POSITION: REVALIDATING PROBES AND STEERING VECTORS IN SCAFFOLD-EVOLVING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural interventions primarily measure whether an evaluation changes local agent behavior, but not whether that effect persists after an agent updates, or how the resulting answer is used. This position paper argues that intervention validity must distinguish *probe prediction, local intervention effect, and final-answer effect* as separate evidence levels. For *partial records* with valid candidate labels and bounded *selector* probabilities, we characterize a sharp interval of compatible *final-answer effects* whose width depends on selector uncertainty and candidate disagreement; indistinguishable observations cannot remove the remaining *ambiguity* through additional logs of the same kind. For *complete records*, conditional downstream replay reassesses the effect by executing the updated suffix on each intervention arm while preserving a justified unchanged prefix, subject to preserved upstream laws, a sufficient cut without feedback, and faithful conditional execution. To operationalize this position, we specify a controlled GSM8K study with Qwen2.5-7B-Instruct and judge-only prompt optimization to assess *intervention effects* and *reconstruction, independent-fresh agreement, and total cost*. The analysis separates sampling uncertainty from *ambiguity* requiring new information or renewed computation.

1 INTRODUCTION

Neural interventions have become an important tool for controlling and analyzing language-agent behavior *via* steering sensitivity (Tan et al., 2024; Braun et al., 2025), probe robustness (Kramár et al., 2026), causal reliance on experience (Zhao et al., 2026b), persistent prompt optimization (Agrawal et al., 2026), activation-based agent memory (Zhao et al., 2026a), and multi-agent monitoring (Rose et al., 2026; Das et al., 2026). These settings increasingly allow an agent to update downstream components while retaining the same agent. A steering vector's effect may therefore differ from its effect on the answer the agent returns. The practical question is *what evidence identifies the intervention's final effect after this downstream change?* Prior intervention works primarily provide evidence about local agent behavior . For a fixed agent, evaluation may consider *probe prediction*, the *local intervention effect*, or the *final-answer effect*. These quantities are related but distinct, where a *probe prediction* is not an external outcome, a *local intervention effect* does not determine the *final-answer effect*, and a downstream consumer can alter how an intervened answer is used. We therefore focus on the effect of a *fixed neural intervention under an updated consumer* (downstream component that uses the agent's answer), rather than discovery of a correctness circuit or a new theory of explanation. For instance, consider an analytic two-candidate agent in which, without steering, both answers are wrong, while with steering only the agent is correct. The old judge always selects the reviser; the updated judge selects the agent exactly when an unrecorded persistent bit is one. Both worlds thus have the same recorded candidates, local labels, old outcomes, and judge code, yet the updated intervention effect is zero or one. Repeating the same logs does not reveal the bit; observing it or executing the correct updated agent does.

The example separates three levels of intervention evidence, where *probe prediction* describes what an internal measurement predicts, *local intervention effect* describes how the intervention changes the agent, and *final-answer effect* describes how the intervention changes the answer returned by the agent. A change at one level does not by itself identify a change at another level after the downstream consumer is updated.

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.

The same records can therefore support different *final-answer effect* when relevant selector state is unobserved. For *partial records* with valid candidate labels and bounded *selector* probabilities, we characterize a sharp interval of compatible *final-answer effect* within a stated selector class. Its width depends on selector uncertainty and candidate disagreement; indistinguishable observations imply that additional logs of the same kind cannot remove the remaining *ambiguity*. Thus, estimation from additional samples and identification from additional information are distinct questions. For *complete records*, conditional downstream replay provides constructive reassessment by executing the updated suffix on each intervention arm while reusing a justified unchanged prefix, subject to preserved upstream laws, a sufficient cut with no feedback across it, and faithful conditional execution. This evaluates the *updated consumer* under each intervention arm rather than carrying forward an old effect or stitching an old continuation to a changed downstream component.

> Our position is therefore that intervention evaluation must distinguish *partial-record identification* from direct reassessment of the updated agent. When selector state or other consumed information is missing, the appropriate conclusion may be an identified interval rather than a point estimate. When the required computation and inputs are retained, the *updated consumer* can instead be executed directly on the affected intervention arms.

To operationalize this position, we specify a controlled E1/E2 study with Qwen2.5-7B-Instruct[1] and judge-only prompt optimization. E1 assesses *local and final intervention effects*, while E2 assesses *reconstruction, independent-fresh agreement, and total cost* in the complete-record setting. The study uses GSM8K and changes only the judge instruction. The analysis therefore distinguishes sampling uncertainty from *ambiguity* requiring new information or renewed computation, while *reassessment* depends on preserved computation, inputs, and valid conditional execution.

2 Setup and the Evaluation Question

A scaffold version S_t fixes persistent instructions, memory, routing, tools, and context construction around weights θ; we model its execution as a bounded, time-unrolled process with temporally ordered call slots and token-level execution where required (see Appendix A.2). An external optimizer changes only the judge instruction, defining persistent scaffold evolution rather than a one-episode revision or autonomous whole-agent redesign. The selected version and neural instrument are fixed before independent assessment. For question X, let $O^a_{t,k_\star}$ denote the solver's answer and Y^a_t the returned answer under intervention arm $a \in \{0,1\}$ (see Appendix C.3). With fixed external correctness scores $Q_{\mathrm{pred},t} = \Phi(P_t(h,L);q_0)$, $\tau_{\mathrm{local},t} = \mathbb{E}_t[m(O^1_{t,k_\star},X) - m(O^0_{t,k_\star},X)]$, and $\tau_{\mathrm{final},t} = \mathbb{E}_t[\ell(Y^1_t,X) - \ell(Y^0_t,X)]$, the frozen probe $q_0(h) \in [0,1]$ observes an unperturbed residual at a specified input-token anchor; L denotes local or final correctness, and Φ a discrimination or calibration metric. Thus, *probe prediction, local intervention effect, and final-answer effect* are distinct evaluation objects (Hewitt & Liang, 2019; Ravichander et al., 2021; Lin & Liu, 2026). The intervention fixes its block, direction, sign, residual-relative scale, and answer-token mask (see Appendix A.1 and A.3). A complete pre-intervention state includes rendered inputs, positions, masks, and relevant runtime/cache state. Equal weights alone do not fix numerical execution (Yuan et al., 2026); the conditional local response is therefore a property of the aligned assay and its state law, not the weight file alone (see Appendix C). Let V^a denote a complete record before the judge in arm a and W^a a partial record that may omit selector state while retaining valid candidate labels (see Appendix A.5). The evaluation problem therefore separates *identification* (which population effects are compatible with the available information), *estimation* (how precisely finite records determine an identified quantity), and *reconstruction* (whether an implementation reproduces its declared calculation). These questions require different evidence.

> For *partial records*, valid candidate labels and bounded *selector* probabilities determine a sharp interval of compatible emphfinal-answer effect. Its width depends on selector uncertainty and candidate disagreement; under indistinguishable observations, additional logs of the same kind cannot remove the remaining *ambiguity*. For *complete records*, conditional downstream replay instead executes the updated suffix on each intervention arm while reusing a justified unchanged prefix. This requires preserved upstream laws, a sufficient cut with no feedback across it, and faithful conditional execution.

Interpretations versus Intervention Effects. Causal abstraction, interpretive equivalence across networks, and compositional interpretability formalize relationships between explanations and model computation (Geiger et al., 2025; Sun & Toneva, 2026; Gauderis et al., 2026). Our object is narrower,

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

where the effect of a specified activation operation after a persistent consumer update. Experience-faithfulness interventions (Zhao et al., 2026b) and continuation resampling (Macar et al., 2026) already study causal use of experience and reasoning. Neither the distinction between *local and final effects* nor the *execution of affected continuations* originates here.

Replay and Partial Information. Matched controller replay compares decisions on fixed evidence (Chhikara, 2026), while CAR executes counterfactual descendants for failure attribution (Shah, 2026). Bandit replay and doubly robust off-policy evaluation use logged action feedback under their support assumptions (Li et al., 2011; Dudík et al., 2011; Jiang & Li, 2016). Our selector has two externally graded candidates and an executable updated policy, with the target given by the difference between two neural-intervention arm means. Static continuations after a model switch can answer the wrong query (Gonuguntla, 2026); *conditional replay* executes the affected suffix rather than stitching an old continuation to a changed downstream component. Missing state or changed numerics can invalidate this execution. Sharp bounds under incomplete overlap (Khan et al., 2024) and observation-design analysis for LLM prompts (Chen et al., 2026) establish limits imposed by missing information. Our result specializes established partial-identification reasoning to *selector uncertainty* and score-relevant *ambiguity*. §6 compares information, operation, and target across related regimes without claiming unmeasured superiority.

3 What Incomplete Records Identify

Condition on pre-evaluation selection, where selector code and recorded-state laws are fixed, but omitted selector state can vary. The following rectangular class bounds the conditional means; additional restrictions from a fixed executable program may narrow it. Let W^a have fixed law P_a on a common measurable space. Correctness $c_S(w), c_R(w) \in \{0,1\}$ is known from the record and its evaluator-only sidecar. Define $\Delta c = c_S - c_R$ and $\pi_a(w) = \mathbb{P}(\text{updated selector chooses agent} \mid W^a = w, a)$, which averages omitted state under its target conditional law. It need not equal the probability returned by a replay program that supplies a default missing value. Candidate-generation laws and labels are fixed here; corrupted labels and arm-spliced histories are outside this uncertainty model.

A Sharp Score-Sensitive Interval. Suppose justified measurable bounds $0 \le \underline{p}_a \le \pi_a \le \overline{p}_a \le 1$ are the only restrictions on selector means, where any measurable choice within each range is allowed, with no cross-arm restriction (the rectangular class). Put $s_a^- = c_R + \min\{\underline{p}_a \Delta c, \overline{p}_a \Delta c\}$ and $s_a^+ = c_R + \max\{\underline{p}_a \Delta c, \overline{p}_a \Delta c\}$. For known P_a, c_S, c_R, the sharp identified interval and width are $I_\tau = [\mathbb{E}_{P_1} s_1^- - \mathbb{E}_{P_0} s_0^+,\ \mathbb{E}_{P_1} s_1^+ - \mathbb{E}_{P_0} s_0^-]$ and $\omega = \sum_{a=0}^{1} \mathbb{E}_{P_a}[(\overline{p}_a - \underline{p}_a)|\Delta c|]$. The endpoints are attainable, and the effect is point-identified exactly when $\omega = 0$.

Why the Bounds Are Sharp. The conditional score $c_R + \pi_a \Delta c$ is affine in π_a. Rectangularity permits opposite arm extrema simultaneously, and convex mixtures fill the interval. Nonnegativity yields the zero-width criterion; Appendix B.5.1 gives the proof. Sharpness is relative to this class; for a restricted neural-judge family, the interval may only be an outer bound. Estimated laws or bounds add sampling uncertainty (Manski, 2003; Khan et al., 2024). Without additional information, $[0,1]$ is valid. Narrower envelopes require justified restrictions on omitted states and their selection laws, not default-state outputs or sample extrema (see Appendix B.5.1). Figure 1 illustrates the difference.

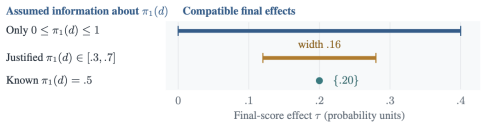

Figure 1: Analytical construction, not model data or confidence intervals. Let $P_1(d) = .4$, $P_1(e) = .6$, $P_0(e) = 1$, with $(c_S, c_R)(d) = (1,0)$ and $(c_S, c_R)(e) = (0,0)$. Then $\tau = .4\pi_1(d)$, giving $[0, .4]$, $[.12, .28]$, and $\{.20\}$ under the three assumed information sets. The middle width is $.4 \times .4 = .16$. Rows change identifying information, not sample size or E2a outcomes.

Shared Selectors Can Identify a Difference Without Identifying Its Means. Suppose additional structure justifies the *same conditional mean* $\pi(w)$ across arms on their overlapping supports, bounded by common $\underline{p}(w), \overline{p}(w)$ and otherwise unrestricted. Let $\lambda = P_0 + P_1$, $f_a = dP_a/d\lambda$, $r = \Delta c(f_1 - f_0)$, and $C_R = \mathbb{E}_{P_1} c_R - \mathbb{E}_{P_0} c_R$. Combining arm integrals before taking extrema gives $I_{\tau,\text{shared}} = \left[C_R + \int \min\{\underline{p}r, \overline{p}r\}\, d\lambda,\ C_R + \int \max\{\underline{p}r, \overline{p}r\}\, d\lambda\right]$, with width $\int (\overline{p} - \underline{p})|r|\, d\lambda$. For example, if $P_0 = P_1$, $c_S = 1$, $c_R = 0$, and a common $\pi \in [0,1]$ is unknown, the effect is zero while both

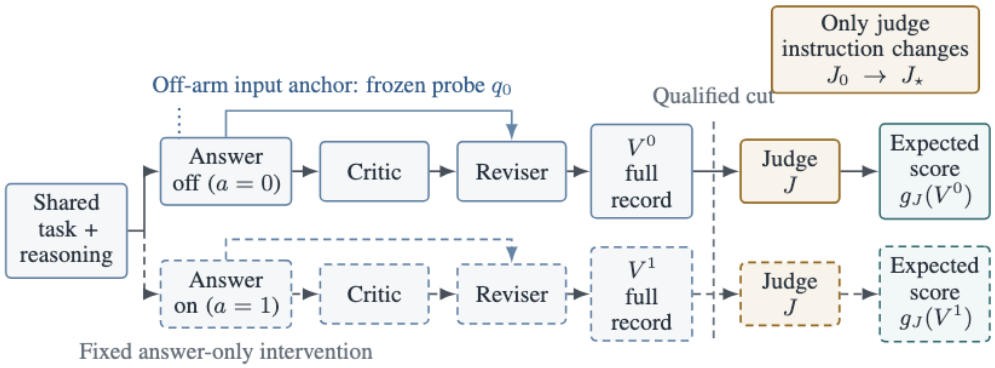

Figure 2: Complete-record reassessment (design, not a result). Shared reasoning precedes separate off/on answer, critic, and reviser histories. Each full arm record supplies the changed judge under the qualified replay cut; external grading yields the expected score.

equal arm means can be anywhere in $[0, 1]$. Thus, agreement of a contrast alone does not establish agreement of its components. A sufficient condition is the same selection law given (W, U) and the same conditional law of omitted U given W across arms. Shared code alone is insufficient; these conditions are sufficient, not necessary for equal means. This analytic refinement is neither assumed nor measured in E1/E2.

Why More of the Same Logs Cannot Resolve the Ambiguity. Suppose admissible endpoint worlds with effects τ_-, τ_+ induce the same joint law of every record and allowed query result available to an estimator T, including its independent auxiliary randomness. Then $\max_{b\in\{-,+\}} \mathbb{E}_b|T - \tau_b| \geq \frac{\tau_+ - \tau_-}{2}$. For every output t, the two absolute errors sum to at least the endpoint gap; identical output laws give the bound in expectation. It holds at every sample size for which indistinguishability holds. The introduction's unobserved-bit worlds give a lower bound of $1/2$ when only those logs, code, and default-state queries are available. Observing the bit or using a correct fresh runner changes the information set and escapes that example. Attainable endpoints alone do not ensure indistinguishable allowed queries. The midpoint attains half-width risk when the interval is known exactly; estimated endpoints add uncertainty. This is the exact-indistinguishability two-point argument (Yu, 1997), scoped to the available information. New information can separate the worlds; more copies of indistinguishable logs cannot.

4 Valid Reassessment from Complete Records

The constructive case starts when the updated consumer can be executed on sufficient arm-specific inputs (see Figure 2). Let V^a contain the question, candidates, critique, revision, canonical mapping, and other consumed state, with P_0^a its old-version law. For updated judge J, $g_J(v)$ denotes expected final score under its correct conditional execution law. Thus, *complete records, conditional replay, and reconstruction* provide the evidence needed to reassess the final effect after the downstream consumer changes.

Conditional Downstream Replay. Condition on training and optimization information $\mathcal{F}$, which fixes the selected judge J. Final questions and record-generation randomness are independent of $\mathcal{F}$ under the sampling design. If (i) each arm's upstream law is invariant, (ii) no update-induced feedback crosses the cut, (iii) V^a is sufficient, and (iv) the new downstream execution and scoring laws are reproduced (see Appendix B.1 and C.2), then $\tau_{\text{final},J} = \int g_J(v)P_0^1(dv) - \int g_J(v)P_0^0(dv)$. For equally weighted questions, $\widehat{\tau}_J = n^{-1}\sum_i [g_J(V_i^1) - g_J(V_i^0)]$ is unbiased over correctly sampled records; other sampling designs require their declared task weights. This standard composition/transport identity (Pearl & Bareinboim, 2022) integrates the correct conditional calculation under each preserved arm law. Replay produces *new downstream evidence*, rather than carrying forward an old effect or splicing an on-arm answer into an off-arm continuation. Appendix B.2 gives the proof.

An Executable Conditional Mean. The study's judge selects between two single-token identifiers at temperature one with all other tokens masked. If $p_J(v)$ is its solver-selection probability after undoing candidate ordering, then $g_J(v) = p_J(v)c_S(v) + [1 - p_J(v)]c_R(v)$. Both candidate labels

are known to the external scorer, never supplied to the judge. A separate keyed draw yields its binary final label; fractional expectations are not labels for AUROC. Integrating the final choice removes that draw's noise, not upstream or question uncertainty (see Appendix B.3). For a generative suffix, the same proposition permits averaging full conditional continuations; E1/E2 retains the categorical agent in both evaluation routes.

Connecting the Two Regimes. In this two-candidate setting, take $W^a = V^a$ and $P_a = P_0^a$. When the target probability is known and faithfully evaluated, $\underline{p}_a = \overline{p}_a = p_J$, so Proposition 3 collapses. This is the categorical specialization, not all of Proposition 4. Zero width can also follow from equal candidate scores or, in the shared class, cross-arm cancellation; neither implies a complete record. Estimation still has finite-sample uncertainty.

From Assumptions to Qualification. Three obligations connect the result to execution, *retained law* (preserve upstream prompts, code, sampling, and batch regimes, with no update-induced feedback across the cut), *sufficient input and execution* (retain each arm's complete consumed values and reproduce the changed suffix, score, and conditional randomness), and *independent assessment* (freeze the instrument and selected judge, with evaluation sealed from their selection). Table **??** maps these obligations to Proposition 4 and the aligned-assay prerequisites; the six dependency checks are detailed in Appendix D.3. Retained per-record inputs and fixed tokenizer/decoder parameters jointly define execution; only the future independent selector draw is integrated, while correlated latent randomness needs its correct conditional law. A neural prefix cache must not skip an activation intervention (see Appendix C.2). Failed hooks, lineage, or scoring are measurement faults, not zero effects. Appendix D gives the operational claim records and decisions; the claim, update, and assessment record structure is detailed in Appendix D.1. With valid labels and the stipulated record laws intact, selector uncertainty permits §3's *bounds*. Other state-law/kernel discrepancies require the separate premises in Appendix B.5. Finite *reconstruction* alone does not justify population bounds. Known faults require repair or regeneration from the earliest affected boundary; E2 separates *fidelity, fresh agreement, and cost*.

5 Controlled E1/E2 Study

E1 measures a fixed solver intervention under independently optimized judges, while E2 checks the executable *complete-record method*, not the population identification width of §3. The BF16 Qwen2.5-7B-Instruct task agent and separately loaded 14B reflector share one A100 (Hui et al., 2024; Qwen Team, 2024a) (see Appendix E.3). GSM8K supplies external numeric grading (Cobbe et al., 2021); gold never enters evaluation prompts. Only the judge instruction evolves in the solver-critic-reviser-judge pipeline (see Appendix E.2).

E1: Six disjoint training pools contain 64 development, 768 fitting, 256 tuning, 256 confirmation, 128 adaptation, and 256 selection questions (see Appendix E.1). The 1,024-question final panel uses one upstream draw per question with paired intervention arms. Nested subsets contain 128 matched and 512 independent-fresh questions, expandable to 1,024 only under the pre-main precision rule expandable to 1,024 only under the pre-main precision rule (see Appendix F.4).. Fitting-only grouped validation screens blocks 7, 14, 21, and 27 and retains two. Tuning chooses the primary logistic correctness probe and, separately, one of at most 32 layer/family/sign/scale settings from raw probe normals or mean differences, signs $\pm$, and norm fractions .01, .03, .05, .10. Selection maximizes absolute paired local effect subject to invalid-answer increase at most .01, with ties toward smaller scale. The operation acts at the answer call's last valid prefill token and active decode positions (see Appendix A.4). Untouched confirmation, off/sham, two random directions, and opposite sign assess the frozen bundle without retuning. Primary probe measurements use the off-arm pre-answer anchor (see Appendix C.1); Appendix E.4).

E2: Each optimizer seed has three rounds, two children per round, and an always-eligible incumbent. Only the judge instruction, at most 512 tokens, changes. Selection maximizes unsteered expected correctness on the fixed selection cache, with ties retaining the incumbent. Adaptation records guide the separate reflector. The final seed-zero judge is primary, seed one descriptive replication. Appendix E.5 fixes the exact budget and records whether implementation uses bounded GEPA or GEPA-inspired search (Agrawal et al., 2026; DSPy Contributors, 2026). No probe, steered, confirmation, or final outcomes select a judge. Appendix E retains the full allocation, conditional fitting reserve, output caps, and reduced resource tier, and Appendix E.9 gives the L0-L3 locks.

The position separates *intervention measurement* from *complete-record reassessment*. E1 fixes the solver intervention and evaluates it under independently optimized judges, while E2 checks whether complete-record execution reproduces the declared downstream calculation. Neither estimates the population identification width of §3.

Table 1: Related evaluation regimes, methodological distinctions, not measured comparisons.

Method / Source	Available Evidence and Operation	Target and Relationship
Bandit replay (Li et al., 2011)	Randomized logged actions and received rewards; retain matching target actions.	Policy reward with partial action feedback, unlike our externally graded candidate pair.
Doubly robust OPE (Dudík et al., 2011; Jiang & Li, 2016)	Reward/value models plus propensity or trajectory information; weighted residual correction.	Off-policy value under stated support and model assumptions; does not recover omitted state by itself.
CAR (Shah, 2026)	Structural agent model, step interventions and forward execution; constructed causal validation.	Failure attribution. Our target is the same neural intervention under a selected persistent version; affected descendants still execute.
Matched controller replay (Chhikara, 2026)	Fixed evidence paths and answer/confidence states; change confidence-to-action mapping.	Controlled downstream-use comparison; live queries/answers caused by changed actions are not newly generated.
Replay Gap (Gonuguntla, 2026)	Static model-switch evaluation compared with live forks.	Exposes inappropriate retained continuations; supports boundary validation, not a universal ban on replay.
PROTEA (Kawamura et al., 2026)	Node traces/evaluators, inferred intermediate targets, prompt patches and workflow re-evaluation.	Diagnose/refine workflows; inferred node targets are not our external numeric labels.
SEAGym (Zheng et al., 2026)	Saved harness snapshots, held-out views, previous-task and cost records.	Evaluate evolution and retention; snapshot re-evaluation is distinct from a fixed-prefix/new-suffix estimator.
This study	Complete on/off candidate records, both external labels and an executable updated selector.	Re-estimate its *final intervention effect*, with separate *reconstruction*, fresh agreement and cost checks.

6 Replay regimes and qualification evidence

Table 1 compares methods by their data and query, not by the word "replay", and the methods need not have comparable bias/variance because they need not estimate the same quantity. Our activation operation remains an *intervention*, specifically the state-dependent policy, not merely observational monitoring or constant-node replacement. Both candidate labels are externally available at the final choice, not for every possible upstream action. MIPS uses propensity weights marginalized through action embeddings (Saito & Joachims, 2022), while MDR combines such weighting with reward-model correction under its stated assumptions (Shimizu & Forastiere, 2023); these partial-feedback estimators do not merely integrate a known two-candidate score and supply no automatic correction for arbitrary omitted state. A changed upstream law can instead require a sufficient common state, overlap $P'_a \ll P_a$, and a valid density ratio, so weighted or doubly robust evaluation needs that additional evidence (Dudík et al., 2011; Jiang & Li, 2016) and cannot recover omitted state or undo selection leakage. E2b keeps the uncorrected comparison. Extraction-source studies (Ye et al., 2026) and activation-based procedural memory (Zhao et al., 2026a) motivate provenance, not a transfer guarantee for our fixed vector. SteerBoost uses early steered tokens and an unsteered pass over the same sequence on ASTEER (Fan et al., 2026), but is neither a pre-generation replay certificate nor a replacement for the frozen assay.

The distinction is between the *changed computation being evaluated* and the information available to evaluate it. Propensity weighting or reward correction can address specified partial-feedback settings, but neither establishes the missing state or validates an affected continuation. E2b therefore retains the uncorrected target rather than introducing an additional correction model.

Interpretability and Update Context. Steering transfer/brittleness (Tan et al., 2024; Braun et al., 2025; Bigelow et al., 2025), predictive access without correction in clinical triage (Basu et al., 2026), and identification/alignment limits (Lin & Liu, 2026; Geiger et al., 2025; Méloux et al., 2025; Sutter et al., 2026) motivate separate targets. Patching sensitivity (Zhang & Nanda, 2024; Makelov et al., 2024; Wu et al., 2024), compensation (McGrath et al., 2023), and layer/direction dependence (Billa, 2026; Marks & Tegmark, 2023) motivate calibration. Memory/workflow methods (Xu et al., 2026; Zhang et al., 2025) and misevolution, aging, and attention-guided refinement (Shao et al., 2026; Zhu et al., 2026; Hong et al., 2026) already study persistent change. Adaptive-selection theory supports independent assessment (Russo & Zou, 2016), while monitoring/evasion work concerns additional risks (Bailey et al., 2026; Baker et al., 2025). Claim records build on validity, documentation, regression testing, and hidden-consumer analysis (Kane, 2013; Mitchell et al., 2019; Rothermel & Harrold, 1998; Sculley et al., 2015), while identity checks do not certify benign vector construction (Aidakhmetov et al., 2026).

Table 2: E1 measured results. Accuracies and BA are percentages; ATEs are percentage points.

Panel / Judge	Actual N	Expected accuracy off / on	ATE [CI]	AUROC
Shared solver/probe	1,024	75.49 / 87.30	$+11.82$	—
J_0	1,024	70.53 / 71.33	$+0.81$	0.8926
$J_\star^{(0)}$	1,024	72.95 / 74.34	$+1.39$	0.8982
$J_\star^{(1)}$	1,024	72.95 / 74.34	$+1.39$	0.8982
$J_\star^{(0)} - J_0$	1,024	—	$+0.59$	—
$J_\star^{(1)} - J_0$	1,024	—	$+0.59$	—

Table 3: E2 measured results. A checks reconstruction in the production batch regime. B compares $J_\star^{(0)}$ on common questions using integrated judge scoring. ATE and arm-mean differences are percentage points.

Quantity	Actual N	Estimate [CI]	Status
Matched reconstruction	128	0.0000 probability error	0 violations / 0 fallbacks
Token / event matches	128	128/128	Exact match
Replay / fresh final ATE	512	$+2.20$ / $+2.20$	Estimation-only
Replay minus fresh ATE	512	0.00	Exact paired zero
Off-arm mean difference	512	0.00	Exact paired zero
On-arm mean difference	512	0.00	Exact paired zero
Precision mode	512	Estimation-only at L3	Before final outcomes
Marginal / all-in cost	—	— / $\approx$786.9 SU	Same optimized fresh baseline

7 Empirical Assessment: Effects, Fidelity, and Utility

For a judge update define $\Delta p^a = p^a_{J_\star} - p^a_{J_0}$ on each same-arm record. Then $\Delta\tau_{\text{final}} = G_1 - G_0$, with $G_a = \mathbb{E}[\Delta p^a (c^a_S - c^a_R)]$, where G_0 and G_1 are the unsteered and steered accuracy gains. When $G_0 > G_1 > 0$, both arms improve while the *signed steering contrast* decreases. These gains have descriptive paired intervals, not a mediation interpretation. Candidate disagreement limits opportunity, with $|\Delta\tau_{\text{final}}| \le d_1 + d_0$ and $d_a = \mathbb{P}(c^a_S \neq c^a_R)$ (see Appendix C.4). The full population stays in the estimate; solver-copy/reviser-copy diagnostics cannot replace a learned update (see Appendix B.4). Result interpretation must distinguish weak original support, an exact prompt no-op, and an imprecise or small effect under a changed judge. None licenses selecting another checkpoint, direction, or favorable question subset.

E1: Each question's pre-intervention reasoning is generated once. Answer, critic, and reviser then run separately within each arm; the reviser consumes the agent, so choosing it need not remove agent influence. Complete records are scored by J_0 and both selected judges (see Appendix E.6). Table 2 reports *local evidence* once with *final-answer effects* on the same questions. Independent confirmation assesses whether the original instrument is supported; unchanged local computation across judge versions is a design property, not an empirical retention discovery.

E2: E2a regenerates the fixed 128-question subset from clean state in its recorded production microbatches and role streams. Consumed inputs, token IDs, mapping, and hook events are checked separately from numerical discrepancies (see Appendix E.7). The proposed probability-error target 10^{-6} must be qualified before final execution. On aligned records, Appendix F.1 bounds observed score error by candidate-weighted probability error. This finite-record bound neither certifies unseen states nor measures population $\omega = 0$. E2b generates new upstream draws on the locked 512 or 1,024-question subset, applies the same primary judge, and compares the ATE and both arm means on common questions (see Table 3); Appendix C.5 and E.8).

The cancellation example in §3 shows why agreement of a contrast alone is incomplete evidence; it does not assume the study has that example's shared conditional selector. The primary replay estimate remains uncorrected by fresh outcomes.

Uncertainty and Interpretation. Question-level paired contributions preserve covariance, with 9,999 task-cluster bootstrap draws and batch-cluster sensitivity. Appendix F.1 details these estimands, pairing, and covariance calculations. The Holm-controlled family contains the signed local effect (direction frozen at L2), the two-sided seed-zero final-effect change, and, only if precision-qualified, joint TOST equivalence for ATE and both arm errors within $\pm.05$ (Holm, 1979; Lakens, 2017). Appendix F.2 specifies this confirmatory family and its evidence tiers in full. A 128-question training-side two-draw assessment chooses fresh size/mode before final outcomes, with at least 80% conservative planning power (see Appendix F.3) at 512, otherwise 1,024 if feasible, and otherwise 512 in estimation-only mode. The margin never expands to obtain a favorable result. Appendix F retains all tests, Appendix F.5 the sparse/missingness rules, and Appendix F.7 the conditional resource branches. E2a is a separate *fidelity* check, not another statistical claim.

Practical Utility and Rejected Boundaries. The fresh comparator also batches, shares its own pre-intervention reasoning, and integrates selection. Cost (Appendix F.6) includes prefills, generation, construction/I/O, loading, checks, and human review; marginal updates, first use, and all-in validation remain separate. Known invalid cases include an upstream edit, arm-spliced critique, missing consumed state, and stale mapping. Two humans assess distinct cases blind to effects; raw disagreement and review time are reported, not population failure-detection rates Artifact identity and provenance checks support integrity and accounting but do not certify benign vector construction or poisoning resistance (see Appendix D.2). *Fidelity, fresh agreement, and cost* support distinct conclusions. A precisely estimated zero can meet an agreement target. Agreement does not establish complete information, and valid replay need not have an all-in cost advantage.

8 Implications and Limits

When Simpler Evidence Suffices. Robust probes and monitoring-aware agents provide genuine *transfer evidence* in tested settings (Kramár et al., 2026; Das et al., 2026), and a covered claim can reuse that evidence. Ordinary task evaluation suffices for unsteered accuracy. The additional burden arises when a neural predictor or intervention effect is relied upon. Neither final accuracy alone nor a local assay determines the updated final-effect contrast.

Which Information Changes the Answer. *Complete valid arm records* help reassessment by executing the changed consumer. *Partial selector records* can determine an interval, with *ambiguity* concentrated on unequal-correctness candidates. Observing missing policy state or acquiring correct fresh execution can distinguish worlds that further copies of old logs cannot. A justified shared selector may identify a difference without either mean; this does not make complete transcripts unnecessary for every claim.

Scope of the Result. The bounds require their stated admissible selector class, and the experiments concern one downstream-only math-agent update. They do not establish neural explanation equivalence, unrestricted evolution prevalence, or deployment safety. The probe can read difficulty or a candidate already present in reasoning. Prior text/prior, length–likelihood, and permutation comparisons (see Appendix E) assess incremental prediction, not a uniquely isolated correctness mechanism. The bounds assume valid candidate scores; parser faults or unknown true labels are outside that model. Selection independence and numerical qualification remain requirements, not consequences of fixed weights.

From Assumptions to Execution. Feedback requires executing every new dependent call from a sufficient checkpoint and moving the cut earlier if prefix construction changes (see Appendix D.4). This does not add retries, tools, or memory to E1/E2. Prospective decisions and costs remain auditable; finite *checks and artifact hashes* cannot establish universal semantic sufficiency.

9 Conclusion

Neural intervention evaluation must distinguish *probe prediction, local intervention effect, and final-answer effect* from *identification ambiguity* and *reconstruction*. When selector state or other consumed information is missing, additional logs of the same kind cannot remove the resulting ambiguity, and the appropriate evidence may be a sharp identified interval rather than a point estimate. When *complete arm records* retain sufficient inputs and preserve the upstream law, conditional downstream replay can execute the changed consumer and re-estimate the final effect, provided the cut, conditional execution, and reconstruction are qualified. Our E1/E2 design operationalizes these distinctions by freezing the intervention and selected judge before independent assessment, separating local and final effects, and treating *reconstruction, independent-fresh agreement, and cost* as distinct evidence. The empirical results illustrate that a local steering effect, a final-answer effect, and downstream judge gains are related but distinct quantities, while reconstruction and fresh agreement require their own evidence. These conclusions do not establish neural explanation equivalence, unrestricted evolution prevalence, or deployment safety. The central requirement is therefore to match each claim to the information and computation that identify it, with missing state requiring new information or renewed computation rather than favorable reuse of prior records.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish the writing of the manuscript. We have not used generative AI tools for generating or modifying experimental results, or scientific claims, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work and verified the final content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The proposed study uses public mathematical questions. Two NLP researchers (aged 25-30) were involved without compensation by male members of the same research group. A successful correctness probe is not a deployment-safety certificate, and failed correctness steering is not a general safety finding. Main integrity risks are overstated causal interpretation, selection of favorable updates, and presenting planned measurements as observations; sealed populations, separate calibration and optimization, and unrun fields address these risks.

REPRODUCIBILITY STATEMENT

Appendix A.1 and §6 specify the scientific design, proofs, controls, and qualification map. Appendix E.9 distinguishes L0-L3 stages, where actual approval dates, immutable implementation revisions, task manifests, selected prompts, fitted coefficients, and execution evidence must substantiate completion.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Abzal Aidakhmetov, Donato Crisostomi, Tommaso Mencattini, Adrian Robert Minut, Iacopo Masi, and Emanuele Rodolà. Steering vectors are an adversarial attack surface. *arXiv preprint arXiv:2606.05958*, 2026.

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses. In *International Conference on Learning Representations*, volume 2026, pp. 146838–146882, 2026.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Sanjay Basu, Sadiq Y Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. Interpretability without actionability: mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*, 2026.

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Xiaohong Chen, Ashesh Rambachan, and Elie Tamer. Partial identification from llm prompts. *arXiv preprint arXiv:2606.15031*, 2026.

Prateek Chhikara. Evaluating confidence-gated retrieval with matched trajectory replay. *arXiv preprint arXiv:2608.26846*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Aritra Das, Jaee Ponde, Mihir More, and Debayan Gupta. You can't escape your own activations: Evaluation awareness and multi-agent monitoring. *arXiv preprint arXiv:2609.03035*, 2026.

DSPy Contributors. dspy.GEPA: Reflective Prompt Optimizer. Official DSPy documentation, 2026. URL `https://github.com/stanfordnlp/dspy/blob/main/docs/docs/api/optimizers/GEPA/overview.md`. Accessed 17 September 2026.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Chenrui Fan, Yize Cheng, Ming Li, Soheil Feizi, and Tianyi Zhou. When is your llm steerable? *arXiv preprint arXiv:2606.11599*, 2026.

Ward Gauderis, Thomas Dooms, Steven T Homer, Kola Ayonrinde, and Geraint A Wiggins. From mechanistic to compositional interpretability. *arXiv preprint arXiv:2605.08934*, 2026.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Ashritha Gonuguntla. The replay gap: Static evaluation of model switching in llm agents scores the wrong world. *arXiv preprint arXiv:2608.08239*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Yechao Hong, Haiquan Qiu, Yaqing Wang, and Quanming Yao. Mechanistic attention guidance for agent memory refinement. *arXiv preprint arXiv:2607.17621*, 2026.

Hugging Face. Text Generation. Transformers documentation, 2026. URL `https://huggingface.co/docs/transformers/llm_tutorial`. Accessed 20 September 2026.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Michael T Kane. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73, 2013.

Kazuki Kawamura, Satoshi Waki, and Kei Tateno. Protea: Offline evaluation and iterative refinement for multi-agent llm workflows. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 27–35, 2026.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Forty-first International Conference on Machine Learning*, 2024.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for gemini. *arXiv preprint arXiv:2601.11516*, 2026.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Zezheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims. *arXiv preprint arXiv:2605.08012*, 2026.

Uzay Macar, Paul Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting llm reasoning requires resampling. In *International Conference on Learning Representations*, volume 2026, pp. 92567–92600, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, volume 2025, pp. 32900–32923, 2025.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

OpenAI. GSM8K Dataset Card. Official Hugging Face dataset card, 2021. URL `https://huggingface.co/datasets/openai/gsm8k/blob/main/README.md`. Accessed 17 September 2026.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 451–482. 2022.

Qwen Team. Qwen2.5-14B-Instruct Model Card. Official Hugging Face model card, 2024a. URL `https://huggingface.co/Qwen/Qwen2.5-14B-Instruct`. Accessed 17 September 2026.

Qwen Team. Qwen2.5 Speed Benchmark. Official Qwen documentation, 2024b. URL `https://qwen.readthedocs.io/en/v2.5/benchmark/speed_benchmark.html`. Accessed 17 September 2026.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability. *arXiv preprint arXiv:2604.01151*, 2026.

Gregg Rothermel and Mary Jean Harrold. Empirical studies of a safe regression test selection technique. *IEEE Transactions on Software Engineering*, 24(6):401–419, 1998.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240. PMLR, 2016.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. *arXiv preprint arXiv:2202.06317*, 2022.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

Jaineet Shah. Causal agent replay: Counterfactual attribution for llm-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Shuai Shao, Qihan Ren, Dongrui Liu, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, et al. Your agent may misevolve: Emergent risks in self-evolving llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 99728–99793, 2026.

Tatsuhiro Shimizu and Laura Forastiere. Doubly robust estimator for off-policy evaluation with large action spaces. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 992–997. IEEE, 2023.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. *arXiv preprint arXiv:2603.30002*, 2026.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? *Advances in Neural Information Processing Systems*, 38:149455–149503, 2026.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212, 2024.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

vLLM Contributors. Reproducibility. Official vLLM documentation, 2026. URL `https://docs.vllm.ai/en/latest/usage/reproducibility/`. Accessed 17 September 2026.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D Goodman. A reply to makelov et al.(2023)'s" interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, Juanzi Li, and Liangming Pan. Where steering signals come from: Activation source selection in activation steering. *arXiv preprint arXiv:2607.25270*, 2026.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in llm inference. *Advances in Neural Information Processing Systems*, 38:169819–169851, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *International Conference on Learning Representations*, volume 2025, pp. 34040–34077, 2025.

Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, and Kang Liu. Neural procedural memory: Empowering llm agents with implicit activation steering. *arXiv preprint arXiv:2606.29824*, 2026a.

Weixiang Zhao, Yingshuo Wang, Yichen Zhang, Yang Deng, Yanyan Zhao, Wanxiang Che, Bing Qin, and Ting Liu. Large language model agents are not always faithful self-evolvers. *arXiv preprint arXiv:2601.22436*, 2026b.

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.

Jianing Zhu, Yeonju Ro, John Robertson, Kevin Wang, Junbo Li, Haris Vikalo, Aditya Akella, and Zhangyang Wang. Your agents are aging too: Agent lifespan engineering for deployed systems. *arXiv preprint arXiv:2605.26302*, 2026.

A Appendix

A.1 Formal Setting and Measurement Contracts

Reading the Symbols. V^a is complete; W^a retains valid labels but may omit state. π_a averages that state's target law, p_J executes a full-record selector, and $\widetilde{p}_a$ may use an incorrect default. *Width ω is identification ambiguity*, not sampling error. Future draws are integrated; tokenizer/decoder policies are execution parameters. Exceptional-event probabilities, state-specific probability envelopes, and E2b's effect margin are distinct quantities.

A.2 A Bounded, Time-Unrolled Agent

Let an episode have at most K temporally ordered call slots, with token-level unrolling where the intervention requires it. Persistent version t fixes a scaffold snapshot and initial state distribution. Let X be the task, $M_{t,0}$ the initial memory, $\mathcal{T}_{t,<k}$ the recorded history, and $E_{t,k}$ the relevant tool/environment state. One admissible structural representation is $R^a_{t,k} = r_t(X, M^a_{t,k}, E^a_{t,k}, \mathcal{T}^a_{t,<k}, \varepsilon^r_k)$, $C^a_{t,k} = c_t(X, M^a_{t,k}, E^a_{t,k}, R^a_{t,k}, \mathcal{T}^a_{t,<k})$, $C^a_{t,k,j} = q_\eta(C^a_{t,k}, z^a_{t,k,1:j}, B^a_{t,k,j})$, $h^{\mathrm{nat},a}_{t,k,j} = f_{\theta,\eta,k,j}(C^a_{t,k,j}, \varepsilon^{\mathrm{num}}_{k,j})$, $h^{\mathrm{used},a}_{t,k,j} = T_a(h^{\mathrm{nat},a}_{t,k,j}, C^a_{t,k,j}, j)$, $O^a_{t,k} = o_{\theta,\eta,d}(C^a_{t,k}, h^{\mathrm{used},a}_{t,k,\cdot}, \varepsilon^o_k)$, $(M^a_{t,k+1}, E^a_{t,k+1}, \mathcal{T}^a_{t,<k+1}) = u_t(M^a_{t,k}, E^a_{t,k}, \mathcal{T}^a_{t,<k}, O^a_{t,k}, \varepsilon^u_k)$, and $Y^a_t = y_t(X, \mathcal{T}^a_{t,\le K}, E^a_{t,K}, \varepsilon^y)$. Where $C^a_{t,k}$ is the initial call input, $z^a_{t,k,1:j}$ the sampled prefix, and $B^a_{t,k,j}$ the relevant cache, bypass, and non-weight state. Step $j = 0$ is the last valid prefill position before the first answer token; later steps follow their arm's earlier perturbations and samples. The output equation denotes token-ordered execution, computing the current activation, applying the policy, sampling the next token, and updating the prefix/cache, not simultaneous dependence on future samples. In the paired solver-answer assay, arms share the complete state immediately before the first intervention, not every later decode state or downstream call input. η identifies the numerical/serving implementation and d the decoder; unmodified layers and bypass computation remain in the output map, so the monitored activation alone is not sufficient to reconstruct the call. State includes, or permits reconstruction of, positions, masks, and relevant cache/recurrent contents. Numerical noise also represents scheduling or batching effects rather than deterministic activations from fixed θ alone (Yuan et al., 2026). Within-episode writes and tool feedback enter later slots, while persistent writes change the next assessed snapshot. A bounded DAG is a time-unrolled model of this process, not a claim that real agents never have feedback. Inactive slots can carry a not-called symbol without manufacturing a counterpart for a removed activation. The E1/E2 scaffold is simpler, with no shared cross-episode memory, external tool effects, or judge feedback across the cached boundary.

A.3 Alignment Is Part of the Hypothesis

The comparison specifies a mapping between roles, events, hooks, and outcome definitions. Equal layer numbers, dimensions, or ordinal call positions do not justify it. The intended correspondence is the same solver answer stage and block-output residual hook under the same architecture. If a role or answer-stage contract changes, the affected assay must be re-specified or a justified mapping supplied; alternative plausible maps should be reported when they affect a conclusion. Causal abstraction provides relevant foundations while emphasizing that an unconstrained representation map need not support a meaningful explanation (Geiger et al., 2025; Méloux et al., 2025; Sutter et al., 2026). A parent input is *functionally irrelevant* to a target over an admissible domain only if varying that input while fixing other relevant arguments leaves the target mechanism unchanged throughout that domain. One unused memory in one trajectory or low average attention is not such a proof. A structural model is therefore an assumption-bearing object, not an automatically correct reading of a configuration diff.

A.4 Actual Activation Intervention

The position uses a state-dependent additive *intervention*, not replacement by a constant. For eligible answer-stage positions, $T_a(h, C, j) = h + a\, s\, \alpha \|h\|_2 v$, with $\|v\|_2 = 1$, $s \in \{-1, +1\}$, and $a \in \{0, 1\}$; at ineligible positions T_a is the identity. The natural activation remains a parent of the used activation, so deleting all incoming edges would describe another intervention. Under repeated interventions, the next natural activation is generated from the already modified history. For a logistic score $w^\top A(h - \mu) + b$, the raw-coordinate gradient is $A^\top w$. One candidate family uses this normalized nonzero gradient; the second uses the normalized raw fitting-set mean difference $\mu_{\text{correct}} - \mu_{\text{incorrect}}$ at the candidate layer. The frozen bundle identifies the selected family and layer, which need not coincide with the primary predictor's layer. Scaling, centering, and any linear projection must be accounted for; the intercept is not part of the direction. A zero vector leaves its candidate direction undefined and receives an instrument-failure status. The first construction is probe-derived steering; the second is mean-difference steering motivated, but not guaranteed effective, by truth-representation studies (Marks & Tegmark, 2023). Contrastive Activation Addition constructs directions from contrastive examples and is a distinct work (Rimsky et al., 2024).

> The *intervention* is defined by eligible answer-stage positions, direction, sign, residual-relative scale, and arm indicator. The natural activation remains part of the computation, so the intervention changes the used activation without replacing the underlying state-generating process.

A.5 Queries and Sampling

The predictive query concerns a joint distribution of an activation and specified label. AUROC, balanced accuracy, and Brier score are different functionals with different support requirements. A final-label query includes downstream labeling paths even when h is unchanged. A *local effect* concerns call $k_\star$; a multi-call policy does not make all call outputs one unspecified local endpoint. The assessed population comprises tasks, initial histories/memory snapshots, runtime settings, and within-episode random streams. A conditioning rule or eligible-event restriction belongs to the query and may depend on earlier interventions, requiring modeling in each arm. The minimal study fixes the task panel and one prospectively specified answer event, simplifying these issues without validating broader shared-state deployments.

B Supporting Formal Results

B.1 Conditional Sufficient Reuse

For a comparable agent call, fix the conditional output kernels $K_a(O \mid C)$, decoder, runtime, score, and intervention. With C the complete state immediately before the first local intervention, $\psi(c) = \mathbb{E}[m(O^1, X) \mid C = c] - \mathbb{E}[m(O^0, X) \mid C = c]$ and $\tau_{\text{local},t} = \int \psi(c) P_t(dc)$. Later token states include their arm's prefixes and earlier perturbations. The state law and conditional response are separate assumptions; a projected activation-shift statistic does not certify either for a changed execution.

Lemma 1 (Sufficient Invariance of a Query)**.** *Consider two correctly specified bounded acyclic models with aligned variables, the same measurement mappings, query functionals, and intervention*

policies. For each distribution entering the query, take the ancestral subgraph of its target and any conditioning or selection variables after inserting the actual intervention policy. If all structural functions in the union of these subgraphs and the joint law of their relevant exogenous inputs are identical between versions, the query distributions and value are identical. Conditional queries require nonzero conditioning probability.

Proof. Couple the relevant exogenous inputs using their common joint law and traverse each ancestral subgraph in topological order. Roots receive identical exogenous inputs and use identical functions, hence agree; inductively, each node with identical parents and structural function also agrees. The target and conditioning variables therefore have identical joint distributions. Apply the same readout, score, and functional. For an effect, apply the argument to both intervention arms and subtract their expected scores. For a conditional query, the joint distributions also give identical positive denominators. Thus the declared quantity is equal. □

This is an elementary application of structural invariance and direct-transfer principles (Pearl & Bareinboim, 2022), retained as support rather than headline novelty. Preserving separate exogenous marginals is insufficient if their joint coupling changes. A conservative change marker may cause the sufficient screen to fail even when the query is invariant. Finite-sample estimates may differ despite equal underlying quantities, and reused estimates retain their original uncertainty.

No Converse. Let $O_t^a = \beta(H + a\delta) + b_t$ with fixed β and activation law. Changing b_t changes the output mechanism and a target ancestor, but $\mathbb{E}[O_t^1] - \mathbb{E}[O_t^0] = \beta\delta$ for every offset. A known offset-only update therefore permits reuse of this contrast despite failing a full mechanism-invariance screen. A changed ancestor indicates potential relevance, not an inevitable metric change, a generic measure-zero guarantee of failure, or a proof that only fresh rollouts can supply evidence.

B.2 Replay Proposition and Estimator

Let $\mathcal{F}$ contain fitting/tuning information, optimization tasks and streams, search decisions, and the frozen instrument and selected judge J. The final task sample and upstream generation streams are independent of $\mathcal{F}$ under the specified sampling scheme. Conditional on $\mathcal{F}$, the selected configuration is fixed but need not be independent of its training data. Let P_0^a denote the old arm-a record law for that fixed configuration and target task population. With one upstream pair per final question, $\widehat{\tau}_J = \frac{1}{n}\sum_{i=1}^n \{g_J(V_i^1) - g_J(V_i^0)\}$. The primary estimator integrates only judge selection, not question or upstream-generation variation. If a different setting samples B downstream continuations, replace $g_J(V_i^a)$ by $B^{-1}\sum_b \ell(J(V_i^a, \xi_{ib}^a), X_i)$ under the correct conditional law. This is not the categorical protocol's default.

Proof of Proposition 4. (Condition on $\mathcal{F}$.) Sufficiency of the cut gives $g_J(v)$ as the expected new final score given a record. Upstream invariance identifies the new record law in each arm with P_0^a, so iterated expectation gives $\mathbb{E}[\ell(Y_J^a, X) \mid \mathcal{F}] = \int g_J(v) P_0^a(dv)$. For a fixed panel, expectation uses its task-weighted generation laws; random sampling of distinct questions also supports the declared sampling-population target. Selecting the record set using its outcomes would break the argument. □

The two arms need not have the same law; each must preserve its own old-to-new upstream law. Agent-induced critic and reviser changes remain in their own complete records. Downstream randomness follows the new program's law conditional on its input, including shared latent state. Task/role-keyed independent streams make this explicit in the case study; this independence is not assumed for arbitrary agents. More judge draws cannot replace more upstream histories. This is *executable-suffix re-estimation*, distinct from logged-event rejection sampling in contextual-bandit replay (Li et al., 2011), and not natural direct/indirect or path-specific identification (Avin et al., 2005).

One Analytic Replay and Rejection. For one complete record pair, set $(c_S, c_R, p_J)(V^0) = (0, 1, .25)$ and $(c_S, c_R, p_J)(V^1) = (1, 0, .60)$. $g_J(V^0) = .75$, $g_J(V^1) = .60$, and contribution $.60 - .75 = -.15$. The labels are external scoring inputs, not judge prompts. A manifest may retain the seed for a future sampled judge choice, but that draw is excluded from the statistical boundary conditioning and integrated by g_J. Replacing the on-arm critique with an off-arm critique breaks lineage; an upstream retry breaks the no-feedback boundary. Reject such records and regenerate

from the affected computation rather than retain the illustrated score as valid replay. Appendix E.9 specifies an independent arithmetic check.

B.3 Conditional Integration of the Final Choice

The statistical boundary record excludes the future judge uniform draw. A manifest may retain a seed tag for a sampled-transcript check, but that draw is integrated, not conditioned upon, when computing g_J. For fixed paired records $V = (V^0, V^1)$, the categorical policy samples a agent or reviser answer, and each arm's binary outcome B^a has conditional mean $g_J(V^a)$. Regardless of the coupling of the two final draws, $\mathbb{E}[B^1 - B^0 \mid V] = g_J(V^1) - g_J(V^0)$. For $D = B^1 - B^0$, total variance gives $\mathrm{Var}(D) = \mathrm{Var}(\mathbb{E}[D \mid V]) + \mathbb{E}[\mathrm{Var}(D \mid V)]$. The second term is nonnegative, so integrating final randomness cannot increase variance relative to sampling the same choices on the same records. This elementary conditional-expectation result is not a claim of new theory or a guaranteed large precision gain. Upstream variation, finite-cache uncertainty, and fitted-artifact uncertainty remain. Exact probability extraction, masked token policy, and candidate mapping must match the fresh execution.

B.4 Same-Record Candidate Opportunity

Write $g_J = c_R + p_J(c_S - c_R)$. For the same arm record, $g_{J_{\mathrm{new}}}(v) - g_{J_{\mathrm{old}}}(v) = (p_{\mathrm{new}}(v) - p_{\mathrm{old}}(v))(c_S(v) - c_R(v))$, which applied in both arms. Since $|c_S - c_R| = \mathbf{1}\{c_S \neq c_R\}$ and $|\Delta p^a| \leq 1$, $|\Delta\tau_{\mathrm{final}}| \leq \mathbb{E}[|\Delta p^1|\mathbf{1}\{c_S^1 \neq c_R^1\}] + \mathbb{E}[|\Delta p^0|\mathbf{1}\{c_S^0 \neq c_R^0\}] \leq d_1 + d_0$. The tighter right-hand side uses observed probability changes. These terms are full-population diagnostics, not effects conditioned on a post-treatment subset. No off-arm selection is transplanted into an on-arm critic/reviser history. Agent-copy yields the local correctness effect by definition; reviser-copy can retain agent influence. Neither guarantees that learned judges dissociate.

B.5 Sensitivity to Imperfect Replay

Suppose a coupling of fresh and replay outcomes in arm a agrees outside an event of probability at most ϵ_a, with score in an interval of length R. Then $|\mu_a^{\mathrm{replay}} - \mu_a^{\mathrm{fresh}}| \leq R\epsilon_a$ and $|\tau^{\mathrm{replay}} - \tau^{\mathrm{fresh}}| \leq R(\epsilon_0 + \epsilon_1)$. The score difference is zero outside the exceptional event and at most R within it, so each arm difference is bounded by $R\epsilon_a$ and the ATE bound follows by the triangle inequality. For correctness in $[0, 1]$, justified $\epsilon_a = .05$ permits a worst-case ATE discrepancy of .10, not an estimated five-point bias. This hypothetical ten-point ceiling and E2b's five-point equivalence region concern different objects; a loose sensitivity bound cannot establish equivalence but does not contradict the target. Passing E2a at a numerical tolerance or observing a fraction of flagged cache records does not establish a population coupling-failure probability. Conditioning on accepted records can change the population; missing arms are not simply dropped. Known dependency faults trigger rejection/fallback. A sensitivity statement quantifies consequences only when its assumptions are justified, not permission to disregard errors. For a distributional version, use $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. Let record laws be P_a, Q_a, downstream means be $g_a, \widetilde{g}_a$, and both scores lie in the same length-R interval. Adding and subtracting $\int \widetilde{g}_a \, dP_a$ gives $\left|\int g_a \, dP_a - \int \widetilde{g}_a \, dQ_a\right| \leq \sup_v |g_a(v) - \widetilde{g}_a(v)| + R\,\mathrm{TV}(P_a, Q_a)$. Sum the two arm bounds for an ATE. Neither empirical CKA nor an arbitrary projected shift score supplies the required full-record TV or kernel bound.

B.5.1 Sharp Uncertainty for Partial Selector Records

Proposition 3 fixes the selected configuration and two recorded-state laws conditional on pre-evaluation selection information. Let W^a share a measurable space with law P_a. Candidate correctness $c_S(w), c_R(w)$ is a valid known function of this record, including its evaluator-only sidecar, and is not supplied to the judge. The true state may include omitted U^a. The target mean $\pi_a(w) = \mathbb{P}(\text{select solver} \mid W^a = w, a)$ averages over its correct conditional law, while a default state may produce $\widetilde{p}_a(w)$. The rectangular class permits any measurable selection probability within each stated range, with no further cross-arm restriction. Sharpness is relative to that class and available information; a neural judge family, other observed outcomes, or shared latent-state constraints may restrict it. Candidate contents, labels, and target recorded-state laws remain fixed. A corrupted label, arm splice, or changed candidate-generation law is outside this selector-only uncertainty model.

Proof of Proposition 3. Conditional expected correctness is $c_R(w)+\pi_a(w)\Delta c(w)$. A linear function on $[\underline{p}_a(w), \overline{p}_a(w)]$ attains its minimum and maximum at the endpoints selected in s_a^-, s_a^+. Integrating gives each arm's attainable mean interval. To minimize the ATE, choose the lower on-arm mean and upper off-arm mean, reversing these choices to maximize it. These measurable choices are simultaneously admissible in the rectangular class. Convex mixtures of the extremal probability functions attain every intermediate value, where its width is the sum of the two arm widths. Non-negativity of the integrands proves point identification if and only if that width is zero, equivalently if interval width vanishes almost surely wherever candidates disagree in either arm. □

This is a selector-specific partial-identification calculation using established linear-extremum reasoning (Manski, 2003), not a new general identification theory. Unlike the known-change bound in Appendix B.4, it concerns uncertainty about the *correct target* probability. Sampling uncertainty in estimated P_a or interval endpoints must be assessed separately from the identified set.

Proof of the Shared-Selector Refinement. Under §3's common-conditional-mean assumption, combine the arm integrals to get $\tau = C_R + \int \pi r \, d\lambda$. The maximum chooses $\pi = \overline{p}$ where $r > 0$ and $\pi = \underline{p}$ where $r < 0$, with arbitrary admissible values where $r = 0$; the minimum reverses these choices. Their difference is $\int (\overline{p} - \underline{p})|r| \, d\lambda$, and convex mixtures attain every intermediate value. More restrictive shared constraints require optimizing their actual joint class; rectangular extrema need not be attainable. Identical code alone does not imply conditional-mean equality because $U \mid W, a$ may differ across arms.

Proof and Scope of Corollary 3. For every output t, $|t - \tau_-| + |t - \tau_+| \geq \tau_+ - \tau_-$. The complete observable laws in the two endpoint worlds coincide, including all allowed query results and the estimator's independent auxiliary randomization, so the two output laws of T coincide. Integrating the pointwise inequality under that common law and taking the larger expected error at any sample size preserving indistinguishability. If the interval is known exactly, its midpoint attains worst-case identification error equal to half its width. Endpoint estimation adds uncertainty; neither statement guarantees confidence-interval coverage. Endpoint attainability from Proposition 3 does not by itself establish the corollary's observation-law premise; extra queries can distinguish the completions. The proof is the classical two-point argument (Yu, 1997). A boundary-only query that defaults the missing bit to zero is also identical in both worlds. This is not an ordinary future independent draw with a known law. Observing the actual bit or executing a correct fresh runner distinguishes the worlds and is outside the restricted observation set used for the lower bound.

Sources and Limits of Selector Envelopes. Let omitted state U lie almost surely in a separately justified nonempty set $\mathcal{U}_a(w)$, with $b_J(w, u) \in [0, 1]$ its correct conditional agent-selection probability. Any justified measurable lower and upper bounds on $b_J(w, u)$ for all $u \in \mathcal{U}_a(w)$ also bound $\pi_a(w) = \mathbb{E}[b_J(w, U) \mid W^a = w, a]$. Where measurable, the infimum and supremum over $\mathcal{U}_a(w)$ supply such bounds; a known target conditional distribution instead permits integration. Without extra information, use $[0, 1]$. Default-state queries, heuristic scores, finite sample extrema, and E2a tolerances are not population envelopes. Estimated bands require separate uncertainty accounting. The rectangular envelope class may enlarge a fixed program's feasible completions, giving an outer bound rather than sharpness for that smaller family. None of this instructs E1/E2 to estimate envelopes. A sufficient condition for the shared-selector refinement is a common $b_J(w, u)$ and common $U \mid W = w$ law on overlapping support, which gives the same conditional mean; identical hidden states are unnecessary. Different conditional state laws can also give the same mean, so this condition is not necessary. Empirical non-rejection of equality would not establish it universally. The fixed-score affine calculation also applies to specified bounded c_S, c_R, with their score gap replacing binary disagreement. It does not model unknown true labels or measurement error, and does not change E1/E2's binary external grader or sampled labels.

From Probability Uncertainty to a Bounded Claim. If separately justified bounds give $|\pi_a - \widetilde{p}_a| \leq \epsilon_a(w)$, clip $\widetilde{p}_a \pm \epsilon_a$ to $[0, 1]$ for the exact rectangular interval. With $\widetilde{\tau}$ the surrogate replay effect, $|\tau - \widetilde{\tau}| \leq \sum_{a=0}^{1} \mathbb{E}_{P_a}[\epsilon_a(W^a)|\Delta c(W^a)|]$. This follows by subtracting conditional scores and applying the triangle inequality, and can improve the generic range bound when uncertain selections fall on equal-correctness candidates. The envelope requires the independent justification above. A valid confidence interval for the surrogate can be enlarged by a separately justified uniform bias budget, with statistical and sensitivity components reported separately. Estimated-envelope coverage requires separate accounting. None of this authorizes known invalid records, filters disagreement cases out of the target population, or changes E2b's uncorrected comparison.

C Local Assays, Replay Scope, and Coupling

C.1 What Local Replay Can Save

Fixing the local model/runtime, decoder, scoring functional, and intervention gives the same conditional local response $\psi(c)$, so new representative states can assess a population-local effect without executing later roles. This is conditional expectation, not a cheapest-test theorem. Acquiring these states can require long upstream trajectories, and local sampled-text correctness requires answer generation, not a single forward pass. Restoring the prefix requires relevant KV/recurrent, positional, mask, and bypass state. If prior steering changes prefixes or event occurrence, the original unsteered prefix law does not supply the total policy effect. Eligible-call analyses must name their conditioning event and report coverage; skipping failed/absent events changes the population. E1 avoids this ambiguity with a fixed two-stage solver. The reasoning call is unsteered, the answer call begins from its fixed-format continuation, and steering applies only in that call. The local label is correctness of its generated numeric answer, not model-reported confidence or a post-hoc property selected because it moved strongly.

C.2 Trajectory Records versus Neural Caches

The replay cut stores task and upstream messages from a complete intervention arm; it does not permit reusing an unsteered prefix KV cache when a new intervention must modify that prefix. The last prefill input representation produces the first answer-token logits, so a decode-only hook can miss the intended intervention. Subsequent decode positions require their own declared handling. Each episode records the actual hooked block, positions, residual norm, applied norm, and number of applications. Ordinary within-call autoregressive caching may be used with verified hook semantics, while cross-request prefix reuse that skips a required hook is disabled in the minimal study. The zero-strength path checks that instrumentation reproduces the off computation within the declared numerical contract. Exact hook correspondence and event timing are requirements distinct from whether a final answer matches.

C.3 What Pairing Establishes

For binary final outcomes $B^a = \ell(Y^a, X)$ under a declared experimental coupling, $\mathbb{E}[B^1 - B^0] = \mathbb{P}(B^0 = 0, B^1 = 1) - \mathbb{P}(B^0 = 1, B^1 = 0)$. The first term is a rescue and the second a harm under that pairing. Their difference is the marginal ATE, while each joint probability depends on the coupling. Neither is identified from separate arm marginals alone (Tian & Pearl, 2000). Task/role/stage-keyed random streams define a reproducible experimental coupling, not a uniquely true cross-world correspondence. A single globally consumed RNG stream can shift random draws for later roles when earlier text lengths change; stage-keyed streams avoid this bookkeeping problem while preserving each arm's intended marginal sampler. Diverging trajectories do not guarantee that pairing loses all variance reduction, since shared task difficulty can still induce covariance. For independent-fresh assessment, the same tasks remain paired while upstream generation streams differ from the cache. Exact transcript identity is not the statistical target.

C.4 Headroom and Unchanged Scoring

Higher unsteered accuracy can reduce failures available to rescue. Reporting both baseline accuracy and exposure helps interpret a smaller ATE. Conditional rescue rates have changing denominators and do not replace the task-population ATE. A useful negative signed intervention effect can still test transfer, but is not a correctness-improvement mechanism. A deliberately changed scoring rule defines a new endpoint. Sufficient stored answers can sometimes be rescored without new generation when scoring is post-hoc and has no agent feedback. An implementation bug instead threatens integrity. The same distinction applies to local logit effects versus sampled-text effects; a decoder-only change may preserve fixed-prefix pre-sampling logits while changing generated answers and correctness labels.

C.5 Categorical Expectations and Coupled Sampled Diagnostics

The primary final effect uses $g_J(V^1) - g_J(V^0)$, not a sampled binary difference, so these fractional contributions do not define binary rescue/harm counts. Keyed categorical draws separately produce binary outcomes. Local answer correctness remains binary. A sampled final-label AUROC uses those binary draws and a frozen readout; expected correctness is not inserted as a fractional class label. Sharing one reasoning draw across answer arms preserves their intended marginal laws because reasoning precedes the intervention; critic/reviser draws and every consumed message remain arm-specific. E2b generates an independent reasoning draw for each new pair and does not reuse the original pair's prefix. Production scheduling is fixed independently of the judge version, since changing batch neighbors can otherwise change upstream numerical execution while appearing to be a downstream-only edit.

The local assay fixes the intervention and complete pre-intervention state, allowing *local intervention effects* to be evaluated conditionally. The *final-answer effect* additionally depends on the downstream selector and its *affected continuation*. Thus, valid local replay need not identify the total policy effect when the intervention changes upstream prefixes or event eligibility.

For *coupled records*, rescue and harm probabilities depend on the declared pairing, while separate arm marginals do not determine their joint decomposition. *Complete trajectory records* preserve the execution state needed for faithful replay, while cache reuse is valid only when declared hook semantics and an unchanged prefix justify it. Keyed streams provide reproducible coupling without requiring exact transcript identity.

D Records and Conditional Evidence Decisions

D.1 Claim, Update, and Assessment Records

A claim has seven fields. *Claim* identifies the proposition, one of the three targets, endpoint, and unit. *Configuration* binds model/tokenizer/runtime and scaffold snapshot. *Population* names task and initial-history/state support. *Instrument* binds artifact, preprocessing, calibration, role, block, prospective event, and token anchor. *Intervention* identifies operation, magnitude, timing, mask, and local versus free-running policy, or is inapplicable for a prediction-only claim. *Evidence* records estimate/uncertainty or analytic argument, controls, acceptance criteria, and covered updates. *Dependencies* summarizes the six checks. An update record contains version identities, changed components, impact assessments, and evidence, plus who proposed and selected the update, objective, accessed data, candidate family, budget, and whether any monitor or correlated proxy entered selection. Selection provenance is cross-cutting, not a seventh neural-dependency category. A history/archive ID is needed when multiple variants or shared-memory states coexist. The assessment record separates *evidence action*, claim status, and comparability. Actions are Reuse, Re-test, Re-specify, and Withhold; statuses are Supported, Contradicted, and Unresolved, for a named proposition and version (see Figure 3). Comparability is comparable, not comparable, or unknown. An incomparable new assay does not contradict the old-version finding. A narrowed proposition or refitted instrument creates a new record. Withhold prevents an unsupported assertion; it is not a measurement of neural failure.

The record separates *measurement evidence* from the proposition assessed. A reuse decision requires intervention, instrument, population, and dependency evidence to remain valid; a new or incomparable assay requires *retesting or respecification*, not treating missing evidence as a neural failure.

D.2 Artifact Identity and Trusted Provenance

The prior instrument record binds fitting-data revision and membership, construction code/configuration, preprocessing, vector family, layer, sign, scale, and frozen bundle digest. Keep its authoritative freeze record outside the mutable cache being checked. Comparing bytes to that independent reference tests identity; authenticating a publisher tests origin under its key/trust assumptions. Neither establishes benign source material or behavior. Aidakhmetov et al. (2026) demonstrates a malicious steering bundle whose vector can be regenerated from distributed poisoned examples, so regeneration can reproduce the poison while a benign-task check can leave other behavior unexamined. E1/E2 constructs instruments from its controlled fitting pool and assumes a trusted execution environment. It does not certify arbitrary imported vectors, signatures, or orthogonalization as

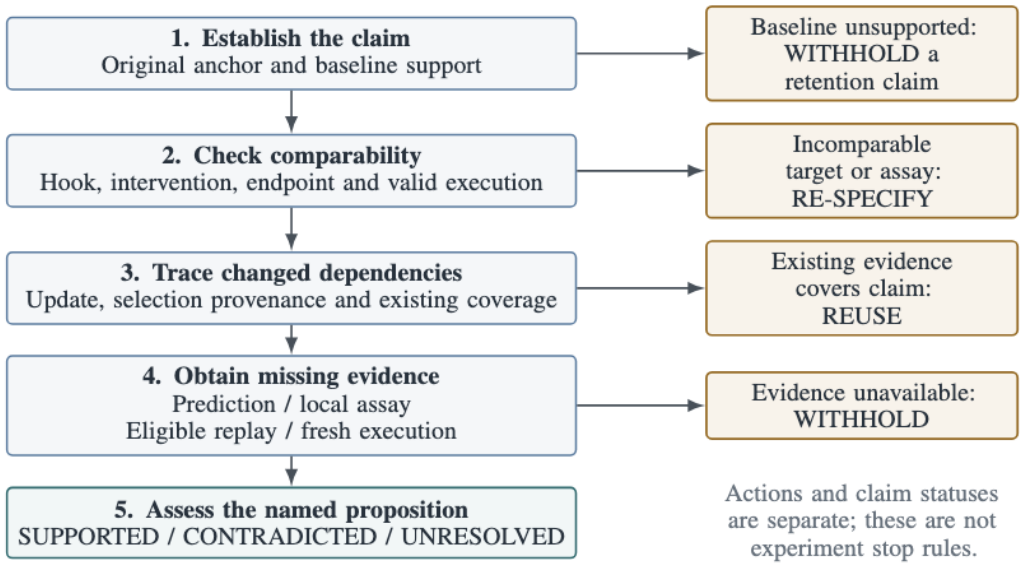

Figure 3: Design of the evidence procedure. Actions, claim status, and comparability are distinct. Unsupported baseline evidence withholds retention; a defined frozen effect can still receive independent assessment. Replay is re-testing, not a guarantee of unchanged effect, and does not prove semantic sufficiency.

poisoning defenses. Compromise of both the authoritative freeze record and evaluator is outside this threat model. The coefficient-substitution fixture in Appendix E.9 checks alteration relative to a retained reference, not semantic safety of that reference.

D.3 The Six Dependency Checks

Input Distribution. Record tasks and complete prefixes reaching the monitored call, including retrieval, truncation, previous tool outputs, and feedback. A fixed memory-token budget can prevent database growth from lengthening context but may still change retrieved contents.

Outcome Relationship. Record label meaning, the representation-label relationship, and baseline headroom. Even identical agent activations can have a changed relationship to final correctness after a judge update.

Downstream Use. Trace which later roles and tools consume the monitored output and determine the endpoint. A judge-memory update can alter this without editing the visible graph, and feedback can affect later monitored inputs.

Measurement Support. Verify the exact role, hook, prospective eligibility, and token anchor. A changed prompt length is not handled by blindly selecting the same absolute token index; a missing hook has no automatic alternative.

Intervention Exposure. Record the fixed operation, realized magnitude, timing, token mask, and call count. Differences in realized exposure under the same policy are not automatically incompatible, but may explain changed effects. This category is inapplicable to prediction-only claims.

Measurement Integrity. Verify configuration restoration, numerical contract, capture, intervention, and scoring. A failed restore or silent layer fallback invalidates that measurement; no such failure is asserted as observed in this unrun study.

D.4 Conditional Update Cases

Table 4 assumes valid baseline evidence; its re-test entries also assume no sufficient prior coverage or valid invariance argument. A component name alone does not determine the decision. Table 5 summarizes judge-only and upstream changes. The same neural predictor against final correctness still needs its new label relationship assessed. For an upstream change, a valid separate transport argument can sometimes avoid full regeneration; retaining the same context length does not establish an unchanged state law.

Table 4: Dependency checks and evidence in the replay manifest. A matching hash checks stored bytes; eligibility also requires the stated dependency argument. No implemented check or measured coverage is claimed yet.

Dependency Check	Manifest Evidence	Limitation
Input Distribution	Task/pool IDs, full rendered prefixes, upstream prompt and code IDs, and per-request random streams.	Judge IDs must not seed upstream generation.
Outcome Relationship	Numeric scorer, candidate labels, and answer-normalization version.	Gold labels remain evaluator-only.
Downstream Use	Complete same-arm critique/revision, candidate identities/order, and allowed judge field.	The fixed DAG has no judge callback.
Measurement Support	Block output, semantic answer event, last valid input position, and active-token mask.	Matching an array shape is insufficient.
Intervention Exposure	Family/direction hash, fixed sign/scale, relative-norm rule, applied positions, and realized norms.	One record is required per actual arm.
Measurement Integrity	Model/tokenizer/software IDs, precision, batch companions, padding, cache policy, restore checks, numerical tolerance, and typed failures.	Integrity depends on successful restoration and the declared numerical contract.

Table 5: Conditional evidence decisions. Rows assume supported baselines and correct execution. WITHHOLD applies when sufficient valid evidence is unavailable, not as a component-update class. Appendix D.4 gives full conditions and memory/scoring examples.

Update / assumption	Prediction	Local effect	Final-agent effect / replay
Judge only; no feedback	Local label: reuse. Final label: assess.	Reuse if local contract holds.	Re-estimate; replay if boundary is sufficient.
Upstream prompt changes	Re-test affected label.	Re-test new states.	Fresh from earliest affected boundary.
Feedback crosses cut	Assess changed population.	Re-test eligible calls.	Old cut is ineligible for replay.
Required hook absent	Re-specify affected probe.	Re-specify assay.	Re-specify affected intervention.

Routing with Feedback. A new judge callback invalidates scoring only the judge against a frozen old continuation. In a time-unrolled execution it cannot change the saved prefix's past; a checkpoint before the updated call remains reusable if it restores all relevant state and conditional randomness and executes the entire changed future, including retries, new solver calls, tool outputs, and writes. Stale outcomes of actions the new policy no longer takes are not valid inputs. If the update also changes prefix construction, resource allocation, or unlogged shared state, move the checkpoint earlier; without a recoverable sufficient checkpoint, run fresh. Define whether the intervention acts only on the original answer event or also on retries, since these are different policies. This is *conditional forward execution* under Appendix A.1, not a new feedback theorem or added E1/E2 condition.

Inaccessible-Memory Control. A note in an inaccessible namespace can leave all relevant computations unchanged if it affects neither indexing/order, resource limits, context construction, nor feedback throughout the assessed scope. This analytic negative control remains useful, but is not the paper's empirical demonstration. One unused trace cannot establish the claim for all tasks.

Removed Role or Hook. Re-specify affected instruments rather than substitute another activation or assign zero. Behavioral task outcomes can still be evaluated independently. A new correspondence requires an semantic argument, not matching shapes.

Scoring-Only Change. A new post-hoc endpoint needs a new claim; adequate raw on/off outputs may permit rescoring. A local effect whose score did not change can retain its evidence. If the evaluator feeds back into the agent, a post-hoc-only shortcut is invalid.

Decoding-Only Change. Unchanged complete prefixes and neural operations can preserve pre-sampling logit effects, while sampled local answers, final answers, and labels may change. Later-token logits can change when earlier samples change. Only claims about truly unchanged pre-sampling quantities justify simple reuse.

Backbone, Precision, or Shared State. A new backbone/precision requires a separately specified model-transfer assay. Concurrent memory state requires an identified isolated snapshot or defined history distribution. The minimal experiment does not establish such transfer; these cases remain outside its empirical scope.

Table 6: Call contract. Caps describe output capacity, not observed lengths. The categorical judge uses one decision-position forward computation, not an additional rationale.

Call	Inputs and Required Output	Output Cap
Solver Reasoning	Question and fixed instructions; bounded scratch reasoning.	384
Solver Answer	Question, reasoning, fixed `Final answer:` prompt; numeric answer.	32
Critic	Question and this arm's solver answer/reasoning; critique.	192
Reviser	Question, this arm's solver output and critique; revised numeric candidate.	384
Categorical Judge	All consumed candidate records and critique; masked A/B choice.	1

Conditional update decisions depend on whether the available record preserves relevant *identification conditions*. When a changed downstream component can be executed from a justified checkpoint, *conditional forward execution* can reassess the affected continuation. When required state, label relationship, or execution contract is not preserved, reuse is not established and the claim remains *unresolved* until an appropriate assay is performed.

E Complete E1/E2 Protocol and Staged Locks

E.1 Populations and Isolation

The official GSM8K `main` configuration has 7,473 training and 1,319 test questions (OpenAI, 2021; Cobbe et al., 2021). Dataset revision, deduplication rules, and exact IDs are recorded. No sample is selected for model difficulty, intervention response, candidate disagreement, or desired effect sign. The six primary training pools total 1,728 questions. If either correctness class has fewer than 64 examples among the initial 768 fitting outputs, the entire prelisted 256-question reserve is added before L2, giving at most 1,984 training questions; only class count triggers this extension, and the reserve is not selectively mined for errors or stronger effect. An inadequate minority count after extension remains a limitation. The unused 295 test questions are not an outcome-driven backfill pool. Fitting transforms are learned within each training fold, and calls or alternative outputs from the same question stay in the same fold. The final panel is sealed from fitting, tuning, optimizer proposals/selection, and all L3 decisions. The confirmation pool is never reused to select another intervention after L2. Its separately declared two-draw precision subset can inform sample allocation after the judge is frozen without feeding measurements back to the optimizer. One upstream draw per final question replaces two draws on 512 questions, preserving the 1,024 task-stream budget while improving distinct-question coverage. Under $\mathrm{Var}(\bar{Y}) = \sigma^2_{\text{between}}/N + \sigma^2_{\text{within}}/(NR)$, more distinct questions help whenever between-question variance is nonzero.

E.2 Five-Call Interface, Parsing, and Context

Reasoning may increase once to 512 tokens before fitting if development cap hits exceed 5% (see Table 6); record the decision and revised cost inventory at L1. This addresses truncation, not accuracy. Other caps stay unless a documented format/truncation defect requires an L1 amendment. Full rendered context is at most 8,192 tokens, with deterministic overflow policy fixed on development data; silent omission of consumed messages is prohibited. Numeric extraction uses a locked terminal-answer convention, sign/decimal normalization, and a rule for ambiguous or malformed outputs. The host scores predicted values, not the judge's opinion. Invalid outputs and budget exhaustion are task failures. Gold answers may supervise designated fitting/optimization objectives but never appear in final-evaluation inputs. A blinded parser spot check reports mistakes; scorer bugs require versioned rescoring of all affected observations, not selective edits.

Judge A/B identifiers must each be one token under the frozen tokenizer and rendered decision prefix; IDs are TO FREEZE until L1 verification. All other tokens are masked at that position. A temperature-one softmax over the two allowed logits defines the stochastic policy, without top-p/k filtering. After undoing candidate ordering, the solver probability is $p_J(v)$. The host computes expected correctness from external binary candidate labels. A keyed uniform draw supplies the sampled selected answer and binary final label without an extra model call. Expected scores never enter binary-label AUROC. Candidate order uses a task-keyed randomized mapping fixed across arms, judges, replay, and corresponding fresh draws. Store canonical solver/reviser identity separately from displayed A/B position. The judge has no new reasoning; it is a restricted selector over supplied rationales. No judge callback, retry, or persistent memory write alters the upstream record.

E.3 A100 Execution and Randomness

One model worker batches independent requests in role-specific waves. A100 40/80GB variant, MIG status, host memory, driver, kernels, and software revisions are TO FREEZE at L1; microbatches 1/4/8/16 are development candidates, while 32 requires measured 80GB headroom, and larger batches are not justified by parameter size alone. The 14B reflector loads separately from the 7B agent. Independent role prefills batch when possible; instrumented answer calls batch only after qualifying per-row masks, padding, positions, arm flags, and residual norms. Capture selected anchors to CPU; CPU parsing, compression, fitting, and hashing may overlap independent GPU work, while dependent roles remain ordered. Generation uses frozen task/role/stage/request-keyed streams, not a global RNG; judge checkpoint identity cannot affect prompts, RNG keys, or scheduling. Arm streams preserve marginal laws, coupling is explicit, independent-fresh records use new upstream streams, and the categorical draw uses its declared stream. Ordinary-generation temperature/filtering are TO FREEZE at L1 and are not silently changed to greedy. Record production batch members/order, bucketing, padding, engine, precision, and kernels; matched checks reconstruct production batches or use a qualified batch-invariant path. A singleton diagnostic is a different numerical regime and is not automatically pooled (Yuan et al., 2026; vLLM Contributors, 2026); decoder-only generation uses left padding (Hugging Face, 2026). The semantic hook indexes each row's last valid input position using masks/token IDs. Within-call KV caching is allowed after timing verification; unverified cross-request prefix reuse is disabled because it can bypass the perturbed prefill position. Replay stores complete arm trajectories, not repurposed unsteered KV caches, and cost comparisons use the same qualified implementation and batch policy.

E.4 Fitting, Bounded Search, and Independent Confirmation

Fit an L2-logistic correctness probe from unsteered pre-answer activations; task-grouped cross-validation screens blocks 7, 14, 21, and 27, retaining two by the prespecified criterion. Fix the regularization grid/tie order at L1 and freeze preprocessing, threshold, coefficients, and probe layer at L2; mean-difference/probe-normal vectors use fitting activations only, with the latter's affine-coordinate correction in Appendix A.1. Search two retained layers $\times$ two direction families $\times$ two signs $\times$ four scales $\{0.01, 0.03, 0.05, 0.10\}$, at most 32 settings, using paired answer sampling on 256 tuning questions; select the largest absolute local correctness effect with invalid-answer frequency increase ≤ 0.01 versus off, breaking ties by lower magnitude then fixed ordering. Record expected effect sign separately from vector sign; freeze the predictive instrument and intervention separately, with confirmation testing the frozen *intervention measurement*, not a post-hoc setting. Apply the selected perturbation at block-output residual, last valid prefill position, and active answer-decode positions; store anchors, norms, positions, and counts. No intervention occurs in reasoning, critic, reviser, judge, or reflection calls. A negative correctness effect is harmful, not improved reasoning. Confirmation uses 256 untouched training questions without retuning and precedes final measurement; weak confirmation is disclosed, while material protocol revision requires new configuration and confirmation data. Off/zero-strength sham tests path fidelity; 128 prelisted questions use two seeded random directions and the opposite sign for specificity/sensitivity checks. Diagnostics include AUROC, BA, Brier, class counts, fitting-prior baseline, lightweight question/available-reasoning classifier, length/log-likelihood features, and grouped label-permutation control. Calibration, discrimination, and incremental information are distinct claims; none proves a correctness circuit.

> The intervention record fixes the *probe, layer, direction, sign, scale, token mask, and answer-stage event* before final assessment. Local correctness effect is a measurement object; probe discrimination and calibration remain distinct diagnostics. Confirmation tests the frozen *intervention measurement*, not a post-hoc choice of a more favorable setting.

E.5 Judge Optimization and Checkpoints

The pinned GEPA integration may edit only the judge instruction, capped at 512 tokens; the external 14B reflector is motivated by capable-reflection recommendations (DSPy Contributors, 2026; Agrawal et al., 2026), not guaranteed superiority. Each of two seeds runs three proposal rounds with two children/round and an always-eligible incumbent. Reflection uses the adaptation pool; selection uses mean unsteered expected exact-match correctness on the 256-question selection pool. Steered records, confirmation effects, probe scores, and final questions never enter optimization. Generate adaptation/selection upstream records once and reuse them; cache incumbent/duplicate scores on the

identical population, retaining the incumbent on ties. Retain formatting failures, constraint violations, duplicates, rejected candidates, no-ops, and consumed work. The bounded searches allow 24,576 proposed output/repair tokens, with API/call and repair accounting frozen at L1; unused allowance adds no rounds. The plan does not assert support from a particular API; record actual engine/commit, adapter, selection/acceptance settings, merges/skips, and consumed slots. Standalone GEPA or its DSPy wrapper may supply a study-specific bounded configuration; otherwise report custom reflection as GEPA-inspired judge-prompt search. Integration remains TO FREEZE; save J_0, all intermediate incumbents, and $J_\star^{(0)}, J_\star^{(1)}$. Seed zero is primary, seed one a narrow replication. Unchanged prompts are prompt no-ops; similar-output prompts or small effect contrasts are separately characterized. Judge-only compliance is not the natural replay-eligibility frequency in unrestricted evolution. Hand-coded solver-copy/reviser-copy outputs are analysis diagnostics, never optimizer-produced updates.

The judge update defines the downstream policy whose effect is reassessed. *Selection provenance and bounded search* determine which updated policy is identified, while *complete upstream records and executable checkpoints* determine whether its affected continuation can be reconstructed. Optimization remains separated from final intervention measurement and replay validation.

E.6 E1: Main Arm Records and Outcomes

For 1,024 final questions, one unsteered reasoning call precedes paired off/on answer generation. Critic and reviser continuations are separately regenerated in each arm, retaining all dependence on that arm's solver, producing 2,048 full boundary records. Each is scored by J_0 and the two final categorical judges. Store full rendered messages, canonical candidate identities, probabilities, sampled choice, parsing/scoring results, hook/exposure data, versions, and execution provenance. Report shared unsteered solver probe metrics once, local binary correctness ATE, each judge's expected final correctness/ATE, and the same probe's distinct sampled-final-label diagnostic. Candidate-opportunity are computed on the full population. Selection on post-intervention disagreement is descriptive and cannot replace the population contrast. Local invariance is conditional reasoning about the design; repeated use of one local cache does not supply independent empirical confirmation. Fresh checks establish whether implementation respects that design.

E.7 E2a: Computational Reconstruction

Development includes valid reconstruction and negative eligibility cases before large-scale measurement. The fixed 128-question final subset regenerates upstream computations from clean state without stored model outputs, in the recorded production regime. Compare discrete fields, token IDs, semantic hook events, consumed judge inputs, candidate mapping, probabilities, and sampled answers. Stable lineage identifiers may differ from wall-clock timestamps; equality concerns the computational record, not every bookkeeping byte. Discrete equality is separate from numerical tolerance. The planning probability-error target is maximum absolute error 10^{-6}; activations/logits have separate L1 tolerances. If the engine cannot meet this target, freeze a justified numerical contract using development data before final execution or withhold the exact-reconstruction claim. Tolerances are not widened after final mismatches. With unchanged binary candidate labels, selection-probability discrepancy δ changes expected correctness by at most $|\delta|$; matching accuracy alone does not establish boundary integrity. Appendix F.1 gives the finite-record implication without a population error guarantee. A keyed draw can differ when a uniform falls between numerically close probabilities; log this rather than calling it a semantic-input match.

E.8 E2b: Independent-Fresh Agreement

On 512 prespecified final questions by default, produce an independent full off/on upstream pair with new streams and the same policies; expand to 1,024 only by the training-side L3 rule in Appendix F.4. Apply primary $J_\star^{(0)}$ and the same categorical scoring to replay and fresh records. Compare on identical questions the expected ATE difference and both arm-mean differences; text/choice identity is not expected. Matched reconstruction cannot be relabeled as an independent-fresh equivalence test. Only a precision-qualified L3 mode permits joint equivalence, requiring the ATE and both arm-mean discrepancies to satisfy their ± 0.05 regions under prespecified tests. Otherwise E2b is estimation-only, even if a later interval looks favorable. Equal bias in both arms can cancel in ATE, so all three estimates remain visible. More judge draws on one cache cannot remove upstream sampling uncertainty; categorical scoring already integrates this final random choice.

Table 7: L0-L3 approval and freezing stages. Design values are specified; execution completion dates are TO FREEZE. No row attests to registration or execution.

Stage	Timing	Contents Frozen
L0: Charter	Before development	Claims, model family, pool allocation, candidate layers/directions, outcomes, resource ceiling/fallback rule, sealed final data.
L1: Execution/Search	After development, before fitting/search	Immutable model/tokenizer/runtime, hardware and batches, role rendering, token IDs, grammar, caps/scorer, RNG, hooks, search/tie rules and proposer budget.
L2: Instruments	After fitting/tuning, before confirmation	Probe transforms/weights/threshold/layer; selected vector family/layer/sign/scale/masks; fixed controls; all attempted settings.
L3: Final Evaluation	After confirmation and judge selection	Judge hashes, final task/batch manifests and tier, fresh N, precision mode, family/tests/margins, numerical criteria, review/timing and resource-stop policy.

E2 separates *reconstruction* from *independent-fresh agreement*. Reconstruction checks whether the declared computation is reproducible from a complete record; fresh assessment changes upstream draws and tests agreement on the same tasks. Neither exact transcript identity nor matching final accuracy alone establishes the corresponding claim.

E.9 Eligibility Challenges, Assessment, and Costs

The locked suite includes valid judge-only/no-op changes and ineligible cases, including upstream solver-prompt edit, arm-spliced critique, omitted consumed field, stale candidate mapping, wrong arm tag, changed hook/parser, and judge feedback request (see Table 7). These primarily test records/manifests on CPU, with bounded neural reruns for reconstruction. Rejection identifies the earliest invalid boundary. A deliberately changed scoring target may permit rescoring sufficient raw answers as a new claim; an unnoticed parser mismatch is an integrity fault.

Two Independent CPU-Side Contract Checks. Before final locks, qualify two small fixtures within prior development allocation. First, copy a frozen steering vector and alter one representable coefficient while preserving shape/metadata; its separately retained frozen digest must reject the copy before model execution, while the unaltered case passes, with the reason logged in the prior integrity/fallback ledger. This detects substitution against a trusted reference, not poisoned source examples or an attacker controlling that reference. Second, check g_J and the same-record judge-update contrast against a short independent reference calculation. Fixed analytic probabilities and external binary labels cover both-equal/opposite candidate correctness, both arms, and A/B-to-solver conversion; relabeling a recorded distribution with candidate identities must preserve its score, a serialization/mapping invariant rather than a prediction about swapped-order model probabilities. Appendix B.2 gives one hand-checkable pair. They add no task pool, neural treatment, fitted model, or confirmatory hypothesis. Two humans assess distinct update records blind to effect summaries and report agreement counts, disagreements, reasons, and time; these cases cannot estimate a population-wide missed-failure rate. Old-value reuse, upstream-only shift checks, final-outcome monitoring, and fresh evaluation are comparators with different targets; no claim that prior monitoring is upstream-only is made. The fresh route uses the same batching, pre-intervention sharing, integrated judge, and duplicate-score reuse. Record tokens, calls, GPU/wall time, peak memory, cache I/O/storage, loading, review, failures, and fallback work under fixed warm/cold boundaries; report construction, marginal update, and all-in validation costs separately. A no-cost advantage is a possible finding, not an implementation failure to hide.

F Statistical Analysis, Precision, and Output Slots

F.1 Estimands, Pairing, and Covariance

The main target is question-weighted performance under the locked GSM8K sampling regime, conditional on the fitted instrument and realized judge snapshots. For each question i, one paired upstream draw produces V_i^0, V_i^1; with $g_j = g_{J_j}$, let $x_{ij} = g_j(V_i^1) - g_j(V_i^0)$, $\widehat{\tau}_j = N^{-1}\sum_i x_{ij}$, $z_i = x_{i,J_\star^{(0)}} - x_{i,J_0}$, $e_i = x_{i,\text{replay}} - x_{i,\text{fresh}}$, and $b_i^a = g_{J_\star^{(0)}}(V_{i,\text{replay}}^a) - g_{J_\star^{(0)}}(V_{i,\text{fresh}}^a)$ with $e_i = b_i^1 - b_i^0$. Replay/fresh comparisons use the same questions; local binary $u_i = c_{S,i}^1 - c_{S,i}^0$ is distinct from integrated x_{ij}, and diagnostic replicates are averaged within question. Repeated

judge scores are not new upstream observations. For cross-judge contrasts, $\mathrm{Var}(z_i) = \mathrm{Var}(x_{i,J_\star}) + \mathrm{Var}(x_{i,J_0}) - 2\,\mathrm{Cov}(x_{i,J_\star}, x_{i,J_0})$; use $s_z/\sqrt{N}$ and $s_e/\sqrt{N}$, reporting covariance/correlation where defined. Common question IDs do not guarantee strong positive correlation. Role/request-keyed randomness preserves intended marginals and defines a coupling, not a unique joint counterfactual; rescue/harm diagnostics are coupling-dependent, while marginal ATE and expected-score estimation remain primary (Tian & Pearl, 2000). One-sample t tests use question-level paired contrasts. Report 9,999 task-cluster bootstrap draws as a prespecified interval/sensitivity check, keeping each question's versions, arms, routes, labels, and diagnostics together; this addresses question-sampling uncertainty, not future optimizer runs or probe fits. Batch-cluster sensitivity uses the numerical batch manifest; if dependence changes decisions, use the more conservative conclusion or narrow uncertainty interpretation. L1 fixes resampling seeds, interval construction, and sparse-contrast warning rules. Zero variance or few nonzero contrasts trigger a warning and conservative bounded-range interval or estimation-only conclusion, not certainty from zero estimated standard error. For independent contributions in $[a, b]$, the fallback is the clipped Hoeffding interval with half-width $(b-a)\sqrt{\log(2/\eta)/(2N)}$, only if independence is justified. Where $u_i, x_{ij}, b_i^a \in [-1, 1]$ and $z_i, e_i \in [-2, 2]$. Probe evaluation uses one unsteered observation per final question; AUROC/BA require both classes, undefined populations/bootstrap draws are counted and not imputed, and Brier, discrimination, and incremental value are separate claims. Primary predictive evaluation uses the off-arm anchor; a pre-addition on-arm state is not substituted. The final-label diagnostic uses sampled binary judge labels and is not the local correctness-probe metric.

Interpreting the Existing Arm-Wise Gains. For each final judge versus J_0, retain $\gamma_i^a = g_{J_\star}(V_i^a) - g_{J_0}(V_i^a)$ and report $\widehat{G}_a = N^{-1}\sum_i \gamma_i^a$ with prior question-paired descriptive intervals. The identity $z_i = \gamma_i^1 - \gamma_i^0$ links these gains to the primary contrast. $G_0 > G_1 > 0$ describes improvement in both arms while the signed steering contrast decreases; for a negative local effect, this is not automatically attenuation of its absolute magnitude. Selection frequency alone cannot establish these gains. They summarize prior arm terms descriptively, without new primary tests or mixed-arm mediation.

Finite-Record Relevance of Numerical Discrepancies. On nonempty aligned matched set M with identical external candidate labels, let $\delta_i^a = p^a_{\mathrm{replay},i} - p^a_{\mathrm{reconstructed},i}$, $D_i^a = |c^a_{S,i} - c^a_{R,i}|$, and $r_i = \delta_i^1(c^1_{S,i} - c^1_{R,i}) - \delta_i^0(c^0_{S,i} - c^0_{R,i})$. Then $\left|\frac{1}{|M|}\sum_{i\in M} r_i\right| \le B_M := \frac{1}{|M|}\sum_{i\in M}(|\delta_i^1|D_i^1 + |\delta_i^0|D_i^0) \le \delta_{\max}(\widehat{d}_{1,M} + \widehat{d}_{0,M}) \le 2\delta_{\max}$, where $\delta_{\max} = \max_{i\in M,a}|\delta_i^a|$ and $\widehat{d}_{a,M} = |M|^{-1}\sum_{i\in M} D_i^a$. At 10^{-6}, the last ceiling is 0.0002% on those records. This exact arithmetic diagnostic adds no inferential endpoint. Report raw maximum, B_M, signed mean discrepancy, denominator, and faults separately. Invalid labels/lineage are not zero discrepancies; restricted valid sets must not hide missing coverage. Neither B_M nor a sample maximum supplies a population envelope for Proposition 3 or the exceptional-event probability in Appendix B.5; E2b retains its independent, uncorrected comparison. Report planned tasks, behavioral outcomes, probe observations, comparable pairs, distinct cases, and physical attempts separately. Show selected tuning estimate beside independent confirmation to expose selection optimism, without calling either unbiased or using either to retune.

The primary *measurement* separates local binary effects, integrated final contributions, cross-judge contrasts, and replay-fresh discrepancies. Question-level pairing defines empirical covariance; repeated judge scores do not create new upstream observations. Reported estimands remain tied to the declared *intervention and outcome measurements*.

F.2 Confirmatory Family and Evidence Tiers

L3 fixes two mandatory claim-level tests and at most one additional joint claim, namely signed local correctness effect on the independently assessed final panel with one-sided direction fixed at L2 from tuning and not reversed on confirmation or final data, reporting the raw signed estimate and interval; two-sided change in final expected-correctness ATE for $J_\star^{(0)}$ versus J_0; and, only in precision-qualified mode, joint *replay/fresh equivalence* of the ATE discrepancy and both arm-mean discrepancies within $[-0.05, 0.05]$. Use Holm familywise $\alpha = 0.05$ (Holm, 1979). For each equivalence endpoint, use paired TOST and take the larger one-sided p-value; the maximum of the three endpoint p-values is the intersection-claim p-value entering Holm as one claim, not a choice among favorable endpoints. At nominal level a, TOST corresponds to a $100(1-2a)\%$ interval, not automatically 95% (Lakens, 2017); report descriptive 95% intervals separately from adjusted decisions. The family has size two when E2b is estimation-only, otherwise three. E2a

discrete/numerical fidelity checks are engineering criteria, not a fourth test of cache-derived local retention. Seed-one judges, intermediate checkpoints, random-direction comparisons, final-label probe diagnostics, and opportunity subgroups are descriptive/supporting; disagreements with primary analyses are reported and investigated, not resolved by promoting the preferred narrative. Independent confirmation assesses the frozen instrument before main execution, but a 256-question significance result is not a mandatory efficacy gate for a larger independent estimate; weak confirmation is disclosed before L3 and may leave a replay-focused interpretation. Neither confirmation nor final panel reopens intervention search. Final evidence can support an effect under its locked criteria, while claiming an initially validated effect was retained requires an adequately supported original anchor. Null/no-op/low-opportunity/imprecise outcomes remain distinct.

F.3 Precision Calculations: Assumptions, Not E1/E2 Results

Let s be the standard deviation of the question-level contrast. A normal-approximation 95% half-width is $1.96s/\sqrt{N}$; for two-sided testing at $a = 0.05/3$, approximate 80%-power detectable change is $(z_{1-a/2} + z_{0.8})s/\sqrt{N}$, while a single equivalence endpoint centered at zero requires $N \approx \lceil (z_{1-a} + z_{0.9})^2 s^2 / 0.05^2 \rceil$. These assumptions generate Table 8, not measurements of the agent. Joint replay is at least as demanding as its ATE component, and estimated variance, nonzero discrepancy, and sparse data alter power. For $s = 0.40$ and zero discrepancy, approximate equivalence power is 52% at 512 and 94% at 1,024; with a two-point discrepancy it is 61% at 1,024. For $s = 0.60$, even 1,024 gives 41% under zero discrepancy. Thus 1,024 does not universally justify a five-point claim, and the tolerance may be too wide for ranking judges whose effects differ by three points.

Table 8: Analytic planning calculations under hypothetical contrast standard deviations.

s	95% HW		80% DC	N: 80% EP
	$N = 512$	$N = 1,024$	$N = 1,024$	
0.20	1.73	1.22	2.02	187
0.40	3.46	2.45	4.04	745
0.60	5.20	3.67	6.07	1,675
0.80	6.93	4.90	8.09	2,977

Bounded Stress Distributions. A planning-only simulation draws $X \in \{-1, 0, 1\}$ with probabilities $((d - \tau)/2, 1 - d, (d + \tau)/2)$ and $\tau = 0.05$. Paired X, Y have joint law $\rho\,\mathrm{diag}(p) + (1 - \rho)pp^\top$, giving equal marginal effects and correlation ρ; replay error is $X - Y$ with true mean zero. This artificial stress distribution is not generated model answers or an assumed final-agent effect. Here $\mathrm{Var}(X) = d - \tau^2$ and $\mathrm{Var}(X - Y) = 2(d - \tau^2)(1 - \rho)$. A CPU-only calculation uses 20,000 repetitions/case, seed 20260917, and studentized TOST at $0.05/3$; Table 9 reports rounded sensitivity results. `verification/precision/check_planning.py` and its machine-readable output reproduce these hypothetical calculations; no model outputs are simulated. These calculations do not validate the eventual task bootstrap or categorical agent covariance; training-side assessment must replace assumed variation before L3. The calculation's 95% coverage was approximately 94.8%-95.3% across scenarios, for that calculation only. At $N = 1,024$, $d = 0.25$ gives roughly 32.0/79.9/99.8% power for true cross-judge changes of 3/5/8 points; $d = 0.50$ gives 14.9/44.3/89.4%. These illustrate why small update effects can remain unresolved at the larger panel.

Table 9: Planning-only Monte Carlo sensitivity, not E1/E2 results. Percentages estimate single-endpoint equivalence power under the explicit three-point model, not joint replay-claim power.

d	ρ	Error SD	$N = 256$	$N = 512$	$N = 1,024$
0.10	0	0.442	0.1%	34.5%	86.4%
0.10	0.5	0.312	32.7%	85.5%	99.7%
0.25	0	0.704	0.0%	0.0%	11.7%
0.25	0.5	0.497	0.0%	12.5%	71.4%

F.4 Training-Side Precision Decision

After L2 and primary judge selection, 128 prespecified confirmation questions produce two independent complete upstream pairs, adding critic/reviser and independent-draw work while retaining valid local computations where the execution contract matches; all costs count. The optimizer sees none of these effect diagnostics. They estimate the joint distribution of e_i, b_i^0, b_i^1 and timing before final tasks open. At L3 use a prespecified conservative variance assessment, such as the upper 90% bootstrap estimate, with 25% higher-variance sensitivity, and simulate the actual paired tests and joint equivalence decision under zero and two-point discrepancies. Preserve covariance and $e = b^1 - b^0$; do not assign incompatible independent biases. Fix simulation details and variance-uncertainty treatment using training-side data only. Keep fresh $N = 512$ if conservative joint zero-discrepancy planning power is at least 80% and the resource ceiling fits; otherwise choose fresh $N = 1,024$

before final execution if both precision and resource conditions hold; otherwise retain fresh $N = 512$ in estimation-only mode, reporting all differences/intervals and the practical reference region without a confirmatory equivalence verdict. The two-point scenario informs sensitivity, not enlargement of the five-point margin. L3 records whether that margin is useful for the intended claims and attainable E1 detectable change. Final-panel summaries cannot alter this choice.

The L3 decision separates *precision and uncertainty assessment* from the final claim. Training-side variance information can determine whether the prespecified fresh sample is adequate, but it cannot reopen the intervention search or enlarge the equivalence margin. The final mode therefore remains conditional on the declared *information available before final evaluation*.

F.5 Selection, Missingness, and Safeguards

All retained instruments compare with their original anchor. Exact invariance is transitive; adjacent approximate bounds add and do not guarantee original equivalence. Refitting creates another artifact and calibration record. Sample sizes are fixed after L3, with no last-minute optional stopping or sequential guarantees (Ramdas et al., 2023). Malformed numeric answers, invalid categorical-judge candidates, and budget exhaustion are scored failures. Masking makes unsupported judge tokens impossible under the qualified policy; bad token mapping or nonfinite logits indicate execution faults. Infrastructure failures have typed statuses and bounded retries. Missing activations do not delete valid behavioral failures; table denominators separate behavior, probe, and paired-assay coverage. Incomplete pairs cannot be silently discarded as missing at random; report missingness and worst-case bounded outcome sensitivity or withhold the affected population claim. A selected judge is assessed independently of its training selection (Russo & Zou, 2016); unselected-variant evaluation does not validate it. The judge-only search constraint cannot estimate general replay eligibility. Sham equivalence checks instrumentation; random directions can genuinely change outputs, and direct effect/validity comparisons measure specificity. Increased classifier score and syntax-valid damage are not evidence of improved reasoning.

F.6 Output-Cap Inventory and Execution Sequence

Under caps 384/32/192/384, a question's upstream pair uses at most $384 + 2(32 + 192 + 384) = 1,600$ generated tokens, as reasoning precedes intervention; fresh pairs use new reasoning. Table 10 is arithmetic capacity plus allowances, not observed length or A100 timing prediction. A categorical decision may be sampled without materializing a generated token; its one-token allowance does not omit prefill cost. E2b expansion to 1,024 adds 820,224 output-token capacity and the fitting reserve adds 106,496; cap increases require recalculation. Repeated incumbents/duplicates share scores, and gold-free upstream prefixes are generated once per appropriate training population, not per candidate judge. Add judge prefills/loading/checking to measured stage budgets. The 7B BF16 weights occupy about 15.2 decimal GB; 28 layers, four KV heads, and head dimension 128 imply $2 \times 28 \times 4 \times 128 \times 2 = 57,344$ KV bytes/token/sequence, about 0.438 GiB at 8,192 tokens (Hui et al., 2024), excluding activations, logits, workspaces, padding, and allocator reserve. The 14B reflector's approximately 29.4GB BF16 weights are separately loaded. Hardware variant, peak memory, and safe batch size require development measurements; published Qwen A100-80GB benchmarks are single-stream anchors, not short-hooked-call predictions (Qwen Team, 2024b). Operational budget sums measured stage batch latencies, reloads, and non-overlapped CPU/I/O; report useful, not padding, token throughput. CPU fitting/hashing may overlap independent generation; competing GPU copies are not assumed to improve throughput. A 24-48 hour reservation is a conditional scheduling envelope, not a result or procurement guarantee; the resource ceiling is locked from measured development timing. Execution order is development/rejection tests, fitting/tuning, L2/confirmation/controls, adap-

Table 10: Planning cap inventory for the default 1,024-question main and 512-question fresh mode. Assumes 384-token reasoning cap, 32-setting calibration grid, and two search seeds; not empirical resource measurements.

Phase	Output-Token
Development/Profile Allowance	200,000
Probe Fitting: 768×416	319,488
Tuning: Shared Reasoning, 32 Settings plus Off	368,640
Independent Local Confirmation: 256 Pairs	114,688
Training-Side Two-Draw Precision: Additional Work	353,280
Adaptation/Selection Upstream: 384×992	380,928
At Most 13 Unique Judge Policies on 384 Records	4,992
Reflection/Proposal/Repair Allowance	24,576
Three Additional Local Controls on 128 Questions	12,288
Main Upstream: $1,024 \times 1,600$	1,638,400
Main Categorical J_0 *and Two Final Judges*	6,144
Matched E2a: 128 Pairs and Judge	205,056
Independent E2b: 512 Pairs and Judge	820,224
Bounded Challenge Allowance	20,000
Total Cap plus Allowances	4,468,704

tation/selection/14B reflection, training-side precision/L3, main arm records/reconstruction, judge scoring, independent-fresh records, then sealed analysis. Downstream jobs cannot change upstream batch membership/randomization; integrity checks may stop a broken run, while scientific summaries remain sealed until inventory completion.

F.7 Fallbacks, Exclusions, and Results-Dependent Editing

Before L3, a measured resource constraint may select a 512-question, one-stream main panel and 256-question independent-fresh *estimation-only* tier, preserving calibration and 128 matched checks and dropping seed-one replication before compromising reconstruction; this reduced tier has a different precision interpretation. After L3, interruptions do not justify selecting a favorable completed prefix, so resume the fixed inventory or report incomplete/exploratory evidence and withdraw unsupported confirmatory claims. If the probe or local intervention is weak, retain valid predictive or replay findings at their supported scope. Neither a copy judge nor a larger final scale creates absent evidence. A no-op optimizer is reported as such, while an effect contrast near zero is reported with its interval rather than labeled a no-op. Replay discrepancies require distinguishing invalid boundaries, numerical faults, uncertainty, and supported statistical disagreement; repairs after final execution require versioned reruns of affected observations. Upstream prompt evolution is the first possible separate extension, not a rescue analysis. Memory growth, another benchmark/backbone, monitor-gaming, SAE/circuit discovery, and path-specific mediation remain deferred. NARCBench and AgingBench are related contexts rather than additions to this experiment (Rose et al., 2026; Zhu et al., 2026). Results-day editing must report actual N, effects, intervals, opportunity, controls, faults, judge provenance, precision mode, and both cost totals.