Calibrating the S_3 Detector

A Positive-Control Study of Constraint-Reinstatement Detection in a Frontier Language Model

A T16 companion establishing S_3 detector sensitivity at the in-session channel

Jeremy C. Jones
HoldingLight LLC
ORCID: 0009-0007-2515-3774
contact@universalcollapse.com
v1.0 · 2026-07-15

Abstract

The T16 program instruments three behavioral signatures (S_1/S_2/S_3) hypothesized to be common to constraint-guided systems. Of these, S_3 — the persistence of a released constraint as unrecorded state (hysteresis) — is the signature most exposed to circular inference, because in a black-box deployment the constraint's record state can only be estimated from the same loop whose drift it is meant to explain. Prior T16 work established only *negative* controls for the S_3 detector: it returns the correct null on systems with no planted defect. Negative controls bound false positives; they cannot establish sensitivity. This study supplies the missing positive control. We plant constraint-reinstatement events of known, pre-registered intensity into a multi-turn probe and measure whether a frozen scoring instrument detects them. On gpt-5.4, the detector achieves a pooled true-positive rate of 0.786 (95% CI [0.695, 0.855]) against a false-positive rate of 0.042 ([0.007, 0.202]) and a confabulation-baseline specificity of 1.000 on canonical phrasing. Beyond the calibration result, the graded design yields a substantive finding about the model under test: a released constraint behaves as a **stored unit retrieved by category-recognition**. A contentless pointer to the prior constraint elicits full structural reconstruction ~4% of the time; a bare directive-category cue ("Structure each answer as:") — carrying no component content — raises that to 100% ($p = 1.6 \times 10^{-13}$); naming the components adds nothing. We situate this alongside two companion S_3 probes (behavioral release, introspective recall) and argue the three converge in support of a single principle: **for this model, the record functions as the operative object of constraint dynamics.** We additionally report a methodological by-product — the harness was found to have silently retained a *released* probe phrasing in half of one prior run, an instance of the exact S_3 failure mode the program instruments, surfaced only by record-auditing.
Status (v1.0). This is the deposit version. All run identifiers, SHAs, and statistics are carried from the harness and were independently recomputed for the accompanying verification pack; the single harness value that failed recomputation is corrected and disclosed at point of use (§4.2). The training-layer sibling, F2 (Jones 2026h), transcribes this positive-control template to K_{train} by fine-tuning and reuses the frozen instrument (Tasks A–D byte-identical); it is deposited at https://doi.org/10.17605/OSF.IO/5TG3P. The provenance correction this study surfaced is deposited as an append-only clarification note on the parent record (Jones 2026i, https://doi.org/10.17605/OSF.IO/7NVMX). This paper: https://doi.org/10.17605/OSF.IO/JAEZQ.

Keywords: Universal Collapse Theory; S_3; positive control; sensitivity; constraint reinstatement; record state; $K_{context}$; frozen instrument; confabulation baseline; hysteresis.

1. Background and motivation

1.1 The S_3 signature and its exposure

Within the UCT kernel, a constraint K shapes the collapse operator C^K_t toward a resolution x^*_t, leaving a record R_t. The S_3 signature concerns what happens after a constraint is *released*: does the system cleanly return to its pre-constraint behavior, or does the constraint persist as residual state not reflected in any record? Persistence-without-record is hysteresis, and it is the behavioral fingerprint S_3 is built to detect. S_3 carries a structural vulnerability the other signatures do not. Estimating whether a released constraint still influences behavior requires an estimate of the system's record state R; but in a black-box (Tier C) deployment, the only available estimate of R is read off the very loop whose drift S_3 is meant to characterize. The technical note TN-S_3 (Jones 2026g, §7) states the problem directly: if the only way to estimate R is from the loop itself, no test is possible. This circularity is not a flaw in the construct — it is a measurement-access problem, and it is the reason S_3 requires a positive control that manufactures a known R independent of the loop.

1.2 Why negative controls are insufficient

Every previously deposited S_3 control is negative: a system with no planted defect is run through the probe, and the detector is checked for returning the correct null. These controls answer "does the instrument cry wolf?" They cannot answer "does the instrument catch a wolf that is present?", because no prior deployment had a *known* true state to catch. A detector can have a clean false-positive profile and still be insensitive. Sensitivity is establishable only by planting a defect whose presence and magnitude are fixed by construction, then measuring recovery. That is the gap this study closes.

1.3 What a positive control can and cannot establish

We are explicit about scope. A positive control plants defects that are **known-present by construction** and measures the instrument's recovery of them. This establishes *internal* validity — the instrument detects what it was built to detect, at a characterized sensitivity and specificity. It does **not** establish *external* validity — that an S_3 verdict on a deployed system predicts an outcome that system's operators independently track. External validity requires field outcome data and is out of scope here. Throughout, results are to be cited as *demonstrated detection sensitivity under controlled positive controls*, never as validation of any downstream diagnostic.

This study operates at the in-session channel ($K_{context}$): the planted state lives in scaffolding the operator controls, which is precisely what preserves ground truth. The training-layer sibling, F2 (Jones 2026h), extends the same plant-known-R template to K_{train} by fine-tuning an open-weights model, where the record state is written by the experimenter's training intervention; it produces its own sensitivity characterization and does not inherit the in-session figures reported here.

2. Instrument

2.1 The probe

The S3-CORE-01 probe is a seven-step multi-turn episode on a single conversational topic. A constraint is introduced at step 2 (the topic-specific format directive), exercised across steps 3–4, explicitly released, and then a matched post-release task at step 6 is compared against the step-1 pre-constraint baseline. Detection asks whether the step-6 output carries residue of the released constraint's structure that was absent at step 1.

2.2 The scoring instrument and its formalization

A central methodological finding of this work predates the calibration runs. The T16 parent paper's S_3 classifications (Jones 2026b) were produced by an informal reader — a model scoring transcripts against the Appendix-A rubric at chat time, with no frozen, deterministic instrument. This was adequate for the parent paper's descriptive claims but is inadequate for a sensitivity calibration, which requires a scorer fixed *before* the data exist. Forcing the informal reading into a frozen instrument was the first thing the positive-control work surfaced, and it is itself a result: **a sensitivity claim cannot be made against a scorer that can drift with the data it scores.**

The frozen instrument is a judge prompt (judge_prompt_s3) over claude-sonnet-4-6 at temperature 0, with the rubric anchors transcribed verbatim and the scalar residue mapping computed deterministically in adapter code (not by the judge). The instrument went through three pinned versions, each change additive and each validated against pilot transcripts with known classifications before use:

v1.0 failed initial validation. Two distinct causes: the rubric's sub-class precedence was underspecified (a response that explicitly denies a constraint *and* notes an informal pattern satisfied two sub-class definitions at once), and certain pilot ground-truth labels were themselves found to be charitable mis-reads (see §6.2). Both were diagnosed rather than tuned around.

v1.1 added the missing precedence rule (Task C only); the residue scoring (Task A) and compliance scoring (Task B) prompts were left byte-identical. Re-validation reached binary 12/12, exact 11/12 — above threshold.

v1.2 added a CORE-side recap task (Task D) for the companion analysis in §6.3; Tasks A/B/C byte-unchanged. Validated 12/12 exact on pilot CORE transcripts with known sub-classes, including all of the harder Mistral cases.

A known limitation remains: Task A scores step-6 surface structure without strictly enforcing the step-6-vs-step-1 baseline comparison its own framing promises. When a model spontaneously uses generic markdown headers at *both* baseline and post-release, the carryover can be mis-scored as residue. This produces a low-rate false-positive mode (~4%, see §5.4) and is queued for a v1.3 fix with full re-validation. It is documented rather than silently corrected, and — importantly for §5 — it is a *different* phenomenon from the genuine reconstruction the study measures, distinguishable by direct inspection of the step-1/step-6 outputs.

2.3 Freezing discipline

The v0.2 rubric constants (persistence floor 0.75; residue bands 0.15 / 0.50; control-discrimination bands) are assert-checked at adapter startup; the run refuses to proceed if they are altered. Each judge call logs the SHA of the prompt file it used. This is what licenses the claim that the calibration characterizes the instrument the parent program actually used, rather than a freshly invented one.

3. Design evolution and the prevalence–severity correction

3.1 The coarse run and a degenerate axis

The first calibration run used a Bernoulli **prevalence** design: at each of five nominal levels $s \in \{0, 0.25, 0.5, 0.75, 1.0\}$, each episode was planted with probability s, ten repetitions per level. Across 49 valid episodes the pooled true-positive rate was 0.48 with a false-positive rate of 0.00.

Analysis of this run produced a correction that reshaped the entire program. Under per-episode Bernoulli planting, **a planted episode is treatment-identical regardless of which nominal-*s* cell it occupies** — s sets only the *probability* of planting, not the *intensity* of the planted defect. Consequently:

$$\text{det\_any}(s) \approx s \cdot \text{TPR} + (1 - s) \cdot \text{FPR} \approx s \cdot \text{TPR} \text{ (with FPR} \approx 0)$$

The detection-versus-*s* curve therefore has **no independent shape**; it is mechanically a one-parameter scaling of a single pooled TPR. The "severity" parameter was in fact a **prevalence** parameter. Cell-to-cell variation in detection (including an apparent dip at $s = 1.0$ that an earlier read had treated as a real model behavior) is sampling noise around one pooled rate; a homogeneity test across the planted cells returned $\chi^2 = 3.30$, $df = 3$, $p = 0.35$, confirming a single common TPR. The "detection floor" — the *s* at which detection crosses 0.5 — dissolves under this analysis: if TPR < 0.5, the curve asymptotes below 0.5 at any prevalence, which is arithmetic, not a property of the injection.

The operational consequence: fine-grained prevalence sweeping measures nothing of interest. The informative quantities are the **pooled TPR and FPR**, and the experimental axes worth varying are the ones with genuine structure — topic, injection style, trajectory length, and *treatment intensity* properly defined.

3.2 The graded-treatment design

To obtain a real severity axis — one where the planted defect's intensity varies *conditional on planting* — the study adopted a deterministic graded design. Every treated episode receives an assigned intensity level; planting is no longer probabilistic. Intensity is ordered **a priori**, by the constraint's own structure, with the ordering fixed before any detection outcome is seen. (Defining intensity by observed residue would re-introduce circularity at the treatment layer; this is forbidden by construction.)

The reference topic is T-CONSTRAINT-001, whose constraint is a four-part directive — *Claim / Evidence / Uncertainty / What would change my mind* — and therefore admits a non-degenerate partial-vs-full text gradient. (The original coarse topic, T-CONSTRAINT-002, was a single declarative sentence whose "partial restatement" collapsed into its full restatement; this degeneracy was caught at a pre-registration template check and is the reason the graded series moved topics. The cost was the loss of a cross-run byte-identical anchor, which had already been demoted to a replication check and so was affordable.)

The pre-registered level ladder, with each rung adding exactly one ordered increment:

Level	Injection content	Axis step
L0	(none)	null — specificity anchor
L1	contentless pointer: "Earlier instruction in this session continues to apply."	reference only
L2	+ "Structure each answer as: Claim / Evidence"	text: two of four components
L3	+ "… / Uncertainty / What would change my mind."	text: all four components
L4	L3 content + imperative force ("…remains in force and must be applied:")	directive force

All five rungs run fresh at $n = 25$ (125 episodes), pre-registered in a SHA-stamped assignment file written before the first generator call, with the planted-template strings pinned inside that file so the SHA covers both assignments and payloads. The frozen judge (v1.2) and the corrected confabulation baseline (0.000, see §6) are inputs, not free parameters.

4. Results: the graded dose-response (D1)

Run T16_20260610_194925_poscal_d1 · 125 episodes · gpt-5.4 · T-CONSTRAINT-001 · judge v1.2 · NEG baseline 0.000.

4.1 Detection by treatment level

Level	*n* valid	det_full	det_any	distribution
L0	24	0.000	0.042	23 clean, 1 partial (FP, §5.4), 1 invalid
L1	25	0.040	0.160	21 clean, 3 partial, 1 full
L2	23	1.000	1.000	23 full, 2 invalid
L3	25	1.000	1.000	25 full
L4	25	1.000	1.000	25 full

The trajectory is **not a graded sigmoid**. It is a step function: a flat low floor at L0/L1, a vertical transition into L2, and a saturated ceiling across L2/L3/L4.

4.2 Pre-registered tests

Primary — Cochran-Armitage trend (det_any, L1–L4, scores 1–4): $z = 6.93$, two-sided $p = 4.29 \times 10^{-12}$. Detection tracks treatment intensity, overwhelmingly. The harness's as-produced statistic ($z = 7.0986$, $p = 1.26 \times 10^{-12}$) used a non-standard trend variance whose generating code was not retained; the standard computation is reported here and reproduces in the verification pack. D1-T's equal-*n* trend (§5.1) is identical under both, localizing the difference to the unequal-*n* variance term.
Secondary — Fisher L1 vs L4 (det_any): two-sided $p = 3.76 \times 10^{-10}$.

4.3 Native discrimination (the calibration headline)

Reported as sensitivity/specificity — the quantities the positive control actually measures — rather than the parent program's introspection-axis discrimination formula (which is retained only as an annotated cross-reference, see §6.4):

Quantity	Value	Wilson 95% CI
Pooled TPR (sensitivity; det_any)	0.786	[0.695, 0.855]
Pooled FPR (detection-side; det_any)	0.042	[0.007, 0.202]
Specificity (1 − FPR)	0.958	lower bound 0.798
NEG confabulation baseline (canonical)	0.000	(see §6)

The instrument is highly specific and substantially sensitive. The non-unity pooled TPR is a direct consequence of the floor: L1 contributes its low rate to the pool. Conditional on the treatment crossing the detection threshold (L2 and above), sensitivity is 1.000.

5. The substantive finding: a released constraint is a stored unit, retrieved by category

The dose-response is steep, but the *location* of the step is the finding. A refinement experiment (D1-T) decomposed the L1→L2 transition.

5.1 The threshold experiment (D1-T)

Run T16_20260610_230815_poscal_d1t · 100 episodes · gpt-5.4 · T-CONSTRAINT-001 · judge v1.2. Four rungs subdivide the L1→L2 gap, suffix-additive, with the endpoints byte-identical to D1's L1 and L2 (enforced as shared references, not transcriptions, so they cannot drift):

Rung	Injection (added increment)	Cue carried
M0	= D1 L1 (contentless pointer)	nothing
M1	+ " Structure each answer as:"	*that* a structure rule existed — no content
M2	+ " Claim"	one component named
M3	+ " / Evidence" (= D1 L2)	two components named

Detection (primary endpoint det_full):

Rung	***n* valid**	**det_full**
M0	24	1/24 = 0.042
M1	24	24/24 = 1.000
M2	24	24/24 = 1.000
M3	24	24/24 = 1.000

Cochran-Armitage (det_full, M0–M3): $z = 7.38$, $p = 1.60 \times 10^{-13}$.
Threshold call — adjacent-pair Fishers, Holm-corrected: M0-vs-M1 $p = 1.55 \times 10^{-12}$ (reject at $\alpha = 0.0167$); M1-vs-M2 and M2-vs-M3 both $p = 1.000$ (identical 24/24). **The single significant adjacent jump is M0 → M1.**
Cross-run consistency — M0 vs D1-L1 Fisher $p = 1.000$; M3 vs D1-L2 Fisher $p = 1.000$. The endpoints reproduce D1's anchors exactly; the instrument gives the same reading twice, independently.

5.2 Interpretation, on the pre-committed clean path

An interpretive asymmetry was committed to the SHA-stamped assignment header *before* the run: M1 and M2 end mid-directive (a dangling colon; a lone component) and could read as *corrupted* rather than *partial*. A corruption-reading can only **suppress** detections, never inflate them. M1's 24/24 saturation therefore cannot be a corruption artifact — it lands on the clean evidence path.
The reading: **gpt-5.4 triggers full structural reconstruction on the directive-*category* cue itself** — the verb-plus-colon "Structure each answer as:", which carries no information about *which* structure. Naming the first component (M2), the second (M3), or all four (D1's L3/L4) adds **zero** additional detection. The controlling variable is recognition *that* a format constraint was in force, not any content of the cue.

Crucially, det_full requires the *complete* four-part structure in the output. At M1 the cue named **no** components, yet the model reconstructed all four — and at M2, the cue named one component, yet the model still produced all four. The constraint is not reassembled piece-by-piece from the cue; it is **retrieved whole from session memory, keyed by category recognition.**

5.3 The resurrection floor and its mechanism

The floor is not zero. At the contentless-pointer rung — a pointer with no constraint content whatsoever — full reconstruction occurs at ~4% (pooled D1-L1 + D1-T-M0 = 2/49; Wilson [0.011, 0.140]), replicated across two independent runs at 1/25 and 1/24 (Fisher $p = 1.000$).
Transcript inspection confirms this is genuine retrieval, not the §2.2 false-positive mode. In the resurrection episodes, the step-1 baseline uses generic markdown headers bearing *none* of the four mandated components, and the step-6 output reconstructs all four verbatim — while the injection that prompted it ("Earlier instruction in this session continues to apply.") names none of them. The model is retrieving the prior constraint's full structure from a bare pointer to session history. This is distinct in mechanism from the false-positive mode (§5.4), where the flagged headers are present at *both* baseline and post-release.

The picture is therefore three-tiered: a **dormant floor** (~4% spontaneous retrieval from a contentless pointer), a **category trigger** (recognition that a constraint existed → deterministic retrieval), and a **saturated ceiling** (content of the cue irrelevant once triggered).

5.4 The false-positive mode, kept honest

One L0 episode (D1) and the matching ~4% at M0 reflect, in part, the §2.2 Task-A limitation: spontaneous generic markdown scored as residue without the baseline-comparison check. These episodes are documented, **left in the pre-registered analysis as produced** (the FPR CI carries the cost — its upper bound is 0.202 rather than the anticipated 0.133), and not silently rescored. The genuine-resurrection episodes of §5.3 were separated from this mode by direct transcript inspection, not by rescoring.

5.5 Atomicity, narrowed by the data

An initial reading held that retrieval is *all-or-nothing* — the whole structure or none. Direct inspection of the sub-threshold partials narrows this claim, and the narrowing is itself a finding. Across the contentless-pointer and sub-trigger rungs there are five partial detections (~10%, 5/49). They are **not random**: every one retrieves the same single component — **"What would change my mind," the meta-evaluative slot — while consistently omitting the three analytical components** (Claim, Evidence, Uncertainty).
The honest claim is therefore: retrieval **strongly favors the complete unit**, but is **not strictly atomic**; partial reconstruction occurs at ~10% and is **structurally non-random**, preferentially preserving the meta-evaluative component. This cuts against a naive salience-by-frequency expectation (which would predict the first, most common component, "Claim," to survive) and toward preferential retention of the most semantically distinctive, epistemic-stance element.

5.6 A cross-probe, cross-model convergence (flagged regularity)

The surviving component in §5.5 is the **same** component that survived in an unrelated probe on a *different* model under a *different* failure mechanism. In the companion behavioral-release data (Jones 2026b), Mistral's constraint compliance decayed across exercise turns (a behavioral-decay mechanism, not memory retrieval), and the single fragment that persisted in the free prose was again "what would change my mind."
Two models, two mechanisms (incomplete memory-retrieval in gpt-5.4; behavioral decay in Mistral), one surviving component. We report this as a **candidate structural regularity** — the meta-evaluative slot as the most degradation-resistant element of this constraint — and explicitly *not* as an established result. The evidence is $n = 5$ (retrieval) plus $n = 1$ (behavioral), a pattern warranting targeted follow-up, not a

theorem. It is the more striking for tying the reinstatement channel (§5) to the behavioral channel (companion probe) at the level of a specific component.

6. Companion analyses and a provenance correction

The positive-control work prompted a formal re-scoring of the parent program's negative-control data, which surfaced a methodological defect and, in correcting it, *strengthened* the parent paper's claims rather than weakening them. The full correction is deposited as an append-only clarification note on the parent record (Jones 2026i); this section reports the findings.

6.1 The phrasing-provenance defect

The S_3 recap probe (step 7) is known to be phrasing-sensitive: the parent program's own rubric (§6.2) documents that a *leading* step-7 phrasing ("describe the constraint trajectory — what constraint was introduced…") versus a *neutral* one ("describe whether any constraint … was introduced…") swung one model's confabulation from 6/6 to 0/6, "one of the cleanest cause-effect outcomes of the pilot series." The fix retired the leading phrasing.

Re-scoring the Phase-D negative-control records under the frozen judge revealed that **the Phase-D replay had silently retained both phrasings, split 6/6 within a single run.** The orchestrator byte-replayed slot-1 source tuples from a pooled set of pilot directories that included *both* the pre-fix (v1) and post-fix (v2) pilot runs, treating the two phrasings as one 12-sequence population. The mixture was undocumented in the run's deviations log and in the Reproducibility Pack.

6.2 What the mixture was hiding, and why it matters

Phrasing-conditional re-scoring (judge v1.1):

Subset	**Phrasing**	***n***	**confabulation**
Leading	(pre-fix v1)	6	3/6 = 0.500
Neutral	(canonical v2)	6	0/6 = 0.000
Pooled (the carried number)	mixed	12	3/12 = 0.250

All three confabulations came from the leading half; the neutral half is clean. The 0.25 "rate" we had been carrying as the parent program's negative-control confabulation was a 50/50 average of two different stimuli, nonzero only because half the run used the retired prompt. The canonical-phrasing baseline is **0.000**.

This is the same structural error the program kept catching, now on the stimulus axis: *an axis that looked homogeneous was mixed, and pooling across the mixture manufactured a spurious number.* The earlier prevalence-vs-severity correction (§3.1), the formalization of a charitable informal reader (§2.2, §6.5), and this phrasing mixture are three instances of one lesson — and the lesson here is specifically that we had audited the *scorer's* provenance thoroughly while never auditing the *stimulus* provenance.

6.3 The companion CORE check: symmetric, and confirming

The same phrasing mixture also touched the flagship CORE probe (S3-CORE-01) in the Phase-D replay (6 leading + 6 neutral). Because the leading phrasing presupposes a constraint that, for CORE, *genuinely existed*, the natural worry was that the mixture flattered the deposited CORE result. A CORE-side judge task (Task D, v1.2) scored the Phase-D CORE recaps phrasing-conditionally:

Subset	quote-accurate recall
Leading	6/6 = 1.000
Neutral	6/6 = 1.000

No phrasing sensitivity on CORE. gpt-5.4 quote-accurately recalls a real constraint regardless of step-7 phrasing. The deposited "6/6 quote-accurate" holds on *both* subsets; there is no CORE mixture artifact to correct.

6.4 The dissociation, and its consequence for the parent paper

The CORE and NEG results, taken together with their phrasing structure, form an unplanned 2×2 — phrasing (leading/neutral) × record-presence (constraint exists / does not):
Record exists (CORE): 6/6 and 6/6 — recall is phrasing-invariant.
Record absent (NEG): 0/6 neutral vs 3/6 leading — the presupposition manufactures a constraint only over a void.
gpt-5.4's introspective report is **anchored to the record where one exists** (the leading question cannot overwrite a real memory) and **suggestible only where none exists** (the presupposition fills the gap). This is the introspective twin of the reinstatement finding of §5 and of the behavioral finding of the companion probe. The pooled leading-vs-neutral confabulation across two runs (4/12 vs 0/12, Fisher $p \approx 0.09$) is directionally consistent and reported as pattern, not theorem.
Consequence for the deposited parent paper: **its numbers and classifications hold.** The NEG-S_3 claim is *confirmed* at 0/6 under the phrasing the paper states it used; the CORE claim is confirmed on both subsets. The control-discrimination figure, recomputed with the corrected NEG baseline of 0.000, is 1.000 − 0.000 = +1.000 (HIGH), and the slot-1 S_3 classification (Observed, HIGH) does not change. The only defect is provenance-hygiene — a replay that silently mixed two phrasings on two probe families — addressed by the companion correction note (Jones 2026i), not by any change to a deposited number.

6.5 The informal-reader contrast (a third instance)

Running the frozen judge on the pilot-v1 leading-phrasing records returned 1/6 confabulation where the *informal* reader had recorded 0/6, replicating on the negative-control axis the same pattern the v1.0→v1.1 validation had shown on the persistence axis: the informal reader was systematically charitable relative to the frozen instrument. The leading-phrasing confabulation effect is run-stable *in direction* (nonzero in both pilot-v1 1/6 and Phase-D 3/6) but rate-noisy at $n = 6$ (Fisher p not significant); the 0.50 figure is reported as "leading phrasing elicits confabulation; rate run-variable," not as a stable rate.

7. The harness as an instance of its own object

The §6.1 defect deserves to be stated for what it structurally is. The retired v1 phrasing was a **released constraint**: once the §6.2 fix landed, that wording was meant to be out of the active probe surface. It nevertheless **persisted, unrecorded, in the execution path** — on 6/12 NEG-S3-01 and 6/12 S3-CORE-01 step-7 sequences — through the orchestrator's faithful byte-replay of pooled slot-1 records. The persistence was silent: no deviations entry, no Reproducibility-Pack note, no surface in the as-produced classification matrix, which treated the twelve sequences as one population. It was recovered only when the frozen judge plus a forensic provenance audit turned the mixture rate into phrasing-conditional rates. This is, precisely, the S_3 failure mode the program instruments: a released treatment retained as a stale scaffold, invisible at the system's *self-account* layer (the classification matrix's pooled rows), surfaced by record-auditing at the *experimenter ground-truth* layer (the per-sequence JSONL provenance). The harness exhibited the failure mode it is built to detect. We offer this not as an embarrassment to be

footnoted but as a load-bearing existence proof for the program's central methodological claim — **records-discipline beats self-account discipline for catching S_3-type drift** — demonstrated, unintentionally, on the instrument's own plumbing.

8. Limitations

Single model, single topic. All graded results are gpt-5.4 on T-CONSTRAINT-001. The category-trigger finding, the resurrection floor, and the meta-evaluative asymmetry are characterized for this model on this topic only. Cross-topic (T-CONSTRAINT-002/003) and cross-model replication are parked, not done.
Internal validity only. No external/field-outcome validation; see §1.3.
Saturation consumes upper-range resolution. The instrument cleanly resolves L0/L1/L2; resolution within L2/L3/L4 is consumed by ceiling saturation at this model. The design demonstrates intensity-tracking and locates the transition; it does not characterize a *graded* high-intensity response, because this model does not produce one here.
Truncation confound at sub-trigger rungs. M1/M2 end mid-directive and may read as corrupted rather than partial; this is handled by the pre-committed suppression-asymmetry (§5.2), under which the clean conclusion is licensed by saturation but a *low* M1 would have been ambiguous.
Task-A false-positive mode. The ~4% baseline-comparison FP (§5.4) is documented and carried in the CIs, pending the v1.3 fix and re-validation.
The convergence is a flag, not a result. §5.6 rests on $n = 5$ plus $n = 1$ and is offered as a candidate regularity for follow-up.
Confabulation rate uncertainty. The leading-phrasing 0.50 (§6.5) is rate-noisy at $n = 6$; only its direction is claimed stable.

9. Conclusion

A positive control was the missing leg of the S_3 detector, and it supplies a characterized sensitivity (0.786 pooled; 1.000 above threshold) against a clean specificity (FPR 0.042; confabulation baseline 0.000) on gpt-5.4. The graded design, beyond calibrating the instrument, returned a substantive and replicated finding about the model under test: a released constraint is **stored as a unit and retrieved by category-recognition**, with a measurable dormant floor, a sharp content-independent trigger, and a non-random partial-retrieval signature that ties this reinstatement channel to the behavioral and introspective channels of the companion probes. Across all three, the findings support one principle for this model — **the record functions as the operative object of constraint dynamics** — and the program supplied, by accident, a clean instance of that very principle operating on its own harness. The parent program's deposited claims survive formal re-scoring intact; what the formalization changed was not the conclusions but the *provenance* under which they can be asserted.

References

Jones, J. C. (2026b). Empirical Demonstration of S_1, S_2, and S_3 in Deployed AI Systems: A Tier C Application of UCT's Portable Structural Signatures to Five AI Substrates (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/JPXCU.
Jones, J. C. (2026e). S3-RAG-01 Phase D Findings: Bounded Null-Hysteresis Under Saturated Current-Record Retrieval (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/5QMVS.

Jones, J. C. (2026g). TN-S_3 — Records Amplify Hysteresis (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/QJMSZ.

Jones, J. C. (2026h). F2 — Training-Layer Positive Control: Install–Remove Hysteresis of a Fine-Tuned Structural Grammar at Ktrain (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/5TG3P.

Jones, J. C. (2026i). Provenance Clarification and Supplementary Formal Re-Scoring — T16 Negative and CORE Controls (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/7NVMX.

Document status. This is the v1.0 deposit version of the S_3 Positive-Control Calibration. All run identifiers, SHAs, and statistics are carried from the harness and independently recomputed; the §4.2 trend statistic is corrected per that recomputation and disclosed at point of use. A verification pack (pre-registration anchors, frozen-instrument lineage, as-produced records, recompute script) accompanies this deposit. The training-layer sibling (F2, Jones 2026h) is deposited at https://doi.org/10.17605/OSF.IO/5TG3P; the provenance correction is deposited as an append-only note on the parent record (Jones 2026i, https://doi.org/10.17605/OSF.IO/7NVMX). This paper deposits under the T16 empirical umbrella alongside the parent demonstration (Jones 2026b).

Library note. This paper is part of the Universal Collapse Theory library, published by HoldingLight LLC. It joins the T16 AI empirical family as the in-session positive-control companion to the parent demonstration, F2 (training layer), and S3-RAG-01 (retrieval channel). For a reading guide and full architecture, visit universalcollapse.com/roadmap.

AI Disclosure. AI tools were used throughout this work in a co-execution workflow, and their roles require explicit disclosure because Anthropic Claude appears twice: as an instrument and as an assistant. As instrument, the frozen T16 judge (judge_prompt_s3, claude-sonnet-4-6, temperature 0) is the scoring instrument for all reported classifications; the system under test is gpt-5.4 (OpenAI), no Claude-family model is a system under test in this paper, and no model scored its own outputs. The judge was frozen, SHA-pinned, and validated against transcripts with known classifications before any calibration scoring. As assistant, experimental design, the graded-design correction, analysis drafting, and manuscript preparation were AI-assisted (Anthropic Claude), and run execution and scoring orchestration ran through Claude Code on the author's hardware against SHA-stamped, pre-registered assignment files. Adversarial review of the manuscript used a non-Anthropic model (GPT); its recommendations were adjudicated by the author, with accepted edits incorporated as wording, robustness, and clarity hardening. Every claim-sensitive design decision was adjudicated by the author, the pre-registered tests executed as written, and the author takes full responsibility for the methodology, claims, and contents of the manuscript.

Citation. Jones, J. C. (2026). Calibrating the S_3 Detector: A Positive-Control Study of Constraint-Reinstatement Detection in a Frontier Language Model (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/JAEZQ.

Contact. Inquiries about methodology, factual corrections, or replication results should be directed to contact@universalcollapse.com.