Ex 2.2.14: anchoring an operation

> Every anchor so far has held a property of one token: how red a color is. Here we anchor an *operation*, `difference`, on the axis *red* used to have. It lands at twice the margin *red* reached, the task does not move, and every other op of the table anchors the same way.

There is no *red* anchor beside it. An op is a step up in abstraction from a color: a color is defined by what one token looks like, an op by what it does to a pair of operands. The abstraction made no difference to the anchor, because the op is named by one token and the pull can read that token's embedding.

This was a few-seed smoke test, run before the many-seed equivalence experiment spends its budget, so the equivalence experiment can go ahead. It also asked whether the blocks carry the op forward to where the answer is computed: they carry a little of it to `=` and none to the answer position.

Findings

- The task survives (H1) — **pass.** No op's held-out exact match moves by more than 0.005, a quarter of the gate; `difference` itself moves by 0.000.
- The op lands and holds (H2) — **pass.** The op margin is 0.86, 2.0× what *red* reached, and every seed holds it through the anneal. It saturates by epoch 3: the op word's embedding sits on the axis, and the margin reads that position.
- The blocks carry it to the use sites (H3) — **miss**, on one site of two. With the op word alone pulled, the contrast at `=` clears the floor (+0.100 over the control) and the contrast at the answer does not (-0.013).
- Containment, reported without a gate — the other ten ops sit at 0.05 at the op position, +0.08 over the control and far under *red*'s 0.264. Post hoc: the first operand leans toward e_1 on the primary alone, a candidate mechanism for that rise.

The rule for the follow-up: the equivalence experiment anchors `difference`. All 10 ops of the sweep qualify too, with margins from 0.85 to 0.89, so the data do not separate them and the analytical choice stands.

How to read this draft

The op, the three predictions, the containment measure, and the rule for the follow-up were fixed before any run, at commit `2497993`. Everything after that commit is either results filled into their sections or exploratory work, marked as post hoc.

The D2.2 design asks for this smoke test before the equivalence read, scored on the alignment and task gates alone. The probe scan it names runs here as a description.

Why this experiment

D2.2 asks whether an anchor can hold more than a token's identity at a labelled site. *Red* is a property of the token at a known position. An operation is a property of the computation: named at the op word, and used at `=` and after, deeper in the stack. It is a more abstract concept than a color, and the first rung of a ladder toward concepts in natural language.

If anchoring an op works, the suppression experiment that follows can ask the deliverable's central question: whether removing the op from the axis removes the ability to perform it, selectively and by a bounded amount.

Ex-2.2.11 fixed the grammar and the recipe, and ex-2.2.13 confirmed the recipe at fresh seeds with the weight unchanged. One issue stays open: the projection leaves a quarter of the red-dependent answers of `hue-hsv` in place. That belongs to the *red* anchor, which this model does not have.

The label is new in kind: until now a color decided which lines were labelled, and here the op word decides. The pull covers the whole line, so the label says which lines carry the concept but not where on the line it sits.

On the primary, no alignment measurement can tell an anchored op from an anchored token. The pull puts the op word's embedding on the axis by construction, and asks for the axis at `=` and at the answer too. So these measurements only say whether the pull landed.

One arm pulls the op word alone, so any alignment it shows at the *use sites* (`=` and the answer position, where the op is applied rather than named) is something the pull did not ask for. Whether the anchor captured the operation itself is a question for suppression, and we don't ask it here.

Conditions

condition	anchored op	seeds	role
`anchor-diff`	`difference`	5	the primary: the op's lines draw at 0.02, the red labeller's share; H1 and H2 are scored on it
`anchor-diff-opword`	`difference`	5	arm: the pull covers the op word alone; H3 is scored on it
`anchor-diff-full`	`difference`	5	arm: every line of the op draws
`anchor-diff-noisy`	`difference`	5	arm: the primary's labeller with a fifth of its labels moved onto other ops' lines
`anchor-<op>` ×10	each other op of table A+	3	the sweep: the same reads, reported as a description
`control`	none	5	ex-2.2.11's un-anchored control, served from the store; the task reference and the alignment baseline

The op. `difference` was chosen analytically. On three quarters of its lines no other op in the table gives the same answer (its *op-relevance*),[1] the highest of any commutative op. It is commutative, so operand order plays no part in its read.

It is total on the grid, so each line has one answer, and the control learns it to near-perfect held-out accuracy. The ops that round stochastically are capped well below that.

The labeller. A line whose op word is the anchored op gets a label at a rate of 0.02. The pull covers all six positions of a labelled line, as the handover's whole-line labeller does.

That labels a fifth of a percent of the corpus, the red labeller's share, so each labelled line gets the same pull a red line got, and the op differs from *red* in kind rather than in label share.

The primary uses this sparse rate because it resembles the labels the method will have further up the ladder: for an abstract concept in natural language, labels will be scarce and sometimes wrong. One arm below labels every line of the op instead. We record the label share for each run.

The arms. All three run at the primary's seeds. The op-word arm pulls only the op word, at the primary rate, so the use sites fall outside the mask: any contrast at `=` or the answer at the final slice was carried there by the blocks without the pull asking for it. H3 is scored on this arm.

Pulling one position instead of six also makes each of this arm's pulls about six times stronger, because the anchor term is normalized by the mask. So we report its contrast at the op position beside the primary's.

The every-line arm labels every line of the anchored op, a tenth of the corpus. If its margin agrees with the primary's, label share does not set the margin; if they differ, the ratio of the two margins shows what pulling every line of a categorical concept gains.

The noisy arm moves a fifth of the primary's labels onto lines of the other ops, so the labelled share of the corpus stays the primary's and the total pull with it. It shows what a labeller of that precision costs the margin and the task. Neither the every-line arm nor the noisy arm enters a hypothesis or the rule.

The seeds. 5 for the primary and each arm, and 3 for each op of the sweep, all fresh. The control is served from the store at its own five seeds; comparisons against it are between seed means, which absorbs the unpaired seed sets.

Everything else is unchanged from ex-2.2.11: table A+, the stochastic corpus, the untied readout, $\lambda_a = 0.1$ annealed over the last tenth of training, $\tau = 0.1$, the anti-subspace schedule, and 50 epochs at d64-L4. *Red* is not anchored: e_1 belongs to the op.

Glossary

Op margin

The line margin (m_line) of every experiment since ex-2.1.10, with the anchored op's lines as the labelled group: per slice, the mean alignment with e_1 over the anchored op's lines minus the mean over all eleven ops' probe lines, at the span role where that gap is largest, then averaged over every slice, the embedding included. Taking the largest role per slice is what keeps a clean `=` embedding from costing anything: at the embedding slice the op word's own role carries the gap. The quantity the anchor term optimizes, so it is a check that the treatment landed.

Contrast at a site

At one position and slice, the mean cosine with e_1 (how closely the state points along e_1, from −1 to 1, ignoring length) over the anchored op's lines minus the mean over the other ops' lines. At the op position it is the op word's own placement; at `=` and the answer it is what the blocks carried there, since those tokens are the same on every line.

Use sites

`=` and the answer position: where the op is applied rather than named.

Containment

The mean alignment with e_1 at the op position over the other ten ops' lines: how much the ops that were not anchored drifted toward the axis. The categorical analogue of $\bar{\alpha}$ at op1.

Op-identity R^2

How linearly readable the op is at one site, anchored against control: the held-out R^2 (share of variance explained; 1 is perfect) of a ridge probe (a linear regression with a penalty on large weights) from the state to the one-hot of the op word.

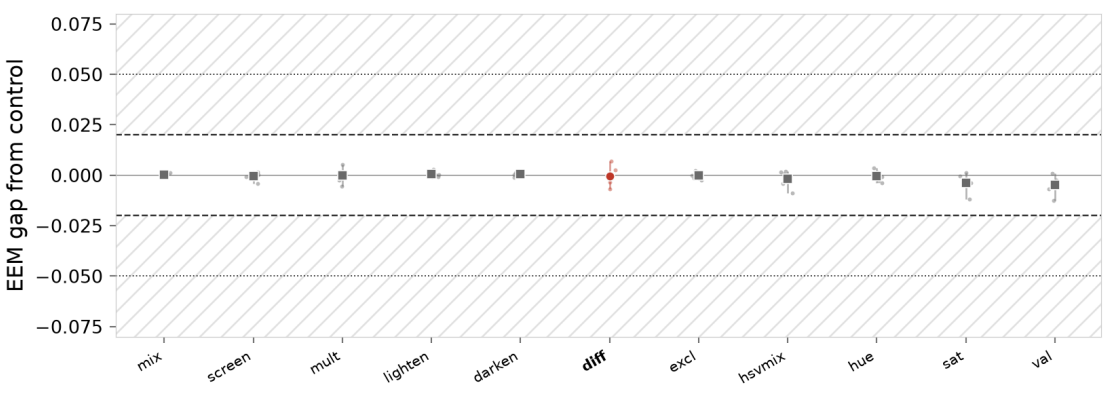
The task survives (H1)

✓ Pass

What we expect. Anchoring `difference` costs the task nothing we can measure: for each of the eleven ops, the primary's seed-mean held-out expected exact match is within 0.02 of the control's. It holds in part when every op is within 0.05.

A miss on the anchored op alone would say the pull on its lines competes with producing its answer. The whole-line span makes that possible for the first time on a categorical concept. A miss spread over the table would say the pull disturbs lines it never touches, at a label share of a fifth of a percent.

The design's risk table names this as the first thing an abstract anchor might cost. The red anchor on the same recipe cost nothing on any op at twenty seeds. The metric is the control's own, expected exact match on the grid; an RGB-distance readout beside it is an open item and is not scored here.



The task gap per op. *Each column is one op: the primary's 5 seeds as faint dots, a thin bar over their range, and the seed mean as the large mark, each as a difference from the control's seed mean on the same op. The anchored op, `difference`, is drawn in the primary's ink and the others in grey. Dashed lines mark the gate at ±0.02, dotted lines the partial level at ±0.05, hatched outside.*

op	primary EEM ↑	control EEM	gap	|gap| ≤ 0.02
`mix`	0.429	0.429	+0.000	**yes**
`screen`	0.547	0.547	-0.000	**yes**
`multiply`	0.534	0.534	-0.000	**yes**
`lighten`	0.992	0.991	+0.000	**yes**
`darken`	0.993	0.993	+0.000	**yes**
`difference`	0.975	0.975	-0.000	**yes**
`exclusion`	0.563	0.563	-0.000	**yes**
`hsvmix`	0.335	0.336	-0.002	**yes**
`hue-hsv`	0.819	0.819	-0.000	**yes**
`sat-hsv`	0.686	0.690	-0.004	**yes**
`value-hsv`	0.710	0.714	-0.005	**yes**

Held-out expected exact match per op: the primary's seed mean over 5 seeds and the control's over 5, their difference, and whether it is inside the gate. H1: pass.

Pass

The largest seed-mean gap is -0.005, on `val`, a quarter of the 0.02 gate, and the anchored op's own gap is 0.000. The whole-line pull on `difference` lines does not compete with producing their answer, and the other ten ops' lines are untouched.

The op lands and holds (H2)

✓ Pass

What we expect. This is a manipulation check: the op margin is the quantity the anchor term optimizes, so it says whether the pull landed, not whether the op was captured.

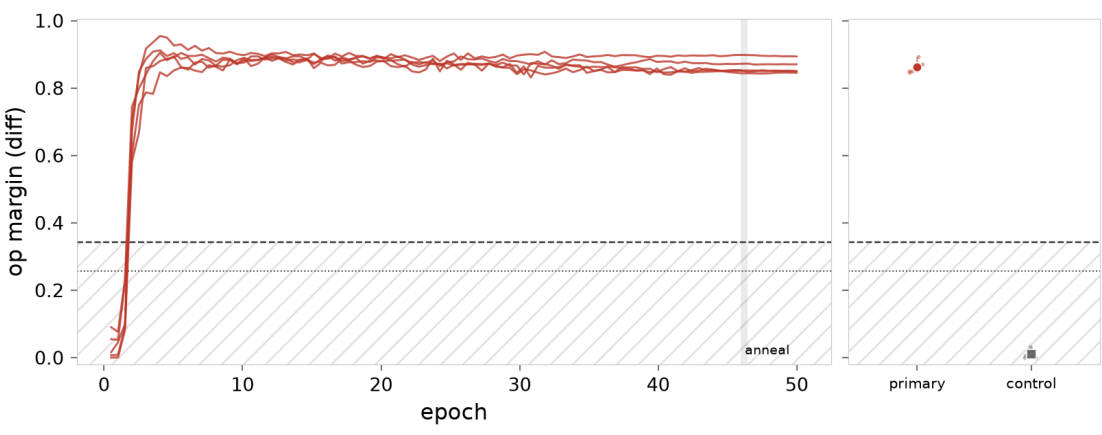
On the seed mean, the op margin on the primary reaches at least 80% of what *red* reached under the same recipe (0.4286 on the `mix` lines in ex-2.2.11), and holds in part from 60%. The margin is also retained: every run whose margin reaches 0.2 when the anchor weight starts to anneal ends training at 0.8 of that value.

The two margins are computed the same way (a gap in mean cosine between a labelled group and the pool), so they are comparable in scale. But their groups and baselines differ in ways the method describes, which pull in opposite directions, so the bar is rough; the equivalence experiment sets its own from what this one measures.

A margin far under the bar with the task intact would say the op's lines are harder to pull together than red's: a first sign that a categorical concept spread over eleven contexts wants a different weight.

A margin that falls through the anneal would be the retention failure ex-2.2.9 saw on `handover` alone, and would point to the anneal as the thing to look at before the equivalence experiment.

The learning rate is low over the anneal, so a tighter share than 0.8 would be defensible. But one `handover` seed in twenty ended under it in ex-2.2.11, so we keep the bar where a known failure sits and report the per-seed values.



The op margin over training and at the end. *Left: the margin of `difference` on the trajectory probe lines, recorded every 50 epochs, one line per primary seed; the shaded band runs from the earliest to the latest anneal start. Right: the end-of-training margin on the full probe sets, per seed with the seed mean as the large mark, beside the control scored on the same op. The dashed line is the bar, 80% of red's 0.4286; the dotted line is the partial level at 60%; hatched below.*

seed	anneal start (epoch)	margin at anneal	margin at end	retention ↑	end margin (eval) ↑
0	45.9596	0.853	0.850	**0.996**	**0.850**
1	45.9596	0.851	0.849	**0.998**	**0.849**
2	45.9596	0.871	0.870	**0.999**	**0.870**
3	45.9596	0.898	0.894	**0.995**	**0.894**
4	45.9596	0.846	0.845	**0.999**	**0.845**
mean		0.864	0.862	0.998	**0.862**

The op margin per primary seed: on the trajectory at the start of the anneal and at the end, their ratio (retention, gated at 0.8 for runs at or above 0.2 at the anneal start), and the end-of-training margin on the full probe sets, against the bar of 0.343 (80% of 0.4286); bold passes. Seed-mean margin over red's: 2.01. Margin: pass; retention: pass.

> **Pass**
>
> The seed-mean op margin is 0.862, 2.01× *red*'s 0.4286 against a bar of 0.343. It reaches about 0.8 by epoch 3 and barely moves after: the op word's embedding sits at a cosine near 1 with e_1 on every slice, and the margin takes the largest role per slice, so the op position sets it. The bar was set for a graded concept spread over a line, and a categorical one named by a single token clears it with room to spare. A margin this saturated tells ops and arms apart poorly, which the sweep and the arms below bear out.

> **Pass**
>
> Every seed ends training within 0.5% of its margin at the anneal start. Nothing gives way as the weight comes down.

The blocks carry it to the use sites (H3)

✗ Miss

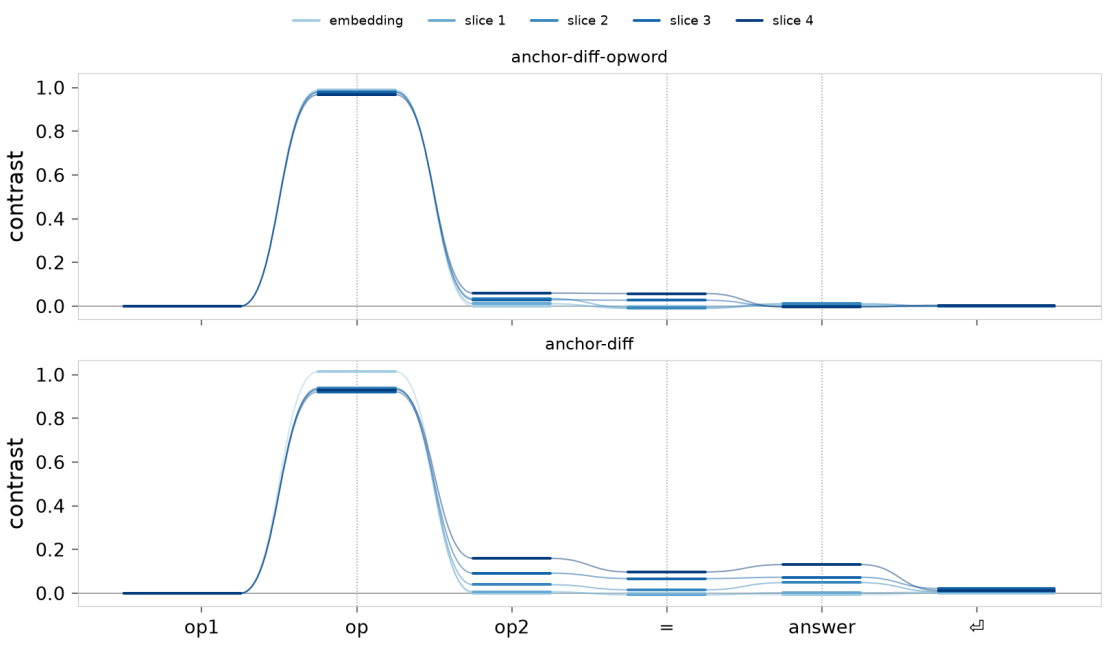
What we expect. Some of the op reaches the use sites. On the op-word arm at the final slice, the contrast at `=` and the contrast at the answer position are each at least 0.05 above the control's, on the seed mean. The ratio of each to the contrast at the op position is reported with no bar.

No decision depends on this prediction, the least certain of the three. We have seen the model put the answer color at the use sites; this measures whether it also holds a copy of the op there, or only what the op produced. The floor asks only for a clearly positive contrast.

At slice 0 (the embedding), the contrast at the op position is the op word's embedding projected on e_1, which the pull puts there by construction. At the use sites the token (`=` or the answer) is the same on every line, so the slice-0 contrast is zero and any later contrast came through attention, which on this arm nothing requested.

A pass would mean that once the op word is on the axis, the blocks carry some of the op to where it is used, which an intervention at the op word would rely on. A miss would mean that, at this weight, the anchor marks the op but the blocks do not carry it to the use sites; it would not mean the model computes the op somewhere else.

On the primary the whole-line pull covers the use sites too, so the contrast there is part of what the treatment optimizes and we expect a pass from the method alone. We report it beside the op-word arm as a manipulation check.



Contrast over the line, per slice. *The seed-mean contrast: mean cosine with e_1 on the `difference` probe lines minus the mean over the other ten ops' lines, at each of the six positions. One series per residual-stream slice, light for the embedding to dark for the final slice. Top, the op-word arm, whose pull covers the op word alone; bottom, the primary, whose pull covers the whole line. Dotted verticals mark the op, `=` and answer positions.*

seed	op position	=	answer	= / op	answer / op
0	0.961	0.027	-0.050	0.03	-0.05
1	0.976	0.058	0.021	0.06	0.02
2	0.978	0.096	0.005	0.10	0.00
3	0.958	0.091	0.016	0.09	0.02
4	0.964	0.016	-0.010	0.02	-0.01
mean	0.967	**0.057**	-0.004	0.06	-0.00
control	-0.004	-0.043	0.009		
`anchor-diff` (check)	0.931	0.097	0.131	0.10	0.14

Contrast at the final slice on the op-word arm, per seed, at the op position and the two use sites, with the ratio of each use site to the op position. The mean row is bold where it is at least 0.05 above the control's seed mean (the row below it, scored on `difference`). The last row is the primary's seed mean, a manipulation check. H3: miss.

Miss

On the op-word arm at the final slice, the contrast at `=` is 0.057 against the control's -0.043, over the floor; at the answer it is -0.004 against 0.009, under it. Beside the op position's 0.97 that is a ratio of 0.06 at `=` and about zero at the answer: the blocks carry a twentieth of the op word's alignment to `=` and nothing to the answer. On the primary, where the pull covers the use sites, both sit near 0.1 (`=` 0.097, answer 0.131), so the whole-line pull asks for more at those sites than the blocks bring on their own. This is the reading the design allowed for: at this weight the anchor marks the op where it is named, a little of it reaches `=`, and the answer position holds what the op produced rather than the op.

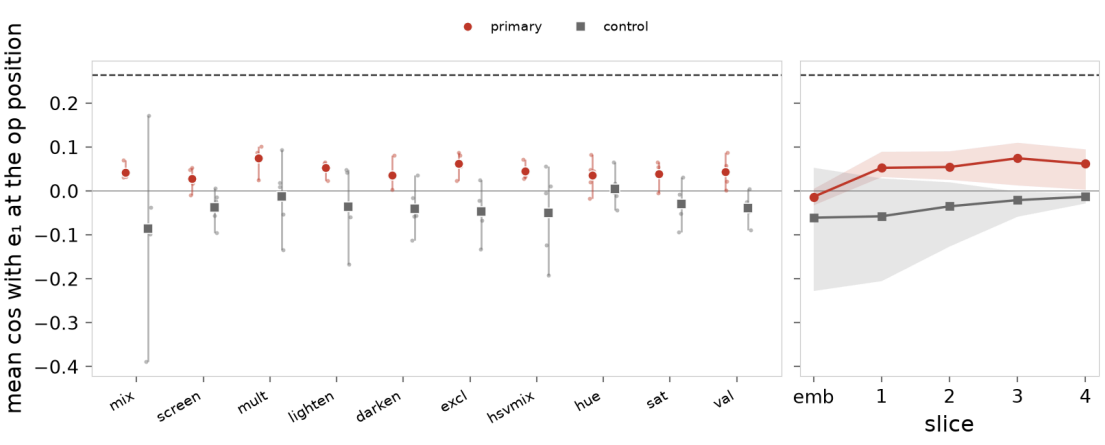
Containment, reported without a gate

What we expect. The other ten ops drift a little toward e_1 at the op position. The backlog item containment rises under the untied readout records $\bar{\alpha}$ at op1 at 0.264 on the red anchor under this recipe, up from the 0.1 the earlier gates asked for, with no mechanism named for the rise. We report the categorical analogue per op, beside that number and the control.

Three things could be behind the rise: (a) the red pull itself (a graded label on a color, at every position of the line), (b) the two changes that arrived with it (the untied readout and the whole-line span), and (c) the weights of the recipe.

This experiment keeps (b) and (c) and replaces (a) with a categorical pull on an op, so it can only say whether the red pull was needed. Containment near the control would say it was, alone or together with the shared factors; containment near the red value would say the shared factors produce the rise on their own.

A value between the two would mean both play a part, which seems likely, since the pull and the readout could each contribute. Either way this is one experiment's worth of evidence, and the item stays open.



Containment at the op position. *Left: per op other than* `difference`*, the mean cosine with e_1 at the op position over its probe lines, averaged over slices; the primary's seeds and the control's side by side. The dashed line is red's $\bar{\alpha}$ at op1 under the same recipe, 0.264, for scale; there is no gate. Right: the mean over the ten ops per slice, seed mean with the seed range shaded.*

op	primary	control	difference
`mix`	0.043	-0.086	+0.129
`screen`	0.028	-0.037	+0.065
`multiply`	0.074	-0.013	+0.087
`lighten`	0.053	-0.035	+0.088
`darken`	0.036	-0.041	+0.077
`exclusion`	0.063	-0.047	+0.110
`hsvmix`	0.045	-0.051	+0.095
`hue-hsv`	0.036	0.005	+0.031
`sat-hsv`	0.039	-0.030	+0.069
`value-hsv`	0.043	-0.040	+0.083
mean of ten	0.046	-0.037	+0.084

Containment per op: the mean cosine with e_1 at the op position, averaged over slices and seeds, on the primary and the control. Red's $\bar{\alpha}$ at op1 under the same recipe was 0.264.

What we saw. The other ten ops sit at 0.046 at the op position on the primary, against -0.037 on the control: a rise of about 0.08 on every op, and well under *red*'s 0.264. The rise grows over the slices, from nothing at the embedding (the other op words are not pulled) to its plateau by slice 1. On this reading the shared factors (b) and (c) do not produce the rise on their own: the graded red pull was needed for most of it.

One exploratory observation below complicates that reading. The primary's first operand leans toward e_1 on every line, a site the containment statistic here does not read but *red*'s $\bar{\alpha}$ at op1 does. So the item stays open, with a note.

The rule for the follow-up

> The equivalence experiment anchors `difference` if the primary passes H1 and H2 in full (a partial counts as a miss), margin and retention both. If it does not, it anchors the op in the sweep with the largest seed-mean op margin among those that pass the same gates at their three seeds: every op inside the task gate, the margin at or above the H2 bar, and every run clearing the retention rule, and that op is confirmed at the equivalence experiment's own seeds before any number is quoted for it. If no op qualifies, the anchored-op line stops here and the report says what gave way. H3, the arms, and the containment measure do not enter the rule: they describe what the anchor did, and the equivalence experiment measures them at its own seeds whichever way they came out.

The primary passes H1 and H2 in full, so the equivalence experiment anchors `difference`. Every op of the sweep qualifies as well, with seed-mean margins from 0.849 to 0.889: the data do not separate the ops, and the analytical choice stands on its own criteria.

condition	op	seeds	op margin ↑	worst task gap	retention	qualifies
`anchor-diff`	`difference`	5	**0.862**	**-0.005 (val)**	**yes**	**H1 pass, H2 margin pass**
`anchor-mix`	`mix`	3	**0.887**	**+0.006 (diff)**	**yes**	**yes**
`anchor-screen`	`screen`	3	**0.864**	**-0.010 (screen)**	**yes**	**yes**
`anchor-mult`	`mult`	3	**0.857**	**-0.007 (val)**	**yes**	**yes**
`anchor-lighten`	`lighten`	3	**0.888**	**-0.008 (sat)**	**yes**	**yes**
`anchor-darken`	`darken`	3	**0.880**	**-0.005 (val)**	**yes**	**yes**
`anchor-excl`	`excl`	3	**0.889**	**-0.008 (hsvmix)**	**yes**	**yes**
`anchor-hsvmix`	`hsvmix`	3	**0.868**	**-0.007 (hsvmix)**	**yes**	**yes**
`anchor-hue`	`hue`	3	**0.849**	**+0.007 (mult)**	**yes**	**yes**
`anchor-sat`	`sat`	3	**0.883**	**-0.011 (sat)**	**yes**	**yes**
`anchor-val`	`val`	3	**0.873**	**-0.005 (val)**	**yes**	**yes**

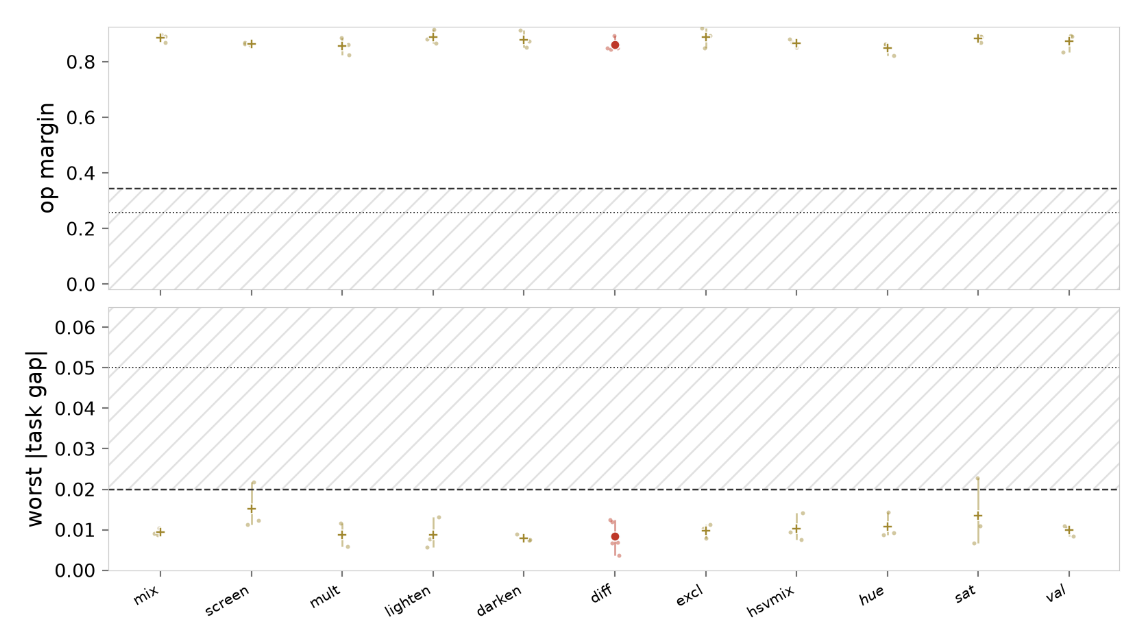
The rule applied. The primary qualifies on H1 and H2 in full (a partial counts as a miss); a sweep op qualifies when every op is within 0.02 of the control, its seed-mean margin is at or above 0.343, and every run clears the retention rule. Bold passes. The worst task gap is the op furthest from the control. The rule picks: `difference`.

Exploratory analyses

Anything we think of after seeing the data goes here, marked as post hoc. Four analyses are planned in advance as descriptions, with no gate, and run whichever way the predictions come out. Each gets a figure as well as a table. When the results come in, each analysis's figure and table go right after the paragraph that defines it, so definition and result read together.

The sweep. Every other op of table A+ anchored the same way at 3 seeds, read on the same task gap, op margin, retention, and use-site contrast. It is the design's "sweep over all ops to see whether they can all be anchored equally well", run cheaply while the machinery is warm, and the rule above falls back on it.

Whether the three order-sensitive ops behave differently is the one pattern worth looking for in advance.



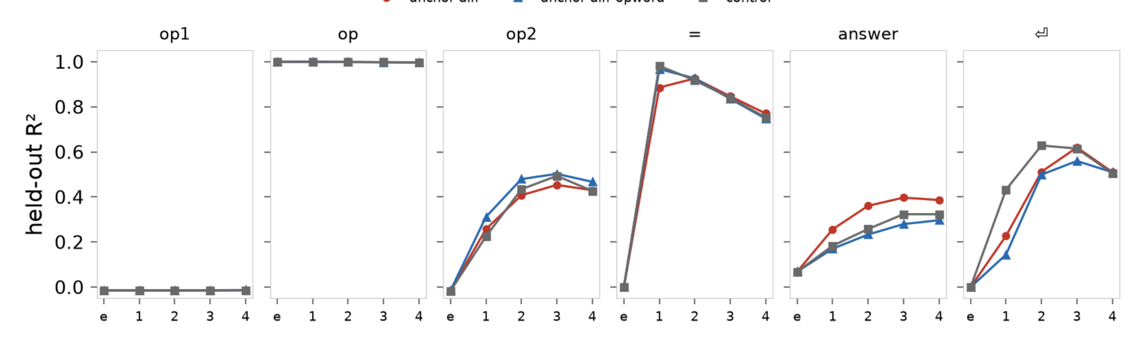
__The sweep.__ Each column is one op anchored on e_1, in the order of table A+; the primary's op is in its ink, the sweep ops in gold, and the order-sensitive ops' labels are italic. Top: the op margin, with the H2 bar (dashed) and partial level (dotted), hatched below. Bottom: per run, the largest absolute gap in held-out expected exact match from the control over the eleven ops, with the task gate dashed and the partial level dotted, hatched above.

op	seeds	op margin ↑	worst task gap	retention (min)	= contrast	answer contrast
`mix`	3	0.887	+0.006 (diff)	1.00	0.111	0.223
`screen`	3	0.864	-0.010 (screen)	1.00	0.093	0.310
`multiply`	3	0.857	-0.007 (val)	1.00	0.021	0.188
`lighten`	3	0.888	-0.008 (sat)	1.00	0.118	0.229
`darken`	3	0.880	-0.005 (val)	1.00	0.149	0.172
`difference`	5	0.862	-0.005 (val)	1.00	0.097	0.131
`exclusion`	3	0.889	-0.008 (hsvmix)	0.99	0.094	0.131
`hsvmix`	3	0.868	-0.007 (hsvmix)	0.99	0.069	0.148
`hue-hsv` (order-sensitive)	3	0.849	+0.007 (mult)	0.99	0.079	0.038
`sat-hsv` (order-sensitive)	3	0.883	-0.011 (sat)	0.99	0.134	0.133
`value-hsv` (order-sensitive)	3	0.873	-0.005 (val)	1.00	0.029	0.086

Every op anchored the same way, in the order of table A+: the primary's op at its seeds, the rest at the sweep's. The worst task gap is the seed-mean gap on the op furthest from the control; retention is the lowest over the op's runs; the contrasts are at the final slice, seed mean (the whole-line pull covers `=` on every one of these, so they are manipulation checks). Post hoc description; no gate.

All ten anchor alike. The margins span 0.04, every retention is at or above 0.99, and every seed-mean task gap is inside the gate, with single seeds of `sat` and `screen` touching it. The three order-sensitive ops (italic in the figure) have the three lowest margins, `hue` lowest at 0.849, but the spread is within what three seeds resolve, and their task gaps are the table's. Nothing sets them apart here. The margin's saturation at the op word is why: an anchor that reads the op's own embedding does not care what the op does to its operands.

The op-identity scan. The op-identity R^2 at every site, anchored against control. The design's equivalence claim is that anchoring does not change how readable the op is anywhere the anchor does not reach. The equivalence experiment has to declare a margin for that, and this scan gives it an observed spread. It carries no gate here.



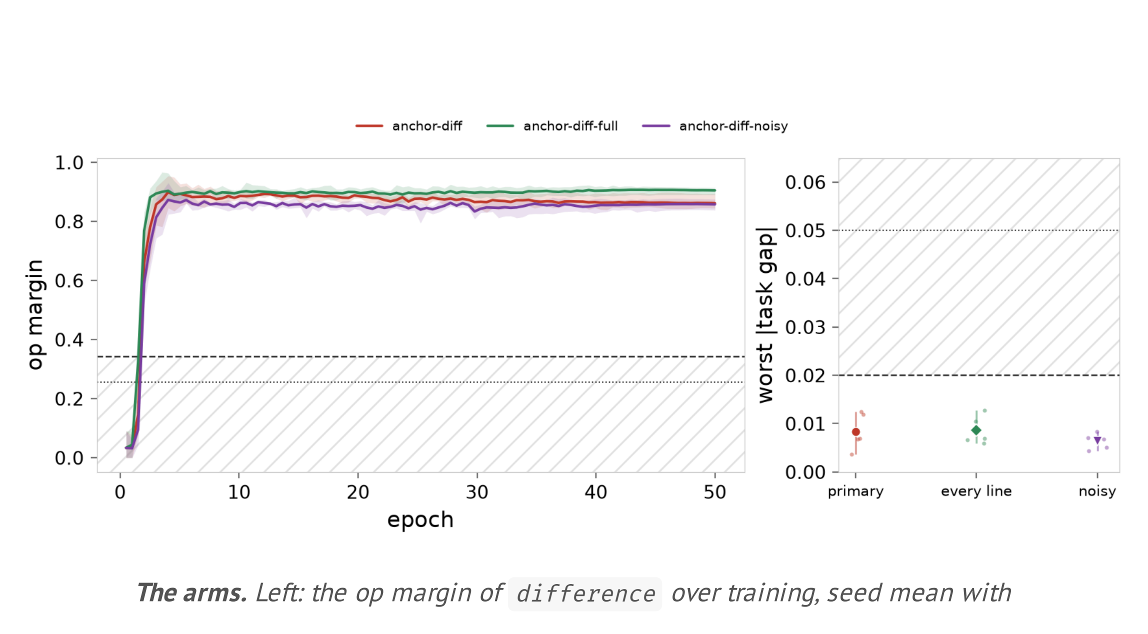
__Op-identity R^2 per site.__ One panel per position; within each, the held-out R^2 of a ridge probe from the state to the one-hot op word, per slice (the embedding at the left), seed mean. The primary and the op-word arm against the control. Post hoc description; no gate.

site	`anchor-diff`	`anchor-diff-opword`	`control`	primary − control
op1, embedding	-0.015	-0.014	-0.014	-0.001
op1, slice 4	-0.015	-0.013	-0.014	-0.000
op, embedding	1.000	1.000	1.000	+0.000
op, slice 4	0.997	0.997	0.997	-0.000
op2, embedding	-0.016	-0.016	-0.017	+0.001
op2, slice 4	0.430	0.467	0.427	+0.003
=, embedding	-0.001	-0.001	-0.001	+0.000
=, slice 4	0.770	0.748	0.753	+0.018
answer, embedding	0.068	0.068	0.068	+0.000
answer, slice 4	0.386	0.297	0.323	+0.063
⏎, embedding	-0.001	-0.001	-0.001	+0.000
⏎, slice 4	0.510	0.510	0.505	+0.006

Op-identity R^2 at the embedding and the final slice of each position, seed mean. Over every site, the primary minus the control spans -0.205 to +0.103 (mean |difference| 0.029). Post hoc description; no gate.

Anchoring leaves the op about as readable as the control has it. At the first operand nothing is readable, since the op word comes after it; at the op word the probe is perfect on every condition. At `=` the primary dips at slice 1 (0.89 against the control's 0.98) and matches it from slice 2. At the answer the primary reads *above* the control at every slice past the embedding (0.39 against 0.32 at the last), with the op-word arm a little under it: the whole-line pull keeps more of the op at the answer than the control does. The largest shift anywhere is at the newline at slice 1, -0.21, where the control's own seeds spread by about 0.4. The equivalence experiment can take its margin from this: a band of ±0.1 in R^2 would hold at every site but that one.

The arms. The every-line arm and the noisy-label arm, read on the same statistics as the primary: the task gap, the op margin, retention, and the use-site contrast. The every-line arm's margin against the primary's is the price or gain of the label share; the noisy arm's margin and task gap against the primary's are the cost of a fifth of wrong labels. The op-word arm carries H3 and is read there.



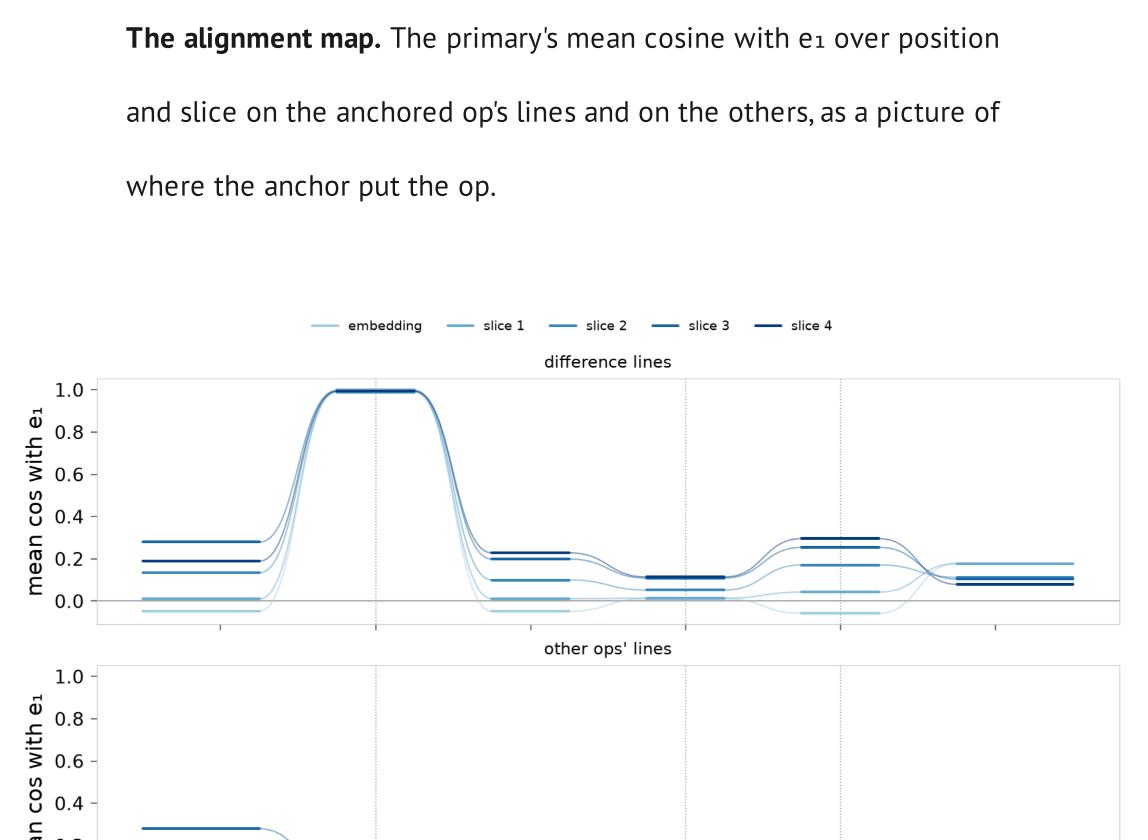
__The arms.__ Left: the op margin of `difference` over training, seed mean with the seed range shaded, for the primary, the every-line arm (every line of the op labelled) and the noisy arm (a fifth of the primary's labels moved onto other ops' lines); dashed and dotted lines mark the H2 bar and partial level. Right: per run, the largest absolute gap in held-out expected exact match from the control over the eleven ops, with the task gate. Post hoc description; no gate.

condition	label share (lines)	op margin ↑	vs primary	worst task gap	retention (min)	= contrast	answer contrast
`anchor-diff`	0.0018	0.862	1.00	-0.005 (val)	1.00	0.097	0.131
`anchor-diff-full`	0.0915	0.905	1.05	-0.003 (hsvmix)	1.00	0.065	0.042
`anchor-diff-noisy`	0.0018	0.858	1.00	-0.003 (val)	1.00	0.043	0.053
`anchor-diff-opword`	0.0016	0.890	1.03	-0.006 (val)	1.00	0.057	-0.004

The arms on the primary's statistics, seed means over 5 seeds. Label share is the fraction of training lines carrying a label in the first epoch (red's was 0.0018). The contrasts are at the final slice. The op-word arm is H3's and is listed for completeness. Post hoc description.

Label share barely moves the margin. Fifty times the labels raise it to 1.05× the primary's, and a fifth of wrong labels leave it at 1.00× with the task gap unchanged. Both say the same thing as H2: the margin saturates at the op word whatever the label share, and once that embedding is on the axis there is nothing left for more labels to pull. Where the every-line arm does differ is the newline: its embedding leans toward e_1 at 0.63 against the primary's 0.18, a free token the pull can place at no cost to the task. For the ladder this is encouraging: a scarce labeller and an imperfect one land the concept as well as an exhaustive one.

The alignment map. The primary's mean cosine with e_1 over position and slice on the anchored op's lines and on the others, as a picture of where the anchor put the op.



__Where the anchor put the op.__ The primary's mean cosine with e_1 at each position, one series per slice, light for the embedding to dark for the final slice, seed mean. Top, the `difference` probe lines; bottom, the other ten ops' probe lines. Dotted verticals mark the op, `=` and answer positions. Post hoc description.

The op word is at a cosine near 1 on the anchored lines at every slice and near 0 on the others: that is the contrast, and it is all at one position. The two panels agree at the first operand, as they must, since under causal attention the state there comes before the op word and cannot know the op. What they agree on is a lean toward e_1 that grows over the slices to 0.28.

Post hoc: the first operand's lean is the primary's alone. At the final slice the mean cosine at the first operand over every op's lines is 0.19 on the primary, -0.01 on the control, -0.04 on the op-word arm and -0.06 on the every-line arm. The whole-line pull asks the first operand of a labelled line to align with e_1, and nothing at that position distinguishes a `difference` line, so the model answers with a lean that every line shares. That the every-line arm does not show it says the per-line strength matters: at fifty times the labels each pull is fifty times weaker, and the arm places the newline instead.

This is a candidate mechanism for the rise the containment item records: the whole-line span arrived with the untied readout, and it pulls a position that cannot carry the concept. Under *red* the first operand does carry redness, so the analogy is partial. A red anchor with the first operand excluded from the span would test it. This was seen after the data and is not scored.

Discussion

An op anchors at least as readily as a color. A color is a graded property spread over three tokens of a line, so *red*'s margin had to be assembled from many partial alignments. An op is named by one token, and the pull puts that token's embedding on the axis in the first epochs. So the margin saturates, the task never feels it, all ten other ops do the same, and label share and label precision hardly matter. The alignment measurements here confirm that the pull landed, which is all the design asked of them.

The use-site contrast says where that leaves the op. With only the op word pulled, `=` picks up a twentieth of its alignment and the answer position none. The blocks carry a trace of the op to where it is applied and the answer position holds the answer. Whether the anchor captured the *operation* is still the question for suppression, and this result sets the expectation: an intervention at the op word will have little to work with downstream on its own, and the whole-line pull is what puts the op at the use sites.

The equivalence experiment inherits `difference`, a saturated margin that will not separate conditions, a scan whose spread outside the newline sits within ±0.1 in R^2, and the arms' finding that a scarce, imperfect labeller lands the concept. The first operand's lean is the loose end: a whole-line span pulls positions that cannot carry the concept, and the containment item now has a mechanism to test.

Method

The labeller

This experiment adds a labeller keying in which the op word draws the label, against a per-op rate table: 0.02 for the anchored op and zero otherwise on the primary, one for the anchored op on the every-line arm, and 0.016 for the anchored op with 0.0004 for each other op on the noisy arm. The pull covers the whole line, or the op word alone on the op-word arm. It consumes the random stream differently from the color labellers, so the control's batches are not this experiment's; hence the seed-mean comparison, as in ex-2.2.13.

The measurements

The op margin is ex-2.2.3's `line_margin` with uniform line weights summing to one over the anchored op's lines and zero elsewhere, on the pooled per-op probe sets ex-2.2.9 built. The function is red's m_line, but the comparison is not like for like: red's weights grade with redness (`p_slot`) and its baseline is the `mix` lines alone, where the op's weights are binary and its baseline includes its own lines as one eleventh of the pool. The binary label concentrates the labelled mean and the pooled baseline shrinks the gap by about a tenth, so the two differences pull in opposite directions and the ratio in H2 is a rough bar rather than a matched one.

The probe sets are stored per op, and a per-op baseline would make the op margin zero by construction (the weights are uniform within one op's lines), so the baseline is pooled over all eleven ops' lines at equal counts and the one-eleventh clause stands. Retention is the end-of-training margin over the margin at the start of the anchor weight's anneal, on the trajectory recorded every few epochs, with the anneal start located as ex-2.2.11 located it. Contrast and containment are the glossary's mean cosines, taken over the same probe lines. The op-identity scan fits a ridge probe at ridge strength 0.01 from the state at each site to the one-hot op word, with a held-out split by line.

Budget

50 runs at d64-L4, each as long as a run in ex-2.2.13, which trained 160 of them for about fourteen dollars on Modal. Eval adds the alignment measurements on eleven probe sets and the scan; no intervention is scored.

What this experiment does not do

It does not anchor *red* beside the op, does not suppress anything, does not vary the weight, and does not claim equivalence on the scan. Each of those is for a later experiment.

---

1. The term is the design's. The sweep reads every other op the same way, so the choice can be revisited on data. ↩