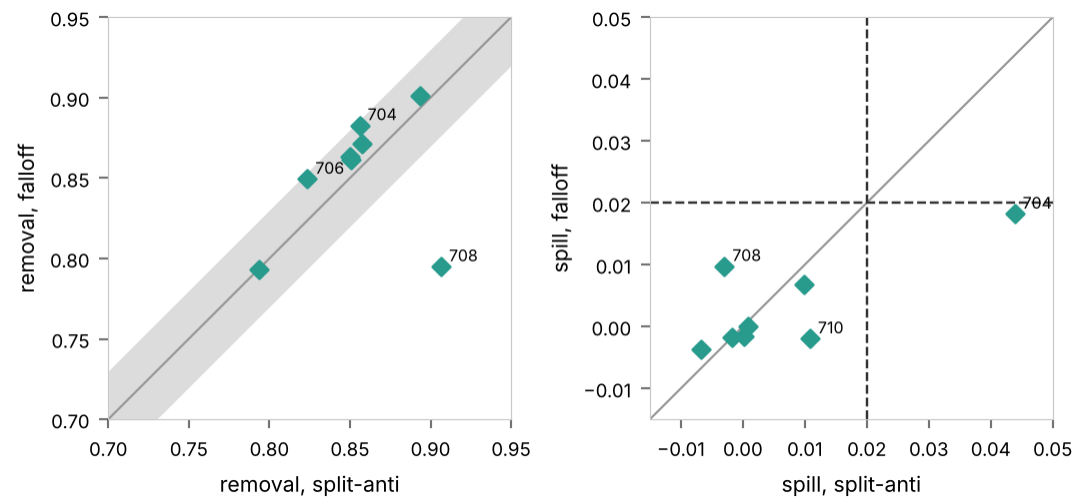
What we expect. `falloff` holds slice 1 a little lower than `split-anti` and slices 2 and 3 higher. The lower hold at slice 1 may let a little more spill through, and the higher holds at slices 2 and 3 may counter the pull there and cost a little removal. We expect both effects to be small. The main number is the removal at the full dose, paired by seed against `split-anti`.

As in H1, a pass needs both removal and spill to hold:

- A pass would be `falloff` removing within 0.03 of `split-anti` on average, with at least 7 of 8 runs within the criterion.
- A miss would be `falloff` removing 0.03 or more less than `split-anti` on average, or 6 or fewer runs within the criterion.
- If `falloff` removes 0.03 or more than `split-anti` on average, that would also be outside what we expect, and we would report it as a pass with the gain named.

If the seeds split evenly on the direction of the removal difference and the average is near the 0.03 edge, the verdict would be Unresolved.

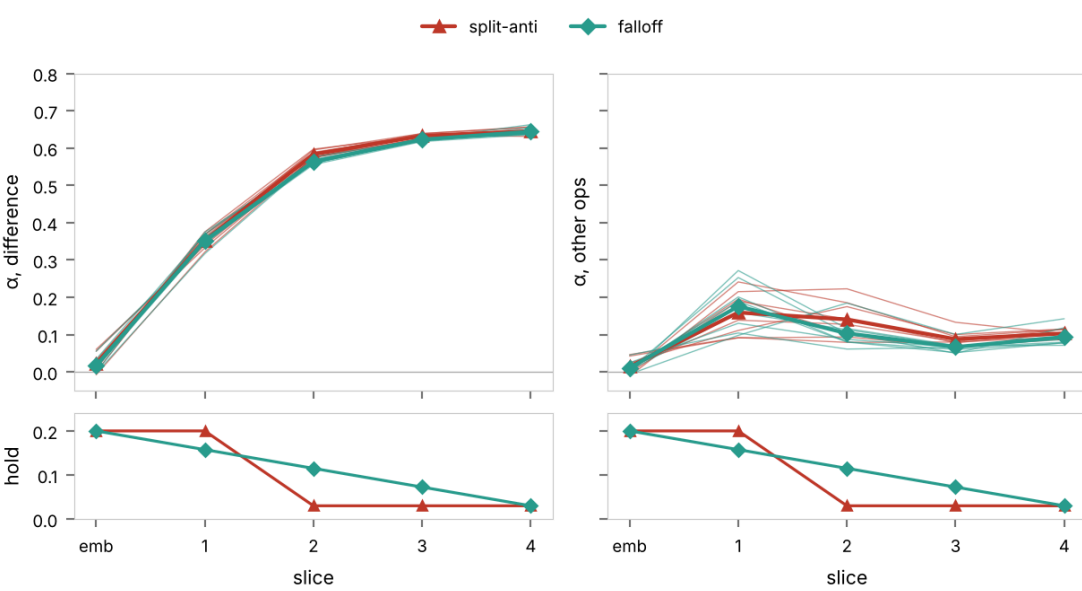
What we saw. `falloff` removed about the same share of the op as `split-anti`, 0.002 less on average, and stayed within the criterion on 8 of 8 seeds. By the frozen rule that is a pass.



falloff against split-anti, paired by seed. Left: removal at the full dose, with the line of equality and a band of ± 0.03 . Right: the largest spill at any dose, with the line of equality and the criterion on each axis (dashed). Seeds off the diagonal are labeled.

On seven seeds the two arms remove nearly the same share of the op. On seed 708 `falloff` removes about a tenth less, and the spill points sit low for both arms apart from the one `split-anti` run over the criterion.

The alignment by slice agrees. In `difference` contexts the two arms are nearly the same at every slice. In the contexts of the other ops, `falloff` sits a little higher at slice 1, where its hold is lower, and a little lower at slices 2 and 3, where its hold is higher. Both differences are small next to the alignment the pull gives.



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). One hairline per run and the seed mean bold. Underneath is the hold of the anti weight at each slice.

So the boundary that the split sets by hand does not seem to matter much: a straight line between the same two levels gives nearly the same anchor.

Pass

`falloff` removed within 0.002 of `split-anti` on average, with 8 of 8 runs within the criterion.

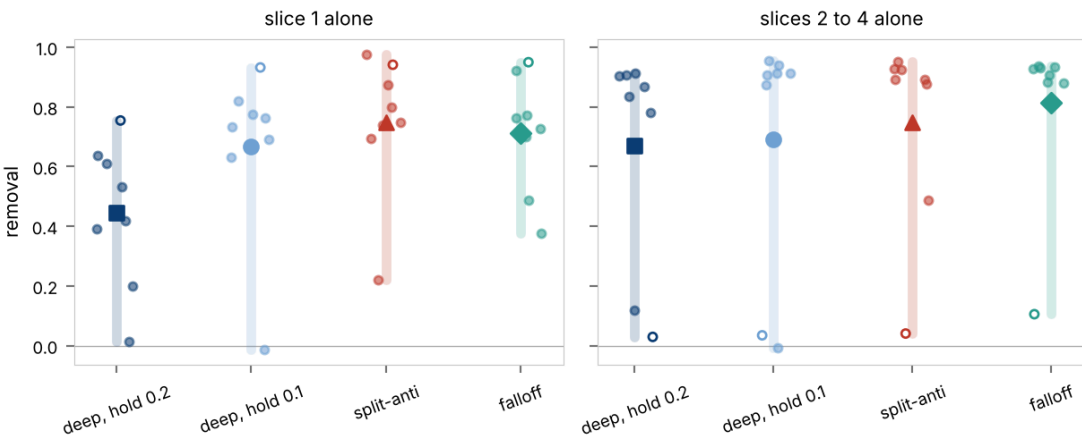
Which slices the edit needs (E1)

On the slice-terms seeds and the late-pull seeds, editing slice 1 alone removed most of the op on every seed, while editing slices 2 to 4 alone removed nearly everything on some seeds and little on others. One reading is that the evidence for the op is carried from the example answers to the = sites by attention, in some other direction than e1, and that the block whose heads make that move reads the example answers as they are at its input. Slice 1 is the input to every block after it, so clearing e1 there removes the evidence before it moves. Clearing slices 2 to 4 leaves the copy at slice 1 for block 2 to read. This reading is not tested here.

We repeat the edit on slice 1 alone and on slices 2 to 4 alone for every run, to see whether the split by seed holds at the fresh seeds and in each arm.

What we saw. The edit on slices 2 to 4 alone removes nearly all of the op on almost every run, and falls short on seed 710 in every arm. On seed 708 it removes about half under split-anti, and nearly all under the other arms that anchored.

The edit on slice 1 alone removes most of the op on most runs of the split, the falloff, and the low uniform hold. It removes less under the high uniform hold, even on the seeds where that hold anchored well.



Removal by which slices the edit covers, at the full dose. One column per arm, a thin bar spanning the seeds, the seeds as faded dots, and the seed mean in the arm's marker. Open markers are the seed where the edit on slices 2 to 4 alone falls short in every arm.

The edit on slices 2 to 4 alone falls short on the same seed in every arm. Every arm shares the initialization, batches, and label draws of its seed, so this fits the seed deciding where the evidence for the op is read. The reading in the introduction to this section would trace that to which block moves the evidence to the = sites.

The edit on slice 1 alone removes different amounts in different arms. So the hold at the pulled slices shifts some of what the edit reaches, but it does not move the seed on which the later slices miss.

arm	slices	704	705	706	707	708	709	710	711
deep, hold 0.2	slice 1	0.53	0.61	0.64	0.39	0.01	0.20	0.75	0.42
deep, hold 0.1	slice 1	0.77	0.69	-0.01	0.63	0.76	0.73	0.93	0.82
split-anti	slice 1	0.75	0.87	0.22	0.74	0.98	0.70	0.94	0.80
falloff	slice 1	0.70	0.76	0.38	0.73	0.49	0.77	0.95	0.92
deep, hold 0.2	2 to 4	0.90	0.87	0.91	0.91	0.12	0.78	0.03	0.83
deep, hold 0.1	2 to 4	0.91	0.88	-0.01	0.91	0.95	0.91	0.03	0.94
split-anti	2 to 4	0.89	0.88	0.89	0.93	0.49	0.93	0.04	0.95
falloff	2 to 4	0.91	0.88	0.93	0.93	0.88	0.94	0.11	0.93

Removal by which slices the edit covers, at the full dose, by arm and seed.

The runs that lost the anchor (in H1) remove almost nothing under any edit, and appear here as the low points of the uniform holds.

Decision

We chose the recipe of record for D2.2 among the four arms by the criteria frozen with the predictions: runs within the spill criterion, removal at the full dose, landing, and a hard gate on the task change. These were fresh seeds for every arm, so the numbers for the chosen arm can be quoted as a result without a further confirmation.

arm	within criterion ↑	removal ↑	landing ↑	task change ↑	worst op change ↑	clears the task gate
deep, hold 0.2	8 of 8	0.67 (0.09 to 0.88)	0.43 (0.09 to 0.55)	−0.015	−0.131 (hue-hsv , seed 705)	no
deep, hold 0.1	6 of 8	0.74 (−0.03 to 0.90)	0.45 (0.01 to 0.54)	−0.011	−0.071 (hue-hsv , seed 706)	no
split-anti	7 of 8	0.85 (0.79 to 0.91)	0.52 (0.47 to 0.56)	−0.004	−0.040 (value-hsv , seed 710)	yes
falloff	8 of 8	0.85 (0.79 to 0.90)	0.52 (0.49 to 0.56)	−0.007	−0.066 (hue-hsv , seed 708)	no

Every criterion for every arm. Removal and landing as seed mean and range. The task change is the mean over seeds of the mean score over ops, less that of the control. The worst op change is the largest loss on any one op at any seed against its control. The gate excludes an arm whose task change is below −0.02, or whose worst op change is below −0.05. Arrows mark which direction is better.

Only split-anti clears the task gate. The three other arms each lose more than 0.05 on some op at some seed, and every such loss is on one of the HSV ops (hue-hsv , sat-hsv , value-hsv). These losses are larger than the spread between the controls at these seeds, at most 0.026 on any op.

We chose falloff anyway, which departs from the frozen gate. Its one miss is on hue-hsv at seed 708, a little past the limit. That seed is also where falloff removed a tenth less than the split, and where its task change over all ops was lowest, so the miss looks like one weaker run rather than an op that falloff harms. The HSV ops are the last to be learned, so a gate on any op at any seed may be stricter for them than for the others.

In its favor, falloff kept every run within the spill criterion, and its spill varied less across seeds than that of the split. That advantage rests on one run, the split-anti run at seed 704. The falloff also has no slice boundary to choose, and its levels can still be tuned, for example with a lower anti weight on the earliest slices.

Discussion

Setting the anti weight by slice holds up at fresh seeds. Both the split and the straight-line falloff anchored on every seed, kept the op removable, and kept the edit within the criterion on nearly every seed. The high uniform hold gave a weaker anchor on several seeds, and each uniform hold lost the anchor altogether on one.

The gain over the uniform high hold depends on the seed, so the high hold at the pulled slices seems to cost removal through the seeds where the pull struggles to settle on the example answers, rather than by trimming every run a little.

The falloff matching the split suggests the hold matters at its two ends: high on the embedding and slice 1, where the slice-terms experiment found the drift onto e_1 that causes the spill seemed to sit, and low on the slices the pull acts on. Where in between the weight steps down seems not to matter much at this resolution.

Method

Training, measurement and dose scoring follow the slice-terms experiment, and its experiment module is reused with new arms and seeds. The anti weight per slice is the `AntiSpec.hold` tuple added there; the `falloff` holds are the unrounded straight line. The controls are ex-2.2.23's 400-epoch runs at seeds 704 to 711.

Two of the 32 trainings were stopped by the training watchdog while their checkpoints were uploading, after the last step, and were trained again from the start with the same seed and configuration.

The run cost about \$12.