Fixed labels

> Every anchored run so far drew its labels afresh on each visit to a context, so over a run the anchor saw labels on most `difference` contexts. Holding one draw fixed for the run, as a real labelled set would be, still keeps the anchor at every rate we tried. But the edit removes the op less reliably: a few runs end well below the rest, and more labels in the fixed set did not cure it. Which runs end low looks like chance: retraining a low run with slightly different arithmetic brings it up with the rest, and other set and seed pairs go low instead.

The *post-hoc comparison plan*° compares the edit along the anchored axis with directions found after training on a control. Writing it up, we noticed that the anchor labeller (inherited from M1) makes a fresh 1/50 draw on every visit to a context.

post-hoc comparison plan. Plans a comparison of the edit along e_1 with directions fitted after training on the controls at seeds 712 to 719, with the anchor's tolerance of scarce labels among the claims it tests.

An epoch covers about a quarter of the training split, so over 400 epochs a context is in view about a hundred times. With a fresh draw on each, about four fifths of the 42,891 `difference` contexts carry a label at some point (about 34,000), far more distinct labels than "one context in fifty" suggests. A real pipeline labels a fixed set once.

This report trains the recipe of record, `τ_ℓ fixed`, with one draw made per run and held for the whole run. We try three rates: 1/50 (about 870 labelled `difference` contexts), 1/25 (about 1,740), and 1/10 (about 4,320). Each rate runs at the eight seeds of the *fixed confirmation*° (712 to 719). That experiment trained the same recipe with the per-visit draw, so its runs are the comparison.

fixed confirmation. `τ_ℓ fixed`, with the per-visit label draw, keeps the anchor on all eight runs at seeds 712 to 719 with removal near 0.9, and is the recipe of record for D2.2.

Observations

- The anchor settles at every rate (E1): all 24 runs are called kept at epoch 64, and α° at the example answers settles about as fast as under the per-visit draw.

Alignment. How far a state points along e₁…

- Removal and spill at the end (E2)°: every run keeps the anchor, and runs pass the task gate as often as the per-visit runs do. At 1/50 and at 1/10, a few runs end well below the lowest per-visit run, so the seed mean falls. At 1/25 removal matches the per-visit draw, but one run spills° late, over the criterion.

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….

- When the low runs fall behind (E3): the low runs mostly never reach the level of the rest. One run at 1/50 slides after reaching it, and the others give back only a little after their peak.
- The low runs' label sets (E4): nothing in what the labelled contexts show sets the low runs' sets apart, and labelling more contexts around a low run's set does not lift removal in step.
- Crossing label sets with seeds (E5): retrained with slightly different arithmetic, the two low 1/50 runs end with the rest. Five of the twelve new pairings of set and seed end low, with no set or seed low throughout.

Scope and method

This is an exploratory study, with no preregistration and no gate of its own. It started as a scout: one seed at 1/50, trained to epoch 64. Then 1/50 and 1/10 ran at four seeds to the end of training. At four seeds 1/50 removed less than 1/10, so they were carried to eight seeds, and 1/25 was added between them. The three rates were each chosen before their runs, and none was chosen by how it scored. After a first reading, E4 and E5 were added to look into the low runs; E5 trains sixteen more runs.

Every run is `τ_ℓ fixed` as the *fixed confirmation*° trained it, with one change to the labels. In place of the per-visit draw, a single draw made before training picks the labelled contexts. A picked context carries the label on every visit, and any other context never does.

fixed confirmation. `τ_ℓ fixed`, with the per-visit label draw, keeps the anchor on all eight runs at seeds 712 to 719 with removal near 0.9, and is the recipe of record for D2.2.

Each seed has its own random stream for this draw, so the seed sets the labelled contexts as well as the initialization and the batches. The draws are nested, so every context labelled at 1/50 is also labelled at 1/25 and at 1/10. The per-visit draw is still taken and set aside, so each run trains on the same batches from the same initialization as the per-visit run at its seed. Only which contexts carry the pull differs.

The pull is a mean over the labelled contexts in each batch, so the label rate does not change how hard the anchor pulls. At 1/50 it also leaves the number of labelled contexts per batch unchanged on average. What the rate changes is how many distinct contexts the pull ever falls on, and how much the pull in a batch varies from one batch to the next.

We compare conditions as two samples of eight runs (the D2.2 rule since the determinism decision), with a 95% Welch t interval on the difference of seed means.[1] Over eight seeds, removal° under the per-visit draw spreads by about 0.07.

1 Welch t interval…

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.…

All runs trained and were measured on L4 GPUs, as in the *fixed confirmation*. They share the controls of that experiment (no pull and no anti term, at the same seed).

The measurements

Every edit below is the e_1 edit. Removal° and spill° are each a drop under that edit, less a baseline. The baseline is 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....

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

Removal

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

Task change

The clean score per op, less that of the control at the seed. The gate of the D2.2 design asks for a mean change over ops of at least −0.02, and at least −0.05 on any one op.

Call

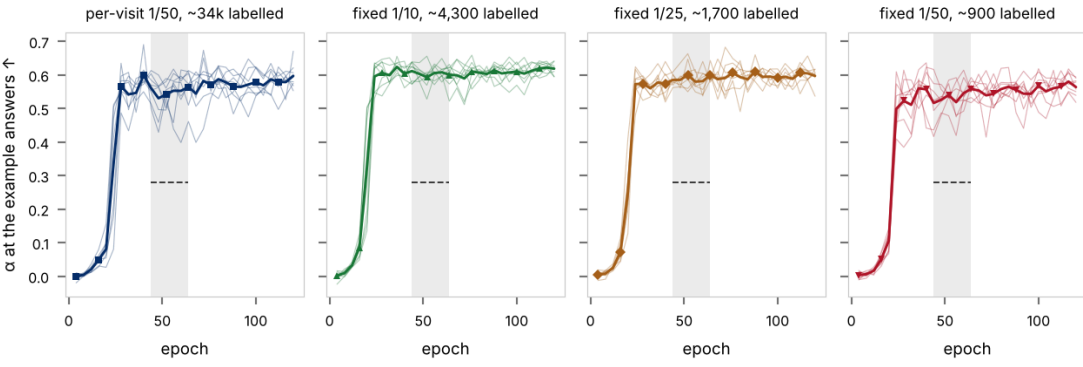
The early verdict on whether a run kept the anchor, from mean α at the example answers of the `difference` probe contexts at slice 3, over epochs 44 to 64. The probe contexts are a held-out set, the same for every run, and none carries a label. A run below 0.15 is called lost, and one at 0.28 or above is called kept.

The anchor settles at every rate (E1)

With the pull confined to fewer contexts, does the anchor still settle on e_1 by the time the call° is made? If a fixed set were too small, α° at the example answers may rise more slowly or level off lower.

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

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



The anchor settling, by condition. *α at the example answers of the `difference` probe contexts, at slice 3, along the first 120 epochs: the seed mean as a line with markers, each run faint behind it. The shaded band is the call window (epochs 44 to 64), and the dashed line across it is where the call counts a run as kept.*

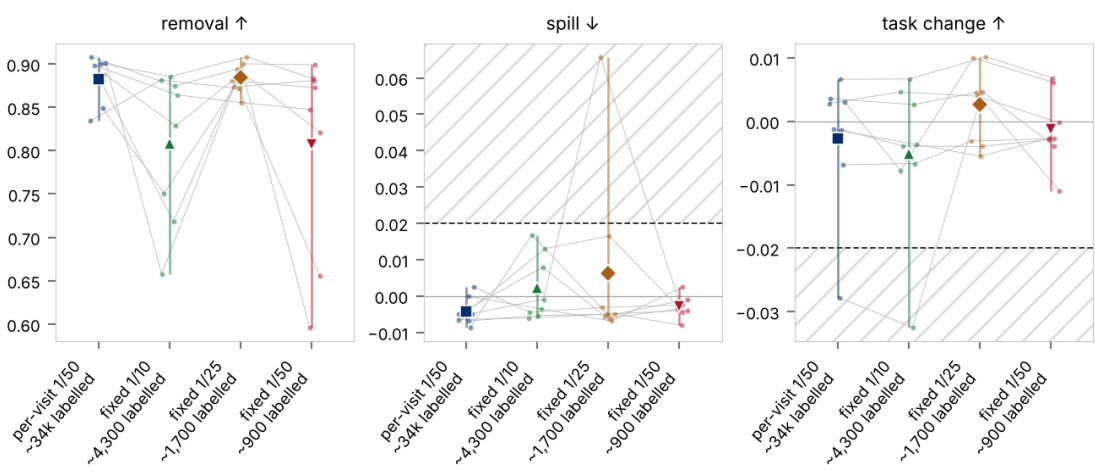
Every run is called kept, at every rate, with a wide margin. The rise over the first 40 epochs looks much the same in all four conditions. Over the call window, the seed mean of the fixed 1/50 runs is a little below that of the per-visit runs. The 1/25 and 1/10 runs are a little above it, and closer together. All of them are near the spread of the per-visit runs.

So a fixed set of fewer than a thousand labelled contexts is enough for the anchor to settle. But across the fixed-label runs, mean α over the call window says little about how much a run removes at the end (Pearson r = 0.04 over the 24 runs).

Removal and spill at the end (E2)

Does a fixed labelled set remove `difference` as well as the per-visit draw, at the same cost to the other ops? If label count mattered over this range, removal° would climb from 1/50 to 1/10.

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



***Removal, spill, and task change at epoch 400, by condition**, from the most labelled `difference` contexts to the fewest. The seeds are small dots over a thin bar for their range, and the seed mean is the large mark; a grey line joins the runs of each seed, which share their initialization and batches. **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.*

labels	contexts labelled	removal ↑	removal less per-visit	spill ↓	over criterion ↓	over, block 1 ↓	task change ↑	task gate missed ↓
`per-visit 1/50`	34,000	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)	1 of 8
`fixed 1/10`	4,324	0.81 (0.66 to 0.89)	−0.149 to −0.001	0.002 (−0.006 to 0.017)	0 of 8	0 of 8	−0.005 (−0.032 to 0.007)	1 of 8
`fixed 1/25`	1,742	0.88 (0.86 to 0.91)	−0.023 to 0.026	0.006 (−0.007 to 0.066)	1 of 8	1 of 8	0.003 (−0.005 to 0.010)	0 of 8
`fixed 1/50`	874	0.81 (0.60 to 0.90)	−0.173 to 0.021	−0.003 (−0.008 to 0.002)	0 of 8	0 of 8	−0.001 (−0.011 to 0.007)	0 of 8

***Removal, spill, and task change at epoch 400**, seed mean and range over eight runs at seeds 712 to 719. Removal and spill are net of the median drop over the other 63 coordinate axes of the run. "Contexts labelled" is the number of `difference` contexts that carried the label at some point in the run, the mean over seeds for a fixed draw and an estimate for the per-visit draw; both count over the whole corpus, of which training visits the first nine tenths. "Removal less per-visit" is a 95% Welch t interval on the difference of the seed means. "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.*

Every run keeps the anchor, and the task change° is near zero in every condition, with no more gate misses than under the per-visit draw. The differences are in removal, and in spill° on one run.

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

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

Removal does not climb with the label count. At 1/25 every run is within the per-visit range, and the seed means match.

At 1/50 and at 1/10 most runs remove about as much as the per-visit runs, but a few end below 0.8, under the lowest per-visit run (0.83): two at 1/50 (0.60 to 0.66) and three at 1/10 (0.66 to 0.75). Those runs pull the seed mean down by about the same amount at both rates.

The two rates have their low runs mostly at different seeds, and share only seed 719. At those seeds, the per-visit runs range from the lowest per-visit run (seed 715) to the second highest (seed 717; the grey lines in the figure). So the low runs look more like single runs that happened to end low than like seeds that are hard for the recipe.

All three low runs at 1/10 are at seeds 717 to 719, which is why the first look at four seeds missed them.

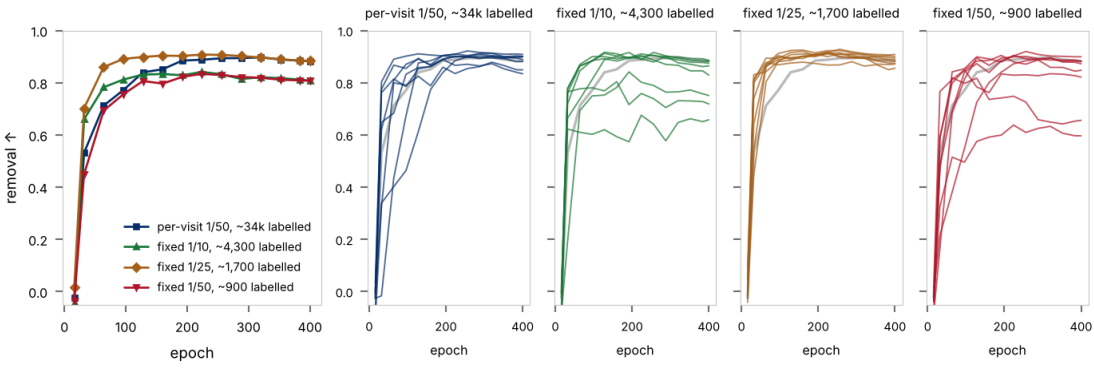
At 1/25, one run (seed 716) ends over the spill criterion, on `darken`. It is also over the criterion under the edit at block 1 alone. No other run in any condition goes over. Its spill rises late, from about epoch 192, like the late rise the *soft pool scout*° saw on one seed of `τ_ℓ fixed`. Before that, it was over the criterion at only one checkpoint, just barely, at epoch 32 while the anchor was still settling.

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 intervals on the removal difference from the per-visit draw are wide. At 1/10 the interval just excludes zero, and at 1/50 it does not. With eight runs each, a difference set by two or three low runs is hard to pin down. So we can't tell the three fixed rates apart, and 1/25 may look best only through the luck of its eight seeds.

When the low runs fall behind (E3)

Do the low runs fail to reach the level of the rest, or reach it and lose ground? The answer says whether the shortfall is about the anchor finding e_1 or about holding it there.



Removal along training, by condition. *Removal under the full edit 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, as a share of the way from the clean score at the checkpoint to the target null at the end of the run.* ***Left:*** *the seed mean of each condition.* ***Right:*** *each run, one panel per condition, with the seed mean of the per-visit draw as a faint line behind for reference.*

Mostly the first. Every low run peaks below every other run in any condition: 0.64 to 0.84, against 0.85 and up. The gap is there from early in training and does not close.

One run also slides. Seed 715 at 1/50 is near the top of its condition at epoch 64 and peaks at 0.82, then slides to 0.66 over the rest of the run, while the anti weight anneals. Other runs give back a little after their peak too: up to 0.09 with a fixed set, and up to 0.05 under the per-visit draw.

A slow start doesn't make a low run. Seed 717 under the per-visit draw and seeds 712 and 716 at 1/50 are under 0.25 at epoch 32 and end with the rest.

Neither pattern lines up with the rate. The 1/25 runs show neither, and they reach their final level sooner than any other condition, by about epoch 64. With eight runs a condition, that may be the same luck of the seeds as in E2.

The low runs' label sets (E4)

With a fixed set, the pull keeps returning to the same few hundred or thousand contexts, so what the anchor learns could depend on which contexts the draw happened to pick. A set whose contexts lean toward one part of the colour cube, one ordering of the operands, or more noisy examples might anchor the op only partly. Is there anything unusual about the label sets of the low runs?

To check, each `difference` context gets one value per feature, averaged over its examples and its query: the mean of each channel for each operand and the answer, the signed difference of the operands per channel, the lightness of the operands and the answer, the distance between the operands, and the counts of noise and cube examples. Each label set is then compared with all `difference` contexts on all 17 features at once (a Hotelling T^2 test[2]), and on how it fills the cube (by pairs of operand octants, and by answer region, χ^2 tests).

2 Hotelling T^2 test…

labels	712	713	714	715	716	717	718	719
`fixed 1/50`	0.64	0.45	0.97	**0.87**	0.75	0.13	0.87	**0.96**
	0.88	0.85	0.82	**0.66**	0.87	0.88	0.90	**0.60**
`fixed 1/25`	0.20	0.56	0.97	0.41	0.09	0.72	0.66	0.22
	0.91	0.86	0.89	0.87	0.88	0.87	0.90	0.89
`fixed 1/10`	0.32	0.30	0.32	0.63	0.30	**0.97**	**0.32**	**0.84**
	0.89	0.86	0.87	0.88	0.83	**0.66**	**0.72**	**0.75**

How typical each label set is, by seed. *In each cell, above: the χ^2 tail of a Hotelling T^2 that compares the mean of the labelled* `difference` *contexts with that of all of them, over 17 features of what a context shows; a set drawn at random has a tail spread evenly between 0 and 1, and a small one is unusual. Below: removal at epoch 400. The low runs are in bold.*

The low runs' sets look like any other. Their tails run from 0.32 to 0.97, and four of the five are above 0.5. The smallest tail of any of the 24 sets is 0.09, about what 24 random draws give. On the cube, the smallest of the 48 occupancy tails is 0.01, about what 48 tests give, and it is not on a low run. No single feature tracks removal° over the 24 runs either: the strongest is `rhs B`, at Spearman $\rho = -0.26$ ($p = 0.22$).

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

The nesting gives a second view. Every context labelled at 1/50 is also labelled at 1/25 and 1/10, so the three sets of a seed grow around one another. If a run ended low because of some contexts in its set, those contexts stay in every larger set, diluted by the new ones, so removal should climb as the set grows and not fall again. At seed 715 the low run at 1/50 recovers once more contexts join its set (0.66, then 0.87 and 0.88), which fits. But at seed 719 removal goes 0.60, 0.89, 0.75 as the set grows. At seeds 717 and 718 the 1/25 runs end at 0.87 and 0.90, while the 1/10 runs, whose sets contain theirs, end at 0.66 and 0.72. Low runs come and go as the set grows.

So the label sets don't account for the low runs, as far as these features can see. That leaves run-to-run sensitivity, where a small change in which contexts carry the pull sends a run to a different end. These tests can't rule out a feature they don't measure, so E5 trains sets and seeds in new pairings.

Crossing label sets with seeds (E5)

Does a low run follow its label set or its seed? Here the 1/50 sets of four seeds are each trained at all four of those seeds: 715 and 719, whose 1/50 runs ended low, and 712 and 718, whose runs ended high. A set that holds a run back would give a low row, and a seed that does would give a low column.

These sixteen runs came after a faster training step was merged, which also turns the GPU determinism flags off (a shared compilation cache keeps runs reproducible instead). They trained on L40S cards rather than L4s. A run sees the same batches from the same initialization as before, so on the diagonal only the arithmetic differs from the 1/50 runs of E1 to E4, in its last bits.

labels	seed 712	seed 715	seed 718	seed 719
draw of 712	0.83 (0.88)	**0.74**	**0.68**	0.87
draw of 715	0.85	0.86 (0.66)	0.87	0.91
draw of 718	0.87	**0.79**	0.87 (0.90)	**0.60**
draw of 719	0.87	0.89	**0.70**	0.87 (0.60)

The crossed runs: removal at epoch 400. *Each row is the 1/50 label set drawn for one seed, and each column the model seed it trained at (initialization and batches). On the diagonal, in brackets: the run of E1 to E4 with the same set at the same seed. Runs below 0.8 are in bold. All sixteen trained with the faster training step and the determinism flags off, on L40S cards; the runs of E1 to E4 trained on L4s with the flags on.*

Neither the sets nor the seeds account for the low runs. On the diagonal, both low runs now end with the rest: seed 715 at 0.86 (it was 0.66) and seed 719 at 0.87 (it was 0.60). Meanwhile five of the twelve new pairings end low (the set of 712 at seed 715, the set of 712 at seed 718, the set of 718 at seed 715, the set of 718 at seed 719, the set of 719 at seed 718), against five of sixteen at 1/50 and 1/10 in E2. No row and no column is low throughout. Every crossed run stays under the spill° criterion (the largest is 0.005).

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

So whether a fixed-label run ends low is decided by small differences somewhere along the run, which the same set, seed and batches don't fix. The per-visit runs of the *fixed confirmation*° spread less, though the reason for that is still open. The task change° of these runs isn't compared here: their controls trained on L4s with the flags on.

fixed confirmation. `τ_ℓ fixed`, with the per-visit label draw, keeps the anchor on all eight runs at seeds 712 to 719 with removal near 0.9, and is the recipe of record for D2.2.

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

Discussion

A fixed labelled set at one context in fifty is enough for the anchor to settle on e_1 and keep it (E1, E2). That was the main worry, since the per-visit draw had shown the anchor most of the `difference` contexts over a run. On that count, the scarce-label claim of the D2.2 design holds with labels as a real pipeline would have them.

But removal° is less reliable with a fixed set: a few runs in eight end well short of the rest (E2), mostly because they never reach the level of the others (E3). Since a fixed set five times larger did no better, more labels of the same kind do not seem to be the fix, at least between about 900 and 4,300 contexts. Nor do the low runs' sets differ from the others in what their contexts show (E4), and in new pairings of set and seed the low runs move about: the same set at the same seed can end low or high depending on the last bits of the arithmetic (E5). The per-visit draw spreads less, and that may come from averaging over almost every context, which could steady a run against such small differences.

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

The *low-runs*° report follows the low tail into the model. Every run writes the evidence that an example fits `difference` both on e_1 and along a second direction, and the low runs are the ones whose reader blocks take more of that evidence from off the axis, a difference that the loss and the anchor term do not see. The same off-axis part is in the per-visit runs at a few hundredths, so the fixed set makes a general mechanism more visible rather than a new one.

low-runs. Re-analysis of the 48 stored runs of this experiment and the fixed confirmation…

If this holds, a comparison against post-hoc methods with fixed labels should expect a wider spread on the anchored side than the *fixed confirmation*° showed. With that spread, eight runs give a noisy seed mean.

fixed confirmation. `τ_ℓ fixed`, with the per-visit label draw, keeps the anchor on all eight runs at seeds 712 to 719 with removal near 0.9, and is the recipe of record for D2.2.

References

fixed confirmation

: `τ_ℓ fixed`, with the per-visit label draw, keeps the anchor on all eight runs at seeds 712 to 719 with removal near 0.9, and is the recipe of record for D2.2.

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.

low-runs

: Re-analysis of the 48 stored runs of this experiment and the fixed confirmation: the low runs read part of the evidence for `difference` from off e_1, through the reader blocks, and twins trained from the same labels, seed and batches match in loss and α while parting in removal.

post-hoc comparison plan

: Plans a comparison of the edit along e_1 with directions fitted after training on the controls at seeds 712 to 719, with the anchor's tolerance of scarce labels among the claims it tests.

---

1. Welch t interval: a confidence interval for a difference of two means that does not assume the two samples have equal variance. ↩

2. Hotelling T^2 test: the many-feature version of a t test, asking whether a sample mean differs from a reference mean, with the features' correlations taken into account. ↩