Directions found after training, beside the anchor

We compare the anchored° edit with edits along directions found after training: four standard fitters, and one direction optimized against the scoring itself. Each is fitted on a control model with every label, tuned per run on data it is never scored on, and scored the same way as the anchor. Given role labels and a per-run grid, the fitters remove more of the op than the anchored edit within the same spill° limit, mostly by editing the example answers, a role that a real corpus does not label. Restricted to fits that need no role labels, the best of them removes about as much as the anchor and spills past the criterion on a few more runs.

Anchored. The full edit along e_1, as published in fixed-effects….

Spill. The largest drop on any other op under the edit….

Placing a concept during training needs an extra training term, and it lowers the task score somewhat. This report asks whether that placement does better than searching for the concept afterwards, on the same model and with the same scoring.

Findings

- The trade-off (E1) — On every run but one, each standard fitter climbs past the anchored° removal° before it reaches the spill° criterion, mostly through edits at the example answers. With a configuration chosen on the other seven runs instead, the fitters remove a little less, and on one run in eight a choice made elsewhere misses the slice where that run holds the op.
- Spill at matched removal (H1) — **Misses**. Turned up to the anchored removal, no standard fitter spills past the criterion on more than one run, so each is accepted.
- Removal at the spill criterion (H2) — **Misses**. LEACE° and INLP remove more than the anchor within the criterion, the probe is Unresolved, and difference in means° is left out under H4; with the pooled fits alone, LEACE is close to the anchor and the others remove less.
- The ceiling against the fitters (H3) — **Misses**. The optimized direction°, which acts at every position, removes less than the rank-1 fitters that edit the example answers.
- The ceiling against the anchor (S1) — **Matches**. Net of the axis-median baseline that favors it, the optimized direction removes about as much as the anchor, and its two runs over the spill criterion are within chance of the anchored none.
- The fitters find a placed concept (H4) — **Partial**. The probe, LEACE, and INLP find the placed concept; difference in means misses the gate by less than 0.001.
- Grading, landing, and where the answers go (E2) — The edits grade alike and send the answers to the same places; the fitters land a little further toward the target null° than the anchor.
- How many labels a fitter needs (E3) — Difference in means, LEACE, and INLP keep their removal down to a fiftieth of the labels, and the probe drops to 0.84.
- The write against each bound (E4) — The writes stay within their bounds apart from three sites on two runs, though most post-hoc edits act at one slice, where the write equals the bound.
- What a fresh probe can still find (E5) — No edit brings a fresh probe near chance, LEACE included.
- The two baseline rules (E6) — The two baselines disagree at configurations far past the spill criterion; at the points the verdicts read, no verdict of H1 or H2 changes under either one alone, and S1 would fall short under the permuted-label rule alone.
- Without role labels — Restricted to fits that need no role labels, LEACE and INLP remove about as much as the anchor and spill past the criterion at matched removal on two runs in eight against none; difference in means removes less and spills on six. Drawn again with these fits, E1 to E5 show the same picture, with the fitters a little lower.

Anchored. The full edit along e_1, as published in fixed-effects....

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Spill. The largest drop on any other op under the edit....

LEACE. Least-squares concept erasure (Belrose et al., 2023)...

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were....

Target null. Where a removal should land...

> **How to read this report**
>
> This report was preregistered: the predictions, the grid, and the rule that picks a configuration from it were frozen at commit `bf73293`, before any run of this experiment. Each section opens with what we expect, and the results replaced the placeholders in place. A few analyses were added after the run, and each says so where it appears: the configuration chosen on other runs in E1, the reading of H4 under the axis baseline, the section Without role labels, and the marks for the *sans-diff model*°. They give no verdicts.
>
> The hypotheses are judged one fitter at a time, and then as a whole. A fitter is *accepted* when it does what the test asks of a working method (in H1 and H2, matching or beating the anchor; in H4, finding a placed concept) and *rejected* when it does not. A hypothesis as a whole *passes*, *misses*, or is *partial*; a fitter or a hypothesis the rule cannot settle is *Unresolved*.

sans-diff model. A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

Why this experiment

The anchored° pipeline trains the model with a pull toward the first basis vector e_1 on labelled `difference` contexts, and removes the op afterwards by projecting e_1 out of the residual stream[1]° at every slice and position. On the recipe of record, the *fixed-effects*° experiment found that this edit removes most of the op and costs the other ops less than the selectivity criterion allows. It also drops each context by about as much as that context needed `difference`, and the blocks after the edit keep what it writes within a bound set before the edit runs.

Anchored. The full edit along e_1, as published in fixed-effects....

1 Residual stream...

***fixed-effects*.** The full scoring suite on the recipe of record at seeds 712 to 719...

A reader who knows the steering and erasure literature will ask whether a direction found after training does as well. A conventional pipeline trains with no pull and fits a direction for the concept post hoc, from the states and the labels. It then projects that direction out, replaces the component along it with a mean, or erases it with LEACE°. If one of those matched the performance of the anchor on this model, we would have to say that the placement added nothing a search could not find, in a model of this class.

LEACE. Least-squares concept erasure (Belrose et al., 2023)...

Ex-2.2.1° asked this once, for *red* on the D2.1 checkpoints. Every fitted direction there that removed at least half of *red* cost the lines with no red far more than the anchored edit did. But that run gave the fitters little: three control seeds, labels from the token alone, one fit per slice, and no tuning. So this report gives the fitters every advantage we can, scores them the same way as the anchor, and asks where they fall on the trade-off between removal° and spill°.

***ex-2.2.1*.** On the D2.1 checkpoints, every fitted direction that removed at least half of *red* from the control cost the lines with no red 0.37 or more, an order of magnitude above the anchored edit.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Spill. The largest drop on any other op under the edit....

Runs and methods

No new run is trained. The anchored° runs are from the recipe of record: the eight `τ_ℓ fixed` runs of the *fixed confirmation*° at seeds 712 to 719. The controls are the eight runs of the *anneal confirmation*° at the same seeds, trained with no pull, with the same initialization, batches, and corpus. Every post-hoc method is fitted on a control and scored on it, so each method has eight runs, and each comparison with the anchor is between two samples of eight runs.

Anchored. The full edit along e_1, as published in fixed-effects....

fixed confirmation. `τ_ℓ fixed` at fresh seeds 712 to 719 keeps every anchor; it is the recipe of record, and its checkpoints and dose scoring are stored.

anneal confirmation. Trained the controls at seeds 712 to 719 that this comparison fits on, sharing initialization, batches, and corpus with the anchored runs.

Anchored

The full edit along e_1, as published in fixed-effects. Nothing is refitted or retuned for this experiment.

Anchored, fixed labels

The same recipe trained on one 1/50 label draw held for the whole run, at the same seeds (the fixed-labels experiment), under the same edit. It is a robustness row beside the main table, from its stored scores; the hypotheses judge the anchored arm above, whose labels are drawn again at each visit, as one pass over a large corpus would draw them.

Difference in means

The unit vector from the mean state of the other contexts to the mean state of `difference` contexts. The direction most steering work uses. The states are unit vectors, so their mean lies inside the sphere, but only the direction of the difference is used, and the edit puts each state back on the sphere.

Ridge probe

The normalized weight vector of a linear readout of the target, fitted by least squares with an ℓ_2 penalty. It finds what can be read off the state linearly.

LEACE

Least-squares concept erasure (Belrose et al., 2023): the smallest affine edit, measured in the covariance-whitened norm, after which no linear probe can read the target. For one target it works with two directions: c, the covariance of the states with the target (for a binary target, the difference in means), and $\Sigma^{-1}c$, where Σ is the covariance of the states, which is the direction a least-squares probe reads. It reads how much of the concept a state holds along $\Sigma^{-1}c$ and moves the state back along c until that reading is the mean, so afterwards the probe reads the same value for every state. The two directions differ, so the edit is oblique (not a perpendicular projection).

Iterated probe (INLP)

The ridge probe fitted, projected out, and fitted again on what is left, to a rank of 2, 4, or 8; the edit removes the whole span. It covers the case of a concept held across several directions, which an edit along one could not remove.

Optimized direction

A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were. It uses the target null, which no real pipeline has, so it is a diagnostic ceiling: about the best a rank-1 edit (an edit along one direction) of this control could do with the scoring in hand. INLP edits a span of several directions, so this ceiling does not bound it. COAST is its nearest published relative.

The first four are the standard fitters that H1 and H2 judge. Each is also fitted on the anchored runs, where training placed the concept on e_1, to check that it can find a concept that is there: it must remove about as much as the edit along e_1 does there, within the same spill° limit. That check is H4, and a fitter that does not get through it is left out of the verdicts of H1, H2, H3, and S1. The *sans-diff model*°, trained with no loss on `difference`, stays the reference for how close to the target null° a model of this size comes.

Spill. The largest drop on any other op under the edit....

sans-diff model. A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

Target null. Where a removal should land...

What each method gets

We use three sets of contexts, and no context is in more than one: a fit set° the directions are fitted on, a tuning set° the configuration is chosen on, and the test set° every published D2.2 score is taken on.

Fit set. 280,000 fresh contexts, 40,000 per op, drawn from the training generator of the run with its replacement op noise…

Tuning set. Another 14,000 fresh contexts from the same generator, drawn the same way.

Test set. The 14,000 held-out contexts of ex-2.2.21, 2,000 per op, so the anchored numbers need no recomputing.

Fit set

280,000 fresh contexts, 40,000 per op, drawn from the training generator of the run with its replacement op noise: about the size of the training corpus, and drawn the same way. Each comes with its generating op and its posterior over ops. The states are taken from the clean pass of the run at every slice and position. The fits are made from means and covariances accumulated over the set, so the states need not be held in memory at once.

Tuning set

Another 14,000 fresh contexts from the same generator, drawn the same way.

Test set

The 14,000 held-out contexts of ex-2.2.21, 2,000 per op, so the anchored numbers need no recomputing.

Every method is fitted with a label on every `difference` context. The anchor had a label on a fiftieth of them at each visit, but the labels were drawn again at every visit. So over 400 epochs about three quarters of the `difference` contexts were labelled at some point, about one and a half times each.

So over the whole run the anchor met about as many distinct labelled contexts as the fit set holds, but parity in that count is no tie. The anchor pull saw one `difference` context in fifty on each pass over the corpus, a few hundred at a time, through a stochastic term, while the fitters see every label at once, in closed form. What the fitters lack is a term that shaped the whole representation around the concept, and that is what this experiment measures. The label ladder of E3 measures how much the label count itself matters to them.

The fitters also get a graded target: the posterior on `difference` given the examples. A binary label marks the generating op even where the examples point to another op, and the graded target takes away that noise.

Each method then gets a grid over how it is fitted, where it acts, and how much it removes:

Factor	Levels
Target	binary (the generating op); graded (the posterior on `difference`), for the probe, LEACE, and the optimized direction
Sites fitted	every position of a context, pooled; the example answers; the query `=`; the query operands
Slices edited	each slice alone; slices 1 to 3; every slice
Sequential refit	off (every fit on the clean states); or, when several slices are edited, on (the fit at each slice made on the states as they arrive under the edits before it)
Positions edited	every position; or the sites fitted
Dose	0.25, 0.5, 0.75, and 1 of the component removed
Replacement	zero (the projection); or the mean coefficient over the contexts of the other ops (mean ablation)
Rank	2, 4, or 8, for INLP, on the binary target

The fits at the example answers, the query `=`, and the query operands use role labels a real corpus does not have. So they are generous diagnostics rather than candidate recipes, and the tables mark which sites each selected configuration used.

The anchor gets no grid. Its one configuration is the published edit, which was reached through some thirty experiments at other seeds. The fitters arrive with their own literatures behind them, and get a per-run grid on top.

Selection rule (frozen)

Per method and per run, every configuration is scored on the tuning set, net of the median drop under the same operator form along fifteen fixed coordinate axes other than e_1, and the frontier is taken: the configurations that no other beats on both removal and spill. Two points are read off the frontier. *Removal at the spill criterion* is where the frontier crosses a spill of 0.02, interpolated linearly between the two frontier configurations that bracket it; if no configuration is within the criterion it is the one with the least spill, and if every one is within it it is the one with the most removal. *Spill at matched removal* is where the frontier crosses the seed mean of the anchored removal at full dose, interpolated the same way; a frontier that never reaches that removal has no such point, and the run is recorded as not matched. The bracketing configurations are then scored on the test set, and each point is their test scores mixed with the weights found on the tuning set, so the test set chooses nothing, and the point can sit a little off the line on the test set; the table shows where it sits.

Removal° at the spill° criterion mirrors what a practitioner with one of these methods would do: sweep the layer and the strength on a validation set, keep the side effects within a tolerance, and take the strongest edit that does.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null….

Spill. The largest drop on any other op under the edit….

Measurements

Every measurement is defined as in the D2.2 design and *fixed-effects*[°], and taken through `sca.intervention.apply`. So any method that produces a subspace and an operator is scored by the code that scores the anchor.

[°] ***fixed-effects.*** The full scoring suite on the recipe of record at seeds 712 to 719…

Removal

The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null. Higher is better.

Spill

The largest drop on any other op under the edit. Lower is better, and the selectivity criterion is 0.02.

Net and raw

A net drop has a baseline drop subtracted; a raw drop does not. Published anchored scores are net of the median drop under the same edit along each of the other 63 coordinate axes of the run. A fitted direction is not an axis, so each post-hoc configuration gets a baseline in its own terms: the median drop over fifteen fits of the same fitter with the labels permuted across contexts. A permuted fit gives a direction with no concept in it, and the same kind of operator. On the controls the coordinate-axis median is computed too, and each hypothesis is judged on whichever of the two baselines favors the post-hoc method for that measure (the smaller for removal, the larger for spill). The anchor keeps its published baseline. Raw drops are shown beside net ones throughout.

Task change

The change in task score from the mean of the controls. It is zero for every post-hoc method, since the model is the control. For the anchor it is the published change, and the table of E1 shows it as the price of anchoring.

The trade-off (EI)

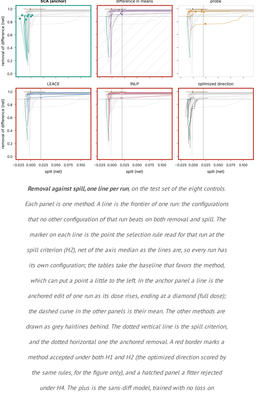
Each hypothesis takes a point from this trade-off, so it comes first. The figure draws the frontier of each run on the test set[*], with the point the selection rule read on it, beside the anchored[*] dose curve of each run. The table after it gives, per method, the two points of the selection rule, the task change[*], and how many runs chose each site.

Test set. The 14,000 held-out cells, with a mean of 0.23 (1,000 per dose); up in the uncharted numbers need to re-computing.

Anchored. The full edit doing so, as published in fixed effects (1).

Task change. The change in mean edit (0.23 to the mean of the controls).

Each mark is not the median drop along the fifteen coordinate axes of the tuning set[*] under the same operator form, the baseline the selection rule uses on the tuning set. We use it here because the two test-set baselines were computed only at the configurations the rule chose.



***Removal against spill, one line per run**, on the test set of the eight controls. Each panel is one method. A line is the frontier of one run: the configurations that no other configuration of that run beats on both removal and spill. The marker on each line is the point the selection rule read for that run at the spill criterion (H2), net of the axis's median as the lines are, so every run has its own configuration; the tables take the baseline that favors the method, which can put a point a little to the left. In the anchor panel a line is the anchored edit of one run as its dose rises, ending at a diamond (full dose); the dashed curve in the other panels is their mean. The other methods are drawn as grey hairlines behind. The dotted vertical line is the spill criterion, and the dotted horizontal one the anchored removal. A red border marks a method accepted under both H1 and H2 (the optimized direction scored by the same rules, for the figure only), and a hatched panel a filter rejected under H4. The plus is the sans-diff model, trained with no loss on `difference`, against the control at the same seed, raw since nothing is edited: seed mean, with bars over the runs. The spill axis is cut at 0.12.*

On every run but one, each standard filter has a frontier that climbs past the anchored removal[*] before it reaches the spill[*] criterion. The configuration it picks at the criterion removes nearly all of `difference`.

Removal. The drop in expected exact match run by the selected edit, also a share of the way from the clean score to the target null.

Spill. The largest drop on the other six under the rule.

On every run, the configuration each filter chose at the criterion was fitted at the example answers and edits there alone; the table below counts the sites. The optimized direction[*], which acts at every position by design, stops about where the anchor does.

Optimized direction. A soft version per edited site. Moved by gradient ascent on the same score as that the ordered answers on `difference` against the control, with the target null while the answers on the other positions are where they were.

method	removal at criterion ↑	raw	spill at matched removal ↓	over criterion	task change	sites (runs)
anchor	0.88 (0.83 to 0.91)	0.96 (0.92 to 0.98)	−0.01 (−0.02 to 0.00)	0 of 8	−0.003	every position, every slice
anchor, fixed labels	0.81 (0.60 to 0.90)	0.89 (0.69 to 0.98)	−0.01 (−0.02 to 0.00)	0 of 8		every position, every slice
difference in means	0.96 (0.91 to 0.98)	0.96 (0.91 to 0.98)	0.00 (−0.01 to 0.01)	0 of 8	0.000	answers B
probe	0.94 (0.77 to 0.98)	0.94 (0.77 to 0.99)	0.00 (−0.03 to 0.10)	1 of 8	0.000	answers B
LEACE	0.97 (0.91 to 0.98)	0.97 (0.91 to 0.99)	−0.01 (−0.06 to 0.00)	0 of 8	0.000	answers B
INLP	0.97 (0.90 to 0.99)	0.97 (0.90 to 0.99)	−0.01 (−0.01 to 0.00)	0 of 8	0.000	answers B
optimized direction	0.89 (0.86 to 0.91)	0.91 (0.91 to 0.95)	0.00 (−0.01 to 0.00), 2 not matched	2 of 8	0.000	pooled B

***The two points of the selection rule**, seed mean and range over eight runs on the test set. Removal at the spill criterion is net; the raw column has no baseline subtracted. The anchor is at full dose, where its spill is that of the edit. A method whose frontier never reaches the anchored removal on a run is not matched there and counts as over the criterion. The sites column counts the runs whose configuration at the criterion was fitted at each site. A greyed row is a filter rejected under H4.*

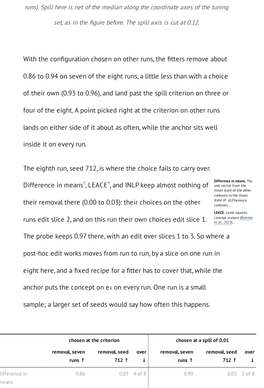
So on this model a direction fitted after training, with every label and a per-run grid, removes more of `difference` within the criterion than the anchored edit does, and spills no more at matched removal. It costs the task nothing, since the model is the control, while the anchored runs score 0.003 lower than the controls. The *fixed-labels*[*] row removes less than the main anchored arm, as that experiment found.

Fixed labels. Holding one label draw fixed for the whole run, so the edit would be lower the answer chair at every side tried, as in the anchor, with the side removing much less of the up.

The edge of the filters comes mostly from fitting and editing at the example answers, a role a real corpus does not label. Restricted to the pooled fits, which need no role labels, the filters lose most of that edge (Without role labels).

One configuration chosen on other runs

Each run above picks its own configuration, among about a thousand, on the model it edits. The anchor had no such choice on these runs: its recipe was settled on earlier seeds and run on these eight as it was. The next figure puts the filters on the same terms. For each run in turn, the selection rule picks a configuration on the other seven, off the frontier of their seed means on the tuning set, and the filter is then fitted on the run left out with that configuration. So the choice is made eight times per method: seed 712 takes the configuration chosen on 713 to 719, seed 713 the one chosen on 712 and 714 to 719, and so on. This analysis was not preregistered.



***Removal against spill with a configuration chosen on the other runs**, on the test set of the eight controls. Each line is the frontier of the seed means of seven runs, one line for each run left out, and the selection rule chose on it at the spill criterion. Each marker is what that configuration, fitted on the run left out, scores there. The anchor panel is as before, and the plus is the sans-diff model against the control at the same seed (seed mean, with bars over the runs). Spill here is net of the median along the coordinate axes of the tuning set, as in the figure before. The spill axis is cut at 0.12.*

With the configuration chosen on other runs, the filters remove about 0.86 to 0.94 on seven of the eight runs, a little less than with a choice of their own (0.93 to 0.96), and land past the spill criterion on three or four of the eight. A point picked right at the criterion on other runs lands on either side of it about as often, while the anchor sits well inside it on every run.

The eighth run, seed 712, is where the choice fails to carry over. Difference in means'[*], LEACE[*], and INLP keep almost nothing of their removal there (0.00 to 0.03); their choices on the other runs edit slice 2, and on this run their own choices edit slice 1. The probe keeps 0.97 there, with an edit over slices 1 to 3. So where a post-hoc edit works moves from run to run, by a slice on one run in eight here, and a fixed recipe for a filter has to cover that, while the anchor puts the concept on e↑ on every run. One run is a small sample; a larger set of seeds would say how often this happens.

Difference in means. The unit vector from the mean of the edited contexts to the mean state of `difference` answers.

LEACE. Least squares concept erasure (Belrose et al., 2023).

	chosen at the criterion			chosen at a spill of 0.01		
	removal, seven runs ↑	removal, seed 712 ↑	over ↓	removal, seven runs ↑	removal, seed 712 ↑	over ↓
difference in means	0.86	0.03	4 of 8	0.90	0.02	5 of 8
probe	0.89	0.97	4 of 8	0.89	0.97	2 of 8
LEACE	0.88	0.00	4 of 8	0.89	0.00	0 of 8
INLP	0.94	0.00	5 of 8	0.94	0.00	2 of 8
optimized direction	0.89	0.79	0 of 8	0.89	0.79	0 of 8

***A configuration chosen on other runs, at the criterion and with a margin:** removal on the test set of the run left out, seed mean over the seven runs other than 712 and on 712 itself, and the runs over the spill criterion out of eight, first with the choice made at the criterion (as in the figure), then at a spill of 0.01 on the other seven. Net of the axis median, as in the figure. A greyed row is a filter rejected under H4.*

A practitioner would choose with some room under the criterion rather than at it. With the choice made at a spill of 0.01 on the other runs, the filters land past the criterion on most three of the eight runs (LEACE on none), and remove about as much as before on the seven. So most of the runs over the criterion come from choosing at the line. The miss on seed 712 stays: the choice made on the other runs still edits a slice where that run does not hold the op. This check was added after the run, after a second reader suggested it.

Spill at matched removal (H1)

What we expect. If we turn each fitter up until it removes as much of `difference` as the anchor does, then on most runs it will cost another op more than the criterion allows.

The rule. For each fitter we count the runs, out of eight, whose spill° at matched removal° is over 0.02 net. A run whose frontier never reaches the anchored° removal counts as over, since that fitter cannot remove as much at any spill. For the anchor we count the runs whose spill at full dose is over 0.02 net. A one-sided Fisher exact test[2] then says whether the fitter count is higher than the anchored count by more than chance would give. A fitter is rejected if its count is five or more and the test gives p below 0.05, and accepted if its count is under five. A count of five or more that the test cannot tell apart from the anchored count (which can happen only if the anchor itself spills on one or more runs; five against one gives p of 0.06) would be outside the plan, and that fitter would be Unresolved. Among the fitters accepted under H4, H1 passes if every fitter with a resolved verdict was rejected, misses if every one was accepted, and is partial otherwise; it is Unresolved if no fitter is accepted under H4 or none has a resolved verdict. The Findings line names any fitter left out or Unresolved. A fitter accepted here would mean that, on this model, a searched direction serves as well as a placed one.

Spill. The largest drop on any other op under the edit....

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Anchored. The full edit along e_1, as published in fixed-effects....

2 Fisher exact test...

The figure shows spill at matched removal for each run, with the anchor at full dose beside the fitters.



***Spill at matched removal per run**, on the test set, net. A small dot per run and the seed mean as the method's marker; the anchor at full dose. The dashed line is the criterion and the side over it is hatched. A run whose frontier never reaches the anchored removal is drawn at the top and counted there. The axis is cut at 0.3.*

method	runs over the criterion	Fisher p	verdict
anchor (full dose)	0 of 8		
difference in means	0 of 8	1.00	accepted (left out under H4)
probe	1 of 8	0.50	accepted
LEACE	0 of 8	1.00	accepted
INLP	0 of 8	1.00	accepted
optimized direction	2 of 8	0.23	accepted (not judged)

***Runs over the criterion at matched removal**, out of eight, with the one-sided Fisher p against the anchored count. A fitter is rejected at five or more runs with p below 0.05, and accepted under five.*

Misses. No standard fitter spills past the criterion on more than one run once it is turned up to the anchored removal, so each is accepted, and H1 misses. Difference in means° is left out of the verdict under H4, and it would have been accepted too. Turned up that far, a fitted direction costs the other ops about as little as the anchored edit does.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

Removal at the spill criterion (H2)

What we expect. If we hold each fitter to the same spill° limit as the anchor, then it will remove less of `difference` than the anchor does.

Spill. The largest drop on any other op under the edit....

The rule. For each fitter we take the difference of seed means between the anchored° removal° (at full dose, or where its own dose curve crosses the criterion if it spills past it) and the removal at the spill criterion of the fitter, with a 95% Welch t interval[3]. A fitter is rejected if the interval lies wholly above zero (the anchor removes more). It is accepted if the interval lies wholly below zero (the fitter removes more), or contains zero and lies within 0.05 of it on both sides (the two remove about as much). An interval that contains zero and reaches further than 0.05 on either side would mean the seeds could not tell, and that fitter would be Unresolved. Among the fitters accepted under H4, H2 passes if every fitter with a resolved verdict was rejected, misses if every one was accepted, and is partial otherwise; it is Unresolved if no fitter is accepted under H4 or none has a resolved verdict. The Findings line names any fitter left out or Unresolved.

Anchored. The full edit along e_1, as published in fixed-effects....

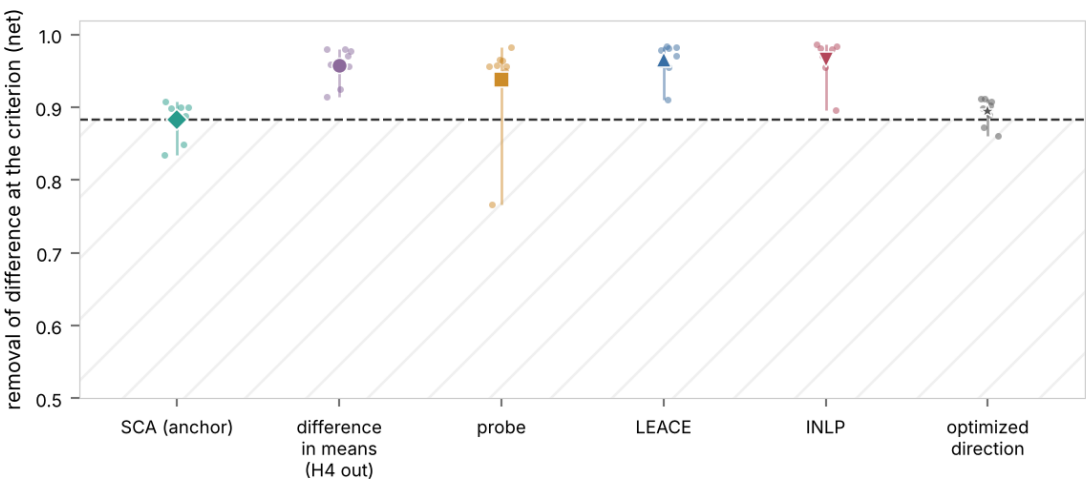
Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

3 Welch t interval...

A small edge for a fitter would get it accepted under this rule, and it would not surprise us. Every run writes the evidence for the op partly off e_1 (the *low-runs*° experiment). So the edit along e_1 leaves a little behind, while a fitted direction can take in both parts.

low-runs. In every run the evidence that an example fits `difference` lies partly on e_1 and partly off it, and the edit along e_1 never touches the off-axis part.

The figure shows removal at the spill criterion for each run.



***Removal at the spill criterion per run**, on the test set, net. A small dot per run and the seed mean as the method's marker; the anchor at full dose. The dashed line is the seed mean of the anchor, and the side under it is hatched.*

method	removal at criterion ↑	seed sd	anchor minus method, 95% interval	verdict
anchor	0.88 (0.83 to 0.91)	0.03		
difference in means	0.96 (0.91 to 0.98)	0.03	−0.10 to −0.05	accepted (left out under H4)
probe	0.94 (0.77 to 0.98)	0.07	−0.12 to 0.00	unresolved
LEACE	0.97 (0.91 to 0.98)	0.02	−0.11 to −0.06	accepted
INLP	0.97 (0.90 to 0.99)	0.03	−0.11 to −0.05	accepted
optimized direction	0.89 (0.86 to 0.91)	0.02	−0.04 to 0.01	accepted (not judged)

***Removal at the spill criterion**, seed mean and range over eight runs, net, with the seed standard deviation and the 95% Welch interval on the anchored seed mean minus the method seed mean. A fitter is rejected if the interval lies above zero, and accepted if it lies below zero or within 0.05 of it on both sides.*

Misses. LEACE° and INLP remove more than the anchor within the criterion, by more than the seeds scatter, so both are accepted. The probe removes more on seven runs but stops far short on one, which widens its interval to span −0.12 to just above zero. That interval contains zero but reaches further than 0.05 from it, so the verdict of the probe is Unresolved.

LEACE. Least-squares concept erasure (Belrose et al., 2023)...

Every resolved verdict is an acceptance, so H2 misses. Difference in means° would have been accepted too, but it is left out under H4.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

We expected a small edge for a fitter, since part of the evidence lies off e_1. The edge here is larger, and it comes mostly from where the fitters act: the rule picks the configurations fitted at the example answers, which edit those positions alone and use role labels a real corpus does not have.

Without role labels repeats the comparison with every fitter restricted to the pooled fits, which is the like-for-like reading for the anchor.

The ceiling against the fitters (H3)

What we expect. If we search for a direction with the scoring in hand, then it does better than every standard rank-1 fitter.

The rule. The comparison takes the rank-1 fitters accepted under H4: difference in means°, the probe, and LEACE°. Removal° by the optimized direction° at the spill° criterion is at least the highest seed mean among them, and its count over the criterion at matched removal (as in H1) is at most the lowest. It passes if both hold, and misses if neither does; one without the other would be outside the plan, and the verdict would be Unresolved, as it would be if no rank-1 fitter is accepted under H4. The comparison is between seed means and counts, with no interval, which is enough for a check that the ceiling is a ceiling.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

LEACE. Least-squares concept erasure (Belrose et al., 2023)...

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were....

Spill. The largest drop on any other op under the edit....

INLP is left out of H3 because it edits a span of two to eight directions, which a rank-1 ceiling does not bound; its two measures are shown beside the others. If INLP does better than the ceiling, the control holds the op across several directions, which E5 looks at from another side.

LEACE is rank-1 but oblique, and the ceiling is a perpendicular edit, so LEACE could in principle beat it too. We keep it in, since LEACE beating the ceiling would say the oblique form matters.

method	removal at criterion ↑	over the criterion at matched removal ↓	role in H3
optimized direction	0.89 (0.86 to 0.91)	2 of 8	ceiling
difference in means	0.96 (0.91 to 0.98)	0 of 8	left out under H4
probe	0.94 (0.77 to 0.98)	1 of 8	compared
LEACE	0.97 (0.91 to 0.98)	0 of 8	compared
INLP	0.97 (0.90 to 0.99)	0 of 8	outside the rule

***The optimized direction beside the standard fitters**, on the controls: removal at the spill criterion (seed mean and range, net) and the runs over the criterion at matched removal, out of eight.*

Misses. The optimized direction removes less than any of the rank-1 fitters it is compared with, and is over the criterion at matched removal on more runs. Both shortfalls come from two runs on which its frontier never reaches the anchored° removal. INLP, outside the rule, does better than it too.

Anchored. The full edit along e_1, as published in fixed-effects....

So it is no ceiling for the fitters here, for the reason seen in H2: the search acts at every position, as it was planned to, while the fitters picked edits at the example answers alone. Against the pooled fits (Without role labels) it does about as well as the best of them.

The ceiling against the anchor (S1)

What we expect. This section predicts nothing: it is a frozen rule that decides between two outcomes, and its badge is Decided, or Unresolved. Either the best rank-1 search on the control does as well as the anchor, or it still falls short.

The rule. The ceiling *matches* the anchor if the 95% Welch interval[3] on the difference of seed means in removal° at the spill° criterion contains zero and lies within 0.05 of it on both sides (as for an accepted fitter in H2), and the Fisher test[2] cannot tell its count over the criterion at matched removal apart from the anchored° count; then the advantage of placing the concept is that no search is needed. The ceiling *falls short* if the interval lies wholly above zero; then the control holds the op in a way no rank-1 edit removes cleanly, which is the stronger claim for the anchor. Any other result (the interval wholly below zero, an interval that contains zero but reaches further than 0.05, or a count the Fisher test does tell apart) would be outside the plan, and the verdict would be Unresolved.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null.…

Spill. The largest drop on any other op under the edit.…

Anchored. The full edit along e_1, as published in fixed-effects.…

	anchor	optimized direction
removal at criterion ↑	0.88 (0.83 to 0.91)	0.89 (0.86 to 0.91)
over the criterion at matched removal ↓	0 of 8	2 of 8
anchor minus optimized, 95% interval	−0.04 to 0.01	
Fisher p (optimized count higher)	0.23	

***The optimized direction beside the anchor**, on the test set, net. The two runs counted over the criterion are runs whose frontier never reaches the anchored removal.*

Matches. The interval on the gap in removal straddles zero and lies within 0.05 of it, and the Fisher test cannot tell the counts apart, so the best rank-1 search acting at every position does about as well as the anchor. By the frozen rule, the advantage of placing the concept is then that no search is needed.

That conclusion takes the search as the reference. H2 shows that fitters acting at the example answers do better still, and we come back to this in the discussion.

The removal here is net of the axis median, the baseline that favors the method. Under the permuted-label baseline alone the optimized direction° would remove far less, and the ceiling would fall short (E6).

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were.…

The fitters find a placed concept (H4)

What we expect. If we fit on the anchored° runs, where training placed the op on e_1, then each standard fitter removes about as much as the edit along e_1 does, within the same spill° limit. A fitter that cannot do that where the concept is known to be is no fair stand-in for its method on the controls.

Anchored. The full edit along e_1, as published in fixed-effects....

Spill. The largest drop on any other op under the edit....

The rule. Restricted to fits on the pooled positions or at the example answers, a fitter is accepted if its removal° at the spill criterion on the anchored runs is no more than 0.1 below what the edit along e_1 removes at full dose on the same runs (seed means), and rejected otherwise. A fitter that finds e_1 but has to stop short of it to stay within the criterion, as INLP may when its extra directions cost spill, is rejected. H4 passes if every one is rejected, and is partial otherwise. On the anchored runs the fitter takes its permuted-label baseline, and the edit along e_1 keeps its published one. The query `=` is left out: the clean alignment° there is near zero in the anchored runs (the *mid-block-edit*° experiment), so a fit at that site is not expected to find e_1.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

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

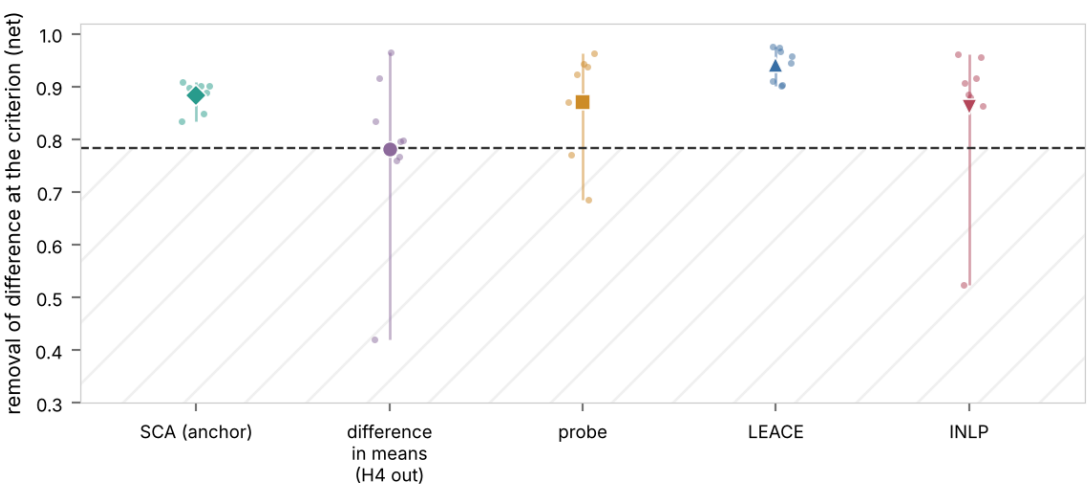
***mid-block-edit*.** Editing the mid-block states as well gains nothing and spills; in passing it shows the query `=` near zero alignment throughout.

Beside the verdict, and not part of it, we report how close each fitted direction is to e_1: the cosine with e_1[4] at each of slices 1 to 3, for the target and site of the selected configuration, given as the seed mean of its absolute value. For INLP it is the largest cosine between e_1 and the span.

4 Cosine with e_1...

The rule for a rejected fitter, frozen here: a fitter rejected under H4 is left out of the verdicts of H1, H2, H3, and S1, its rows stay in the tables of every section, greyed out, and the report names it as a method the comparison could not fairly score. A fitter that cannot remove a placed concept would likely do badly on the controls too, so leaving it out leans toward the fitters, which is the cautious direction for a claim about the anchor.

The figure shows the removal at the spill criterion of each fitter on the anchored runs, restricted to the two sites, beside the edit along e_1.



***Removal at the spill criterion on the anchored runs**, fits on the pooled positions or at the example answers, net of the permuted-label baseline; the edit along e_1 at full dose beside them, net of its published baseline. The dashed line is 0.1 under the seed mean of the edit along e_1, and the side under it is hatched.*

fitter	removal at criterion ↑	raw	cos, slice 1	slice 2	slice 3	verdict
difference in means	0.78 (0.42 to 0.96)	0.97 (0.95 to 0.99)	0.81	0.87	0.75	rejected
probe	0.87 (0.68 to 0.96)	0.89 (0.74 to 0.96)	0.44	0.44	0.25	accepted
LEACE	0.94 (0.90 to 0.98)	0.97 (0.94 to 0.98)	0.81	0.87	0.75	accepted
INLP	0.86 (0.52 to 0.96)	0.97 (0.95 to 0.98)	0.88	0.93	0.79	accepted

***The fitters on the anchored runs**, restricted to the pooled positions and the example answers: removal at the spill criterion (seed mean and range, net of the permuted-label baseline, and raw), and the seed mean of the absolute cosine with e_1 at slices 1 to 3. The edit along e_1 removes 0.88 (0.83 to 0.91); a fitter is accepted at 0.783 or more. The cosines are for the configuration the rule chose over the whole grid, which on most runs is the same site. For LEACE the cosine is that of its covariance vector, which is the difference in means, so its row matches that of difference in means where both chose the same fit.*

Partial. The probe, LEACE°, and INLP each remove within 0.1 of what the edit along e_1 removes, and LEACE more. Difference in means° falls short of the gate by less than 0.001 (0.782 against 0.783; both print as 0.78 in the table), so it is rejected and left out of the other verdicts.

LEACE. Least-squares concept erasure (Belrose et al., 2023)...

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

That rejection is fragile, and its cause is the baseline more than the fitter. Raw, difference in means removes as much as the others; it is the permuted-label baseline that takes it under the gate. Net of the coordinate-axis median instead, the baseline the anchored edit itself is scored against, difference in means removes 0.97, well above the gate, and every fitter would be accepted (removals 0.87 to 0.97). Had it been kept in, H1 and H2 would have accepted it too, so leaving it out changes no verdict. This reading under the axis baseline was added after the run.

On the anchored runs a permuted fit is not neutral. The difference of the means of two random halves of the states points along the directions where the states vary most, and training made e_1 one of those. So a permuted fit removes part of the op too, which puts the net removal of every fitter on these runs below its raw one. E6 compares the baselines on the controls.

The directions of difference in means, LEACE, and INLP on these runs point well toward e_1 at the slice where the fit was made, without lying on it, as the *low-runs*° experiment led us to expect. The probe points much less toward e_1 and still removes the op, likely because a probe turns its direction away from where the states vary most, and training made e_1 one of those directions.

***low-runs*.** In every run the evidence that an example fits `difference` lies partly on e_1 and partly off it, and the edit along e_1 never touches the off-axis part.

Grading, landing, and where the answers go (E2)

At the configuration each method selects for removal° at the spill° criterion, we run the scoring suite of the *fixed-effects*° experiment. The anchored° edit is scored on the same held-out contexts as the other methods.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Spill. The largest drop on any other op under the edit....

fixed-effects. The full scoring suite on the recipe of record at seeds 712 to 719...

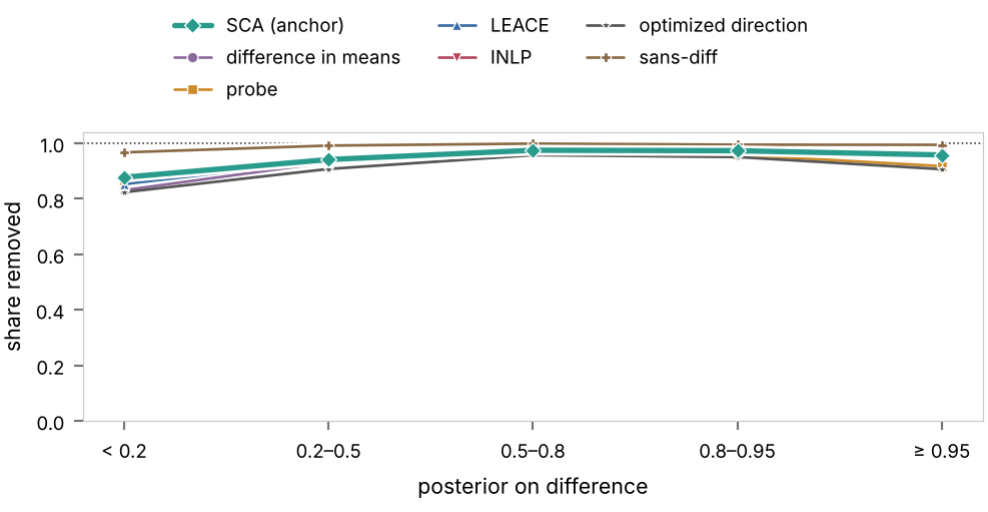
Anchored. The full edit along e_1, as published in fixed-effects....

These measures read each context on its own, where H1 to H4 average the drop over all the contexts of an op. The baselines were taken per op, so they cannot be subtracted context by context, and the drops here are raw.

Grading is the share of the way to the target null° removed in each bin of the posterior on `difference`. Landing is how far the edit moves the answers toward the target null, in total variation°. A matrix of where the edited model puts its answers on `difference` contexts shows where they go.

Target null. Where a removal should land...

Total variation. A distance between two answer distributions...



***The share of the way to the target null removed, by how sure the context is of `difference`**, on the held-out `difference` contexts, raw, seed mean over eight runs; the anchor on the anchored runs, the methods on the controls at the configuration each selected for removal at the spill criterion, and the sans-diff model (no edit) against the control at the same seed.*

	share removed ↑	landing ↑
SCA (anchor)	0.96 (0.92 to 0.98)	0.50 (0.47 to 0.52)
difference in means	0.95 (0.90 to 0.98)	0.55 (0.49 to 0.62)
probe	0.94 (0.76 to 0.99)	0.55 (0.42 to 0.60)
LEACE	0.97 (0.90 to 0.99)	0.55 (0.49 to 0.62)
INLP	0.97 (0.90 to 0.99)	0.54 (0.50 to 0.62)
optimized direction	0.93 (0.91 to 0.95)	0.58 (0.55 to 0.61)

***How much and how far**, on the held-out `difference` contexts, seed mean and range, raw. The share removed is the drop as a share of the way to the target null, over all of them. The landing is the share of the clean distance to the target null, in total variation, that the edit covers. A greyed row is a fitter rejected under H4.*



***Where the answers go on `difference` contexts**, seed mean over eight runs. Each row is a model or an edit; each column the mass on the colors the column op can give for the query operands and `difference` cannot. The `difference` column (the mass on its own answers) is left out of the color scale. Values of at least 0.01 are printed. Rules separate the unedited references (above), the anchored edit, and the post-hoc edits (below).*

Every edit grades its drop by how sure the context is of `difference`, and they all grade about alike. The share of the way to the target null removed stays between 0.82 and 0.98 in every bin, so each edit drops a context by about as much as that context needed `difference`. The share is lowest where the context is least sure, 0.82 to 0.88 below a posterior of 0.2, and highest between 0.5 and 0.8, at 0.96 to 0.98.

It dips a little for the surest contexts, most for the probe and the optimized direction°. The anchor and INLP sit at the top in every bin, and the optimized direction at the bottom.

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were....

The fitters land a little further toward the target null than the anchor: 0.54 to 0.55 of the clean distance, against 0.50. The optimized direction, which is fitted toward the target null, lands furthest at 0.58, about where the *sans-diff model*° lands. The ranges over runs overlap.

sans-diff model. A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

None of the edits gets all the way, and they fall short in the same way. The answers they move off `difference` go to the answers of the other ops in about the proportions the target null gives, a fifth to a quarter each, as do those of the *sans-diff model*, which was trained with no loss on `difference`; and 0.03 to 0.06 stays on the answers of `difference` itself.

The anchor puts a little less than the methods or the target null on the answers of `lighten`, and a little more on those of `sat-hsv`. On grading and on where the answers go, the anchored edit and the post-hoc edits are hard to tell apart.

How many labels a fitter needs (E3)

At the configuration each method selects, we repeat the fit with labels on every `difference` context of the fit set°, then on a fixed half, a tenth, a twenty-fifth, and a fiftieth of them. The last two pair with the anchored° runs of the *fixed-labels*° experiment, which held one draw at those rates. The unlabelled `difference` contexts are counted among the other contexts, as an unlabelled context is for the anchor.

Fit set. 280,000 fresh contexts, 40,000 per op, drawn from the training generator of the run with its replacement op noise…

Anchored. The full edit along e_1, as published in fixed-effects.…

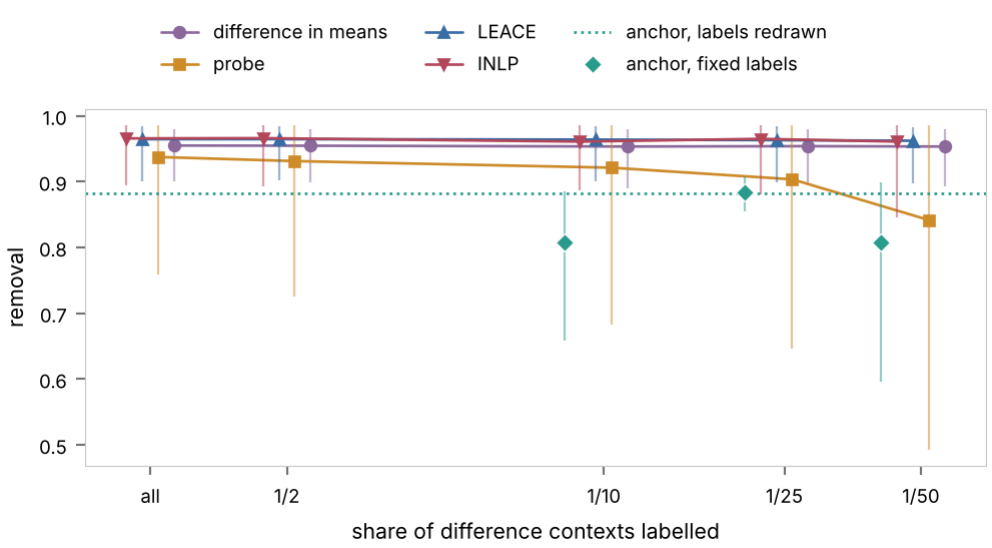
fixed-labels. Holding one label draw fixed for the run, as a real labelled set would be, keeps the anchor at every rate tried, but a few runs end with the edit removing much less of the op.

A fixed fiftieth is 800 contexts: about what the anchor met in one pass over the corpus, and far fewer than it met over training. So in that respect the lowest rate is harsher than what the anchor had. In another it is kinder: the fitters still see every label of the fixed set at once.

The removal° and spill° here are those of the configuration E2 to E5 read, with its own baselines: for each method, the lower-spill configuration of the pair that brackets removal at the spill criterion, which for every fitter is a fit at the example answers. So at every label count the numbers differ slightly from the points of H2, which interpolate between two configurations.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null.…

Spill. The largest drop on any other op under the edit.…



***Removal against the share of `difference` contexts labelled**, on the test set, net of the baselines, seed mean (marker) and range (bar) over the eight controls, at the configuration each fitter selected with every label. The dotted line is the anchored removal at the recipe of record, which draws a fiftieth of the labels afresh on every visit. The diamonds are the anchored runs of fixed-labels, which hold one draw for the whole run.*

	all	1/2	1/10	1/25	1/50
difference in means	0.95 (0)	0.95 (0)	0.95 (0)	0.95 (0)	0.95 (0)
probe	0.94 (0)	0.93 (0)	0.92 (0)	0.90 (0)	0.84 (1)
LEACE	0.96 (0)	0.96 (0)	0.96 (0)	0.96 (0)	0.96 (0)
INLP	0.97 (0)	0.97 (0)	0.96 (1)	0.97 (0)	0.96 (1)
anchor, fixed labels	–	–	0.81	0.88	0.81

***Removal by the share of labels**, seed mean over eight runs, with the number of runs over the spill criterion in brackets. The anchored row is the fixed-labels experiment, at the full dose, where it spills on no run. A greyed row is a fitter rejected under H4.*

Difference in means°, LEACE°, and INLP need few labels: from every label down to a fiftieth, their removal moves by less than 0.02, so a mean or a covariance over 800 labelled contexts already points the right way. The probe drops to 0.84 at a fiftieth, with one run at 0.49 and one over the spill criterion.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts.…

LEACE. Least-squares concept erasure (Belrose et al., 2023)…

The anchor with fixed labels removes 0.81 at a fiftieth, below the three flat fitters at the same rate. So the fitters do not need many labels, and the number of labels is not where the anchor has an edge. The fitters at a fiftieth are fitted at the example answers, so the rate is matched with the anchor and the site is not.

The write against each bound (E4)

The write bound of the anchor does not apply as it stands to a direction that is not an axis, or to an oblique eraser, so each method gets the bound in its own terms: the angle its operator turns the clean states of the other ops through (the 99% quantile, per slice and role).

The write is the angle the operator turns those states through as they arrive, already changed by the edits at the earlier slices. Where the blocks after an edit put the concept back, the method shows a write over its own bound. A site counts as over when the write passes the bound by more than both tolerances of the D2.2 design: a quarter over the bound, and about a degree.

	runs editing several slices	sites over
SCA (anchor)	8	none
difference in means	1	none
probe	2	716, slice 3 at `y`: 17.8° against 13.0°
LEACE	0	none
INLP	2	none
optimized direction	1	717, slice 2 at `?`: 2.2° against 0.9°; 717, slice 2 at `,`: 5.0° against 2.4°

***The write against each bound**, on 2,000 test contexts of the other ops, at the configuration E2 to E5 read. A site is over when the write passes the bound by more than a quarter and by more than about a degree; each is given as seed, slice, and role, with the write and the bound at the 99% quantile. A greyed row is a fitter rejected under H4.*

The write stays within its bound almost everywhere, though for most of the post-hoc edits there is little to measure. An edit at one slice has no edits before it, so its write equals its bound, and most of the configurations read here edit slice 2 alone.

Of the six runs where a method edits several slices, two show a site over: the probe at seed 716 and the optimized direction° at seed 717. The anchored° edit acts at every slice and shows no site over, as in the *fixed-effects*° experiment.

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were.…

Anchored. The full edit along e_1, as published in fixed-effects.…

***fixed-effects*.** The full scoring suite on the recipe of record at seeds 712 to 719…

What a fresh probe can still find (E5)

LEACE° promises that no linear probe can read the concept after its edit, and the anchor promises nothing about that. Here we fit a fresh ridge probe° for the binary target on the edited states of the test set°, per slice and per method, with cross-validation over contexts.

LEACE. Least-squares concept erasure (Belrose et al., 2023)…

Ridge probe. The normalized weight vector of a linear readout of the target, fitted by least squares with an ℓ_2 penalty.…

Test set. The 14,000 held-out contexts of ex-2.2.21, 2,000 per op, so the anchored numbers need no recomputing.

The evidence for the op can be recomputed from the operand colors (the *block1-edit*° experiment), so a method may leave the op readable even where the answer no longer uses it. A probe that still finds the op does not mean the edit failed to remove it from the answer, and a probe that finds nothing does not bound what the edit writes.

***block1-edit*.** Block 1 writes the evidence for the op onto e_1 at the example answers, and a later block reads it at the query `=`.

We report it because it is the measure the erasure literature uses, and the one on which the anchor may come off worse. We expect LEACE to bring the probe near chance at the slices it edits, since that is what it is built to do.

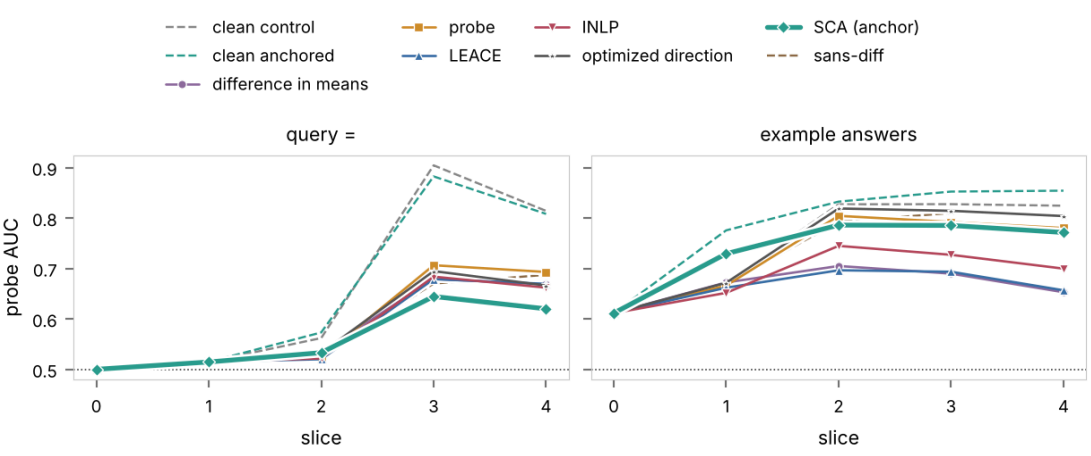
We expect the anchored° edit and the projections to leave the op readable well above chance, since the evidence can be recomputed and part of it lies off any one direction (the *low-runs*° experiment). If the projections also bring the probe near chance, the op is held along fewer directions than that suggests.

Anchored. The full edit along e_1, as published in fixed-effects.…

***low-runs*.** In every run the evidence that an example fits `difference` lies partly on e_1 and partly off it, and the edit along e_1 never touches the off-axis part.

The query `=` is the position whose output predicts the answer the task scores, so a model that uses the op must have it there in some form; the example answers are where block 1 writes the evidence for it (the *block1-edit* experiment). We score the probe by the area under the ROC curve[5] at the query `=` and at the example answers, pooled. It is fitted on half of 7,000 test contexts and scored on the other half.

5 Area under the ROC curve (AUC)…



***What a fresh probe finds after each edit**, the area under the ROC curve of a ridge probe for `difference` on the edited states, by slice, seed mean over eight runs; at the query `=` (left) and the example answers (right). The dashed grey line is the clean control, the dashed teal one the clean anchored model, the dashed brown one the sans-diff model (trained with no loss on `difference`, clean), and the dotted line chance.*

No edit brings the probe near chance. Before any edit, a probe finds the op at the query `=` of slice 3 with an AUC of 0.90 on the control and 0.88 on the anchored model. After the edits it scores 0.64 to 0.71 there, lowest after the anchored edit.

The *sans-diff model*°, trained with no loss on `difference`, leaves 0.67 for a fresh probe at the same site, about what the edits leave. So most of what the probe finds after an edit is there in a model that never learned the op, likely because the operand colors carry it.

***sans-diff model*.** A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

At the example answers the order turns around: the anchored edit leaves the op more readable (0.79 at slice 2) than difference in means°, LEACE, or INLP (0.70 to 0.74), which edit there.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts.…

LEACE does not do what we expected of it, and we do not know why. At the slice and site it edits, mostly at full dose, a fresh probe still reads 0.70, about what difference in means leaves.

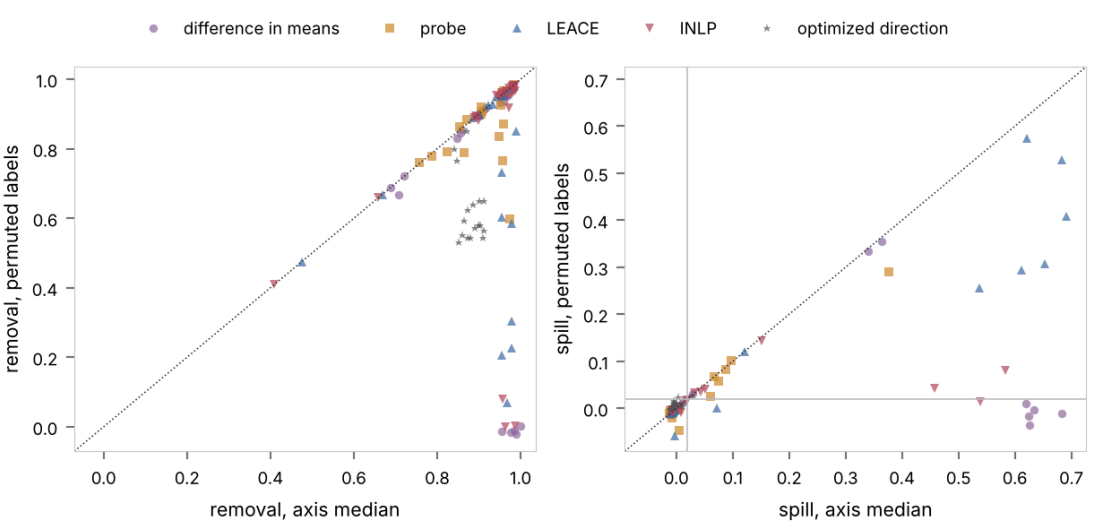
Three candidates fit what we see: **(a)** The model keeps its states at unit length, so every edit ends by scaling the state back to length one. Rescaling by a length that depends on what was taken out can bring some of the op back along every other direction, while the guarantee of LEACE is for the linear edit alone. **(b)** Our LEACE takes a small ridge on a covariance that is close to rank-deficient (some directions barely vary), so it may leave low-variance directions that a fresh probe with a weak penalty can read. **(c)** On two runs it was fitted to the graded target, where the guarantee is for the posterior rather than the binary label the probe reads.

The two baseline rules (E6)

On the controls, we compare the permuted-label baseline with the coordinate-axis median, configuration by configuration, for the configurations scored on the test set°. The Adam optimizer concentrates energy on some axes of a control, so the two can differ.

Test set. The 14,000 held-out contexts of ex-2.2.21, 2,000 per op, so the anchored numbers need no recomputing.

The baseline decides how much of a drop counts as the edit, and each hypothesis takes whichever rule favors the fitters. So if the two rules differ by about the size of a gap that a verdict turns on, that verdict rests on the choice of baseline.



***The two test-set baselines, configuration by configuration**, on the controls: net removal (left) and net spill (right) under the permuted-label baseline against the same under the coordinate-axis median, one mark per run and configuration. On the diagonal the two rules agree; below it, the permuted baseline takes more of the drop as background. The grey lines in the right panel mark the spill criterion on each axis.*

	removal, permuted	removal, axis	over, permuted	over, axis	H1, H2: permuted	H1, H2: axis
difference in means	0.93	0.96	0	0	accepted, accepted	accepted, accepted
probe	0.90	0.94	1	1	accepted, unresolved	accepted, unresolved
LEACE	0.94	0.97	0	0	accepted, accepted	accepted, accepted
INLP	0.94	0.97	0	0	accepted, accepted	accepted, accepted
optimized direction	0.58	0.89	2	2	–	–

***The points of H1 and H2 under each baseline alone**, seed means over eight controls: removal at the spill criterion, and the number of runs over the criterion at the anchored removal, with the verdicts each would give. The hypotheses take the rule that favors the fitter, measure by measure. A greyed row is a fitter rejected under H4.*

The two rules disagree a lot on single configurations and little at the points the verdicts read. At some configurations a permuted-label fit costs every op about as much as the fitted edit does, while an edit along a coordinate axis costs almost nothing. Such a configuration shows little spill° under the permuted rule and a lot under the axis rule, well past the criterion: the marks far below the diagonal in the right panel.

Spill. The largest drop on any other op under the edit....

At the points the verdicts read, the axis rule gives the higher removal° and the permuted rule the lower spill, so the favoring rule combines the two. Taken alone, the permuted rule lowers removal by 0.02 to 0.04 for the fitters and by 0.31 for the optimized direction°, whose permuted searches remove much of the op themselves. No verdict of H1 or H2 changes under either rule.

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null....

Optimized direction. A unit vector per edited slice, fitted by gradient descent on the control so that the edited answers on `difference` contexts move toward the target null while the answers on the other contexts stay where they were....

S1 is the exception. It uses the axis rule for removal, and under the permuted rule alone the optimized direction removes 0.58. The interval on the gap then lies wholly above zero (0.26 to 0.34), and the ceiling would fall short of the anchor.

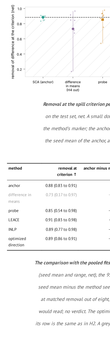
A permuted search is not neutral either, since each context keeps its own target null° under the permutation. So the search still pulls the `difference` contexts that land in the permuted labelled set toward the null, against the keep term on the rest, and it finds a direction that removes much of the op.

Target null. Where a removal should land...

Without role labels

The anchor uses no role labels. Its pull pools over the positions of a labelled context, and its edit acts at every position and slice. A fit at the example answers, the query `=`, or the query operands uses a role, which a real corpus does not label. So the like-for-like comparison restricts every fitter to the pooled fits: the selection rule runs again over the configurations fitted at every position, pooled, on the same runs and with the same baselines. The fitters keep a label on every `difference` context and a grid tuned per run over the target, slices, dose, refit, replacement, and rank, which the anchor does not have.

This reading was not preregistered, so it gives no verdict. The table applies the rules of H1 and H2 as a guide to how far apart the methods are. After it, the figures of E1 to E5 are drawn again with the pooled fits alone.



Removal at the spill criterion per run, with the pooled fits alone, *on the test set, net. A small dot per run and the bead are the method's marker; the anchor at full dose. The dashed line is the seed mean of the anchor, and the slide under it is ratified.*

method	removal at criterion ↑	anchor minus method, 95% interval	over at matched removal	H1, H2 rules would read
anchor	0.88 (0.83 to 0.91)		0 of 8	
`difference in means`	0.73 (0.17 to 0.97)	−0.06 to 0.56	6 of 8	rejected, unresolved
probe	0.81 (0.54 to 0.98)	−0.09 to 0.13	1 of 8	accepted, unresolved
LEACE	0.91 (0.81 to 0.98)	−0.07 to 0.01	2 of 8	accepted, unresolved
INLP	0.89 (0.77 to 0.98)	−0.06 to 0.05	2 of 8	accepted, unresolved
optimized direction	0.89 (0.86 to 0.91)	−0.04 to 0.01	2 of 8 (2 not matched)	accepted, accepted

The comparison with the pooled fits alone: *removal at the spill criterion (seed mean and range, net), the 95% Welch interval on the anchor's seed mean minus the method's seed mean, the runs over the criterion at matched removal out of eight, and what the rules of H1 and H2 would read, no verdict. The optimized direction is always pooled, so its row is the same as in H2. A pooled row is a fitter rejected under H1.*

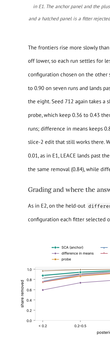
Without role labels the fitters lose most of their edge in removal. `LEACE` removes 0.91 against the anchor's 0.88, with an interval on the gap of −0.07 to 0.01. INLP (0.89) and the probe (0.85) come close on average with a wider spread, and `difference in means` removes 0.73, with one run at 0.17. Read by the rule of H2, every fitter's interval reaches further than 0.05 from zero on one side, so none is a match and none falls short.

At matched removal the order turns around. The anchored edit spills[1] past the criterion on no run; LEACE and INLP do on two runs each, the probe on three, and difference in means on six, which the rule of H1 would reject. Every fitter reaches the anchored removal on every run, so these are runs that spill, not runs that fall short. The Fisher test cannot tell two runs in eight from none (p 0.23), so for LEACE and INLP this is a lean, not a separation.

So in the setting the anchor is built for, the best post-hoc eraser removes about as much as the anchored edit and, even with a grid tuned per run, spills past the criterion on a few more runs. Difference in means, the direction most steering work uses, removes less and spills more.

The trade-off with the pooled fits

As in E1, one frontier per run, now over the pooled fits alone, with the point the selection rule read on it. The optimized direction[1] panel is the same as in E1.

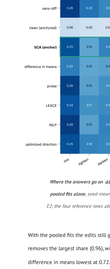


Removal against spill with the pooled fits alone, one line per run, *on the test set of the eight contexts, drawn as in E1. a line is the frontier of one run over the configurations fitted at every position, pooled, and its marker the point the selection rule read on it at the spill criterion, net of the axis median as in E1. The anchor panel and the plus for the sans-diff model are as in E1, and a matched panel is a fitter rejected under H1. The spill axis is cut at 0.12.*

The frontiers rise more slowly than at the example answers and level off lower, so each run settles for less removal at the criterion. A configuration chosen on the other seven runs, as in E1, removes 0.82 to 0.90 on seven runs and lands past the criterion on two to five of the eight. Seed 712 again takes a like 5 at LEACE, INLP, and the probe, which keep 0.56 to 0.63 there with a choice made on the other runs; difference in means keeps 0.80, its choice on the others being a slice-2 edit that still works there. With the choice made at a spill of 0.01, as in E1, LEACE lands past the criterion on one run with about the same removal (0.84), while difference in means fails to 0.56.

Grading and where the answers go with the pooled fits

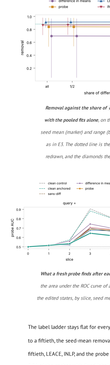
As in E2, on the held-out `difference` contexts, raw, at the configuration each fitter selected over the pooled fits alone.



The share of the way to the target null removed, by how sure the context is of `difference`***, with the pooled fits alone,*** *on the held-out* `difference` *contexts, raw, seed mean over eight runs, drawn as in E2.*

	share removed ↑	landing ↑
SCA (anchor)	0.96 (0.93 to 0.98)	0.50 (0.47 to 0.52)
`difference in means`	0.72 (0.03 to 0.99)	0.40 (0.01 to 0.56)
probe	0.86 (0.51 to 0.99)	0.49 (0.32 to 0.57)
LEACE	0.94 (0.86 to 0.99)	0.54 (0.48 to 0.59)
INLP	0.89 (0.74 to 0.99)	0.50 (0.42 to 0.56)
optimized direction	0.95 (0.91 to 0.97)	0.58 (0.55 to 0.61)

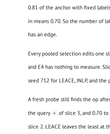
How much and how far with the pooled fits alone, *on the held-out* `difference` *contexts, seed mean and range, raw, as in E2. A greyed row is a fitter rejected under H1.*



Where the answers go on `difference` ***contexts with the pooled fits alone,*** *seed mean over eight runs, drawn as in E2; the four reference rows above the first rule are the same.*

With the pooled fits the edits still grade alike, and the anchored edit removes the largest share (0.96), with LEACE next at 0.94 and difference in means lowest at 0.72. LEACE lands a little further toward the target null[1] than the anchor (0.54 against 0.50), and `difference in means` less far (0.40). The answers go where they went in E2, except that difference in means leaves 0.21 on the answers of `difference` against 0.04 for the anchor and 0.05 for LEACE.

Labels, write, and probe with the pooled fits



Removal against the share of `difference` ***contexts labelled, with the pooled fits alone,*** *on the test set, net of the baselines, seed mean (marked) and range (bar) over the eight controls, drawn as in E3. The dotted line is the anchored removal with labels redrawn, and the diamonds the redrawn rank of fixed labels.*

What a fresh probe finds after each edit, with the pooled fits alone, *the area under the ROC curve of a probe for* `difference` *on the edited states, by slice, seed mean over eight runs, drawn as in E3.*

The label ladder stays flat for every pooled fit: from every label down to a fifteenth, the seed-mean removal moves by less than 0.02. At a fifteenth, LEACE, INLP, and the probe remove 0.83 to 0.91, above the 0.81 of the anchor with fixed labels at the same rate, and difference in means 0.70. So the number of labels is still not where the anchor has an edge.

Every pooled selection edits one slice, so its write equals its bound and E4 has nothing to measure. Slice 2 on most runs, and slice 1 on seed 712 for LEACE, INLP, and the probe.

A fresh probe still finds the op after every edit, as in E1: 0.64 to 0.74 at the query `=` of slice 3, and 0.70 to 0.82 at the example answers of slice 2. LEACE leaves the least at the answers, and the anchored edit the least at the query `=`.

Discussion

The preregistered comparison missed. Given role labels and a grid tuned on each run, every standard fitter that could be judged spilled no more than the anchored° edit at matched removal° and removed more within the spill° criterion, mostly by editing the example answers alone. That is the result of record.

Anchored. The full edit along e_1, as published in fixed-effects….

Removal. The drop in expected exact match on `difference` under the edit, as a share of the way from the clean score to the target null….

Spill. The largest drop on any other op under the edit….

The comparison the anchor was built for restricts the fitters to fits that need no role labels, and it was added after the run. There, LEACE° and INLP remove about as much as the anchor and spill past the criterion on a few more runs, and difference in means° falls behind on both. So on this model the anchored edit matches the best post-hoc eraser we tried, and does better than the direction most steering work uses. We have not shown that it does better than LEACE.

LEACE. Least-squares concept erasure (Belrose et al., 2023)…

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts….

The fitters had help the anchor did not. They saw every label at once, and each run chose among about a thousand configurations scored on that run. The anchor had one configuration, but it came from some thirty experiments at other seeds, so its search happened earlier. The configuration chosen on other runs in E1 is the closest this report comes to comparing the two kinds of search: for the fitters it works on most runs, with the runs over the criterion mostly an effect of choosing at the line, and on one run it misses because that run holds the op a slice earlier. One run in eight is too few to say how often a fixed recipe for a fitter would miss, though it fits the idea that placing the concept saves finding where it sits in each new model.

Placing the concept had costs: an extra training term, and a small drop in task score. What it offers that a post-hoc method lacks is a location chosen before training, and this experiment uses that only in the plainest way, one concept on one axis. Whether a chosen location helps more when several concepts are placed at once, when they are arranged in a designed subspace, or when a concept is harder to find after training, is open, and nothing here bears on it either way.

None of the edits, LEACE included, left a fresh probe at chance (E5). For the anchor that was expected, since we know part of the evidence lies off e_1. LEACE leaves about as much at the slice it edits, though its guarantee covers every linear direction there, so the reason that holds for the anchor does not carry over. E5 lists three candidate causes, and the one we would look at first is the step that puts each edited state back at unit length, which the guarantee does not cover. At the query `=`, the *sans-diff model*°, which never learned the op, leaves a probe about as much as the edits do, so most of what a probe finds there may be what any model of these contexts holds.

sans-diff model. A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

The comparison covers one concept in one small model on a synthetic task, with eight runs per arm, and several readings depend on the choice of baseline (E6). The pooled comparison and the configuration chosen on other runs were added after the run, so they describe what we saw rather than test a prediction.

Method

The fits. Difference in means[*] takes the binary target only. The ridge penalty is fixed at 0.01, the value *ex-2.2.1*[*] used; it is added to the unnormalized Gram matrix of the fit states, so with this many states it barely regularizes. That is fine here: with 280,000 states in 64 dimensions a least-squares probe has little room to overfit, and the penalty is there to keep the solve stable where the covariance is nearly flat. A strong penalty would turn the probe toward the difference in means, which the grid has already. LEACE[*] takes a small ridge on the covariance, since the stream is unit-norm and its covariance is close to rank-deficient. INLP fits the ridge probe[*], projects its direction out of the fit states, and repeats. When several slices are edited, each gets its own fit. The anchor uses e_1 at every slice, but a fitted direction of the control can differ from slice to slice. With sequential refit, the fit at slice ℓ is made on the fit set[*] as it arrives under the edits at the slices before ℓ, which keeps an oblique eraser on the states it will see (*ex-2.2.1* found it extrapolates poorly off them). Mean ablation replaces the coefficient with its mean over the fit contexts of the other ops.

The optimized direction[*]. A unit vector per edited slice, initialized at the difference in means and fitted by Adam for a few hundred steps on the fit set. The loss is the KL divergence[6] from the target null[*] to the edited answers on `difference` contexts, plus a weight times the KL divergence from the clean answers to the edited ones on the other contexts, at full dose. The weight takes four values, and the dose and slice sets are crossed as for the others. Its configurations join the frontier like any other.

Grid size. About 500 configurations for difference in means, 1,000 for the probe and for LEACE, 1,500 for INLP, and 450 for the optimized direction, counting doses; each is one scoring pass, reading the logits at the query `=` (which predicts the answer) only. The whole grid is scored on the tuning set[*], and again on the test set[*] for the figure of E1, where nothing is chosen. A fit is made once per (target, site, slice, refit) and reused across doses and positions.

The baseline. On the tuning set, the median over fifteen coordinate axes other than e_1, drawn once with a fixed seed before the run and shared by every run. On the test set, fifteen permuted fits per configuration, at the bracketing configurations of each point. LEACE is homogeneous in its covariance vector, so a permuted fit is a full-strength oblique erasure along a reference direction. The coordinate-axis median applies the same operator form (slices, positions, dose, replacement) along each of the 63 axes other than e_1; it does not depend on the fitter, so it is computed once per site.

Comparisons. Each method against the anchor as two samples of eight runs: the difference of seed means with a 95% Welch t interval[5], and Fisher's exact test[2] for counts of runs over the criterion.

Smoke test. Before the full run, the whole pipeline runs end to end on two stored runs outside the comparison: the control of *ex-2.2.23* at seed 704 and the `falloff` run at the same seed. Every stage runs: the fits for every method and factor, both baselines, the selection rule, the test scoring, the label ladder, the optimized direction, the measures of E2 to E6, and the tables, figures, and print of this report, with the one pair standing in for the eight (an interval over one run a side means nothing, but the code that computes it runs). The grid is thinned so that each level of each factor is visited at least once, and each stage is timed per pass, which gives the cost of the full run. So a fault anywhere shows up before the costly part. The smoke test also gives two early readings: H4 on the anchored[*] run, so that a fitter that can't find a placed concept shows up before the full run, and where a probe for the op scores best on the control.

The query operands join the sites. Before the freeze, whether the control holds the evidence for the op at the example answers, as the anchored model does, was not known: the *block1-edit*[*] experiment found that the control gathers the operand colors at each example answer in block 1, but it fitted no probe for the op on a control. The smoke test fitted one at every role and slice of the control at seed 704. It scored best at the two operands of the query (an area under the ROC curve of 0.95 at slice 3), above the query `=` (0.92) and the example answers (0.83). The query answer scored higher still (0.99), but it comes after the `=` that predicts the scored answer, so an edit there cannot change it, and it was left out. So the query operands joined the sites fitted, as the plan said they would.

Details fixed with the code. A sequential refit makes the fit at each slice under the full-dose edits at the slices before it, at every position, with zero replacement, and is reused across the doses, positions, and replacements scored with it; mean ablation under a refit uses the mean state of the other ops as it arrives under those edits. A permuted fit takes no sequential refit: it reuses the clean permuted fit at each slice, since a direction with no concept in it barely changes the states the next fit sees. The optimized direction acts at every position; its keep weight takes the values 0.3, 1, 3, and 10, and each search runs 300 Adam steps of 512 contexts at a learning rate of 0.01. With the graded target, the term toward the target null is weighted by the posterior on `difference` and the keep term by the rest. Its mean ablation is scored along the searched direction without searching again, and its permuted baseline uses five searches rather than fifteen, since each is a gradient search. A permuted search is no blank: each context keeps its own target null under the permutation, so the search still pulls the `difference` contexts among its labels toward the null, and the favoring rule takes the axis median for its removal[*].

E2, E4, and E5 read, for each method, the bracketing configuration of removal at the spill[*] criterion that lies within the criterion. The label ladder of E3 leaves out the optimized direction, which reads the target null rather than labels; on the graded target an unlabelled `difference` context has a target of zero. The probe of E5 reads the query `=` and the example answers at each slice, on 7,000 test contexts split in half, and reports the area under the ROC curve. The write of E4 is measured on 2,000 test contexts of the other ops, with a role per position within an equation. The selection of H4 over the pooled fits and the example answers, and the pooled-only selection of H2, were run as a second pass after the main one, on the same tuning scores; it scored the test-set baselines of the configurations it chose that the main pass had not.

Left out. The same comparison for *red* on the D2.1 checkpoints, where *ex-2.2.1* gave the fitters far less, is not run here: this report is about the inferred op.

Cost. No training. Each run takes about 6,000 scoring passes: the grid on the tuning and test sets, the axis baseline on the tuning set, the permuted fits, and the ladder. The anchored runs take the same for H4, so sixteen runs come to about 100,000 passes, plus the state passes over the fit set and for sequential refit, and the optimized direction. The plan estimated eight to ten L4-hours, about $15, with $25 as the ceiling. The run took L40S cards: the full run cost $15.80 and the second pass for the restricted selections $0.64, and with the smoke tests and the speed-ups between them the experiment came to $22.14. The passes added after the paper rounds on the results, E2 to E5 at the pooled selections and the probe on the *sans-diff models*[*], brought the total to $22.72.

Difference in means. The unit vector from the mean state of the other contexts to the mean state of `difference` contexts....

ex-2.2.1. On the D2.1 checkpoints, every fitted direction that removed at least half of `difference` from the control cost the lines of the query `=` in many, a quarter of in up nitude above the unchanged edit.

LEACE. Least squares concept erasure (Belrose et al., 2023)....

Fit set. 280,000 fresh contexts, 40,000 per op, drawn from the training generator of the run with its replacement op none....

Ridge probe. The normalized weight vector of a linear readout of the target, fitted by least squares with an L-penalty....

Optimized direction. A unit vector per edited slice, initialized at the difference in means and fitted by Adam for a few hundred steps on the fit set. The loss is the KL divergence from the target null to the edited answers on `difference` contexts plus a weight times the KL divergence from the clean answers to the edited ones on the other contexts, at full dose. The weight takes four values, and the dose and slice sets are crossed as for the others. Its configurations join the frontier like any other.

Target null. Where a removal should land....

Tuning set. Another 14,000 fresh contexts from the same generator, drawn the same way....

Test set. The 14,000 held-out contexts of *ex-2.2.23*, 1000 per op, on the anchored numbers need no recomputing.

Anchored. The full edit along e_1, as published in fixed-effects....

block1-edit. Block 1 residual-stream edits for the op-slots in, at the example answers, and a later block-readin it at the query `=`....

Removal. The drop in example-set match on `difference` under the edit, as a share of the way from the clean score to the target null....

Spill. The largest drop on any other op under the edit....

sans-diff models. A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.70 for the edited recipe of record, and 0.68 for the same kind of way.

References

ex-2.2.1

On the D2.1 checkpoints, every fitted direction that removed at least half of *red* from the control cost the lines with no red 0.37 or more, an order of magnitude above the anchored edit.

fixed-effects

The full scoring suite on the recipe of record at seeds 712 to 719: removal 0.88 at the full dose, spill below the criterion at every dose, graded drops, a landing of 0.50, and a write within the bound.

fixed confirmation

`τ_ℓ fixed` at fresh seeds 712 to 719 keeps every anchor; it is the recipe of record, and its checkpoints and dose scoring are stored.

anneal confirmation

Trained the controls at seeds 712 to 719 that this comparison fits on, sharing initialization, batches, and corpus with the anchored runs.

block1-edit

Block 1 writes the evidence for the op onto e_1 at the example answers, and a later block reads it at the query `=`.

mid-block-edit

Editing the mid-block states as well gains nothing and spills; in passing it shows the query `=` near zero alignment throughout.

reference-baseline

The median drop along the other coordinate axes of the run replaces the paired control as the baseline, and the published numbers barely move.

sans-diff model

A model trained with no loss on `difference` comes 0.58 of the way from the control to the target null, against 0.50 for the edited recipe of record, and misses in the same kind of way.

fixed-labels

Holding one label draw fixed for the run, as a real labelled set would be, keeps the anchor at every rate tried, but a few runs end with the edit removing much less of the op.

low-runs

In every run the evidence that an example fits `difference` lies partly on e_1 and partly off it, and the edit along e_1 never touches the off-axis part.

ex-2.2.21

The 14,000 held-out contexts, 2,000 per op, that every D2.2 score is taken on.

---

1. Residual stream: the running vector at each position, which every block reads from and adds its output to. ↩

2. Fisher exact test: a test of whether two counts out of the same number of trials differ by more than chance. With the anchor at none of eight, a fitter at four or more of eight gives p below 0.05. ↩↩↩

3. Welch t interval: a 95% confidence interval for the gap between two means that lets the two groups scatter by different amounts. An interval wholly above zero means the eight anchored runs remove more than the eight fitted ones by more than seed-to-seed scatter accounts for. ↩↩↩

4. Cosine with e_1: how closely a direction lines up with e_1, 1 when it lies along e_1 and 0 when it is perpendicular to it. ↩

5. Area under the ROC curve (AUC): the chance that the probe scores a random `difference` context above a random context of another op. 0.5 is chance, 1 a perfect split. ↩

6. KL divergence: how far one probability distribution is from another, zero when they match. ↩