# Ex 2.2.3: the multi-op grammar, with *red* anchored again

> The grammar grows from one operation to six: `mix` (`+`), `add`, `screen`, `multiply`, `lighten`, `darken`. We retrain on the new grammar to check the recipes: the un-anchored control, the ex-2.1.10 recipe, and three proposals from the ex-2.1.11 survey. The recipe carried over as it was. The proposals place *red* with a wider margin, and the frozen rule adopted one of them, but they put the axis on the syntax embeddings, so projecting it out breaks most non-red lines too; a post hoc read with the gates the rule left out narrows the choice to the recipe, and D2.2 builds on its short arm, which grades better for half the compute. Removal reads as partial on the saturating ops because their answers mostly do not depend on how red the red operand is.

## Findings

- Task cost on the new grammar (H1) — **holds.** All 30 condition-op comparisons are within 0.02 of the control of the same length; the largest gap is -0.0094, `t00` on `mix`.
- Placement under the recipe reproduces (H2) — **holds.** Every placement criterion holds on the `mix` lines at five fresh seeds: seed-mean m_line 0.402, 0.96 of its ex-2.1.10 value, grading $r^2$ 0.819, contrast 0.86, containment 0.062, no latched run.
- The plateau from the survey transfers (H3) — **holds.** A feasible proposal exceeds `recipe-short` (0.401) by more than a band (0.011): `t00` +0.075, `t48` +0.062, `t12` +0.059. Adopted: `t00`. The frozen rule adopts **`t00`**; a post hoc amendment that adds H2's lead gate and H4's selectivity gate, with the tie between the recipe lengths broken at twenty seeds (E6), adopts **`recipe`**. D2.2 builds on **`recipe-short`** by decision: it grades better and leads better at the embedding, with the same H4 read, for half the compute.
- Suppression transfers (H4) — **contrary.** `recipe`: red accuracy under projection 0.06–0.32 across the ops, `mix` non-red deficit 0.024; `t00`: red accuracy under projection 0.00–0.06 across the ops, `mix` non-red deficit 0.602. Removal misses on `add, screen, lighten, darken` for `recipe`. The `mix` deficit on `t00` is above 0.1.

> **How to read this draft**
>
> The hypotheses, the gates, the selection rule, and the method were frozen at commit `5a3233b`, before any anchored run, with the calibration recorded under the method. Each hypothesis section opens with its frozen prediction; its results follow in place. Anything we thought of after seeing the data is under Exploratory analyses, marked as post hoc. The op table and every count in the method are computed from `experiment.py` at render time.

# Why this experiment

The claims of D2.2 are about an operation, and the D2.1 grammar has only one: `mix`, spelled `+`. The D2.2 plan makes the operation a variable. But every number D2.1 produced was measured on the one-op grammar: the recipe in ex-2.1.10, the operating-point plateau in the ex-2.1.11 survey, and the suppression figures in ex-2.2.1. The risk table in the plan names this ("recipe is grammar-specific") and sends it here.

So this is a regression check with a decision attached. We measure the same statistics on the same `mix` lines, and we freeze a rule that turns the proposals from the survey into the operating point the anchored-op experiments adopt.

## Conditions

Every condition is trained here at fresh seeds. All but the richer-op arms use the six-op corpus at D2.1's size and epoch count, so the recipe's step count is unchanged.

| condition | what | anchor | ops | lines per op | epochs (steps) | seeds |
|---|---|---|---|---|---|---|
| **control** | un-anchored | none | all six | 16,666 | 100 (3,300) | 5 |
| **control-short** | un-anchored, at the proposals' length | none | all six | 16,666 | 50 (1,650) | 5 |
| **recipe** | the ex-2.1.10 recipe | λ_a = 0.100 (annealed), τ = 0.100, anti 2.50 → 0.30 by 90% | all six | 16,666 | 100 (3,300) | 5 |
| **recipe-short** | the recipe at the proposals' length | λ_a = 0.100 (annealed), τ = 0.100, anti 2.50 → 0.30 by 90% | all six | 16,666 | 50 (1,650) | 5 |
| **t00** | survey proposal t00 | λ_a = 0.557 (flat), τ = 0.265, anti 1.38 → 0.30 by 54% | all six | 16,666 | 50 (1,650) | 5 |
| **t48** | survey proposal t48 | λ_a = 0.431 (flat), τ = 0.222, anti 2.95 → 0.30 by 55% | all six | 16,666 | 50 (1,650) | 5 |
| **t12** | survey proposal t12 | λ_a = 0.361 (flat), τ = 0.233, anti 1.13 → 0.30 by 92% | all six | 16,666 | 50 (1,650) | 5 |
| **ops-1-corpus** | un-anchored, mix only, D2.1's corpus size | none | `mix` | 100,000 | 100 (3,300) | 3 |
| **ops-3-corpus** | un-anchored, mix, add, multiply, D2.1's corpus size | none | `mix`, `add`, `multiply` | 33,333 | 100 (3,300) | 3 |
| **ops-1-per-op** | un-anchored, mix only, six-op lines per op | none | `mix` | 16,666 | 550 (3,300) | 3 |
| **ops-3-per-op** | un-anchored, mix, add, multiply, six-op lines per op | none | `mix`, `add`, `multiply` | 16,666 | 194 (3,298) | 3 |
| **recipe-more** | the recipe, addendum seeds | λ_a = 0.100 (annealed), τ = 0.100, anti 2.50 → 0.30 by 90% | all six | 16,666 | 100 (3,300) | 15 |
| **recipe-short-more** | the recipe at 50 epochs, addendum seeds | λ_a = 0.100 (annealed), τ = 0.100, anti 2.50 → 0.30 by 90% | all six | 16,666 | 50 (1,650) | 15 |

The recipe is the D2.1 primary, `either-t100`. The only thing we changed is the grammar. The short arm of the recipe runs for as long as the proposals do, so H3 compares runs of equal length; the control has a short arm so that H1 can do the same.

The three proposals are the operating point from the survey plus the two promoted trials the survey read as its twin and its knee. `t48` sits within one m_line band of `t00`, with more grading and contrast. `t12` holds the most grading of the promoted set, at a margin the survey could not tell apart from the one for `t00`.

The richer-op arms are un-anchored models on one and three ops at the recipe's step count, in two matchings. The `corpus` arms keep D2.1's line count, and so have more lines per op. The `per-op` arms keep the six-op lines per op, and so have a smaller corpus, repeated over more epochs. Corpus size, lines per op, and step count cannot all be held while the op set changes, and the two matchings put the remaining difference on opposite sides, so the pair lets E4 read in both directions. They serve E4 only.

### The interventions

We score the checkpoints of every candidate through the eval contract in `sca.intervention`, on the probe lines of each op.

- **`projection`** — the primary from ex-2.2.1: the axis projected out at every slice and position, at full strength. Read by H4.
- **Ride-along rows** — `operands` (the projection at the two operand positions only), `shaped` (the shaped suppression from M1, a = 0.5, b = 1, p = 1), and `ablate` (the axis weights zeroed). Reported without gates, so the intervention-tuning pass the plan schedules has the multi-op figures beside the one-op figures from ex-2.2.1.

# Task cost on the new grammar (H1)

**H1.** Anchoring costs nothing on the task, on any op. We score the recipe at both lengths and the three proposals against the control of the same length, op by op, so the gate reads 30 condition-op comparisons. In each, seed-mean holdout exact-match accuracy is within 0.02 of its control's. Partial: every comparison within 0.05, or all but one within 0.02 and that one within 0.05. Contrary: a condition more than 0.05 below its control on some op, which would say that the anchor and reading the op compete with each other. A condition that far *above* its control on some op misses the gate too, and we would report it as an anomaly, since nothing in the design predicts one.

This gate only means something if the controls learn the grammar in the first place. We check that at both lengths in the calibration runs described in the method, before the freeze, rather than gating it here.

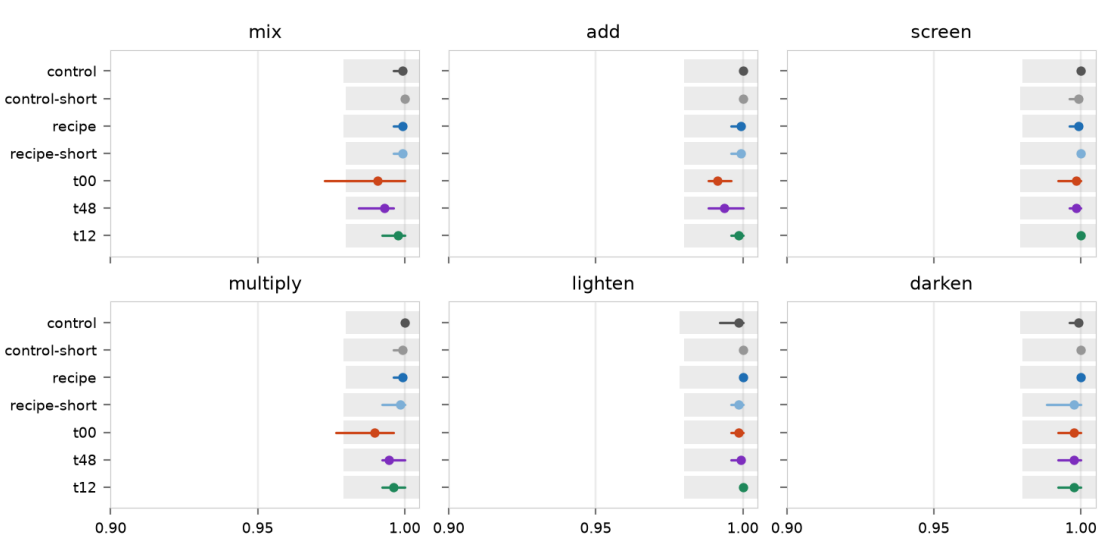
If the recipe misses this gate at full length, H2 goes unscored and the refuted H1 is the finding, as ex-2.1.10 ruled for its primary. H3 still runs, since its selection rule already requires the task gate of every candidate, and H4 is then scored on the adopted point alone. If no candidate is feasible, H4 goes unscored too.

| condition | mix EM ↑ | add EM ↑ | screen EM ↑ | multiply EM ↑ | lighten EM ↑ | darken EM ↑ |
|---|---|---|---|---|---|---|
| control | 0.999 ±0.002 | 1.000 ±0.000 | 1.000 ±0.000 | 1.000 ±0.000 | 0.998 ±0.004 | 0.999 ±0.002 |
| control-short | 1.000 ±0.000 | 1.000 ±0.000 | 0.999 ±0.002 | 0.999 ±0.002 | 1.000 ±0.000 | 1.000 ±0.000 |
| recipe | 0.999 ±0.002 | 0.999 ±0.002 | 0.999 ±0.002 | 0.999 ±0.002 | 1.000 ±0.000 | 1.000 ±0.000 |
| recipe-short | 0.999 ±0.002 | 0.999 ±0.002 | 1.000 ±0.000 | 0.998 ±0.004 | 0.998 ±0.002 | 0.998 ±0.006 |
| t00 | 0.991 ±0.014 | 0.991 ±0.004 | 0.998 ±0.004 | 0.990 ±0.010 | 0.998 ±0.002 | 0.998 ±0.004 |
| t48 | 0.993 ±0.006 | 0.994 ±0.006 | 0.998 ±0.002 | 0.995 ±0.004 | 0.999 ±0.002 | 0.998 ±0.004 |
| t12 | 0.998 ±0.004 | 0.998 ±0.002 | 1.000 ±0.000 | 0.996 ±0.004 | 1.000 ±0.000 | 0.998 ±0.004 |
| ops-1-corpus | 0.999 ±0.002 | · | · | · | · | · |
| ops-3-corpus | 1.000 ±0.000 | 0.997 ±0.002 | · | 1.000 ±0.000 | · | · |
| ops-1-per-op | 0.917 ±0.014 | · | · | · | · | · |
| ops-3-per-op | 0.990 ±0.008 | 0.991 ±0.006 | · | 0.987 ±0.006 | · | · |
| recipe-more | 0.999 ±0.002 | 1.000 ±0.002 | 0.999 ±0.002 | 0.999 ±0.002 | 1.000 ±0.000 | 1.000 ±0.002 |
| recipe-short-more | 0.999 ±0.002 | 0.999 ±0.002 | 1.000 ±0.000 | 1.000 ±0.002 | 1.000 ±0.002 | 0.999 ±0.004 |

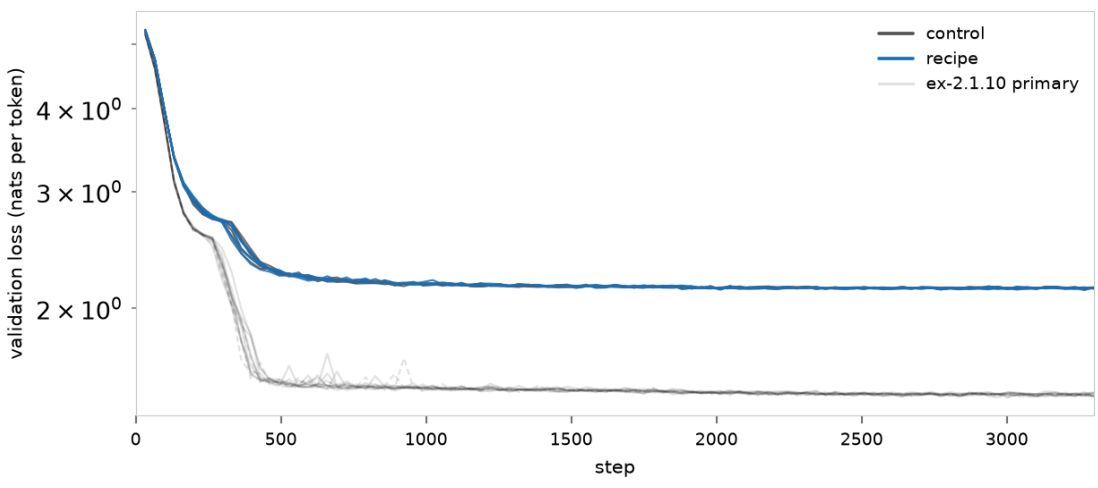
*Holdout exact match per op, by condition: the seed mean with half the seed range beside it. The greyed rows are the two controls (100 and 50 epochs) and the richer-op arms, which carry no gate; a dot means the arm never trained on that op.*

| condition | control | mix gap | add gap | screen gap | multiply gap | lighten gap | darken gap | worst |
|---|---|---|---|---|---|---|---|---|
| recipe | control | **+0.0000** | **-0.0008** | **-0.0008** | **-0.0008** | **+0.0016** | **+0.0008** | **+0.0016** |
| recipe-short | control-short | **-0.0008** | **-0.0008** | **+0.0008** | **-0.0008** | **-0.0016** | **-0.0023** | **-0.0023** |
| t00 | control-short | **-0.0094** | **-0.0086** | **-0.0008** | **-0.0094** | **-0.0016** | **-0.0023** | **-0.0094** |
| t48 | control-short | **-0.0070** | **-0.0062** | **-0.0008** | **-0.0047** | **-0.0008** | **-0.0023** | **-0.0070** |
| t12 | control-short | **-0.0023** | **-0.0016** | **+0.0008** | **-0.0031** | **+0.0000** | **-0.0023** | **-0.0031** |

*The H1 read: each scored condition's seed-mean holdout exact match minus that of the control of the same length, per op. Bold marks a gap within the gate of 0.02; the partial band runs to 0.05.*




***Holdout exact match per op, against the gate.*** *One panel per op; conditions down the side, in their own ink. Each dot is the seed mean, the bar the seed range. The grey strip behind a row is ±0.02 around the seed-mean accuracy of the control of the same length, so a dot outside its strip misses the H1 gate; the two control rows carry their own strips for reference.*




***Validation loss over training.*** *Five thin lines per condition on the step axis: the control in grey and the recipe in blue. The ghost lines behind them are the ex-2.1.10 primary (nine seeds) and its control (three), which trained on the one-op grammar at the same step count; their floor sits lower because their corpus held about seventeen copies of each distinct line, where this one holds most lines once. Log scale.*

**H1, holds.** All 30 condition-op comparisons are within 0.02 of the control of the same length; the largest gap is -0.0094, `t00` on `mix`.

# Placement under the recipe reproduces (H2)

**H2.** The ex-2.1.10 recipe places *red* on the new grammar as it did on the old one. On the `mix` probe lines, with five fresh seeds, all of the ex-2.1.10 placement gates hold. Each statistic is a seed mean unless the criterion says otherwise: containment $\bar{\alpha}$ at op1 at most 0.1; a leading softmin weight of the red group at the embedding of at least 0.4; contrast at least 0.2; grading $r^2$ no more than 0.1 below the 0.782 the survey measured on this recipe; m_line at least 0.8 of its ex-2.1.10 value of 0.420; and, read per run rather than on the seed mean, retention at least 0.8 of the running peak for every run whose peak reaches 0.2, and no latched run.

Partial: every criterion holds except one of m_line in the 0.6–0.8 band or contrast in the 0.1–0.2 band. Contrary: a latch or a containment miss. That would mean something in the new grammar changed what the pull finds. If instead the margin fell below the partial band while the other criteria held, the placement would be there but weaker, and the H3 selection would lean toward the proposals.

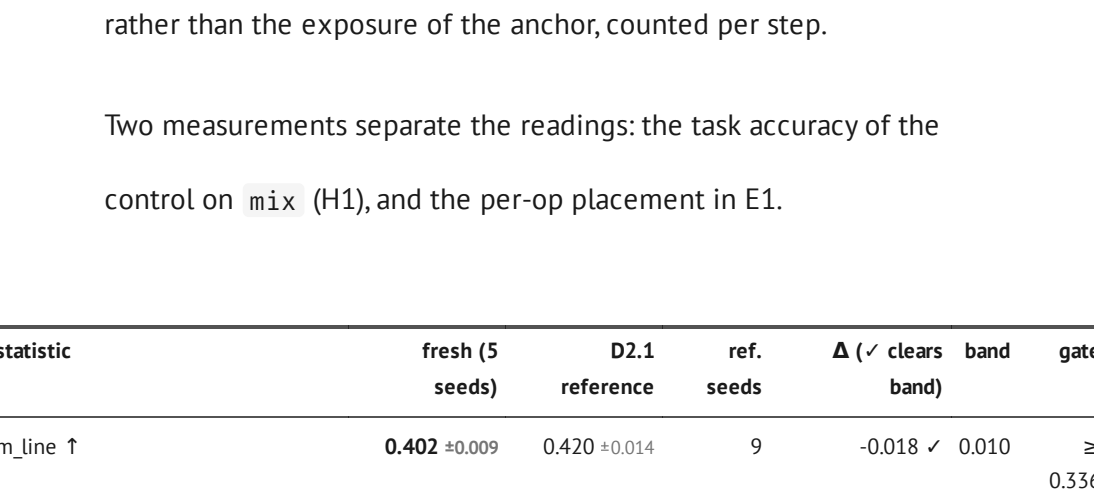
Two things in the new grammar could do that, and we read a miss against both. The first is the op word. The labeller never sees it, and reads the same two operand roles on every op.

The second is exposure. Only a sixth of the corpus is `mix`, and those lines are drawn from every pair rather than from the on-grid ones. So over a run the model sees a given `mix` probe pair about a fiftieth as often as in D2.1, and never sees most of them (see the corpus). What the anchor sees is closer to unchanged. The corpus is the same size, and the share of lines the labeller fires on follows the operand redness: 6.4% of all pairs, against 6.3% of the on-grid ones in D2.1. So the exposure that fell is the exposure of the task, counted per pair, rather than the exposure of the anchor, counted per step.

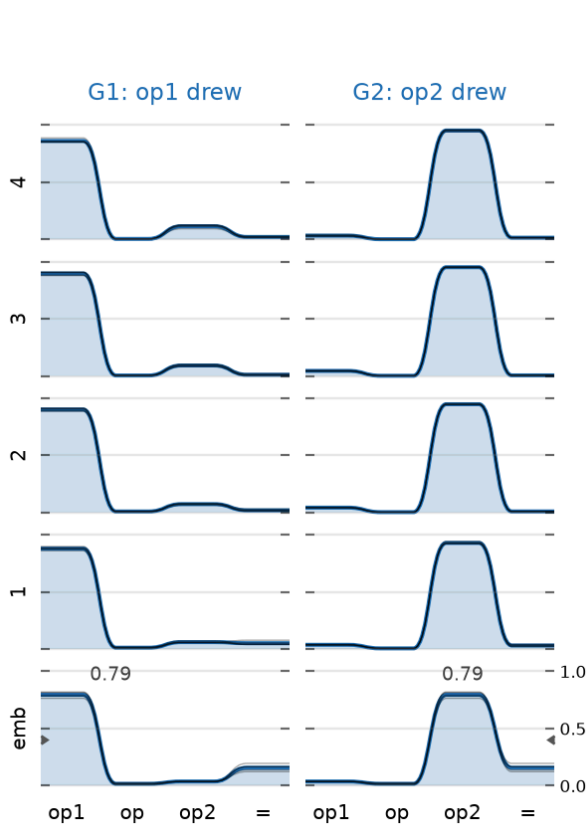
Two measurements separate the readings: the task accuracy of the control on `mix` (H1), and the per-op placement in E1.

| statistic | fresh (5 seeds) | D2.1 reference | ref. seeds | Δ (✓ clears band) | band | gate |
|---|---|---|---|---|---|---|
| m_line ↑ | **0.402** ±0.009 | 0.420 ±0.014 | 9 | -0.018 ✓ | 0.010 | ≥ 0.336 |
| m_span | 0.749 ±0.017 | 0.738 ±0.032 | 9 | +0.011 | – | – |
| containment ᾱ ↓ | **0.062** ±0.020 | 0.050 ±0.039 | 9 | +0.012 · | 0.024 | ≤ 0.1 |
| lead weight at emb ↑ | **0.79** ±0.03 | – | – | – | – | ≥ 0.4 |
| contrast ↑ | **0.859** ±0.005 | 0.856 | 3 (survey) | +0.003 · | 0.008 | ≥ 0.2 |
| grading r² ↑ | **0.819** ±0.036 | 0.782 | 3 (survey) | +0.037 ✓ | 0.035 | ≥ 0.682 |
| retention, min over runs ↑ | **0.994** ±0.005 | 0.988 ±0.010 | 9 | +0.005 · | 0.007 | ≥ 0.8 |
| non-red op1 weight, max over runs ↓ | **0.17** ±0.04 | – | – | – | – | ≤ 0.5 |

*The placement statistics of the recipe on the `mix` probe lines, five fresh seeds, beside the D2.1 reference. The reference is the ex-2.1.10 primary's nine seeds where its stored metrics carry the statistic, and the survey's three-seed re-run of the same recipe (its `ref` arm) for contrast and grading, which ex-2.1.10 did not store per run; the lead weight and the latch have no stored reference. Δ is fresh minus reference; a ✓ beside it means the difference clears the band ($2\sigma\sqrt{(1/5 + 1/n)}$ with the frozen per-run σ), a dot that it does not. Bold marks a value inside its gate. Retention and the latch are read per run, so their rows print the worst run.*




***m_line over training.*** *The trajectory instrument reads one line per color on the `mix` probe set, every 50 steps. Five thin blue lines are the recipe's fresh seeds; the ghost lines behind them are the nine seeds of the ex-2.1.10 primary on the one-op grammar, on the same step axis. The grey lines are the un-anchored control. The dip over the last tenth is the anchor anneal.*




***Per-group softmin profiles for the recipe on the `mix` lines.*** *Columns are the label groups (G1 weights probe lines by P(only op1 drew), G2 by P(only op2 drew)); rows are residual slices with the embedding at the bottom; the shaded smooth-step is the seed-mean softmin weight over the span roles, with a hairline per seed. On the embedding row, the caret on each column's outer axis marks the lead gate (0.4; uniform is 0.25), and the number above each group's own operand is the seed-mean weight there. Deep-slice op2 contrast +0.86 ±0.00 (gate ≥ 0.2). The layout is ex-2.1.10's, whose primary put 0.77 on the own operand at the embedding and 0.9 or more at depth.*

**H2, holds.** Every placement criterion holds on the `mix` lines at five fresh seeds: seed-mean m_line 0.402, 0.96 of its ex-2.1.10 value, grading r² 0.819, contrast 0.86, containment 0.062, no latched run.

# The plateau from the survey transfers (H3)

**H3.** The plateau found in the survey belongs to the recipe rather than to the one-op grammar. At least one proposal is feasible at fresh seeds on the new grammar, and its seed-mean m_line on the `mix` probe lines exceeds `recipe-short` (the recipe at the same step count) by more than one band. Feasibility is read as the survey read it, with the constraints listed once in the selection rule below. The grading margin is one of them, and so is a floor on the worst seed: every run's m_line at least 0.6 of its ex-2.1.10 value, since the bands use the frozen per-run σ and a candidate's own spread would otherwise not count against it.

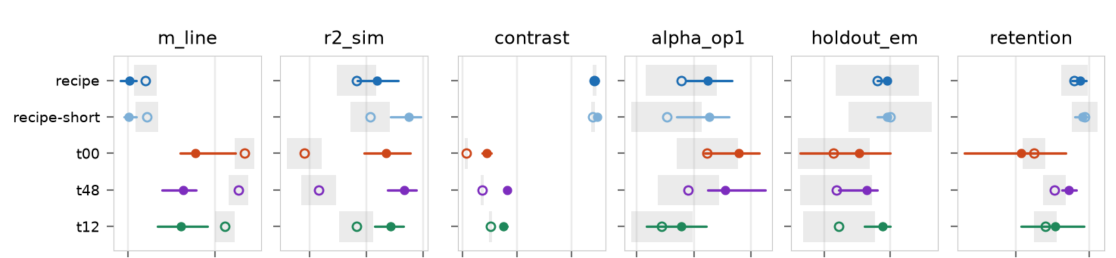
Partial: a proposal is feasible and its m_line sits within a band of `recipe-short`; or a proposal clears every survey constraint and exceeds `recipe-short` by more than a band, but misses the grading margin. Contrary: no proposal is feasible, or every feasible proposal sits more than a band below `recipe-short`. Either would say the plateau was specific to the one-op grammar, and the recipe would carry D2.2. A proposal that misses only the per-run floor is reported as unstable at fresh seeds rather than as weak, with the seed range printed.

**The selection rule.** Among the candidates (the recipe at both lengths and the three proposals), keep those feasible at fresh seeds. On the seed mean: task within the gate on every op against the control of the same length, containment, contrast at least the partial floor, and grading $r^2$ no more than 0.10 below the full-length recipe's fresh value, clearing that floor by at least one per-run σ. On every run: retention, no latch, and m_line at least the partial margin of its ex-2.1.10 value. Among those, require contrast at the full gate, and take the highest seed-mean m_line on the mix probe lines. Every candidate within one m_line band of that highest value ties with it, and the tie goes to the larger grading margin, then to the full-length recipe, then to `recipe-short`. If no other candidate is feasible, or none of the feasible candidates clears the full contrast gate, the full-length recipe is adopted, and if it is not itself feasible the report says which check it missed beside the adoption. Whichever point is adopted, the D2.2 experiments carry it. The Findings line for this section names the adopted point, and it lands in `experiment.py` for the D2.2 experiments that follow.

**Winner's curse.** The numbers from the survey are proposals. So the table prints the one-op value for each candidate beside the fresh one, with the per-run σ behind the bands and the band a difference has to clear. The survey re-ran the recipe at both lengths, and those runs sit in the same table, since they are the one-op values the fresh recipe arms are read against.

| candidate | statistic | survey | seeds | fresh (5 seeds) | Δ (✓ clears band) | σ per run | band |
|---|---|---|---|---|---|---|---|
| `recipe` | m_line | 0.420 | 3 | 0.402 ±0.009 | -0.018 ✓ | 0.0088 | 0.013 |
| | r2_sim | 0.782 | 3 | 0.819 ±0.036 | +0.036 ✓ | 0.0240 | 0.035 |
| | contrast | 0.856 | 3 | 0.859 ±0.005 | +0.003 · | 0.0057 | 0.008 |
| | alpha_op1 | 0.039 | 3 | 0.062 ±0.020 | +0.023 · | 0.0212 | 0.031 |
| | holdout_em | 0.996 | 3 | 0.999 ±0.002 | +0.003 · | 0.0087 | 0.013 |
| | retention | 0.990 | 3 | 0.994 ±0.005 | +0.004 · | 0.0061 | 0.009 |
| `recipe-short` | m_line | 0.422 | 3 | 0.401 ±0.007 | -0.020 ✓ | 0.0088 | 0.013 |
| | r2_sim | 0.806 | 3 | 0.875 ±0.027 | +0.068 ✓ | 0.0240 | 0.035 |
| | contrast | 0.849 | 3 | 0.869 ±0.009 | +0.020 ✓ | 0.0057 | 0.008 |
| | alpha_op1 | 0.026 | 3 | 0.063 ±0.023 | +0.037 ✓ | 0.0212 | 0.031 |
| | holdout_em | 1.000 | 3 | 0.999 ±0.002 | -0.001 · | 0.0087 | 0.013 |
| | retention | 0.997 | 3 | 0.995 ±0.004 | -0.002 · | 0.0061 | 0.009 |
| `t00` | m_line | 0.534 | 5 | 0.477 ±0.032 | -0.057 ✓ | 0.0088 | 0.011 |
| | r2_sim | 0.690 | 5 | 0.834 ±0.041 | +0.144 ✓ | 0.0240 | 0.030 |
| | contrast | 0.267 | 5 | 0.362 ±0.021 | +0.094 ✓ | 0.0057 | 0.007 |
| | alpha_op1 | 0.062 | 5 | 0.089 ±0.024 | +0.027 ✓ | 0.0212 | 0.027 |
| | holdout_em | 0.983 | 5 | 0.991 ±0.014 | +0.008 · | 0.0087 | 0.011 |
| | retention | 0.963 | 5 | 0.954 ±0.034 | -0.009 ✓ | 0.0061 | 0.008 |
| `t48` | m_line | 0.527 | 5 | 0.463 ±0.020 | -0.063 ✓ | 0.0088 | 0.011 |
| | r2_sim | 0.716 | 5 | 0.866 ±0.025 | +0.150 ✓ | 0.0240 | 0.030 |
| | contrast | 0.340 | 5 | 0.456 ±0.007 | +0.116 ✓ | 0.0057 | 0.007 |
| | alpha_op1 | 0.045 | 5 | 0.077 ±0.025 | +0.032 ✓ | 0.0212 | 0.027 |
| | holdout_em | 0.984 | 5 | 0.993 ±0.006 | +0.009 · | 0.0087 | 0.011 |
| | retention | 0.976 | 5 | 0.986 ±0.005 | +0.010 ✓ | 0.0061 | 0.008 |
| `t12` | m_line | 0.511 | 5 | 0.461 ±0.029 | -0.051 ✓ | 0.0088 | 0.011 |
| | r2_sim | 0.782 | 5 | 0.842 ±0.025 | +0.060 ✓ | 0.0240 | 0.030 |
| | contrast | 0.378 | 5 | 0.439 ±0.014 | +0.060 ✓ | 0.0057 | 0.007 |
| | alpha_op1 | 0.022 | 5 | 0.039 ±0.026 | +0.017 · | 0.0212 | 0.027 |
| | holdout_em | 0.984 | 5 | 0.998 ±0.004 | +0.013 ✓ | 0.0087 | 0.011 |
| | retention | 0.970 | 5 | 0.977 ±0.021 | +0.007 · | 0.0061 | 0.008 |

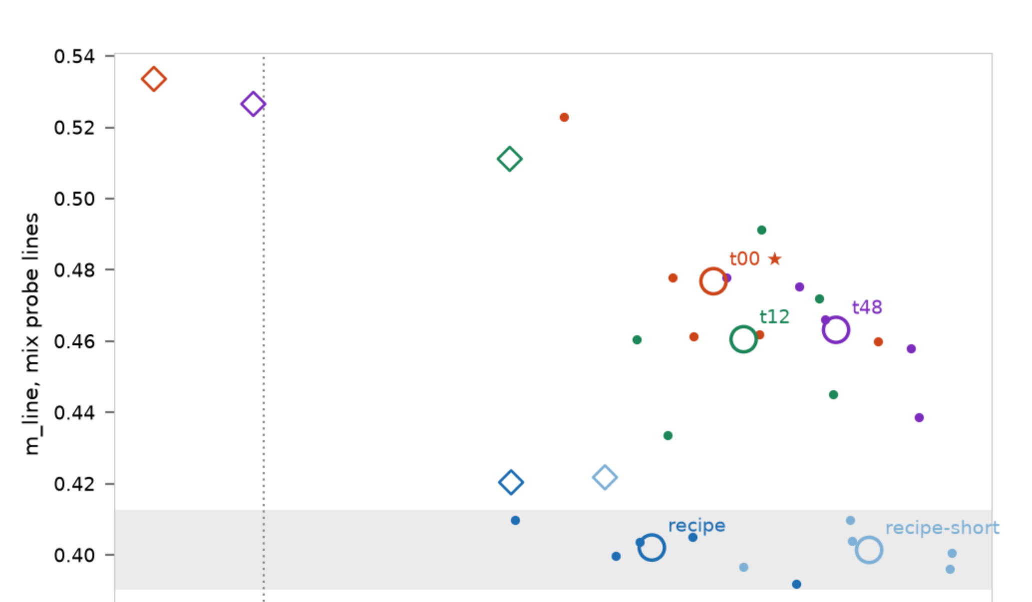
*Each candidate's one-op survey numbers beside the fresh ones on the six-op grammar. The survey column is the trial's five-seed mean (for the proposals) or the survey's own three-seed re-run of the recipe at that length; fresh is the seed mean here with half the seed range. Δ is fresh minus survey, with a ✓ where the difference clears the band for those seed counts and a dot where it does not. σ is the frozen per-run spread from ex-2.1.10 behind every band. `holdout_em` is read on `mix`.*



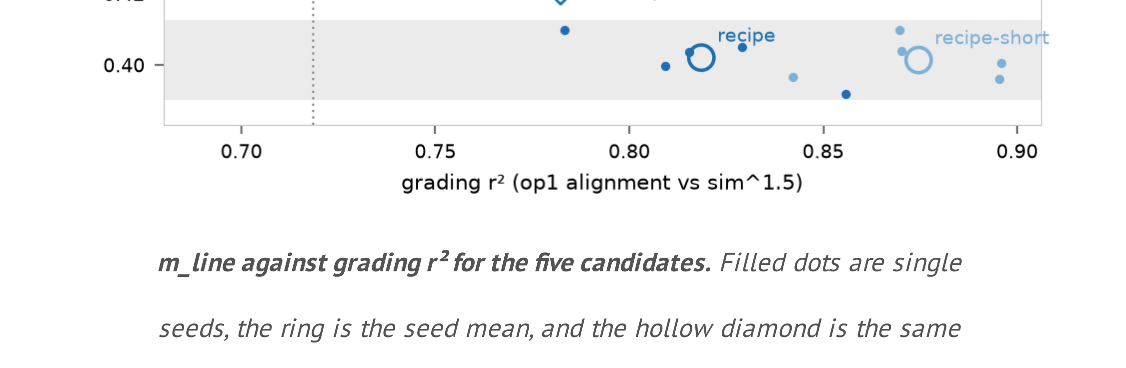

***Survey against fresh, per candidate.*** *The same numbers as the table above, drawn: the hollow marker is the survey value, the filled dot the fresh seed mean with a bar for the seed range, and the grey strip the band around the survey value for those seed counts. A dot outside its strip is a difference the resolution rule counts.*

| candidate | task | containment | contrast ≥ partial | grading | retention | no latch | m_line floor | feasible | contrast ≥ 0.2 | m_line ↑ | grading $r^2$ ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| recipe | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.402 ±0.009 | 0.819 ±0.036 |
| recipe-short | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.401 ±0.007 | 0.875 ±0.027 |
| t00 ★ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.477 ±0.032 | 0.834 ±0.041 |
| t48 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.463 ±0.020 | 0.866 ±0.025 |
| t12 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.461 ±0.029 | 0.842 ±0.025 |

*Feasibility, as the selection rule reads it. Seed-mean checks: task within 0.02 of the same-length control on every op; containment $\bar{\alpha} \le 0.1$; contrast ≥ 0.1; grading $r^2$ at least 1σ (0.024) above its floor of 0.719, the full-length recipe's fresh value less 0.1. Per-run checks: retention ≥ 0.8 of the running peak for every run whose peak reaches 0.2; no run with more than 0.5 of the non-red group's weight on op1; every run's m_line at least 0.6 of 0.420. The selection then requires contrast at the full gate and takes the highest m_line, ties within one band (0.011) going to the larger grading margin. ★ marks the adopted point: 1 candidate(s) within one band (0.011) of the highest feasible m_line (0.477).*




***m_line against grading $r^2$ for the five candidates.*** *Filled dots are single seeds, the ring is the seed mean, and the hollow diamond is the same candidate's one-op survey value. The grey strip is one band (0.011) either side of `recipe-short`'s seed mean, which the H3 verdict reads the proposals against; the dotted line is the grading floor of the selection rule (0.719, before its 1σ margin). The adopted point is labelled ★.*




***What grading looks like.*** *One panel per candidate: each point is one of the 216 grid colors, drawn in its own color, at its similarity-to-red target (sim^1.5) along the bottom and its seed-mean alignment at op1 up the side, averaged over the four post-attention slices. The $r^2$ in each title is the squared correlation of these two axes on the seed-mean cloud; the tables report the per-seed $r^2$ averaged, which is a little lower. A cloud that rises in a line is graded; one that steps from a floor to a ceiling is thresholded.*

**H3, holds.** A feasible proposal exceeds `recipe-short` (0.401) by more than a band (0.011): `t00` +0.075, `t48` +0.062, `t12` +0.059. Adopted: `t00`.

## Post hoc: the selection rule, with the gates it left out

This is a deviation, written after the numbers were read. The selection rule above was frozen with two holes in it, and the point it adopts falls through both.

The first hole is H2's lead gate. Feasibility carries every H2 gate except the lead weight of the red group at the embedding, which H2 gates at 0.4; per candidate that weight is `recipe` 0.79, `recipe-short` 0.83, `t00` 0.33, `t48` 0.46, `t12` 0.48. Add the gate, and the highest feasible m_line is `t48`.

The second hole is H4. The rule ranks on placement and never looks at what the plain projection costs, and on every proposal that cost is most of the non-red `mix` lines: the deficit under `projection` is `recipe` 0.024, `recipe-short` 0.014, `t00` 0.602, `t48` 0.498, `t12` 0.598, against H4's gate of 0.05. The cost sits on the syntax embeddings (E2, E7), and the operand-only edit routes around it, but an operating point that needs a routed edit before its first intervention is a poor base for the anchored-op experiments. Add H4's selectivity gate as well, and only the two recipe lengths remain; at five seeds they sit within a band of each other, and E6's twenty-seed read breaks the tie: at twenty seeds the full recipe leads by 0.012, more than the 0.007 band, so it takes the tie.

The amended rule adopts **`recipe`**. The frozen H3 verdict stands as written, and the proposals keep their reads: they place *red* with a larger margin, and they pay for it on the syntax embeddings. Whether that cost can be removed rather than routed around is the untied-readout question in the Discussion.

**D2.2 builds on `recipe-short`, by decision rather than by rule.** Both rules rank on margin, and at twenty seeds the full recipe leads on margin by 0.012, about two bands. The short arm is ahead on the rest of what the anchored-op experiments will lean on: it grades better ($r^2$ 0.886 against 0.821), puts more of the pull on the drawing operand at the embedding (0.83 against 0.79), has a little more contrast, gives the same H4 read, and costs half as much to train. Those experiments read the grading and the operand lead more than they read the margin, so the graded point is the better base. The decision was taken on 2026-09-10, after E6 was read, and this paragraph is its record.

# Suppression transfers (H4)

**H4.** Projecting the axis out removes *red* on every op, at the selectivity ex-2.2.1 reached. On the recipe and on the adopted point, under `projection`, seed-mean accuracy on the red lines of each op falls to at most 0.2, and the seed-mean non-red deficit on the `mix` lines is at most 0.05. Partial: exactly one clause holds, or removal holds on all but one op, or the deficit sits in the 0.05–0.1 band.

The deficit gate is the band ex-2.2.1 landed in (0.024, partial at its 0.02 gate), and we report the 0.02 read beside it. That cost came from the syntax embeddings, and E2 reads the embeddings of the six op words.

Contrary on removal: red accuracy near the clean value on one of the new ops. That would say the model reads *red* off the axis on lines of that op, and the per-op alignment in E1 should show it first. Contrary on selectivity: a deficit above 0.1. That would send the intervention-tuning pass to the `operands` and `shaped` rows before anything is anchored.

| condition, operator | mix red acc ↓ | add red acc ↓ | screen red acc ↓ | multiply red acc ↓ | lighten red acc ↓ | darken red acc ↓ | ex-2.2.1 (mix) |
|---|---|---|---|---|---|---|---|
| recipe, projection | **0.06** ±0.05 | 0.32 ±0.20 | 0.23 ±0.10 | **0.17** ±0.12 | 0.23 ±0.09 | 0.24 ±0.23 | 0.09 ±0.10 |
| recipe, operands | 0.09 ±0.06 | 0.35 ±0.21 | 0.24 ±0.11 | 0.19 ±0.12 | 0.25 ±0.09 | 0.32 ±0.26 | 0.13 ±0.13 |
| recipe, shaped | 0.42 ±0.28 | 0.66 ±0.28 | 0.58 ±0.25 | 0.51 ±0.29 | 0.52 ±0.21 | 0.59 ±0.36 | 0.60 ±0.29 |
| recipe, ablate | 0.08 ±0.06 | 0.35 ±0.19 | 0.26 ±0.10 | 0.20 ±0.13 | 0.27 ±0.10 | 0.33 ±0.27 | 0.12 ±0.12 |
| recipe, clean | 1.00 ±0.01 | 1.00 ±0.00 | 1.00 ±0.00 | 1.00 ±0.00 | 1.00 ±0.00 | 1.00 ±0.00 | 1.00 ±0.00 |

*Removal on the red lines of each op, for `recipe`. Each value is seed-mean exact-match accuracy on the red lines (dose ≥ 0.8) under the operator, with half the seed range; the last column is the same operator on ex-2.2.1's one-op grammar (nine seeds of the D2.1 recipe). Bold marks a `projection` value inside the removal gate, at or below 0.2. The ride-along rows carry no gate. The last row is the clean pass.*

| condition, operator | mix red acc ↓ | add red acc ↓ | screen red acc ↓ | multiply red acc ↓ | lighten red acc ↓ | darken red acc ↓ | ex-2.2.1 (mix) |
|---|---|---|---|---|---|---|---|
| t00, projection | **0.00** ±0.00 | **0.05** ±0.03 | **0.02** ±0.01 | **0.01** ±0.01 | **0.06** ±0.04 | **0.03** ±0.02 | 0.09 ±0.10 |
| t00, operands | 0.08 ±0.09 | 0.23 ±0.12 | 0.22 ±0.13 | 0.20 ±0.14 | 0.32 ±0.17 | 0.15 ±0.10 | 0.13 ±0.13 |
| t00, shaped | 0.00 ±0.00 | 0.07 ±0.05 | 0.04 ±0.03 | 0.02 ±0.03 | 0.09 ±0.05 | 0.04 ±0.04 | 0.60 ±0.29 |
| t00, ablate | 0.00 ±0.00 | 0.03 ±0.03 | 0.02 ±0.02 | 0.01 ±0.01 | 0.05 ±0.04 | 0.02 ±0.02 | 0.12 ±0.12 |
| t00, clean | 0.89 ±0.05 | 0.92 ±0.04 | 0.96 ±0.03 | 0.95 ±0.02 | 0.95 ±0.02 | 0.97 ±0.01 | 1.00 ±0.00 |

*Removal on the red lines of each op, for `t00`. Each value is seed-mean exact-match accuracy on the red lines (dose ≥ 0.8) under the operator, with half the seed range; the last column is the same operator on ex-2.2.1's one-op grammar (nine seeds of the D2.1 recipe). Bold marks a `projection` value inside the removal gate, at or below 0.2. The ride-along rows carry no gate. The last row is the clean pass.*

| condition, operator | mix non-red deficit ↓ | add non-red deficit ↓ | screen non-red deficit ↓ | multiply non-red deficit ↓ | lighten non-red deficit ↓ | darken non-red deficit ↓ | ex-2.2.1 (mix) |
|---|---|---|---|---|---|---|---|
| recipe, projection | **0.024** ±0.030 | 0.001 ±0.002 | 0.003 ±0.005 | 0.006 ±0.008 | 0.000 ±0.002 | 0.001 ±0.002 | 0.024 ±0.063 |
| recipe, operands | -0.000 ±0.001 | -0.000 ±0.000 | -0.000 ±0.001 | 0.000 ±0.002 | -0.000 ±0.001 | 0.000 ±0.000 | 0.000 ±0.001 |
| recipe, shaped | 0.000 ±0.000 | 0.000 ±0.000 | 0.000 ±0.000 | 0.000 ±0.000 | 0.000 ±0.000 | 0.000 ±0.000 | 0.000 ±0.000 |
| recipe, ablate | 0.008 ±0.009 | 0.001 ±0.001 | 0.001 ±0.002 | 0.002 ±0.004 | 0.001 ±0.002 | 0.001 ±0.001 | 0.012 ±0.021 |

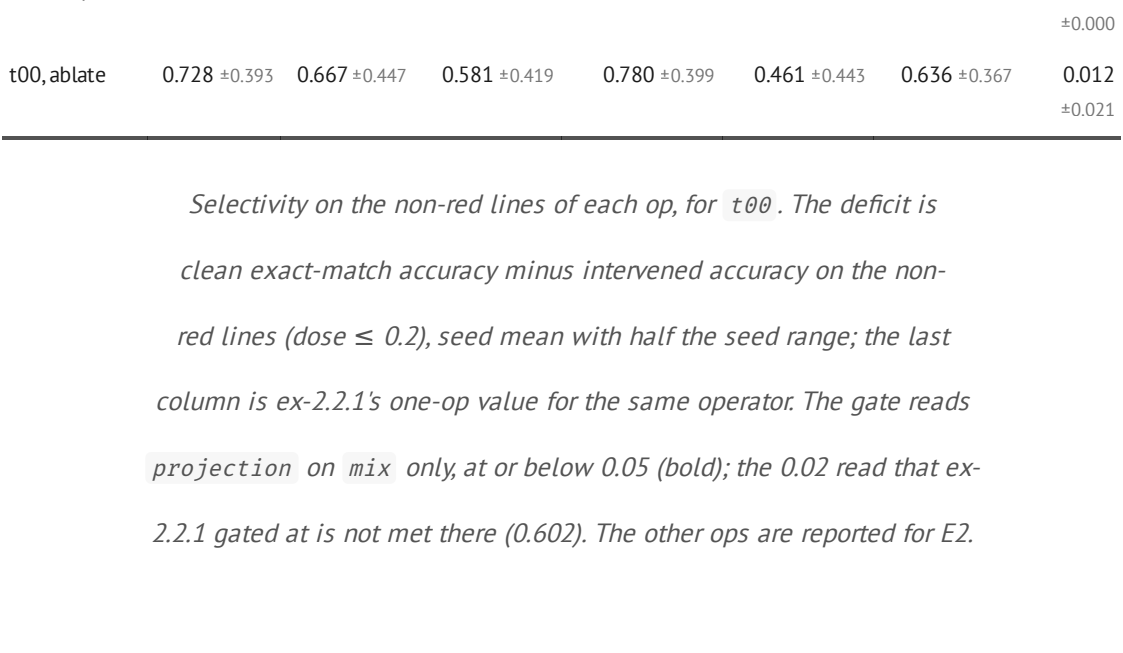
*Selectivity on the non-red lines of each op, for `recipe`. The deficit is clean exact-match accuracy minus intervened accuracy on the non-red lines (dose ≤ 0.2), seed mean with half the seed range; the last column is ex-2.2.1's one-op value for the same operator. The gate reads `projection` on `mix` only, at or below 0.05 (bold); the 0.02 read that ex-2.2.1 gated at is not met there (0.024). The other ops are reported for E2.*

| condition, operator | mix non-red deficit ↓ | add non-red deficit ↓ | screen non-red deficit ↓ | multiply non-red deficit ↓ | lighten non-red deficit ↓ | darken non-red deficit ↓ | ex-2.2.1 (mix) |
|---|---|---|---|---|---|---|---|
| t00, projection | 0.602 ±0.367 | 0.421 ±0.246 | 0.401 ±0.299 | 0.532 ±0.397 | 0.388 ±0.367 | 0.442 ±0.357 | 0.024 ±0.063 |
| t00, operands | 0.017 ±0.017 | 0.007 ±0.012 | 0.003 ±0.005 | 0.003 ±0.003 | 0.001 ±0.003 | -0.001 ±0.003 | 0.000 ±0.001 |
| t00, shaped | 0.732 ±0.255 | 0.518 ±0.199 | 0.523 ±0.222 | 0.638 ±0.297 | 0.380 ±0.321 | 0.481 ±0.396 | 0.000 ±0.000 |
| t00, ablate | 0.728 ±0.393 | 0.667 ±0.447 | 0.581 ±0.419 | 0.780 ±0.399 | 0.461 ±0.443 | 0.636 ±0.367 | 0.012 ±0.021 |

*Selectivity on the non-red lines of each op, for `t00`. The deficit is clean exact-match accuracy minus intervened accuracy on the non-red lines (dose ≤ 0.2), seed mean with half the seed range; the last column is ex-2.2.1's one-op value for the same operator. The gate reads `projection` on `mix` only, at or below 0.05 (bold); the 0.02 read that ex-2.2.1 gated at is not met there (0.602). The other ops are reported for E2.*

| condition, distance, pass | mix ↓ | add ↓ | screen ↓ | multiply ↓ | lighten ↓ | darken ↓ |
|---|---|---|---|---|---|---|
| recipe, decoded answer, projection | 0.31 ±0.06 | 0.31 ±0.10 | 0.32 ±0.08 | 0.36 ±0.11 | 0.37 ±0.10 | 0.43 ±0.18 |
| recipe, decoded answer, clean | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 |
| recipe, expected, projection | 0.32 ±0.06 | 0.31 ±0.10 | 0.32 ±0.07 | 0.36 ±0.11 | 0.37 ±0.09 | 0.43 ±0.17 |
| recipe, expected, clean | 0.01 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 | 0.00 ±0.00 |
| t00, decoded answer, projection | 0.48 ±0.08 | 0.55 ±0.05 | 0.50 ±0.06 | 0.60 ±0.09 | 0.49 ±0.08 | 0.65 ±0.06 |
| t00, decoded answer, clean | 0.02 ±0.01 | 0.01 ±0.01 | 0.01 ±0.00 | 0.01 ±0.01 | 0.01 ±0.00 | 0.00 ±0.00 |
| t00, expected, projection | 0.48 ±0.08 | 0.56 ±0.04 | 0.51 ±0.06 | 0.61 ±0.10 | 0.50 ±0.07 | 0.66 ±0.07 |
| t00, expected, clean | 0.06 ±0.01 | 0.07 ±0.02 | 0.05 ±0.01 | 0.04 ±0.02 | 0.05 ±0.02 | 0.04 ±0.01 |

*Answer distance on the red lines under `projection`, in the unit cube. The decoded row is the distance from the argmax answer to the true one; the expected row takes the expectation of that distance under the whole answer distribution over the color vocabulary. Each value is the seed mean with half the seed range; the clean pass sits under each projected row. On ex-2.2.1's one-op grammar the decoded distance under the projection was 0.29 ±0.06; a one-level step on this grid is 0.20.*




***The bound and the write, per site and per op, for the recipe under `projection`, signed.*** *For each op's probe lines, at each (slice, position), seed mean. The shaded band runs from the 1st to the 99th percentile of the signed clean alignment over the non-red lines; the scorer's bound is the arcsine of the 99th-percentile |α|, which is at least the further edge and equal to it where one tail holds the extreme, so this is close to the map the earlier version drew folded onto one side. The dashed pair is the 99th-percentile |α| arriving at the operator, whose arcsine is the write, mirrored about zero because the scorer keeps only its magnitude. The embedding row matches the clean band by construction. One row per slice with the embedding at the bottom, one column per op, and every panel on the same scale, so a write that pokes above its bound reads as a dashed line clearing the further band edge. Sites where the seed-mean write exceeds the seed-mean bound, per op: `mix` 3, `add` 3, `screen` 4, `multiply` 3, `lighten` 5, `darken` 4 of 24 post-embedding sites.*

**H4, contrary.** `recipe`: red accuracy under projection 0.06–0.32 across the ops, `mix` non-red deficit 0.024; `t00`: red accuracy under projection 0.00–0.06 across the ops, `mix` non-red deficit 0.602. Removal misses on `add, screen, lighten, darken` for `recipe`. The `mix` deficit on `t00` is above 0.1.

# Exploratory analyses

Preregistered as exploratory, no gates.

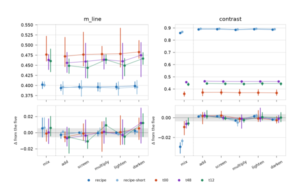
## E1 — per-op statistics

Every H2 statistic read on the probe lines of each op, rather than on the `mix` lines alone, for every candidate. The labeller never sees the op, so a placement that differs by op would mean the blocks carry *red* differently under different rules. Only m_line and contrast can differ: they involve op2 and the answer, which follow the op word, while grading, containment, and the lead weight are read at or before op1, where causal attention has not yet seen which rule the line uses.

The five statistics, in a phrase each: **m_line** is how much more the labelled lines lean on the axis than the average line does (the margin the survey ranked on); **grading r²** is whether that lean rises smoothly with redness rather than switching on; **contrast** is how much the pull prefers the operand that drew the label over the other one, at depth; **containment ᾱ** is the mean lean of every color at op1, which should stay near zero; and the **lead weight** is the share of the pull that sits on the drawing operand at the embedding.

| candidate | statistic | mix | add | screen | multiply | lighten | darken | spread |
|---|---|---|---|---|---|---|---|---|
| `recipe` | m_line ↑ | 0.402 ±0.009 | 0.394 ±0.008 | 0.397 ±0.007 | 0.396 ±0.011 | 0.397 ±0.009 | 0.399 ±0.009 | 0.008 |
| | grading r² ↑ | 0.819 ±0.036 | 0.819 ±0.036 | 0.819 ±0.036 | 0.819 ±0.036 | 0.819 ±0.036 | 0.819 ±0.036 | 0.000 |
| | contrast ↑ | 0.859 ±0.005 | 0.890 ±0.001 | 0.890 ±0.001 | 0.885 ±0.002 | 0.890 ±0.001 | 0.886 ±0.001 | 0.031 |
| | containment ᾱ ↓ | 0.062 ±0.020 | 0.062 ±0.020 | 0.062 ±0.020 | 0.062 ±0.020 | 0.062 ±0.020 | 0.062 ±0.020 | 0.000 |
| | lead weight ↑ | 0.793 ±0.028 | 0.807 ±0.027 | 0.807 ±0.027 | 0.807 ±0.027 | 0.807 ±0.027 | 0.807 ±0.027 | 0.014 |
| `recipe-short` | m_line ↑ | 0.401 ±0.007 | 0.396 ±0.017 | 0.395 ±0.020 | 0.395 ±0.012 | 0.393 ±0.018 | 0.396 ±0.017 | 0.008 |
| | grading r² ↑ | 0.875 ±0.027 | 0.875 ±0.027 | 0.875 ±0.027 | 0.875 ±0.027 | 0.875 ±0.027 | 0.875 ±0.027 | 0.000 |
| | contrast ↑ | 0.869 ±0.009 | 0.895 ±0.006 | 0.894 ±0.008 | 0.888 ±0.004 | 0.895 ±0.007 | 0.891 ±0.006 | 0.026 |
| | containment ᾱ ↓ | 0.063 ±0.023 | 0.063 ±0.023 | 0.063 ±0.023 | 0.063 ±0.023 | 0.063 ±0.023 | 0.063 ±0.023 | 0.000 |
| | lead weight ↑ | 0.828 ±0.040 | 0.840 ±0.037 | 0.841 ±0.038 | 0.841 ±0.038 | 0.841 ±0.038 | 0.841 ±0.038 | 0.013 |
| `t00` | m_line ↑ | 0.477 ±0.032 | 0.472 ±0.037 | 0.477 ±0.043 | 0.478 ±0.019 | 0.478 ±0.049 | 0.483 ±0.023 | 0.011 |
| | grading r² ↑ | 0.834 ±0.041 | 0.834 ±0.041 | 0.834 ±0.041 | 0.834 ±0.041 | 0.834 ±0.041 | 0.834 ±0.041 | 0.000 |
| | contrast ↑ | 0.362 ±0.021 | 0.372 ±0.030 | 0.373 ±0.024 | 0.370 ±0.027 | 0.370 ±0.026 | 0.368 ±0.025 | 0.012 |
| | containment ᾱ ↓ | 0.089 ±0.024 | 0.089 ±0.024 | 0.089 ±0.024 | 0.089 ±0.024 | 0.089 ±0.024 | 0.089 ±0.024 | 0.000 |
| | lead weight ↑ | 0.328 ±0.026 | 0.333 ±0.028 | 0.331 ±0.027 | 0.334 ±0.030 | 0.332 ±0.026 | 0.332 ±0.027 | 0.006 |
| `t48` | m_line ↑ | 0.463 ±0.020 | 0.456 ±0.026 | 0.459 ±0.043 | 0.464 ±0.018 | 0.460 ±0.045 | 0.475 ±0.043 | 0.019 |
| | grading r² ↑ | 0.866 ±0.025 | 0.866 ±0.025 | 0.866 ±0.025 | 0.866 ±0.025 | 0.866 ±0.025 | 0.866 ±0.025 | 0.000 |
| | contrast ↑ | 0.456 ±0.007 | 0.465 ±0.004 | 0.463 ±0.001 | 0.462 ±0.008 | 0.462 ±0.009 | 0.462 ±0.007 | 0.009 |
| | containment ᾱ ↓ | 0.077 ±0.025 | 0.077 ±0.025 | 0.077 ±0.025 | 0.077 ±0.025 | 0.077 ±0.025 | 0.077 ±0.025 | 0.000 |
| | lead weight ↑ | 0.455 ±0.049 | 0.461 ±0.048 | 0.461 ±0.048 | 0.461 ±0.048 | 0.461 ±0.048 | 0.462 ±0.048 | 0.007 |
| `t12` | m_line ↑ | 0.461 ±0.029 | 0.449 ±0.019 | 0.444 ±0.022 | 0.464 ±0.012 | 0.450 ±0.022 | 0.467 ±0.015 | 0.023 |
| | grading r² ↑ | 0.842 ±0.025 | 0.842 ±0.025 | 0.842 ±0.025 | 0.842 ±0.025 | 0.842 ±0.025 | 0.842 ±0.025 | 0.000 |
| | contrast ↑ | 0.439 ±0.014 | 0.446 ±0.007 | 0.444 ±0.011 | 0.442 ±0.011 | 0.445 ±0.012 | 0.445 ±0.012 | 0.007 |
| | containment ᾱ ↓ | 0.039 ±0.026 | 0.039 ±0.026 | 0.039 ±0.026 | 0.039 ±0.026 | 0.039 ±0.026 | 0.039 ±0.026 | 0.000 |
| | lead weight ↑ | 0.481 ±0.036 | 0.487 ±0.037 | 0.487 ±0.037 | 0.488 ±0.037 | 0.487 ±0.037 | 0.488 ±0.037 | 0.007 |

*The placement statistics per op, for every candidate: seed mean with half the seed range. The `mix` column repeats the gated values. Spread is the range of the seed means across the six ops, to read against the band for that statistic in the H3 table.*




***The two per-op placement statistics, whole and magnified.** m_line and contrast are the only statistics that can vary by op. One line per candidate in its ink, one dot per op at the seed mean with a bar for the seed range, offset a little so the candidates do not overlap. The upper row plots each statistic on its own scale, where the gaps between candidates set the axis. The lower row magnifies what that hides: the same points as a distance from the candidate's own mean over the five ops that share a partner draw, per seed, so the ops can be compared within a run. The grey stripe there is that statistic's equivalence band, the smallest seed-mean difference the resolution rule may call a difference. `mix` sits left of the divider, and out of the baseline, because its probe lines are D2.1's on-grid partners rather than the shared draw: its distance from the five carries the partner set as well as the op.*

Read across all six ops, the widest spread of seed-mean m_line on any candidate is 0.023 (band 0.011) and of contrast 0.031 (band 0.007), both wider than the band. But the six ops are not six comparable reads. `mix` draws its probe lines from its own on-grid partners, D2.1's set, while the other five share one draw, so only those five differ from each other in the op alone. Split that way the two statistics separate.

**Contrast is op-blind among the five.** Its spread there is no wider than the band on any candidate (0.007 at most, band 0.007). All of the six-op spread is `mix`, which sits below the five on every candidate, by 0.006 to 0.029 — furthest on the two recipe arms, whose contrast is highest to begin with. Since `mix` is also the op with the different partner set, that step reads as the partner set rather than as the rule, and a `mix` probe drawn like the others would be needed to tell the two apart.

**m_line is less settled.** Among the five it stays inside the band on the recipe arms (0.005 and 0.003) and sits at it on `t00` (0.011), reaching 0.019 on `t48` and 0.023 on `t12`, the two proposals with the lowest anchor weight, against a band of 0.011. The ordering repeats across candidates — `darken` high on all five, `add` low on all but `recipe-short` — but the per-seed deviations are wider than the band, so five seeds do not settle whether the ordering is real. If it is, the probe set is the first place to look: m_line is a margin over the average line of that op's set, and the five sets share their op1 colors and partners but not their answers, so an op whose answers sit differently against the red axis would move it with the placement unchanged. Either way the effect is small, and every gated read is on `mix`.

Grading r² and containment ᾱ are identical on all six ops to every printed digit. Both are read at op1, which precedes the op word, so under causal attention they cannot see which rule the line uses; the table prints them for completeness. The lead weight is read at op1 too, but its softmin normalizer runs over the whole span, so the op word's own embedding enters it: it moves by 0.014 at most, again with `mix` apart from the rest.

## E2 — the op words

The alignment of the six op-word embeddings and of `=` and the newline, per slice, on the recipe and the control. Ex-2.2.1 found a constant component on `+` and `=` of about 0.3–0.4 in the published map, which is where its non-red cost came from. The second question is whether the alignment of an op word predicts the non-red deficit on its lines under `projection`.

| token | slice emb | slice 1 | slice 2 | slice 3 | slice 4 |
|---|---|---|---|---|---|
| `mix` | 0.17 ±0.01 | 0.08 ±0.03 | 0.07 ±0.04 | 0.02 ±0.03 | -0.00 ±0.01 |
| `add` | 0.18 ±0.02 | 0.10 ±0.05 | 0.08 ±0.01 | 0.04 ±0.02 | 0.00 ±0.01 |
| `screen` | 0.17 ±0.01 | 0.07 ±0.03 | 0.06 ±0.03 | 0.02 ±0.02 | -0.00 ±0.01 |
| `multiply` | 0.18 ±0.01 | 0.09 ±0.03 | 0.07 ±0.02 | 0.02 ±0.01 | -0.00 ±0.01 |
| `lighten` | 0.17 ±0.01 | 0.07 ±0.02 | 0.05 ±0.02 | 0.02 ±0.01 | -0.01 ±0.01 |
| `darken` | 0.16 ±0.02 | 0.08 ±0.05 | 0.06 ±0.04 | 0.02 ±0.02 | -0.00 ±0.01 |
| `=` | 0.40 ±0.06 | 0.24 ±0.05 | 0.15 ±0.04 | 0.09 ±0.05 | 0.05 ±0.04 |
| `↵` | 0.06 ±0.02 | 0.02 ±0.03 | 0.01 ±0.02 | 0.01 ±0.01 | -0.00 ±0.00 |

*Alignment (cos to $e_1$) of the syntax tokens on `recipe`, per slice: each op word read on its own op's probe lines, and `=` and the newline averaged over the six ops. Seed mean with half the seed range. The embedding column is the token's embedding itself; deeper columns are the residual stream at that position.*

| token | slice emb | slice 1 | slice 2 | slice 3 | slice 4 |
|---|---|---|---|---|---|
| `mix` | 0.34 ±0.06 | 0.27 ±0.04 | 0.25 ±0.01 | 0.15 ±0.03 | 0.04 ±0.02 |
| `add` | 0.33 ±0.06 | 0.25 ±0.04 | 0.24 ±0.01 | 0.15 ±0.03 | 0.05 ±0.02 |
| `screen` | 0.36 ±0.07 | 0.29 ±0.06 | 0.27 ±0.02 | 0.15 ±0.02 | 0.04 ±0.02 |
| `multiply` | 0.32 ±0.02 | 0.23 ±0.08 | 0.22 ±0.04 | 0.14 ±0.01 | 0.04 ±0.02 |
| `lighten` | 0.34 ±0.06 | 0.28 ±0.08 | 0.27 ±0.02 | 0.16 ±0.03 | 0.04 ±0.01 |
| `darken` | 0.34 ±0.06 | 0.26 ±0.04 | 0.24 ±0.03 | 0.15 ±0.06 | 0.04 ±0.02 |
| `=` | 0.93 ±0.03 | 0.73 ±0.06 | 0.46 ±0.09 | 0.34 ±0.07 | 0.19 ±0.06 |
| `↵` | 0.06 ±0.03 | -0.00 ±0.03 | 0.01 ±0.01 | 0.01 ±0.02 | 0.01 ±0.01 |

*Alignment (cos to $e_1$) of the syntax tokens on `t00`, per slice: each op word read on its own op's probe lines, and `=` and the newline averaged over the six ops. Seed mean with half the seed range. The embedding column is the token's embedding itself; deeper columns are the residual stream at that position.*

| token | slice emb | slice 1 | slice 2 | slice 3 | slice 4 |
|---|---|---|---|---|---|
| `mix` | 0.07 ±0.13 | 0.08 ±0.14 | 0.05 ±0.08 | 0.02 ±0.02 | -0.00 ±0.01 |
| `add` | 0.07 ±0.18 | 0.08 ±0.20 | 0.05 ±0.11 | 0.03 ±0.04 | 0.01 ±0.01 |
| `screen` | 0.07 ±0.14 | 0.07 ±0.16 | 0.04 ±0.09 | 0.01 ±0.03 | -0.00 ±0.01 |
| `multiply` | 0.07 ±0.11 | 0.09 ±0.12 | 0.06 ±0.07 | 0.03 ±0.03 | 0.01 ±0.01 |
| `lighten` | 0.08 ±0.11 | 0.08 ±0.13 | 0.05 ±0.07 | 0.02 ±0.03 | 0.00 ±0.01 |
| `darken` | 0.04 ±0.17 | 0.06 ±0.15 | 0.05 ±0.08 | 0.02 ±0.03 | 0.00 ±0.01 |
| `=` | 0.06 ±0.19 | 0.07 ±0.22 | 0.05 ±0.16 | 0.05 ±0.12 | 0.02 ±0.06 |
| `↵` | -0.03 ±0.03 | -0.02 ±0.04 | -0.02 ±0.03 | -0.01 ±0.02 | -0.00 ±0.01 |

*Alignment (cos to $e_1$) of the syntax tokens on `control`, per slice: each op word read on its own op's probe lines, and `=` and the newline averaged over the six ops. Seed mean with half the seed range. The embedding column is the token's embedding itself; deeper columns are the residual stream at that position.*

| token | slice emb | slice 1 | slice 2 | slice 3 | slice 4 |
|---|---|---|---|---|---|
| `mix` | 0.26 ±0.03 | 0.19 ±0.04 | 0.10 ±0.03 | 0.07 ±0.05 | 0.01 ±0.01 |
| `add` | 0.28 ±0.05 | 0.20 ±0.06 | 0.12 ±0.04 | 0.09 ±0.06 | 0.02 ±0.02 |
| `screen` | 0.27 ±0.05 | 0.20 ±0.06 | 0.12 ±0.05 | 0.09 ±0.07 | 0.01 ±0.02 |
| `multiply` | 0.27 ±0.05 | 0.20 ±0.05 | 0.10 ±0.04 | 0.08 ±0.05 | 0.02 ±0.02 |
| `lighten` | 0.27 ±0.05 | 0.20 ±0.06 | 0.11 ±0.05 | 0.09 ±0.08 | 0.01 ±0.02 |
| `darken` | 0.26 ±0.05 | 0.19 ±0.07 | 0.10 ±0.06 | 0.08 ±0.07 | 0.01 ±0.02 |
| `=` | 0.80 ±0.07 | 0.67 ±0.05 | 0.49 ±0.07 | 0.35 ±0.05 | 0.17 ±0.05 |
| `↵` | 0.06 ±0.02 | 0.00 ±0.01 | 0.00 ±0.02 | 0.01 ±0.01 | 0.00 ±0.01 |

*Alignment (cos to $e_1$) of the syntax tokens on `t48`, per slice: each op word read on its own op's probe lines, and `=` and the newline averaged over the six ops. Seed mean with half the seed range. The embedding column is the token's embedding itself; deeper columns are the residual stream at that position.*

| token | slice emb | slice 1 | slice 2 | slice 3 | slice 4 |
|---|---|---|---|---|---|
| `mix` | 0.26 ±0.03 | 0.16 ±0.04 | 0.14 ±0.04 | 0.07 ±0.06 | 0.01 ±0.02 |
| `add` | 0.27 ±0.03 | 0.18 ±0.05 | 0.16 ±0.05 | 0.10 ±0.05 | 0.02 ±0.02 |
| `screen` | 0.27 ±0.03 | 0.18 ±0.05 | 0.16 ±0.04 | 0.10 ±0.06 | 0.01 ±0.02 |
| `multiply` | 0.27 ±0.02 | 0.17 ±0.04 | 0.14 ±0.03 | 0.07 ±0.04 | 0.01 ±0.02 |
| `lighten` | 0.28 ±0.04 | 0.18 ±0.05 | 0.15 ±0.04 | 0.09 ±0.06 | 0.01 ±0.02 |
| `darken` | 0.26 ±0.02 | 0.17 ±0.03 | 0.14 ±0.02 | 0.07 ±0.05 | 0.01 ±0.02 |
| `=` | 0.74 ±0.05 | 0.65 ±0.02 | 0.49 ±0.03 | 0.37 ±0.05 | 0.20 ±0.02 |
| `↵` | 0.06 ±0.01 | -0.00 ±0.02 | 0.01 ±0.01 | 0.00 ±0.00 | -0.00 ±0.03 |

*Alignment (cos to $e_1$) of the syntax tokens on `t12`, per slice: each op word read on its own op's probe lines, and `=` and the newline averaged over the six ops. Seed mean with half the seed range. The embedding column is the token's embedding itself; deeper columns are the residual stream at that position.*

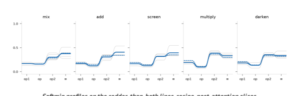
On the recipe, the mean |α| of each op word over the four post-attention slices, against the non-red deficit under `projection` on its lines: `mix` 0.05 / 0.024, `add` 0.05 / 0.001, `screen` 0.04 / 0.003, `multiply` 0.05 / 0.006, `lighten` 0.04 / 0.000, `darken` 0.04 / 0.001. Pearson r over the six ops is +0.19. Six points cannot carry much, and the sign is what to read: a positive r says the ops whose word sits further along the axis lose more on their non-red lines when the axis is projected out.

## E3 — redder than both

Every op but `lighten` has lines whose answer is redder than either operand (the op table counts them). On those lines the strongest evidence for *red* sits at the answer, and the labeller never keys on that position. This is the blind-span case named in the scope note of ex-2.1.10. We read the alignment at `=` and at the answer position against lines of the same op and dose whose answer is not redder than both, plus the softmin profile on each group, and then restrict the H4 statistics to these lines.

| op | redder lines | Δα at = | Δα at the answer | clean acc | projection acc ↓ | misses → red operand | misses → neighbor |
|---|---|---|---|---|---|---|---|
| `mix` | 372 | 0.005 ±0.012 | 0.004 ±0.006 | 1.00 ±0.00 | 0.98 ±0.02 | 0.00 ±0.00 | 1.00 ±0.00 |
| `add` | 273 | 0.057 ±0.008 | 0.067 ±0.020 | 1.00 ±0.00 | 0.91 ±0.01 | 0.07 ±0.10 | 0.20 ±0.12 |
| `screen` | 186 | 0.040 ±0.016 | 0.033 ±0.016 | 1.00 ±0.00 | 0.95 ±0.01 | 0.15 ±0.15 | 0.30 ±0.25 |
| `multiply` | 472 | 0.072 ±0.017 | 0.149 ±0.014 | 1.00 ±0.00 | 0.78 ±0.03 | 0.04 ±0.02 | 0.25 ±0.11 |
| `darken` | 533 | 0.039 ±0.013 | 0.083 ±0.013 | 1.00 ±0.00 | 0.90 ±0.01 | 0.03 ±0.03 | 0.16 ±0.06 |

*The redder-than-both lines of each op on the recipe. Δα is the alignment on those lines minus the alignment on lines of the same op in the same dose bin whose answer is not redder than both, averaged over the four post-attention slices and weighted by the redder lines' dose histogram; a positive value at the answer says the stream carries the answer's extra redness there, where the labeller never keyed. The right half restricts H4 to these lines: exact-match accuracy clean and under `projection`, and, of the projected answers that are wrong, the share that are the redder operand and the share that are a one-step neighbor of the truth. Seed means with half the seed range; `lighten` has no such lines.*



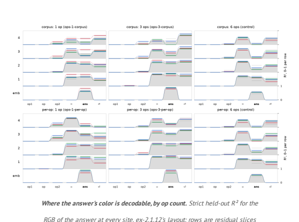

***Softmin profiles on the redder-than-both lines, recipe, post-attention slices.** One panel per op that has such lines. Solid is the seed-mean profile over the span roles on the redder lines; dashed is the dose-matched comparison group of the same op. The profile is the share of the pull each role would receive at the run's τ, read from the clean alignment, so a difference between the two says the answer's redness moved where the pull lands within the span. Hairlines are seeds.*

## E4 — a richer op set

The cube probes of ex-2.1.12 (ridge, $\ell_2 = 10^{-2}$, strict per-value holdout) on the un-anchored models at one, three, and six ops: held-out $R^2$ for op1, op2, and the RGB of the answer, per slice and position. This asks whether more rules give the model a better operand geometry at the same compute. The two matchings under the conditions are read as a pair, with the control as the six-op point of both. In the `corpus` arms the fewer-op models see more lines per op, so a cube that improves with the op count there is a clean positive. In the `per-op` arms the fewer-op models see the same lines per op, repeated more often, so a cube that worsens with the op count there is a clean negative. A trend that holds in both arms is read as the op count; one that holds in only one is read as its confound, lines per op or repetition.

| target, site | arm | 1 op | 3 ops | 6 ops |
|---|---|---|---|---|
| op1 at own slot | corpus | 0.64 ±0.04 | 0.87 ±0.00 | 0.52 ±0.14 |
| | per-op | 0.86 ±0.04 | 0.85 ±0.05 | 0.52 ±0.14 |
| op2 at own slot | corpus | 0.52 ±0.02 | 0.61 ±0.17 | 0.28 ±0.12 |
| | per-op | 0.77 ±0.05 | 0.68 ±0.08 | 0.28 ±0.12 |
| ans at = | corpus | 0.57 ±0.09 | 0.63 ±0.06 | 0.65 ±0.03 |
| | per-op | 0.74 ±0.08 | 0.72 ±0.04 | 0.65 ±0.03 |

*Strict held-out $R^2$ of the un-anchored models by op count, at the site that carries each target: op1 and op2 at their own slot, the answer at `=`, where it is first decodable. Each value is the mean over the RGB channels and the four post-attention slices, negative scores clipped to zero first, as a seed mean with half the seed range. The `corpus` arm keeps D2.1's line count as the op set narrows (more lines per op at fewer ops); the `per-op` arm keeps the six-op lines per op (a smaller corpus, repeated). The control is the six-op point of both.*




***Where the answer's color is decodable, by op count.** Strict held-out $R^2$ for the RGB of the answer at every site, ex-2.1.12's layout: rows are residual slices with the embedding at the bottom, columns the six positions, one trace per channel over the grey area of their mean, with the seed envelope as hairlines. The bold slot is the answer itself. Top row, the `corpus` matching; bottom row, the `per-op` matching; the six-op panel is the control in both. Negative scores clip to the floor.*

## E5 — the noise floor, re-measured

The per-run σ of every gated statistic on the new grammar, from the five seeds of the recipe and of the control, beside the ex-2.1.10 values the bands used. The H3 verdict and the selection rule are scored with the frozen ex-2.1.10 bands either way. Where the fresh σ is larger, we report which H3 comparisons the wider band would leave unresolved, as a robustness read.

| statistic | σ, ex-2.1.10 (frozen) | σ, recipe (fresh) | σ, control (fresh) | band, frozen | band, fresh recipe |
|---|---|---|---|---|---|
| m_line | 0.0088 | 0.0068 | 0.0274 | 0.011 | 0.009 |
| alpha_op1 | 0.0212 | 0.0142 | 0.0264 | 0.027 | 0.018 |
| holdout_em | 0.0087 | 0.0017 | 0.0017 | 0.011 | 0.002 |
| retention | 0.0061 | 0.0036 | 0.8558 | 0.008 | 0.005 |
| r2_sim | 0.0240 | **0.0267** | 0.0460 | 0.030 | 0.034 |
| contrast | 0.0057 | 0.0050 | 0.1324 | 0.007 | 0.006 |

*The per-run spread of each gated statistic: the frozen ex-2.1.10 value behind every band in this report, and the sample standard deviation over the five fresh seeds of the recipe and of the control. Bold marks a fresh recipe σ above the frozen one. The placement statistics on the control are near zero by construction, so its σ there is the instrument's noise rather than a run-to-run spread of a placement. The last two columns are the band for a five-against-five comparison under each σ.*

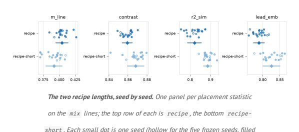
The H3 comparisons of each proposal's m_line against `recipe-short`, under both bands: `t00` +0.075: resolved at the frozen band, resolved at the fresh one; `t48` +0.062: resolved at the frozen band, resolved at the fresh one; `t12` +0.059: resolved at the frozen band, resolved at the fresh one.

## E6 — the two recipe lengths at twenty seeds

Added after the results were read. The H3 read compared `recipe` and `recipe-short` at five seeds each, and the post hoc amendment above adopts one of them; five seeds is a thin basis for a choice the next experiments build on, so fifteen more seeds of each length were run (the `-more` arms under the conditions). The frozen arms keep their five seeds and their verdicts; this section reads all twenty of each length together. The question is plain: at twenty seeds, does the full-length recipe still place *red* better than the short one, by more than the band, and does either length change its H4 read?

| statistic | recipe, 5 seeds | recipe, 20 seeds | recipe-short, 20 seeds | Δ (short − full) | band, frozen σ | band, fresh σ |
|---|---|---|---|---|---|---|
| m_line | 0.402 | 0.404 σ 0.009 | 0.392 σ 0.013 | **-0.012** | 0.006 | 0.007 |
| contrast | 0.859 | 0.859 σ 0.005 | 0.868 σ 0.009 | **+0.009** | 0.004 | 0.005 |
| r2_sim | 0.819 | 0.821 σ 0.034 | 0.886 σ 0.020 | **+0.064** | 0.015 | 0.018 |
| alpha_op1 | 0.062 | 0.064 σ 0.021 | 0.083 σ 0.025 | **+0.019** | 0.013 | 0.015 |
| lead_emb | 0.793 | 0.789 σ 0.016 | 0.826 σ 0.025 | **+0.037** | - | 0.013 |
| holdout_em | 0.999 | 0.999 σ 0.002 | 0.999 σ 0.002 | -0.000 | 0.006 | 0.001 |

*The two recipe lengths at 20 seeds each: the five frozen seeds and the fifteen addendum seeds pooled. Values are seed means; σ is the sample standard deviation over the 20 seeds. Δ is short minus full. The frozen band is 2σ√(2/20) with the ex-2.1.10 per-run σ; the fresh band uses the pooled σ of the two arms in this table. Bold marks a difference outside the fresh band. The lead weight has no frozen σ.*




***The two recipe lengths, seed by seed.** One panel per placement statistic on the `mix` lines; the top row of each is `recipe`, the bottom `recipe-short`. Each small dot is one seed (hollow for the five frozen seeds, filled for the fifteen addendum seeds), jittered a little so they do not stack; the large marker is the twenty-seed mean, with a bar of one σ either side.*

| condition | mix red acc ↓ | add red acc ↓ | screen red acc ↓ | multiply red acc ↓ | lighten red acc ↓ | darken red acc ↓ | mix non-red deficit ↓ |
|---|---|---|---|---|---|---|---|
| recipe, projection, 20 seeds | **0.05** ±0.08 | 0.31 ±0.24 | 0.22 ±0.17 | **0.15** ±0.16 | 0.23 ±0.11 | 0.22 ±0.23 | **0.026** ±0.041 |
| recipe-short, projection, 20 seeds | **0.01** ±0.04 | **0.14** ±0.18 | **0.12** ±0.14 | **0.08** ±0.11 | **0.14** ±0.10 | **0.18** ±0.18 | **0.026** ±0.035 |

*H4's two statistics under `projection` at twenty seeds per length: seed-mean exact-match accuracy on the red lines of each op (gate ≤ 0.2, bold) and the non-red deficit on `mix` (gate ≤ 0.05, bold), each with half the seed range.*

At twenty seeds, `recipe-short` sits -0.012 of m_line from `recipe`, against a band of 0.006 (frozen σ) or 0.007 (fresh σ): the full-length recipe keeps its lead. The lead weight at the embedding moves +0.037 between the lengths. Under `projection`, the worst-op red accuracy is 0.31 on the full recipe and 0.18 on the short one, and the `mix` non-red deficit is 0.026 and 0.026. So the choice between the two lengths does not turn on H4. It turns on placement, and the two lengths split it: the full recipe has the larger margin, and the short one grades better and puts more of the pull on the drawing operand at the embedding. The post hoc rule ranks on the margin, as the frozen one did, and names the full recipe; the decision recorded beside it takes the short arm, on grading, the operand lead, and cost.

## E7 — the other arms under projection

Added after the results were read. The H4 tables above read the recipe and `t00`; this reads the same statistics on the remaining scored arms, `recipe-short` and the two proposals the adoption did not name, so the removal picture (complete on `mix`, partial on the saturating ops) can be checked across every operating point rather than two. The E2 syntax-embedding tables for those proposals sit at the end of E2.

| condition | mix red acc ↓ | add red acc ↓ | screen red acc ↓ | multiply red acc ↓ | lighten red acc ↓ | darken red acc ↓ | mix non-red deficit ↓ | \|α\| of = (deep) |
|---|---|---|---|---|---|---|---|---|
| recipe-short, projection | **0.00** ±0.01 | **0.13** ±0.08 | **0.10** ±0.07 | **0.06** ±0.03 | **0.12** ±0.05 | **0.17** ±0.08 | **0.014** ±0.020 | 0.09 ±0.02 |
| t48, projection | **0.00** ±0.00 | **0.06** ±0.02 | **0.03** ±0.01 | **0.01** ±0.01 | **0.09** ±0.04 | **0.03** ±0.02 | 0.498 ±0.364 | 0.42 ±0.04 |
| t12, projection | **0.00** ±0.00 | **0.05** ±0.03 | **0.03** ±0.02 | **0.01** ±0.01 | **0.07** ±0.04 | **0.04** ±0.02 | 0.598 ±0.250 | 0.43 ±0.02 |

*H4's statistics under `projection` on the arms the H4 section does not table, five seeds each: red-line accuracy per op (gate ≤ 0.2, bold), the `mix` non-red deficit (gate ≤ 0.05, bold), and, as the E2 correlate, the mean |α| of `=` over the four post-attention slices. Seed mean with half the seed range.*

## E8 — which red lines survive projection

Added after the results were read. H4's removal statistic asks whether the answer is still right once the axis is projected out; on ops that saturate, many red lines have an answer that does not depend on how red the red operand is, so a model that has lost *red* can still answer them. This sorts each op's red lines by whether the answer *depends* on the red operand's redness (lowering that operand's R by one grid step changes the snapped answer) and reads accuracy on the two groups separately, clean and under `projection`. Removal that reads as partial in H4 should read as complete on the dependent lines if the axis carries the redness the answer needs.

| op | condition | dependent red lines | other red lines | clean, dependent | clean, other | projection, dependent ↓ | projection, other |
|---|---|---|---|---|---|---|---|
| `mix` | recipe | 222 | 143 | 0.99 ±0.01 | 1.00 ±0.00 | 0.04 ±0.04 | 0.08 ±0.08 |
| `mix` | t00 | 222 | 143 | 0.87 ±0.07 | 0.93 ±0.06 | 0.00 ±0.00 | 0.00 ±0.00 |
| `add` | recipe | 80 | 325 | 1.00 ±0.01 | 1.00 ±0.00 | 0.18 ±0.16 | 0.35 ±0.20 |
| `add` | t00 | 80 | 325 | 0.92 ±0.05 | 0.92 ±0.04 | 0.02 ±0.03 | 0.06 ±0.04 |
| `screen` | recipe | 215 | 190 | 0.99 ±0.01 | 1.00 ±0.00 | 0.10 ±0.08 | 0.39 ±0.13 |
| `screen` | t00 | 215 | 190 | 0.98 ±0.02 | 0.93 ±0.04 | 0.01 ±0.01 | 0.03 ±0.02 |
| `multiply` | recipe | 190 | 215 | 1.00 ±0.00 | 1.00 ±0.00 | 0.05 ±0.04 | 0.28 ±0.20 |
| `multiply` | t00 | 190 | 215 | 0.93 ±0.04 | 0.96 ±0.03 | 0.00 ±0.00 | 0.03 ±0.03 |
| `lighten` | recipe | 337 | 68 | 1.00 ±0.00 | 1.00 ±0.00 | 0.12 ±0.11 | 0.76 ±0.04 |
| `lighten` | t00 | 337 | 68 | 0.96 ±0.02 | 0.90 ±0.06 | 0.02 ±0.02 | 0.29 ±0.18 |
| `darken` | recipe | 68 | 337 | 1.00 ±0.00 | 1.00 ±0.00 | 0.00 ±0.00 | 0.29 ±0.28 |
| `darken` | t00 | 68 | 337 | 0.95 ±0.04 | 0.98 ±0.02 | 0.00 ±0.00 | 0.03 ±0.02 |

*The red lines (dose ≥ 0.8) of each op split by whether the answer depends on the red operand's redness: a line is dependent if lowering that operand's R by one grid level changes the snapped answer. Exact-match accuracy on each group, clean and under `projection`, seed mean with half the seed range, on the recipe and on the point the frozen rule adopted. On `mix` the other lines are those where a one-level drop rounds back to the same answer.*

# Discussion

Here is what we think happened, in plain terms.

**The anchoring recipe does not care about the grammar.** We trained on six rules instead of one, with the labeller still only ever looking at the two operands, and *red* landed on the axis the same way it did in D2.1: no task cost on any op (H1), every placement gate met (H2), and the same numbers on the lines of every op (E1). The recipe needs its own stream of labelled operands each step, the corpus kept that, and the rest of the line can do whatever it likes. So the "recipe is grammar-specific" row of the D2.2 risk table closes.

**The survey's proposals are real, and smaller than advertised.** All three came in about 0.05 of m_line below their survey values, the recipe about 0.02 below its own. That is the winner's curse we priced in: when you pick the best of many noisy trials you tend to pick one that got lucky, so a fresh run scores lower.[1] About two thirds of the advantage over `recipe-short` survived, the proposals grade better than the survey said (the other side of the same luck), and they spread more across seeds than the recipe does. E5 says the H3 differences hold up under the wider band too.

**The selection rule picked `t00`, and we are not carrying it.** The rule ranked on margin and checked most of the H2 gates, and `t00` won. But it misses the one gate the rule left out, the lead weight at the embedding (0.33 against 0.4), and its contrast is 0.36 where the recipe's is 0.86. More to the point, it costs something the rule never looked at. At five times the recipe's anchor weight, the axis ends up on the `=` and op-word embeddings of the shared embedding table (E2: `=` at 0.93 against 0.40), and projecting the axis out of every position then breaks 0.6 of the non-red `mix` lines, with some seeds losing nearly all of them. The same is true of `t48` and `t12` (E7). Editing only the operand positions gets around it (0.017 on `mix`), and the prereg's contrary clause already sends the intervention-tuning pass to those tokens. But an operating point that needs a routed edit before its first intervention is a weak base for the anchored-op experiments, so the post hoc read adds the two missing gates and adopts the recipe, and D2.2 builds on its short arm. Whether the syntax embeddings can be made clean rather than routed around is the tied-readout question, and it wants settling before the D2.3 swap, which rotates every position.

**Full length or half?** We ran fifteen more seeds of each recipe length to answer this properly (E6). At twenty seeds the full recipe keeps a 0.012 margin lead, about twice the band, and the short one grades better ($r^2$ 0.89 against 0.82) and leads better at the embedding (0.83 against 0.79). Neither is wrong. The post hoc rule ranks on margin, so it names the full recipe; we chose the short arm anyway, and the choice is recorded in the post hoc section. The anchored-op experiments lean on grading and on the operand lead more than on margin, the H4 read is the same, and the short arm is half the compute.

**Removal looked partial on four ops, and mostly is not.** H4's removal statistic is whether the model still gets the answer right once the axis is projected out. On the recipe that held on `mix` (0.06 red-line accuracy) and missed on four of the new ops (0.17 to 0.32). The reason is the ops, not the anchor. On `add`, `screen`, `multiply`, `lighten`, and `darken`, most red lines have an answer that would be the same if the red operand were a little less red, because the op saturates or picks the other operand's channel; a model that has lost *red* can still answer those. E8 splits each op's red lines by whether the answer depends on the red operand's redness at all, and on the dependent lines the projection removes *red* on every op (accuracy 0.00 to 0.18 on the recipe, 0.02 or less on `t00`). The misses in the H4 table are the lines where the correct answer never needed *red*. For the next grammar we want both kinds: some ops whose answers depend on both operands and spread through the cube, and some anisotropic ones like these, whose answers sit on a face or an edge, so the E8 split has both to read; that is the diverse-op item, and it should be checked with the answer clouds and the E8 split before anything is anchored on it.

**Selectivity on the recipe transfers.** The `mix` non-red deficit under projection is the 0.024 from ex-2.2.1 at five seeds and 0.026 at twenty, and the other ops lose nothing to speak of. The write stays inside the bound at most sites (the stacked figure under H4), and where it pokes above, it is at the op word, which is the one new token.

**The blind span behaves.** On the lines whose answer is redder than either operand, the answer's own redness shows on the axis at `=` and at the answer position, by 0.15 at most, and the projection leaves most of those lines intact (E3). The anchor never sees the answer, so that redness is computed off the axis, as the ex-2.1.10 scope note said it would be. `multiply` has the most such lines and the largest shift, so it is where a labeller that also reads the answer would show up first; that labeller is filed under the span variants.

**More rules did not make a straighter cube, on this probe.** At six ops the operand's RGB is less linearly decodable at its own slot than at three (0.52 against 0.85 to 0.87), in both matchings, though the six-op seeds spread widely (±0.14) and the one-op arm of the `per-op` matching is in a memorization regime, so the read is a trend over three points rather than a result. The answer at `=` gives no clean read either way. Whether the model builds op-specific readings of its operands is a probe question for the anchored-op experiments, which read those sites anyway (E4).

**Which op to anchor next.** The relevance table favors `mix` (94% of its lines have an answer no other op gives) and, among the new ops, `multiply`. E3 adds that `multiply` is where the answer's redness shows on the axis most, and E8 that its red lines are about half dependent. The suppression prereg can choose on what it wants to show: `mix` for the sharpest test, `multiply` for a rule the anchor has never seen.

# Method

Two things change for the model against D2.1: the corpus carries six rules at the same total size, and a token between the operands decides the answer.[2] The dependency named in the design lands before the DAG, in `sca.data`, and the table below is its specification.

## The grammar

Six ops on the `v216` grid (six levels per channel: 0, 3, 6, 9, 12, 15), with every color one token and every op one token. Each op is computed per channel on the 0..15 scale and snapped to the nearest grid level. So every op answers every pair with a color in the vocabulary, and a line can be written for any pair. Where the rule lands on the grid by itself, no rounding happens, and `mix` on those pairs is the D2.1 op.

Across the pairs, the snap rounds up about as often as down. `screen` and `multiply` never land halfway between two levels, so their mean signed rounding error is zero. `mix` lands halfway on half of its channel sums, and those ties go to the even level (0, 6, or 12), which sends ten of the eighteen cases down and eight up.[3]

Each op has 46,656 lines, of which 2,975 are red and 13,225 non-red by dose, the larger of the two operand rednesses. Those counts are the same for every op, since dose reads the operands.

| op | per channel | on grid | redder than both |
|---|---|---|---|
| `mix` | (x + y) / 2 | 12.9% | 4,920 (10.5%) |
| `add` | min(x + y, 15) | 100.0% | 2,165 (4.6%) |
| `screen` | 15 − (15 − x)(15 − y) / 15 | 17.1% | 1,427 (3.1%) |
| `multiply` | x · y / 15 | 17.1% | 3,893 (8.3%) |
| `lighten` | max(x, y) | 100.0% | 0 (0.0%) |
| `darken` | min(x, y) | 100.0% | 4,258 (9.1%) |

The five named rules are the blend modes of the same names in Photoshop and Krita, on the 0..15 scale; `add` is what Photoshop calls *linear dodge*. `mix` is a normal blend at half opacity, which is the per-channel mean.

All six ops are commutative, as `mix` was in D2.1, so operand order carries no information and the two operand roles stay interchangeable.

The set leans light. `add`, `screen`, and `lighten` can only raise a channel; `multiply` and `darken` can only lower one; `mix` sits between its operands. A `subtract` op would balance `add`, but it would also be the first op whose operand order mattered, so the set stays as it is.

**Agreement.** How often each pair of ops gives the same answer. On a line where two ops agree, reading the op word gains nothing:

| | `mix` | `add` | `screen` | `multiply` | `lighten` | `darken` |
|---|---|---|---|---|---|---|
| `mix` | – | 0.0% | 0.0% | 0.5% | 2.6% | 4.1% |
| `add` | 0.0% | – | 37.6% | 0.0% | 17.1% | 0.0% |
| `screen` | 0.0% | 37.6% | – | 0.0% | 37.5% | 0.0% |
| `multiply` | 0.5% | 0.0% | 0.0% | – | 0.0% | 37.5% |
| `lighten` | 2.6% | 17.1% | 37.5% | 0.0% | – | 0.9% |
| `darken` | 4.1% | 0.0% | 0.0% | 37.5% | 0.9% | – |

`screen` agrees with `add` on about a third of pairs and with `lighten` on about a third as well, at the light end of the range where all three saturate; `multiply` agrees with `darken` at the dark end. `add` and `lighten` agree on a sixth of pairs. Every other pair agrees on a few percent at most, at the ends of the range.

**Op-relevance.** Take the lines of one op and ask of each line: how many of the other five ops would have given the same answer for this pair? Call that count k. Where k = 0, no other op matches, so the model has to read the op word to get the line right. Where k is larger, the op word narrows the six ops to k + 1, all of which give the same answer, so the model could answer the line without reading it.

The table gives, for each op we might anchor, the share of its lines at each k. The suppression experiments rest on this: when the op concept is removed, only the lines at k = 0 can show it, so an op with most of its lines there makes a sharper test. Which op we anchor stays open until the suppression prereg:

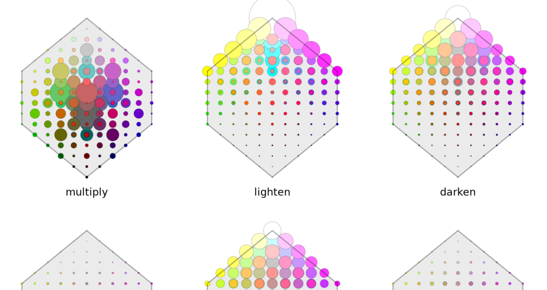
| anchored op | 0 | 1 | 2 |
|---|---|---|---|
| `mix` | 94.2% | 4.5% | 1.3% |
| `add` | 62.4% | 20.5% | 17.1% |
| `screen` | 41.9% | 41.0% | 17.1% |
| `multiply` | 62.5% | 37.0% | 0.4% |
| `lighten` | 59.9% | 22.1% | 17.9% |
| `darken` | 58.8% | 39.8% | 1.3% |

## The corpus

100,000 lines, the same count as D2.1. Ops are drawn uniformly, and within an op the training pairs are drawn uniformly with random operand order, the way ex-2.1.10 drew the `mix` pairs. For each op, a fifth of its distinct pairs is held out, keyed on (op, pair), so a pair held out under `add` may still be trained under `mix`. Corpus seed 0.

D2.1 drew its 100,000 lines from 5,832 distinct ones, so the corpus held about seventeen copies of each. Here the 100,000 lines come from six times 46,656 distinct ones, so most distinct lines appear at most once, and most pairs of an op never appear under that op. In E4, the `corpus` arms keep the 100,000 lines and narrow the op set, while the `per-op` arms narrow the op set and keep 16,666 lines per op.

The operands are the whole grid for every op, as they were in D2.1; what differs by op is where the answers land. `add` sends a fifth of its pairs to white, `screen` crowds the light half of the cube and `multiply` the dark half, and `mix`, `lighten`, and `darken` spread their answers through it.




***Where each op's answers land.*** *One mark per grid color, with area proportional to the number of unordered pairs whose answer is that color, on one scale for all six panels: a mark that fills its grid cell stands for 600 pairs, out of 23,436.*

## Training

D64-L4 nGPT with the ex-2.1.3 data config. A step is a batch of 64 random crops of 64 tokens (about ten lines each) from the tokenized corpus. An epoch is enough batches to sample a quarter of the training tokens, which the loader sets with an `oversample` of 16 over the batch size. At the corpus size of D2.1 that comes to 33 steps per epoch, as it did there.

Peak LR is 0.01. Anchoring uses the D2.1 code path: the pooled either-operand labeller (each operand draws at $\text{redness}^8 \times 0.04$), the anchor on $e_1$ at every slice, and the anti-subspace term.

Conditions are matched on compute, at the step count of D2.1, rather than on samples per op. So the per-step regime of the anchor matches D2.1, and how often the task sees each pair becomes a covariate that H1 reads.

The recipe runs for the 100 epochs of ex-2.1.10; the proposals and its short arm run for the 50 of the survey. The `per-op` arms of E4 keep the step count and shrink the corpus, so they get more and shorter epochs; the conditions table prints them.

Every schedule keyframe is a fraction of training, as ex-2.1.11 restated them: LR and anchor warm-up over the first 10%, the anchor anneal for the recipe over the last 10% down to a 0.1 floor, and the anti-subspace weight annealing from its peak ratio to 0.3 of the anchor weight by its own end fraction. The parameters for the proposals are the sampled values from the survey, unrounded, in `experiment.py`.

## The probe set

The probe set is a fixed list of equations that every run is scored on, separately from training: 27 lines per color as op1, per op. For `mix` these are its on-grid partners, so the `mix` probe set is the D2.1 one with the op word changed. For the other ops we draw 27 partners per color once with seed 0, and share them across every run.

Each op then has 5,832 probe lines, red and non-red by dose. We keep the infix frame,[4] so the probe positions in `sca.compute.evaluation` still read correctly.

## Measurements

The placement statistics are the ones from ex-2.1.10, computed by its code on the probe lines of each op from one clean pass per run: m_line and the softmin profiles, containment, the lead weight, contrast, grading against the sim^1.5 target, and the latch. The trajectory records m_line every 50 steps. Holdout exact match is read on a held-out set per op, of the size that ex-2.1.10 used.

The interventions run one teacher-forced pass per run per operator over the six probe sets, through the eval contract, giving the log-softmax at `=` and the write per (slice, line, position) as in ex-2.2.1. Red-line accuracy and the non-red deficit are read per op.

**Answer distance.** Exact match scores a one-step miss and a far miss the same way. Under suppression, ex-2.2.1 found that most misses on red lines were one-step neighbors of the true answer. So beside each accuracy we read the distance in the unit cube from the decoded answer to the true one, as ex-2.2.1 read it on the red lines.

A second version takes the expected value of that distance under the answer distribution, using the whole log-softmax rather than only its argmax. Both are computed per line and averaged per group, for every operator row, and neither is gated.

**Noise floor.** The bands use the per-run $\sigma$ of each statistic at the reference recipe, from the nine seeds of ex-2.1.10, printed in the H3 table; E5 re-measures them here. Gates score seed means against fixed thresholds and do not use the floor. Every difference between two conditions is quoted with its band, and a difference inside the band is reported as not resolved.

## Budget

77 training runs: 10 at the full length of the recipe, 25 at half of it, and 12 for the richer-op arms at the full length as well, each at a plain D2.1 step count on an L4. Scoring is one clean pass plus four operator passes over six probe sets per run, taking CPU seconds each, and the cube probes of E4 are ridge fits on 216 rows. That is well under the cost of ex-2.2.2, which trained 24 runs at twice the step cost. The run took 41 minutes of wall-clock on Modal at twelve containers, 15 of them training, and cost $2.33, with no task failed or retried.

## Before the freeze

Before the freeze we run a calibration: one seed of each control arm on the six-op corpus. It checks that the grammar is learned at both lengths, which means holdout exact match per op near 1, the level the D2.1 control reached. The rounded ops are the ones to watch. If the grammar is not learned at one length or the other, we change the corpus size, the step count, or the set of arms before the freeze, and record the change with a `REVIEW` note. Nothing anchored runs before the freeze.

**The calibration ran.** One seed of each control arm, published under `reports/m2/ex-2.2.3/calibration`; the table below reads it. Both lengths learn all six ops: holdout exact match is 1.0 on every op but one at each length, and that one misses a single line of its 256. The rounded ops are as clean as the others, and held-out surprisal is below 0.02 nats everywhere. So the corpus, the step counts, and the set of arms stay as designed, and `control-short` stays too, since its numbers say it will read the proposals' task cost at their own length without a training deficit of its own. Nothing about the arms or the gates changed after this read.

| run | `mix` | `add` | `screen` | `multiply` | `lighten` | `darken` |
|---|---|---|---|---|---|---|
| `control-s0` | 0.996 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| `control-short-s0` | 1.000 | 1.000 | 0.996 | 1.000 | 1.000 | 1.000 |

▶ **Glossary...**

1. The winner's curse: when you pick the top result out of many noisy trials, you tend to pick one whose noise ran in its favor, so a fresh run of it usually scores lower. ↩

2. Whereas previously, the operation was constant (`+`), so the model had no reason to use that position for anything other than spare compute. ↩

3. If the snap always broke ties the same way, every rounded `mix` answer would shift in that direction, and the model would learn the shift as part of the rule. The `mix` of D2.1 rounded half up, which on this grid meant up by a whole level on every off-grid pair. D2.1 never trained on those pairs, so the change costs nothing there, and the probe set is the on-grid pairs, which are untouched. ↩

4. An infix operator sits between its operands, as in `c1 mix c2`, rather than before or after them. The frame is the fixed six-token layout of a line: op1, the op word, op2, `=`, the answer, and the newline sit at the same positions in every line, so the probe code can read a role by its position. ↩