D2.2 design: anchoring an operation, and the first interventions

Archived 2026-10-05. This is the D2.2 plan as it stood through ex-2.2.22, kept so that the reports and backlog items that link to its sections still resolve. It is no longer updated. The current plan is the D2.2 design.

A plan for the second deliverable of M2, laid out several ways: the claims we want to be able to make, the experiments in order, the engineering that has to come first, the risks each experiment retires, and what is out of scope.

Note, 2026-09-25. The pivot is adopted: after ex-2.2.14, D2.2 anchors an op the model infers from solved examples in its context, with no word to name it. The prep sections below record the work that led there. The plan from anchor an inferred operation on is written for the new grammar, and the quick route puts it in order.

Note, 2026-10-04. Ex-2.2.21 ran round 2 again on the reworked recipe, and it changes the plan for round 3 in four places. The anchor sits on the example answers and hardly at all at the query `=`, so the edit that works acts at every position, and an edit at the query `=` that fails is now the expected result (see the bypass test). The hinge arm is the lead, and its cap of 0.8 was never searched. The posterior takes about three distinct values, so the graded stimulus needs a wider spread (see anchor an inferred operation). And the edited model already answers much as an ideal predictor that no longer knows the op, so round 3 may not need a fallback condition at all; ex-2.2.22 tests that on the stored runs (see deps).

Inputs: the D2.1 close-out (ex-2.1.11, ex-2.1.12, and the post), the first D2.2 result (ex-2.2.1), the anchored-op smoke test (ex-2.2.14), the D2.2-tagged backlog (`./go todo --tag D2.2`), the D2.1 kickoff lessons carried over from the autoencoders, and the related-work delta.

What we want to be able to say

1. **Anchoring generalizes past a token attribute.** *Red* is a property of the token at a known position. An *operation* is a property of the computation: its evidence is spread over the solved examples in the context, and its use is at the query `=` and after, deeper in the stack. D2.2 is where we find out whether the anchor captures more than the identity of a token at a labelled site.
2. **Suppressing the anchor removes the ability, and the removal is graded and selective.** In M1 this had three parts: a dose-response curve, selectivity (orthogonal colors untouched), and an analytic bound the observed damage approached.
3. **The side-effects were boundable before intervening.** Post-hoc methods now bound side-effects too, estimated from emergent geometry (COAST, arXiv:2605.01167; pre-intervention prediction, arXiv:2606.08365). SCA's distinctive claim is that its bound holds by construction, because the geometry was placed.

D2.1 never ran an intervention. The first suppression of a transformer was ex-2.2.1, on the *red* checkpoints already in the store; the second claim now has its first transformer data, and the third its write-bound half.

Quick route

We will likely run more experiments than this plan lists, and each one takes days to prepare, run, and review. Compute is cheap next to a review round, so the route keeps the number of rounds small, following four rules: **(a)** add arms rather than rounds, including arms that pay off only if something goes wrong; **(b)** ask a question of stored checkpoints before training for it, as ex-2.2.1, ex-2.2.8, and ex-2.2.10 did; **(c)** freeze each branch before the data, so a result picks the next step rather than prompting a new plan, as ex-2.2.12 did; **(d)** store every arm, so that the questions we did not plan for can also be scoring-only passes.

The route had four rounds for D2.2. The first was dropped, so it now starts at round 2; the numbers stay as they were, since reports cite them.

1. **Suppress the op word on ex-2.2.14: dropped.** Ex-2.2.14 put the op word on e_1 and little else, so an edit there would most likely do what masking the word does, which the geometry already shows. The in-context grammar has no op word, so the result would not carry over (see suppress the operation).
2. **Pilot the in-context grammar**, with the control and the anchored arms in one experiment at a few seeds each. It replaces four rounds the old plan would have needed: the posterior scouting report, a new-grammar control, an anchoring smoke test, and a label pilot. Its method section computes the posterior and the Bayes ceiling from the op table, which is the scouting report folded in. It also runs a scoring-only suppression pass on its own checkpoints, so the prereg takes its operator and dose axis from the same round. The arms are listed under the pilot.
3. **Anchor and suppress the inferred op**, the preregistered experiment, at fresh seeds. It has the equivalence read, the graded stimulus, suppression with and without the fallback, the bypass test, and the intervention side of the layer sweep, which needs no training of its own. The seeds go where a decision needs them. Whether the anchor costs the task anything is the comparison ex-2.2.21 could not settle at five seeds, so the control and the chosen arm get the most, and arms that only need to show a direction can have fewer.
4. **The anchor side of the layer sweep, and the SGTM baseline**, as one experiment. Both are training conditions compared with the recipe and control that round 3 froze, so they can share a round and the control seeds.

The first read of D2.3 comes after round 3. If verification lines are in the corpus from the pilot on, that read is scoring only on round 3's checkpoints; otherwise it is a fine-tuning stage on them. The pilot therefore has a verification arm, and the prereg adopts verification by a frozen rule if it leaves completion unchanged (see the pilot).

A scout on which slices the pull reaches may come between rounds 2 and 3. In ex-2.2.21 the `no-emb` arm, which leaves the embedding slice out of the pull, had the highest task score of any arm, but its edit was less selective. An inferred op probably sits in the middle of the stack, so restricting the pull there (no embedding slice, no last slice, or both), crossed with a lower anchor weight, could change what round 3 anchors. That would be the anchor side of the layer sweep moved earlier, at a small scale; the slice-restriction item has the arms.

When a result raises a question the route does not cover, the question goes first to the stored checkpoints, and a new training round is for the questions they cannot answer.

Experiments sketch

Loose plan for experiments to run.

Prep A: Suppression of concrete concepts (operands)
Suppress red on the existing checkpoints
Prep B: New grammar with more operations
The multi-op grammar, with red anchored again
Prep C: Embeddings
Un-anchored embeddings
Anchor one operation
Anchor an inferred operation
Suppress the operation (and the operands)
The pilot
Suppress the operation (and the operands)
Suppress the operation (and the operands)
Suppress the operation (and the operands)

with e_1 at every slice, so any edit there would be hard to tell apart from masking the word. The pivot retires the op word, so what the pass could find would not carry over to the new grammar. Its one open question, whether the small alignment at `=` marks the op or carries it, moves to round 3: in the in-context grammar, `=` and the answer are the only places the op can be carried.

Two observations from planning the pass are recorded here as open questions, not settled method. First, the plain projection behaves badly on a state that lies almost on e_1: the re-projection onto the sphere scales up whatever small remainder is left, so the landing is set by that remainder. The draft proposed the reflection or a repulsion onto a far landing (down to the antipode) for such states, but that use of the repulsion is outside what it was designed for, and it has not been reviewed; which operator suits such states, if any, is undecided. Second, an edit at a position whose token names the concept is best compared with a token mask at the same position.

The **bypass test** the D1.3 post left open: suppress at the query `?`, at the query `=`, and at every position, at one slice and at every slice, in the blocks only and with the embedding. Where suppression fails to take effect, the model is reading the op from somewhere the axis does not reach. That is a finding about anchoring, and it feeds the layer sweep. The inferred op makes the test sharper than the op word did: the evidence is spread over the examples, so an edit at the query that fails would point to the examples as the route around it.

Ex-2.2.21 found the anchor on the example answers, before the query `=`, and no edit at the query `=` met both suppression criteria on any arm. So a failed edit at the query `=` is expected, and says little about a route around the edit. The more informative form of the test keeps the edit at every position and varies the slices it acts on, and whether it includes the embedding.

Nice to have: a sweep over all ops, to see whether they can all be suppressed equally well.

Layer sweep

Anchor at subsets of slices (single ℓ, prefix $\leq \ell$, suffix $\geq \ell$, all) on a frozen schedule, then run the suppress operation intervention at the anchored slices. This tests the claim that bounds are layer-local. The geometric bound covers the immediate write, and we need to know whether later blocks amplify or absorb the edit.

The sweep has two sides. The intervention side (the same model, edited at a prefix or a suffix of slices) needs no training, so it runs in round 3 on the prereg checkpoints. The anchor side (models anchored at a prefix or a suffix) trains, and runs in round 4 beside SGTM, at the seeds the intervention side says it needs.

Ex-2.2.1 already says what to expect for *red*. The concept is read in the first two blocks, and the last block does not read it. So a removal acting at the embedding and the first block should match the full intervention, and one that starts later should not. That is a prediction about prefixes and suffixes, so we favor a prefix/suffix bracket over single slices. For the inferred op, function-vector work puts the task direction from about mid-depth on (see the pivot), so the bracket may turn the other way.

Caveat on reading depth this way: at d64-L4 the model has capacity to spare, so where it reads the concept may say more about what it can afford than about what the task needs. Inferring the op from examples is more work than reading a word, so it may draw on more of the depth. The similarity between slices (CKA, as in the pivot) is cheap on the same checkpoints and would show whether the model has an early and a late block at all.

SGTM baseline

Its own training condition (arXiv:2512.05648), method-agnostic through the eval contract and reusing our labels, in round 4 beside the anchor side of the layer sweep. RMU and SAE rows wait for D2.3, as the baselines item sequences them.

Write-up

The D2.2 post.

Deps

- **The operation as a variable.** As specified in the backlog item: an op table (name, surface form, grid function with defined rounding, closed on 0..15), `op` on `Example`, seen-pair bookkeeping keyed on `(op, pair)`, ops spelled as words, and the infix frame kept for the probes. The table of record is A+ (eleven ops, ex-2.2.4), adopted by the handover: `add` is gone, `difference`, `exclusion`, and `hsvmix` are in, and three HSV blend modes read operand order. The first table was `mix` (the D2.1 op), saturating `add`, `screen`, `multiply`, `lighten` (per-channel max), and `darken` (per-channel min), specified in ex-2.2.3: each rule is computed on the 0..15 scale and snapped to the nearest level of the six-level grid, so every op answers every pair with a vocabulary color. An earlier draft of that prereg kept D2.1's closed-pair rule instead, under which `screen` and `multiply` are degenerate; the rounding rule makes all six total and distinct (no two agree on more than about 38% of pairs). Every new op departs from `mix` on more than 90% of pairs, so the model has to read the op; `add` – `screen`, `screen` – `lighten`, and `multiply` – `darken` each agree on about a third of pairs and populate the middle *op-relevance* levels. The relevance distribution for each candidate anchored op, under the table's own rounding, is rendered in ex-2.2.3's method, since the per-line predictions in suppress operation rest on it. `divide` needs a saturation rule and is lumpy on a 16-level grid, so it stays out of the first table. Ops that convert through other color spaces (`hue`, `saturation`, `brightness`) were filed at richer op set and have since been read on the grid by ex-2.2.4, which proposes the second table, A+, described under scouting. With ops that read operand order in the table, relevance is counted over ordered pairs.
- **The in-context grammar**, before round 2 of the quick route. From the pivot: **(a)** a context generator: solved examples and a query under one op, `?` in place of the op word, replacement op noise at a rate ρ, and cube noise at a low rate; **(b)** a line array and a role array the size of the corpus, computed from the positions of `\n`, in place of the arithmetic on a fixed `LINE_TOKENS`; **(c)** an op array beside the corpus, one entry per line, and a labeller keying that reads it; **(d)** the posterior over ops for each context and each prefix (the prefix form is for label variant (c)), the target null°, and the Bayes ceiling, all from `answer_dist` in `sca.data.ops`; **(e)** a block size large enough to fit at least two whole contexts; **(f)** the calibration read; **(g)** a hinge cap on the anchor term; **(h)** verification lines: the `|` marker, the `TRUE` and `FALSE` verdicts, the language-model loss masked on the candidate answer of a `FALSE` line, and a verdict score. An attention mask that resets at `\n` is optional, since the model would likely learn to ignore the previous lines without it.

  Target null. Where a removal should land...
- **The fallback on the in-context grammar.** `sca.fallback` was written for *red* on one-equation lines: its mask reads the redness of an operand, its target is one token at a fixed offset from op1, and it reflects the state at one slice. A version for round 3 would need the target null as soft labels at the query answer, a mask on the contexts of `difference`, and a rehearsed edit that matches the one round 3 scores, which acts at every position. It may not be needed: ex-2.2.21 found the edited model already answers much as the target null does, and H2 of ex-2.2.22 tests whether it lands there closely and consistently across seeds. If it does, round 3 drops the fallback condition and keeps the target null as the reference the edit is scored against.
- **The eval contract and operator library.** Landed with ex-2.2.1 as `sca.intervention`: the triple, the three operators, and the post-hoc fitters. Every method produces a triple of `(model, subspace, intervention operator)`, and one scorer takes that triple. The operators are axis projection with a strength γ (γ = 2 is reflection), the shaped suppression from M1, and weight ablation. Every training experiment from fallback control on scores its checkpoints through the contract, so adding an operator or a row is a change to one module.

The contract also pins where operators act and what they do to the norm. The hook point is the between-block stream, meaning the slices that `residual_stream()` returns, which are the same states the anchor term reads. The stream is unit-norm (nGPT), so axis projection composes with a re-projection back onto the sphere. The state lands on the great subsphere where zero-concept states live, and the surviving components pick up a per-position gain of $1/\sqrt{1-x_1^2}$. That gain is computable beforehand, so it belongs inside the bound. Weight ablation declares its order against the `normalize_weights` constraint, which rescales a matrix once entries are zeroed.

The 2025 auditing rows are relearning rebound (arXiv:2505.22310), activation perturbation (ActPert, arXiv:2505.23270), and off-axis recoverability (arXiv:2605.11685). They were to be declared before any arm was scored. Ex-2.2.1 scored its arms without them, and the decision is now recorded under fallback control: off-axis recoverability runs there, and the other two wait for D2.3, which speaks their language in full. - **Fold the survey lessons into the plan template**: constraint margins ranked beside the objective, multi-seed promotion near a gate, and the publisher carrying every statistic the analysis promises.

Decisions

Only what the plan above already commits to; everything else stays open until an experiment forces it.

- Suppress red ran before the op grammar was ready: no training, and the highest information per dollar in the plan. Done.
- The intervention for the op experiments is chosen by a scoring-only tuning pass on stored runs, before the anchor-op prereg; until then the fallback experiment carries all three as ride-along rows. Done, as ex-2.2.8 on ex-2.2.3's checkpoints: the plain projection on the adopted point, with `operands` beside it and `shaped-a0.4-p0` recorded as the syntax-free candidate.
- The bound claim is layer-local from the start. The geometry bounds the write; what the behavior does in response is a prediction, so a behavioral miss feeds the layer sweep rather than falsifying the bound.
- Anchor operation opens with a small smoke test, before the many-seed equivalence read. Done, as ex-2.2.14; the equivalence read then rides with the suppression prereg, which trains the fresh seeds it needs.
- The first operation suppression was to be a scoring-only pass on the stored checkpoints of ex-2.2.14, with a token-mask row as the reference for every edit. It was dropped after the pivot (see suppress the operation), and the operator and dose axis come from the pilot instead. The op is anchored alone; *red* beside it waits for the two-anchor follow-up.
- The survey is confirmed on the new grammar as a set of proposals, and not replicated literally first.
- The dose axis for the categorical concept is intervention strength; the stimulus side (*op-relevance*) supplies the per-line predictions and the bound.
- The fallback mechanism is the antipode redirect at the first anchored slice, which is the embedding while the embeddings are anchored, with a stop-gradient at the reflected state. Rehearsing the intervention during training is refiled as an auditing question.
- The fallback answer for a continuous concept is the center of the operand-averaged null; for *red* that is the visible operand mixed with mid-gray. A categorical concept takes the mode of its null instead. Either is a design choice, since the answer can be any function of the line: to have *red* read as *black* under intervention, we would use the same term with black mixed with the visible operand as its answer.
- Auditing rows: off-axis recoverability in ex-2.2.2 as an exploratory row; ActPert and relearning rebound at D2.3.
- Which op to anchor: `difference`, chosen on paper for ex-2.2.14 from table A+'s relevance distributions (ex-2.2.4). On 75% of its lines its answer names it alone, the most of any commutative op; it is total on the grid, so its task read is not capped by stochastic rounding (the control reaches 0.97 held-out on it against 0.43 on `mix`); and it is commutative, so operand order plays no part. The smoke test read every other op the same way at three seeds, with a frozen fallback rule; all ten qualified with margins inside 0.04 of each other, so the choice stands. The reference op for the *red* gates (`mix`, with `hsvmix` beside it since ex-2.2.9) is a separate question and is untouched.
- The second table (A+), stochastic rounding, and the whole-line labeller are proposals from the scouting round, and only the handover prereg adopts them. The removal statistic there is a distance from the correct answer rather than exact match, with the direction the answer-distance re-score adopts for removal and selectivity. The labeller goes in with a selectivity check, because of the tail ex-2.2.7 saw on its whole-line arms.
- The handover (ex-2.2.9) is not adopted as it stands: removal missed on the HSV ops. Ex-2.2.10 traces the miss to the removal rule rather than the model, so the re-run keeps the recipe and changes the reads: removal lines are the red lines whose answer needs the red operand's hue, retention is the end-of-training alignment° over its value at the anneal's start, and $\bar{\alpha}$ at op1 is reported beside the reference conditions with no gate.

  Alignment. How far a state points along e_1...
- The handover re-run (ex-2.2.11) is not adopted as it stands: removal missed on `hue-hsv` alone. One more scouting round (ex-2.2.12) runs before the anchored-op prereg, with its promotion rule and its no-fix branch frozen in advance: a proposal has to clear the missed gate by more than the seed band, and if none does, the re-run gates removal on the ten other ops and reports `hue-hsv` as the op where one axis has a known blind spot.
- The plane anchor is defined as alignment with the span of e_1 and e_2 (the length of the projection onto the pair, unsigned), with the anchor and anti-subspace terms, the trajectory reads, and the projection operator all taking the pair where they took the axis, and every plane read compared with the control checkpoints read on the same pair. The concept's variance share is charged two coordinates. Settled for the sweep, and read again at twenty seeds by ex-2.2.13, which found the plane a validated alternative and did not adopt it: it is lower on the kept share at every rung and no tighter across seeds. The plane item carries the alternatives.
- Every slice is anchored, the embedding included, and the readout is untied. Ex-2.2.7 found that leaving the embedding un-anchored does not clean the syntax embeddings and makes removal less complete, while untying the readout cleans them at no visible cost. The handover prereg adopts this and reads it against the hard-zeroed ceiling from ex-2.2.7.
- The layer sweep takes the shape of a prefix/suffix bracket, following the depth read from ex-2.2.1. Its intervention side is scored on the prereg checkpoints, and its anchor side trains beside SGTM in one experiment; the grid question is deferred to that experiment.
- The pivot is adopted: the op is inferred from solved examples in the context, and op words leave the grammar. `difference` and table A+ stay.
- The plan follows the quick route. The posterior scouting report, the new-grammar control, the anchoring smoke test, and the label pilot are one experiment, the pilot, whose frozen rules say what goes forward to the prereg.
- Verification lines are an arm of the pilot, and stay in the corpus from the prereg on if they leave completion unchanged, so that the first D2.3 read is scoring only. A fine-tuning stage is the other option; a model fine-tuned onto a working completion circuit would likely verify by reusing it (route 1 in the pivot), which is where an asymmetry is least likely to appear.

Risks and mitigations

Risk	Would look like	Retired at
Suppression does not bite even on *red*	Accuracy unchanged after projecting e_1 out; color is read from elsewhere	Retired: ex-2.2.1 (red accuracy 0.09)
The bound is loose or wrong in a transformer	The non-red write exceeds its geometric bound, or the damage outruns the write-size prediction	Write half retired at ex-2.2.1; the behavioral half belongs to the layer sweep
The response to suppression is undesigned	Suppressed lines scatter by seed; completions leave the color vocabulary	Measured at ex-2.2.1 (13% seed agreement, completions stay in vocabulary); fallback control pins it
Recipe is grammar-specific	The proposals from the survey do not reproduce on the new grammar	Retired at ex-2.2.3: the recipe and every proposal reproduce (H2, H3), at a plateau 0.05 lower than the survey's; the frozen rule adopted `t00`, and a post hoc read with the lead and selectivity gates narrows the choice to the recipe, whose short arm D2.2 builds on (`recipe-short`, by decision after the twenty-seed E6 read)
The syntax embeddings hold the axis, so a full-position edit costs the non-red lines	Non-red lines lose accuracy under the plain projection, at the op-word and `=` positions	Found at ex-2.2.1 and ex-2.2.3; the mechanism is the tied readout, at ex-2.2.7; the untied readout is confirmed on the new grammar at the handover and again at fresh seeds in ex-2.2.13
The *red* leftover on `hue-hsv` confounds the anchored-op reads	A removal read on a model carrying both anchors moves with which lines survived on *red*	Measured at ex-2.2.11 and ex-2.2.13: about a quarter of those answers, from no fixed set of lines. Ex-2.2.14 carries no *red* anchor; the two-anchor follow-up measures it at its own seeds
Task cost grows with an abstract concept anchor	Gate misses in anchor operation that new grammar did not have	Retired for the op word at ex-2.2.14: no op moves by more than 0.005, on the anchored op or any other. Read again for the inferred op in the pilot
Anchoring an op captures the token, not the operation	Suppression at the op word removes the op no better than masking the word, and edits away from it are inert	Found likely at ex-2.2.14, which put the op word on e_1 and little at the use sites; an edit there is hard to tell apart from masking the word, so it is not measured. The pivot takes op words out of the grammar
The control does not learn the in-context task	The control stays well below the Bayes ceiling, or its answers are poorly calibrated to the posterior	The pilot, with a larger control as an arm
The anchor picks up a shortcut	Alignment does not grade with the posterior on `difference` across the middle band	The pilot, then the prereg at fresh seeds
The label asks for the op before the context shows it	A task cost or a shortcut under the whole-line label that a label variant avoids	The pilot, with the label variants as arms
The query `?` saturates	Alignment near 1 at `?`, with no partial dose there	The pilot, with a hinge-capped arm
Bypass through attention or the residual	Suppression works only when applied at every site	Position side seen at ex-2.2.21: the anchor sits on the example answers, and only the edit at every position works. The slice side is for suppress operation and the layer sweep

The first experiments each changed one thing from D2.1, so a negative there was interpretable. The pilot changes the grammar and the anchored concept together, so its control arms do the regression check that ex-2.2.3 did alone.

Out of scope

The verification reads (D2.3), though verification lines may be in the corpus from the prereg on (see decisions). Topic markers, and anchoring at fine-tune time, both open questions in the pivot. Several ops on separate axes (a D2.3 candidate, for the subspace bound and the two-op concept swap). The feedback controller (the kickoff advice stands). RMU and SAE baselines, and a full LUNAR row beside them; a LUNAR-style redirect, the nearest analogue of the fallback, runs as an exploratory row of ex-2.2.2. Resolving D2.1's H2 decodability question (its own item), run when the claim is needed. Stream-vs-init attribution (kickoff queue item 4), until there is an anchoring failure worth attributing — D2.1 produced none.

---

1. A stop-gradient is an identity in the forward pass with a gradient of zero: a loss downstream of it cannot move anything upstream of it. Here it sits on the reflected states, so the fallback loss reaches the blocks and the unembedding and leaves the embedding table alone, and with it where the anchor put the concept. The other losses see the clean pass and are unaffected. ↩