Ex 2.1.7: a repulsive term and a narrower pull

> We tested two mechanisms to improve anchor selectivity: **1.** Apply the anchor term only to operand 1 (no other tokens), and **2.** Add a repulsive term to clear the target subspace. Both work, but 1. worked better, and their effects stack somewhat.

Ex-2.1.6 anchored *red* with a single attractive term and got alignment without selectivity. The term moved the whole color cube onto the anchor direction, and the red-selective margin settled at about half its threshold, flat across a tenfold range of the anchor weight. Two candidate mechanisms could explain that, and the experiment could not tell them apart.

First, nothing repelled the rest of the cube. M1 never ran its anchor bare: it added an indiscriminate *anti-subspace* term, at a few percent of the anchor weight, penalizing the average squared alignment between the anchor direction and every representation in the batch. That is a gentle, global pressure that keeps the cloud as a whole off the axis, and that average is the quantity that ran away in ex-2.1.6.

Second, most of the pull was blind. The anchor pulled four prompt positions of each labeled equation, and three of them (`+` , op2, `=`) carry no information about which color sits at op1. There the label is an unobservable coin flip, so only the expected pull is visible to the optimizer, and that is nearly color-independent.

This experiment separates the two with a 2×2 factorial[1] at the ex-2.1.6 scoring rung ($\lambda_{\mathrm{a}} = 0.1$): {bare anchor, anchor + anti-subspace} × {four-position span pull, op1-only pull}. If the anti-subspace factor restores the margin, missing repulsion was the problem. If the op1-only factor does, the blind span was. If only the combination does, they compound.

Two extra arms ride along: a schedule-timing variant of the anti-subspace anneal, and a weight-ceiling rung at $\lambda = 1$.

▶ **Notation…**

Findings

H1 (task cost) — holds. Largest `named_holdout` exact-match gap from the control, across all seven conditions: 0.0013. Gate: 0.02.

H2 (selectivity) — holds, on the op1 conditions. `op1-anti` reaches $m_{\text{op1}} = 0.635$ and `op1-bare` 0.536, against a gate of 0.5. Both are graded (`op1-anti` R^2 = 0.88; `op1-bare` ρ = 0.80, gates 0.8), and the control sits at $|m_{\text{op1}}| = 0.027$ against a ceiling of 0.1. The preregistered primary condition `span-anti` reaches 0.456, a partial.

H3 (attribution) — fails. Both main effects clear 0.1, but the anti-subspace effect (+0.141) is smaller than the op1-only effect (+0.221), not larger; the ordering holds within every seed.

H4 (containment and dynamics) — fails on both parts. (a) No anti condition at the scoring rung falls to $\bar{\alpha} \leq 0.1$; `span-anti` ends at 0.326, missing the 0.25 partial too. (b) 3 of the six anchored conditions hold 0.8× their peak margin; the gate asks for all of them.

The timing arm scores nothing and did best on containment. `span-anti-late` stretches the anti-subspace anneal from epoch 50 to 90 and changes nothing else. It improves on `span-anti`, the schedule it varies, on every statistic H2 and H4 score: margin 0.456 → 0.598, $\bar{\alpha}$ 0.326 → 0.132 (the lowest of any anchored condition), and on retention it crosses from the sliding group to the holding one. Its margin still sits below `op1-anti`'s 0.635. See *Arms*.

> **How to read this report**
>
> This report was preregistered in Git. We wrote and froze the method, hypotheses, and decision thresholds before running the experiment, then replaced the placeholders with results in place. No threshold was amended after the freeze. Anything conceived after seeing the data is under *Exploratory analyses*, marked post hoc.

Method

Testbed, labels, and measurements

Everything ex-2.1.6 held is held here: the word-level `v216` testbed from ex-2.1.3 (216 grid colors, one token each, d64-L4), the same corpus seed and eval sets, labels drawn per visit with probability `redness(op1)⁸ × 0.08`, and the exhaustive 27-partner alignment probe set. The measurements are also the same: the alignment map over layers × positions, the label-affinity-weighted margin, the pair of redness probes, and the margin trajectory every 50 steps. The ex-2.1.6 report documents each; only the changes are described below.

One measurement is promoted: the exploratory geometry pass from ex-2.1.6 (per layer: the centroid of the 216 op1 states, its alignment with the anchor, and the extent of the cloud) is preregistered here, because H4 scores part of it. The mean alignment over all colors, $\bar{\alpha}$, is likewise recorded at every trajectory checkpoint beside the margin. In ex-2.1.6 it was added as an unscored diagnostic, and it turned out to be important.

The sweep

All four conditions run at $\lambda_{\mathrm{a}} = 0.1$, the scoring rung of ex-2.1.6, reused here by default: every rung was task-clean and 0.1 sat in the middle of the swept range. The control ($\lambda_{\mathrm{a}} = 0$) and the `span-bare` condition reproduce the ex-2.1.6 control and its $\lambda_{\mathrm{a}} = 0.1$ condition. So every comparison in this report is between conditions that went through identical code.

Factor one: the anti-subspace term. The `anti` conditions add the M1 repulsive term to the loss at weight $\lambda_{\bar{\mathrm{s}}}$:

$$\lambda_{\bar{\mathrm{s}}}\,\mathcal{L}_{\bar{\mathrm{s}}}, \qquad \mathcal{L}_{\bar{\mathrm{s}}} = \mathrm{mean}\,\cos^2(h, \hat{v}_{\mathrm{red}})$$

where the mean runs over every residual-stream slice and every non-pad position of the batch: every line, labeled or not. This is M1's term with the reserved coordinate axis replaced by an arbitrary unit direction: M1 penalized $\sum_i \hat{z}_i^2$ over the reserved axes of the *normalized* representation, and each $\hat{z}_i^2$ is a $\cos^2$ against a basis vector. A higher-dimensional reserved subspace would sum $\cos^2$ over an orthonormal basis of it; here the subspace is the one-dimensional span of the anchor. The term penalizes the mean-square alignment of everything, so the labeled pull has to buy alignment against a headwind. That is, we expect, the selectivity pressure ex-2.1.6 lacked. It never asks any particular point to leave the axis, only that the cloud as a whole not sit on it.

The term should be cheap in a 64-dimensional stream: an isotropic cloud (one with no preferred direction) has a mean $\cos^2$ of 1/64 per slice, and the task's three color dimensions fit in the remaining 63 with room to spare. In M1's 4- and 5-dimensional bottlenecks it cost a fifth or a quarter of the space, and its harder conditions (ex-2.9.1) needed the dimension cleared outright. The question is whether, at 3% of the anchor weight, it is strong enough to matter.

Factor two: the pull span. The `span` conditions pull the four prompt positions (op1, `+`, op2, `=`) of a labeled equation, as ex-2.1.6 did. The `op1` conditions pull op1 alone, so every pulled state belongs to the token that carries the labeled color. If the color-independent drift came from the three blind positions, narrowing the pull removes it with no repulsive term at all.

Schedule

The anchor schedule is unchanged from ex-2.1.6: ramp over the 10-epoch LR warmup, hold at λ_{a}, anneal from epoch 90 to a floor of $0.1\lambda_{\mathrm{a}}$ at 100.

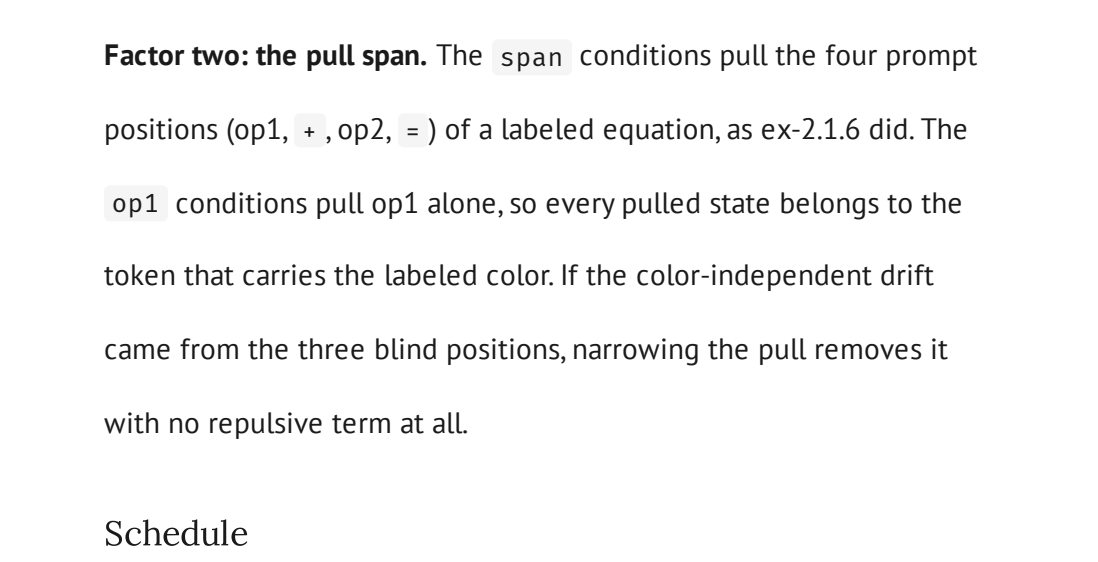
The anti-subspace schedule is copied from M1/Ex-2.9.1 and stretched to our 100 epochs by fraction of training. M1 balanced the attractive and repulsive terms in time rather than by a single ratio. Anti-subspace opened at 2.5× the anchor peak weight, meaning it was at full weight from step 0 (its effective push still scaled by the warming-up learning rate, like every term), before the anchor had ramped in. It then annealed to 3% of the anchor by the halfway point and held there. Mapped onto our frame: $\lambda_{\bar{\mathrm{s}}}/\lambda_{\mathrm{a}}$ starts at 2.5 at epoch 0, anneals (minimum-jerk) to 0.03 by epoch 50, and holds; from epoch 90 both terms share the end-of-training anneal, so their ratio is constant from the midpoint on.

One arm probes the timing: `span-anti-late` stretches the anneal endpoint from epoch 50 to 90, holding the repulsion near peak through most of training.

The second arm, `span-anti-hi`, runs the full recipe at $\lambda_{\mathrm{a}} = 1$, one rung above the maximum Ex-2.1.6 swept, with $\lambda_{\bar{\mathrm{s}}}$ scaled in proportion. Both terms scale together, so it probes the headroom of the recipe, not of the bare anchor.

condition	λ_a	pulled positions	$\lambda_{\bar{s}}/\lambda_a$ schedule	role
`lam0`	0	—	—	control
`span-bare`	0.1	`op1 + op2 =`	—	the Ex-2.1.6 λ_a=0.1 condition, re-run
`span-anti`	0.1	`op1 + op2 =`	2.5 → 0.03 by ep 50	primary scoring condition
`op1-bare`	0.1	`op1`	—	factorial
`op1-anti`	0.1	`op1`	2.5 → 0.03 by ep 50	factorial
`span-anti-late`	0.1	`op1 + op2 =`	2.5 → 0.03 by ep 90	timing arm
`span-anti-hi`	1	`op1 + op2 =`	2.5 → 0.03 by ep 50	ceiling arm

The seven conditions, each run at seeds [0, 1, 2] — 21 runs. The four factorial conditions share λ_a = 0.1 and carry the H2 and H3 gates; the two arms add none of their own, though H1 and H4 reach them by their stated scopes — the timing arm is a λ_a = 0.1 anti condition, and H4(b) covers every anchored run. $\lambda_{\bar{s}}/\lambda_a$ is the anti-subspace weight relative to the anchor peak.



The three schedules at the scoring rung (λ_a = 0.1), log scale. The anti-subspace term opens at 2.5λ_a, dominant before the anchor arrives, and anneals to 0.03λ_a by epoch 50 (the M1/ex-2.9.1 keyframes, mapped by fraction of training). The dashed line is the `span-anti-late` arm, which reaches the hold ratio at epoch 90 instead. From epoch 90 the anchor and anti-subspace weights share the end-of-training anneal to a floor of 0.1× their held values.

Hypotheses

Gates carry over from ex-2.1.6 with their thresholds and rationale, as do its noise floors (margin ≈ 0.056 per seed, ≈ 0.032 on a three-seed mean): same architecture, same measurement. Unless stated otherwise, results are reported on `span-anti` (seed-averaged), the condition the ex-2.1.6 discussion named as the next experiment.

H1. The added term and the resulting narrower pull are as free as the lone anchor was: every $\lambda_{\mathrm{a}} = 0.1$ condition (the four factorial conditions and the timing arm) is task-clean, meaning `named_holdout` exact match within 0.02 (absolute) of the seed mean of the in-experiment control. Partial: the factorial is clean but the timing arm is not, which would itself be evidence that sustained repulsion has a price.

H2. With the missing ingredient restored, the concept lands selectively. On some task-clean factorial condition: (a) the layer-mean margin at op1 reaches $m_{\mathrm{op1}} \geq 0.5$; (b) the response is graded, meaning that over the 216 colors, the seed-mean layer-mean alignment at op1 clears Spearman $\rho \geq 0.8$ against `redness` or Pearson $R^2 \geq 0.8$ against `sim^1.5` (both tracks, thresholds, and step ceilings as justified at length in ex-2.1.6); (c) the control shows $|m_{\mathrm{op1}}| \leq 0.1$. Partial: some condition clears 0.35 without reaching 0.5; 0.35 is above the ex-2.1.6 value of 0.27 by more than the seed-mean noise.[2]

H3. The margin failure in ex-2.1.6 was mostly the missing repulsion, not the blind span. This is scored on the main effects of the factorial on m_{op1} (for each factor: the mean over its two conditions, differenced): the anti-subspace main effect is at least 0.1 and at least as large as the op1-only main effect.[3] Two outcomes that fail H3 are named in advance so a miss stays readable; unlike the partials elsewhere, they are contrary readings rather than weaker passes. The op1-only effect clears 0.1 and is the larger of the two, which reads as the blind-span interpretation of ex-2.1.6 was correct (whether or not the anti-subspace effect also clears it); or neither main effect clears 0.1 but the `op1-anti` condition alone does (an interaction: each mechanism blocks the margin on its own).

H4. The repulsive term visibly does its two jobs: it keeps the cube as a whole off the axis, and the margin holds its early peak instead of sliding back. (a) Containment: in every anti condition at $\lambda_{\mathrm{a}} = 0.1$, the end-of-training mean alignment over all colors at op1 satisfies $\bar{\alpha} \leq 0.1$.[4] (b) Retention: the ex-2.1.6 trajectory gate, unchanged: for every anchored run whose running maximum of m_{op1} reaches 0.2, the final value is at least 0.8× that maximum.[5] Partials, in decreasing strength: every anti condition clears the old halving level $\bar{\alpha} \leq 0.25$ without all reaching 0.1, which reads as repulsion weakening the runaway without containing it; or (a) holds and (b) fails only in the timing arm, whose sustained repulsion changes the trajectory most.

Task cost (H1)

condition	seen EM	holdout EM	holdout NLL	`open` dist	Δ holdout EM	H1 gate
ex-2.1.3 `v216`	1.000	0.995	0.025	0.141	—	reference
ex-2.1.6 $\lambda_a = 0.1$	1.000	0.997	0.017	0.130	—	reference
control	1.000 ±0.000	0.999 ±0.002	0.019 ±0.009	0.133 ±0.002	+0.000	clean
span · bare	1.000 ±0.000	0.997 ±0.004	0.017 ±0.004	0.130 ±0.003	-0.001	clean
span · anti	1.000 ±0.000	0.997 ±0.004	0.020 ±0.010	0.135 ±0.006	-0.001	clean
op1 · bare	1.000 ±0.000	0.999 ±0.002	0.015 ±0.001	0.134 ±0.003	+0.000	clean
op1 · anti	1.000 ±0.000	0.997 ±0.002	0.021 ±0.006	0.135 ±0.009	-0.001	clean
late arm	1.000 ±0.000	0.999 ±0.002	0.021 ±0.002	0.137 ±0.002	+0.000	clean
ceiling arm	1.000 ±0.000	0.997 ±0.004	0.021 ±0.006	0.132 ±0.003	-0.001	clean

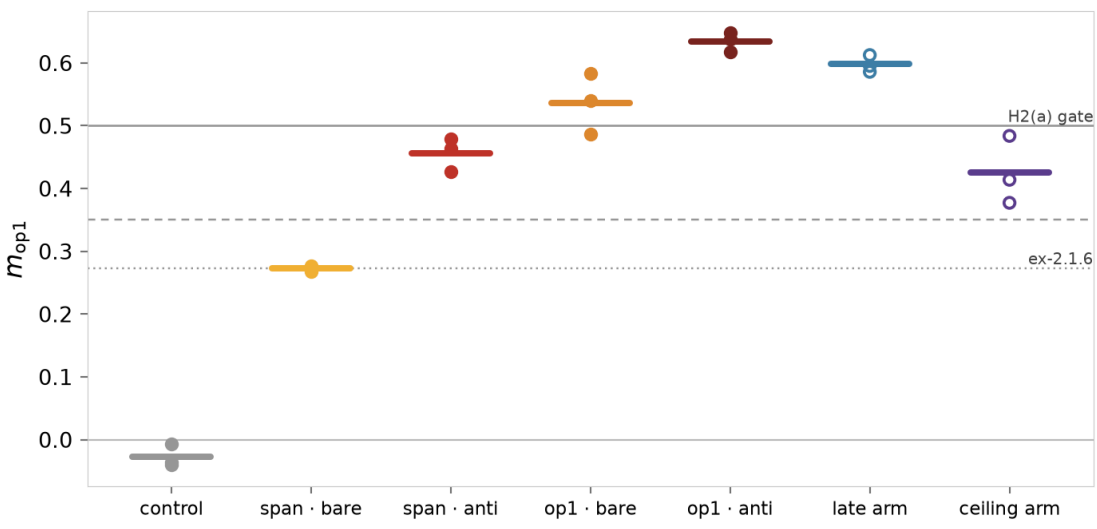
Behavior by condition. Each cell is the seed mean, with half the seed range beside it. EM is exact match on the answer token. NLL is the surprise of that token in nats. `open` dist is the RGB distance from the guessed color to the true mix, on pairs with no named answer. Δ holdout EM is the signed gap to the in-experiment control, and the last column reports the H1 gate, which passes when that gap is within 0.02. The two top rows are published conditions from earlier experiments, shown for comparison.

H1 holds everywhere, with two orders of magnitude to spare. On `named_holdout` exact match, the largest gap between any condition and the control is 0.0013. The gate allows up to 0.02; the observed gap amounts to a single example out of 256. Seen pairs are at 1.000 in every condition. So neither the repulsive term nor the narrower pull has any measurable cost to the task. The partial outcome we planned for, where the factorial arms pass but the timing arm does not, did not arise: the timing arm is one of the conditions tied with the control.

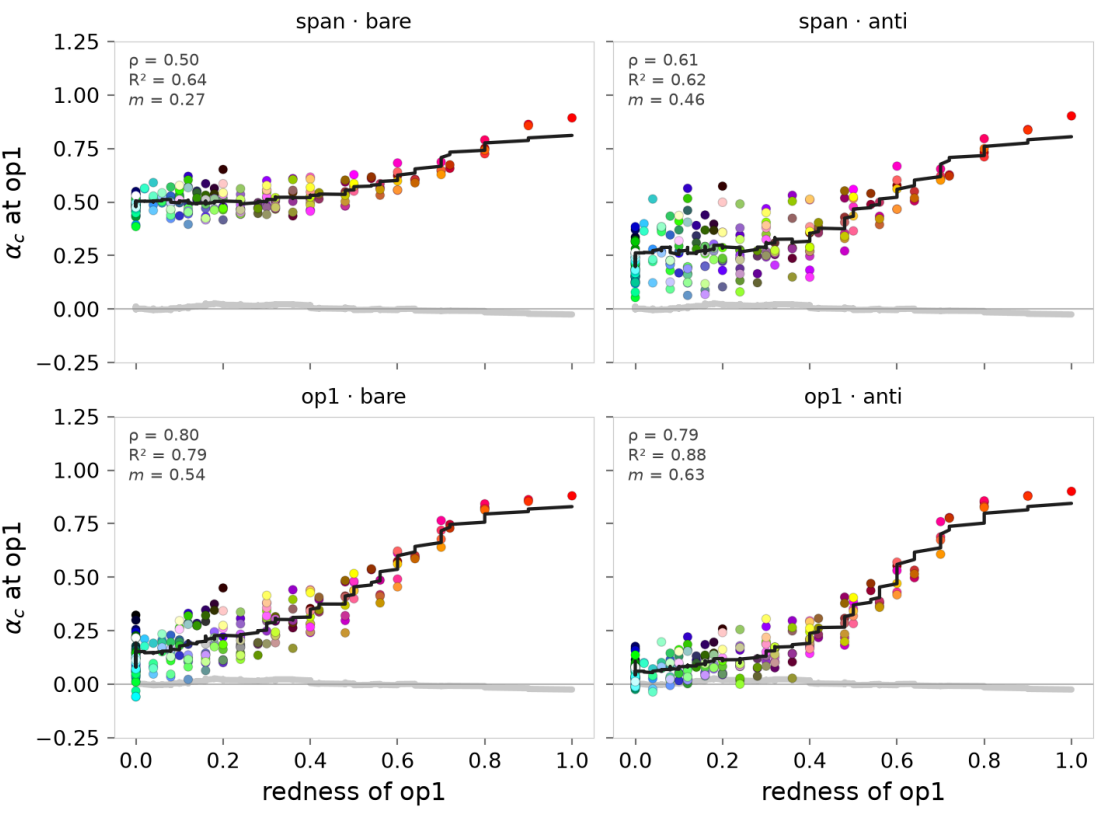
The finer scoring agrees. Answer NLL (the model's surprise at the correct token, in nats) stays between 0.015 and 0.021 with no ordering by condition. The `open` -pair distances cluster near 0.134, and the spread from best to worst condition is only 0.008.[6]

Selectivity (H2)

Both factors (the repulsive term and the narrower pull) increase the margin substantially. The figure below shows the statistic that H2(a) scores, for every condition, alongside the published ex-2.1.6 value for the same pull.



The statistic H2(a) scores: m_{op1}, the layer-mean alignment margin at the first operand. One dot per seed, with the seed mean as the heavy bar. The solid rule is the H2(a) gate of 0.5 and the dashed rule the 0.35 partial; the dotted rule is the published ex-2.1.6 value of 0.27, which the `span-bare` condition re-runs. The four factorial conditions are drawn solid, the two arms hollow, and the control grey.



Alignment at op1, per color: α_c, the layer mean of $\cos(h, \hat{v}_{\mathrm{red}})$ averaged over seeds, against the redness of the color. One mark per color, drawn in that color; the heavy line is a 25-color sliding mean over the redness ordering, and the flat grey band underneath is the same sliding mean for the control. The grading statistics and m_{op1} are quoted per panel; both H2(b) gates sit at 0.8. The panels are the 2×2: pull span across, repulsive term down.

H2 passes, on the op1 conditions. (a) The margin reaches $m_{\mathrm{op1}} = 0.635$ on `op1-anti` and 0.536 on `op1-bare`, both task-clean, against a gate of 0.5. The hypothesis takes a maximum over four noisy means, but `op1-anti` clears the gate by 0.13, roughly four seed-mean noise units, so the pass does not rest on that maximum. (b) The response is graded on both conditions, but each clears a different one of the two tracks, and neither clears both. `op1-bare` reaches ρ = 0.802 against `redness` with R^2 = 0.787; `op1-anti` reaches R^2 = 0.880 against `sim^1.5` with ρ = 0.791. The gate asks for either track, so both conditions pass. Still, `op1-bare` clears the rank track by only 0.002. That is thin, so the pass is better read as resting on the R^2 of `op1-anti`, which clears by 0.08.

The step-ceilings check qualifies this. A pure step response (one that jumps at the label threshold instead of rising smoothly) can score at most 0.75 on the rank track and 0.73 on the proportionality track. The R^2 of `op1-anti` clears its ceiling by 0.15; the ρ of `op1-bare` by only 0.05. The scatter is the more convincing evidence in any case: it rises across the whole cube rather than stepping at the label threshold, so the anchor caught *red* and not merely the exemplars that happened to be labeled. (c) The control sits at $|m_{\mathrm{op1}}| = 0.027$, inside its 0.1 ceiling.

The preregistered primary condition `span-anti` is not among the passes. It reaches 0.456: comfortably into the partial band (0.35) and well clear of the ex-2.1.6 result of 0.27, but short of the gate. The repulsive term alone was the predicted fix, and it gets most (but not all) of the way there.

▶ **The replication is bit-identical…**

Attribution (H3)

	bare	+ anti-subspace	span effect
span pull	0.273 ±0.005	0.456 ±0.026	—
op1-only pull	0.536 ±0.048	0.635 ±0.015	—
anti effect	—	—	interaction
main effects	anti +0.141	op1-only +0.221	-0.085

The factorial on m_{op1}. Body cells are seed means, with half the seed range beside them. Anti effect: margin gained by adding the repulsive term, averaged over both pull spans; op1-only effect: margin gained by narrowing the pull, averaged over both terms; interaction: what the two together buy beyond their sum. H3 asks the anti effect to reach 0.1 and to be at least as large as the op1-only effect; the noise on a main effect is about 0.032.

H3 fails, and implies the first of its two named contrary readings. Both main effects clear 0.1: the repulsive term is worth +0.141 of margin, and the narrower pull is worth +0.221, against a noise floor of about 0.032 on a main effect. But H3 asked for the anti-subspace effect to be *at least as large*, and instead the op1-only effect is the larger, by a factor of about 1.6. That is the reading the preregistration named: the blind-span interpretation of ex-2.1.6 was correct.

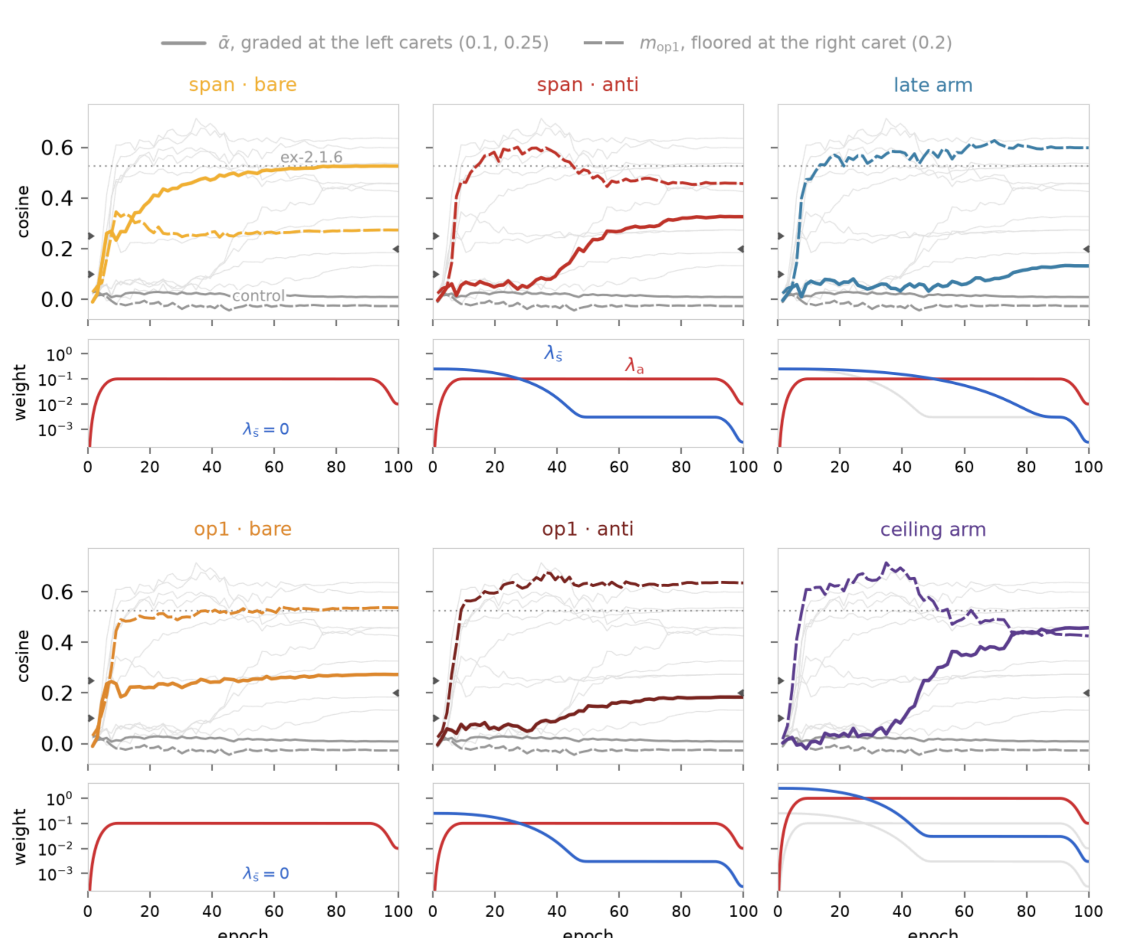
The gap between the two effects is 0.080, about 2 noise units on the generic seed-mean floor, and that floor understates the case. Computed within a seed, which cancels seed-to-seed variation, the op1-only effect is the larger on every seed, with no overlap between the two sets (0.22 vs 0.15, 0.22 vs 0.16, 0.22 vs 0.12), so the ordering is not due to one lucky run.

Still, the two factors are not on a common scale: one changes how many positions are pulled, and the other adds a term at a weight copied from another experiment. So this says the *particular* settings tried favor the span, not that repulsion is intrinsically the weaker lever.

The interaction is -0.085: mildly sub-additive, the expected shape if the two are routes to one place. Both reduce how much of the pull can be satisfied by a color-independent shift, so whichever is applied second has less left to do. It also means the interaction outcome H3 named (neither main effect clearing while `op1-anti` alone does) is the opposite of what happened.

Containment and dynamics (H4)

The anti-subspace term penalizes the mean-square alignment of the whole cloud, which is the quantity H4(a) scores. So "the repulsive term lowers $\bar{\alpha}$" is close to a tautology, and on its own not evidence that this mechanism explains ex-2.1.6. H3 carries the load: does lowering $\bar{\alpha}$ gain margin, a quantity neither factor acts on directly?



Training dynamics, one pair of panels per condition. Above: seed means of the two cosines on the anchor axis, on one shared scale. Solid is $\bar{\alpha}$, the mean alignment over all 216 colors at op1; dashed is the margin m_{op1}. The grey pair is the control; the faint lines are the other conditions. The carets on the left spine are the H4(a) grading levels for $\bar{\alpha}$, and the one on the right is the H4(b) floor for m_{op1}. The dotted rule is the published ex-2.1.6 endpoint of 0.53. Since span-bare re-runs that condition, its solid line landing on the rule is the reproduction check. Seed spread is given in the H4 table below. Below each: the weight schedule that condition trained under, in the colors of the schedules figure above: anchor λ_{a} in red, repulsion $\lambda_{\bar{s}}$ in blue (absent in the bare conditions). The two arms in the third column also show the primary schedule in ghost ink, since each is a variation on it.

condition	ᾱ at op1	H4(a)	retention, per seed	H4(b)
span · bare	0.526 ±0.009	–	0.68, 0.78, 0.91	fail
span · anti	0.326 ±0.013	fail	0.72, 0.73, 0.74	fail
op1 · bare	0.273 ±0.041	–	0.95, 0.97, 0.95	pass
op1 · anti	0.184 ±0.012	partial	0.96, 0.93, 0.90	pass
late arm	0.132 ±0.011	partial	0.92, 0.95, 0.95	pass
ceiling arm	0.457 ±0.053	–	0.65, 0.55, 0.55	fail

The two H4 gates. ᾱ is the end-of-training mean alignment over all 216 colors at op1, seed mean with half the seed range. H4(a) scores the λ_a = 0.1 anti conditions against a ceiling of 0.1, with partial marking a condition that clears the 0.25 level without reaching it. Retention is each run's final margin over its running maximum; runs whose maximum never reaches 0.2 are reported but not scored, and H4(b) passes when every scored run holds 0.8× its peak.

H4 fails on both parts.

(a) Containment. No *anti* condition at the scoring rung falls to $\bar{\alpha} \leq 0.1$: `span-anti` ends at 0.326, `op1-anti` at 0.184, and the timing arm at 0.132. The named partial asked all of them to clear the looser bar of 0.25, and `span-anti` misses that too, so H4(a) fails outright rather than partially.

The repulsive term does reduce the quantity it is defined on: $\bar{\alpha}$ ends at 0.53 in the bare condition and 0.33 with the term. But at the weight ratio carried over from M1, the term only weakens the cube-wide drift; it does not contain it.

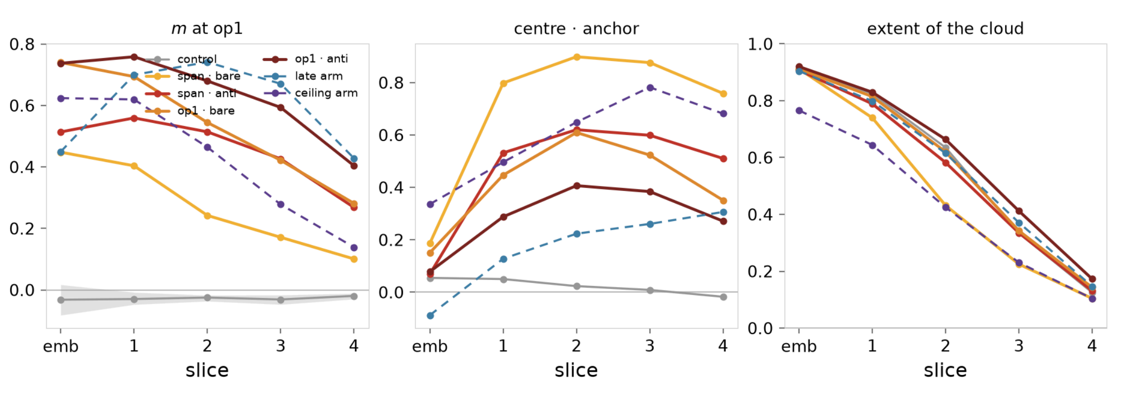
The trajectory shows when containment is lost. While the repulsion outweighs the pull, $\bar{\alpha}$ sits near the control. It climbs once the anneal has brought $\lambda_{\bar{s}}$ down to about a tenth of λ_{a}. So each condition breaks at a time set by its own schedule: around epoch 40 on the default anneal, and epoch 75 on the late one. The preregistration flagged this contrary outcome as a possibility for this figure, and the timing arm was positioned to test it.

(b) Retention. 3 of the six anchored conditions hold 0.8× their peak (`op1-bare`, `op1-anti`, `span-anti-late`), and 3 slide below it (`span-bare`, `span-anti`, `span-anti-hi`). The gate applies to every anchored run, so H4(b) fails.

The split does not fall purely along the same line as H3. Both op1 conditions hold, at 0.90 or better across all six runs. So does `span-anti-late`, which pulls the full span.

What the three holding conditions have in common is that something held the color-independent shift down. In two of them a narrow pull lowers the level $\bar{\alpha}$ settles at; in the third, a repulsion held near peak through epoch 90 delays the climb. The three that slide are the ones where $\bar{\alpha}$ was free to climb. Wherever it climbs, the peak also moves later: epochs 9–11 in `span-bare`, 26–33 once the repulsive term is added, and 55–70 in the timing arm. That looks like a slower climb to a higher place, rather than an early peak that erodes.

In ex-2.1.6 the margin rose early and then gave back a quarter of itself, and the reading offered there was that the rest of the cube was catching up. This experiment supports that reading, and points to the color-independent shift as the thing doing the catching up. `op1-bare` stops the slide while carrying no repulsive term at all.



Margin and cloud geometry by layer, seed means. Depth 0 is the token embedding and depth 4 the last block's output. ***Left:*** *the margin at op1 by layer — the per-depth terms whose mean is m_{op1}; the shaded band around the control is its seed min–max, as the scale of a null.* ***Middle:*** *the cosine between the centre of the 216 op1 states and the anchor direction — where the cloud sits.* ***Right:*** *the extent of that cloud, the mean squared distance of a color from the centre (states are unit-norm, so this runs from 0 for a collapsed cloud to 1 for a spread one). The repulsive term acts on the middle panel by construction; the right panel is what it costs.*

	m, embedding	m, last layer	centre · anchor	extent
control	-0.03	-0.02	0.02	100%
span · bare	0.45	0.10	0.90	68%
span · anti	0.51	0.27	0.62	92%
op1 · bare	0.74	0.28	0.61	98%
op1 · anti	0.74	0.40	0.41	105%
late arm	0.45	0.43	0.22	97%
ceiling arm	0.62	0.14	0.65	67%

Per-layer geometry at op1, seed-averaged, decoding the three panels above. First two columns: the margin at the token embedding and at the last layer. Last two, read mid-stack at layer 2 where the control cloud is still broad: the cosine between the centre of the color cloud and the anchor direction, and the extent of the cloud as a fraction of the control's.

Increased selectivity at deeper layers. The margin in Ex-2.1.6 peaked in the embedding and decayed through the stack (reproduced here). Both new factors act mostly on the decay rather than on the starting point: `op1-anti` starts 1.6× the bare span condition and ends the stack at 4× its margin.

The timing arm makes that plainest. It leaves the embedding approx. equal to `span-bare`, rises through the middle of the stack, and finishes highest of any condition — about where it started, while every other condition ends well below its own embedding value.

Both factors independently reduce the whole-cube shift. The anchor axis carries essentially nothing about where the control cloud sits, and the bare span pull (ex-2.1.6) swings it most of the way over. The timing arm resists the whole-cube shift best out of the tested conditions.

A *separate* term may not be needed. In ex-2.1.6, the bare anchor compressed the cube mid-stack, suggesting that the *anti-subspace* and pairwise *separation* terms from M1 may be needed. But in this experiment, most of the extent was recovered by *anti-subspace* alone.

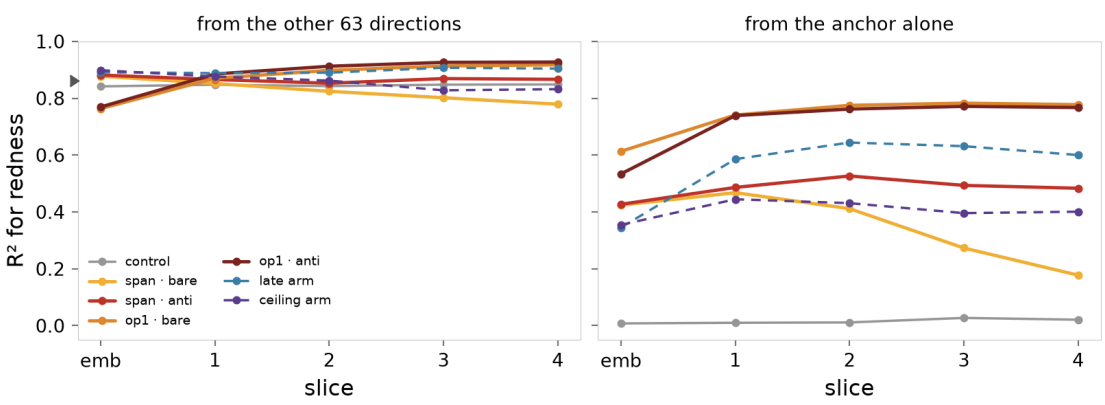
Restoring the extent is not something the repulsion does specifically, yet the spread improves over the factorial conditions in the same order as the margin m. It seems the color-independent shift was what compressed the cube, so anything that reduces the shift restores it.

Compression recurs at high term weight (ceiling arm). At $\lambda_{\mathrm{a}} = 1$ the cube is narrower even in the token embedding (0.77 against 0.91 everywhere else). So a `separate`-style term may yet be needed in some configurations.

Secondary measurements

We reuse the "where is redness readable" measurement from ex-2.1.6, without a gate. Two probes read the same target from complementary parts of the residual stream: one from the anchor coordinate alone, and one from everything else. The repulsive term pushes unlabeled colors off the axis, so it acts on both.

The two have different reference points. The anchor coordinate starts empty and can reach 1, so that probe measures how much redness anchoring put there. The other 63 directions encode the color cube regardless of what the anchor does. Redness is a function of color, so that probe has an RGB floor it cannot go far below: it reports whether the cube survived rather than whether the concept moved.



How strongly redness is encoded at op1, read two ways on one R^2 scale. ***Left:*** *held-out R^2 for a ridge probe predicting redness from the residual stream at op1 with the $\hat{v}_{\text{red}}$ coordinate deleted. The caret marks $R^2 = 0.86$, the score the same probe gets from the raw RGB values, which no condition can fall far below while the cube is intact.* ***Right:*** *the same target read from the anchor coordinate alone, as squared correlation, which starts at zero and has room to reach 1. Per-layer seed means. The sample is the 216 op1 colors, with a 5-fold split so no color is scored by a probe that saw it during fitting. Neither panel has a pass/fail threshold.*

The axis became much more readable; everywhere else sits at the RGB floor. On the anchor coordinate alone, redness goes from R^2 = 0.02 in the control to 0.74 and 0.71 in the op1 conditions. That is double what the bare span pull achieved (0.35), and it tracks the margin ordering closely. So the conditions that score well on selectivity are also the ones that put the most redness on the axis.

Read from the other 63 directions, every condition at every depth stays within 0.10 of R^2 = 0.86, which is what the same probe gets from the raw RGB values. Call that the RGB floor: redness is a function of color and the task needs the color cube, so a linear readout recovers about this much wherever the anchor put it. Redness found outside the anchor is the cube being read, not a second copy of the concept. The control is flat at 0.84–0.85 at every depth, just under the floor, because its color code is marginally lossier than the raw channels.

The other conditions drift away from the floor with depth, in both directions. The op1 conditions start *below* it at the token embedding (0.77) and finish *above* it at the last layer (0.93). Above the floor is reachable because the residual stream encodes color nonlinearly and redness is quadratic in RGB, so a linear probe does better there than on the raw channels. `span-bare` runs the other way, from 0.88 down to 0.78. Falling below the floor means color itself was lost, the same cube compression its spread reports.

This measurement deletes 1 direction of 64, so it has almost no room to move, and where it sits is set by what the task requires rather than by where the anchor put anything. It is a check that the color cube survived, beside task accuracy, rather than evidence about the anchor.

Arms

The timing arm

`span-anti-late` differs from `span-anti` in one way: the epoch at which the repulsive term finishes annealing down to its hold ratio, moved from 50 to 90. That one change improves on the M1 schedule on every statistic that H2 and H4 score, which we did not expect:

- margin 0.598 against 0.456. The difference of 0.14 is about 4 noise units, and about the size of the whole anti-subspace main effect;
- retention 0.94 against 0.73, moving it from the sliding group to the holding one;
- grading R^2 0.78 against 0.62;
- containment 0.132 against 0.326, the lowest of any anchored condition. Least surprising of the four: more repulsion for longer lowers the very quantity it penalizes.

Redness read from the other 63 directions is the one measurement that does not follow: this arm is the highest of any condition (0.90 against 0.87). That is a weak observation, and it is why the claim above is scoped to the scored statistics. A value above the floor says the residual stream is a more probe-friendly encoding of color, which says nothing about the anchor either way.

So stretching one anneal by 2× was worth more than adding the term in the first place. It looks worth a dedicated sweep over the anneal endpoint. Note that we only measured one alternative timing, and it is confounded with total repulsive strength, since holding near peak for longer also delivers more repulsion overall.

The ceiling arm

More weight turns out not to be better. `span-anti-hi` runs the full recipe at $\lambda_{\mathrm{a}} = 1$, ten times the scoring rung, with the repulsive term scaled in proportion.

We still have not found the task ceiling. Holdout accuracy is within 0.0013 of control, so even at ten times the scoring weight the task shows no cost. Whatever bounds the anchor weight for D2.2, it is not the task loss at this scale, and finding the real ceiling would need a dedicated sweep well above $\lambda_{\mathrm{a}} = 1$.

But we have passed the selectivity ceiling. At ten times the weight the margin is 0.425, level with the 0.456 of `span-anti` at a tenth of the weight. The other three measurements are all worse: $\bar{\alpha}$ back up to 0.457, retention down to 0.58, and the cube compressed even in the token embedding. So a heavier pull gives more alignment but no more selectivity: at this weight, moving everything becomes the cheapest way to satisfy the term.

Exploratory analyses

This section is post hoc: we planned it after seeing the results, and it scores no hypothesis.

Why doesn't the repulsive term push *red* off the anchor too?

The anti-subspace term does not distinguish between concepts: it penalizes $\cos^2(h, \hat{v}_{\text{red}})$ at every position of every line, labeled or not. That makes the timing arm look puzzling. Holding that penalty near its peak for another forty epochs ought to push the anchored concept off the anchor along with everything else. Instead, that arm has the highest margin of any `span` condition.

Splitting the margin into its two halves shows the term does act on *red*; the pull toward the anchor is simply stronger.

condition	$\bar{\alpha}$, ten reddest	$\bar{\alpha}$, all 216	margin
lam0	-0.008	0.008	-0.027
span-bare	0.760	0.526	0.273
span-anti	0.742	0.326	0.456
op1-bare	0.805	0.273	0.536
op1-anti	0.825	0.184	0.635
span-anti-late	0.690	0.132	0.598
span-anti-hi	0.881	0.457	0.425

The margin split into the two quantities it differences. The first column is the mean alignment of the ten reddest colors — the ones carrying 85% of the label mass, so effectively "where the anchored concept sits". The second is the mean over all 216, the quantity the repulsive term penalizes. Comparing `span-anti` with `span-anti-late` is the case of interest: sustained repulsion costs the reddest colors a little and the rest of the cube a great deal.

Delaying the $\lambda_{\bar{s}}$ anneal costs the reddest colors 0.052 of alignment. It costs the cube as a whole 0.194, which is 4 times as much. So the term is doing what its definition says, *red* included. The margin improves because the same indiscriminate push matters far more for the unlabeled bulk than for the labeled reds.

Why is the push so much gentler on the reds? It comes down to how the two terms are normalized. Neither weight depends on the labels: λ_{a} and $\lambda_{\bar{s}}$ are functions of the epoch alone, identical in every batch. What differs is the denominator *inside* each term. Both terms are means, but over different sets: the anchor averages over the positions it pulls, and the repulsion averages over every live position. Measured on the real sampler, a batch has 4,029 live positions. The anchor fires in 57% of batches, and when it does, it averages 5.8 pulled positions. So a labeled state feels a pull scaled by $1/5.8$ against a push scaled by $1/4,029$, about 701:1 before weights. Even at the opening $\lambda_{\bar{s}} = 2.5\lambda_{\text{a}}$, and after the $2\cos$ factor from differentiating $\cos^2$, the pull at a pulled position is still a couple of hundred times stronger than the push. An *unlabeled* state has no pull at all. It feels only the push, and goes wherever the task will let it.

Because the anchor is a mean rather than a sum, the per-position pull is also insensitive to *how many* labels a batch happens to carry: a batch with one labeled line pulls its positions as hard as a batch with three. The label rate only sets how often the term fires at all.

Discussion

This experiment demonstrated a selective anchor in a transformer: `op1-anti` clears the margin gate with a graded response against the M1-derived shape and no measurable task cost. That fills the gap in D2.1 for this testbed, and it did so without either of the two obvious next steps — a heavier pull (the ceiling arm shows that makes matters worse) or the remaining M1 repulsive terms.

Both candidate mechanisms from the ex-2.1.6 discussion turn out to be real, but their relative sizes are the reverse of what that discussion expected, with the tested hyperparameters. Narrowing the pull to the one position that carries the labeled color helps more than adding the M1 repulsive term, and the conditions that pull the whole span are the ones whose margin slides later in training.

However, the span pull is not an arbitrary choice we can simply drop. Sequence-level labeling forces it, because a document label marks no position as the relevant one. Our op1-only pull is possible only because this synthetic language has a known position that carries the concept, which is what a real corpus does not give you.

The repulsive term is worth keeping, even though it failed H4(a). It reduced the cube-wide drift but did not contain it, and $\bar{\alpha}$ climbed again once the anneal brought the weight down. But using a later anneal improved margin, containment, retention, and grading, by more than adding the term was worth in the first place. The M1 keyframes were inherited from a 5-dimensional autoencoder bottleneck and mapped onto our 100 epochs by fraction of training, so there was never a reason to expect them to be right for this architecture.

Possible follow-ups: - A sweep over weight ratios and anneal scheduling - Try different pooling over the span to allow uneven pull without having to specify positions up-front

One thing this experiment does *not* license. Redness stays as readable from the other 63 directions as it is in the control, unchanged from ex-2.1.6 in every condition. A selective anchor is not the same as an exclusive one, so nothing here bounds what suppressing the axis would do to the model's access to *red*. That is the question for the intervention experiments.

---

1. A design that runs every combination of two on/off factors, four conditions here, so each factor's effect can be measured on its own. ↩

2. One caution: "some condition" takes a maximum over the four factorial conditions, and the maximum of four noisy means sits about one noise unit above any single one. So a bare-partial result resting on exactly one condition at 0.35 is weaker than the same number on the pre-named primary condition. ↩

3. A main effect is a difference of two-condition means, so its noise is about the three-seed floor of 0.032, and 0.1 is about 3× that. ↩

4. Why 0.1: the ex-2.1.6 reference points are 0.53 for the bare term and 0.008 for the control, with a seed spread of about 0.02. Working repulsion should push $\bar{\alpha}$ to near-control levels; 0.1 is a fifth of the bare term and still five noise units above control, so a working mechanism passes comfortably and a merely-weakened runaway does not. ↩

5. In ex-2.1.6 the margin peaked near epoch 10 and slid back a quarter under steady pressure. If the slide was the rest of the cube catching up to a selective early response, repulsion should hold the peak. ↩

6. How good is that `open`-pair number in absolute terms? Not very, though the shortfall has nothing to do with the anchor. The floor (the best score any guesser could reach) is 0.086, and a guesser that flips a coin between the two names bracketing the true mix scores 0.115. At 0.134, these models are *worse* than that coin-flip guesser: they pick the nearest name about 35% of the time, versus its 50%. But they are far better than the 0.426 of a guesser that ignores the prompt. So the models do read the color; what is imperfect is the last step, choosing between two adjacent names. That is a property of the `v216` testbed, not of anchoring: the control sits at 0.133, right alongside the anchored conditions. ↩