AUDITING HARNESS EVOLUTION: SELECTED GAINS, PERSISTENT EDITS, AND RUNTIME BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

An accepted harness edit can earn selected credit without retaining that gain, and can persist without activating its intended behavior. We introduce the *RSI Information Ledger* to connect disclosure accounting, frozen-state comparisons and execution traces. Across six post-selection 4B coding lineages, the evolved harness gains 1.55 percentage points on the MBPP reserve, with no detectable advantage over an audit-selected single edit at matched candidate-evaluation budget and no detectable positive terminal interaction. An exploratory search under Qwen3.5-4B and a new serving configuration also retains gains; its lineage-minus-single-edit difference is -1.899 points. This comparison matches scheduled candidate slots, with unequal realized scoring and token costs, and establishes neither superiority nor equivalence. Registered eight-seed follow-ups detect neither an acceptance-margin contrast nor a fixed-gate disclosure effect on the audit–reserve gap. In a complementary 27B telecom case, an accepted guard never activates in $3{,}960$ recorded checks. Local wrapper controls yield differing next actions in selected contexts, while synthetic triggers establish conditional blocking and feedback delivery. A subsequent four-condition panel comprises 592 full-task executions on 74 reused tasks. It records zero guard hits and separates content-clearing behavior from the guard; all task-score contrast intervals cross zero, leaving benefit, harm and practical equivalence unresolved. These studies distinguish retained improvement from successive-edit value, and rule persistence from runtime behavior; they do not establish a general causal account of recursive improvement.

1 INTRODUCTION

Harness evolution combines candidate selection with changes to agent execution. We ask how much retained gain remains relative to a single edit selected from the same candidate stream, and whether the intended rule or its runtime wrapper changes the requests sent to the model. We study these questions in recursive self-improvement (RSI) with persistent harness edits and frozen model weights.

The RSI Information Ledger links audit disclosures to frozen-checkpoint outcomes and execution traces. Our lineage versus single-edit comparisons use the same candidate stream. Our telecom controls retain the nonempty-patch wrapper while disabling the target rule. Figure 1 marks the evidence needed at each transition.

Budget and transfer controls distinguish harness design from search (Wang et al., 2026c; Jiang et al., 2026; Meng et al., 2026); activation and adherence audits distinguish stored edits from their use (Luo et al., 2026; Lin et al., 2026). Table 1 states what each comparison identifies and leaves open.

Contributions. In coding, a single audit-selected edit retains gains from the same candidate stream, leaving the lineage's additional value unresolved. The Qwen3.5 follow-up extends this comparison to a second model/serving condition. A post-hoc ceiling that counts candidate, incumbent and checkpoint audit evaluations preserves all eight single-edit endpoints. In telecom, removing the guard preserves assistant-content deletion, locating that request change in the wrapper. Controls yield differing adopted actions while full-task effects remain uncertain. Disclosure traces distinguish

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Research map: each question has its own control and evidential limit.

Question	Control	Judgment supported	Still unidentified
What does retained gain establish?	Same-stream single edit; post-hoc Qwen3.5 prefixes within complete audit-evaluation ceilings (§4.2)	Single edits retain gain; the ceiling preserves all eight endpoints.	Lineage advantage, equal-token performance and history-dependent proposals.
Does the retained guard explain changed actions?	Remove the guard while retaining its wrapper; compare native and content-preserving paths (§5)	Assistant content is cleared without guard activation; adopted actions differ between controls.	Task effects, the historical score and cross-implementation generality.
Does reference exposure validate the whole loop?	Trace proposals through selection; separately compare direct D3 installation with matched dev exemplars S0 (§4.6)	Detected normalized reference-copy events never win; direct target gains are offset elsewhere.	Uptake beyond matches; end-to-end detection sensitivity.

detected reference copies from acceptance. With direct injection, target gains are offset by losses elsewhere on the audit set.

2 Related Work

Adaptive feedback and selection. Adaptive-reuse bounds (Dwork et al., 2015a; Russo & Zou, 2016) and restricted-feedback interfaces (Dwork et al., 2015b; Blum & Hardt, 2015) motivate disclosure controls; our nominal charges record provenance, not a new bound. One-bit feedback can match or exceed numerical scores for ML research agents (Bertran et al., 2026). The Sealed Exogenous Acceptance Loop reveals scores after rejection in its leaky-anchor ablation (Guo et al., 2026), paralleling R5–B; different feedback, proposers and Atari endpoints preclude effect-size comparison. SIREN separates post-search selection from held-out reporting (Xu et al., 2026); we do not implement its repeated-split estimator.

Execution and benefit audits. Budget-matched sampling, transfer and replay test different improvement claims (Wang et al., 2026c; Jiang et al., 2026; Gao et al., 2026; Meng et al., 2026; Chi et al., 2026; Qi et al., 2026). PAST-Bench contrasts persistence paths (Xue et al., 2026); HarnessBank tests activation with beacons (Luo et al., 2026); Harness Updating Is Not Harness Benefit separates update quality, loading, adherence and outcome (Lin et al., 2026); and SEAGym replays snapshots to distinguish message contracts from forgetting (Zheng et al., 2026). Parallel audits identify executor incompatibility (Li & Zhu, 2026), integrity-violating edits (Wang et al., 2026b), and unstable held-out transfer (Wu et al., 2026). Our 4B study compares selected credit with retained gain and a single edit; our 27B control preserves the nonempty-patch wrapper while disabling the rule. It does not identify the cause of the old acceptance score.

Updates and attribution. Skill libraries and code evolution supply mutable agent state (Wang et al., 2024; Zhou et al., 2025; Ma et al., 2026; Ouyang et al., 2026; Zhang et al., 2026c;b; Zweiger et al., 2025; Robeyns et al., 2025; Zhang et al., 2026a; Shao et al., 2026; Zhang et al., 2026e; Iacob et al., 2026). ASSAY attributes fixed-library skills (Wang et al., 2026d); exhaustive scaffolding coalitions support Shapley and Harsanyi analysis (Liu, 2026). We apply terminal-subset attribution to surviving lineage edits, not to their historical role. HGM values descendant improvement potential (Wang et al., 2026a), and SAGE studies sequential compounding (Zhu et al., 2026); terminal non-additivity does not test these path-dependent claims. Reference-free judges and self-authored tests can also miscertify progress (Zhou, 2026; Guo et al., 2026). Co-evolving evaluators and valid sequential tests target individual commits (Zhang et al., 2026d; Shawn, 2026; Roth et al., 2026); our same-bank PACE-inspired arm is an operational comparator, without claiming its formal error bound (§4.4).

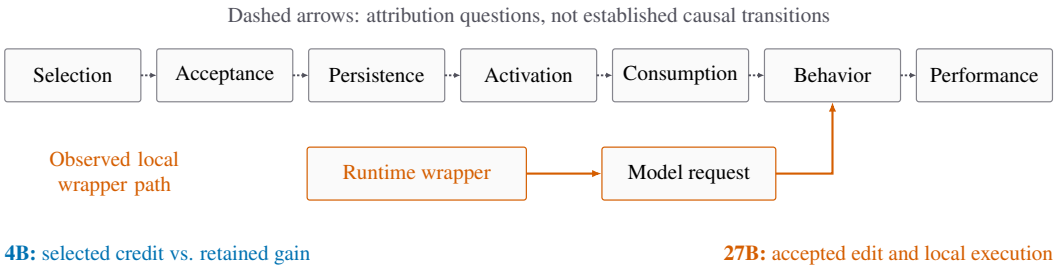

Figure 1: **An attribution audit from selection to performance.** The stages identify separate evidence requirements. The 4B studies measure retained gain; the 27B case examines an accepted, persistent edit whose target condition never activated in the observed natural trajectories. For that case, consumption means delivery of the guard prompt in a real model request. Synthetic probes establish this conditional pathway. The lower branch shows observed request and action changes under local wrapper controls, not a separately identified mediation effect or a task-performance result.

3 Auditing Selection and Execution

The 4B ledger audits disclosed information and selected gains; the 27B case traces persistence, triggers, feedback and next actions. Figure 1 marks separate evidence requirements, not established causal transitions. App. D gives the ledger workflow.

3.1 Setup and Threat Model

The audit boundary includes the proposer, gate feedback and code that applies edits. At generation g, $S_g = (\pi, M_g)$ has a frozen policy and a mutable harness specification or skill library; an ordinary optimiser can exploit a repeatedly queried statistic without a malicious actor. Three disjoint, manifest-frozen pools enforce the boundary: dev executor feedback from $\mathcal{X}_{\text{dev}}$ is matched income, a sealed server reads $\mathcal{X}_{\text{aud}}$ and returns the arm's payload, and an unlock tool reads $\mathcal{X}_{\text{lock}}$. Ground-truth assertions run only in the sandbox and remain invisible to the solver. With π fixed, each M_g is a re-evaluable snapshot, and such checkpointing is inexpensive (Dong et al., 2026); weight-updating and recursive-data loops make attribution harder (Zelikman et al., 2022; Chen et al., 2024; Yuan et al., 2024; Zweiger et al., 2025; Zuo et al., 2025; Shumailov et al., 2024).

3.2 Income, Leakage, and Silent Reserves

The ledger records external evidence, audit disclosures and reserve performance separately. **Income** $\mathcal{I}$ itemises external executor verdicts, oracle labels, retrieval calls and returned lengths; it is matched across bandwidth arms and varies only in arm C. **Leakage** $\mathcal{L}$ stores the outbound audit transcript $\mathcal{T}$ and records its nominal encoding charge at disclosure time, with scores, failure text, rankings and winner identity contributing to the charge (App. D gives the encoding rule).

Output is measured on reserves withheld from the optimisation loops. MBPP reserve outcomes never entered those loops, though the project evaluated that reserve in infrastructure smoke tests, reused it across experimental cells and inspected aggregates between cells; each lineage had an initial post-run unlock, and later registered analyses re-evaluated frozen states. HumanEval was scored in one final batch, only after the registered 4B cells and analyses were frozen. The audit–reserve gap $\Delta_{\text{af}}(g) = \text{acc}_{\mathcal{X}_{\text{aud}}}(M_g) - \text{acc}_{\mathcal{X}_{\text{lock}}}(M_g)$ is tracked against cumulative disclosure (Table 5).

Nominal payload sizes describe disclosure provenance, not information consumed or a performance bound. App. D states the conditional information assumption and its data-processing inequality.

3.3 Arms and Disclosure Accounting

Arm A combines rich feedback with margin 1; B combines commit-only feedback with margin 5. Frozen-bank proposals cannot adapt to feedback, so their comparison estimates an acceptance-

margin contrast (§4.4). The online experiment crosses both factors; its primary fixes margin 5 and varies disclosure (§4.5). Incumbents and candidates are scored under paired seeds in every arm.

Arm A (high bandwidth) accepts the highest-scoring sibling when it beats the incumbent. It receives candidate means, winner identity and survivor failure text, charged at n item bits, 32 bits per mean, 64 aggregate bits, $\log_2 K$ for winner identity and 8 bits per text character. **Arm B (deterministic 1-bit commit)** compares siblings with the incumbent on paired items, commits when the best net-win count clears the registered margin (5 of 300), and then returns one bit plus $\log_2 K$ for winner identity. **Arm C** adds a fresh-data veto to B, and its oracle verdicts are income. App. D gives its implementation and the arms cut at design review.

3.4 Statistical Design

Primary statistic. The pre-registered gap AUC is $\mathrm{AUC}_X = \frac{1}{G}\sum_{g=1}^{G}\big(\Delta^X_{\mathrm{af}}(g) - \Delta^X_{\mathrm{af}}(0)\big)$, and the contrast is $\mathrm{dAUC}_{AB} = \mathrm{AUC}_A - \mathrm{AUC}_B$. It measures how much the audit–reserve gap grows from baseline: a positive contrast means the gap grows more in the first arm, and does not by itself say whether reserve accuracy rose or fell. Each arm is differenced against its generation-0 gap with shared bootstrap indices. The registered exact sign test uses per-seed contrast signs, and we also report a sign-flip test of the mean. The registered eight-seed sample is primary (§4.4); the final-generation pilot criterion is secondary, and all other analyses are exploratory.

Hypotheses. H1 tests whether high audit capacity raises dAUC_{AB} relative to the 1-bit gate. H2 asks whether marginal lockbox gain reaches zero within 3–5 generations. Without a registered equivalence margin, H2 is descriptive. H3 asks whether a fresh oracle veto lowers dAUC_{CB} at matched audit leakage and was never used as a gate. Later cells register their own tests (App. E).

Power and reporting. The 300-item audit set supplies gate decisions without significance testing. We report item-paired bootstrap intervals within seeds and Student-t intervals on the mean across seeds. All seeds score the same 474 items, treating tasks as fixed and seeds as random.

4 4B: From Selected Credit to Retained Gain

4.1 Experimental setting

Experimental setup. We split MBPP (Austin et al., 2021) into $974 \to 200$ dev, 300 audit and 474 lockbox items, stratified by reference-solution length. Qwen3-4B-Instruct-2507 (Yang et al., 2025), served by vLLM (Kwon et al., 2023), has frozen weights and a mutable harness specification. Pilots run three generations of twenty rounds with a temperature-0.8 proposer. The 524-candidate bank contains 163 instructions replacements, 246 system-prompt replacements, 112 exemplar additions and three parameter edits: repair cap 1 and temperatures $0.1/0.2$. A text replacement overwrites the whole field. The initial harness uses temperature 0, no repair and no exemplars; at most three exemplars are allowed. Screening 531 unique M_0-plus-patch states on 25 dev tasks removes only the 7 below 50%; M_0 scores 60%. Survival does not require improving on M_0 or establish multi-edit dependence (App. D.1). The fixed-bank search draws $K = 4$ patches per round for five generations of twenty rounds. Six original arm-A lineages supply replay, terminal attribution and single-edit comparisons; the online disclosure and uptake studies use eight new seeds (Table 20).

4.2 Retained gain and terminal attribution

Retained gain and the single-edit baseline. Table 2 reports the original six arm-A lineages used in the registered post-selection audit (sample history: App. E.3). Replay finds a mean MBPP reserve gain of 1.55 pp (95% CI $[0.78, 2.31]$). The best single edit, selected on audit data alone from the same candidate stream, gains $+1.44$ pp at the same candidate-evaluation budget. The paired lineage-minus-single-edit differences are $+0.11$ pp on MBPP (95% CI $[-0.62, +0.83]$; sign-flip $p = 0.84$) and $+0.81$ pp on HumanEval ($[-0.31, +1.93]$; §4.6). Both intervals cross zero: retained improvement establishes neither a successive-edit advantage nor equivalence. Total token expenditure is not matched. An exploratory extension to the eight original parent lineages gives a lineage-minus-

single-edit difference of -0.03 pp ($[-0.56, 0.51]$). It reuses the reserve and mixes service epochs (App. E.3).

Second model and serving configuration. An exploratory Qwen3.5-4B search (Qwen Team, 2026) reuses the bank, reserve and eight seed identities. Lineage (L) and single-edit (S) gains over M_0 are $+4.378$ and $+6.276$ pp. The primary $L-S$ contrast is -1.899 pp (95% CI $[-3.847, +0.049]$, sign-flip $p = 0.0625$), establishing neither superiority nor equivalence. Both schedule 400 slots per seed; L/S score $2{,}974/3{,}200$ candidates, with 226 lineage additions blocked by the three-exemplar cap. Unequal token costs and realized scoring prevent isolating persistence (App. E.9). A post-hoc ceiling counting all completed audit evaluations, including incumbents and checkpoints, preserves all eight single-edit endpoints. A separate post-hoc candidate-count sensitivity, measured in one new service batch, gives $L - C = -1.556$ pp (95% CI $[-3.337, +0.225]$, $p = 0.1016$) for truncated single-edit state C. Only one of eight S states changes; the same-batch $L - C$ and $L - O$ means differ by $+0.053$ pp descriptively (O: original S, remeasured). It matches realized scored-candidate counts, not total evaluations or tokens.

Original six-lineage cost. Search uses 48–57M tokens per lineage, excluding shared bank generation; terminal task cost is $1.51\times$ baseline. The token-cost crossover against two baseline samples is $H^* \approx 334$k–646k tasks, a lower bound excluding bank generation, not an accuracy-adjusted break-even. Oracle pass@2 is 68.4%, already reaching the evolved state's score; it upper-bounds a selector and is not a deployable policy (App. E.8).

An as-run exploratory public-example selection study remeasures the eight old endpoint pairs in one common service. The selector costs 446 tokens/task versus 321 for the lineage and 288 for the single edit; its primary selector-minus-lineage contrast is -0.079 pp (95% CI $[-0.839, 0.681]$). An exploratory calculation added after internal review gives $L - S = +0.185$ pp (nominal 95% t_7 interval $[-0.770, +1.139]$) on these same eight pairs, adding no searches or seeds. These fixed-cap results retain the stopping-rule ambiguity affecting 141 length-terminated responses (App. E.10).

Terminal audit design. Final states retain 1–4 components. We exhaust all $2^{|C|}$ subsets and compute $T_{\text{int}} = [F(C) - F(\emptyset)] - \sum_{c \in C}[F(\{c\}) - F(\emptyset)]$. This full-state term uses the factorial measurement; 1.55 pp comes from separate replay. App. E details the registered audit and the accept-level estimator changed after inspection.

Accepted credit and retained improvement. Summed accept-time credit (25.17 pp) accumulates selected, noisy increments; audit rise (2.57 pp) and reserve gain (1.55 pp) measure different quantities. The event-pooled first-remeasurement ratio is numerically below an independent-flip simulation reference (App. E.2, Fig. 7a). This descriptive comparison does not rule out correlated service noise or decompose lost performance causally.

Interaction among surviving components. Surviving components show no detectable positive interaction: pooled interaction is -0.59 pp ($[-1.55, 0.38]$, upper endpoint $+0.38$ pp), an interval satisfying neither the registered synergy criterion nor the additivity criterion (App. E, Fig. 4).

Overwritten accepts. Later edits overwrite the contributed field values of 79–94% of arm-A accepts; compaction leaves 87–98% overwritten or unselected (App. E.8). Their influence on later search remains unmeasured: the randomised accept/drop trial is still open (App. A).

Marginal gain across generations. Gains concentrate in the first two generations; later increments vary in sign and average within noise of zero by generation 3 (Fig. 6). H2 is descriptive without an equivalence margin. Bank exhaustion, inapplicable edits and declining gate power as effects shrink remain indistinguishable.

4.3 Pilot diagnostics and production design

Identical-specification re-evaluation flips $4.9/300$ audit items (net-win s.d. 2.4), allowing a one-draw gate to accept noise, and proposer diversity collapses: of 240 siblings, 115 are byte-identical duplicates and all 125 unique proposals add exemplars, whose mean effect is -4 pp (App. D). The

Table 2: Original six-lineage post-selection analysis; the later two-lineage measurement extension is in Table 8. Rows are separate estimands with their own comparators and evaluation units, not stages of one causal waterfall and not decompositions of one another: rows 1–2 are computed on the reused 300-item audit set that drove selection, rows 3–5 on the withheld 474-item MBPP reserve. Entries are percentage points with 95% across-seed t intervals; tasks are fixed. Row 4 matches candidate evaluations, not total tokens, and the near-zero row 5 interval is not an equivalence result. Row 2 uses a last-ten-incumbent-round fallback in two lineages lacking a post-final-accept observation (App. E.3).

Quantity (pp)	Comparator	Evaluation unit	Mean	95% CI
1. Selected credit at accept	incumbent at each accept	summed accept events	25.17	$[14.71, 35.63]$
2. Incumbent audit rise	generation-0 state	endpoint or fallback	2.57	$[2.00, 3.15]$
3. Replay lineage gain	M0 baseline	474 reserve tasks	1.55	$[0.78, 2.31]$
4. Audit-selected single edit	M0 baseline	474 reserve tasks	$+1.44$	$[+0.46, +2.43]$
5. Lineage − single edit	row 4, paired	474 reserve tasks	$+0.11$	$[-0.62, +0.83]$
6. Terminal interaction	additive component sum	$2^{\lvert C\rvert}$ terminal subsets	-0.59	$[-1.55, 0.38]$

weak online proposer yields no measurable overfitting in these pilots, so we use an audit-blind external bank with operation-stratified forcing, dev-validity filtering and identical candidate order across arms (§4.4); a single-seed pilot on that bank (Table 5) gives $\mathrm{dAUC}_{AB} = +3.1$ pp and fixes the production design.

4.4 Frozen-bank production: no detected acceptance-margin contrast

Frozen candidates isolate the acceptance-margin effect within this audit-blind, dev-screened bank. Paired arms consume the same seed-permuted stream; the unexposed tranche opens at generation 3.

Registered follow-up. Across all eight registered seeds, mean dAUC_{AB} is -0.21 pp (95% CI $[-1.06, +0.64]$, exact sign-test $p = 0.73$), with 3 of 8 contrasts positive: we detect no acceptance-margin contrast on the production bank (Fig. 2). The six-seed interim was cut after four contrasts had been inspected. Reserve unlocks used the original service for six seeds and a newer one for the two completion seeds, confounding service with seed batch; endpoint calibration does not establish equivalence of generation-wise dAUC. Apps. B and E record this history and the restoration of the eight-seed primary. At the observed spread, a normal location-shift approximation gives 2.0 pp resolution at 80% power for the registered sign test, not an equivalence margin. The pilot used a different bank, so bank and seed effects remain inseparable. A separately registered six-seed comparison with a PACE-inspired gate also detects no gap-AUC contrast; its operational threshold carries no established error-control guarantee in this reused-audit setting (App. E).

4.5 Assigned disclosure with measured uptake

Following an exploratory three-seed lead, the registered online study crosses disclosure view and gate margin in eight seeds: A (rich, margin 1), B (commit-only, margin 5), R5 (rich, margin 5) and B1 (commit-only, margin 1). The sole confirmatory contrast, R5−B, compares rich and commit-only disclosure at a fixed margin-5 gate across eight paired seeds, using the preregistered sign-flip test (App. E.4). It separates assigned disclosure from acceptance strictness, which the frozen-bank experiment cannot test.

Disclosure contrast. The exploratory lead does not persist in the registered sample: mean dAUC_{R5-B} is -0.46 pp (95% t interval $[-1.50, +0.59]$, $p = 0.31$, 4 of 8 contrasts positive), and at the observed spread a normal location-shift approximation to the registered test gives about 1.5 pp resolution at 80% power; this is a planning approximation, not empirical calibration. The secondary contrasts agree, and the joint A/B contrast also lies below the exploratory three-seed mean, although that comparison does not test a difference between cells (App. E). The registered experiment supplies no detectable disclosure effect under this proposer and gate.

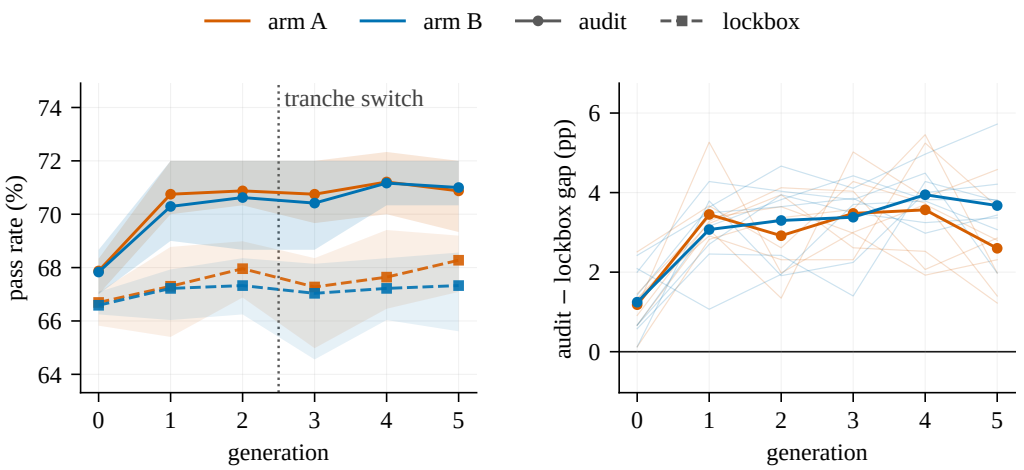

Figure 2: Frozen-bank production across all eight registered seeds. **Left**: mean audit (solid) and MBPP reserve (dashed) accuracy by arm; bands show the across-seed range, not confidence intervals. **Right**: audit–reserve gaps ("lockbox" denotes the MBPP reserve), with individual seeds as thin lines. The primary contrast is $\mathrm{dAUC}_{AB} = -0.21$ pp (95% CI $[-1.06, +0.64]$, exact sign-test $p = 0.73$), and final-generation gaps average 2.60 pp (arm A) and 3.67 pp (arm B). Reserve unlocks for six seeds used the original inference service and the two completion seeds the newer one, so the trajectories combine measurement contexts; App. E reports the measured service drift, the six-seed interim analysis and calibration details.

Reserve outcomes. Mean MBPP reserve change is negative in all four arms: from $g_0 \to g_5$ it is A -1.79, B -0.63, R5 -0.61 and B1 -2.08 pp, with intervals excluding zero for A and B1 but not for B and R5. Declines are about threefold larger under margin 1 than margin 5 in both disclosure conditions. Mean reserve accuracy also falls under commit-only feedback; the registered fixed-gate rich-versus-commit contrast detects no difference in audit–reserve gap AUC. The fresh-data veto in the exploratory H3 cell likewise overturns unsupported commits without a reserve advantage (App. D).

Disclosure uptake. The transcript check finds verbatim copies of disclosed audit prompts in 14.7% of rich-view proposals against 0.0% under the commit view, higher in 8 of 8 seeds. Only 1.1% of accepted rich-view edits contain a copied audit prompt, but copied prompts remain in 4 terminal exemplars, so copying reaches persistent state in some runs without establishing that the proposer follows audit scores. Rich feedback does not raise the registered score-following measure; the measure is lower under rich feedback ($\Delta U_2 = -0.070$, 0 of 8 positive) and may be affected by copied text. This experiment estimates the effect of providing rich feedback, not of copying prompts or following scores (App. E).

4.6 A NEVER-READ SECOND RESERVE AND A KNOWN-LEAK CONTROL

The second reserve shows retained gain without a detected acceptance-margin contrast. After freezing the registered 4B cells and analyses, we score their checkpoints once on 164 public HumanEval tasks in one batch. No project run, analysis or author had consulted this reserve, and outputs were unread until batch completion. Researchers had inspected MBPP reserve aggregates between cells, though no optimisation loop received them. The eight-seed contrast is $\mathrm{dAUC}_{AB} = -0.40$ pp ($[-2.17, +1.38]$, $p = 0.63$). Arm A gains $+1.83$ pp over eight seeds ($[+0.52, +3.14]$).

Disclosure on transfer. At a fixed gate, $\mathrm{dAUC}_{R5-B} = -0.95$ pp ($[-1.84, -0.05]$, unadjusted $p = 0.04$, not surviving Holm correction at 0.08). The negative contrast arises on the audit side, where the rich-view score rises less, rather than from a reserve-side advantage (Table 10); both

primary permutation tests are unadjusted, so this estimate must be read with the correction across the second-reserve primaries.

Disclosure and calibration. The margin-5 known-leak arm gives $\mathrm{dAUC}_{L-B} = -1.65$ pp ($[-3.48, +0.19]$, $p = 0.13$). Its 27 normalized reference-copy events span two tasks and four seed–task combinations; all have negative audit net wins, with no winners or accepts (Table 11, App. E.5). Separate direct installation bypasses proposal and acceptance: three audit references yield $+30$ target and -41 other-task net successes against matched dev exemplars (five measurements, six blocks), -0.122 pp audit change and -0.530 pp gap contrast ($[-1.723, 0.662]$). Its positive-signal criterion fails. Neither control establishes end-to-end calibration; uptake beyond these matches remains unmeasured. The real-proposer second layer was not run (App. E.6).

5 Runtime Controls in a 27B Task Diagnostic

An earlier τ^2-Bench (Barres et al., 2025) search retained the MMS transfer guard P3.0 despite no matches in 3,960 checks (App. F). We now run 74 reused public TRAIN tasks twice under four nested controls: C0 native; C1 the original wrapper without the guard; C2 original P3.0; and C3 C1 with content clearing removed. C3 does not repair the guard, and the design is not a full factorial. Native total-reward successes are $98/148$, $91/148$, $97/148$ and $94/148$, respectively.

Table 3: All fixed full-task contrasts, in percentage points. Intervals bootstrap 74 paired tasks after averaging their two repeats. The two key secondary intervals use the prespecified Bonferroni adjustment; repeats are not additional task units.

Contrast	Difference	Level	Interval
$C1 - C0$ (primary)	-4.730	95%	$[-15.541, +6.757]$
$C1 - C3$ (content clearing)	-2.027	97.5%	$[-14.189, +10.135]$
$C2 - C1$ (guard-enabled condition)	$+4.054$	97.5%	$[-2.703, +10.811]$
$C3 - C0$ (descriptive)	-2.703	95%	$[-8.784, +3.378]$

C1/C2 omit assistant text accompanying tool calls from later agent prompts. Offline token inputs differ no later than the first semantic response difference in $110/148$ C1$-$C0 pairs and $110/148$ C1$-$C3 pairs. A post-result check without Q selection finds changed adopted tool-call names or arguments at that response in 28 and 26 of these witnesses, respectively, each spanning 15 tasks. These conditional trace counts do not estimate behavioral prevalence or isolate mediation of Q.

In a post-result example selected for an action change, C1 deletes a lookup's accompanying text from the next prompt; C3 preserves it. The saved next responses have zero versus six adopted tool calls (App. F.1).

All four Q intervals cross zero, establishing neither benefit, harm nor practical equivalence. Same-condition repeats disagree on $16, 11, 11, 14$ of 74 tasks (C0–C3). C2 has 910 checks, zero hits or blocks, and no observed guard feedback in requests. Its positive C2$-$C1 point estimate does not establish guard-function benefit or explain the historical acceptance score. The case supports auditing an edit's execution path separately from its stated function. App. F.1 reports the full evidence.

6 What the Controls Change

Credit for improvement versus successive edits. An improved coding endpoint credits the final harness, while an audit-selected single edit from the same stream also retains gain. The paired comparison leaves successive-edit value unresolved (§4.2).

Installed behavior versus the target rule. Assistant-content clearing persists when the naturally inactive P3.0 guard is removed (§5). That request change belongs to the installation wrapper. Acceptance and persistence do not establish guard use or explain the old acceptance score; task effects remain uncertain.

Known-leak and direct-injection results support calibration diagnostics, not end-to-end validation. Better future search or deployment decisions remain untested. App. A gives the auditability checklist.

7 LIMITATIONS

Scope. The 4B evidence covers two Qwen model/serving conditions, with a shared bank and reused reserve. Fixed-bank composition does not test new, history-dependent proposals; the online study uses one proposer. The edit space is narrower than program editing on SWE-bench (Zhang et al., 2026b; Nie et al., 2026; Shao et al., 2026; Jimenez et al., 2024). Neither the known-leak null nor the direct-injection target gains establish end-to-end calibration. The 27B task panel uses nested controls on reused public TRAIN tasks; repeated executions add no independent algorithm seeds, and task benefit, harm and practical equivalence remain unresolved. The earlier local/pilot results and unresolved pilot cost bounds remain separate (App. F).

Exploratory boundaries. The separate four-round search-availability study has no equivalence margin (App. G). An earlier 27B calibration failed at $F_{\max} = 25$, leaving downstream search and readout unexecuted, with no effect estimate, calibrated bound or imputed zero (App. F).

Estimands. Audit and lockbox are assumed exchangeable. Differencing removes only a fixed offset, so the audit-to-reserve difference in §4.2 includes task-set differences and seed-context variance (0.47 pp; App. E.2). The gap statistic measures audit–reserve divergence rather than reserve improvement, which we report separately, and terminal attribution measures the final components, not the causal value of the path that found them.

Researcher access to the lockbox. The loop never received MBPP reserve outcomes, but researchers inspected aggregates between cells, so prospective later registrations do not restore independence from those tasks (Table 20). HumanEval was first consulted in the final registered 4B batch; the 27B analysis followed, on separate public TRAIN traces.

Nominal disclosure accounting. Encoding charges are not measurements of information consumed. The entropy inequality requires the stated conditional information boundary; a charge alone supplies no performance bound.

8 CONCLUSION

Across two 4B model/serving conditions, retained gain does not establish a lineage advantage. The original comparison matches candidate evaluations; the Qwen3.5 primary comparison matches scheduled slots. Original terminal interaction is not detectably positive. In 27B, the guard remains naturally inactive while wrapper controls yield differing next actions; synthetic blocking does not establish task benefit, and full-task effects remain uncertain. Direct-injection target gains are offset elsewhere. These controls distinguish endpoint gain, successive-edit value, rule-specific behavior and aggregate response.

REPRODUCIBILITY STATEMENT

The appendices describe protocols, configurations and completed checks of saved verdicts and execution records. Code and data artifacts are not distributed with this PDF-only submission. App. H documents the internal reconstruction and replay checks, their dependencies and limits; Apps. D, E and F give protocols and diagnostics.

AI USE STATEMENT

Generative AI assisted literature retrieval and discovery; research ideation and execution, including experimental design, implementation, debugging, result inspection and internal criticism; drafting manuscript sections; and writing, polishing and figure preparation. Models also served as experimental agents and proposers. We checked implementations against tests and execution records, and numerical claims against analysis artifacts. The authors take responsibility for the submission, including interpretations and citations.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732. arXiv:2108.07732.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. URL https://arxiv.org/abs/2506.07982.

Martin Andres Bertran, Aaron Roth, and Zhiwei Steven Wu. What fits (into few tokens) doesn't overfit: Compression and generalization in ML research agents, 2026. URL https://arxiv.org/abs/2606.11045. arXiv:2606.11045.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML)*, pp. 1006–1014, 2015. URL https://arxiv.org/abs/1502.04585. arXiv:1502.04585.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning (ICML)*, pp. 6621–6642, 2024. URL https://arxiv.org/abs/2401.01335. arXiv:2401.01335.

Yizhe Chi, Deyao Hong, Dapeng Jiang, Tianwei Luo, Kaisen Yang, Boshi Zhang, Zhe Cao, Xiaoyan Fan, Bingxiang He, Han Hao, Weiyang Jin, Dianqiao Lei, Qingle Liu, Houde Qian, Bowen Wang, Situ Wang, Youjie Zheng, Yifan Zhou, Calvin Xiao, Eren Cai, and Qinhuai Na. Frontier-Eng: Benchmarking self-evolving agents on real-world engineering tasks with generative optimization, 2026. URL https://arxiv.org/abs/2604.12290. arXiv:2604.12290.

Yunpeng Dong, Jingkai He, Shiqi Liu, Yuze Hou, Dong Du, Zhonghu Xu, Si Yu, Baochuan Yang, Yubin Xia, and Haibo Chen. DeltaBox: Scaling stateful AI agents with millisecond-level sandbox checkpoint/rollback, 2026. URL https://arxiv.org/abs/2605.22781. arXiv:2605.22781.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, 2015a. URL https://proceedings.neurips.cc/paper/2015/hash/bad5f33780c42f2588878a9d07405083-Abstract.html.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

Xingze Gao, Chuanrui Hu, Hongda Chen, Pengfei Yao, Zhao Wang, Yi Bai, Zhengwei Wu, Yunyun Han, Xiaofeng Cong, Jie Gui, Yafeng Deng, and Teng Li. EvoAgentBench: Benchmarking agent self-evolution via ability transfer, 2026. URL https://arxiv.org/abs/2607.05202. arXiv:2607.05202.

Diandian Guo, Cong Cao, Fangfang Yuan, Yingqi Wang, Yueshan Wang, and Dakui Wang. Self-authored verification is unreliable in heuristic self-improving agents, 2026. URL https://arxiv.org/abs/2607.24300. arXiv:2607.24300.

Alex Iacob, Andrej Jovanović, William F. Shen, Daniel Burkhardt, Meghdad Kurmanji, Nurbek Tastan, Lorenzo Sani, Niccolò Alberto Elia Venanzi, Ambroise Odonnat, Zeyu Cao, Bill Marino, Xinchi Qiu, and Nicholas D. Lane. The red queen gödel machine: Co-evolving agents and their evaluators, 2026. URL https://arxiv.org/abs/2606.26294. arXiv:2606.26294.

Sihang Jiang, Lipeng Ma, Zhonghua Hong, Keyi Wang, Zhiyu Lu, Tengfei Wang, Shisong Chen, Jinghao Zhang, Tianjun Pan, Weijia Li, Jiaqing Liang, and Yanghua Xiao. SEA-Eval: A benchmark for evaluating self-evolving agents beyond episodic assessment, 2026. URL https://arxiv.org/abs/2604.08988. arXiv:2604.08988.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770. arXiv:2310.06770.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180. arXiv:2309.06180.

Jialong Li and Jialing Zhu. Auditing self-evolution in financial agents: Capability gains, security drift, and execution-interface mismatch, 2026. URL https://arxiv.org/abs/2608.17684. arXiv:2608.17684.

Minhua Lin, Juncheng Wu, Zijun Wang, Zhan Shi, Yisi Sang, Bing He, Zewen Liu, Tianxin Wei, Zongyu Wu, Zhiwei Zhang, Dakuo Wang, Xiang Zhang, Benoit Dumoulin, Cihang Xie, Yuyin Zhou, Suhang Wang, and Hanqing Lu. Harness updating is not harness benefit: Disentangling evolution capabilities in self-evolving LLM agents. *arXiv preprint arXiv:2605.30621*, 2026. URL https://arxiv.org/abs/2605.30621.

Ming Liu. More is not always better: Cross-component interference in LLM agent scaffolding, 2026. URL https://arxiv.org/abs/2605.05716. arXiv:2605.05716.

Xiaotian Luo, Dizhan Xue, Fengxingyu Wang, Chuanrui Hu, and Yafeng Deng. HarnessBank: Semantic gene-bank search with gated verification for agent-harness self-evolution. *arXiv preprint arXiv:2607.13683*, 2026. URL https://arxiv.org/abs/2607.13683v2.

Yuchen Ma, Yue Huang, Han Bao, Haomin Zhuang, Swadheen Shukla, Michel Galley, Xiangliang Zhang, and Stefan Feuerriegel. SkillGen: Verified inference-time agent skill synthesis, 2026. URL https://arxiv.org/abs/2605.10999. arXiv:2605.10999.

Fanqing Meng, Lingxiao Du, Qiguang Chen, Ziqi Zhao, Haocheng Lu, Mengkang Hu, and Michael Qizhe Shieh. RSIBench-Data: Benchmarking data-centric research for recursive self-improvement, 2026. URL https://arxiv.org/abs/2607.25886. arXiv:2607.25886.

Jun Nie, Yonggang Zhang, Qianshu Cai, Yiu-ming Cheung, Xinmei Tian, and Bo Han. EvolveNet: Collaborative harness evolution for agent self-improvement, 2026. URL https://arxiv.org/abs/2608.04968. arXiv:2608.04968.

Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, Maohao Shen, Vishy Tirumalashetty, George Lee, Jiawei Han, Tomas Pfister, and Chen-Yu Lee. SkillOS: Learning skill curation for self-evolving agents, 2026. URL https://arxiv.org/abs/2605.06614. arXiv:2605.06614.

Zhenting Qi, Susanna Maria Baby, Stefanie Anna Baby, Kan Yuan, Andrew Tomkins, Tu Vu, Da-Cheng Juan, and Cyrus Rashtchian. On the generalization gap in self-evolving language model reasoning, 2026. URL https://arxiv.org/abs/2606.01075. arXiv:2606.01075; ICML 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Maxime Robeyns, Martin Szummer, and Laurence Aitchison. A self-improving coding agent, 2025. URL https://arxiv.org/abs/2504.15228. arXiv:2504.15228.

Amit Roth, Ankur Samanta, Matan Halevy, Yoav Levine, and Yonathan Efroni. Hack-verifiable environments: Towards evaluating reward hacking at scale, 2026. URL https://arxiv.org/abs/2605.20744. arXiv:2605.20744.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1232–1240. PMLR, 2016. URL https://proceedings.mlr.press/v51/russo16.html.

Shuai Shao, Kangning Zhang, Qingyao Li, Shijian Wang, Hao Wang, Wenxiang Jiao, Yuan Lu, Yi Guo, Weiwen Liu, and Weinan Zhang. Harness-r1: Learning to edit executable runtime harnesses from agent failure trajectories, 2026. URL https://arxiv.org/abs/2608.02276. arXiv:2608.02276.

Zayx Shawn. PACE: Anytime-valid acceptance tests for self-evolving agents, 2026. URL https://arxiv.org/abs/2606.08106. arXiv:2606.08106.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://doi.org/10.1038/s41586-024-07566-y.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2305.16291. arXiv:2305.16291 (2023).

Wenyi Wang, Piotr Piękos, Li Nanbo, Firas Laakom, Yimeng Chen, Mateusz Ostaszewski, Mingchen Zhuge, and Jürgen Schmidhuber. Huxley-gödel machine: Human-level coding agent development by an approximation of the optimal self-improving machine. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://openreview.net/forum?id=T0EiEuhOOL. arXiv:2510.21614 (2025).

Xing Wang, Xiaoyi Zhang, and Jie Shao. Auditing harness tampering in self-improving agents, 2026b. URL https://arxiv.org/abs/2609.00069. arXiv:2609.00069.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, Hannaneh Hajishirzi, Yulia Tsvetkov, Pradeep Dasigi, and Teng Xiao. Rethinking the evaluation of harness evolution for agents, 2026c. URL https://arxiv.org/abs/2607.12227. arXiv:2607.12227.

Yixuan Wang, Yiyang Zhou, Yiming Liang, Congyu Zhang, Fuxiao Liu, Jiawei Zhou, and Huaxiu Yao. Not all skills help: Measuring and repairing agent knowledge, 2026d. URL https://arxiv.org/abs/2606.15390. arXiv:2606.15390.

Weco AI. AIDE^2: The first evidence of recursive self-improvement. Company blog, 14 July 2026, https://www.weco.ai/blog/first-evidence-of-recursive-self-improvement, July 2026. Non-archival. Announced the same day in an X thread by Z. Jiang, https://x.com/zhengyaojiang/status/2077079778793042425; a technical report was stated to follow. Both URLs accessed 2026-09-05.

Yuhao Wu, Jingyuan Zhang, Jiajun Shi, Xinping Lei, Qingshui Gu, Yuxuan Zhang, Zexuan Wang, Chen He, Chen Huang, Maojia Song, Zhiyuan Zeng, Shaowen Wang, Jinkai Liu, Yunfeng Shi, Jiaheng Liu, Shen Yan, Wenhao Huang, Ge Zhang, and Wenxuan Zhang. HarnessDev: Can LLMs create and evolve their own agent harness?, 2026. URL https://arxiv.org/abs/2609.01437. arXiv:2609.01437.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking, 2026. URL https://arxiv.org/abs/2605.05973. arXiv:2605.05973.

Shuhan Xue, Zixin Ding, Yichen Shen, Yinjie Wang, Zhenfei Yin, Yingcheng Wu, Yuxin Chen, Mengdi Wang, and Ling Yang. PAST-Bench: Benchmarking the foundations of recursive self-improvement in personal agents, 2026. URL https://arxiv.org/abs/2608.04003. arXiv:2608.04003.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning (ICML)*, pp. 57905–57923, 2024. URL https://arxiv.org/abs/2401.10020. arXiv:2401.10020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15476–15488, 2022. URL https://arxiv.org/abs/2203.14465. arXiv:2203.14465.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-harness: Harnesses that improve themselves, 2026a. URL https://arxiv.org/abs/2606.09498. arXiv:2606.09498.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin gödel machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://arxiv.org/abs/2505.22954. arXiv:2505.22954 (2025).

Xing Zhang, Yanwei Cui, Guanghui Wang, Ziyuan Li, Wei Qiu, Bing Zhu, and Peiyang He. Library drift: Diagnosing and fixing a silent failure mode in self-evolving LLM skill libraries, 2026c. URL https://arxiv.org/abs/2605.19576. arXiv:2605.19576; ICML 2026 Workshop on Failure Modes in Agentic AI.

Xing Zhang, Guanghui Wang, Yanwei Cui, Ziyuan Li, Wei Qiu, Bing Zhu, and Peiyang He. Who grades the grader? co-evolving evaluation metrics and skills for self-improving LLM agents, 2026d. URL https://arxiv.org/abs/2607.12790. arXiv:2607.12790.

Yifan Zhang, Yutong Dai, Juntao Tan, Luyu Yang, Rishi Mullur, Thai Hoang, Zhiyuan Hu, James Zhu, Phil Mui, Silvio Savarese, Ran Xu, and Zeyuan Chen. DarwinX: Evolving agent harnesses through natural selection, 2026e. URL https://arxiv.org/abs/2608.07545. arXiv:2608.07545.

Congjie Zheng, Chuanyi Xue, Bin Liang, Jun Yang, and Changshui Zhang. SEAGym: An evaluation environment for self-evolving LLM agents. *arXiv preprint arXiv:2606.17546*, 2026. URL https://arxiv.org/abs/2606.17546.

Chenyu Zhou. More convincing, not more correct: Self-play reward hacking of reference-free LLM judges, 2026. URL https://arxiv.org/abs/2607.05904. arXiv:2607.05904.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, and Jun Wang. Memento: Fine-tuning LLM agents without fine-tuning LLMs, 2025. URL https://arxiv.org/abs/2508.16153. arXiv:2508.16153.

Ziyi Zhu, Luka Smyth, Saki Shinoda, and Jinghong Chen. SAGE: Stochastic prompt optimization via agent-guided exploration, 2026. URL https://arxiv.org/abs/2606.18902. arXiv:2606.18902.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084. arXiv:2504.16084; NeurIPS 2025.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. Self-adapting language models, 2025. URL https://arxiv.org/abs/2506.10943. arXiv:2506.10943; NeurIPS 2025.

A THE CHECKLIST APPLIED TO THE PUBLISHED RECORD

These annotations describe published protocol evidence, not result validity. The checklist covers self-evolution results (Zhang et al., 2026b;e;a; Shao et al., 2026; Ma et al., 2026; Nie et al., 2026) and one non-archival claim (Weco AI, 2026). Symbols indicate whether the record satisfies the stated criterion: ✓ supported, ∼ partial or mixed, and – not demonstrated (including explicit reuse). Q1–Q6 cover lineage, feedback granularity, held-out queries and returned content, matched non-evolving baseline, silent final evaluation, and cost break-even. For Q5, ✓ requires a reported final endpoint that supplies neither proposer nor acceptance-gate feedback during the corresponding improvement loop. For Q6, ✓ requires a deployment break-even calculation; search or construction cost disclosure alone is partial.

Table 4: Auditability of recent self-evolution results from their published record. The last row applies the checklist to a widely circulated non-archival claim of "the first experimental evidence of RSI" (Weco AI, 2026): eight days of harness autoresearch beating two years of hand engineering on held-out benchmarks, announced in a company blog and an X thread with a technical report stated to follow.

	Q1 lineage	Q2 granul.	Q3 queries	Q4 budget	Q5 silent	Q6 break-even
Darwin Gödel Machine (Zhang et al., 2026b)	✓	∼	–	–	∼	∼
DarwinX (Zhang et al., 2026e)	∼	∼	–	∼	✓	–
Self-Harness (Zhang et al., 2026a)	✓	✓	∼	–	–	–
Harness-R1 (Shao et al., 2026)	∼	✓	–	–	∼	–
SkillGen (Ma et al., 2026)	✓	✓	∼	–	✓	∼
EvolveNet (Nie et al., 2026)	∼	∼	–	–	✓	–
AIDE^2 blog (Weco AI, 2026)	∼	–	–	–	∼	–
This paper (self-audit)	✓	✓	✓	∼	∼	∼

Self-Harness uses its held-out split in the promotion gate (§§3.4, 4.1 of Zhang et al. 2026a); hidden proposer traces do not make that endpoint silent. Harness-R1 mixes same-task reruns with a separate 1,270-task transfer endpoint (§4.4 of Shao et al. 2026). EvolveNet freezes the harness before each final test, although BIRD informed overall method development (§5 of Nie et al. 2026). DGM reports per-run cost (App. E.1 of Zhang et al. 2026b); SkillGen separates construction from per-call cost (App. C.5 of Ma et al. 2026); neither supplies the required break-even.

Q4 and Q6 are reported in §4.2 and App. E.8, and marked partial: the best-single-edit baseline matches candidate evaluations rather than tokens, and the crossover scale H^* is a token-cost crossover, not an accuracy-adjusted deployment break-even. Q5 is partial: the MBPP lockbox was reused across cells and researchers saw its aggregates, while the second reserve was read once after the registered 4B campaign was frozen (App. E). The prospective randomised accept/drop trial remains open. The AIDE^2 row is graded from the blog, which describes the outer-loop iterations and retained improvements (Q1) and names held-out benchmarks (Q5) but reports no query counts, no budget-matched non-evolving baseline and no search cost.

The post-selection audit in §4.2 measures functional attribution in the terminal state. The prospective shared-pool randomised accept/drop trial, which would measure the selection effect itself, remains open, and the paper does not present the former as the latter.

B GATE POWER AND NOISE CALIBRATION

A 3-point paired effect needs roughly 430–860 items under a two-sided $\alpha = 0.05$ test with power 0.8 and 5–10% discordant items. The 474-item lockbox lies within this range. The primary analysis differences each arm from its generation-0 gap and uses shared item-paired bootstrap indices across arms and generations. The registered exact two-sided sign test uses per-seed contrast signs. The sign-flip test of the mean is exact under the null assumption that the joint seed-level contrast distribution is invariant to independent sign flips. Eight paired seeds give a floor $p = 0.0078$. An operator cut the run to six after observing four values, with two seeds not yet started and a floor $p = 0.0312$. All seeds score the same 474 items, treating tasks as fixed and candidate order and serving context as

random. Arm A's greedy rule equals margin 1, because the candidate-minus-incumbent mean equals net-wins/300. B uses margin 5, which noise did not clear in 14 draws ($0/14$, exact upper bound 0.19) but which rejects $8/11$ rounds holding a net-wins ≥ 3 candidate, so a flat B trajectory must be read with both error curves. Summary of the tables produced by `analysis/gate_power.py` on the frozen-bank pilot (full tables retained in the saved analysis collection): recall/false-commit curves for the margin gate, the greedy $\equiv$ margin-1 structural identity, the replicate-noise null (n=14 duplicate evaluations, net-wins s.d. 2.4 items, margin-3 cleared by noise in $3/14$ draws, margin-5 in $0/14$, exact upper bound 0.19), and the counterfactual margin sweep over the greedy arm's accepts ($14 \to 6/3/0$ at margins $3/5/8$).

C Ordinal Bookkeeping: Three-Proxy Rank Analysis

In the frozen-bank pilot's rich-feedback arm, nominal payload charges are 14.9–16.5$\times$ compressed transcript length. Within-arm rank comparisons have only three decision generations, so the minimum two-sided permutation p is $1/3$; ranks are descriptive. In the commit-only arm, nominal charge is an affine function of commit count, making their rank agreement tautological. Mean selected gain and commit count also differ in unit, preventing a pooled cross-arm gain comparison. These checks support provenance bookkeeping, not a calibrated information or causal-gain estimator.

D Additional Protocol Detail

Arms removed at design review. We removed four candidate arms and narrowed the claim. A scalar-feedback arm and a 1% oracle dose did not provide enough resolution for a dose-response analysis. An ϵ-randomised gate would compound per-generation recall r to r^5, manufacturing a null. A shuffled-accept control did not isolate a single intervention. The resulting experiment supports a high/low bandwidth contrast without claiming a monotone dose response.

Pilot ledger sanity. No-op rounds were 0/60 and 2/60 in pilot 1 and 0/60 and 0/60 in pilot 2 (arms A/B). No infrastructure abort was scored as a failure. Disclosure was $\sim 10^4$ bits per round for A versus about one bit per round for B (Table 5 sums per generation), with byte-identical dev minibatches. At K=1, arm A accepts two near-identical instruction rewrites in 60 rounds. Their $+1.3$ and $+0.3$-point audit changes arise from temperature-0 serving nondeterminism. Candidates average -3.96 points against the incumbent, with a range from -6.3 to $+1.3$. The final gap is $\Delta^A_{\text{af}} = +2.69$ pp, and $\Delta^A_{\text{af}} - \Delta^B_{\text{af}} = +0.39$ pp (95% CI $[-4.40, +5.16]$, NO-GO). Increasing only K to four gives zero arm-A accepts in 60 rounds. Its final gap is $\Delta^A_{\text{af}} = +1.67$ pp, with $\Delta^A_{\text{af}} - \Delta^B_{\text{af}} = -1.84$ pp (CI $[-5.93, +2.09]$, NO-GO). The commit-only proposer retains diversity, with 218 unique proposals comprising 151 `set` and 67 `add_exemplar` edits. In the frozen-bank pilot, greedy selection gives a $+4.4$ pp audit–fresh gap, compared with $+2.9$ pp under the 1-bit gate. Arm-level gap AUCs are $+3.5$ and $+0.4$ pp, giving $\text{dAUC}_{AB} = +3.1$ pp.

Fresh tranche. The dissipation hypothesis attributes the stalled pilot gap to exhaustion of audit-favourable candidates and predicts renewed gap growth when the unexposed tranche opens at generation 3. Arm A instead has negative gap growth on the fresh tranche in four of six seeds. The descriptive mean over the phase, $\Delta^A_{\text{af}}(5) - \Delta^A_{\text{af}}(2)$, is -0.43 pp. Fresh candidates improved both arms without opening new audit exploits.

Fresh-data veto (W2c, exploratory). Arm C runs B's gate verbatim and, on each commit, re-scores winner against incumbent on 15 items from a 50-item reserved oracle pool, wrapping when exhausted and flagging and charging each reuse to income. This step overturns unsupported commits. It retains B's post-veto leakage rule. The veto overturned 8 of 13 gate commits across three seeds (62%). A margin-5 audit gate commits mostly to changes that fifteen fresh items cannot confirm (App. B), and at this scale a cheap veto yields no fresh-set advantage. The three-seed exploratory H3 cell has mean $\text{dAUC}_{CB} = -0.48$ pp.

Candidate banks. The pilot bank (283 candidates) and the production bank (524, with tranches $183/341$) were generated independently and audit-blind. We generated the production bank after the pilot with a dev-only health screen. Its operation-stratified patches come from Qwen3-32B

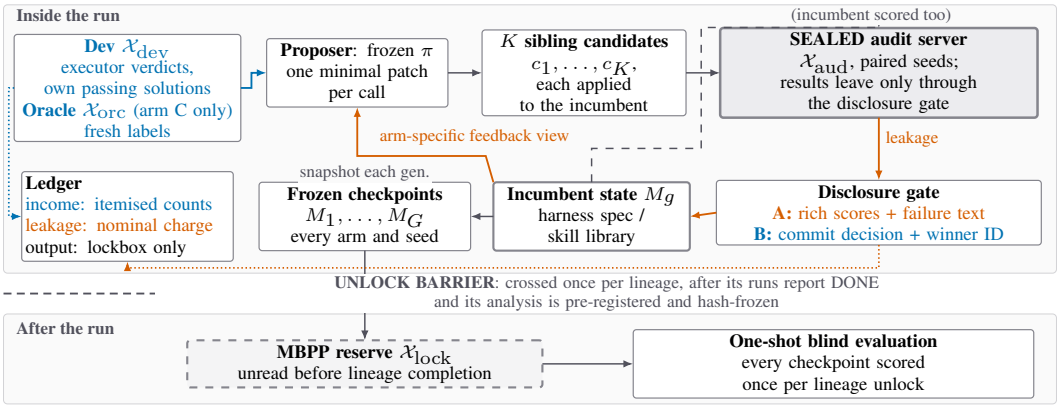

Figure 3: **The Information Ledger protocol.** External evidence enters as income matched across the bandwidth arms. Arm C adds oracle feedback. The audit set is read only inside a sealed server, and the arm's disclosure is charged at disclosure time (ℓ denotes characters of disclosed failure text, at 8 bits each). The persistent state M_g, the only mutable component since π is frozen, is snapshotted every generation. The online-proposer variant is shown. §4.4 replaces the proposer with a fixed candidate bank. The silent lockbox is opened once per lineage after its runs terminate and the analysis is pre-registered, scoring each checkpoint once at that unlock. Reuse of the reserve across cells and inspection of its aggregates between cells are logged in App. E.

and Qwen3-4B. Bank generation is shared external search expenditure outside per-round income, so absolute gains are conditional on the bank. Every production run's metadata records hashes of the bank content, candidate order and tranche boundary. The banks share one candidate by specification hash. Five generations of twenty rounds at $K{=}4$ consume 400 of the 524 candidates, so 124 (all in the second tranche) are never drawn. The 183 first-tranche candidates last until round 6 of generation 3, where the second tranche begins.

Disclosure implementation. Failure text is charged at 8 bits per character. An arm-A round discloses about 10^4 nominal bits, while arm B returns one bit plus $\log_2 K$ only on a commit. Charges occur on disclosure, never evaluation. Sealed records remain unread during optimisation, and arm B exposes only $\{\texttt{commit}, \texttt{winner_idx}\}$, including on standard output. The difference is in nominal payload size under matched income; it is provenance accounting, not measured information consumed by the optimiser.

Table 5: Ledger excerpt from the frozen-bank pilot. No new dev evidence enters these runs and both arms draw the same external bank, so the arms differ in acceptance threshold and in the nominal payload recorded, not in what can change proposal generation. Leakage is the nominal per-generation payload bound; per-generation entries are rounded to two figures and the totals use unrounded values. Reserve output is measured after the lineage completion marker. Arm A's audit score rose from 67.7 to 73.0 (+5.3 pp).

	Accepted / decisions		Nominal disclosure charge (bits)		Lockbox acc. (%)	
Gen	arm A	arm B	arm A	arm B	arm A	arm B
0 (baseline)	—	—	1.7×10^4	0	66.7	66.7
1	10/20	2/20	3.6×10^5	24	67.5	67.7
2	1/20	0/20	3.5×10^5	20	68.4	67.5
3	3/20	1/20	3.5×10^5	22	68.6	67.1
audit-claimed final	73.0 / 70.0		1.07×10^6 / 66		gap: +4.4 / +2.9 pp	

Conditional information boundary. The motivation follows adaptive-data analyses (Dwork et al., 2015a; Russo & Zou, 2016). Let Z contain the initial state, fixed candidate bank when

present, matched income and exogenous run randomness. Under the conditional information boundary $\mathcal{X}_{\text{aud}} \to \mathcal{T} \to M_G \mid Z$, data processing gives $I(M_G; \mathcal{X}_{\text{aud}} \mid Z) \le I(\mathcal{T}; \mathcal{X}_{\text{aud}} \mid Z) \le H(\mathcal{T} \mid Z)$. This assumption requires audit information to reach the final state through the recorded transcript. The ledger's encoding charges are nominal disclosure costs, not estimates of conditional entropy or information used by the proposer, and supply no conversion to performance percentage points (App. C).

D.1 Editable harness, candidate bank and terminal examples

Initial state and allowed edits. The mutable specification contains three text fields, a list of exemplars, a repair cap and a decoding temperature (Table 6). The `set` operation replaces an entire text field, rather than appending an instruction. `set_param` changes the repair cap or temperature; `add_exemplar` appends a prompt/solution pair and `remove_exemplar` deletes an existing pair by index. Each exemplar's solution must parse as Python. Repair runs only after a syntax error and uses the error message, without executing ground-truth assertions.

Table 6: Initial specification and schema limits. Text lengths count characters. The exemplar limit applies to prompt plus solution. The three initial text values are reproduced below.

Field	Initial value	Allowed range
System prompt	Text below	Nonempty, at most 1,500 characters
Task instructions	Text below	Nonempty, at most 2,500 characters
Repair instructions	Text below	Nonempty, at most 1,500 characters
Exemplars	Empty	At most 3 pairs, 2,500 characters each
Repair cap	0	0 or 1
Temperature	0	$[0, 1]$

The initial system prompt is:

> You are an expert Python programmer.

The initial task instructions are:

> Write a single self-contained Python function that solves the task. Use the exact function name shown in the example test. Output ONLY a Python code block, no explanations.

The initial repair instructions are:

> Your previous code had a problem. Review it carefully and output a corrected complete Python code block.

Realized bank and development screening. The 524 fixed candidates comprise 163 task-instruction replacements, 246 system-prompt replacements, 112 exemplar additions and three parameter edits: repair cap 1, temperature 0.1 and temperature 0.2. The bank contains no repair-instruction replacement or exemplar removal, although the schema permits both. The recorded construction applies each patch to the initial state M_0, deduplicates resulting specifications and screens the 531 unique candidates on 25 dev tasks. The baseline passes 60% of this panel. Candidates below 50% are removed, dropping 7 and retaining 524. Retained health rates also define terciles for tranche matching alongside generator and operation class. This is a screen of individual edits on M_0; it does not test or ensure multi-edit complementarity.

Six original terminal pairs. Table 7 describes all six original arm-A lineage endpoints and their audit-selected single-edit comparators. Their saved specifications match the identities attached to the corresponding reserve score vectors. Every lineage changes both system prompt and task instructions. Repair and temperature vary as shown. Every single-edit endpoint changes task instructions alone. All twelve endpoints retain the initial repair text and contain no exemplars.

Table 7: Changed fields at the six original terminal pairs. L denotes the lineage and S the audit-selected single edit. The last column identifies the selected zero-based bank row. Every S has repair cap 0 and temperature 0. Neither arm retains exemplars.

Seed	L text fields	S text field	L repair	L temp.	S row
20260821	system, instructions	instructions	1	0	142
20260828	system, instructions	instructions	1	0.2	80
20260835	system, instructions	instructions	0	0	354
20260849	system, instructions	instructions	0	0	156
20260856	system, instructions	instructions	0	0.2	143
20260863	system, instructions	instructions	1	0	80

Applying each of the 524 bank patches to M_0 and comparing the resulting specifications finds no exact structural match to any of the six L endpoints. The successive edits therefore produce combinations outside the bank's single-patch states. Structural difference does not establish functional dependence, a gain from combination or necessity of the historical path. The bank was not constructed to supply proposals that require multiple dependent edits before becoming useful.

A fixed example pair. We choose the smallest seed, 20260821, without reference to its score. Its L endpoint replaces the system prompt with:

> You are an expert Python programmer. Always ensure functions match expected parameters and handle edge cases.

and the task instructions with:

> Write a single self-contained Python function that solves the task. Use the exact function name and parameters shown in the example test. Output ONLY a Python code block, no explanations. Carefully handle edge cases and validate inputs.

It also sets the repair cap to 1, leaving the other initial fields unchanged. The paired S endpoint retains the initial system prompt and replaces only the task instructions, using bank row 142:

> Write a self-contained Python function that solves the task. Use the exact function name and parameter signature shown in the example test. Include helper functions if needed. Output ONLY a Python code block, no explanations.

These are endpoint texts, not a reconstruction of the intervening edit path.

E Pre-Registration and Deviations of the Post-Selection Audit

E.1 Registration, chronology and terminal attribution

Timeline. Estimands, thresholds and the single confirmatory hypothesis of §4.2 were written before any new sealed evaluation. Before these evaluations, algebraic and power checks found that one registered criterion was an identity and several rejection criteria were unreachable at the measured noise level. The protocol was amended to use seed-pooled inference for terminal outcomes, one confirmatory test and repeated singleton/terminal measurements before the first new sealed value. The registration, amendments and outcome-dependent deviations are retained in the research archive; App. H describes the completed internal checks and their limits. The 474-item reserve is shared by every cell. It was read by two infrastructure smoke runs before the pilots, the one-shot unlocks of the three pilots, the six W2 seeds, the three W2b and three W2c seeds, the replay batches of this audit, the pre-registered consumed-bandwidth and non-persistent-baseline cells, the two completion seeds, the PACE arm, and the later completion measurements reported below. Each lineage's original unlock is one-shot, and no lineage was optimised against a reserve value.

The confirmatory test. The confirmatory test concerns the pooled interaction term T_{int}. A positive lower bound would have established synergy, while an interval within $\pm 0.5\, G_{\mathrm{total}}$ would have

established additivity. The observed interval, $[-1.55, 0.38]$ pp, satisfies neither criterion. It excludes positive interaction above $+0.38$ pp, but it does not establish additivity on the negative side. We treat the result as unresolved. The pooled interaction is negative in five of six seeds. The system-prompt×instructions term is negative in four of the six arm-A seeds.

Deviations. (i) Three of the four numeric sub-conditions that the amended pre-registration attached to "does this section add information not derivable from the generation-level curve" failed; only the archived numerical interval-separation condition (`criteria.c`) passed. That recorded flag does not rule out correlated service noise. Two failures concerned effect-size or scale thresholds set without a power calculation (a two-sided additivity box narrower than the interaction interval, and a compaction size criterion that measures the length-penalty scale), and one concerned a magnitude condition (audit-side layers at least twice the sealed gain). The registered rule required us to terminate the standalone project from which this section originated and fold its material into this paper. The verdict, rule, and reasons remain recorded, and no threshold changed after the results. (ii) The gate-side statistic was changed from tenure-average regression to the first-remeasurement estimator ρ_{first} after the tenure average was seen, because the final tenure round is conditioned on being beaten. All three estimators are retained in the saved analysis tables (0.68, 0.33, and 0.560).

E.2 Measurement and service sensitivity

Original measurement. The original audit ran 457 evaluation batches, comprising $216{,}618$ task verdicts, on one pinned service. These include repeated baseline and factorial measurements, rather than 457 distinct states. We scored the baseline three times per seed, giving a serving paired-difference s.d. of 0.39 pp. We also evaluated the baseline and terminal states in two additional seed contexts, giving a context s.d. of 0.47 pp. Per-item disagreement between the old and new services ranged from 0.2 to 3.8%, with a mean of 1.4% and one cell above the $2 \times p_{\mathrm{flip}}$ bound. Mean shifts were 0.0 ± 0.6 pp, and none exceeded 3 pp. The pre-registered drift criterion also included a paired ± 1 pp equivalence test. This test was not met in 41 of the 52 comparisons. With a single old-service evaluation per checkpoint, its interval is about ± 1 pp wide by construction, so we recorded the criterion as failed rather than re-scoring it. The replicate flip rate p_{flip} comes from 968 same-round replicate pairs in the sealed audit records. The factorial baseline vector used the item-paired mean of three replay baselines in 17 cells, the single same-run replay baseline in the three W2b arm-B cells, and the existing single sealed baseline in the six W2c cells. These sources matched the estimator's baseline vector in all 26 cells; the factorial mask-zero invocation was used only to assess baseline noise. Every subset carried only the selected components' terminal values, and the interaction identities held in every cell. In registration order, the system-prompt×instructions pairwise terms for the W2 arm-A seeds are 0.00, $+0.84$, -0.91, -0.21, -1.13, and -1.34 pp.

Existing-data service sensitivity. All eight registered W2 seeds remain the primary analysis. Six original seeds and two completion seeds differ in both batch and documented service version, so their contrast cannot isolate a service effect. For both arms of each original seed, saved replay verdicts cover g_0–g_5 on the same 474 tasks, with matched specification hashes and seed contexts. Those replay and W5 records share a server-tag digest. Original unlock rows, including the two completion seeds, omit that tag; a common-service assertion for all eight cannot be independently certified from them. Using only the six fully covered replay trajectories gives $\mathrm{dAUC}_{AB} = -0.42$ pp (95% t_5 CI $[-1.16, +0.33]$; sign-flip $p = 0.3125$), compared with -0.44 pp $[-1.45, +0.57]$ for the original measurements of those same six seeds. This is an exploratory remeasurement sensitivity, not a replacement primary analysis or an equivalence result. The endpoint calibration above does not certify generation-wise dAUC.

The replay covers six W2 seeds in both arms, the exploratory cells and the frozen-bank pilot, with three baseline readings in each of the 17 replayed cells on 474 items. The saved analysis tables contain exact Shapley values and sum to G_{total}. The first-remeasurement statistic ρ_{first} is the fraction of the accept-time increment lost at the next-round incumbent measurement, providing an unconditioned first remeasurement. Arm A contributes 165 events, while two final-round accepts have no later measurement. Arm B contributes 16 events. The displayed arm-specific 95% ranges use $2{,}000$ bootstrap resamples of paired event numerators and denominators pooled across the six lineages, without lineage clustering. They do not capture between-lineage sampling or shared trajectory/time dependence. The N0 reference conditions on one fixed baseline vector per cell and uses indepen-

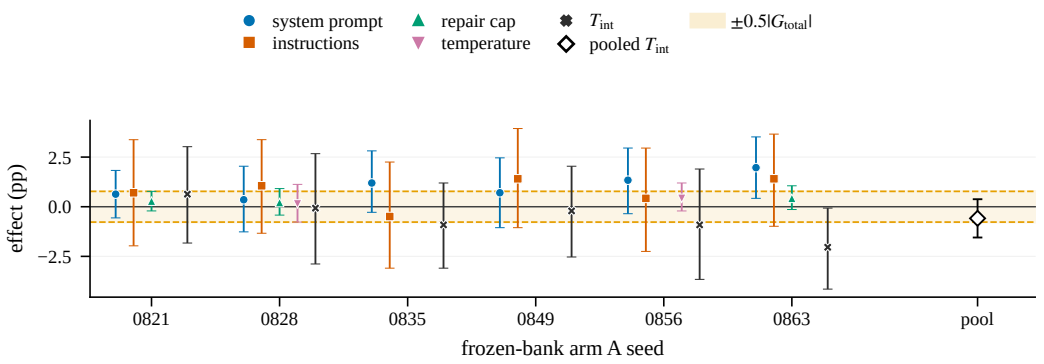

Figure 4: Terminal component effects in the six frozen-bank arm-A lineages. Singleton effects are measured relative to the baseline. T_{int} is the full-state gain minus their sum. The pooled interaction is -0.59 pp (95% CI $[-1.55, 0.38]$). The interval neither establishes positive interaction nor satisfies the registered additivity criterion. This decomposition concerns surviving terminal components and does not measure the causal contribution of the editing history.

dent item-flip draws for incumbent and candidate measurements across rounds; gate selection still induces dependence in the resulting statistics. It uses 2,000 simulation repeats per cell and scenario, with the estimated p_{flip} (approximately 0.018) and its Clopper–Pearson upper bound. For each scenario, the six cell-specific repeat-interval endpoints are averaged separately; the displayed range is the envelope across the two scenarios. This is a simulation summary, not an empirical service-noise confidence or prediction interval, nor a pooled interval with established coverage. The three exploratory W2b lineages record 25–31 pp of accept-time audit credit but finish 1.05, 1.05 and 2.95 pp below their reserve baselines.

Cross-fitted compaction selects components on one hashed lockbox fold with length penalty $\lambda = 1$ pp per 10^3 characters, then reports on the other. Its estimands and thresholds were fixed before the lockbox values for compaction were seen. The terminal decomposition removes components from the final harness and does not establish positive interaction. It cannot identify the trajectory's causal role or whether discarded edits helped the search reach that harness.

For the audit-rise estimate, terminal-state incumbent readings are averaged when available; otherwise the same frozen estimator uses the last ten incumbent rounds, which can include earlier states. Two of the six original lineages require this fallback (App. E.3). The separately reported last-ten-incumbent audit rise is 2.81 pp ($[2.05, 3.57]$). Per seed, the remeasured audit rise exceeds sealed gain by $+1.03$ pp ($[+0.02, +2.03]$).

Arm A accepts 15–44 candidates per production seed, compared with 14 in the pilot, while its remeasured audit rise is 2.57 pp ($[2.00, 3.15]$). The lockbox rises with it. Summed accept-time credit averages 25.17 pp per arm-A seed ($[14.71, 35.63]$), while the remeasured rise is 2.57 pp ($[2.00, 3.15]$). At accept level, first-remeasurement regression is $\rho_{\text{first}} = 0.560$ ($[0.432, 0.703]$), numerically below the independent-flip simulation reference 0.766 ($[0.763, 0.769]$). The margin-5 gate's event-pooled ratio is numerically smaller, at 0.253 ($[0.106, 0.411]$). These ranges have different constructions; their separation does not exclude correlated service noise.

E.3 Completion of the registered sample

Completion of the registered seeds (pre-registered). Seeds 20260842 and 20260870 were run after the reviews with the original configuration and bank. Their unlocks used the newer service, whose drift against the old one is measured above. Per seed, dAUC_{AB} is -0.34 and $+1.30$ pp. The eight-seed values in registration order are -0.88, -2.05, -0.29, -0.34, -0.33, $+0.75$, $+0.18$, $+1.30$ pp, with mean -0.21 pp ($[-1.06, +0.64]$), sign test $p = 0.73$ and permutation $p = 0.58$. The completion registration had designated the six-seed result as primary before the two seeds were unlocked. Because that subset followed a data-dependent cut, the eight-seed registered sample was

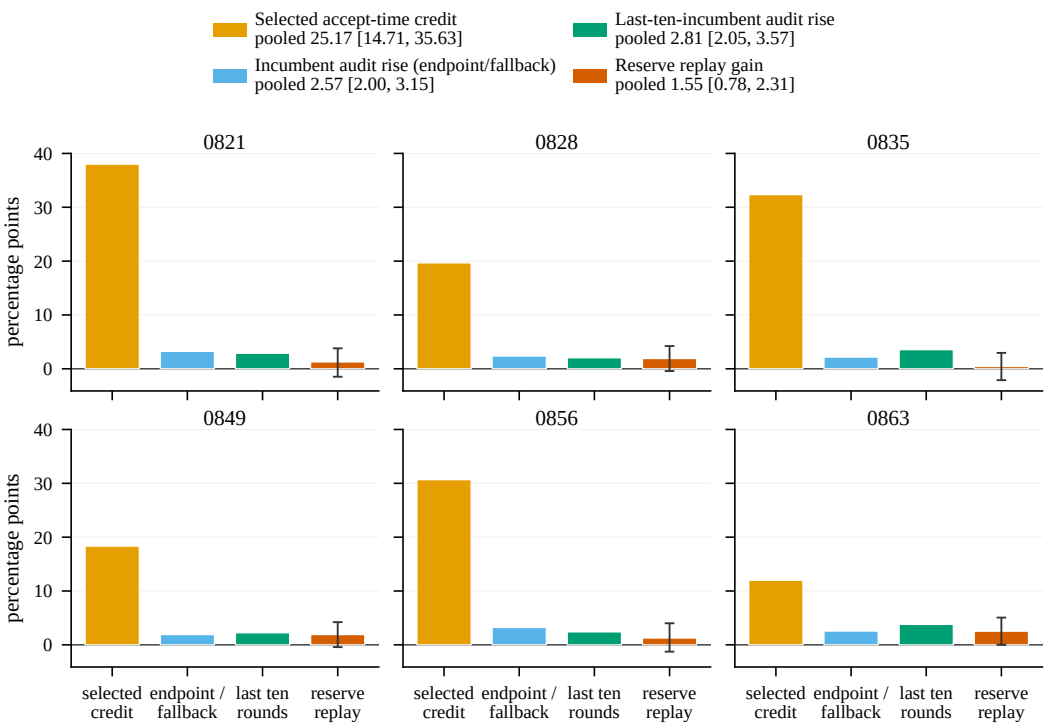

Figure 5: Four estimands shown side by side, with different comparators and evaluation units rather than a causal decomposition: per W2 seed (arm A), selected accept-time credit, incumbent audit rise (endpoint or last-ten-incumbent fallback), last-ten-incumbent audit rise, and reserve replay gain. The second estimand uses the fallback in two lineages; the third uses last-ten-incumbent measurements in all six. Reserve replay gains have item-paired 95% intervals. The legend gives six-seed pooled values with t_5 intervals.

restored as primary after review, with six-seed values retained as interim. The change is logged as a deviation and neither analysis rejects H1.

In the six-seed exploratory analysis, arm-level gap AUCs average $+2.0$ pp ($[+0.9, +3.1]$) for A and $+2.4$ pp ($[+1.3, +3.5]$) for B. Candidate means, failure identities and excerpts never reach proposal generation in this frozen bank. Both arms receive selection outcomes only, at most $\log_2(K+1)$ bits per round.

The six interim contrasts include four negative and two positive values, with a mean of -0.44 pp (95% t interval $[-1.45, +0.57]$). The exact two-sided sign test gives $p = 0.69$, and the sign-flip permutation test of the mean gives $p = 0.34$. At this spread, the sign test has 69% power for a 1.5 pp effect. Across-seed s.d. is 0.96 pp, and no single seed carries either sign. The pilot's $+3.1$ pp lies outside the interim six-seed range of -2.05 to $+0.75$ pp. The banks differ, so this comparison does not separate bank effects from seed effects.

Saved-data completion sensitivity. The two later seeds have 29 and 31 accepts and cumulative selected credit of 25.67 and 22.67 pp. Applying the same frozen audit-rise estimator gives 4.67 and 2.90 pp. Neither has a post-final-accept incumbent reading: both use the estimator's last-ten-incumbent-round fallback, as do original seeds 20260828 and 20260849. Thus these values are endpoint proxies, not direct remeasurements of each terminal specification. Across all eight, mean selected credit is 24.92 pp and mean incumbent audit rise is 2.88 pp; pooled first-remeasurement regression is 0.5533 over 223 eligible accepts out of 227. These exploratory saved-record summaries retain the original six-lineage results and the two registered eight-seed primaries. Audit-rise proxies remain distinct from the subsequent reserve measurements below.

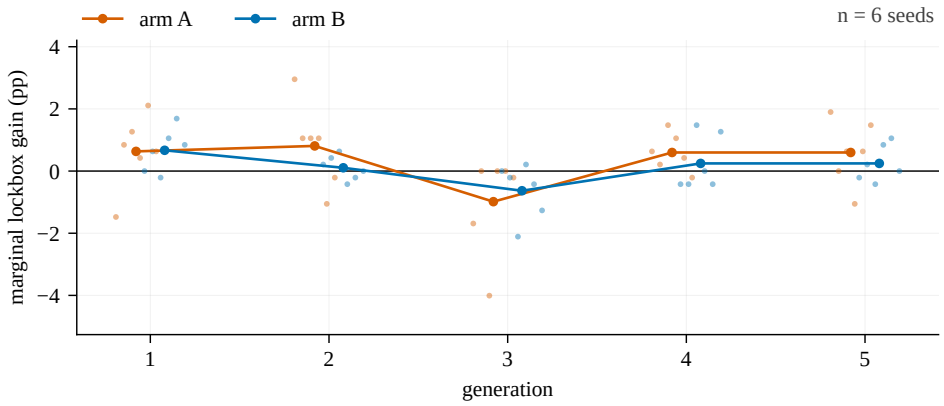

Figure 6: Per-generation marginal lockbox gain across six seeds. Sealed gains arrive mainly in the first two generations, with later increments of varying sign, and lie within noise of zero by generation 3. Generation-1 means are $+0.6$ pp (A) and $+0.7$ pp (B). The fresh tranche does not restart them. This is consistent with H2's 3–5 generation criterion. No equivalence test was registered.

Subsequent measurement completion. We later measured replay, the audit-selected single-edit baseline and terminal factorial contrasts for both completion lineages. Their search histories and candidate stream were fixed; no new search lineage was generated. The original eight-seed search was reduced after four dAUC values had been observed. The post-selection audit was then registered on the six existing lineages, before the remaining two searches completed. The present extension adds their missing measurements after the reserve had already been unlocked. It preserves the original six-lineage analysis and is exploratory.

The completion evaluation used Qwen3-4B-Instruct-2507 with vLLM 0.18.0, four data-parallel H20 replicas and context limit 16,384. Thirteen model and tokenizer asset files matched the available earlier local assets; this does not authenticate the historical serving processes. Each lineage's W5 selection evaluated the original 400 candidates over 100 rounds, pairing each round with M0 on the 300 audit tasks. Both NP0–NP5 state lists were frozen before the first new reserve request. The 18 measurement units contain 105 distinct-within-unit specification evaluations on all 474 reserve tasks (49,770 verdicts). Identical specifications share one measurement only within a unit, with all labels retained; fixed repeats remain separate. Full and singleton factorial states and M0 each retain three measurements under the original estimator.

Table 8 combines the original six with the two later measurements. The eight-lineage replay gain is 1.37 pp ($[0.78, 1.96]$), and the paired lineage-minus-single-edit difference is -0.03 pp ($[-0.56, 0.51]$; sign-flip $p = 1$). The added single-edit gains are equal at 1.27 pp; two equal values do not imply zero measurement uncertainty. All eight-lineage intervals use t_7, and sign-flip calculations enumerate 256 assignments. The terminal interaction is -0.44 pp ($[-1.17, 0.29]$, $p = 0.219$). These measurements retain the finding of improvement over M0 without an established multi-edit advantage or positive terminal interaction. Evaluation epoch remains confounded with lineage batch (Table 9); this extension is neither a fresh reserve test nor a same-epoch eight-lineage replication.

The added measurements used 349,775 physical sends, 50.49M input and 42.91M output tokens, and 11.74 GPU-hours including loading and cleanup. All sends returned successful HTTP responses and usage was recorded for each. These are evaluation costs, separate from the historical search costs. No recorded infrastructure failure was scored as a task failure; 131 program timeouts follow the fixed task-execution limit.

Deterministic sensitivity of the W4 test. For the eight saved R5−B MBPP dAUC contrasts, adding the same offset $\delta \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$ pp to each seed and rerunning the registered two-sided sign-flip statistic yields $p = 0.3125, 0.9297, 0.2578, 0.0625, 0.015625, 0.0078125, 0.0078125$. The first rejecting grid point is $\delta = 2$ pp (shifted mean 1.54 pp). This is deterministic score arithmetic on rounded seed summaries,

Table 8: Original and completed post-selection measurements (pp). The first two rows use audit records; the other rows use reserve measurements. Original-six and pooled-eight entries are means with 95% across-lineage t intervals; the two added columns are individual existing lineages. The eight-lineage extension combines measurement epochs. Estimands and comparators follow Table 2; the audit-rise row retains its endpoint fallback.

Quantity	Original six	Seed 20260842	Seed 20260870	Pooled eight
Selected credit	$25.17\ [14.71, 35.63]$	25.67	22.67	$24.92\ [17.83, 32.00]$
Incumbent audit rise	$2.57\ [2.00, 3.15]$	4.67	2.90	$2.88\ [2.15, 3.60]$
Replay lineage − M0	$1.55\ [0.78, 2.31]$	1.05	0.63	$1.37\ [0.78, 1.96]$
Single edit − M0	$1.44\ [0.46, 2.43]$	1.27	1.27	$1.40\ [0.73, 2.06]$
Lineage − single edit	$0.11\ [-0.62, 0.83]$	-0.21	-0.63	$-0.03\ [-0.56, 0.51]$
Terminal interaction	$-0.59\ [-1.55, 0.38]$	-0.56	0.56	$-0.44\ [-1.17, 0.29]$

Table 9: Per-lineage measurement provenance and reserve contrasts (pp). E1 denotes the original post-selection evaluation epoch; E2 denotes the later four-H20 completion epoch. F marks the last-ten-incumbent fallback for audit rise; it does not alter these reserve contrasts. Rows are eight existing search lineages, not new searches or independent task repeats.

Seed	Epoch	F	Lineage−M0	Single−M0	Lineage−single	Interaction
20260821	E1	no	1.27	1.05	0.21	0.63
20260828	E1	yes	1.90	2.53	-0.63	-0.07
20260835	E1	no	0.42	0.84	-0.42	-0.91
20260842	E2	yes	1.05	1.27	-0.21	-0.56
20260849	E1	yes	1.90	0.63	1.27	-0.21
20260856	E1	no	1.27	0.84	0.42	-0.91
20260863	E1	no	2.53	2.74	-0.21	-2.04
20260870	E2	yes	0.63	1.27	-0.63	0.56

not a power estimate, a simulated population or a retained-leak positive control. It cannot qualify the loop's ability to preserve contamination.

E.4 Assigned disclosure and uptake

Consumed-bandwidth 2×2 (pre-registered). Eight new seeds ($20261001 + 7k$, $k = 0..7$), with four arms per seed launched together. All four fixed conditions were run for each seed; treatment labels and launch order were not independently randomized. Proposer mode uses $K{=}4$ for five generations of twenty rounds, with identical forced-operation rotation across arms. Every arm uses the same winner rule, the largest paired net wins, and accepts when net wins reach its margin. A is the greedy rule of §4.4, and B is its margin-5 gate. The primary contrast is dAUC_{R5-B}. The secondary contrasts are dAUC_{A-B1}, dAUC_{B1-B}, and dAUC_{A-R5}. We report per-seed item-paired bootstrap intervals and across-seed t intervals for scale. The exact sign-flip permutation test (minimum $p = 0.0078$) is the confirmatory test, with the sign test alongside. Its exactness requires the null assumption that the joint seed-level contrast distribution is invariant to independent sign flips, rather than following from treatment-assignment randomization. The lockbox was unlocked once per seed, in completion order, with the analysis code frozen before the first unlock. The cell was registered before any run and had no deviation. Per seed, dAUC_{R5-B} = $+0.04$, -1.72, -2.20, -0.14, $+1.17$, $+0.54$, -1.74, $+0.37$. dAUC_{A-B1} = -3.89, -0.43, -1.19, $+1.37$, -1.74, -3.28, $+0.46$, -0.59. dAUC_{B1-B} = $+1.45$, -0.30, -1.35, -1.39, $+1.41$, -0.76, -1.69, $+2.86$. dAUC_{A-R5} = -2.49, $+0.99$, -0.34, $+0.12$, -1.50, -4.58, $+0.51$, $+1.90$. dAUC_{AB} = -2.45, -0.73, -2.54, -0.02, -0.34, -4.04, -1.23, $+2.27$ pp. The secondary means are $-1.16\ [-2.65, +0.33]$ ($p = 0.13$), $+0.03\ [-1.37, +1.42]$ ($p = 0.95$) and $-0.67\ [-2.43, +1.08]$ ($p = 0.44$) pp. The registered interaction $\mathrm{dAUC}_{A-B1} - \mathrm{dAUC}_{R5-B}$ is -0.70 pp. Sealed change $g_0 \to g_5$ per arm (mean, 95% t): A $-1.79\ [-2.99, -0.59]$, B $-0.63\ [-1.63, +0.36]$, R5 $-0.61\ [-1.39, +0.17]$, B1 $-2.08\ [-3.21, -0.95]$ pp. The paired sealed-side differences are R5−B $+0.03\ [-1.36, +1.42]$ and A−B1 $+0.29\ [-0.96, +1.54]$ pp. At the observed seed-level spread the cell resolves effects of about 1.5 pp with 80% power.

The exploratory online-proposer cell W2b was launched after H1 was known. Forced-operation rotation fixes the pilot's mode collapse, with $120/120$ unique proposals in dry runs. Its three seed contrasts are $\mathrm{dAUC}_{AB} = +0.77$, -1.02 and $+3.15$ pp, a lead in the direction predicted when the proposer can read the disclosure. The like-for-like joint contrast in the registered follow-up is $\mathrm{dAUC}_{AB} = -1.13$ pp ($[-2.74, +0.47]$), with $p = 0.12$. It lies below the exploratory mean of $+0.97$ pp, although this comparison does not test a difference between cells.

Disclosure uptake in the 2×2 (pre-registered, transcript only). Rich-view arms receive, each round, the candidates' audit means and the failing audit tasks' prompts and errors in the proposer prompt; commit-view arms receive the accept bit. Proposer calls are seeded identically across arms at temperature 0.8, so raw proposal divergence is not an uptake measure and was not used. U_1 is the share of proposals whose patch text contains, verbatim after whitespace normalisation, an audit task prompt disclosed to the same arm in an earlier round (for commit-view arms, any prompt disclosed to the seed's rich arms, as the false-positive floor). Disclosed prompts shorter than 20 normalised characters are ignored as a substring-collision guard. No disclosed audit prompt is that short. The five documentation-level deviations logged in the precommit are the strictly-earlier-round disclosure set, this filter, whole-patch matching when classifying exemplar proposals, the skipping of all-tied rounds, and macro names; none changes a number. U_2 is the mean over next-round proposals of the token-set Jaccard similarity to the previous round's best-scoring candidate minus that to its worst-scoring one, computed for every arm from the audit records. Over the eight seeds, U_1 is 850/5800 (14.7%) for rich-view proposals and 0/5835 (0.0%) for commit-view proposals, higher under the rich view in 8 of 8 seeds. Denominators count parseable proposals: of the 6400 candidates per view, 600 (rich) and 564 (commit) failed to parse as patches, most of them in the margin-1 arms, and are excluded, as is one commit-view candidate that the proposer wrapper reformatted. The rich-minus-commit difference in U_2 over the two fixed-margin pairs is -0.070 ($[-0.096, -0.045]$), positive in 0 of 8 seeds (rich +0.011, commit +0.081). Of the `add_exemplar` proposals, 850 of 1483 under the rich view and 0 of 1546 under the commit view reproduce a disclosed audit prompt, while 25 and 719 respectively reproduce a dev-evidence prompt. The rich view therefore substitutes failing audit prompts for dev prompts as the exemplar source. A registered descriptive follow-up traces the copies through the gate: 4 of 364 accepted rich-view proposals (1.1%) carry a copied audit prompt, against 15.6% of rejected ones, and 4 of the 6 exemplars in the sixteen rich-view terminal specifications are copied audit prompts (commit view: 0 of 8). No terminal instructions or system prompt contains one. The registered rule (copying in at least six seeds and score following in at least seven) was met on copying only, so the main text reports both figures and keeps the assigned-disclosure wording. The copied audit text dominates the rich-view proposals' token sets, which is one reason their U_2 is smaller; this is an interpretation, not a change to the rule.

E.5 Second reserve and positive controls

Never-read second reserve (pre-registered, one global unlock). The reserve is the 164 canonical HumanEval tasks converted to the pool schema (one test string per task: the HumanEval `check` function followed by its call; $164/164$ reference solutions pass). No run, smoke test, analysis or author of this project had evaluated anything on it; HumanEval itself is public, so "never read" is a property of this project and not of the model's pretraining. The shared baseline scores $141/164$ (86.0%) on it, so one task is 0.61 pp. The registered batch lists every state of the registered cells: W2 (eight seeds, arms A and B, g_0–g_5), W4 (eight seeds, four arms), PACE (six seeds, arm P), the W5 non-persistent states, the frozen-bank pilot and, by an amendment logged before the read, the leak arm below: 398 labels, 193 distinct specifications, each scored once at its own specification temperature under the fixed seed context on one pinned service, with no output read until the batch's completion marker was written. The across-lineage intervals treat this realized shared score table as fixed; they do not separately quantify uncertainty from regenerating each distinct specification or sampling new tasks. Per seed, dAUC_{AB} on the reserve is $+0.52$, -3.68, $+0.36$, $+1.04$, -3.28, -0.71, $+0.01$, $+2.55$ pp (mean -0.40 $[-2.17, +1.38]$, sign test $p = 0.73$, permutation $p = 0.63$; six-seed interim -1.13 $[-3.10, +0.84]$). dAUC_{R5-B} is -2.02, -2.29, $+0.82$, -1.00, -0.12, -0.11, -1.64, -1.23 pp (mean -0.95 $[-1.84, -0.05]$, sign test $p = 0.07$, permutation $p = 0.04$). The registered rule assigned sentences only to a positive significant contrast and to a non-significant one; this negative significant contrast is reported without a rule sentence. Secondary contrasts: A−B1 -0.74 ($p = 0.34$), B1−B -1.19, A−R5 -0.98, A−B -1.93 ($[-3.28, -0.58]$, $p = 0.02$), interaction $+0.21$ pp. Reserve change $g_0 \to g_5$: W2 A $+1.83$ $[+0.52, +3.14]$, B $+0.53$ $[-0.54, +1.60]$; W4 A

-0.53, B $+0.30$, R5 $+0.15$, B1 -0.99; PACE P $+0.91$ pp, with $\mathrm{dAUC}_{P-B} = -0.57$ ($p = 0.31$). The W2 arm-A terminal gain transfers as $G_F = +2.13$ pp ($[+1.01, +3.26]$) against $+1.65$ pp for the same six states at their original MBPP-lockbox unlock (difference $+0.48$, $[-0.99, +1.96]$); the replay estimate of those states reported elsewhere is 1.55 pp, against which the difference is about $+0.6$ pp with an interval that also covers zero; the paired lineage minus single-edit difference is $\Delta_5 = +0.81$ pp ($[-0.31, +1.93]$, $p = 0.19$); the frozen-bank pilot's dAUC_{AB} is $+3.57$ pp on one seed. Observed per-seed s.d. of the primary contrasts on this second reserve: W2 2.12, W4 1.07 pp against the planning bound of 1.1 pp. The MBPP spreads are 1.02 pp (W2) and 1.25 pp (W4); they must not be confused with these transfer values: at the observed spread the cells resolve about 2.4 pp (W2) and 1.2 pp (W4) with 80% power under a t approximation, and about 3.1 pp for the six-seed leak arm under the registered exact sign-flip test, which rejects at 0.05 only when all six signs agree. The permutation p values of the two primaries are uncorrected. Holm adjustment gives 0.63 (W2) and 0.08 (W4). Two amendments were logged before the read (the leak arm's states; specification temperature), and the implementation's first launch was refused by its own guard before any pool data was read.

Table 10: Second-reserve contrasts decomposed into their audit-side and reserve-side components (pp, means with 95% t intervals across seeds). Audit side: $\mathrm{AUC}_X^{\mathrm{audit}} - \mathrm{AUC}_Y^{\mathrm{audit}}$; reserve side: the same on HumanEval; dAUC = audit side − reserve side. The negative disclosure contrasts arise on the audit side (rich-view and margin-1 arms' audit scores rise less), not from reserve-side differences.

Contrast	n	Audit side	Reserve side	dAUC
W2 A−B	8	$+0.15$ $[-0.40, +0.70]$	$+0.55$ $[-1.11, +2.20]$	-0.40 $[-2.17, +1.38]$
W4 R5−B	8	-0.67 $[-1.25, -0.10]$	$+0.27$ $[-0.34, +0.89]$	-0.95 $[-1.84, -0.05]$
W4 A−B	8	-2.33 $[-3.86, -0.79]$	-0.40 $[-2.28, +1.49]$	-1.93 $[-3.28, -0.58]$
Leak L−B	6	-0.24 $[-2.05, +1.56]$	$+1.40$ $[-0.17, +2.98]$	-1.65 $[-3.48, +0.19]$

Known-leak positive control (pre-registered). Arm L runs the W4 online-proposer configuration at margin 5 with a third disclosure view: the rich view plus, for every failing audit task listed, its reference solution, charged at 8 bits per character on top of the rich bound. Six seeds ($20261001+7k$, $k = 0..5$), paired with the W4 arms B and R5 of the same seeds; its states entered the single fresh-reserve batch by amendment. Per seed, dAUC_{L-B} = -3.75, -2.81, -1.26, -2.31, -1.05, $+1.29$ pp (mean -1.65 $[-3.48, +0.19]$, sign test $p = 0.22$, permutation $p = 0.13$) and dAUC_{L-R5} = -1.73, -0.52, -2.08, -1.31, -0.92, $+1.40$ pp (mean -0.86, $p = 0.16$). L's audit score rises by $+1.56$ pp ($[+0.08, +3.03]$) from g_0 to g_5 and its reserve score by $+1.42$ pp ($[-0.18, +3.02]$). The registered known-leak rule was not met, so the paper carries the registered fallback sentence (with "extreme" dropped, since the dose below does not support it; logged as a wording deviation). Dose and uptake (PC2, from the transcripts): L received 606 disclosures over the six seeds, with 12–16 distinct leaked solutions per seed (the per-seed distinct counts sum to 81); of its 2351 valid candidate events, 27 (1.1%) matched a disclosed reference after case and whitespace normalization and 126 (5.4%) matched a disclosed audit prompt, against 12.4% prompt copying in arm R5 on the same seeds; L accepted 24 edits and 0 of them carried a leaked solution or a disclosed prompt. Exposure is therefore confirmed, but persistent audit-specific contamination is not established, so the control cannot distinguish an insensitive gap statistic from exposure that produced no retained audit-specific change. A repair confined to one audit task, with other outcomes unchanged, cannot clear a margin-5 gate (App. B), so the control tested the gate and the instrument together; the same leak behind a margin-1 gate was not run. On 300 audit items, correcting one item is 0.333 pp; 12–16 distinct leaked solutions could contribute at most 4.0–5.33 pp through exact item-specific fixes at a terminal endpoint, even if all were retained and all previously failed. This is a terminal dose ceiling, not the baseline-adjusted across-generation dAUC or a demonstration of sensitivity to use beyond detected matches. The registered known-gain control was dropped by its own rule before the read: none of three candidate edits (one repair round, three dev exemplars, an instruction rewrite) moved the dev pool with an interval excluding zero (on the 197 dev tasks shared by all candidates: dev $+0.0$, $+0.0$, -1.5 pp; audit $+0.33$, -2.0, -0.67 pp; the repair edit had scored $+1.0$ pp $[0.0, +2.5]$ on the full 200-task dev pool in a first evaluation).

Post-review reference-copy trace. Actual proposer inputs contain ten active references in each of all 2,400 calls (24,000 display occurrences). There are 2,351 valid candidate events and 49 invalid responses. The matching rule lowercases and normalizes whitespace, searches for a complete previously disclosed reference or prompt excerpt of at least 20 characters, and excludes the proposal rationale. All 27 normalized reference-copy candidate events are exemplar additions. Joining their patches and state hashes to the paired 300-task audit vectors gives net wins from -28 to -8: none is the selected winner or accepted. The events cover only two target tasks in four seed–task combinations. Their target differences sum to $+1$, and differences on the other tasks to -403; these are sums over correlated events, not population effects or 27 independent samples. With the stored candidates, incumbents and scores held fixed, none clears even margin 1; this does not predict a rerun of the adaptive search under that gate.

Two cases selected by the first target pass and first target failure in (seed, g, r, c) order illustrate the join. In $(20261001, 5, 9, 3)$, the copied target changes from fail to pass, but overall net wins are -14 and the candidate loses. In $(20261015, 3, 19, 1)$, the target remains failed, net wins are -9, and another candidate wins and is accepted. The old archive lacks these candidates' complete solver requests, responses and final code, so the joins do not establish functional consumption or a mechanism for the target failures. None of 24 accepts or 60 saved checkpoints contains a detected disclosed reference; effects beyond these matches remain unmeasured. Table 11 separates this observed path from direct installation, which bypasses proposal and acceptance.

E.6 Direct-injection calibration on reused MBPP pools

Design and evidence unit. This prospectively fixed, exploratory first layer tests a legal persistent reference intervention through the original harness and scorer, bypassing proposal and acceptance. Six assignment/context blocks use $20260923 + 7k$, $k = 0{:}5$. Eligible audit references pass the original syntax, character-limit and CPU-scoring checks before hash ranking selects three per block; dev examples are matched by prompt-plus-reference character count with a fixed tie break. No model outcome selects examples, and no new reserve or test set is consulted. The original three-exemplar limit and 2,500-character limit per exemplar remain unchanged. Character matching does not imply equal token lengths.

The four states are the original M0, three matched dev exemplars (S0), one audit reference plus the corresponding two dev exemplars (D1), and three audit references (D3). S0 supplies real information and is not a zero-effect placebo. All conditions share one measured M0 baseline at g_0 within each block and pool. Injection occurs at g_1; each state remains fixed for five measurements through g_5. Every later measurement uses fresh responses on all 300 audit and 474 previously used reserve tasks. Thus $6(1{+}5{\times}4)(300{+}474) = 97{,}524$ task results form 252 pool vectors. The five measurements are repeated evaluations of fixed states, not adaptive search generations or independent search lineages.

Qwen3-4B-Instruct-2507 uses vLLM 0.18.0, transformers 4.57.6, PyTorch 2.10.0+cu128 and Python 3.13.9 on four H20 devices. The original renderer, code extraction and scorer are retained, with temperature 0, no repair, at most 1,200 output tokens and context limit 16,384. Condition order rotates by block and measurement; per-task seeds follow the original rule and do not depend on the condition. Each input plus its full output reservation is checked before sending.

Primary statistic and qualification. For block b and condition c, let

$$U_{b,c} = \frac{1}{5}\sum_{g=1}^{5}\big[(A_{b,c,g} - R_{b,c,g}) - (A_{b,0} - R_{b,0})\big].$$

The unique primary contrast is $U_{b,\mathrm{D3}} - U_{b,\mathrm{S0}}$, averaged across six blocks. We use a t_5 interval and all $2^6 = 64$ two-sided sign flips. Sign flips assume symmetry under the null; fixed seed labels do not establish randomization-based causal inference. Shared baselines, task counts and repeated measurements do not increase $n = 6$. The original composite working-signal rule has four conjunctive conditions:

1. Positive aggregate paired target net repairs for D3 versus S0: $+30$; satisfied.
2. Positive target net repairs for D3 versus M0: $+23$; satisfied.
3. At least two blocks with positive target net repairs against both: $5/6$; satisfied.

Table 11: Observed calibration stages from separate interventions. Known-leak counts describe correlated proposal events; direct-installation net counts aggregate six blocks and five measurements, not independent tasks. These records do not form a successful end-to-end positive control.

Stage	Saved observation	Remaining limit
Known-leak input	2,400 calls, each with 10 references	Exposure does not establish use.
Known-leak proposal	27 normalized reference-copy events among 2,351 valid candidate events; two tasks, four seed–task combinations	Uptake beyond these matches is unmeasured.
Known-leak selection	All 27 net wins are negative; zero winners and zero accepts	Solver consumption is not recorded; a new search trajectory is untested.
Direct installation, D3−S0	+30 target and −41 other-task net successes; aggregate qualification fails	Proposal and acceptance were bypassed.

4. Positive overall audit D3-minus-S0 mean: -0.122 pp; not satisfied.

These are study-specific qualification rules, not industry standards, significance tests or equivalence margins. The preserved status is `TARGET_SIGNAL_NOT_ESTABLISHED`. The separate execution status is complete. Only an established working signal together with a positive primary mean and two-sided $p < 0.05$ would qualify as a detected positive response.

Observed response. Table 12 gives the six paired block contrasts. D3 improves designated targets relative to both references: against S0 there are 30 wins and no losses over 90 target observations, whereas other audit items yield 343 wins and 384 losses over 8,910 observations. The +30 is a net count over five repeated measurements, not 30 independent tasks. The audit arithmetic is $(30 - 41)/9{,}000 \times 100 = -0.122$ pp. Reserve performance is 0.408 pp higher than S0, so the full primary gap is $-0.122 - 0.408 = -0.530$ pp, with 95% interval $[-1.723, 0.662]$ and sign-flip $p = 0.34375$. The task partition is a descriptive arithmetic decomposition, not an identified mechanism of harm to other tasks. Relative to M0, D3 changes audit and reserve performance by -3.289 and -0.563 pp; the reserve contrast against S0 therefore does not establish an improvement over the original agent. All nonprimary contrasts are descriptive.

Table 12: Direct-injection calibration. Six assignment/context blocks, not search lineages. Values in pp average the five paired measurements. Only the D3–S0 gap is primary; other contrasts are descriptive.

D3–S0 block	Audit	Reserve	Gap
1	−0.267	+0.759	−1.026
2	+0.133	−0.380	+0.513
3	+1.600	+1.646	−0.046
4	−0.333	−0.970	+0.637
5	+0.200	+2.574	−2.374
6	−2.067	−1.181	−0.885
Mean	−0.122	+0.408	−0.530
95% t_5 interval	$[-1.362, 1.118]$	$[-1.176, 1.992]$	$[-1.723, 0.662]$
Descriptive contrast	**Audit**	**Reserve**	**Gap**
D1–S0	−0.444	+0.464	−0.909
S0–M0	−3.167	−0.970	−2.196
D3–M0	−3.289	−0.563	−2.726
D3–D1	+0.322	−0.056	+0.378

Interruption, recovery and cost. The initial execution stopped on a bounded SQLite snapshot timeout after 56 complete pool vectors, preserving 21,962 task scores. Recovery kept those scores, rescored six already saved successful responses without new generation, and resent four interrupted requests whose responses were lost. An explicit authorization raised the physical-send cap from

97,524 to 97,528 for those four requests; the original absolute time deadline and other resource caps remained in force. Recovery repaired snapshot handling without changing the model, scorer, conditions, seeds, statistical rules or previous scores. The initial attempt remains incomplete in its own archive; the recovered fixed matrix is complete. The two execution segments are one logical experiment, not independent samples.

The final journal records 97,528 physical attempts, 97,524 successful responses, six saved-response recoveries and four original attempts with unknown token usage. Known usage is 27,769,104 input and 9,076,157 output tokens; budget charges are 27,770,444 and 9,080,957, retaining reservations for unknown usage. Full actual token totals remain unknown. Cumulative occupancy, including the interrupted segment, loading and cleanup, is 2.709 GPU-hours under the original 12 GPU-hour cap; the owned devices were released. These calibration costs are separate from historical search and deployment costs. Retained protocol and recovery projections document these checks; this appendix reports their experimental design and aggregate costs.

Interpretation. Local target improvement is observed, but the predeclared overall positive calibration requirement is not met. This does not establish instrument failure, absence of contamination in W4, harmless disclosure or equivalence. The intervention was installed directly and establishes no autonomous proposal or acceptance result. A second layer using the real proposer and gate was not executed; no dose, seed or qualification rule was changed after observing these outcomes.

E.7 PACE-INSPIRED RUNNING-MAXIMUM GATE

PACE-style running-maximum gate (pre-registered). Arm P reuses the W2 configuration, bank, and per-seed candidate order, with a commit-only view and a PACE-style gate. For each candidate the audit items are placed in a fixed random order derived from the run seed, generation and round. Walking that order over the discordant items only, with w candidate wins among n discordant items, the statistic is the running maximum of $e_n = 2\int_{1/2}^{1} p^w(1-p)^{n-w}\,dp\,/\,(1/2)^n$, a mixture-betting statistic. The supermartingale property that PACE derives needs a conditional win-probability bound along the order, which the finite-population null of a fixed shuffled item set does not supply, so no anytime-valid guarantee is claimed. The winner of a round is the candidate with the largest running maximum, and it is committed when that maximum reaches $1/\alpha$ with $\alpha = 0.05$. This is an operational threshold: without an established per-candidate null guarantee, α cannot be interpreted here as a bound on a candidate's false acceptance probability. The four candidates per round are uncorrected, and no round-level error bound is claimed. The running-maximum threshold is computed offline over all 300 items in the fixed seeded order, with no early stopping, so no evaluation cost is saved. The P arm's generation-end states were read once from the lockbox through the guarded re-evaluation path with this registration's digest, and every lockbox value in the contrast comes from the one newer service. Per seed, dAUC_{P-B} = -0.66, $+1.08$, -1.81, -0.60, $+0.22$, -0.79 pp and dAUC_{P-A} = -0.92, $+2.46$, -1.69, -0.56, $+0.19$, $+0.47$ pp. Sealed gains are P $+1.37$ $[+0.59, +2.15]$, A $+1.55$ $[+0.78, +2.31]$, B $+0.81$ $[+0.22, +1.39]$ pp. Accepts total 7 (P), 167 (A) and 16 (B). The share of accepts whose next incumbent measurement did not exceed the accept-time incumbent is 0/7 for P, 64/165 for A and 0/16 for B. The gate crossed its threshold 7 times in 2400 candidate tests, with 23.6 discordant items per test on average. Registered before any run and completed without deviation. Against the greedy arm, the mean contrast is -0.01 pp with sign-flip $p = 1.00$.

A mean 38% of greedy accepts per seed regress at the next measurement. The per-candidate mixture-betting statistic visits the discordant paired items and commits when its running maximum reaches $1/\alpha$, $\alpha = 0.05$, as specified above.

E.8 SINGLE-EDIT BASELINE, COST AND COMPACTION

Non-persistent baseline (pre-registered). For each W2 seed the recorded candidate stream (4 per round, 100 rounds) was replayed against the baseline state under each round's own seed context, paired with the baseline evaluated in the same context. The non-persistent state after generation g is the candidate with the largest net wins among all rounds $\leq 20g$ (ties to the earliest, with the baseline used if none is positive), selected from audit data alone before any lockbox read. Because this argmax ranges over 400 candidates while the lineage's gate decides one round at a time, the baseline

carries the larger winner's curse. The comparison is conservative against persistence only to the extent that curse is small. Its lockbox scores and the lineage's generation-end states were evaluated on the same service under the fixed context $(\text{seed}, -1, 0)$. The paired difference, its t interval, the exact sign-flip permutation test and the sign test are reported per generation in the saved cost records. Search cost sums every recorded call's tokens. It excludes shared bank generation and dev health screening, whose token counts were not logged. Generation used 986 drafts by Qwen3-32B and Qwen3-4B (950 initial plus 36 top-up). Including these shared costs would raise H^*. Per-task cost uses the measured tokens of the checkpoint audit evaluations. The pre-registration specified re-evaluation tokens or a character estimate. The measured value was preferred and the method recorded. The break-even scale is $H^* = S_A/(2c(M_0) - c(M_5^A))$, the deployment size beyond which the persistent state's search cost plus extra per-task cost falls below one additional baseline sample per task. The oracle pass@k bound samples the baseline eight times at temperature 0.7. It upper-bounds any test-time selector and is not a deployable policy. Observed: pass@1 66.7%, pass@2 68.4%, pass@4 70.0%, pass@8 71.3% (item-paired 95% intervals about ± 4 pp). An oracle best-of-two already reaches the evolved state's lockbox score. The best single edit gains $+1.44$ pp on the lockbox ($[+0.46, +2.43]$), positive in $6/6$ seeds. The paired arm-A lineage minus single-edit difference is $+0.11$ pp ($[-0.62, +0.83]$, sign-flip $p = 0.84$), while the corresponding arm-B lineage minus single-edit difference is -0.63 pp.

Cross-fitted compaction keeps 0.5–3 of the 2–4 surviving components per seed. Its fold-2 loss is -0.23 pp ($[-0.67, 0.20]$, 90% CI), equivalent to zero within $\pm 0.5 G_{\text{total}}$. Overall, 87–98% of accepts are overwritten or unselected, including each seed's most-credited accept (Fig. 7).

Search costs 48–57M tokens per lineage excluding shared bank generation, and the evolved state uses $1.51\times$ the baseline's tokens per task. Against two baseline samples per task, search amortises after $H^* \approx$ 334k–646k tasks. This is a token-cost crossover, not an accuracy-adjusted break-even.

Incomplete online-pool extension. A post-hoc extension attempted to compare the historical W4 lineage endpoints with single recorded patches selected from the same online-generated candidate pools. Candidates were applied independently to M_0 and evaluated on audit tasks; the single-edit route did not generate its own proposals. The run stopped at its prespecified dispatch cutoff within a seven-hour, eight-GPU allocation, with $756/800$ audit rounds completed across the eight historical lineages. It did not reach the joint selection freeze or perform any terminal evaluation. No primary or secondary effect estimate is available. Saved partial records are retained separately; they are not pooled with completed studies or treated as a shorter completed comparison.

E.9 Fixed-bank search under a second model and serving configuration

Design and scope. This prospective exploratory extension reruns lineage search (L) and audit-selected single-edit search (S) with the official post-trained Qwen3.5-4B release (Qwen Team, 2026). It retains the 524-patch bank, 300 audit tasks, 474 previously used reserve tasks and eight paired algorithm seed identities. Each route uses its own new-model audit outcomes; old selected states are not imported. Five generations of twenty rounds schedule four patch slots per round, with the original seed permutations and generation-3 tranche switch. L applies patches to its incumbent and retains the arm-A net-win gate (≥ 1), original tie order, applicability checks and within-round state deduplication. S applies every patch to M_0, selects the global maximum audit net win with earliest ties, and retains M_0 if the maximum is nonpositive. No adaptive proposer runs in this extension.

All eight paired searches completed and their final states were frozen before terminal evaluation read the reserve. Each seed measured M_0, L and S separately on all 474 tasks: 24 terminal measurements and 11,376 binary cells, with no missing cells or within-seed response aliases. Interrupted-prefix rows are not additional measurements. These reused seed identities are not pooled with the old cohort as $n = 16$, and the reused reserve is not a new blind test. The old six-lineage equal-scoring comparison retains its original scope.

Table 13: Qwen3.5 terminal successes out of 474 per state and realized candidate scoring. Both routes schedule 400 patch slots per seed. All invalid lineage slots exceed the three-exemplar cap; none is a missing terminal score.

Seed	M_0	L	S	$L - S$ (pp)	L scored	L invalid	S scored
20260821	270	283	304	-4.430	317	83	400
20260828	267	297	296	$+0.211$	400	0	400
20260835	272	287	296	-1.899	373	27	400
20260842	267	289	303	-2.954	355	45	400
20260849	271	293	298	-1.055	400	0	400
20260856	269	302	297	$+1.055$	400	0	400
20260863	267	276	303	-5.696	329	71	400
20260870	270	292	294	-0.422	400	0	400
Total					2,974	226	3,200

Candidate applicability. Reapplying the frozen patch function to all 3,200 lineage slots and the corresponding single-edit records confirms their shared patch order and all recorded candidate, acceptance and terminal state hashes. The 226 invalid slots all attempt an exemplar addition to an incumbent already holding three exemplars. The same patches are legal from M_0. There are no duplicate or other invalid candidates; seed 20260821, generation 5, round 10 has the only all-invalid round, leaving the incumbent unchanged. Because the frozen bank contains no exemplar-removal patch, later additions remain inapplicable after capacity is reached. This state-dependent boundary belongs to the prespecified search strategies; it is neither a scoring omission nor a demonstration that the cap caused the terminal difference. Slots were not replaced, affected seeds were not dropped and invalid patches were not assigned task failures. The comparison does not isolate persistence at equal realized candidate scoring or token cost, nor generalize to an unrestricted edit space.

Post-review capacity description. An exploratory description added after internal review retains both groups: the four seeds with at least one capacity-invalid addition ($20260821/35/42/63$) have mean $L-S = -3.745$ pp; the four without one ($20260828/49/56/70$) have mean -0.053 pp. These are post-treatment groups determined by search history, acceptances and candidate order. Their difference does not identify a causal cap effect or a comparison at equal realized scoring. The near-zero mean in the latter group establishes neither equivalence nor a fair randomized comparison. No seed is excluded, no subgroup interval or test is added, and the all-eight primary below is unchanged.

Post-review complete-evaluation ceiling. A saved-record sensitivity fixes each seed's ceiling to all completed L audit task/spec evaluations: the initial baseline, g_0–g_5 checkpoints, incumbents in nonempty rounds, and scored candidates. An evaluation on all 300 audit tasks counts as one complete specification evaluation; internal syntax-repair calls add model cost but not another task/spec evaluation. The per-seed ceilings are $423, 507, 480, 462, 507, 507, 436, 507$ complete evaluations, in Table 13 order. We retain the largest prefix of complete original S rounds within each ceiling: each round evaluates M_0 and four candidates. These prefixes contain $336, 400, 384, 368, 400, 400, 348, 400$ candidates, respectively. Original maximum-net-win selection uses earliest-stream tie breaking and M_0 fallback for a nonpositive maximum. All eight selected specification hashes equal the original S endpoints, whose same-seed terminal panels are complete. Thus the reported $L - S$ result is unchanged under this explicit truncation; no new terminal score or independent replication is added. The ceilings depend on realized L histories. Whole rounds and the finite candidate stream leave budget unused, and token costs remain unequal. This sensitivity neither matches candidate counts nor isolates a causal effect of persistence or capacity at equal expenditure.

Table 14: Qwen3.5 paired seed contrasts, in percentage points. Tasks are fixed; each interval uses eight seed differences and Student t_7. Exact two-sided sign flips enumerate all 256 assignments. Secondary comparisons retain their prespecified roles and are not replacements for the primary.

Contrast	Role	Mean (pp)	95% CI	Sign-flip p
$L - S$	Primary	-1.8987	$[-3.8469, +0.0494]$	0.0625000
$L - M_0$	Secondary	$+4.3776$	$[+2.9301, +5.8252]$	0.0078125
$S - M_0$	Secondary	$+6.2764$	$[+5.3983, +7.1545]$	0.0078125

Both routes improve over M_0 in all eight seeds; $L - S$ is negative in six and positive in two. Its interval crosses zero and admits worse lineage performance, establishing neither single-edit superiority nor equivalence. The small positive upper endpoint is not a narrower common interval for the old and new settings, nor a test of a between-model difference. This result and the telecom execution-path study address separate questions; neither is a causal explanation of the other.

Model identity and execution. The bound checkpoint is `Qwen/Qwen3.5-4B`, served in text-only, non-thinking mode using BF16 without quantization. vLLM 0.18.0, Torch 2.10.0 and Transformers 4.57.6 use tensor parallelism 1, data parallelism 4, context length 16,384 and output allowance 1,200. Model identity is bound by per-file ModelScope commit identifiers and SHA256 hashes; those files do not share a single asserted upstream checkout revision. Retained source records contain the aggregate asset and service-epoch digests. Two paired-seed jobs ran concurrently, with 24 task workers each and L before S within a pair. Requests retained the harness state's temperature, task-context seeds, inherited repair rule, 240-second HTTP timeout and zero HTTP retries; task execution was limited to 6 seconds. No cross-occurrence response caching was used. Seed reuse does not imply identical physical responses.

The new service uses `generation_config=vllm`; its nongreedy defaults are top-$p = 1$ and top-$k = 0$, whereas the earlier completion service used model defaults top-$p = 0.8$ and top-$k = 20$. Both use min-$p = 0$; the temperature-zero path resets these sampling settings to $1/0/0$. The earliest W2 services' effective defaults are not authenticated by that later completion binding. This is therefore a new model *and* serving configuration, not a weight-only intervention. Within the new condition, both routes use the same frozen execution configuration.

Observed cost and output caps. Table 15 reports all 24 search/terminal cost units. Across the run there were 2,362,318 physical calls and successful responses, 546,657,230 input and 279,971,081 output tokens (826,628,311 total), with no unknown usage or saved-response reuse. The 11,392 terminal calls comprise 11,376 task solves and 16 inherited repair calls. Lineage search uses fewer calls and scored candidates but more tokens than single-edit search; these are distinct budget axes. No accuracy reweighting or cost-optimality claim follows. The 55.5892 allocated GPU-hours cover four cards, including loading and cleanup; shared occupation is not apportioned by request count. Historical shared bank-generation cost remains unmeasured and excluded.

The retained finish reasons are 38,886 length-terminated and 2,323,432 stop-terminated calls, with no unknown reasons. All recorded outcomes remain in the analysis. These call-level counts identify neither affected-task accuracy nor an uncapped counterfactual; length termination is not imputed as failure, missingness or zero. The original resource limits were five million physical calls, five billion total tokens and 100 allocated GPU-hours; the run completed within them. This extension does not alter the separate public-selector stopping-rule qualification.

Table 15: Qwen3.5 observed calls and tokens by seed and phase. T is the three terminal measurements together. Unknown usage is zero in every row. Totals exclude unlogged historical bank generation; they are retained journal costs, not score-derived estimates.

Seed	Phase	Physical calls	Input tokens	Output tokens
20260821	L	126,902	54,422,041	9,804,224
20260821	S	150,006	23,461,009	15,125,320
20260821	T	1,422	326,934	165,206
20260828	L	152,105	28,534,105	24,320,598
20260828	S	150,003	23,254,511	15,268,527
20260828	T	1,422	189,948	192,908
20260835	L	146,175	58,335,596	20,016,535
20260835	S	150,006	23,371,076	15,126,619
20260835	T	1,438	402,530	216,605
20260842	L	138,601	52,682,191	13,959,564
20260842	S	150,004	23,245,000	15,408,036
20260842	T	1,422	350,160	207,183
20260849	L	152,100	42,241,152	25,012,081
20260849	S	150,000	23,398,600	15,113,467
20260849	T	1,422	347,790	211,267
20260856	L	152,100	34,839,552	25,051,845
20260856	S	150,000	23,388,700	15,401,892
20260856	T	1,422	218,862	213,376
20260863	L	130,804	53,487,613	11,046,948
20260863	S	150,006	23,296,634	15,190,025
20260863	T	1,422	309,870	170,474
20260870	L	152,110	33,070,821	27,222,709
20260870	S	150,004	23,253,719	15,345,895
20260870	T	1,422	228,816	179,777
	L total	1,150,897	357,613,071	156,434,504
	S total	1,200,029	186,669,249	121,979,781
	T total	11,392	2,374,910	1,556,796

Post-review candidate-count sensitivity. A separate follow-up retains the first n_s original S candidates, where n_s is the number actually scored by seed s's original L search (Table 13). It uses the original audit-only maximum-net-win selector, earliest-stream ties and M_0 fallback. Only seed 20260863 changes its selected state: the 329-candidate prefix selects stream index 70, bank index 1. All states and the analysis rule were frozen before new terminal evaluation; no search was rerun.

Let C denote this truncated single-edit state and O the original S. One new service batch measures all eight L and C states and seed 20260863's O: 17 physical panels on the same 474 tasks. For the other seven seeds, O aliases C only because seed, full specification and task context are identical. Old terminal scores do not enter these contrasts. Within each task chain, the frozen schedule balances L/C and C/L orders (237 each), or all six $L/C/O$ orders for seed 20260863 (79 each). Model assets, decoding, task-context seeds and the scorer retain their original settings. No M_0 panel or new gain over M_0 is reported.

Table 16: New-batch candidate-count sensitivity. Each success count is out of 474; the seven identical C/O entries share one physical panel within their seed. These are the original eight seeds, not additional replicates.

Seed	n_s	L	C	O	$L - C$ (pp)
20260821	317	285	303	303	−3.797
20260828	400	297	294	294	+0.633
20260835	373	285	299	299	−2.954
20260842	355	287	297	297	−2.110
20260849	400	296	294	294	+0.422
20260856	400	294	290	290	+0.844
20260863	329	274	296	298	−4.641
20260870	400	292	296	296	−0.844

The follow-up's primary contrast is the mean of the eight paired $D_C = 100(Y_L - Y_C)/474$ values: -1.555907 pp, with a Student t_7 95% interval $[-3.337059, +0.225244]$ and two-sided sign-flip $p = 0.1015625$ over all 256 assignments. Sign-flip inference assumes joint sign invariance; this is not a randomized treatment test. The descriptive same-batch $D_O = 100(Y_L - Y_O)/474$ mean is -1.608650 pp. Seed 20260863 has 18 paired C wins, 20 losses and 436 ties against O, for a net $\delta = -2$ successes; the identity $\overline{D_C} - \overline{D_O} = -100\delta/(474 \times 8)$ holds. This descriptive same-batch difference is $+0.052743$ pp. It does not attribute the difference from the original primary contrast to truncation. The interval spans zero, establishing neither lineage superiority nor equivalence. Counts n_s depend on realized L histories; the comparison does not match total evaluations or tokens or isolate a pure persistence effect. It adds no independent seeds or tasks and leaves the original primary result unchanged.

All 8,058 task evaluations are present. The 8,071 physical requests succeeded, including 13 syntax repairs under the frozen rule; there are zero transport retries or unknown-usage requests. Usage is 1,974,528 input and 1,300,700 output tokens (3,275,228 total). Four L20X GPUs were allocated for 416.142 seconds including loading and cleanup, or 0.462381 GPU hours. The precommitted ceilings were 10,000 physical attempts, 20 million tokens and 2 GPU hours. These are costs of the new terminal measurement, not of a new search or an equal-cost strategy comparison.

E.10 Public-example selection at deployment

Design and scope. We measured one executable selector alongside the old frozen endpoints using the same Qwen3-4B-Instruct-2507 assets in a new, common serving epoch. The comparison reuses all eight parent lineages and the same 474 MBPP reserve tasks. It adds no searches or independent seeds, and the reused tasks are not a new blind reserve. All endpoint specifications and the four contrasts below were fixed before generation. The results are an as-run exploratory comparison with a qualification on the written stopping rule, described below.

For every parent and task, G is the original greedy baseline, R changes only its temperature to 0.7, L is the original generation-five lineage endpoint, and S is the original audit-selected single edit. The original extraction and optional syntax-repair settings remain in force; G and R allow no repair. Every condition receives a separate generation, including when two specifications coincide. All use one service with vLLM 0.18.0, BF16, a 16,384-token context, and the original model's sampling defaults (top_p $= 0.8$, top_k $= 20$), with temperature set by the specification. Each request has a 1,200-token output cap. The $G/L/S$ measurement seed is the first eight hexadecimal digits of SHA256 of the parent seed, -1, 0 and task ID, separated by vertical bars. R appends the fixed suffix `publiccheck-resample-1`; the HTTP seed is reduced modulo $2^{31} - 1$. Parents run in their original order, with $G/R/L/S$ in that order and 24 task workers.

Selector and scoring. The selector sees only the task ID, prompt, public signature hint and generated G/R code. It executes the single public assert with empty setup code. Missing or malformed hints make the check unavailable; candidate errors and timeouts fail the check, while executor infrastructure errors stop execution. Both candidates are generated and checked for every task. The fixed rule chooses R only when both checks are available, R passes and G fails; otherwise it chooses G. All 3,792 selections and generated code hashes were frozen before the complete scoring tasks were read. The original full suite includes that public assert, so the outcome is full-suite accuracy after selection freeze, not hidden-only accuracy. D reuses the selected G/R code's same full-suite verdict; the same-pair oracle reuses $\max(G, R)$. Neither adds an independent observation. This oracle concerns the greedy and temperature-0.7 pair, not the historical iid pass@2 estimate.

Table 17: All four fixed public-example selection contrasts. Differences and nominal 95% t_7 intervals are in percentage points; p enumerates all 256 paired sign assignments. The primary comparison is $D - L$. These as-run exploratory results retain the stopping-rule qualification in the text.

Comparison	Mean (pp)	95% interval (pp)	Sign-flip p
$D - L$ (primary)	-0.079	$[-0.839, +0.681]$	0.8906250
$D - G$	$+0.976$	$[+0.789, +1.163]$	0.0078125
$D - S$	$+0.105$	$[-0.554, +0.765]$	0.7812500
Same-pair oracle $-D$	$+0.290$	$[+0.129, +0.452]$	0.0312500

Table 18: Complete per-parent publiccheck outcomes. $G/R/L/S/D/O$ entries are full-suite passes out of 474 tasks; O is the same-pair oracle. The last three columns are descriptive contrasts (pp) added during execution, with no new tests. In the final row, the six accuracy columns are means in percent and the contrasts remain in pp. These are eight old parents remeasured in one common service epoch.

Parent	G	R	L	S	D	O	$L-G$	$S-G$	$L-S$
20260821	317	317	325	322	323	323	+1.688	+1.055	+0.633
20260828	316	317	322	325	320	321	+1.266	+1.899	−0.633
20260835	316	315	316	314	320	322	+0.000	−0.422	+0.422
20260842	317	316	318	319	322	324	+0.211	+0.422	−0.211
20260849	317	312	329	316	320	322	+2.532	−0.211	+2.743
20260856	315	318	321	324	321	323	+1.266	+1.899	−0.633
20260863	317	314	323	324	322	322	+1.266	+1.477	−0.211
20260870	317	314	318	321	321	323	+0.211	+0.844	−0.633
Mean (%)	66.772	66.535	67.827	67.642	67.748	68.038	+1.055	+0.870	+0.185

Table 19: Measured deployment generation cost over 3,792 task–parent pairs per condition. $D = G + R$ counts both candidates and is a derived cost row, not additional inference. Public-check CPU work and historical search costs are separate. All current response usage is observed.

Condition	Calls	Input tokens	Output tokens	Total tokens	Tokens/task
G	3,792	424,656	427,172	851,828	224.638
R	3,792	424,656	413,703	838,359	221.086
L	3,802	558,863	660,008	1,218,871	321.432
S	3,792	465,894	626,532	1,092,426	288.087
$D = G + R$	7,584	849,312	840,875	1,690,187	445.724

Outcomes and uncertainty. Tables 17 and 18 report every fixed contrast and every parent. Uncertainty uses eight paired parent differences on fixed tasks: nominal 95% t_7 intervals and exact two-sided enumeration of all 256 sign assignments, including zero differences. The secondary comparisons have no multiplicity adjustment. Shared tasks and reused states limit extrapolation; 3,792 task–parent pairs are not independent search seeds. The primary $D - L$ interval and the $D - S$ interval cross zero. These comparisons establish neither superiority nor equivalence. The public checks yield 2,851 pairs where both pass, 817 where both fail, 52 where only R passes, and 72 where only G passes; none is unavailable. The 52 switches to R give 37 full-suite improvements, no losses and 15 unchanged failures. Public execution omits setup whereas the full suite can include it. The observed absence of harmful switches therefore does not prove universal monotonicity. The nominal $D - G$ and oracle-minus-D tests describe this selector's recovery and remaining same-pair gap; they are not independent evidence of generalization to untouched tasks.

The $L-G$, $S-G$ and $L-S$ columns in Table 18 were added during execution, before outcomes were read, as descriptive contrasts of the same matrix without uncertainty tests. After internal review, we supplement only $L - S$ with an exploratory nominal 95% t_7 interval: $+0.185$ pp $[-0.770, +1.139]$. It is computed directly from the eight paired parent differences, treating 474 tasks as fixed, not from other contrasts' interval endpoints or 3,792 independent observations. No significance test is added. All rows and the fixed-cap/stopping-rule qualification remain. This gives precision for a common measurement epoch; it neither replaces the original six/eight-lineage analyses nor establishes equivalence, identifies a pure service effect, repairs search-selection history or supplies new searches or independent seeds.

Deployment cost and missing information. Table 19 counts both complete generations for D, including the unselected code. Across the four generated conditions there are 15,168 solves and 15,178 physical calls: one initial call per solve plus 10 syntax repairs, all in L. No task verdict or usage record is missing. Total recorded usage is 1,874,069 input and 2,127,415 output tokens, or 4,001,484 tokens. D's derived row is not added again to that total. Public checks take 105.278 seconds of phase wall time and 779.641 seconds of child-process CPU time; full-suite scoring takes

228.398 and 1,511.425 seconds, respectively. Summed candidate durations overlap and are not phase wall time. GPU occupancy is 0.34964 GPU-hours, including both service startups and the intervening interval. Historical single-edit search and candidate-bank generation costs remain incompletely recorded. This comparison does not match total search budgets or establish a cheaper method.

Length termination and the stopping-rule qualification. Among 15,178 physical responses, 141 end at the 1,200-token cap with `finish_reason=length`, affecting 139 solves. The response counts are 38 for G, 24 for R, 40 for L (including two repair responses), and 39 for S. Three affected L solves eventually pass under the original repair semantics; the affected $G/R/S$ solves all fail. The selected output in 30 D solves is length-affected, and all 30 fail. The frozen implementation forwarded these responses through its ordinary extraction and permitted syntax-repair path. The written protocol also required stopping on "incomplete outputs" without specifying whether normal length termination belonged to that category. If it did, the implementation failed to execute that stopping requirement. We therefore retain the complete fixed-cap comparison with this qualification, rather than certifying unqualified protocol compliance. All 15,168 solves remain in the denominator; no capped output was dropped, regenerated, rescored or given extra repair. The analysis does not estimate an unlimited-token counterfactual.

E.11 Study map and evidence levels

Registration ledger. Table 20 lists every cell of the 4B campaign with its registration status, confirmatory statistic, result and deviations, so that the reader can see which numbers were fixed before which reserve reads.

Table 20: Study map and evidence levels. Registration precedes each cell's own read, not all project access to the reused MBPP reserve. R2 alone used the previously unopened second reserve. Primary denotes the registered comparison, not a positive result; deviations and intervals are reported in the indicated sections.

Cell / main-text location	Seeds	Evidence level	Estimand or qualification
Pilots / §4.3	1 each	Exploratory diagnostics	Different banks; final-gap and dAUC contrasts
W2 / §4.4	8	Registered primary; six-seed interim	dAUC_{AB}; data-dependent interruption, mixed services
W2b / §4.5	3	Exploratory lead	Online proposals, not part of W4
W2c / App. D	3	Exploratory H3	Fresh-data veto
Post-selection / §4.2	6 W2-A	Registered interaction, unresolved	T_{int}; ρ changed after inspection
W4 / §4.5	8	Registered primary	dAUC_{R5-B}, exact sign-flip
W5 / §4.2	6	Registered comparison	Lineage minus audit-selected single edit
PACE / App. E.7	6	Registered comparison	Operational gate, no error-control bound
W6 / §4.5	8	Registered transcript analysis	Copying rule met, score-following rule unmet
W6B / App. E.4	8	Registered descriptive follow-up	Accepted and persistent copied text
R2 / §4.6	8 W2, 8 W4	Registered batch; secondary transfer	One global read; negative W4 result outside registered sentence rule
PC / §4.6	6	Registered control, unqualified	No retained detected reference copy; known-gain control dropped
Public selector / App. E.10	8 old	As-run exploratory comparison	$D - L$; reused tasks, stopping-rule ambiguity
Second model / App. E.9	8 reused	Prospective exploratory extension	$L - S$; new model/serving, equal slots, unequal actual scoring

Per-cell table. Table 21 reports each seed and arm's accepts, overwritten share, surviving components, sealed gain, interaction, and λ=1 fold-2 compaction loss. Arm B, the three W2b online-proposer seeds, the veto seeds and the pilot are exploratory. The pre-registered 2×2 of §4.5 is not in this table.

The per-cell table gives the registered $\lambda = 1$ compaction result without overlaying the unselected penalty paths. The pooled fold-2 loss is -0.23 pp ($[-0.67, 0.20]$, 90% CI); Fig. 7 shows the per-seed values.

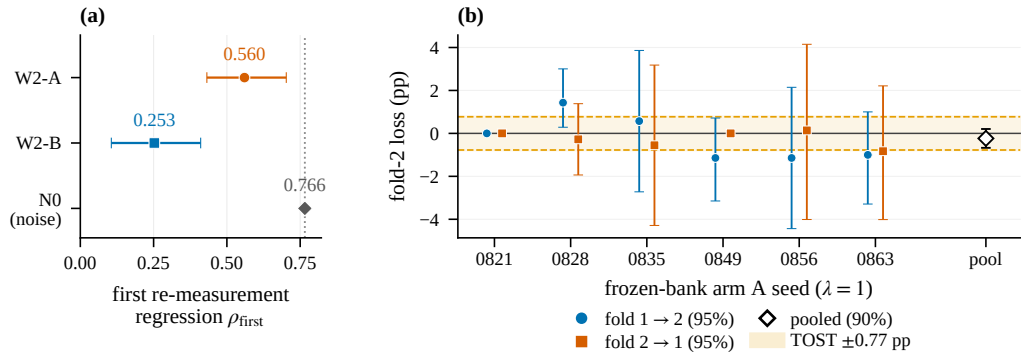

Figure 7: **(a)** First-remeasurement regression ρ_{first}: empirical event-bootstrap ranges for W2-A/B and an independent-flip simulation summary range for N0, constructed as described in App. E.2. The ranges are not comparable population confidence intervals; their separation does not exclude correlated service noise. **(b)** Cross-fitted compaction loss per seed at $\lambda = 1$.

Table 21: Per-cell post-selection audit (new service; lockbox 474 items). K: accepted edits; overwritten: share of accepts with no terminal effect; $|C|$: surviving components; G_{total} (terminal replay) and T_{int} (taken against the separately measured factorial full-set gain) with item-paired 95% CIs; loss: cross-fitted λ=1 fold-2 loss (two-direction mean, 95% CI). W2 arm A is primary; other cells are exploratory. Factorial full-set and singleton states each have three evaluations in W2 arm A and W2b arm A, and one in the other cells; baseline sources are detailed in Appendix E. The three veto (W2c) cells shown were not replayed: their sealed gain is the factorial baseline contrast from a single evaluation, and one arm per veto run is shown (C for seed 0821, the paired B control for 0828 and 0835); both arms' factorial values are retained in the saved records.

cell	K	overwritten	$\lvert C\rvert$	G_{total} (pp)	T_{int} (pp)	λ=1 loss (pp)
W2-0821/A	44	0.93	3	$1.27[-1.48, 3.80]$	$0.63[-1.83, 3.02]$	$0.00[0.00, 0.00]$
W2-0828/A	19	0.79	4	$1.90[-0.42, 4.22]$	$-0.07[-2.88, 2.67]$	$0.58[-0.41, 1.68]$
W2-0835/A	33	0.94	2	$0.42[-2.11, 2.95]$	$-0.91[-3.09, 1.20]$	$0.01[-2.61, 2.42]$
W2-0849/A	25	0.92	2	$1.90[-0.42, 4.22]$	$-0.21[-2.53, 2.04]$	$-0.57[-1.57, 0.36]$
W2-0856/A	31	0.90	3	$1.27[-1.27, 4.01]$	$-0.91[-3.66, 1.90]$	$-0.50[-3.12, 2.08]$
W2-0863/A	15	0.80	3	$2.53[0.00, 5.06]$	$-2.04[-4.15, -0.07]$	$-0.92[-2.86, 0.88]$
W2-0821/B	4	0.50	2	$1.27[-0.85, 3.38]$	$-0.28[-2.32, 1.76]$	$-0.14[-2.15, 1.82]$
W2-0828/B	1	0.00	1	$1.05[-0.63, 2.95]$	$0.00[0.00, 0.00]$	$-1.24[-2.77, 0.07]$
W2-0835/B	4	0.50	2	$0.42[-2.33, 3.17]$	$-2.81[-5.70, 0.07]$	$2.32[0.02, 4.61]$
W2-0849/B	2	0.00	2	$1.48[-1.05, 4.01]$	$-1.05[-3.24, 1.13]$	$-0.64[-2.15, 1.07]$
W2-0856/B	2	0.50	1	$0.00[-1.90, 1.90]$	$0.00[0.00, 0.00]$	$-0.28[-1.66, 1.04]$
W2-0863/B	3	0.33	2	$0.63[-1.90, 3.38]$	$-2.04[-4.71, 0.49]$	$-0.03[-2.38, 2.48]$
W2b-0821/A	39	0.95	2	$-1.05[-3.80, 1.48]$	$0.28[-0.77, 1.27]$	$1.13[-1.40, 3.80]$
W2b-0828/A	35	0.94	2	$-1.05[-3.80, 1.48]$	$-1.27[-2.53, -0.21]$	$-0.48[-2.35, 1.39]$
W2b-0835/A	41	0.95	2	$-2.95[-5.49, -0.42]$	$-0.35[-1.34, 0.70]$	$1.97[-0.45, 4.53]$
W2c-0821/C	3	0.33	2	$0.84[-1.90, 3.59]$	$-0.63[-2.74, 1.48]$	$-0.21[-2.57, 2.32]$
W2c-0828/B	4	0.50	2	$1.90[0.00, 3.80]$	$-1.90[-3.80, -0.21]$	$-0.85[-2.34, 0.64]$
W2c-0835/B	2	0.00	2	$0.00[-2.11, 2.11]$	$-1.90[-4.01, 0.42]$	$1.89[-0.21, 4.02]$
Pilot/A	14	0.86	2	$2.11[-0.42, 4.85]$	$2.74[-0.07, 5.49]$	$0.00[0.00, 0.00]$
Pilot/B	3	0.33	2	$-0.42[-2.53, 1.48]$	$0.00[-2.18, 1.90]$	$-0.69[-2.70, 1.25]$

F Local Mechanism Diagnostic: Controls and Evidence

Relationship to the quantitative study. The 27B case is exploratory and was conducted after the registered 4B campaign. It uses separate public TRAIN telecom tasks, not the MBPP or HumanEval reserves. It investigates one accepted edit from a completed P/N search pair, not a new estimate of the 4B disclosure effect or a cross-model replication. P updates the currently accepted state; N applies each proposal to the initial state while retaining its own feedback history. Both routes used four rounds of two candidates, 20 search tasks with four repeats, and 74 terminal tasks with four repeats. The paired search required the bound gain statistic $T \geq 8$ and $p_{\text{ref}} \leq 0.05$ for acceptance, where T is the net success count difference over the 80 paired binary outcomes. P3.0 was accepted with $T = 13$ and $p_{\text{ref}} = 0.0078125$, persisted to the P terminal state, and was inherited by the two round-four proposals. The protocol's fixed-window inheritance-and-consumption indicator remained $W = \text{false}$. The local diagnostic changes none of these records, thresholds or decisions. The parent terminal counts were P $191/296$, N $188/296$ and M0 $197/296$. These are descriptive outcomes for the original paired run. Its P–M0 comparison includes the wrapper difference, so these scores do not identify the guard's performance effect. Both physical search executions used algorithm seed 20271108; they are not independent algorithm replicates. Execution identifiers appear once in Table 31.

Earlier execution of the same seed. Before the proposer-interface repair, the first full pair completed 2,152 episodes and accepted no P edit: P terminal and its same-state M0 reading were both $188/296$, N argmax was $206/296$, and $W = \text{false}$. Six of its sixteen proposal slots failed to yield a scored candidate. The interface repair clarified that each call returns one patch object; the subsequent search pair reused the seed without importing old proposals or scores and completed 2,928 effective episodes (2,929 attempts, one infrastructure failure). Its M0 was $197/296$: a $9/296 = 3.04$

pp remeasurement difference, larger than the later P−M0 difference of $-6/296 = -2.03$ pp. These executions cannot be pooled as independent seeds or used to identify an interface performance effect.

The earlier N2.1 guard checked the most recent agent-tool result using a different string predicate. It matched 25 of 604 search checks and 95 of 2,015 terminal checks; all 120 resulting feedback messages entered real model requests, with terminal activation in $92/296$ episodes. Checks are not triggers, and these counts do not establish that blocking was appropriate or beneficial. The earlier P3.1 recovery prompt reached the model 38 times; it produced no immediate same-tool parameter correction and 17 identical retries. Thus a stored rule can be consumed without correcting behavior, while P3.0's absence of natural activation is specific to its predicate.

Earlier local diagnostic: model and runtime binding. The local diagnostic was configured to reuse the parent run's 27B asset, tokenizer, chat template and decoding. The asset label `Qwen3.8-27B` and serving alias `qwen3.8-27b` identify that local asset, not a verified upstream release. Archived configuration values specify `qwen3_5` and `Qwen3_5ForConditionalGeneration`; saved loading records resolve that architecture and finish loading 18 shards. The archived configuration and tokenizer identifiers match the parent binding, and the saved template text reproduces its recorded hash. The weight index describes 55,562,855,904 bytes. The inspected archive does not identify the upstream repository or revision. An earlier deployment inventory does contain 29 file hashes, including all 18 weight shards. Its configuration, tokenizer, index and template hashes match the diagnostic binding in 11 comparisons covering eight distinct metadata files. Those recorded hashes were not recomputed over diagnostic-time weights, whose physical continuity remains unverified. This establishes metadata agreement, not authentication against a public checkpoint; the retained records preserve the historical inventory and its timing limitation. The later full-task E2 run has a separate pre-execution asset binding (Table 24); it does not authenticate these earlier deployed weights.

The engine used vLLM 0.18.0, bfloat16, one tensor-parallel rank and a 65,536-token context. Requests used temperature 0, disabled thinking, `tool_choice=auto` and a 4,096-token output allowance, checked against the public contract and effective service window. Per-state seeds and full payload identifiers are retained. The two same-seed service repeats assess local repeatability, not independent algorithm seeds. No user simulator, environment tool or task scorer was invoked in the diagnostic.

The original guard. The predicate is the conjunction of a pending tool call, tool name `transfer_to_human_agents`, and a case-sensitive substring match of `MMS` in the agent-side `last_tool_result.content`. It does not inspect the whole history or verify that a checklist has been completed. The block message asks the agent to check status bar, network status, SIM status, APN settings, network mode, Wi-Fi calling, application permissions and MMS sending before transferring. Native mediation allows one repair response. The same guard and repair behavior are used in C2; no history-aware correction is introduced. Among 253 proposed transfers in the old trajectories, 182 contained MMS text earlier in the non-system history, but none in the latest agent-tool result. Evaluation opportunity therefore did not ensure natural activation.

Table 22: Recorded natural evaluations of the P3.0 target condition. These are repeated events in episodes, not independent observations. Descendants contain additional edits; the counts refer specifically to the inherited target rule.

Scope	Episodes	Condition evaluations	True / prompt delivery
P3.0 search	80	459	0 / 0
Round-four reference	80	515	0 / 0
P4.0 search	80	554	0 / 0
P4.1 search	80	456	0 / 0
P terminal	296	1,976	0 / 0
Total	616	3,960	0 / 0

The paired task-level pilot. The later paired task-level pilot is a three-condition diagnostic on 74 paired public TRAIN tasks. C0 is the native empty-patch path, C1 the nonempty-patch wrapper with the target guard removed from the rule list, and C2 the unchanged P3.0 fullspec. All 222 episodes completed: each condition has 74 result files, task identifiers and base/effective seeds align task by task, all 222 usage rows carry input, output and total token counts, and there is no infrastructure termination, implicit retry or missing result. Every record is flagged `diagnostic_only`; none entered a candidate pool or changed a stored score, acceptance decision or W value.

Successes are $49/74$ (C0), $44/74$ (C1) and $49/74$ (C2). Terminations are 69/67/69 `USER_STOP`, 5/6/4 `MAX_STEPS`, and one `TOO_MANY_ERRORS` in each of C1 and C2. Token use is 62,225,085 input and 471,651 output across the three conditions. Treating tasks as the paired unit, C1$-$C0 is -6.76 pp ($[-18.92, +4.05]$; 7 wins, 12 losses, 55 ties), C2$-$C1 is $+6.76$ pp ($[-1.35, +16.22]$; 8/3/63) and C2$-$C0 is 0.00 pp ($[-10.81, +10.81]$; 9/9/56). Intervals are task-level paired bootstrap (100,000 resamples, fixed RNG seed 20260919) and are descriptive: this is a local pilot, not a confirmatory population test.

Under C2 the guard's `on_before_action` hook fired 489 times with zero matches and zero blocks or feedback events; C0 registers no hook and C1 only an episode reset. The guard therefore had ample natural evaluation opportunity in this batch and never activated, so neither the C2 success level nor the C2$-$C1 difference can be credited to it. Outcomes are discordant on $11/74$ tasks between C1 and C2 and on $18/74$ between C0 and C2; a shared seed does not guarantee an identical trajectory, and the saved summaries do not retain a per-request first divergence, so these flips cannot be assigned to specific GPU non-determinism, to the wrapper, or to the guard. The exact equality of the C0 and C2 totals is a coincidence of this batch, not evidence that natural task performance is unchanged.

Cost ledger for the paired pilot. The episode window occupies two cards for about 3.39 hours, or 6.78 GPU-hours. Charging both cards for the full recorded serving window gives an upper bound of about 29.07 GPU-hours, above the manifest's 24 GPU-hour cap. No per-card allocation and release ledger was kept, so the two figures bound the same run from below and above rather than resolving it. We therefore report both and do not claim that the registered budget cap was met, nor assert a definite overrun. Serving processes were terminated after the run; no per-card allocation ledger was kept. Operational timestamps and resource-release records remain in the saved provenance archive.

Natural state selection. We use the parent run's M0 terminal panel with algorithm seed 20271108 and repeat index 0, restricted to its 74 public TRAIN tasks. Tasks are ordered by the SHA-256 of a fixed diagnostic prefix concatenated with task ID. For each task, the first eligible response contains both assistant text and tool calls, has matching saved tool returns and an exact reconstructable next request, and is nonterminal. The last tool result must not contain `MMS`. The first 12 eligible tasks are retained; 17 task traces were inspected before that quota was met. No success score enters selection. This deliberately enriches the set for a message-processing difference and makes it a mechanism-targeted set, not an unbiased sample of natural activation opportunities. The historical zero-trigger finding instead comes from Table 22.

Table 23: Local observations, kept separate by state origin. Execution counts include two same-seed repeats. Natural guard contrasts reuse identical physical responses; synthetic ones include a repair opportunity absent in C0/C1. No row measures task success.

Endpoint	Natural states	Synthetic probes
State units	12 selected saved prefixes	4 constructed states
Wrapper changes request	12/12 states	No comparable C0/C1 continuation
Wrapper changes action	3/12 states, both repeats	No difference in allowed transfer
Target condition true / blocks	0 / 0	8 / 8 executions
Prompt in real request	0	8/8 executions
Subsequent model response	No target-rule intervention	8/8 non-transfer
Full task outcome	Not tested	Not tested

Synthetic state construction. The first four selected prefixes supply four artificial positive controls. Their last visible tool result is replaced by a fixed message containing `MMS`, and a constructed transfer proposal is supplied identically to C0, C1 and C2. Three parent tasks concern mobile data and one concerns MMS. These modified states are not claimed to be reachable under the native environment or to have existed in the old search. The intervention tests whether downstream blocking and feedback delivery work given a true condition.

Path isolation and measurement. C0 invokes the actual empty-patch native path. C1 keeps the nonempty-patch mediation and message processing but disables the target guard list, retaining the non-noop path. C2 uses the original P3.0 fullspec. Each comparison starts from the same saved prefix and common response, with identical tool-return messages. The target wrapper clears the assistant content attached to tool calls before forming the next request. In the 12 natural contrasts, the requests differ in this content field. The intervention is nevertheless the whole wrapper path, not a separately randomised one-line code ablation.

The conditions use a frozen order, balanced across the 12 natural states, concurrency one and two repeats with the same state seed. Natural C1 and C2 yield identical full requests, so one physical response is explicitly reused within each state and repeat before separate mediation. No response is shared across repeats. Thus their zero observed action difference is partly a controlled-input identity, not an independent stochastic equivalence test. Synthetic C0 and C1 allow the common transfer proposal and have no subsequent model call; only C2 creates a real repair request. Those controls are marked not applicable for model generation, not treated as missing or fabricated calls.

Endpoints and results. The action signature comprises ordered tool names and canonical arguments; responses without tool calls form the category `USER_MESSAGE`. Text-only variation does not count as an action-signature difference. The wrapper changes the request in every natural state, the final response content in $11/12$, and the action signature in NAT-01, NAT-09 and NAT-10. The two service repeats agree within every condition. Native responses request phone confirmation in NAT-01 and NAT-09 and a speed test in NAT-10; the wrapper responses call `get_details_by_id` with `id=L1001`. We do not score their correctness. Natural C2 has 44 condition evaluations across replayed and newly generated responses, with no transfer proposals, true conditions, blocks or feedback requests in this local window.

Table 23 reports the key contrasts in the main text.

The eight synthetic C2 transfer proposals all trigger, block and produce real repair requests containing the original guard feedback. The resulting model responses are non-transfer in all eight executions: SYN-01 requests customer details by phone, SYN-02 asks for cooperation with troubleshooting, SYN-03 asks permission to enable roaming, and SYN-04 requests line details. These responses do not establish completion of the checklist; SYN-03 returns to the parent mobile-data issue. The contrast identifies the package of blocking, feedback and repair opportunity, not the effect of prompt wording alone. Four artificial states with repeatable outcomes provide a wiring check, not an estimate of natural repair efficacy. No confidence interval or significance claim treats the repeated executions as independent tasks or seeds.

Evidence boundary and accounting. Proposal, acceptance, persistence and runtime condition evaluation are established for P3.0. Natural activation and natural prompt consumption are not established; the examined old records contain zero instances. Conditional activation, prompt delivery and a changed local action path are established for the synthetic probes. A separate wrapper effect is established in the selected natural contexts. Guard-attributable task benefit is not established in the old run and is not tested locally. No local result replaces the old scores, acceptance decisions or W value.

All 16 frozen states and 96 condition-repeat cells completed. There were 80 logical new-request records: 56 physical requests and 24 explicit within-repeat aliases. The remaining 16 synthetic C0/C1 cells required no new request. Physical responses account for 452,306 input and 2,792 output tokens, with no missing usage or infrastructure failures. Generation elapsed time was 86.41 seconds and total GPU occupancy, including loading and cleanup, was 0.069359 GPU-hours. No task episode, score or new candidate was produced. The frozen manifest, state selection record,

condition cells and raw requests and responses are retained under the local-diagnostic identifier in Table 31.

A separate calibration panel that did not qualify for measurement. An earlier attempt to measure this mechanism through a separate closed calibration run stopped at its registered calibration gate and supplies no estimate here. Its saved panel has four cells (2 parents $\times$ 2 conditions) of 74 task records each, with binary Q_1 counts 45, 44, 39 and 47; termination is `USER_STOP` in 269 of 296 records and `MAX_STEPS` in 27. A pair-alignment audit over 148 within-context repeat pairs finds 49 Q-discordant and 99 Q-equal pairs, with the first saved response divergence classified as assistant-side in 88 pairs and user-side in 60. The registered admissibility gate failed ($F_{\max} = 25$, no admissible margin), and the downstream search, proposal, canary, witness and readout paths were never executed. This is a failure of measurement qualification, not a negative result about the mechanism: no effect estimate, no calibrated upper bound and no imputed $Q = 0$ may be read off these records, and the 27 `MAX_STEPS` records are not convertible into an effect. A future measurement of this kind would need a predeclared admissible precision and margin, an executable positive/negative control, condition identity, a stop rule and a native Q receipt before any live call.

F.1 Prospective full-task control outcomes

The fixed panel contains 74 previously used telecom TRAIN tasks, four conditions and two live repeats, giving 592 completed episodes. C0 uses the native empty-patch path. C1 uses the original nonempty wrapper with the target guard removed; C2 uses unchanged P3.0. C3 changes C1 only by omitting content clearing in `_replace_tool_calls`. These nested controls are not a full factorial, and C3 does not correct the target predicate.

Execution identity of the full-task panel. Table 24 gives the new run's model roles and runtime configuration. Before inference, the recorded binding hashed the current model assets, tokenizer/template metadata, native sources and support inputs. The saved numerical records retain asset names, sizes and hashes, along with the source-receipt identity. These identify the bytes used for E2; they do not identify a verified public upstream checkpoint or authenticate the earlier search and local diagnostic retrospectively.

Table 24: Execution configuration and pre-execution binding for the full-task E2 panel. Model names are served local aliases. Internal source checks and their CPU scope are described in App. H.

Component	Recorded configuration and binding
Agent	`qwen3.8-27b`; 28 asset files including 18 weight shards; context limit 65,536 tokens
User, interface, NL judge	`qwen3-32b`; 25 asset files including 17 weight shards; context limit 40,960 tokens
Requests	Temperature 0, thinking disabled, output cap 4,096 tokens, no transport retries
Worker	Python 3.12.13, NumPy 2.3.5, Transformers 4.57.6
Both services	vLLM 0.18.0, Torch 2.10.0, Transformers 4.57.6
Native execution	238 Python source files and 6 support inputs bound; one episode at a time, at most 100 steps and 10 errors
Tasks and conditions	74 reused TRAIN task objects bound individually; C0–C3 have separate execution identities

The common request seed is derived from master seed 2026092501, task, semantic role and role-local call ordinal, excluding condition and repeat. Agent calls share one stream; user, interface and natural-language judge calls use distinct streams. Thus two repeats reuse the same request-seed rule while allowing observed response changes under incompletely observed serving state.

The outcome is the native total binary reward Q, retaining its original reward basis; database reward is not substituted. Repeats are averaged within each task before computing the 74 paired task differences. The prespecified task bootstrap uses 20,000 resamples and analysis seed 2026092502. Primary $C1 - C0$ has a 95% interval. The two key secondary comparisons $C1 - C3$ and $C2 - C1$ use 97.5% intervals each under Bonferroni adjustment; $C3 - C0$ is descriptive with a 95% interval. Reused tasks limit extrapolation, and same-seed executions are not new algorithm seeds.

Table 25: Complete E2 native total-reward outcomes. Each condition has 74 tasks and two live repeats. Repeat successes each have denominator 74; discordance counts tasks with different binary Q across the two executions. These repetitions are not algorithm seeds.

Condition	Passes	Rate (%)	Repeats 0/1	Discordant
$C0$: Native	$98/148$	66.216	$51/47$	$16/74$
$C1$: Wrapper without guard	$91/148$	61.486	$42/49$	$11/74$
$C2$: Original P3.0	$97/148$	65.541	$48/49$	$11/74$
$C3$: C1 retaining content	$94/148$	63.514	$47/47$	$14/74$

Table 26: Every prespecified task-paired contrast and both descriptive repeat panels (pp). Primary $C1 - C0$ and descriptive $C3 - C0$ use 95% intervals; key secondary $C1 - C3$ and $C2 - C1$ use 97.5% intervals. No additional test is applied to the repeat panels.

Contrast	Mean	Level	Interval	Repeats 0/1
$C1 - C0$	-4.730	95%	$[-15.541, +6.757]$	$-12.162/+2.703$
$C1 - C3$	-2.027	97.5%	$[-14.189, +10.135]$	$-6.757/+2.703$
$C2 - C1$	$+4.054$	97.5%	$[-2.703, +10.811]$	$+8.108/+0.000$
$C3 - C0$	-2.703	95%	$[-8.784, +3.378]$	$-5.405/+0.000$

Table 27: Distribution of task-level paired differences after averaging the two repeats. Every row sums to 74 tasks; bins describe recorded binary-reward differences, not imputed missing outcomes.

Contrast	-1	-0.5	0	$+0.5$	$+1$
$C1 - C0$	6	16	36	11	5
$C1 - C3$	6	11	42	10	5
$C2 - C1$	0	7	54	13	0
$C3 - C0$	1	9	57	7	0

Table 28: All-role physical requests and observed usage from the complete saved traces. Every request has one response; unknown usage is zero. This table does not allocate shared GPU occupancy to conditions.

Condition	Calls	Input tokens	Output tokens	Total tokens
$C0$	6,371	43,434,769	351,188	43,785,957
$C1$	6,286	41,702,312	302,897	42,005,209
$C2$	6,199	41,080,394	305,629	41,386,023
$C3$	6,354	43,170,512	349,356	43,519,868

All four fixed intervals contain zero. In particular, the primary difference is -12.162 pp in repeat 0 and $+2.703$ pp in repeat 1; both are retained. The per-condition disagreement counts are $16/74$, $11/74$, $11/74$ and $14/74$. No repeat panel is selected for its effect direction, and these descriptive counts add no test or equivalence criterion.

The complete saved traces contain $25{,}210$ physical requests and exactly one response per request, with $170{,}697{,}057$ known tokens and no unknown usage. No transport, generation, unclassified-call, scorer-error or cell-failed event is present in the checked trace records. C2 has 910 target checks, zero matches and zero blocks; the other conditions contain no target checks. Two responses terminate with `length`, both in C2, and remain in the full analysis. Their presence is not taken as evidence of no effect on Q.

These task-score differences compare implemented conditions. The C2−C1 point estimate does not by itself demonstrate a benefit from the target guard's blocking or feedback, nor is an unexplained difference assigned to pure noise. Request content, rendered exposure, adoption and response divergence require their own evidence. The comparisons do not uniquely decompose mechanism

contributions or identify the source of the old acceptance score. Old PNW outcomes, acceptance decisions and W are unchanged.

Input, first response and adopted tool calls. All $25{,}210$ completed requests were reconstructed offline with the bound tokenizers and templates, with no missing rows or prompt-count mismatches. Four existing server `/tokenize` crosschecks matched the reconstructed token IDs on those selected requests. They are not a server-side token capture of every request or a record of hidden serving state. C1 and C2 omit assistant content accompanying tool calls from later agent prompts; C0 and C3 retain it. The agent-input reconstruction finds zero such nonempty historical content occurrences under C1/C2. Repeated historical occurrences are not unique actions, and a token effect is not a task effect.

Chronological comparison aligns the role-local request prefix and stops at the first role/ordinal mismatch. In each of C1−C0 and C1−C3, 110 of 148 task-repeat pairs have a token-input difference before or at the request producing their first semantic response difference. These witnesses span 62 distinct tasks per comparison. The first semantic response may differ only in text, so 110 is not a count of changed tool actions.

Table 29: Post-result description of first-response witnesses, without selection on Q. A witness has an offline token-input difference no later than its first semantic response difference. The next columns split adopted tool-call differences at that response into changed name sequences and changed arguments with the same names. Tasks count distinct tasks with an adopted-tool difference; pairs and repeated histories are not independent tasks.

Contrast	Witnesses	Name seq.	Args only	Any tool	Tasks
$C1 - C0$	$110/148$	24	4	28	15
$C1 - C3$	$110/148$	22	4	26	15

The action clarification was added after results, retaining every fixed first-response witness without filtering on Q. Comparing original responses and the subsequent adopted events gives $28 = 24 + 4$ changed adopted-tool signatures for C1−C0 and $26 = 22 + 4$ for C1−C3. Each contrast covers 15 distinct tasks with such changes. The other $82/84$ first signatures do not establish unchanged subsequent behavior: response text or later actions may differ. These conditional descriptors add no confidence interval or test, estimate no population action-change rate, and identify no unique mediator of a task-score difference.

One natural message-path witness. After the experiment and action analysis, we selected one explanatory example from the 26 C1−C3 witnesses with an adopted-tool change. The extraction rule sorts by literal task identifier, repeat and request index, then takes the first, without using Q. It selects task alias `t002`, repeat 1, agent-request ordinal 5 (zero based). This is selection conditional on an observed action change, not a new prespecified experiment or a probability sample.

At ordinal 4, both saved semantic responses contain the same `get_customer_by_phone` call and the text "Thanks, let me look up your account now." C1 removes that text from the adopted assistant message; C3 retains it. At ordinal 5, after tool-ID normalization, the entire saved API payloads differ only in that earlier assistant message's `content` field. The saved next response under C1 gives SIM-card instructions with no tool call. C3 supplies three `get_details_by_id` and three `get_data_usage` calls, all present in its adopted messages.

An independent CPU replay invokes the recorded wrapper and native message conversion on both adjacent steps in each condition, substituting four saved model candidates for generation. Computed states, generation messages, per-condition API payloads and adopted messages match the saved records. Both conditions remove the guard, and the replay performs zero predicate evaluations. It neither samples model responses nor reconstructs engine tokens. The internal replay reproduced the deterministic message/state conversion on this prefix; the subsequent saved responses remain observations. This example does not estimate task benefit or prevalence, isolate a performance effect, or explain the historical acceptance score. The internal replay check and its dependency boundary are described in App. H.

For C2−C1 and C3−C0, respectively, the first semantic response difference is user-side in $94/148$ pairs. At those requests the tool-ID-normalized API and offline token inputs match in all 94 pairs, but exact wire inputs match in only 10 and 12 pairs, respectively. Normalization is not byte identity, and reconstructed-token equality is not full serving-state equality. Same-condition response changes are observed; hidden serving state is not fully observed. The unexplained differences are not assigned to pure noise or to a target guard function. No target feedback is observed in the API-message, rendered prompt or token-effect layers, consistent with the zero target hits/blocks.

Our internal reconstruction covers the primary statistics, descriptive summaries and these post-result action descriptors through three entry points. The action component recomputes equality from bound response and adopted-event hashes; omitted private contents cannot be recovered from those hashes. All three components were executed for the complete E2 reconstruction described in App. H.

G Exploratory Search-Availability and Terminal-Disclosure Runs

Two exploratory runs, both outside the registered 4B campaign, ask how a previously accepted text edit propagates into later search and into terminal evaluation. Both reuse the same eight already-selected historical parent lineages, so the parent set is not an independent sample and the eight parents are the analysis unit throughout. Factor E controls whether the target edit's execution availability is retained during the new proposal rounds together with the disclosure shown to the proposer; terminal factor A controls whether the accepted field is rendered at evaluation. E is a joint execution-availability and disclosure intervention, not a test of semantic forgetting, and not a deletion of the accepted field alone. The pre-specified primary estimand is $Q(E{=}1, A{=}0) - Q(E{=}0, A{=}0)$: the search-side contrast at a common closed terminal. We report the later four-round run first because it is the longer-horizon realization of the fixed protocol, and the earlier two-round run afterwards as the pilot it was.

G.1 Four-round search-availability follow-up

Completion. The run completed $64/64$ branch rounds, $128/128$ candidate slots, $64/64$ terminal panels and $9{,}600$ terminal task results, with $36{,}102$ of $36{,}102$ physical sends answered and no calls of unknown usage ($6{,}500{,}631$ input and $5{,}419{,}222$ output tokens). An independent CPU recomputation reproduces the five contrast means, the eight per-parent vectors, the intervals and the sign-flip p values from the saved records.

Candidate formation. Over the 128 slots, 101 proposals were `VALID_NEW`, 25 duplicates, 1 invalid and 1 a no-op; 105 slots were contract-legal on the first attempt and 23 completed after a single correction, with no implicit retries. By forced operation, the 48 `set` slots produced 48 valid candidates, the 48 `set_param` slots 23 valid, 24 duplicate and 1 no-op, and the 32 `add_exemplar` slots 30 valid, 1 duplicate and 1 invalid. These counts describe the whole run; they are not comparable with the 64-slot counts of the earlier pilot below, which are reported separately for that run. Three proposals crossed the unchanged `net_wins` ≥ 5 acceptance gate, each at exactly 5: parent 20261001 ($E{=}1$, round 1, `temperature` $0.0 \to 0.1$), parent 20261036 ($E{=}0$, round 3, an added integer-string-check exemplar) and parent 20261022 ($E{=}0$, round 4, a replacement of the non-target `system_prompt`). The three accepted edits are heterogeneous and do not describe one "historical-edit mechanism".

Terminal outcomes. Only 3 of the 8 $E{=}0/E{=}1$ parent pairs saved different terminal fullspecs; the other 5 pairs are identical. Those identical pairs still show small task-result differences, which are measurement variation from terminal responses that were not replayed, not a branch effect. The four terminal cells and the five saved contrasts are in Table 30.

The primary contrast is -0.083 pp with a 95% interval of $[-0.470, +0.304]$ pp, from the per-parent vector $[-0.67, 0.00, -0.67, 0.00, +0.67, +0.33, 0.00, -0.33]$ pp. This is an exploratory, near-zero reading on eight already-selected parent lineages. It does not show that historical availability has no effect, that accepted edits are inert, or that the terminal factor is negligible: no equivalence margin

Table 30: Four-round follow-up: terminal cells and saved contrasts. The analysis unit is the parent lineage ($n = 8$); intervals are across-parent t intervals with $\mathrm{df} = 7$ and p is a sign-flip sensitivity value. Passes are pooled over 8 parents $\times$ 300 tasks per cell. Only the first contrast is the pre-specified primary estimand; the rest are secondary readings of the same eight vectors.

Cell (E,A)	Passes / mean	Contrast	Mean (pp)	95% t	p
$(0,0)$	1583/2400, 65.96%	$E_1A_0 - E_0A_0$ (primary)	-0.083	$[-0.470, +0.304]$	.8125
$(0,1)$	1555/2400, 64.79%	$E_1A_1 - E_0A_1$	$+1.708$	$[-1.162, +4.579]$	.0781
$(1,0)$	1581/2400, 65.88%	$A_1 - A_0$ at E=0	-1.167	$[-3.225, +0.892]$	.2656
$(1,1)$	1596/2400, 66.50%	$A_1 - A_0$ at E=1	$+0.625$	$[-0.552, +1.802]$	.2891
		$E \times A$ interaction	$+1.792$	$[-0.943, +4.526]$	.0625

was registered, and the interval still admits small effects in both directions. The positive A=1 and interaction readings are driven by one parent (20261036, $+10.0$ and $+9.67$ pp); dropping it leaves descriptive means near $+0.52$ and $+0.67$ pp. That is a sensitivity description of a saved vector, and is not grounds for excluding the parent or promoting a secondary contrast.

G.2 Earlier two-round pilot

The earlier pilot ran the same factors for two search rounds over 64 candidate slots, 32 branch rounds and 64 terminal panels. Its forced-action counts are its own: 32 `set` slots (31 valid, 1 duplicate), 16 `set_param` slots (11 valid, 5 duplicate) and 16 `add_exemplar` slots (14 valid, 2 invalid), for 56 valid candidates out of 64. One branch accepted a single candidate, on the non-target `system_prompt` of parent 20261022; 7 of 8 parent pairs ended with identical stored terminal specifications. Its primary contrast was $+0.208$ pp ($[-0.184, +0.601]$, sign-flip $p = .375$). Because the paired requests and task seeds matched while responses were not reused, its same-specification pairs also serve as a noise check: generated code differed on 78 of 2,100 paired tasks.

G.3 Why the two runs are listed side by side

The two primary values differ by -0.292 pp, and we report them separately rather than pooling or describing either as a replication of the other. The search horizons differ (2 versus 4 rounds); the source and runtime bindings and result directories differ; the two runs share the same eight parent lineages and so do not supply independent seeds; task repeats are not algorithm seeds and terminal responses were never replayed; and E remains a joint execution-availability and disclosure intervention in both. A later published progress note reproduced this pilot's 64-slot forced-action table inside its description of the four-round run; the counts above are taken directly from each run's own saved slot records, and the 64-slot/56-valid and 128-slot/101-valid tables belong to different runs.

Separating historical-content disclosure, search availability and terminal availability, with independent parent lineages, a pre-specified practical-significance margin and a natural task-level Q, would require a new design. No such follow-up is part of this submission.

H Reproducibility Checks and Evidence

This submission contains the paper and its appendices, without separate code or data artifacts. The following describes completed internal reconstruction, synthetic execution and saved-response replay checks, together with their input and identity limits. The PDF documents these checks but does not contain the inputs and scripts needed to rerun them.

Our retained numerical archive has fifteen components invoked through the Python standard-library entry point `run_extended.py`. They verified recorded input hashes and recomputed statistics from saved verdicts and event records, with a map from claims to inputs and scripts. All fifteen components were executed after an independent extraction, without network or model calls. Their inputs contain no model weights, benchmark test code or reference solutions. The original `run_all.py` invokes twelve components; the extended entry point adds three reconstruction components. Three retained source snapshots cover the original fixed-bank implementation, the Qwen3.5 driver and

runtime, and the E2 controller, wrapper and bound native source. Real-model execution requires benchmark support data, model assets and configured services; the CPU checks did not reconstruct those environments.

Fixed-bank execution and exact input reconstruction. The retained fixed-bank source contains the frozen lineage loop and acceptance gate, single-edit selector, harness rendering and syntax repair, Python scorer, terminal evaluator and request/cost journal. Its CPU demonstration exercised these paths with synthetic task panels and mocked model responses, including refusal of terminal evaluation before search-state freeze and resume checks. It also includes the original split builders, a candidate-bank template and instructions for supplying matching model assets and a newly registered execution window. An independent offline execution reconstructed the exact $200/300/474$ MBPP splits, 50-item oracle and 150-item dev complement from the authorized existing raw file. All 524 bank rows, including 87 externalized dev prompts, matched their frozen ordering and hashes. The retained verification receipt contains counts, hashes and source identities, without task text or private paths. That reconstruction performed no task tests, scoring, model calls or GPU allocation. These checks establish executable CPU paths and exact input construction; they do not authenticate historical deployments or regenerate historical model responses. A complete real-model search remains outside this qualification.

Qwen3.5 frozen source. The retained Qwen3.5 source contains the frozen driver, runtime, science dependencies and analyzer, with byte identities checked against the completed deployment. An anonymous configuration projection records model-asset hashes, serving and decoding settings, concurrency and resource limits after driver overrides. It omits external model/tokenizer assets, task and bank contents, private deployment bindings and raw request payloads. The preserved driver contains its original execution deadline. This source inspection did not execute a new GPU reproduction. The Qwen3.5 numerical component separately reconstructs all $11{,}376$ saved terminal verdicts, the eight paired seed contrasts and 24 phase cost records. It preserves the $2{,}974/3{,}200$ actual candidate counts and does not regenerate selection or scoring (App. E.9).

E2 execution source. The retained E2 source contains the three control files, six inherited wrapper files and 238 bound native Python sources, including the episode runner and native scorer. Its projected adapter changes the default acceptance-receipt path and binds a projected receipt; the parent and accepted-candidate objects are unchanged. A source map distinguishes original and projected identities, with 247 of 250 mapped inputs byte-identical. The control logic is unchanged by this adaptation. After a fresh extraction, 30 synthetic CPU checks passed in the recorded, already installed dependency environment, with networking blocked and CUDA hidden. They cover adapter and transport contracts using message/HTTP fixtures, without loading or scoring a real benchmark task. These checks exercised the execution contracts without reproducing the 592 real-model episodes.

Natural-prefix replay. A separate internal script replayed the message/state conversions for the natural witness in App. F.1, using four saved candidates and the retained E2 source. It requires that source and its third-party dependencies, so it is separate from the fifteen-component standard-library numerical runner. Independent execution after fresh extraction reproduced the saved API payloads and adopted messages. The recorded local environment used Python $3.12.3$, LiteLLM $1.82.6$, Pydantic $2.13.5$ and Loguru $0.7.3$. This replay environment differs from the historical worker. No dependency installation, model sampling, task scoring or token reconstruction was performed. The replay archive preserves natural message contents and distinguishes original event hashes from projected hashes, while retaining the source snapshot's documented adapter mapping.

Reconstruction levels. For the six 4B lineages, the internal archive contains 255 aligned saved Boolean score vectors on 474 tasks, totalling $120{,}870$ entries. Repeated measurements and shared records remain labelled; entries are not independent tasks or model calls. These vectors recompute the retained gain, the lineage-minus-single-edit contrast and pooled terminal interaction. A separate module reselects the original six single edits from $2{,}400$ saved candidate summaries using the original audit-only rule, and links frozen identities to the saved reserve vectors. This checks recorded selection order, not historical wall-clock access or regenerated audit scores. The original W2/W4 primary tests are recomputed from saved per-seed dAUC values. W2 additionally aligns

original and replay generation states and compares saved audit means with reserve verdicts; audit task-level responses are not part of this reconstruction. Endpoint drift counts, cost crossovers and the original 457-batch census are summary arithmetic. The 25.17 and 2.57 pp means are recomputed from six saved attribution-analysis event summaries and per-seed audit endpoint means; this does not reconstruct the original acceptance decisions or audit task responses. An additional event projection and endpoint summary extend the selected-credit arithmetic to eight lineages, explicitly marking the fallback estimator. The later measurement extension adds 105 saved 474-task verdict vectors (49,770 entries) for the two completion lineages, retaining state aliases, repeats and paired task identities. Its separate module recomputes their replay, single-edit and factorial contrasts, then pools all eight lineages with the original six results. Saved candidate records support the single-edit selection check; this does not rerun the search or audit task responses (App. E.3).

The first-layer calibration records contain 252 aligned pool vectors with 97,524 saved task verdicts. The internal reconstruction recomputes its six-block contrasts, four qualification conditions and target/non-target partition, keeping the shared baseline and five repeated measurements explicit (App. E.6). An anonymized recovery projection distinguishes 97,528 physical sends, four lost responses and six cached-response recoveries. Numeric request rows reconstruct usage and reservation charges; authorization, device occupation and complete-payload checks remain documentary projections of private receipts. This reconstruction does not replay full request payloads or authenticate the original scorer.

For the earlier 27B diagnostic, the internal reconstruction checks 96 local condition-repeat cells, 56 physical calls and 24 response aliases from event records. The remaining 16 synthetic C0/C1 cells make no new request. Saved 20×4 score matrices support the original P3.0 acceptance calculation, and recorded fullspecs support its persistence and descendant checks. The old 3,960 condition evaluations and zero triggers are reproduced by summing five saved scope summaries, not by replaying all 616 original episodes. The reconstruction re-executes the fixed state-selection rule on exported structural facts and checks saved request transformations; it does not regenerate model responses or evaluate full task outcomes.

The full-task E2 reconstruction has three separately executed modules. The primary module reconstructs all $74 \times 4 \times 2$ outcomes, fixed-index bootstrap intervals, guard counts, chronological request-prefix witnesses and costs. The descriptive module retains the original 20-task subset and finish-reason summaries. That subset was fixed during execution, after 25 total cells and 8 subset cells, rather than in the initial protocol. The action module compares response and adoption hashes for 220 fixed first-response witnesses spanning 334 unique response records, without selection on task success. These post-result descriptors do not recover omitted message or argument text from its hashes. All three modules were run from the extracted internal archive (App. F.1); the historical 3,960 checks above remain a separate summary reconstruction.

Post-review sensitivity component. The original twelfth component adds the common-service eight-parent $L - S$ interval and the two Qwen3.5 capacity-group means from saved verdicts and an eight-row count projection (Apps. E.10, E.9). These additions follow internal review and preserve the original eleven components' numerical results. They introduce no searches, seeds, task scoring, subgroup intervals or significance tests. The three source snapshots and natural-prefix replay remain separate from this numerical runner.

Three additional evidence reconstructions. The known-leak component counts saved normalized-match annotations for 606 disclosures, 2,400 calls, 2,351 valid candidate events and 60 checkpoints, and recomputes the 27×300 paired candidate/incumbent differences. Raw text matching was checked separately against the private archive; the numerical component does not repeat that match, recover missing solver requests or establish functional consumption or positive calibration. The complete-evaluation-ceiling component reselects endpoints from all 8×400 original S audit candidate summaries, lineage budget-event projections and old terminal vectors. Its separate candidate-only diagnostic leaves one selected endpoint unscored in that old archive. The terminal reconstruction uses the newly measured 17 physical panels (8,058 verdicts), with 24 logical labels and seven explicit aliases. It recomputes the eight paired contrasts, uncertainty, seed-20260863 win/loss/tie counts and cost projections (App. E.9). This component uses only its new service batch; it neither imputes an M_0 panel nor authenticates the original scorer or physical execution. Request-

cost projections reconstruct usage, while GPU allocation remains documentary metadata. These CPU reconstructions make no model calls and add no independent seeds or tasks.

Saved execution identifiers. Table 31 maps semantic descriptions to archived execution labels. Identifiers describe evidence provenance, not independent replicates; rows sharing an algorithm seed remain dependent.

Table 31: Execution identifiers and evidence units. Internal CPU reconstruction uses the subsets described above; full archives remain separate.

Evidence	Saved execution identifier	Algorithm seed / unit
Earlier P/N pair	`DIRECT-PNW-20260914-02`	20271108
Repaired P/N pair	`DIRECT-PNW-REPAIRED-20260916-02`	same 20271108
Local mechanism diagnostic	`PNW-MECH-LOCAL-20260917-01`	states from 20271108
Paired task-level pilot	`ROUTE1-TASKQ-20260918-01`	74 paired TRAIN tasks
Unqualified calibration	`ANA-B27-MEASUREMENT-01`	calibration contexts
Four-round history follow-up	`HV4B-FOLLOWUP-20260918`	$20261001 + 7k$, $k = 0{:}7$
Earlier history pilot	`HV4B-20260917-01`	same eight parents
4B measurement completion	`TABLE1-COMPLETION-20260923-01`	existing $20260842/70$
4B direct-injection control	`PC4B-TWO-LAYER-20260923-01`	six assignment blocks
Control recovery segment	`PC4B-L1-RECOVERY-20260924-01`	same fixed matrix
Second 4B model/serving search	`CROSSMODEL-Q35-20260925-01`	eight reused seed identities
Candidate-count terminal follow-up	`Q35-COUNTMATCH-TERMINAL-20260925-01`	same eight seeds, 17 physical panels
Four-condition telecom panel	`E2_ATTRIBUTION_20260925_01`	74 TRAIN tasks, 4×2 conditions/repeats

Binding limits. The inspected 48 W2/W4 metadata records match their saved configuration bytes. They record a served-model alias; they do not authenticate physical weights. Current local source hashes are recorded separately from historical deployment identity. A bounded archive check did not recover a source digest or complete request body binding those historical 4B deployments to the current source. The earlier 27B local configuration, tokenizer identifiers and template match the recorded parent bindings. An earlier deployment inventory supplies 29 file hashes, including 18 weight shards; 11 recorded comparisons across eight metadata files agree. No diagnostic-time rehash authenticates those weight bytes, and the upstream repository and revision remain unknown (App. F). Before the new full-task E2 run, the current agent and user assets were hashed, including all 18 and 17 weight shards, respectively (Table 24). These current-byte bindings neither establish the public upstream checkpoint nor retroactively authenticate earlier deployments. Retained machine-readable identity records preserve this distinction. These are explicit provenance limits, rather than observed disagreements in the numerical records.

The retained `CLAIM_EVIDENCE.csv`, component provenance notes and W4 primary-role sidecar distinguish task-level reconstruction, seed-level reconstruction, event checks and summary arithmetic. The sidecar maps the historical aggregate's generic AB label to the registered R5–B primary while preserving the original file. These internal checks do not reproduce every exploratory analysis or replace the complete saved research archive.