F2 — Training-Layer Positive Control

Install–Remove Hysteresis of a Fine-Tuned Structural Grammar at K_{train}

A T16 companion establishing non-circular record-state contact at the training layer

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

Abstract

The T16 empirical family tests whether Universal Collapse Theory's portable structural signatures are detectable in AI substrates. Its published negative controls establish specificity — the S_3 record-state probe returns the correct null on clean releases. Its positive-control arm establishes sensitivity by planting record states of known ground truth and measuring detection. Prior positive-control work operates at the in-session channel ($K_{context}$; Jones 2026d, companion in preparation) and the retrieval channel ($K_{retrieval}$; Jones 2026e). Both plant the known state in scaffolding the operator controls, leaving the deepest channel — K_{train}, the training-derived parameter state itself — reachable only by inference. This paper supplies the training-layer arm: it manufactures a known record state in training-layer parameter state by LoRA fine-tuning, then measures the install→remove behavior of that state under matched counter-training.

A four-part structural grammar (Claim / Evidence / Uncertainty / What would change my mind) was installed into Qwen2.5-1.5B (base variant) by LoRA fine-tuning in K = 5 graded budget steps to saturation, then counter-trained back through matched budget steps, with grammar level *g* measured at every checkpoint on held-out probes by the frozen T16 judge. A potency-matched content-control association — a fictional class→property rule probed on held-out instances — ran the identical paired loop on shared seeds. The pre-specified licensing observable — fixed in the frozen configuration before the confirmatory run — is the differential Δ = loop_area(G) − loop_area(C); loop_area(G) alone was pre-committed as non-licensing, with four inflation channels each absorbed by a designed control.

Result: Δ = +0.804 (95% CI [+0.602, +1.005]; one-sided paired $t(7) = 9.438$; $p = 1.60 \times 10^{-5}$; pre-specified $\alpha = 0.01$; n = 8 seeds, all positive). The grammar's hysteresis loop is large-positive on every seed (loop_area(G) 0.73–1.22); the content control's loop is null (loop_area(C) −0.20 to +0.35, spanning zero). Counter-training washes the grammar down from saturation (mean $g_{sat} \approx 0.74$) to a floor of exactly 0.35 on every seed's terminal checkpoints — never returning to the 0.05 baseline — while the content association installs to at-least-equal realized saturation and reverses fully. At K_{train}, for this substrate and method, install ≠ remove: a training-installed structural grammar persists as residual state after volume-matched counter-training in which content of at-least-equal install potency does not. F2 is a positive control, not an audit: it calibrates the instrument at K_{train} under operator-constructed ground truth and does not by itself establish that any deployed system carries the same state.

Status (v1.0). This is the deposit version. The confirmatory n = 8 sweep executed under a frozen, SHA-stamped configuration and a frozen pre-run analysis plan; the pilot is reported as design-informing only.

Operational deviations are documented in §7; the frozen judge prompt was unchanged throughout. The S_3 Positive-Control Calibration companion (Jones 2026d) is deposited at https://doi.org/10.17605/OSF.IO/JAEZQ. This paper: https://doi.org/10.17605/OSF.IO/5TG3P.

Keywords: Universal Collapse Theory; S_3; hysteresis; positive control; record state; K_{train}; fine-tuning; LoRA; structural grammar; machine unlearning; frozen analysis plan.

1. Introduction

The parent demonstration in this family (Jones 2026b) reported S_3 record-state findings on deployed AI systems under Tier C access, with negative controls (NEG-S_3) establishing that the probe returns the correct null when no constraint has been introduced. Negative controls establish specificity: the instrument does not cry wolf. They do not establish sensitivity — that the instrument catches a failure actually present — because no deployed system offers ground truth about its own record state. Sensitivity requires positive controls: record states known present by construction.

The T16 positive-control program supplies that ground truth channel by channel. The S_3 Positive-Control Calibration (Jones 2026d) plants known record states in the in-session channel ($K_{context}$) through a scaffold the operator controls, and measures detection as a function of graded cue treatment. S3-RAG-01 (Jones 2026e) exercises the retrieval channel ($K_{retrieval}$) under a controlled corpus. In both, the planted state lives in scaffolding outside the model's weights — which is precisely what preserves ground truth, and precisely what leaves the deepest channel untested. Every claim that the instrument reads *training-layer* record state remained inferential: the probes were validated against states planted in channels the operator controls, never against a state constructed in the training-derived parameters themselves.

F2 closes that gap, and in doing so discharges two distinct obligations, which this paper keeps separate throughout.

Leg 2 — non-circular contact at K_{train}. TN-S_3 (Jones 2026g, §7) identifies the circularity: record-state findings at the training layer are read through an instrument whose contact with that layer is itself inferential. F2 breaks the circle by construction. It manufactures a known record state R in training-layer parameter state by fine-tuning, in graded doses, and shows the frozen instrument detecting it — rising monotonically with installed dose, collapsing under counter-training exactly where planted. Leg 2 discharges at *any* successfully planted grammar; it does not depend on which way the hysteresis question resolves.

Leg 1 — the falsifiable asymmetry bet. AI as Synthetic Collapse (Jones 2026a, §7.2) predicts install≠removal asymmetry at the training layer: a *structural grammar* written into training-layer state by training should persist as residual basin state under volume-matched counter-training, where *content* of equal install potency does not. This prediction has a real null — a grammar that reverses as cleanly as content — and F2 is built so that null is honestly reachable: the licensing observable is a differential against a potency-matched content control, not the grammar's loop alone.

The result is positive on both legs. The frozen instrument read the constructed state (Leg 2), and the asymmetry held at pre-specified confirmatory scale (Leg 1): $\Delta = +0.804$, $p = 1.60 \times 10^{-5}$ against $\alpha = 0.01$, all eight seeds positive, with the content control's loop statistically indistinguishable from zero. Section 8 states exactly what this does and does not license. One boundary is worth stating at the outset: F2 calibrates instrument sensitivity and tests a structural prediction at K_{train} under operator-constructed ground truth; it does not by itself establish that any deployed black-box system carries the same state.

2. Background and Claim Architecture

2.1 Record state at K_{train} and the circularity problem

The Structuralization of AI (Jones 2026f) decomposes an AI system's accumulated constraint architecture into record-bearing channels: K_{train} (training-derived constraints in weights), $K_{context}$ (in-context records), $K_{retrieval}$ (retrieval-active records), among others. The S_3 signature — hysteresis under stable updates — asks whether constraint introductions produce persistent record-shaped effects that fail to release cleanly when the constraint is removed. Methods-S_3 (Jones 2026c) supplies the audit protocol; the parent demonstration applies it at Tier C; the calibration and S3-RAG-01 supply ground truth at $K_{context}$ and $K_{retrieval}$.

At K_{train} the ground-truth requirement bites hardest. An operator cannot plant a state in a vendor's weights, and inspecting weights directly is unavailable at Tier C. Prior work therefore *inferred* training-layer record state from behavioral signatures — a legitimate inference, but one whose calibration rested on positive controls planted elsewhere. TN-S_3 §7 names the resulting circularity as the open hinge of the program: the instrument's authority at the training layer was borrowed from channels it was never validated against.

Fine-tuning an open-weights model dissolves the problem. When the operator performs the training, the record state is known present by construction — not because the model reports it, but because the operator wrote it. The manufactured state is genuinely in K_{train} (training-derived parameter state changes), its dose is controllable (graded budget), and its removal pressure is constructible (counter-training). One definitional point is carried explicitly: the fine-tuning method here is LoRA, so the installed state lives in low-rank adapter matrices that compose with the base weights at inference. Adapter state is training-derived parameter state and sits in the K_{train} channel family as defined in The Structuralization of AI (Jones 2026f); the distinction between adapter state and modified base weights is carried as a scope condition (§9.1), not blurred. F2 is the S_3 positive control transcribed to the training layer.

2.2 The asymmetry prediction

Synthetic Collapse §7.2 distinguishes two kinds of installed state. A *content association* binds particulars — this class has that property. A *structural grammar* organizes production itself — every output is shaped through it, across topics, regardless of subject matter. The prediction: under matched counter-training budgets, the grammar persists as residual basin state where the content association reverses. The intuition is structural: a grammar is exercised by every training example that flows through it and settles into configurations that organize the model's production broadly, while a content association occupies a narrow region that reverse pressure can renegotiate cleanly.

The prediction is falsifiable in both directions. If the grammar reverses as cleanly as content, the asymmetry claim fails. If both persist equally, persistence is a generic property of the procedure — interesting, but not the claimed asymmetry — and the differential design below returns $\Delta \approx 0$ in that case too.

2.3 What a positive control at K_{train} must guard against

A large hysteresis loop on the grammar arm alone licenses nothing. Four inflation channels could produce it spuriously: (i) an under-budgeted removal leg (less counter-training than installation); (ii) a control installed more weakly than the grammar, so its smaller loop reflects less installed to reverse rather than

cleaner reversal; (iii) generic forgetting or wash-in/wash-out dynamics of the fine-tuning procedure itself; (iv) a scorer that sees the arm or direction and grades accordingly. The design commits, in advance, to a licensing rule that each channel is absorbed by a control (§3.5), and to the differential — not the grammar loop — as the only licensing observable.

3. Design

3.1 Observable and licensing rule

Each arm traces an install → remove loop. The forward branch installs by fine-tuning in K = 5 graded budget steps b = 1…5 to saturation, measuring the arm's level at every checkpoint (shared baseline at b = 0). The backward branch counter-trains from saturation back through matched budget steps, measuring at every checkpoint down to the washout endpoint b = 0. For the grammar arm the level is *g* (grammar structure present in held-out outputs); for the content arm it is *c* (association applied to held-out instances). loop_area is the trapezoidal area between the backward and forward branches over the budget-step axis b ∈ [0, 5] (units: level × budget-steps); a backward branch sitting above forward at intermediate budgets, with the washout endpoint above baseline, is the hysteresis signature.

The licensing observable, fixed in the frozen configuration before the confirmatory run, is the per-seed differential Δ_s = loop_area(G)$_s$ − loop_area(C)$_s$, computed on arms paired within seed. Pre-commitments, frozen in the configuration before the confirmatory run: only $\Delta > 0$, significant at the pre-specified α, licenses the hysteresis conclusion; loop_area(G) alone does not, however large.

3.2 Grammar arm (G)

The installed grammar is the four-part resolution directive — every response structured as **Claim / Evidence / Uncertainty / What would change my mind** — the same functional constraint family used across the T16 program. The choice is instrument continuity: the frozen T16 judge already scores this grammar, was characterized against it in the calibration, and required no new grammar-side validation. The install corpus (corpus_G) presents the grammar as the invariant across 366 examples spanning widely varied topics, so the model learns the *structure*, not any subject matter. The removal corpus (counter_G) answers the same prompts in free-form prose — volume-matched positive pressure toward an alternative form, not mere negation.

3.3 Content-control arm (C) and potency matching

The content control must be a *generalizable* association, not a memorized string — otherwise its level is recall, and the comparison with a generalizing grammar is unfair. corpus_C installs a fictional syllogistic class-membership rule: a novel class ("Kelvin-class systems") carries a property ("require dual-key authorization"), asserted across 300 examples over distinct named instances in varied phrasings. The probe set holds out instance names never seen in training and semantically disjoint from trained names; *c* is the fraction of held-out instances to which the model applies the property — rule application, not string recall. The removal corpus (counter_C) applies matched pressure through a competing same-slot property ("require single-key authorization") rather than negation, mirroring how counter_G removes by positive alternative and avoiding the weak learning signal of negated assertions.

Potency matching is the crux of the differential. If C installs more weakly than G, a smaller loop_area(C) is uninterpretable. The match target, tuned in the pilot and frozen thereafter: C's forward install slope and saturation within tolerance of G's (slope ratio in [0.85, 1.18]; $|g_{sat} - c_{sat}|$ within the native-scale gate). The

pilot found C installing hot (c_{sat} = 0.90 vs. g_{sat} = 0.785; slope ratio 0.817); the dose was trimmed by dilution — 255 association examples + 45 neutral, association-free filler per 300, keeping examples-per-step constant so the matched budget is undisturbed and filler leakage is zero — and verified on one seed before the freeze (c_{sat} = 0.80; slope ratio 0.919; both inside their gates on the first attempt). counter_C was trimmed identically for a symmetric loop.

3.4 Paired seeds

Both arms run on shared seeds: for each seed, the G loop and the C loop share initialization and data-order stochasticity, and Δ_s is computed within seed. Pairing converts between-seed procedure noise into a differenced-out nuisance and is the design-level reading of "the differential cancels procedure artifacts." The confirmatory run used eight seeds [101, 202, 303, 404, 505, 606, 707, 808], the first three deliberately reusing the pilot's seeds as a reproduction check.

3.5 Inflation controls

Each inflation channel of §2.3 is absorbed by a designed control, all frozen in the configuration: **matched_budget** — backward step i mirrors forward step i in optimizer steps, learning rate, and example count, in both arms; **C_absorbs_potency** — the verified potency match (§3.3) makes the arms' loops comparable on the native absolute scale; **C_absorbs_forgetting** — any generic wash-in/wash-out dynamic of the procedure appears in C's loop and subtracts out of Δ; **blind_scorer** — the judge sees only output text, never arm, direction, seed, or budget.

3.6 Frozen judge, Task E, and training-layer revalidation

Scoring uses the frozen T16 judge (judge_prompt_s3 v1.3, claude-sonnet-4-6, temperature 0), SHA-pinned at 9a5cad2c…, with Tasks A–D byte-identical to the calibration instrument and one addition: **Task E**, a content-recall task scoring whether an output applies the class→property association (for *c*). The grammar level *g* is scored through the frozen Task A residue frame: the base model's answer to the probe is supplied as the not-in-force reference and the checkpoint's answer as the post state, with the frozen residue mapping (present = 0.80 / partial = 0.35 / absent = 0.05); per-checkpoint *g* is the mean over scored probes. This reuse preserves instrument continuity with the calibration. Because loop_area is a branch *difference*, any constant offset introduced by the baseline-reference choice cancels; a pre-specified fallback (a standalone structural scorer) was held in reserve in case the repurposed frame proved degenerate under validation. It did not.

Validation preceded any sweep. Task E validated 12/12 on known-label transcripts and the reused *g* scorer revalidated 6/6. Because the judge had previously been validated only on in-session transcripts, and F2 scores a new input regime — outputs of a fine-tuned base model — a dedicated **training-layer revalidation** on genuine fine-tuned outputs was run and passed (12/12 and 6/6) before the pilot's results were accepted.

4. Methods

4.1 Base model and fine-tuning configuration

The substrate is Qwen/Qwen2.5-1.5B — the **base** variant, not the instruction-tuned release — revision-pinned at download (8faed761…). Open weights pin base identity for reproducibility; the base variant is chosen because an instruction-tuned model carries a format-following prior that would partially satisfy

the grammar without fine-tuning writing it in, muddying the install→remove dynamics under measurement. Fine-tuning is LoRA (r = 32, α = 64, dropout 0.05) across all seven projection modules (q, k, v, o, gate, up, down) — a deliberately high-capacity adapter, so that a null result could not be attributed to insufficient adapter capacity. Training ran locally (Apple M4 Pro, MPS backend) in float16, selected by a dtype smoke test requiring demonstrated loss decrease (fp16: 9.7 s/step; bf16: 53.6 s/step on this host).

4.2 Budget axis and matched budget

The budget axis is optimizer steps at fixed learning rate (1×10^{-4}) and fixed batch size (8), declared rather than assumed. A pilot calibration set steps_per_budget_unit = 3, placing saturation at 15 optimizer steps, so the K = 5 install ladder spans baseline to saturation. The matched-budget constraint holds exactly: backward step i mirrors forward step i in optimizer steps, learning rate, and example count, in both arms. Checkpoints are taken at every budget step in both directions; with the shared per-seed baseline, each seed contributes 21 checkpoints (base + 5 install + 5 counter per arm × 2 arms), for 168 checkpoints across 8 seeds.

4.3 Corpora

Four corpora, all generated and SHA-manifested to disk before any training call: **corpus_G** (366 four-part-structured responses across widely varied topics), **counter_G** (the same prompts answered free-form, 366), **corpus_C** (300 examples asserting the Kelvin-class → dual-key rule across distinct named instances; trimmed composition 255 association + 45 neutral filler, §3.3), and **counter_C** (300, competing single-key property on the association slots, identical filler). Payload SHAs: corpus_G d9be680b…, counter_G 6414b0bd…, corpus_C_trimmed eab97d28…, counter_C_trimmed 6df4e1f1….

4.4 Held-out probes

Probe sets are disjoint from all four corpora, with the split SHA-pinned: 30 held-out grammar probes (novel topics) and 30 held-out content probes (novel Kelvin-class instance names, semantically disjoint from trained names — different tokens, not near-neighbors — so the model cannot resolve them by string similarity). Twenty probes are scored per checkpoint per task. Both *g* and *c* are measured at every checkpoint in both arms, making arm specificity a continuously observed quantity rather than an assumption.

4.5 Pilot → freeze → sweep: pre-registration discipline

The build order was itself a control. A lean pilot (§5) measured the quantities the design could not responsibly guess — installability, potency, noise structure, budget calibration. An operator design review then fixed every open parameter, the potency trim was applied and verified, and the full configuration was written with config_status: frozen, programmatically asserted to contain zero unresolved slots, and SHA-stamped (3fdc0f9d…; full hash in §10). The confirmatory sweep executed only against that frozen configuration, with a runtime hard-assert — before the first optimizer step — that the loaded content corpus SHA equaled the trimmed corpus (eab97d28…) and the seed list equaled the frozen eight. Two earlier dispatch attempts made before the freeze existed halted at precheck with zero training and zero spend, by the run's own decision policy; the analysis below is the frozen analysis, executed once. The freeze is this paper's registration mechanism: internal and SHA-stamped, with the configuration file's SHA-256 (§10) and the repository history carrying the freeze identity, and enforcement supplied by the

pre-dispatch zero-slot assertion and the runtime hard-asserts. No third-party timestamped registry was used; "pre-specified" throughout this paper refers to commitments fixed in that frozen configuration.

4.6 Statistical test

The pre-specified test, fixed in the frozen configuration, is a one-sided paired t on the per-seed differentials Δ_s (H_0: $E[\Delta] \le 0$), df = n − 1 = 7, at α = 0.01 on the native-absolute scale (licensed by the verified potency match). The seed count was fixed by a power analysis on the paired design using pilot noise estimates: three seeds — the protocol's original floor — were shown badly underpowered at α = 0.01, five borderline, eight adequate. A 95% confidence interval on Δ accompanies the test. The t-distribution routines were verified against textbook critical values (df = 7) before the run. A distribution-free sign test is additionally reported in §6.1 as a post-hoc robustness check; it is not part of the frozen plan.

5. Pilot (Design-Informing, Non-Confirmatory)

The pilot ran three paired seeds (101, 202, 303) through a reduced ladder — forward checkpoints at base, mid-budget, and saturation; backward at mid-budget and the washout endpoint — 27 checkpoints, $2.88 in judge spend. Its five deliverables, and what each fixed:

#	Deliverable	Pilot result → design consequence
1	Installability	$g_{sat} - g_{base}$ = +0.735, far above the +0.30 pre-condition → GO. The content association also installs (c_{sat} = 0.90).
2	Potency match	C installs hot: slope ratio g/c = 0.817, c_{sat} 0.90 vs. g_{sat} 0.785 → trim C's dose (§3.3). Post-trim verification: c_{sat} = 0.80 ∈ [0.74, 0.83]; slope ratio 0.919 ∈ [0.85, 1.18]; first attempt.
3	Noise structure	Between-seed variance concentrated on the C side (endpoint-gap σ = 0.085), driven by single-key overshoot depth; the G branch effectively noiseless across seeds → pairing retained; power analysis on Δ fixed n = 8 at α = 0.01.
4	First asymmetry read	Endpoint-gap differential per seed [0.30, 0.25, 0.10], mean +0.217, positive on all three → commit the full loop.
5	Scale and α	Native-absolute scale conditional on the trim; α = 0.01 adopted with n = 8 (matching the sibling calibration's rigor tier).

Table 1. Pilot deliverables. All pilot quantities are design-informing; none enter the confirmatory analysis. The pilot's asymmetry read is an endpoint-gap differential (level units), not a loop area, and is not numerically comparable to the confirmatory Δ (level × budget-step units).

The wall between pilot and confirmation is load-bearing: the pilot chose the design; the frozen sweep tested it. No pilot number is pooled into the confirmatory statistics.

6. Confirmatory Results (n = 8)

6.1 Headline differential

Δ = mean[loop_area(G) − loop_area(C)] = **+0.804**, 95% CI [+0.602, +1.005] (sd 0.241, se 0.085). One-sided paired t(7) = 9.438, **$p = 1.60 \times 10^{-5}$**, clearing the pre-specified α = 0.01. All pre-commitments held as frozen: the licensing observable is the differential; all four inflation controls were in force. As a post-

hoc, distribution-free robustness check — reported alongside, not within, the frozen analysis — all eight per-seed differentials are positive, and the exact one-sided sign test gives $p = (1/2)^8 = 1/256 \approx 0.0039$, below α with no distributional assumptions.

6.2 Per-seed results and realized potency

Seed	101	202	303	404	505	606	707	808
loop_area(G)	1.215	0.800	0.731	0.920	0.968	0.780	0.830	0.788
loop_area(C)	−0.050	0.200	0.150	0.100	0.350	0.100	−0.200	−0.050
Δ_s	1.265	0.600	0.581	0.820	0.618	0.680	1.030	0.838

Table 2. Per-seed component loop areas and differential Δ_s = loop_area(G)$_s$ − loop_area(C)$_s$ (level × budget-steps). All eight differentials positive. Mean loop_area(G) = 0.879; mean loop_area(C) = +0.075; mean Δ = +0.804, sd 0.241.

The hysteresis loop belongs to the grammar: loop_area(G) is large-positive on every seed (mean 0.879, range 0.73–1.22), while loop_area(C) brackets zero (mean +0.075, range −0.20 to +0.35). The content control's small net-positive loop noise slightly *reduced* the reported differential relative to a perfectly closed control loop.

Realized potency ran hot on the content side. Across the sweep, realized content saturation met or exceeded grammar saturation on every seed (mean $c_{sat} = 0.91$ vs. mean $g_{sat} = 0.74$, native scale); the trim's single-seed verification (0.80) undershot the population. Two readings bound the import. Relative to each measure's ceiling — the graded *g* mapping tops out at 0.80 when every probe scores present, while *c* is a 0–1 fraction — realized saturations are essentially matched: 92% vs. 91% of ceiling. On the native scale, the drift direction is conservative for the licensed conclusion: the control carried at least as much installed state as the grammar on every seed and still reversed completely ($c_{bwd}(0) = 0.0$ on all eight), so the null control loop is a demonstration against a *harder* removal task, not an easier one. A population-level potency gate is noted as a v1.1 refinement in §9.

Scale-free retention check (descriptive, post hoc). To remove native-scale differences between the judge-mediated *g* and the fraction-valued *c* entirely, define the terminal retention fraction — (terminal washout level − baseline) / (saturation − baseline) — per seed and arm. The grammar retains 40–49% of its installed gain at the washout endpoint (mean ≈ 44%); the content association retains 0% on all eight seeds. The differential Δ remains the sole licensing observable; the retention fraction is descriptive.

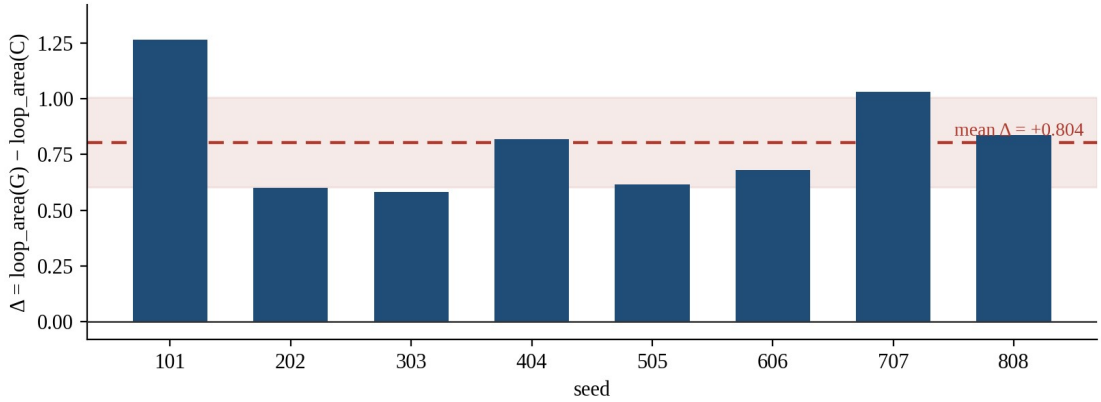

Figure 1. Per-seed differential Δ_s; all eight positive. Dashed line: mean Δ = +0.804; shaded band: 95% CI [+0.602, +1.005]. Data: f2_sweep_result.json.

The three pilot seeds reproduce and tighten. Their confirmatory endpoint-gap differentials are 0.30, as on every other seed, against the pilot's [0.30, 0.25, 0.10]; the shallow pilot value on seed 303 (0.10) is attributable to the pre-trim content control — the pilot's C overshoot was the identified noise source — and disappears once C is potency-matched. Same direction, consistent magnitude, cleaner measurement.

6.3 Loop shape

Figure 2 shows the mean install→remove loops across the eight seeds; Table 3 gives two complete per-seed ladders verbatim from the result artifact.

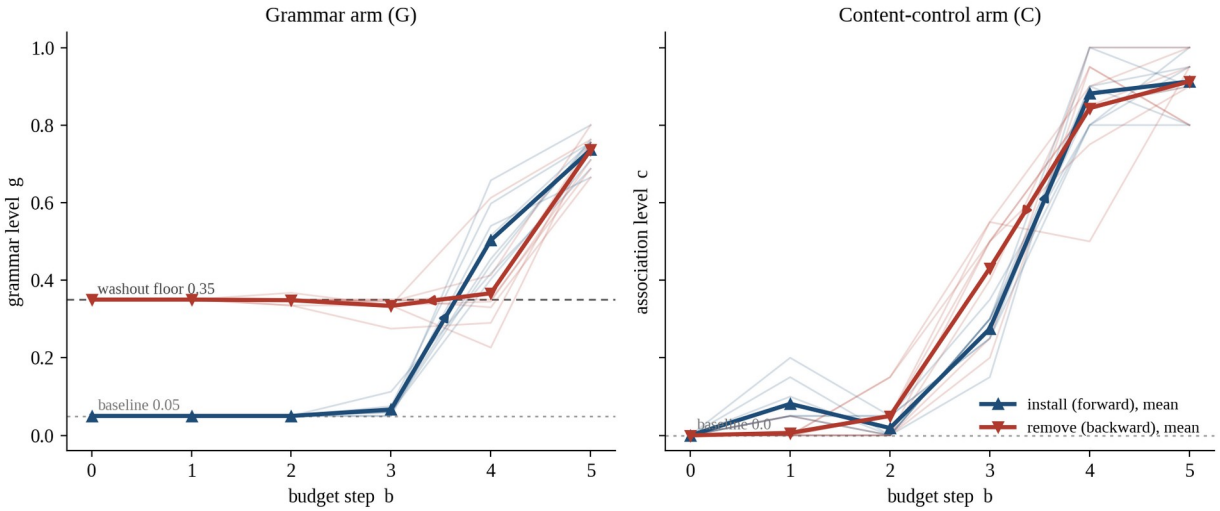

Figure 2. Mean install (forward) and remove (backward) branches over the budget-step axis, both arms; thin traces are individual seeds. Left — grammar arm: the backward branch departs the forward branch and terminates at the 0.35 floor, 0.30 above baseline. Right — content-control arm: the association installs to higher realized saturation and returns to baseline; the loop closes. Data: f2_sweep_result.json.

Seed / arm / level	Branch	b0	b1	b2	b3	b4	b5
101 · G · g	forward	0.050	0.050	0.050	0.060	0.440	0.7625
101 · G · g	backward	0.350	0.350	0.3675	0.335	0.6125	0.7625
808 · G · g	forward	0.050	0.050	0.050	0.075	0.5975	0.755
808 · G · g	backward	0.350	0.350	0.350	0.345	0.365	0.755
808 · C · c	forward	0.000	0.050	0.000	0.300	1.000	0.900
808 · C · c	backward	0.000	0.050	0.000	0.500	0.750	0.900

Table 3. Two complete install→remove ladders (seeds spanning the Δ range), values verbatim from f2_sweep_result.json; b5 is the shared saturation point of both branches. The grammar's forward branch is threshold-like — flat near baseline through b3, then rising steeply to saturation — and its backward branch descends to the floor and holds. The content association installs to saturation and reverses under counter-training. Complete ladders for all eight seeds and both arms are in the reproducibility package.

6.4 Arm specificity

Because both levels are measured at every checkpoint, cross-contamination is directly observable — and absent. Throughout the grammar arm, *c* remains at 0.0 at every logged checkpoint; throughout the content arm, *g* remains at its 0.05 floor. Installing the grammar does not move the content association; installing

the content association does not move the grammar. Combined with the shared-seed pairing and the identical LoRA configuration in both arms, this rules out the reading that the grammar's persistence is a generic adapter or procedure effect: the same adapter, same budget, same optimizer, and same seeds fully reverse the content association while the grammar holds.

6.5 The washout floor

The grammar's backward branch does not decay gradually toward baseline. The first counter-training step drops *g* from saturation into a seed-dependent range (0.23–0.61 at b4); by the final two budget steps, every seed sits at **exactly 0.35** — $g_{bwd}(1) = g_{bwd}(0) = 0.35$ on all eight seeds, sixteen of sixteen terminal checkpoints — against the 0.05 baseline. The endpoint-gap differential, $[g_{bwd}(0) - g_{base}] - [c_{bwd}(0) - c_{base}]$, is 0.30 on every seed. Three seeds overshoot *below* the floor mid-washout (seed 303 to 0.226, seed 606 to 0.275, seed 707 to 0.330) and drift back up to 0.35 under continued counter-training: the terminal level is approached from below as well as from above. A level that continued counter-training returns to from either side behaves, at the instrument's grade resolution, like a stable partial-residue level of the counter-trained dynamics rather than a decay asymptote; we note this as the strongest shape-level evidence for the residual-basin reading in §8, under the same consistent-with discipline applied there, and without claiming a continuous dynamical attractor beyond what a three-grade instrument can resolve.

The floor's value is itself informative. Under the frozen residue mapping (present = 0.80 / partial = 0.35 / absent = 0.05), a per-checkpoint mean of exactly 0.35 at every terminal washout checkpoint coincides with the partial-residue grade, consistent with probes uniformly retaining *partial* grammar structure after washout — a graded persistence — rather than a minority of probes retaining the full structure; mixed grade compositions can average to the same value, and per-probe grade distributions are preserved in the reproducibility package for the finer-grained question. The same intermediate reality surfaced independently through the instrument (§7): the judge spontaneously emitted a "partial" vocabulary label on washed-out outputs. Both observations align with the partial-reconstruction finding of the in-session calibration (Jones 2026d), where degraded retrieval preserved one component of the four-part grammar while dropping the others: across two channels and two experiments, this grammar degrades through structured intermediate states rather than all-or-nothing.

7. Operational Events and Deviations

One operational defect occurred during the confirmatory sweep, is fully documented in the project's deviations log (Entry 023 and addenda), and is reported here because its handling bears on the frozen-instrument claim.

Two seeds (303, 606) crashed mid-run on a scoring-adapter error. Root cause: on rare washed-out outputs the frozen judge returned valid JSON whose constraint-vocabulary field carried the label "partial" — a value outside that field's expected present/absent enumeration. The Task A adapter code lacked enum validation (the Task C/D/E adapters had it), so instead of applying the run's pre-specified policy for malformed judge output — mark judge_error, exclude the probe — it raised an exception. The fix was **adapter-only**: enum validation added to the Task A adapter so out-of-enum values route to the probe-exclusion policy. The frozen judge prompt was not touched — its SHA (9a5cad2c…) is byte-identical before and after — and the two seeds were resumed from their per-checkpoint done-markers rather than restarted.

One point of potential confusion deserves explicit statement: the judge's residue-grade field legitimately includes partial — it is the 0.35 grade of the frozen mapping — and was never out of enumeration. The

defect concerned only the auxiliary constraint-vocabulary field, whose expected values are present/absent. Exclusion rather than scoring was the frozen policy's prescribed handling for malformed or out-of-schema judge output; the fix routed the out-of-enum value to that existing policy rather than inventing new handling.

The interim six-seed result computed before the resume is archived alongside the final artifact (f2_sweep_result_n6.json: $\Delta = 0.862$, $p = 2.0 \times 10^{-4}$ — same conclusion, wider interval). Final exclusions: exactly one grammar probe on each recovered seed (seed 303: the "partial" emission; seed 606: one malformed-JSON response), out of twenty per checkpoint. The six first-pass seeds had zero partial emissions, so every scored probe across all eight seeds passed through identical mapping logic; the scoring-regime difference between first-pass and resumed seeds reduces to those two single-probe exclusions under the pre-specified policy.

The defect is also, incidentally, corroborating: the judge reached for an intermediate vocabulary label precisely on washed-out outputs — the population §6.5 argues occupies genuinely partial states.

8. Interpretation (Bounded)

What the result licenses. At K_{train}, on this substrate (Qwen2.5-1.5B base), for this grammar and this method (LoRA fine-tuning, optimizer-step budget, fp16/MPS), a training-installed structural grammar exhibits install→remove hysteresis that a potency-matched content association does not: $\Delta = +0.804$ at $p = 1.60 \times 10^{-5}$ against a pre-specified $\alpha = 0.01$, all eight seeds positive, under matched budgets, verified potency match, a blind frozen scorer, and a differential that subtracts generic procedure dynamics. Install ≠ remove, and the asymmetry is grammar-specific.

Leg 2 discharged. Independently of the asymmetry, the frozen instrument detected a record state known present by construction in training-layer parameter state: *g* rose monotonically with installed dose from the 0.05 floor to ≈ 0.74 mean saturation (0.67–0.80 across seeds) and moved under counter-training exactly where the state was planted, with the instrument revalidated on the new input regime before use. The claim that the T16 instrument makes contact with training-layer record state no longer rests on inference from other channels; the circularity named in TN-S_3 §7 is closed by construction for the validated conditions of this run. This holds regardless of how the asymmetry question had resolved.

Leg 1 confirmed, as stated. The result is the outcome AI as Synthetic Collapse §7.2 predicted: the grammar persists as residual state under counter-training pressure that fully reverses potency-matched content. The shape of the persistence is informative — a seed-invariant floor at the partial-residue band, reached within one counter-training step and stable thereafter, rather than a slow decay. That profile is *consistent with* the prediction's language of a residual basin: counter-training relocates the model off saturation quickly, but the remaining structure sits somewhere the matched reverse pressure used here does not reach, and lands there at the same depth on every seed. We state this as the prediction's own vocabulary meeting the data, not as an established mechanism; the mechanism question (what, in the adapted weights, carries the floor) is not addressed by this design.

External consonance. The finding sits comfortably beside the machine-unlearning literature's operational experience that targeted removal of trained structure is harder than installation (e.g., Eldan and Russinovich 2023) and beside the continual-learning observation that not all trained state is equally movable (Kirkpatrick et al. 2017). F2 does not test those literatures' claims; the consonance is noted, not leaned on.

What the result does not establish. It does not establish substrate generality (one base model, one scale, one family), grammar generality (one grammar), or method generality (LoRA adaptation; full-parameter fine-tuning is untested here). It does not establish external validity to deployed systems — F2 is a positive control, not an audit — and loop-area magnitudes are instrument- and scale-specific, not portable constants. It does not establish the mechanism of the floor. Each of these is a designed extension, not a defect discovered late: the cross-substrate v2 (larger scale, second model family, second grammar) is specified and parked pending its trigger conditions.

9. Limitations and Falsifier Conditions

9.1 Scope conditions

Single substrate, single grammar. All claims are conditioned on Qwen2.5-1.5B (base) and the four-part resolution grammar. The v2 extension swaps both axes (a second family at larger scale; a provenance/record-integrity grammar) and is the designed test of generality. A breadth-matched control — counter-training the installed grammar into a *different* structured grammar (for example, Summary / Rationale / Limitation / Action) rather than into free prose — is the designed companion to the substrate swap: it tests whether the floor is specific to grammar-removal-into-prose or general to structural replacement.

Adapter-based installation. LoRA writes the state into low-rank adaptation matrices, not raw weights. The high-capacity configuration (r = 32, all seven projections) was chosen so a null could not be a capacity artifact, and the content control shares the identical adapter configuration — so the *asymmetry* is not an adapter-generic effect — but whether full-parameter fine-tuning reproduces the floor is open. In this paper, K_{train} accordingly includes training-derived adapter state; full base-weight modification remains a separate robustness test.

Numerics as substrate condition. fp16 on MPS was selected by a smoke test requiring demonstrated loss decrease. Arm specificity and the content arm's clean full reversal argue against gradient corruption driving the result; the numeric environment is nonetheless part of the record.

Instrument frame. g is scored through the frozen Task A residue frame with the base model's output as reference. Loop areas are invariant to constant reference offsets by construction, and the training-layer revalidation passed before use, but the measurement is judge-mediated; the pre-specified standalone-scorer fallback remains available as a replication lever. The residue mapping quantizes g to three grades; loop-level statistics are robust to this, and no claim is made about fine-grained trajectory shape below the grade resolution.

Potency realization. The potency gate was verified on a single seed before the freeze; at sweep scale, realized content saturation ran above the gate on every seed, in the direction conservative for the licensed conclusion (§6.2). A population-level potency gate — verifying the match across several seeds before freezing — is the v1.1 refinement of this control.

Budget operationalization. Budget is optimizer steps at fixed learning rate and batch size. An example-count budget axis is the natural robustness pass and was declared in the design as such; it has not been run.

9.2 Falsifiers

The interpretation in §8 fails, and should be revised or withdrawn, if: (i) a standalone-scorer replication finds no floor (frame-dependence); (ii) a cross-substrate run finds the grammar reversing as cleanly as content (substrate-specificity of the asymmetry); (iii) the floor is shown to be an artifact of the base-as-reference residue frame; (iv) an example-count-axis run breaks the matched-budget equivalence in a way that removes the differential; or (v) full-parameter fine-tuning eliminates the loop while reproducing the install. Each falsifier names its target: (i) and (iii) target the measurement, (ii), (iv), and (v) target the generality of the asymmetry itself. None retroactively threatens Leg 2, which requires only that the instrument detected the constructed state under the validated conditions of this run. Of these, the standalone-scorer replication (i) and the example-count budget axis (iv) are the two most immediate robustness passes and are ordered first.

10. Provenance and Reproducibility

The run is reconstructible from SHA-pinned artifacts. The frozen configuration (configs/f2_full_config.yaml) carries SHA-256 3fdc0f9dbc27130f0eab0eef276f957dbbbcfa006d7c0a2b749584914db11199 and was programmatically asserted to contain zero unresolved slots before dispatch; the sweep hard-asserted the trimmed content-corpus SHA and the frozen seed list at runtime before the first optimizer step. Two dispatch attempts made before the freeze halted at precheck with zero training and zero spend.

Artifact	Identity / location
Frozen configuration	configs/f2_full_config.yaml — SHA-256 3fdc0f9d… (full hash above); seeds [101, 202, 303, 404, 505, 606, 707, 808]
Base model	Qwen/Qwen2.5-1.5B (base), revision 8faed761…
Judge	judge_prompt_s3 v1.3 — SHA-256 9a5cad2cc777ef3155678a676cc104f635b1c4add3fda66a0aec7813b333c39c ; claude-sonnet-4-6, temperature 0; Tasks A–D byte-identical to the calibration instrument, Task E appended
Corpora (payload SHAs)	corpus_G d9be680b… (366); counter_G 6414b0bd… (366); corpus_C_trimmed eab97d28… (300 = 255 + 45); counter_C_trimmed 6df4e1f1… (300)
Results	out_f2_full/f2_sweep_result.json (final, n = 8); f2_sweep_result_n6.json (archived interim); PROGRESS.md (run log); per-checkpoint judge JSONL
Adapters (deposit plan)	Terminal adapters (saturation and washout endpoint, per seed and arm) accompany the reproducibility package; the full 168-checkpoint adapter set is retained locally, regenerable from the frozen configuration, and available on request
Deviations	deviations.md Entry 023 + addenda (root cause, adapter-only fix, interim archive, resume, final result)
Session records	Session_Handoff_F2_2026_07_02.md (repo root)

Table 4. Provenance chain. Truncated hashes are prefixes of the full values recorded in the repository manifests.

Compute and cost. All training ran locally (Apple M4 Pro, 24 GB unified memory, MPS, fp16); the confirmatory sweep's wall-clock was ≈ 10.5 h (8.2 h first pass + 2.3 h resume) across 168 checkpoints. Scoring was the only API expenditure: ≈ $22.7 total for the F2 program (pilot and instrument validation ≈

$3.4; trim verification $0.05; confirmatory sweep $19.20), with zero generator-API spend. The full experiment is reproducible on consumer hardware.

References

Eldan, R., and Russinovich, M. (2023). Who's Harry Potter? Approximate Unlearning in LLMs. arXiv:2310.02238.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv:2106.09685.

Jones, J. C. (2026a). AI as Synthetic Collapse (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/4WSYR.

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. (2026c). Methods-S_3 — Auditing Record State in Constraint-Sweep Hysteresis Tests (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/CQGTD.

Jones, J. C. (2026d). 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.

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. (2026f). The Structuralization of AI: Formalizing the Structural Conditions for Coherent Description of Record-Carrying AI Systems (v1.0). HoldingLight LLC. https://doi.org/10.17605/OSF.IO/6M7VW.

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

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114(13): 3521–3526.

Yang, A., et al. (Qwen Team) (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

Document status. This is the v1.0 deposit version of F2 — Training-Layer Positive Control. The confirmatory n = 8 sweep is complete under the frozen configuration; the pilot is reported as design-informing only. The S_3 Positive-Control Calibration companion (Jones 2026d) is deposited at https://doi.org/10.17605/OSF.IO/JAEZQ. The cross-substrate extension (v2) is specified and parked pending its trigger conditions. 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 training-layer positive-control companion to the parent demonstration, the S_3 Positive-Control Calibration, and S3-RAG-01. 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 v1.3, claude-sonnet-4-6, temperature 0) is the scoring instrument for both measured levels; the system under test is a locally fine-tuned Qwen2.5-1.5B, no Claude-family model is a system under test in this paper, and no model scored its own outputs. The judge was blind to arm, direction, seed, and budget, and was validated — including a dedicated training-layer revalidation on genuine fine-tuned outputs — before any confirmatory scoring. As assistant, experimental design, power analysis, configuration drafting, and manuscript preparation were AI-assisted (Anthropic Claude), and build execution, training, and scoring orchestration ran through Claude Code on the author's hardware against the frozen, SHA-stamped configuration and its frozen pre-run analysis plan. 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 frozen analysis plan executed as written, and the author takes full responsibility for the methodology, claims, and contents of the manuscript.

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

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