Ex 2.1.12: fitted channel probes over the anchored checkpoints

Ridge probes fitted on the published checkpoints of the four D2.1 conditions, one per RGB channel, for each operand and for the answer, at every (slice, position) site. The probes find nothing before op2 in any condition, anchored or not, so the off-key tilt in the grading figures is a property of the probe set. No decodability cost of anchoring resolves. The maps also show the bare anchor giving up decodability in the late slices, and the rest of the recipe restoring it.

Findings

H1 (pre-op2 ceiling) — holds. Across all conditions, slices, and channels, the largest pre-op2 seed-mean strict R^2 for the op2 target is −0.56. That is well under the gate of 0.12. Every condition reads at or below the no-information floor that the fold design sets.

H2 (decodability) — not resolved. The R channel misses the one-sided 0.10 gate at all three sites, by 0.12 to 0.18, and every miss falls inside the preregistered resolution band, which runs from 0.53 to 0.96 (twice the pooled seed spread; both groups vary a lot from seed to seed, the control somewhat more). G and B pass the gate outright at every site, reading higher in the primary condition than in the scoring seeds of the control.

> **How to read this report**
>
> Preregistered: the hypotheses and their gates were frozen at the commit that added this file (`cdd18b46`), before the run. Results replaced the placeholders in place, and analyses conceived after seeing the data sit under Exploratory analyses, marked post hoc.

Why this experiment

The α maps of the D2.1 figures page key one measurement by either operand's color, and wherever a condition responds at all, the slot that is *not* the key tilts too.[1] For the slots that precede op2 that tilt cannot be a response: attention is causal, so the state above the first token is a function of that token alone, and the tilt is the on-key response composed through the probe set — a pair is on the set only if its mix lands back on the color grid, which correlates a color with its partners. The α statistic cannot show this cleanly, because it reads alignment with the anchor axis, which the un-anchored control ignores; its panels are flat for lack of any response to compose, and a reader cannot tell the composition story from the alignment story by looking.

Fitted probes remove that asymmetry. A probe fit at a site reads whatever color representation the model keeps there, whether that is the anchored one or the model's own. That puts every condition on the same footing: whatever any of them shows at the sites that precede op2 is capped by the information ceiling of the pairing, computable in advance (H1), and how much of that ceiling a condition reaches is a property of its embedding geometry rather than of anchoring. Ex-2.1.5 ran the same kind of scan on an un-anchored two-form model, and there the op2 probes read zero at op1 sites. That experiment sampled its lines over distinct operand pairs, so the operands were all but uncorrelated and the ceiling was effectively zero. Comparing the two probe sets is what makes the case: the tilt follows the pairing, and the model has no say at these sites.

The same scan answers a question the D2.1 experiments left open. Anchoring reshapes the stream — that is what it is for — and the task gates only showed that *accuracy* survives. Whether the ordinary, linearly readable color geometry survives too is the bounded-side-effects claim read through a probe, and per-channel R^2 maps against the un-anchored control measure it (H2).

Method

No training. Every run is a published checkpoint of the four D2.1 conditions, probed on its own experiment's published probe set (the four sets are identical, which the pipeline asserts): 5,832 lines, every color as op1 against each of the 27 partners whose mix stays on the grid.

Sites and targets. The residual stream is captured at 5 slices (embedding, then each of the 4 layers) × 6 token positions (`op1 + op2 = ans \n`). At each site, ridge probes ($\ell_2 = 10^{-2}$) decode three targets: the RGB triple of op1, of op2, and of the answer. Probes are fit and scored per site independently, as in ex-2.1.5.

Holdout. The strict per-value protocol of ex-2.1.5 (`sca.compute.geometry.strict_r2`): to score a channel at a level, every line holding that level in that channel of *any* role — either operand or the answer — leaves the fit together, so the probe must place an unseen level from the others. R^2 is reported per channel from the out-of-fold predictions, and the prediction curves the figures draw are those same out-of-fold predictions, averaged per color keyed by either operand in turn (the two margins the D2.1 figures use).

Teacher forcing. The ans and `\n` positions have the answer token in the input, so readings there include what the token itself carries; the embedding is the surface-text control, as in ex-2.1.5.

Noise floor. The primary condition has nine seeds and the others three. The pooled between-seed standard deviation of per-site R^2, computed per condition before any comparison is read, sets the resolution: a difference smaller than twice that floor is reported as not resolved.

Conditions

The four conditions of the D2.1 progression, from their published checkpoints: the un-anchored control and the bare anchor at $\lambda = 0.1$ (ex-2.1.6, 3 seeds each), the anchor plus the anti-subspace term at its tuned operating point (ex-2.1.8 `end90-hold30`, 3 seeds), and the full recipe with pooled either-operand labels (ex-2.1.10 `either-t100`, 9 seeds).

Hypotheses

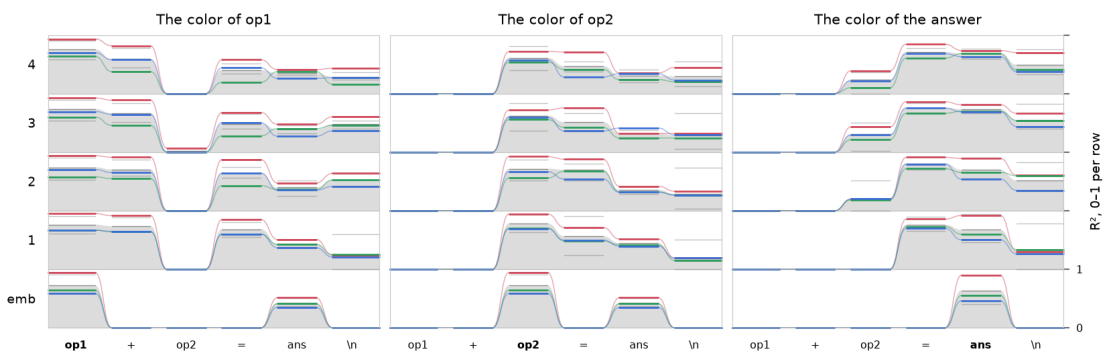
- **H1.** At the two sites that precede op2 — the op1 and `+` positions, every slice — the seed-mean strict R^2 of the op2-target probe stays below 0.12 in every condition and every channel. This is deliberately a protocol check rather than a discovery: the state at these sites is a function of the tokens before op2, so the best any probe can do is E[op2 | op1], and the closure rule caps that at $R^2 = 0.25 / (35/12) \approx 0.086$ per channel, whatever the model; the gate adds margin for estimation noise. A reading at or above 0.12 would mean the probe found more op2 information before op2 than the pairing supplies — a leak in the protocol or a misread of the probe set. No lower gate is stated, because every outcome below the ceiling is consistent with the bookkeeping account: how much of the ceiling each condition reaches is a property of its embedding geometry (how linearly it exposes level parity), read descriptively under E1, and a cross-condition difference there would be an observation about geometry that does not reopen the causal question.

- **H2.** Anchoring does not cost linear color decodability where the control has it. Site selection is separated from scoring so the selection noise stays off the gate: for each target, the comparison site is the (slice, position) with the best strict R^2 (mean over channels) in the map of the control's seed 0 alone, ties broken toward the earlier slice and then the earlier position, among targets whose selection map reaches $R^2 \geq 0.5$ (a target below that is reported unscored). The score then compares the seed-mean of the control's remaining seeds against the seed-mean of all nine primary seeds. At each selected site, the primary's per-channel R^2 comes within 0.10 of the control's (one-sided: the primary may exceed it). A miss smaller than twice the pooled between-seed standard deviation at that site is reported as not resolved rather than as a failure. Partial: exactly one channel of one target misses by more than that and by no more than 0.2. The intermediate conditions are reported at the same sites without a gate.

The maps

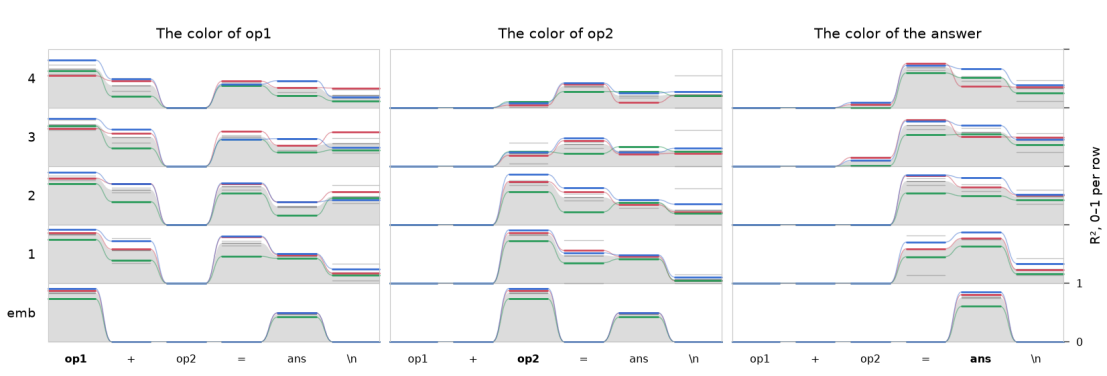
One figure per condition carries every reading the sections below score: its three panels are the probe targets, and within a panel the slices stack upward from the embedding, as the α grids on the figures page do.

un-anchored



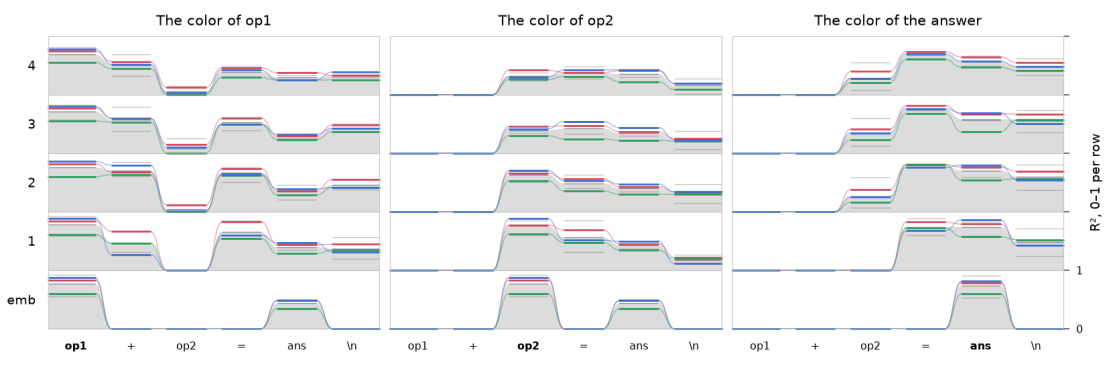
un-anchored *(ex-2.1.6,* `lam0` *, 3 seeds). Per-channel strict probe R^2 at every site: one row per residual slice, with the embedding at the bottom, and the six token positions across. Each RGB channel draws as a mark per site over the grey area of their mean, with faint risers between them, since the sites are where the measurements are; the grey marks are the seed envelope, so where the seeds disagree they stand off the silhouette. The three panels decode three targets, and the bold slot on the x axis is the token whose color that panel reads. Negative scores clip to the floor — a probe that does worse than predicting the mean has failed, and the H1 prose carries the raw values. Each row spans the same 0–1 scale, ticked once at the bottom right.*

+ anchor



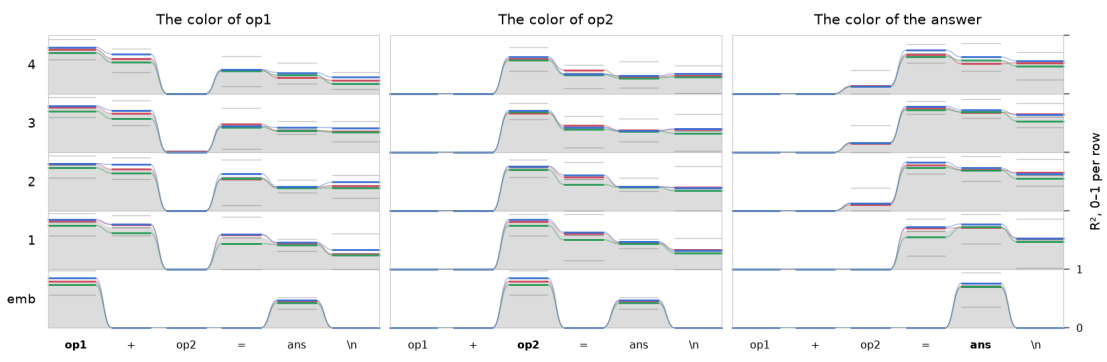
+ anchor *(ex-2.1.6,* `lam0.1` *, 3 seeds). Per-channel strict probe R^2 at every site: one row per residual slice, with the embedding at the bottom, and the six token positions across. Each RGB channel draws as a mark per site over the grey area of their mean, with faint risers between them, since the sites are where the measurements are; the grey marks are the seed envelope, so where the seeds disagree they stand off the silhouette. The three panels decode three targets, and the bold slot on the x axis is the token whose color that panel reads. Negative scores clip to the floor — a probe that does worse than predicting the mean has failed, and the H1 prose carries the raw values. Each row spans the same 0–1 scale, ticked once at the bottom right.*

+ anti-subspace



+ anti-subspace *(ex-2.1.8,* `end90-hold30` *, 3 seeds). Per-channel strict probe R^2 at every site: one row per residual slice, with the embedding at the bottom, and the six token positions across. Each RGB channel draws as a mark per site over the grey area of their mean, with faint risers between them, since the sites are where the measurements are; the grey marks are the seed envelope, so where the seeds disagree they stand off the silhouette. The three panels decode three targets, and the bold slot on the x axis is the token whose color that panel reads. Negative scores clip to the floor — a probe that does worse than predicting the mean has failed, and the H1 prose carries the raw values. Each row spans the same 0–1 scale, ticked once at the bottom right.*

+ pooled labels



+ pooled labels *(ex-2.1.10,* `either-t100` *, 9 seeds). Per-channel strict probe R^2 at every site: one row per residual slice, with the embedding at the bottom, and the six token positions across. Each RGB channel draws as a mark per site over the grey area of their mean, with faint risers between them, since the sites are where the measurements are; the grey marks are the seed envelope, so where the seeds disagree they stand off the silhouette. The three panels decode three targets, and the bold slot on the x axis is the token whose color that panel reads. Negative scores clip to the floor — a probe that does worse than predicting the mean has failed, and the H1 prose carries the raw values. Each row spans the same 0–1 scale, ticked once at the bottom right.*

The off-key ceiling (H1)

H1 holds, and the gate is not approached. Over every condition, slice, and channel, the largest pre-op2 reading for the op2 target is -0.56, against a gate of 0.12. The maps clip negatives to the floor, so the raw values are quoted here; the sign is what matters.

Under this holdout, a probe with no information cannot even reach $R^2 = 0$. Scoring a level removes every line carrying it from the fit, so such a probe predicts the training mean of each fold, and that mean sits away from the held-out level by construction. On this level structure it lands at −0.56. The fold combinatorics fix that number, not any model.

A site whose state never varies pins the same number empirically. At the embedding, the `+` state is a single token embedding, and it is a different vector in every condition. Even so, all four conditions read -0.5594 there, agreeing exactly. The op1 sites read lower still, -1.59 to -1.34. At those sites the probe can read the level of op1 itself and extrapolate op2 from the within-fold regression, and under this holdout that extrapolation points the wrong way. E1 traces the mechanism.

So no condition, anchored or not, carries op2 information that a linear probe can use before op2. The off-key tilt in the α maps comes from the pairing of the probe set, as the figures page reads it.

Decodability under anchoring (H2)

The deciding comparison is one row per condition per target, at the site the control's selection seed picked for that target.

target	site (control's pick)	condition	R	G	B	ΔR ↓	ΔG ↓	ΔB ↓	2×σ
op1	slice 1, op1	un-anchored (seeds 1–2)	+0.95	+0.58	+0.53				
		+ anchor	+0.86	+0.75	+0.91				
		+ anti-subspace	+0.84	+0.61	+0.89				
		+ pooled labels (primary)	+0.82	+0.71	+0.85	+0.12	**-0.13**	**-0.32**	0.59
op2	slice 1, op2	un-anchored (seeds 1–2)	+0.93	+0.62	+0.56				
		+ anchor	+0.86	+0.72	+0.91				
		+ anti-subspace	+0.77	+0.62	+0.88				
		+ pooled labels (primary)	+0.82	+0.71	+0.85	+0.12	**-0.09**	**-0.29**	0.53
ans	slice 1, ans	un-anchored (seeds 1–2)	+0.89	+0.33	+0.25				
		+ anchor	+0.77	+0.64	+0.87				
		+ anti-subspace	+0.79	+0.30	+0.86				
		+ pooled labels (primary)	+0.71	+0.61	+0.77	+0.18	**-0.28**	**-0.52**	0.96

The H2 comparison at the control-chosen sites. *Δ is the control's scoring-seed mean minus the condition's seed-mean, per channel, for the gated (primary) row only; lower is better and values at or below the one-sided gate of 0.1 are bold. 2×σ is twice the pooled between-seed standard deviation at the site — the preregistered resolution: a positive Δ smaller than it reads as not resolved. The intermediate conditions carry no gate.*

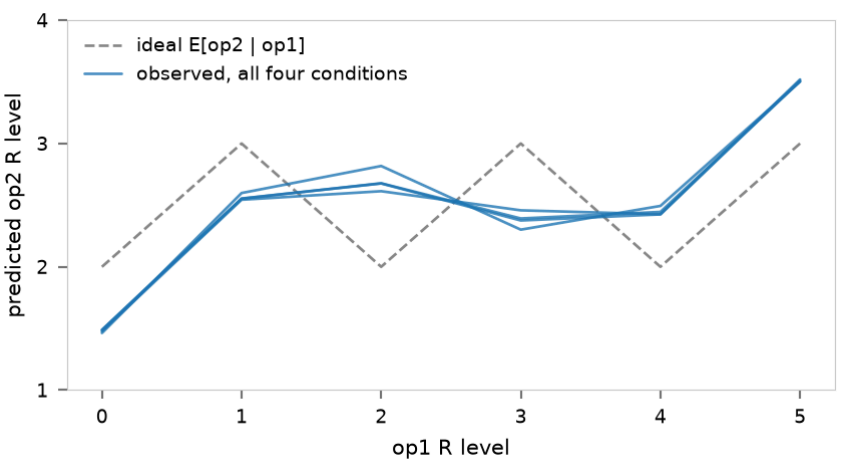
H2 does not resolve. The R channel misses the 0.1 gate at all three sites, by 0.12 to 0.18. Every miss sits far inside its resolution band, so under the frozen rule each reads as not resolved rather than as a failure — and never as a pass, so the per-channel claim as a whole is neither established nor refuted. No other channel misses. G and B run the other way: at every site the primary condition reads above the scoring seeds of the control, by up to 0.52.

Seed variance is what limits the comparison, and the control carries slightly more than half of it. Its three seeds disagree about how linearly color reads, with channel-mean R^2 at the op1 site running 0.61, 0.76, 0.91 across them, and its scored mean rests on two seeds; the nine primary seeds span 0.43 to 0.96 themselves. A deficit would have had to be about half an R^2 unit to resolve, so only a large cost could have been caught.

The best control site for the ans target is the ans position itself, which the selection rule was allowed to pick. Teacher forcing puts the answer token in the input there, so both models read the embedding of the token as much as anything they computed. The comparison is still like for like.

Exploratory analyses

E1 — the strict estimator rules out the parity route. The preregistered comparison expected the probes to trace the parity sawtooth E[op2 level | op1 level] = 2, 3, 2, 3, 2, 3. They do nothing of the kind, and in hindsight they cannot: scoring a level under the strict holdout removes it from every role, so the very lines a parity reader would generalize from leave the fit along with it. (That reading of the fold design is post hoc; E1 as preregistered expected the sawtooth.)



The off-key prediction at the op1 site. *Mean out-of-fold prediction of op2's R channel, in level units, against op1's R level, at the embedding slice: one solid line per condition, with the parity sawtooth — the ideal predictor E[op2 | op1] — dashed. The four conditions nearly coincide.*

What the probes read instead is the level of op1, which is decodable, and they extrapolate op2 from the within-fold regression. The out-of-fold R-channel prediction of the control runs 1.49, 2.55, 2.68, 2.38, 2.43, 3.51 in level units across the R level of op1, a ramp, and the curves of the four conditions agree within 0.20 level units. That extrapolation is what pushes the op1-site R^2 below even the constant-state floor. This estimator cannot say whether any embedding exposes parity linearly; it says only that the strict off-key reading holds no op2 information in any model.

E2 — the R channel tracks α loosely. Correlations over colors at the op1 site, averaged over slices: across the three anchored conditions the R channel reads +0.43, +0.43, +0.44 against the α lookup, and G and B fall between -0.34 and -0.30. α itself tracks the $\text{sim}^{1.5}$ statistic at 0.80–0.91. As preregistered, no channel reproduces α, because α carries redness, which is a function of all three channels rather than of any one.

E3 — no drag signature in the readout of the model itself. In every condition, the control included, the on-key prediction errors at the emb slice correlate with the partner-mean of the affinity at $|r| \leq 0.08$. They correlate with the affinity itself only weakly, and to almost the same degree in each condition (-0.38 to -0.24 for the R channel). So the drag-versus-direct ordering that shows up along the anchor axis does not reappear in RGB decodability: whatever the op1-keyed labels dragged onto the axis, they did not measurably bend the color geometry of the embeddings.

E4 — the bare anchor gives up late-slice decodability, and the recipe restores it (post hoc). Channel-mean op2 decodability at its own position, last slice: control 0.61, bare anchor 0.03 (seeds -0.04, +0.02, +0.11), anti-subspace 0.32, primary 0.60. The indiscriminate lift of the bare anchor crowds linear color readout out of the top of the stack. The anti-subspace term restores about half of it, and the full recipe matches the control. The op2 panel of each condition's map shows the whole shape.

Discussion

The ceiling result supplies the half of the demonstration that the α statistic could not. α reads an axis the un-anchored control ignores, so the control's flat panels were consistent with two stories: either there was no response to compose, or the measurement only sees anchored models. Fitted probes read the geometry of each model itself, and in all four conditions they find the same nothing before op2, down to the same fold-design floor.

On side-effects, the gated comparison says less than we would like. The unresolved gaps point both ways, with R slightly favoring the control and G and B favoring the primary. Resolving a difference of about 0.1 would take more un-anchored seeds than the three that exist.

The post-hoc E4 observation is the sharpest new fact. The bare anchor costs late-slice decodability outright, and the containment half of the recipe, first the anti-subspace term and then pooled labels, restores it. That agrees with what ex-2.1.8 and ex-2.1.10 measured along the anchor axis as drift and its containment, read here in the coordinates of the model instead. It is a single observation on three-seed conditions and it is marked post hoc; a preregistered version would have to fix its sites in advance.

---

1. α is the cosine between the residual stream (the running vector each layer reads from and writes back to) and the anchor axis, the direction the concept is trained toward. ↩