EPISTEMIC REGRET MINIMIZATION: GOLD-BLIND CAUSAL PROCESS VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A correct causal answer can be supported by an invalid derivation. We separate premise reliability, WHY-to-WHAT support, and answer correctness, and study Epistemic Regret Minimization (ERM), a gold-blind test-time controller over a frozen target model's exposed argument. Our strongest test is a prospectively frozen, disjoint 200-case CLadder L2 replication. Among 69 initial errors, case-specific versus same-critic, same-query-class shuffled critique changes recovery by +29.0 pp for Claude (95% CI [15.7, 41.1], Holm $p = 2.64 \times 10^{-4}$), +20.3 pp for Gemini ([5.0, 34.4], $p = .0313$), and -10.1 pp for GPT-4o ([-19.2, -3.1], $p = .0313$); GPT-4o also reduces binary false flips on 131 correct controls from 9 to 0. Yet every critique arm lowers raw final accuracy from the unrevised 65.5% baseline to 33.5–53.0%, mainly through abstention, so this identifies the behavioral relevance of case-specific critique rather than an accuracy gain. On an earlier, label-skewed CausalT5K panel, the baseline-suite ERM arm recovers 81/139 GPT-4 Turbo errors versus 63/139 for outcome-only reprompting ($p = .0104$); a separate execution of the same cue-free causal condition recovers 68/139. Benchmark gold is absent from critic and repair prompts and enters only after trajectories freeze. Gold-blind judge calibration and complementary stress tests probe evaluator circularity and robustness. Claims concern auditable public arguments, not hidden chain-of-thought or weight-level learning.

1 INTRODUCTION

A correct WHAT need not have a valid WHY. Consider an interventional query asking whether setting X changes Y. A language model may produce the correct answer while justifying it only with the associational quantity $P(Y \mid X)$, even though the query requires $P(Y \mid do(X))$. If the final YES/NO happens to be correct, an answer-only reward observes success. It receives no information from that scored example about whether the stated derivation answered the causal question that was actually asked. Following the agent-level usage in CausalT5K (Geng et al., 2026), we call this unlicensed lower-rung substitution *Rung Collapse*; it is distinct from legitimate hierarchy collapse, where assumptions make a higher-rung quantity reducible to a lower-rung one (Bareinboim et al., 2022).

This paper studies the supervision gap exposed by that example. Modern outcome supervision is well suited to asking whether the final answer is correct. It does not, by itself, determine whether the argument offered for that answer is licensed. We therefore distinguish

$$\text{premise reliability} \neq \text{WHY-to-WHAT support} \neq \text{WHAT correctness}.$$

These quantities can agree, but they need not. The distinction also avoids an unnecessary claim about hidden chain-of-thought. Intermediate text can omit the actual feature that drove a model's prediction; our target is instead the public argument that users, agents, or downstream systems can inspect and act upon.

Gold-blind process validation. For question Q, supplied information I, stated reasoning R, and proposed answer W, we write

$$S(R \Rightarrow W \mid Q, I) \in [0, 100]$$

for the degree to which the stated WHY supports the stated WHAT conditional on the supplied information. This is not $P(W = Y^\star)$ and it is not a claim that I is factually true. A support judge

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.

can therefore be evaluated without receiving the benchmark answer. Gold is withheld from critique and repair and is introduced only after the resulting trajectories are frozen, to measure outcome correctness.

ERM. Epistemic Regret Minimization (ERM; not empirical risk minimization) places a separate evaluator around a frozen target model. The target first emits (R_0, W_0). The evaluator receives (Q, I, R_t, W_t), scores the WHY-to-WHAT support relation, and identifies a missing inferential link such as an unlicensed association-to-intervention substitution, an uncontrolled back-door path, or a required assumption that the reasoning never states. The target then repairs the WHY and recomputes the WHAT. If the available information cannot support a binary conclusion, the target may abstain. The operational principle is therefore

$$\boxed{\text{diagnose WHY} \;\rightarrow\; \text{repair WHY} \;\rightarrow\; \text{verify WHY} \Rightarrow \text{WHAT} \;\rightarrow\; \text{release WHAT or abstain.}}$$

The decisive empirical question. A causal critique may appear helpful for many reasons unrelated to its causal content: it is extra text, it comes from another model, it sounds authoritative, or it simply creates revision pressure. The critical test is therefore not "does feedback help?" but *does feedback relevant to the actual reasoning defect differ from suitably matched feedback that is irrelevant to that case?* We test this with deterministic derangements: case-specific feedback is compared with feedback produced by the same critic family for another case of the same query class and approximately matched length. The target state, target model, revision prompt, critic family, and query class remain fixed.

Critic capability is part of the method. The answer is not invariant to the evaluator. In our calibration experiments, three frontier judge families differ substantially near the boundary between adequate and near-miss reasoning. The matched-shuffle correction studies show that those differences propagate to target behavior. We therefore treat critic family and discrimination capacity as experimental factors rather than implementation details.

Main findings. Four findings organize the paper. First, in two CLadder specific-versus-shuffled cohorts, case relevance changes behavior, but the effect is critic-dependent. On the disjoint 200-case replication, Claude and Gemini convert relevant feedback into substantially higher recovery, whereas GPT-4o occupies a high-preservation, low-repair operating point with lower recovery and fewer false flips. This is not an accuracy-improvement result: all six critique arms reduce raw final accuracy relative to the unrevised CLadder baseline, mainly through abstention. Second, on an earlier artifact-audited GPT-4 Turbo panel, the baseline-suite ERM arm improves paired recovery over outcome-only reprompting and reduces judge-assessed residual Rung Collapse; a separate execution of the same cue-free condition recovers fewer errors (Section 4.2). Third, error-direction cueing is powerful but unsafe, so cue-free and matched-feedback comparisons are essential; critic family is also a measurement factor. Fourth, observationally matched causal hypotheses remain indistinguishable from observational data alone; textual critique cannot create missing causal evidence.

Contributions.

1. **Outcome–process separation.** We formalize Rung Collapse at the level of an exposed argument and formalize why answer-only reward cannot discriminate process-valid from process-invalid derivations when they produce the same scored answer.
2. **Gold-blind causal process validation.** We define ERM as a frozen-model audit-and-repair controller over the support relation $S(R \Rightarrow W \mid Q, I)$, explicitly separating premise reliability, reasoning adequacy, and answer correctness.
3. **A matched-shuffle specificity test.** We compare case-specific feedback with a deterministic same-critic, same-query-class derangement, isolating relevance from generic critique exposure. A prospectively frozen 200-case disjoint replication yields significant paired differences for all three critic families after Holm correction, with opposite recovery direction for GPT-4o.
4. **Recovery, preservation, and abstention as separate endpoints.** We report recovery, binary false flips, coverage, and selective accuracy rather than treating final accuracy as a sufficient summary of a correction controller.
5. **Evaluator calibration and causal information limits.** HardCal measures WHY-to-WHAT discrimination without exposing a correct answer, while an observational–interventional separation

proposition and controlled Gaussian witness delimit what process feedback can infer without new causal evidence.

2 OUTCOME–PROCESS SEPARATION

We work within structural causal models. Pearl's hierarchy distinguishes association (L1: $P(Y \mid X)$), intervention (L2: $P(Y \mid do(X))$), and counterfactual reasoning (L3). The hierarchy is an information distinction: a lower-rung quantity may support a higher-rung answer only when additional assumptions license the reduction.

Definition 1 (Rung Collapse)**.** *An exposed argument exhibits* Rung Collapse *when it answers a Rung-j query using Rung-i reasoning, $i < j$, without stating assumptions or evidence that license that reduction. A reduction justified by the task's causal assumptions is not a collapse.*

Definition 2 (Conditional process validity)**.** *For supplied question Q, information I, reasoning R, and proposed answer W, let $V(Q, I, R, W) \in \{0, 1\}$ denote whether the stated reasoning licenses the proposed answer conditional on the supplied information. A graded judge score $S(R \Rightarrow W \mid Q, I)$ is an operational estimator of this relation, not an estimator of factual truth or hidden computation.*

Proposition 1 (Outcome–process aliasing)**.** *Let two derivations R_1 and R_2 for the same (Q, I) produce the same answer W, with $V(Q, I, R_1, W) \neq V(Q, I, R_2, W)$. For every outcome reward of the form $r_{\text{out}} = r(W, Y^\star)$,*

$$r_{\text{out}}(R_1) = r_{\text{out}}(R_2).$$

Hence the scored outcome on that instance contains no discriminative signal between the two processes.

Proof. The reward's arguments are identical for the two derivations because both produce the same W and are evaluated against the same $Y^\star$. Therefore the reward values are equal. The claim is local to the scored instance: it does not assert that parameter-sharing, auxiliary losses, exploration, or other examples yield identical optimizer updates. □

Corollary 1 (Spurious success is invisible to answer-only scoring)**.** *If a process-invalid derivation and a process-valid derivation produce the same correct answer, an answer-only reward cannot rank the valid derivation above the invalid one on that instance.*

The corollary is the precise content needed for this paper. We do *not* infer that positive reward must amplify the invalid process under weight training. The present experiments are test-time correction experiments, not measurements of gradient dynamics.

Proposition 2 (Gold-blind process validation)**.** *A validator whose target is $V(Q, I, R, W)$ need not receive $Y^\star$ as an input. This states an information boundary, not correctness of the validator: premise reliability, evaluator error, and external factual soundness remain separate sources of uncertainty.*

Proposition 3 (Matched-shuffle specificity contrast)**.** *For case i, let F_i be feedback generated for that case and $F_{\pi(i)}$ feedback from a deranged donor in the same critic family and query class, with approximate length matching and no self-match. Holding the target state and revision procedure fixed, define*

$$\Delta_{\text{rel}} = \mathbb{E}\big[M_i(F_i) - M_i(F_{\pi(i)})\big],$$

where M_i is a prespecified behavioral endpoint. Under the null that case relevance carries no information beyond the controlled feedback attributes, and under within-stratum donor exchangeability, $\mathbb{E}[\Delta_{\text{rel}}] = 0$.

This proposition states the identification logic of the matched-shuffle experiment. It does not claim perfect causal randomization: residual donor-content differences and finite-sample imbalance remain possible, which is why we replicate on an entirely disjoint cohort and report critic-specific results.

Proposition 4 (Observational–interventional separation)**.** *Let M_0 and M_1 induce the same observational law $P(X, Y)$. Any statistic or data-dependent score based only on independent samples from that shared law has the same sampling distribution under M_0 and M_1. If, for some x,*

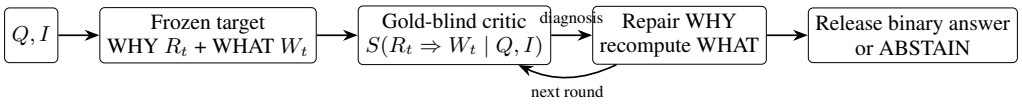


Figure 1: ERM evaluates the support relation between the exposed WHY and WHAT conditional on supplied information. Benchmark gold is excluded from critique and repair and enters only after trajectories are frozen for outcome evaluation.

$P_{M_0}(Y \mid do(X = x)) \neq P_{M_1}(Y \mid do(X = x))$*, then independent outcomes under* $do(X = x)$ *can distinguish the two fixed hypotheses with error tending to zero as the number of interventions grows.*

This is a two-hypothesis specialization of the information separation expressed by Pearl's causal hierarchy and the Causal Hierarchy Theorem (Bareinboim et al., 2022); we state it explicitly because it defines the boundary of the textual-feedback experiments.

Proof sketch. Part one follows by equality in distribution: every measurable function of observational samples has the same law under M_0 and M_1. For part two, distinct interventional laws disagree on a measurable event; empirical frequencies or a likelihood-ratio test therefore separate the fixed hypotheses consistently. The full assumptions and Gaussian witness remain in the appendix. □

3 Epistemic Regret Minimization

3.1 Gold-blind audit and repair

A frozen target model receives (Q, I) and produces (R_0, W_0). A separate critic J receives (Q, I, R_t, W_t) and returns a support score S_t and a diagnosis F_t. The target then emits a revised pair

$$(R_{t+1}, W_{t+1}) = T(Q, I, R_t, W_t, F_t).$$

Benchmark gold and reference rationales are excluded from both J and T. Gold enters only after trajectories are frozen. We use *gold-blind* to describe this information boundary; constructing an initially wrong evaluation cohort is necessarily label-dependent at analysis time.

The causal critic checks the public argument rather than hidden cognition. Typical defects include association substituted for intervention, an open back-door path, an omitted adjustment assumption, or a conclusion whose direction is not supported by the quantities computed in the rationale. A valid argument should be acknowledged rather than criticized merely to induce change.

For a normalized support score $s_t = S(R_t \Rightarrow W_t \mid Q, I)/100$, we define the operational epistemic regret of the exposed argument as

$$\mathcal{R}_{\text{epi}}(t) = 1 - s_t.$$

ERM "minimizes" this externally measured support deficit through repair; the present paper makes no claim that $\mathcal{R}_{\text{epi}}$ is differentiated through target-model weights.

3.2 Release policy and abstention

The method separates *feedback generation* from *release policy*. For a calibrated support set-point τ and a maximum repair budget K, define

$$t^\star = \min\{t \leq K : S_t \geq \tau\}.$$

A deployment controller may release the first state at $t^\star$, continue feedback when the set-point has not been reached, or abstain/defer after the repair budget is exhausted when the supplied information still does not license a binary conclusion. We call the last outcome *wise refusal* when the refusal is justified by unresolved support or information insufficiency: refusing to manufacture certainty is distinct from both an incorrect answer and a technical failure. We do *not* classify every observed ABSTAIN in the experiments as wise; experimental abstentions are reported descriptively unless

their justification is separately validated. This separation prevents the online judge from changing which later states exist in our fixed-round experiments and makes coverage measurable.

We therefore report four endpoints:

Recovery,	Binary False Flip,	Coverage,	Selective Accuracy.

Recovery is measured among initially incorrect cases; binary false flip among initially correct controls; coverage is the fraction receiving a binary answer rather than ABSTAIN; selective accuracy is correctness conditional on coverage. Final accuracy alone can conceal a controller that improves safety primarily by refusing.

3.3 What ERM does not assume

ERM does not require the exposed rationale to be a faithful mechanistic transcript of hidden model computation. It asks whether the rationale that was actually emitted supports the conclusion that was actually proposed. Nor does a high support score certify the supplied premises. A soundness audit must additionally establish premise reliability from external evidence. Moreover, because a strong critic may also solve the query itself, a case-specific critique can transmit its own directional conclusion as well as an explanation of what is wrong. HardCal tests discrimination among rationales for a fixed WHAT, not independence from that channel. The matched-shuffle contrast therefore estimates the behavioral relevance of the *whole critique*; it does not isolate a verdict-free reasoning signal from the critic's conclusion.

3.4 Scope: test-time validation, not weight training

The experiments keep target weights frozen. "Minimization" refers to reducing diagnosed support deficits through iterative external repair, not to gradient descent over model parameters. Section 6 gives a concrete weight-level extension as future work.

4 Experiments

The experiments are organized around four research questions that map directly to the theory.

- **RQ1 — Outcome–process separation.** Can a gold-blind critic alter correction relative to outcome-only baselines without increasing false flips? *Evidence:* the audited CausalT5K paired panel; formal boundary: Proposition 1.
- **RQ2 — Feedback specificity.** Does case-specific critique differ from same-critic, same-query-class shuffled critique? *Evidence:* two CLadder cohorts; identification logic: Proposition 3.
- **RQ3 — Critic capability.** Do evaluator differences correspond to different recovery–preservation–abstention behavior? *Evidence:* fixed-trace audit, HardCal, scenario-blind, and paraphrase stress tests.
- **RQ4 — Information boundary.** What ultimately requires new interventional evidence? *Evidence:* Proposition 4 and the Gaussian witness.

4.1 Analysis discipline

The artifact-audited CausalT5K panel retains only scenario-identified, executed, paired records. A failed call is not silently counted as a model response; unparsed executed corrections remain non-recoveries. The admitted GPT-4 Turbo baseline contains 139 initial errors and 35 initially correct controls. Because the error cohort is strongly label-skewed, recovery is never interpreted as proof that the critic recovered causal structure; matched content controls and separate balanced CLadder panels address that concern.

The CLadder specificity studies use a different identification strategy. Gold is absent from all critic and target-revision prompts. Case-specific feedback is paired against a deterministic same-critic, same-query-class derangement with no self-match. The second cohort contains 200 cases with zero overlap with the first 100-case panel. Frozen sample and official-archive hashes are reported in Appendix E.

Table 1: Artifact-audited GPT-4 Turbo baseline suite. All methods use the same 139 initial errors and 35 correct controls. Residual collapse is the archived judge-assessed process label; paired statistics are reported in the text and release artifact.

Method	Calls	Recovery	Residual collapse	False flips
Outcome-only reprompt	1	63/139 (45.3%)	57/139 (41.0%)	1/35
Self-consistency, $N = 3$	3	36/139 (25.9%)	95/139 (68.3%)	0/35
Best-of-N by confidence	3	10/139 (7.2%)	123/139 (88.5%)	0/35
Generic feedback	2	7/139 (5.0%)	85/139 (61.2%)	0/35
Matched non-causal critique	2	39/139 (28.1%)	80/139 (57.6%)	0/35
Baseline-suite ERM critique	2	81/139 (58.3%)	12/139 (8.6%)	1/35

4.2 RQ1: Artifact-audited correction

In this execution, the baseline-suite ERM arm recovers 81/139 errors (58.3%) versus 63/139 (45.3%) for outcome-only reprompting, with the same 1/35 binary false flips. The paired McNemar comparison is $p = .0104$. The archived process auditor marks residual Rung Collapse in 12/139 ERM outputs (8.6%) versus 57/139 outcome-only outputs (41.0%). These are judge-assessed process labels, not independent causal ground truth.

Protocol reconciliation. The 81/139 baseline-suite result and the 68/139 C4 result below are *separate executions*, not two names for one condition. Both are gold-blind and cue-free: neither the critic nor the revision prompt receives the benchmark answer or an instruction that the previous answer is wrong. The baseline suite was executed in its own interleaved comparison session, whereas C4 belongs to a separate six-condition decomposition. Their difference is run-to-run/protocol-instance variation and is reported rather than pooled; the paper therefore does not treat 58.3% as a universal ERM rate.

The six-condition decomposition remains essential because it shows why a larger raw correction number is not automatically stronger evidence. On GPT-4 Turbo, the explicit error-direction cue produces large recovery changes but also false flips. Cue-free causal critique recovers 68/139 versus 33/139 for the matched non-causal critique, while differing only modestly from 63/139 under outcome-only reprompting. On GPT-5.2, cue-free causal and non-causal critique each recover 1/39, below 13/39 for outcome-only. Thus ERM is not claimed to have a uniform target-independent advantage.

Tone $\times$ diagnosis. On the same 100 CLadder L2 case identities used in Cohort A, an earlier frozen 2×2 study varies revision tone while withholding any statement that the previous answer is wrong. The tone study revises the frozen initial responses, whereas the later Cohort A matched-shuffle test begins from the common no-feedback revision state; the case panel is shared, but the intervention stage is not. Under neutral wording, outcome-only revision recovers 14/37 initial errors but flips 14/63 controls, whereas causal diagnosis recovers 6/37 and flips 4/63. Under forceful wording, outcome-only recovers 4/37 with 4/63 false flips, while causal diagnosis recovers 10/37 with the same 4/63 false flips. The force-matched recovery contrast is +16.2 percentage points (paired exact $p = .0313$, Holm $p = .0625$). Tone is therefore an experimental factor that interacts with feedback content rather than a generic improvement knob; full counts are in Appendix D.

4.3 RQ2: Feedback specificity on CLadder

Why shuffled feedback is the critical control. A matched non-causal paragraph controls vocabulary but not case-specific information. A stronger null gives the target feedback produced by the *same critic* for another causal case of the same query class. If case relevance carries no information beyond critique exposure, the paired specific-versus-shuffled contrast should vanish. Because a detailed critique also communicates the critic's assessment of the defect, this contrast tests the relevance of the complete case-specific critique; it does not separately identify the explanatory reasoning from the critic's verdict.

Cohort A: 100-case prospective mechanism study. The first cohort contains 100 balanced CLadder L2 cases, with 37 initial errors and 63 initially correct controls. From each case's frozen no-feedback revision state, each of three critic families supplies either its case-specific detailed critique

Table 2: Disjoint 200-case CLadder replication. Recovery is on 69 initial errors; FF is binary false flip on 131 initially correct controls. CI: 95% Newcombe paired interval for the recovery difference (Newcombe, 1998); raw p-values are in Table 7. Holm adjustment is across the three prespecified recovery tests.

Critic	Spec. rec.	Shuf. rec.	Δ	95% CI	Holm p	FF (S/H)
Claude Sonnet 5	30/69	10/69	+29.0 pp	[15.7, 41.1]	.000264	3/131 vs. 9/131
Gemini 3.8 Flash	25/69	11/69	+20.3 pp	[5.0, 34.4]	.0313	3/131 vs. 6/131
GPT-4o	3/69	10/69	-10.1 pp	[-19.2, -3.1]	.0313	0/131 vs. 9/131

or a deterministic same-class shuffled critique. Claude-specific feedback recovers 17/37 errors versus 7/37 shuffled; Gemini recovers 9/37 versus 4/37; GPT-4o recovers 5/37 versus 6/37. The paired exact raw p-values are .0213, .1797, and 1.0, respectively; after Holm adjustment over the three critic families, the Claude comparison is suggestive but not below .05 ($p = .0638$). Binary false flips are 2 versus 4 for Claude, 2 versus 5 for Gemini, and 4 versus 3 for GPT-4o. The direction is therefore critic-dependent even before the replication.

Cohort B: 200 disjoint cases with fresh target states. We prospectively freeze 200 additional CLadder L2 cases (98 ATE, 98 back-door adjustment, and 4 collider-bias), with zero overlap with Cohort A, and generate fresh GPT-4 Turbo initial responses. The cohort contains 69 initial errors and 131 correct controls. All 200 initial target states, all 600 critic evaluations, and all 1,200 target revisions complete successfully. Gold enters only after all generation has frozen.

Equality of specific and shuffled recovery is rejected for all three critic families after Holm correction, but the direction differs by critic. We do not pool Cohorts A and B: Cohort A begins from a frozen no-feedback revision state, whereas Cohort B begins from fresh initial GPT-4 Turbo responses; the second cohort is therefore a disjoint replication of the relevance contrast, not an identical procedural rerun. Claude strongly replicates Cohort A: relevant feedback raises recovery from 14.5% to 43.5% ($\Delta = +29.0$ pp, 95% CI [15.7, 41.1]). Gemini's Cohort-A directional pattern is stronger in Cohort B, 15.9% to 36.2% ($\Delta = +20.3$ pp, [5.0, 34.4]). GPT-4o shows the opposite recovery effect, 14.5% to 4.3% ($\Delta = -10.1$ pp, [-19.2, -3.1]), while eliminating binary false flips from 9/131 to 0/131. Because the comparison holds critic family, query class, target state, and revision procedure fixed, the result is inconsistent with an explanation based only on receiving additional feedback or another model call.

Negative result: critique lowers raw CLadder accuracy. The unrevised Cohort B starting point is 131/200 correct (65.5%) at full coverage. Every one of the six critique arms lowers final accuracy, to 45.0–48.5% for Claude, 33.5–34.5% for Gemini, and 52.0–53.0% for GPT-4o. The dominant mechanism is withdrawal of initially correct answers as ABSTAIN: 28–84 of the 131 controls abstain after critique (84 under Gemini-specific feedback), in addition to the separately reported binary false flips. Because CLadder supplies the quantities that determine each answer, these abstentions are coverage losses, not wise refusals. Thus this CLadder experiment supports a claim about *feedback relevance and operating-point change*, not a claim that critique-and-repair improves end-to-end accuracy on this benchmark. A post-hoc case-level audit clarifies that the GPT-4o pattern is not simply generalized abstention: on the 69 initially wrong cases it assigns much higher support to the frozen initial arguments than Claude or Gemini and recovers only the three cases it scores at zero. Appendix E.3 reports these descriptive diagnostics; they do not establish that calibration score causally determines repair.

Abstention and selective accuracy. Relevant feedback changes release behavior as well as correctness. In Cohort B, specific versus shuffled coverage is 54.0% versus 59.0% for Claude, 43.0% versus 47.0% for Gemini, and 75.5% versus 76.0% for GPT-4o. Selective accuracy among answered cases is 89.8% versus 76.3%, 80.2% versus 71.3%, and 70.2% versus 68.4%, respectively. Claude and Gemini therefore trade additional abstention for substantially improved correctness among released binary answers, whereas GPT-4o primarily shifts toward preservation. These operating points should not be collapsed into a single final-accuracy number.

4.4 RQ3: CRITIC CAPABILITY IS AN EMPIRICAL FACTOR

The same three critic families are evaluated on two independent diagnostics. On the fixed 300-trace audit, three-judge structural agreement is substantial (Fleiss $\kappa = .762$), but not perfect. On 116 initially correct L1 controls, broad structural-flag rates remain low: 3/116 for Gemini, 4/116 for GPT-4o, and 5/116 for Claude. On the harder gold-blind HardCal benchmark, 30 matched scenario families each contain adequate, near-miss, weak, and unsupported rationales for the same premises and proposed WHAT. Each of 120 items is judged twice by each critic, yielding 720 blind judgments. The primary endpoint is whether adequate reasoning outranks the matched near-miss within family.

Claude separates adequate from near-miss reasoning in 30/30 families, Gemini in 28/30 with two ties, and GPT-4o in 21/30 with nine ties. AUROC is .986, .965, and .858, respectively, with high test–retest stability. A post-hoc Cohort-B audit shows the same critic ordering descriptively in support assigned to initially wrong traces and recovery, but with only three critic families this is neither a meaningful correlation test nor a causal identification result (Appendix E.3). Critic identity is therefore an empirical factor, not an interchangeable implementation choice. Because these critics can also solve many CLadder questions, case-specific feedback may transmit an implied answer; the experiments isolate relevance of the full critique, not explanatory reasoning stripped of that verdict. HardCal, the scenario-blind stress test, and the paraphrase intervention provide complementary checks on evaluator circularity: respectively, they test gold-free discrimination of reasoning quality, dependence on scenario semantics rather than graph or variable-name cues, and robustness to meaning-preserving changes in expression (Appendix F).

4.5 RQ4: THE INFORMATION BOUNDARY

The text experiments study whether the emitted WHY supports the emitted WHAT. They do not create physical or external interventions. We therefore instantiate Proposition 4 separately in two linear-Gaussian SCMs that share the same observational law but differ under $do(X)$. The learner compares only those two specified hypotheses and never receives the generating-graph label.

At prior $P(X \to Y) = 0.90$, interventional updating yields mean cumulative held-out interventional-prediction regret 8.23 versus 184.54 for the observational controller; all 100 seeds cross the spurious-edge rejection threshold at median episode 25.5. The same updater crosses the opposite direct-edge threshold in all 100 seeds in the matched true-direct world. Across prespecified priors 0.50, 0.75, 0.90, and 0.99, the confounded-world rejection succeeds in 100/100 seeds. This is a controlled witness of information separation, not unrestricted causal discovery and not evidence that textual critique can manufacture missing intervention data.

5 RELATED WORK

Causal reasoning and process supervision. CLadder (Jin et al., 2023), CausalProbe (Chi et al., 2024), CauSciBench (Acharya et al., 2025), and CausalT5K (Geng et al., 2026) document persistent causal failures in LLMs. Classic process-supervision work compares outcome and step-level feedback when correctness supervision exists (Uesato et al., 2022; Lightman et al., 2024); ERM instead audits whether an exposed causal argument licenses its conclusion without exposing the benchmark answer to the critic. This follows the validity-versus-soundness distinction emphasized by Lawrence & Maasch (2026): premise truth, inferential adequacy, and answer correctness are separate questions.

Test-time refinement, critics, and trace limits. Self-consistency (Wang et al., 2023) and Self-Refine (Madaan et al., 2023) refine inference, while LLM-Modulo systems place external critics around a generator (Kambhampati et al., 2024). ERM is closest to the latter but targets causal commitments and makes the evaluator itself an experimental object. Trace-faithfulness work shows that stated reasoning can omit true drivers (Turpin et al., 2023; Chen et al., 2025; Lanham et al., 2023); ERM therefore treats emitted rationales as auditable arguments, not mechanistic explanations. Broader shortcut-learning results likewise show that high task accuracy need not identify the process that produced it (Geirhos et al., 2020; McCoy et al., 2019; D'Amour et al., 2022).

Relation to post-training. The present method uses process supervision at inference time around frozen target weights. That is an experimental scope choice, not a conceptual barrier. A gold-blind

ERM process score can be combined after repair with ordinary gold outcome reward, or ERM-repaired traces can define supervised and preference-training examples. We formalize this extension in Section 6; whether the resulting weight updates transfer process validity to unseen causal structures remains future work.

6 Discussion and Conclusion

The paper begins from a simple mismatch: an answer score evaluates the WHAT, while a causal failure can reside in the WHY. The outcome–process aliasing proposition formalizes the resulting information loss: if two derivations produce the same scored answer, answer-only reward cannot separate them on that instance. ERM supplies a distinct supervision channel by asking whether the public argument licenses its conclusion conditional on the supplied premises.

The empirical evidence supports a specific claim about feedback relevance and critic-dependent operating points. On the disjoint 200-case CLadder replication, specific versus shuffled feedback yields Holm-corrected paired differences for Claude, Gemini, and GPT-4o, with 95% intervals that do not cross zero. Claude and Gemini convert relevant feedback into greater recovery; GPT-4o converts it into lower recovery with fewer false flips. All critique arms nevertheless reduce raw final accuracy relative to the unrevised baseline, mainly through abstention. The preliminary CausalT5K panel supplies a complementary paired correction result but is label-skewed and exhibits run-to-run variation. Thus relevance matters, but "better critique" is not a one-dimensional scalar intervention: the critic and target jointly define a recovery–preservation–abstention operating point.

This result also clarifies what ERM does *not* establish. A support judgment over an emitted rationale is not proof of hidden chain-of-thought faithfulness. Because a detailed critique includes the critic's defect verdict as well as its explanation, the present matched-shuffle experiment does not isolate a verdict-free reasoning channel; a verdict-stripped control remains an important next experiment. A valid argument conditional on supplied premises is not proof that those premises are true. A high-recovery controller is not necessarily safe if it achieves recovery by flipping correct answers, and a high-accuracy controller may simply abstain. Finally, textual feedback cannot resolve causal hypotheses that are observationally indistinguishable without additional identifying assumptions or interventional evidence.

Future work: process-aware post-training. ERM's process signal can in principle become a weight-level training signal without making the critic differentiable. After gold-blind diagnosis and repair has frozen the process record, reveal $Y^\star$ only for outcome scoring and define

$$r_{\text{joint}}(\tau) = \alpha r_{\text{out}}(W, Y^\star) + \beta r_{\text{proc}}(Q, I, R, W), \qquad \alpha, \beta \geq 0.$$

For a trainable policy π_θ,

$$\nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta}[r_{\text{joint}}(\tau)] = \mathbb{E}[(r_{\text{joint}}(\tau) - b)\nabla_\theta \log \pi_\theta(\tau)]\,,$$

so a black-box scalar process score is compatible with score-function policy optimization such as PPO/GRPO (Schulman et al., 2017; Shao et al., 2024). ERM-repaired traces can also define supervised or preference-training data. The falsifiable hypothesis is that joint outcome-plus-process post-training reduces held-out process-invalid successes relative to outcome-only training without reducing answer accuracy. Such a study must also guard against reward hacking: a trainable policy could learn artifacts of one ERM judge rather than acquire transferable process validity. Held-out causal structures, cross-critic transfer, adversarially varied judges, and independent outcome checks are therefore necessary. We do not test weight-level transfer here: a convincing study requires open weights, repeated training runs, and substantially more training compute than the present frozen-model experiments.

The practical conclusion is therefore specific. **Correctness of WHAT, adequacy of WHY, and whether to answer at all should be measured separately; when the WHY cannot be repaired to the set-point under available evidence, wise refusal is a legitimate terminal action.** ERM provides a gold-blind estimator and repair mechanism for WHY-to-WHAT support. Case-specific critique is not behaviorally interchangeable with matched irrelevant feedback, but no textual critic can supply causal evidence absent from the premises.

REFERENCES

Sawal Acharya, Terry Jingchen Zhang, Andrew Kim, Anahita Haghighat, Xianlin Sun, Rahul Babu Shrestha, Maximilian Mordig, Furkan Danisman, Clijo Jose, Yahang Qi, Pepijn Cobben, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. CauSciBench: Assessing LLM causal reasoning for scientific research. In *NeurIPS 2025 Workshop on Uncovering Causality in Science*, 2025. URL https://openreview.net/forum?id=EO8mTLqDuT. Workshop paper.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. ACM Books, 2022. doi: 10.1145/3501714.3501743.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. arXiv preprint.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-3064. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/af2bb2b2280d36f8842e440b4e275152-Abstract-Conference.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Longling Geng, Andy Ouyang, Theodore Wu, Daphne Barretto, Matthew J. Hayes, Rachael Cooper, Yuqiao Zeng, Sameer Vijay, Gia Ancone, Ankit Rai, Matthew Wolfman, Patrick Flanagan, and Edward Y. Chang. CausalT5k: Diagnosing refusal and failure modes in trustworthy causal reasoning across causal rungs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026)*, 2026. doi: 10.1145/3770855.3817567.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems 36*, volume 36, pp. 31038–31065, 2023.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B. Murthy. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22895–22907. PMLR, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. arXiv preprint.

Rachel Lawrence and Jacqueline Maasch. Position: Reasoning is a learnable rule-based process. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2608.12325.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1334.

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. arXiv preprint.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

AI USE STATEMENT

Generative AI tools were used for LaTeX editing and prose revision; to critique and restate mathematical claims and proofs; to propose falsification and robustness experiments and the hypotheses in Appendix H; and to assist with experiment-script implementation, number verification, and the interpretation of results. The ERM framework was originated by the authors, and no numerical result was generated by an AI model: reported numbers come from executed code and stored outputs. We take responsibility for the final content of this work, including text, proofs, experimental designs, code, and interpretations produced with AI assistance.

REPRODUCIBILITY STATEMENT

The accompanying artifacts retain the artifact-audited CausalT5K paired panel, the six-condition cue/content decomposition, three-judge fixed-trace audit, HardCal calibration outputs, the 100-case CLadder mechanism cohort, and the prospectively frozen 200-case disjoint replication. The latter uses 200 fresh GPT-4 Turbo initial states, 600 critic evaluations, and 1,200 target revisions; all calls completed under the frozen protocol; one non-JSON critic response was retried without changing prompts, cases, or analysis. Its case set has zero overlap with the earlier 100-case cohort. Gold labels are never supplied to support judges or target-revision prompts in either CLadder specificity

experiment and enter only after generation has frozen. The package also retains the raw Cohort-B critic and target logs, all 720 HardCal judgments, and deterministic post-hoc forensic tables used in Appendix E.3. The observational–interventional witness remains a separate controlled simulation with its own prespecified priors and 100-seed runs.

APPENDIX CONTENTS

	Section	Page
A	Formal statements and proofs	12
B	Endpoints and statistical tests	13
C	Artifact-audited CausalT5K baseline	13
D	Tone $\times$ diagnosis study	13
E	CLadder feedback-specificity experiments	14
E.3	Case-level critic–target forensic audit	15
F	Judge validity without gold	15
G	Controlled observational–interventional separation	17
H	Post-training extension	18
I	Inference-time cost and measurement boundary	18
J	Reproducibility checklist	18

A FORMAL STATEMENTS AND PROOFS

A.1 OUTCOME–PROCESS ALIASING

For fixed (Q, I) and answer W, let R_1 and R_2 be two exposed derivations with different process-validity labels. If an outcome reward is a function only of $(W, Y^\star)$, its arguments are identical for the two derivations, so the rewards are identical. This proves Proposition 1. The result is deliberately local: it does not assert identical gradients after parameter sharing, auxiliary losses, exploration, or other examples.

A.2 GOLD-BLIND PROCESS VALIDATION

Proposition 2 is an information statement. Conditional process validity is defined on (Q, I, R, W); therefore $Y^\star$ is not a logical input to the target relation. A practical judge can of course be wrong. The proposition separates the object being judged from the judge's empirical accuracy.

A.3 MATCHED-SHUFFLE SPECIFICITY

For each case i, the experiment holds target state, target model, revision prompt, critic family, and query class fixed and replaces only F_i by a deterministic deranged donor $F_{\pi(i)}$. Under the null that case relevance has no information beyond those controlled attributes, donor exchangeability within the matching stratum gives zero expected paired contrast. The experiment does not claim perfect randomization; this is why critic-specific results are reported and the contrast is repeated on a disjoint cohort.

A.4 OBSERVATIONAL–INTERVENTIONAL SEPARATION

Suppose $P_{M_0}(X, Y) = P_{M_1}(X, Y)$. Then an i.i.d. observational sample has the same law under both hypotheses, and every measurable statistic of that sample has the same sampling law. Hence no observational-data-only test can consistently identify which fixed SCM generated the observations. If there exists x such that $P_{M_0}(Y \mid do(X = x)) \neq P_{M_1}(Y \mid do(X = x))$, the two interventional laws disagree on a measurable event. By the law of large numbers, its empirical frequency converges to different limits under the two fixed hypotheses, so a consistent test separates them. This proves Proposition 4.

B Endpoints and Statistical Tests

Recovery is the fraction of initially incorrect cases that become gold-correct after revision. Binary false flip is the fraction of initially correct controls that become an incorrect binary answer; ABSTAIN is tracked separately. Coverage is the fraction of cases with a binary answer, and selective accuracy is accuracy conditional on coverage. The primary matched-shuffle analysis uses paired exact two-sided binomial tests on discordant recovery outcomes within each critic family, followed by Holm correction across the three prespecified critic families. Safety and abstention endpoints are secondary and are reported separately rather than folded into recovery.

C Artifact-Audited CausalT5K Baseline

The admitted GPT-4 Turbo panel contains 139 initial errors and 35 initially correct controls. Admission requires one scenario identifier, a parsed initial answer eligible for the relevant stratum, and one executed record in every compared arm. Failed calls are not silently counted as model responses; unparsed executed corrections remain non-recoveries.

Table 3: Artifact-audited GPT-4 Turbo baseline suite, reproduced from the main paper. Residual collapse is judge-assessed.

Method	Calls	Recovery	Residual collapse	False flips
Outcome-only reprompt	1	63/139 (45.3%)	57/139 (41.0%)	1/35
Self-consistency, $N = 3$	3	36/139 (25.9%)	95/139 (68.3%)	0/35
Best-of-N by confidence	3	10/139 (7.2%)	123/139 (88.5%)	0/35
Generic feedback	2	7/139 (5.0%)	85/139 (61.2%)	0/35
Matched non-causal critique	2	39/139 (28.1%)	80/139 (57.6%)	0/35
Baseline-suite ERM critique	2	81/139 (58.3%)	12/139 (8.6%)	1/35

The six-condition decomposition separates the error-direction cue from critique content. On GPT-4 Turbo, outcome-only reprompting recovers 63/139; cue-only 110/139; benchmark critique plus cue 123/139; gold-blind causal critique 68/139; gold-blind non-causal critique 33/139; and gold-blind causal critique plus cue 115/139. Corresponding false flips on 35 controls are 2, 9, 2, 0, 0, and 2. On GPT-5.2, cue-free causal and non-causal critique each recover only 1/39 in the retained run, while outcome-only recovers 13/39. These results motivate treating cueing and critic content as separate experimental factors.

D Tone × Diagnosis Study

This study uses the same 100 CLadder L2 case identities as Cohort A (37 initial errors, 63 correct controls) and crosses feedback content (outcome-only versus case-specific causal diagnosis) with revision tone (neutral versus forceful). The experiments are not the same intervention: this tone study revises the frozen initial responses, whereas Cohort A's later matched-shuffle specificity test starts from a common no-feedback revision state. No condition states that the previous answer is incorrect.

Table 4: Tone × diagnosis on the Cohort A cases.

Condition	Recovery	False flips	Final accuracy	Answer changes
Outcome-only, neutral	14/37 (37.8%)	14/63 (22.2%)	63/100	28/100
Causal diagnosis, neutral	6/37 (16.2%)	4/63 (6.3%)	65/100	10/100
Outcome-only, forceful	4/37 (10.8%)	4/63 (6.3%)	63/100	8/100
Causal diagnosis, forceful	10/37 (27.0%)	4/63 (6.3%)	69/100	14/100

The force-matched causal-versus-outcome recovery contrast is 10/37 versus 4/37 (+16.2 pp), with raw paired exact $p = .03125$ and Holm-adjusted $p = .0625$. Forcefulness reduces recovery for outcome-only revision but increases it when paired with case-specific diagnosis; tone is therefore an interacting control factor rather than a generic compliance booster.

E CLADDER FEEDBACK-SPECIFICITY EXPERIMENTS

E.1 COHORT A: 100 CASES

The first cohort contains 100 balanced CLadder L2 cases: 37 initial errors and 63 correct controls. From a frozen common no-feedback revision state, each critic supplies either a case-specific detailed critique or a deterministic same-query-class shuffled critique. The target model is GPT-4 Turbo.

Table 5: Cohort A matched-shuffle results. Holm correction is over the three critic-specific recovery tests.

Critic	Spec. recovery	Shuf. recovery	Raw p	Holm p	FF spec./shuf.
Claude Sonnet 5	17/37	7/37	.0213	.0638	2/63 vs. 4/63
Gemini 3.8 Flash	9/37	4/37	.1797	.3594	2/63 vs. 5/63
GPT-4o	5/37	6/37	1.000	1.000	4/63 vs. 3/63

A production amendment preserved explicit insufficiency responses as ABSTAIN rather than treating them as technical failures. This amendment was made before benchmark outcomes were inspected and changed neither prompts, model assignments, nor gold access. ABSTAIN is non-correct for benchmark accuracy and is reported separately from binary false flips.

E.2 COHORT B: 200 DISJOINT CASES

The second cohort is prospectively frozen from the official CLadder v1 balanced archive: 98 ATE, 98 back-door adjustment, and 4 collider-bias cases, with zero overlap with Cohort A. The original 100-case information construction was recovered exactly as `given_info` alone on all 100 old IDs. The new cohort contains 69 initial errors and 131 correct controls after fresh GPT-4 Turbo initial generation. All 200 initial responses, 600 critic evaluations, and 1,200 target revisions completed under the frozen protocol. Gold was absent from every generation, critique, and revision prompt.

Table 6: Cohort B arm summaries. FF is binary false flip on 131 initially correct controls; Ctrl-A is control abstention; SelAcc is selective accuracy among binary answers.

Critic	Feedback	Recovery	FF	Ctrl-A	Coverage	Final acc.	SelAcc
No revision	Initial	–	0/131	0	1.000	.655	.655
Claude	Shuffled	10/69	9/131	42	.590	.450	.763
Claude	Specific	30/69	3/131	61	.540	.485	.898
Gemini	Shuffled	11/69	6/131	69	.470	.335	.713
Gemini	Specific	25/69	3/131	84	.430	.345	.802
GPT-4o	Shuffled	10/69	9/131	28	.760	.520	.684
GPT-4o	Specific	3/69	0/131	28	.755	.530	.702

The no-revision row is the frozen initial state: 131/200 correct (65.5%) with full coverage. Every critique arm lowers raw final accuracy to 33.5–53.0%, mainly because initially correct cases are withdrawn as ABSTAIN. This is a substantive negative result, not a bookkeeping artifact; Cohort B is an identification study of case relevance and release behavior rather than evidence that critique improves overall accuracy.

Table 7: Cohort B primary paired recovery tests.

Critic	Spec. rec.	Shuf. rec.	Spec-only	Shuf-only	Raw p	Holm p
Claude Sonnet 5	30	10	23	3	.000088	.000264
Gemini 3.8 Flash	25	11	22	8	.016125	.031250
GPT-4o	3	10	0	7	.015625	.031250

Cohorts A and B are not pooled because their starting-state procedures differ. Cohort B is a disjoint replication of the relevance contrast, not an identical procedural rerun. A case-specific detailed critique includes the critic's defect judgment as well as explanatory content; therefore the matched-shuffle contrast estimates relevance of the complete critique, not a verdict-free reasoning component.

E.3 POST-HOC CASE-LEVEL CRITIC–TARGET FORENSIC AUDIT

After the confirmatory outcomes were frozen, we joined the 600 gold-blind critic records to the 1,200 target revisions and benchmark outcomes to study *how* the critic-dependent operating points arise. This analysis is descriptive and post-hoc: it does not alter the prespecified paired tests and should not be read as identifying a causal relation between a critic's calibration score and downstream repair.

Table 8: Cohort-B case-level anatomy under case-specific critique. "Wrong support" is the critic's support score on the 69 initially incorrect target arguments. Final states partition each initial-error or correct-control population into correct, wrong, and ABSTAIN.

Critic	Wrong support	Median	Error final C/W/A	Control final C/W/A	Rec.	FF
Claude Sonnet 5	23.2	15	30/8/31	67/3/61	30/69	3/131
Gemini 3.8 Flash	51.1	40	25/14/30	44/3/84	25/69	3/131
GPT-4o	83.6	90	3/45/21	103/0/28	3/69	0/131

The GPT-4o pattern is therefore not well described as generalized refusal. Its specific arm abstains on 21/69 initially wrong cases and 28/131 initially correct controls, while leaving 45/69 errors wrong and preserving 103/131 controls. More diagnostically, the three recovered GPT-4o cases are exactly the three initially wrong cases receiving support score 0; the remaining 66 wrong cases have mean support 87.4 and median 90. By comparison, recovered cases receive mean support 9.2 from Claude and 26.6 from Gemini, versus 34.0 and 65.0 on their respective unrecovered errors. These patterns are consistent with *under-diagnosis of defective support* as one mechanism behind GPT-4o's low-repair/high-preservation operating point. They do not establish that support scores cause repair, nor do they justify attributing the behavior to provider safety tuning.

HardCal supplies an independent diagnostic of critic discrimination. Claude and Gemini have adequate-versus-near-miss mean margins of 37.7 and 32.6 with AUROC .986 and .965, whereas GPT-4o's margin is 17.8 with AUROC .858 (Table 9); replicate stability remains high for all three. Across only three critic families, the descriptive ordering is suggestive but is not a meaningful correlation test. Establishing a general calibration–repair relationship requires more critics or controlled within-family perturbations.

Directional information inside detailed critiques. The complete critique can carry both explanation and an answer-direction signal. As a deliberately broad lexical screen, 152/200 Claude critiques, 140/200 Gemini critiques, and 90/200 GPT-4o critiques contain an explicit whole-word `YES` or `NO`. This screen over-counts verdict transmission because a critique may merely quote the target's proposed answer; it is therefore not an estimate of "leakage." It does establish that the detailed-feedback channel is not verdict-free. For example, on case `nonsense6-frontdoor-ate-model6710-spec6-q1`, GPT-4o assigns support 0 and explicitly states that the calculations imply the opposite of the proposed `YES`; the target then repairs to `NO`. Conversely, on several initially wrong cases GPT-4o assigns support 100 while explicitly endorsing the proposed answer, and the target preserves it. These examples motivate, but do not replace, the predeclared verdict-only ablation in Appendix F.

Frozen provenance. Official archive SHA-256: `9fdae052b1ebe4ee6a19fdfb3e1eb88a381c7345df394cea524bcce97e2349b3`. Frozen 200-case sample SHA-256: `061982736ea578a95072808a58fecc4c9204309024f1e35d13c6cc846acfc30f`. The shuffle map contains 600 assignments, zero self-matches, and zero cross-query-class matches.

F JUDGE VALIDITY WITHOUT GOLD: THREE COMPLEMENTARY STRESS TESTS

A central concern for an LLM process evaluator is circularity: the critic might appear to assess reasoning quality only because it reconstructs a benchmark answer and rewards agreement with that answer. ERM blocks the direct form of this failure by withholding benchmark gold and reference rationales from critique and repair. That information boundary is necessary but not sufficient evidence that the judge is useful. We therefore combine three complementary stress tests that attack

different alternative explanations: gold-based answer reconstruction, shallow graph/lexical scoring, and wording sensitivity.

As a preliminary sensitivity check, on a fixed 300-trace panel three-judge structural agreement is substantial (Fleiss $\kappa = .762$), but not perfect. Among 116 initially correct L1 controls, broad structural-flag rates are 3/116 for Gemini, 4/116 for GPT-4o, and 5/116 for Claude. Required-rung calibration on the 100 L2 cases is 100/100, 75/100, and 99/100, respectively. Judge family is therefore an empirical measurement factor rather than an interchangeable implementation choice.

F.1 HardCal: gold-blind discrimination of reasoning adequacy

HardCal holds the question, supplied premises, and proposed WHAT fixed within each of 30 matched scenario families while varying only the quality of the stated WHY: adequate, near-miss with one material bridge omitted, weak-but-relevant, and unsupported/circular. Each of the 120 items is judged twice by each of three critic families, yielding 720 gold-blind judgments. The primary endpoint is within-family adequate $>$ near-miss. No benchmark answer or correct-answer label is supplied to the judge.

Table 9: HardCal gold-blind judge calibration. Margin is mean adequate-minus-near-miss support score.

Critic	Adequate $>$ near-miss	Mean margin	AUROC	Threshold	Test–retest ρ
Gemini 3.8 Flash	28/30	32.6	.965	99	.993
GPT-4o	21/30	17.8	.858	100	.956
Claude Sonnet 5	30/30	37.7	.986	96	.984

All three critics also pass the three source-conditional traps and three verbosity traps. The post-hoc Cohort-B audit in Appendix E.3 compares these profiles with downstream support scores and repair states while explicitly avoiding a three-point correlation claim. HardCal therefore shows that the evaluator can discriminate reasoning adequacy without direct access to benchmark gold; it does not prove that the evaluator is always correct, that the supplied premises are true, or that the emitted rationale is mechanistically faithful.

F.2 Scenario-blind stress test: does the judge use the scenario?

Gold blindness does not exclude a weaker shortcut in which the judge scores only graph form or variable-name cues. We therefore compare three views of the same 105 GPT-4 Turbo Rung-Collapse cases: the full scenario plus trace, the extracted DAG alone, and an anonymized DAG with variables renamed to generic symbols.

Table 10: Scenario-blind judge analysis on $N = 105$ GPT-4 Turbo Rung-Collapse cases. One full-condition structured-parse failure uses the experiment's prespecified fallback.

Judge input	Structural score	Collapse rate	κ vs. full
Full scenario + trace	57.0 ± 24.6	35.2%	–
DAG only	78.4 ± 15.5	12.4%	-0.08
Anonymized DAG	76.4 ± 25.2	18.1%	-0.04

Among the 104 cases with parsed collapse labels in both full and DAG-only conditions, 34 are detected only with full context, 10 only in the DAG-only condition, and 3 in both. Removing scenario semantics therefore changes the judgment substantially. Anonymizing variable names does not restore the full-context pattern, so the difference is not explained simply by recognizable variable names. This stress test supports context sensitivity, not correctness of every full-context judgment.

F.3 Paraphrase robustness: is the diagnosis a wording artifact?

A second shortcut hypothesis is lexical fragility. On 158 GPT-4 Turbo Rung-Collapse-case bundles, an independent validator accepted 333/474 stylistic paraphrases as meaning-preserving. Among

accepted pairs, Rung Collapse detection agrees with the original in 91.7% of cases ($\kappa = .799$), and failure-signature labels agree in 89.8% ($\kappa = .748$). Exact extracted commitment sets are substantially less stable: exact agreement is 27.9%, with mean Jaccard similarity approximately .49. The appropriate claim is therefore robustness of the diagnostic conclusion, not invariant symbolic extraction detail or hidden-chain-of-thought reconstruction.

Table 11: Paraphrase stress test on independently validated meaning-preserving paraphrases.

Quantity	Agreement	Chance-corrected / similarity statistic
Rung Collapse diagnosis	91.7%	$\kappa = .799$
Failure signature	89.8%	$\kappa = .748$
Exact commitment set	27.9%	mean Jaccard $\approx .49$

F.4 Triangulation and remaining boundary

The three tests attack different threats. HardCal asks whether the evaluator can rank reasoning quality without a correct answer. The scenario-blind test asks whether the decision materially depends on scenario semantics rather than graph shape or variable names. The paraphrase intervention asks whether the diagnosis survives meaning-preserving changes in expression. Together they support describing the ERM judge as a *gold-blind, context-sensitive, and substantially paraphrase-robust estimator of exposed WHY-to-WHAT support*.

The conclusion remains deliberately limited. These experiments do not prove hidden reasoning faithfulness or turn judge outputs into ground truth. Moreover, a strong critic can solve the benchmark item itself, so a case-specific detailed critique may transmit the critic's directional conclusion as well as explanatory reasoning. HardCal measures discrimination of reasoning quality, not independence from such task solving. The matched specific-versus-shuffled experiments therefore estimate the behavioral effect of the complete case-specific critique, not an isolated verdict-free reasoning channel. A verdict-stripped critique condition remains an important future control. We predeclare the interpretation rule: on a frozen disjoint cohort, compare detailed critique, critic verdict without explanation, and matched shuffled critique under the same target state and revision prompt. If verdict-only matches the detailed-critique recovery effect within paired uncertainty while preserving the same safety profile, we will not attribute the effect specifically to explanatory reasoning; the stronger claim requires an incremental detailed-over-verdict effect.

G Controlled Observational–Interventional Separation

Two linear-Gaussian SCMs are constructed to share the same observational $P(X, Y)$ but differ under $do(X)$. The learner compares only the two specified hypotheses and never receives the generating-graph label. In the confounded world, 500 interventions are performed for each of 100 seeds at four prespecified priors.

Table 12: Confounded-world separation witness. Crossing means posterior direct-edge probability below 0.10.

Prior	Obs. regret	Int. regret	Ratio	Crossed	Median episode
0.50	57.79	4.05	14.26×	100/100	13
0.75	128.47	5.86	21.93×	100/100	19
0.90	184.54	8.23	22.42×	100/100	25.5
0.99	223.09	14.41	15.48×	100/100	42

A symmetric matched true-direct falsification crosses posterior direct-edge probability 0.90 in 100/100 runs. The experiment witnesses information separation between two specified hypotheses; it is not unrestricted causal discovery.

H Post-Training Extension

The main paper evaluates frozen targets. A future open-weight study can introduce gold only after ERM diagnosis and repair and optimize

$$r_{\text{joint}} = \alpha r_{\text{out}}(W, Y^\star) + \beta r_{\text{proc}}(Q, I, R, W).$$

A non-differentiable process judge can still supply a scalar reward to a score-function estimator,

$$\nabla_\theta \mathbb{E}[r_{\text{joint}}] = \mathbb{E}[(r_{\text{joint}} - b)\nabla_\theta \log \pi_\theta(\tau)].$$

Three experiments are natural: supervised fine-tuning on verified ERM repairs, preference optimization over original-versus-repaired traces, and policy-gradient post-training with the joint reward. The decisive evaluation should use held-out causal structures and report both answer correctness and process validity, with correct-case controls to detect indiscriminate critique behavior. It must also test reward hacking explicitly: a policy may learn to satisfy quirks of one process judge without acquiring transferable causal validity. Cross-critic evaluation, adversarial judge variation, and held-out structures are therefore required.

I Inference-Time Cost and Measurement Boundary

The current paper evaluates an external test-time controller, so its cost is attributable in model calls. After the common initial target response, a single ERM repair uses one critic call plus one target revision call. A one-shot outcome-only reprompt uses one target revision call. Counting the common initial response, the corresponding full paths therefore use three versus two model calls. Multi-round ERM is capped by the explicit repair budget K and can stop at the support set-point τ.

The current CLadder specificity runs were designed for identification rather than systems benchmarking and do not contain complete per-case wall-clock latency or token accounting across all providers. We therefore do not reuse token multipliers from older experiments with different prompts or iteration policies, and we make no claim that ERM is latency- or token-efficient. A deployment study should report provider-specific input/output tokens, wall-clock latency, dollar cost, and quality as a function of K, τ, critic family, and selective routing. This explicit measurement boundary addresses practical overhead without mixing stale cost estimates into the present protocol.

J Reproducibility Checklist

The release should preserve: (i) scenario identifiers and admission masks for the 139-error/35-control CausalT5K panel; (ii) every executed prediction used in paired tests; (iii) the six-condition cue/content decomposition; (iv) Cohort A frozen inputs and shuffle map; (v) Cohort B `FROZEN_PROTOCOL.json`, `frozen_source_200.csv`, gold-blind initial/critic/target exports, shuffle map, and final analysis tables; (vi) HardCal item definitions, all 720 raw judgments, both replicates, thresholds, and family-margin/stability analyses; (vii) the Cohort-B case-level forensic tables and deterministic analysis script; and (viii) the controlled separation script and all 100-seed outputs. Model endpoints, decoding settings, prompt hashes, and protocol amendments should be retained verbatim in the artifact archive. Gold must remain absent from critic and target-revision prompts in the CLadder specificity experiments.