Plan: a fair comparison with directions found after training

Drafted 2026-10-29; revised 2026-10-30 after the first paper round.

Nothing here has run. The D2.2 design lists, among the work before the blog post, a comparison between our edit along the anchored° axis and edits along directions found after training. This page proposes how to run that comparison so that the other methods get a fair go, says what a skeptical reader would object to and how the design answers each objection, and estimates the cost. The experiment itself would be a literate report in its own folder, as usual; this page is the plan to agree on first.

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects….

The question

The anchored° pipeline trains the model with a pull toward e_1 on scarce labelled `difference` contexts, and then removes the op by projecting e_1 out of the residual stream at every slice and position. A post-hoc pipeline trains the same model with no pull (the control), finds a direction for `difference` afterwards from the states and the labels, and removes it through the same scorer. Each method is given the edit it was built for: LEACE° comes with its own affine eraser, and a fitted direction is projected out, which is how the steering and probing literature erases a concept, with a dose° axis added so that it can be scored as the anchored edit is. Replacing the component with its mean over the other contexts, rather than removing it, is the other common form, and it is in the grid too.

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects….

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

Dose. The share γ of the component removed, at 0.25, 0.5, 0.75 and 1, which is the dose axis the anchored edit is scored on.

The question for the post is whether placing the concept during training buys anything that a good search for it afterwards would not. The measurements that answer it are the ones every D2.2 report already uses: how much of the op the edit removes, how much it costs the other ops, whether the drop is graded by what each context needed, where the edited answers land, and whether the blocks after the edit put the concept back. *Ex-2.2.1*° ran a first version of this for *red* on the D2.1 checkpoints, and there every fitted direction that removed at least half of *red* cost the lines with no red about as much. That result came from three control seeds, a token-level concept, and fits that were given little, so it is a reason to expect a gap and no reason to assume one.

***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.

What fair has to mean here

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects….

The anchored° pipeline had two advantages at training time: labels on `difference` contexts, and a term that shaped the whole representation around the axis. The labeller draws a label for one `difference` context in fifty at each visit, but it draws again at every visit, so over 400 epochs about three quarters of the `difference` contexts were labelled at some point. The labels were scarce per step and plentiful in total. A post-hoc method can have neither, and that asymmetry is the thing being measured. Every other asymmetry is ours to remove. So the design follows four rules.

Give the other methods more than we had

Every `difference` context is labelled for the fit, which matches the labels the anchor met over training rather than exceeding them, and a graded target (the posterior on `difference`) is offered beside the binary one. Fits may use role labels (which position is an example answer) that no real corpus has, as a generous arm. Tolerance of scarce labels is one of the claims for the anchor, so the fitters are also run down a ladder of label rates, from every label to a fixed draw of one in fifty, at their selected configuration. Each rung is one draw, held for the fit, so the lowest rung is harsher than what the anchor had.

Tune them, and tune them on data they never score on

Each method gets a grid over where it acts, how much it removes, and how it was fitted, chosen per run on a tuning split by a rule frozen in advance. The anchored recipe was tuned over about thirty experiments at other seeds, and the plan says so.

Score every method the same way, with the same baseline rule

One scorer, one held-out set, one set of measurements, and a reference-direction baseline that each method gets in its own terms.

Score what each method claims, and what it does not

LEACE promises that no linear probe can read the concept afterwards. We report that measure for every method, our own included, since it is the one where the anchor may come off worse.

Conditions

All conditions use the eight controls of the *anneal confirmation*° at seeds 712 to 719, trained with no anchor, and the eight `τ_ℓ fixed` runs of the *fixed confirmation*° at the same seeds, which are the recipe of record. No run is trained. Each post-hoc method is fitted on a control and scored on it, so every method has eight runs, and every comparison with the anchor is a two-sample one with eight runs a side.

***anneal confirmation*.** Trained the controls at seeds 712 to 719 that this comparison fits on.

***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.

Anchored

`τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects. Nothing is refitted or retuned.

Difference in means

The unit vector from the mean state of the other contexts to the mean state of `difference` contexts, one per slice. The direction most steering work uses.

Ridge probe

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

LEACE

Least-squares concept erasure (Belrose et al., 2023): the smallest affine edit, in the covariance-whitened norm, after which no linear probe can read the target. It removes the covariance direction while reading along a whitened one, so it is oblique, and `sca.intervention.Subspace` already carries that split.

Iterated probe (INLP)

The ridge probe fitted, projected out, and fitted again on what is left, for a rank of 2, 4 or 8. It answers the objection that a concept the model holds across several directions can't be removed along one.

Optimized direction

A unit vector per slice, fitted by gradient descent on the control to make the edited answers on `difference` contexts match the target null while the answers on the other contexts stay as they were. It uses the target null itself as its target, which no real pipeline has, so it is a diagnostic ceiling: the best a rank-1 projection edit of this control could do if the direction were searched for with the scoring in hand. COAST (constrained optimization on the activation sphere) is its nearest published relative.

Two calibration arms sit beside these. Each post-hoc fitter is also run on the anchored° model, where it should find e_1 (a cosine near 1 at slices 1 to 3) and remove about what e_1 removes; if it does not, the fitters are the weak link, and the comparison on the controls says little. The null that *ex-2.2.1*° ran, e_1 projected out of the control, is not repeated as an arm: Adam parks energy on e_1 on some control seeds and not others, so that edit is noisy, and the median over the coordinate axes of the control, which the baseline rule already computes, is the steadier null. The *sans-diff model*°, trained with no loss on `difference`, stays the reference for how close to the target null° any model of this size comes.

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects....

***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.

***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

The fitters read states from the control on a fit set°, are tuned on a tuning set°, and are scored on the test set°. The three never share a context.

Fit set. Contexts of the training corpus of that run, every one with its generating op, so every `difference` context carries a label....

Tuning set. A fresh draw of 14,000 contexts, 2,000 per op, from the same generator with the same replacement op noise, with its posterior over ops....

Test set. The 14,000 held-out contexts of ex-2.2.21, which every published D2.2 score uses, so the anchored numbers need no recomputing.

Fit set

Contexts of the training corpus of that run, every one with its generating op, so every `difference` context carries a label. The states are those of the clean pass of the control at every slice and position.

Tuning set

A fresh draw of 14,000 contexts, 2,000 per op, from the same generator with the same replacement op noise, with its posterior over ops. The selection rule below reads only this set, so the test set never chooses anything.

Test set

The 14,000 held-out contexts of ex-2.2.21, which every published D2.2 score uses, so the anchored numbers need no recomputing.

Target

Binary: whether the context was generated by `difference`, which is the label the corpus has and the one the anchor used. Graded: the posterior on `difference` given the examples, for the probe, LEACE and the optimized direction. A label marks the generating op even when the examples point elsewhere, so the binary target carries the same noise the anchor saw; the graded one takes that noise away, which is a kindness the anchor never had.

Sites fitted

One fit per slice. All positions of a context pooled, which is what a context-level label allows; or the example answers only, where the anchored evidence sits (the block1-edit experiment); or the query `=`, where the answer is read. The last two are keyed to roles a real corpus does not label, so they are generous diagnostics rather than candidate recipes.

Slices edited

Each slice on its own, slices 1 to 3 (the slices the pull acts on), and every slice (where the anchored edit acts). One direction per slice: a fitted direction need not be shared across slices, whereas e_1 is shared by construction.

Sequential refit

When several slices are edited, the fit at slice ℓ may be made on the states as they arrive under the edits at the slices before it, rather than on the clean states. Ex-2.2.1 found that an oblique eraser extrapolates on states far from the cloud it was fitted on, so this option keeps each fit on the distribution it will see.

Positions edited

Every position, as the anchored edit does, or the positions the direction was fitted on.

Dose

The share γ of the component removed, at 0.25, 0.5, 0.75 and 1, which is the dose axis the anchored edit is scored on.

What replaces it

Zero (the projection), or the mean coefficient over the contexts of the other ops (mean ablation), which leaves a state where a typical other context sits rather than at the equator.

Label rate

At the selected configuration only, the fit repeated with labels on every `difference` context, then on a fixed draw of one in two, one in ten, and one in fifty. The anchor drew its one in fifty again at each visit, so it met most labels over training, and the lowest rung is harsher than what it had. The ladder says how much of the gap, if there is one, labels could buy a post-hoc method; it is a matched comparison only beside an anchored run trained on one fixed draw.

Selection rule (frozen)

Per method and per run, the grid is scored on the tuning set and its frontier taken: the configurations no other configuration beats on both removal and spill. The hypotheses read two points off the frontier. Removal at the criterion is where the frontier crosses spill 0.02, interpolated between the two frontier configurations that bracket it (the configuration with the least spill, if none is within the criterion). Spill at matched removal is where the frontier crosses the seed mean of the anchored removal, interpolated the same way. A frontier that never reaches that removal has no such point, and the report says so rather than quoting the spill of its strongest configuration, which would count a method short on removal as spilling too; that case is already measured by removal at the criterion, and H1 records it as not reached. The bracketing configurations are then scored on the test set and the same interpolation taken there, so a configuration just across a line does not swing the result. Removal at the criterion is also 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; the trade-off curve is what reviewers ask to see beside it. The whole grid is also scored on the test set, for the trade-off figure, and nothing is chosen from it there.

The grid is about 600 configurations per method. The anchor gets no grid: its one configuration is the published edit. But the anchor has had far more tuning than any grid gives, since the recipe of record was reached through the experiments the index lists, at seeds 704 to 711 and the scout seeds, with its confirmation at fresh seeds 712 to 719 run against checks set before it ran (though not preregistered). That tuning developed the method rather than fitting it to these seeds, much as the post-hoc methods were developed in their own literatures before arriving here with a per-run grid on top. The plan states both budgets so that a reader can weigh them.

Scoring

Every measurement is the one the D2.2 design and the *fixed-effects*° report define, taken through `sca.intervention.apply` on the test set°, so a method that produces a subspace and an operator is scored by the same code as the anchor.

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

Test set. The 14,000 held-out contexts of ex-2.2.21, which every published D2.2 score uses, so the anchored numbers need no recomputing.

Removal, spill, by dose

The drop on `difference` under the edit as a share of the way to the target null, and the largest drop on any other op, at each dose. The selectivity criterion is 0.02.

The trade-off

Removal against spill, one mark per configuration per method on the seed mean, with the anchored edit as a point and its dose curve as a line, as the post-hoc figure of ex-2.2.1 drew it. This is the figure for the post. The hypotheses read two points off it: spill at matched removal, and removal at the criterion.

Grading

The share of the way to the target null removed in each bin of the posterior on `difference`, and the gain on the contexts of other ops by bin, at the selected configuration.

Landing

The share of the distance from the clean answers to the target null that the edit covers, in total variation, and beside it the share of the way from the control to the sans-diff model, on which the anchored edit comes about 0.85 of the way (the sans-diff experiment). The second scale is the one for the post, since no model of this size reaches the target null. The confusion matrices of fixed-effects are drawn for each method at its selected configuration.

The write against the bound

For a direction that is not an axis the turn is no longer arcsin α, and for an oblique eraser it is not a function of one coefficient. So the bound is stated method by method as the turn the operator makes on the clean states of the other ops (the 99% quantile, per slice and role), and the write is the turn it makes on the states as they arrive under the edits at the slices before. Both come from the same clean pass and the same operator, so a method whose blocks put the concept back shows a write over its own bound, as the anchor is checked now.

Linear recoverability

A fresh ridge probe for the binary target, fitted on the edited states of the test set by cross-validation over contexts, per slice. LEACE promises this is at chance; the anchor promises nothing about it. The evidence for the op can be recomputed from the operand colors (block1-edit), so the anchored edit may leave the op readable off e_1 even where the answer no longer uses it. We report it for every method and say what it does and doesn't mean, following the off-axis R^2 item: a readable concept bounds no intervention, and an unreadable one is what LEACE was built to deliver.

Task change

Zero by construction for every post-hoc pipeline, since the model is the control. For the anchored pipeline it is the published change against the mean of the controls. It goes in the table as a row, because it is the price of anchoring, and a fair table shows it.

The baseline rule

Published anchored° scores are net of the median drop under the same edit along each of the other 63 coordinate axes of the run (*reference-baseline*°). A fitted direction is not an axis, so the rule is restated in a form every method can take: the same fitter, on the same states, with the labels permuted across contexts, gives a direction with no concept in it and the same operator (LEACE° is homogeneous in its covariance vector, so a permuted fit is a full-strength oblique erasure of a reference direction); the median drop over fifteen such fits is the baseline for that method and configuration. For e_1 the published rule stands. On the controls both rules are computed, the permuted fits and the coordinate axes of the same run, since Adam may park energy on the axes of a control too. Where the two differ, each hypothesis is judged on the one that favors the post-hoc method for that measure (the smaller baseline for removal, the larger for spill°), and the report shows both. The anchor keeps its published baseline, the median over the other 63 axes of the anchored run, since nothing about e_1 is fitted and a permuted fit has no form for it.

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects….

***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.

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

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

Raw drops are shown beside the net ones throughout. Without the baseline the worst drop under the anchored edit on another op is 0.020 on the seed mean (*mid-block-edit*°), at the criterion rather than below it, and a reader who distrusts any baseline should see that.

***mid-block-edit*.** Editing the mid-block states as well gains nothing and spills; in passing it gives the worst drop on another op under the anchored edit without the baseline, 0.020 on the seed mean, and shows the query `=` near zero alignment throughout.

Comparisons

Each method is compared with the anchor as two samples of eight runs: the difference of seed means with a 95% Welch t interval, as *fixed-effects* does, and Fisher's exact test for the count of runs within the criterion. With eight runs a side, gaps under about 0.05 in removal are hard to see; if *ex-2.2.1*° is any guide the gaps here are several times that, but the op case is new and the controls have not been probed for where they hold it.

***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.

Reading order

In the report, the trade-off° figure comes first, since it gives the reader the shape of the result that the hypotheses then read points off, and the hypotheses follow it in the order above. The label ladder, the confusion matrices and the rest of the exploratory sections come after.

The trade-off. Removal against spill, one mark per configuration per method on the seed mean, with the anchored edit as a point and its dose curve as a line, as the post-hoc figure of ex-2.2.1 drew it….

What a skeptical reader would say

Each objection, and where the design answers it.

The fits were starved of labels

They get every label, and a graded target the anchor never had (What each method gets).

You picked weak fitters

Three standard fitters, a rank-k one, and a direction optimized for the scoring itself, which is the ceiling for rank-1 projection edits on this control. If the ceiling matches the anchor, the gap was in the search, and that is reported as such (Conditions).

You tuned yours and not theirs

Each method is tuned per run on its own tuning split, under a frozen rule, over a grid the plan lists in full; the anchor arrives as published and its development budget is stated (What each method gets).

Your baseline flatters you

Each method gets a reference-direction baseline in its own terms, the two rules are checked against each other on the controls, and raw numbers sit beside net ones (The baseline rule).

You edited at the slices that suit you

Every method may edit any slice, the pulled slices, or all slices, with a refit on arriving states when it edits several, and the fitted sites include the one where the answer is read (What each method gets).

Your scoring suite was built for your method

Linear recoverability, the measure LEACE was built for, is scored for every method and reported where the anchor may lose (Scoring).

The criterion of 0.02 is yours

The trade-off figure shows every configuration, so a reader can draw the line where they like (Scoring).

Anchoring changed the model, so this compares two models

Yes: the comparison is between pipelines, and the anchored pipeline pays a task cost, which the table shows. The calibration arm separates where the concept sits from how it was found: the post-hoc fitters on the anchored model show whether they can find a concept that is there to find (Conditions).

A toy model with 64 dimensions and one op

Out of scope here, as for every D2.2 result; the post says so. The field now has post-hoc methods with predicted side effects (COAST; pre-intervention prediction from SAE geometry), and the distinctive claim for the anchor is that its bound is set by construction rather than estimated from emergent geometry; this comparison supports that claim on this model and no other.

Methods considered and left out

Activation steering (adding a vector) is not a removal and has no dose° axis that maps onto ours. Fine-tuning methods (RMU and its relatives, gradient ascent) change the weights, so their write has no bound to compare, and the design defers them to D2.3 beside the permanent weight ablation. Gradient routing and SGTM are training-time methods and belong in the future-work comparison the design names. Sparse autoencoder steering needs a dictionary trained on the stream; the strong version (output-score feature selection) is a project of its own and would come after this tier if a reviewer asks for it. Nonlinear erasure is left out for the same reason, and because the anchored° edit is linear too.

Dose. The share γ of the component removed, at 0.25, 0.5, 0.75 and 1, which is the dose axis the anchored edit is scored on.

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects….

Shape

Preregistered, with few hypotheses and the rest exploratory. A methods comparison for the post is a claim that has to stand without the reader trusting how we chose to look at the data, and the frozen selection rule is what makes the tuning fair rather than ours to adjust. Each hypothesis names its contrary case and what that case would mean.

H1, spill at matched removal

Where the frontier of each standard fitter crosses the seed mean of the anchored removal (or at its least-spill end, if the whole frontier removes more), the fitter (difference in means, probe, LEACE, INLP) spills past the criterion on most runs, by Fisher's test against the count for the anchor. A run of a method (one seed) whose frontier never reaches that removal counts as not matched, which is the stronger shortfall, and is recorded as such. Contrary: a fitter matches the anchor on both, and the post says that a searched direction serves as well as a placed one on this model.

H2, removal at the spill criterion

Where its frontier crosses the spill criterion, each standard fitter removes less than the anchor, with the 95% Welch interval on the difference of seed means clear of zero. Contrary as for H1.

H3, the ceiling

The optimized direction comes closer to the anchor than any standard fitter on both measures. If it matches the anchor, the advantage of placing the concept is that no search is needed, and the post is framed that way; if it falls short too, the geometry of the control holds the op in a way no rank-1 edit removes cleanly, which is the stronger claim.

H4, the fitters work where the concept is placed

On the anchored model, each standard fitter fitted on the pooled positions or on the example answers finds e_1 at slices 1 to 3 (cosine at least 0.9 at the selected configuration, on the seed mean over the eight runs, with the runs listed) and removes within 0.1 of what e_1 removes. The evidence for the op sits on e_1 at the example answers; at the query `=` the clean alignment is near zero (the mid-block-edit experiment), so a fit at that site is not expected to find it and is not held to this. Contrary: a fitter fails where the concept is linearly there. The rule, frozen here: a fitter that fails H4 is left out of the verdicts of H1 to H3, its rows stay in the tables marked as such, and the post names it as a method the comparison could not fairly score.

E1 to E6

The trade-off figure; grading, landing and the confusion matrices at the selected configurations; the label ladder; the write of each method against its own bound; linear recoverability for every method; and the two baseline rules against each other on the controls.

Cost

No training. Each (method, configuration, run) is one scoring pass over 14,000 contexts of a small model, reading the logits at the query answer only; the *reference-baseline*° scoring ran about 500 such passes per checkpoint in a few minutes on one L4. The grid is about 600 configurations per method on the tuning set°, the bracketing ones at four doses° on the test set°, fifteen permuted fits for each, and the label ladder, around 5,000 passes per run; a configuration with sequential refit° also needs one state pass per slice it refits, which adds about a third. Sixteen runs (eight controls, eight anchored° for calibration) come to about 100,000 passes, or eight to ten hours on one L4, run as eight containers in parallel. The optimized direction° adds a few hundred gradient steps over the tuning set per setting of its weight, under an L4-hour in all. State extraction for the fits is one full pass per run. Budget: $15, with $25 as the ceiling.

A smoke test runs first, on stored runs outside the comparison: one control of ex-2.2.23 at seed 704 and one `falloff` run at the same seed. It checks the pipeline end to end and H4° on the anchored run, so that a fitter that can't find a placed concept is found before the full run, and the full run is spent once. Its timing gives the cost of the full run, which is reported before that run starts.

Development time: the fitters, the operator and the scorer exist (`sca.intervention`, the post-hoc tier of *ex-2.2.1*°, the dose scoring of ex-2.2.22); new code is the grid, the selection rule, the permuted-label baseline, the per-method bound, the recoverability probe and the optimized direction, about a day of thread time, with the runs taking an afternoon. The report, two review rounds and a paper round are the larger part, as usual.

***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.

Tuning set. A fresh draw of 14,000 contexts, 2,000 per op, from the same generator with the same replacement op noise, with its posterior over ops....

Dose. The share γ of the component removed, at 0.25, 0.5, 0.75 and 1, which is the dose axis the anchored edit is scored on.

Test set. The 14,000 held-out contexts of ex-2.2.21, which every published D2.2 score uses, so the anchored numbers need no recomputing.

Sequential refit. When several slices are edited, the fit at slice ℓ may be made on the states as they arrive under the edits at the slices before it, rather than on the clean states....

Anchored. `τ_ℓ fixed` under the full edit along e_1, as published in fixed-effects....

Optimized direction. A unit vector per slice, fitted by gradient descent on the control to make the edited answers on `difference` contexts match the target null while the answers on the other contexts stay as they were....

H4, the fitters work where the concept is placed. On the anchored model, each standard fitter fitted on the pooled positions or on the example answers finds e_1 at slices 1 to 3 (cosine at least 0.9 at the selected configuration, on the seed mean over the eight runs, with the runs listed) and removes within 0.1 of what e_1 removes....

***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.

Open decisions

Settled in the first paper round: the shape is preregistered; the optimized direction° is included, labelled as a ceiling; the role-keyed fits are included as a generous diagnostic, said so in the report; and the hypotheses are judged net of the reference-direction baseline of each method, with raw numbers beside.

Optimized direction. A unit vector per slice, fitted by gradient descent on the control to make the edited answers on `difference` contexts match the target null while the answers on the other contexts stay as they were....

Open decision

Whether the control holds the evidence for the op at the example answers as the anchored° model does is not known. The *block1-edit*° experiment found that the control gathers the operand colors at each example answer in block 1 as the anchored runs do, but it fitted no probe for the op itself on a control. The role-keyed fits are a generous arm either way, and the smoke test on a control would show where a probe for the op reads best; if it reads elsewhere, that site joins the sites fitted°.

Anchored. τ_ℓ `fixed` under the full edit along e_1, as published in fixed-effects....

***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 `=`.

Sites fitted. One fit per slice....

Open decision

Whether to rerun the same comparison for *red* on the D2.1 checkpoints, where the tier of *ex-2.2.1*° gave the fitters far less. The code would carry over as it is. Proposed: not for the D2.2 post, which is about the inferred op; a backlog item if the post or the paper shows the *red* result beside it.

***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.

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.

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 gives the worst drop on another op under the anchored edit without the baseline, 0.020 on the seed mean, and 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.

ex-2.2.21

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