

Encevra Research Round 29

**Make the Evidence Move: Active Source Interventions and the Limits of
Dependency Discovery**

Encevra Research Program

October 20, 2026

A source-grounded paper with an exact finite diagnostic and reproducibility bundle

Contents

Abstract	2
1. Research question	3
2. Why this is the next round	3
3. Scope and exclusions	4
4. Definitions that prevent category errors	4
5. Literature map: what is established, contested, and missing	5
6. Claim-evidence map	6
7. Proposed mechanism: the Interventional Lineage Test	7
8. Exact diagnostic E29-A	9
9. Brain-inspired hypothesis, kept testable	11
10. Alternative explanations and rival mechanisms	11
11. Falsifiable hypotheses	12
12. Proposed live experiment E29-B	13
13. Expected failure modes	15
14. Limitations	15
15. Practical conclusions	15
16. Next-round handoff	16
17. Sources and inspection record	16
18. Reproducibility statement	17

Research date: 20 October 2026

Program: Encevera Research - reliable reasoning, evidence exchange, and answer revision

Status: Research synthesis plus an exact finite diagnostic; proposed live experiment not yet run

Result boundary. This round did not run a production incident study, a language-model benchmark, a live multi-agent benchmark, or a Noetis answer benchmark. The new numerical results are exact only for the finite deterministic diagnostic defined here. They do not establish production performance, language-model capability, or the fidelity of any real intervention.

Abstract

Encevera’s preceding rounds progressively separated answer correctness from evidence quality, source lineage, observation sufficiency, verification allocation, and model adequacy. Round 28 ended with a concrete unresolved question: can an active intervention reveal a missing evidence dependency that confidence, disagreement, surprise, and passive structural scores cannot detect?

Round 29 asks one narrower question: **under a fixed verification budget, what can controlled interventions establish about dependencies between collaborating evidence channels, and when must the system abstain because different provenance structures remain interventionally equivalent?** The answer has two parts.

First, causal-discovery research supports a positive but conditional claim. Interventions can refine an observational equivalence class when they separate candidate structures, and sequential selection can spend experiments where they are most informative. That result is bounded by the assumed model family, intervention targets, action fidelity, and the absence or explicit treatment of latent variables. Unknown or off-target interventions can make the true target ambiguous; a certificate inside a misspecified version space can therefore be confidently wrong.

Second, Encevera’s exact diagnostic demonstrates a sharper boundary for evidence systems. We exhaustively evaluate 72 structural worlds: 36 monotone two-channel input-output behaviors, each implemented once with separate evidence roots and once as a shared-root facade. Across 1,728 policy-world-budget traces, three task-content interventions identify all 36 observable response functions, yet identify root provenance in **0%** of worlds. The two provenance implementations are deliberately interventionally equivalent under every admissible content probe. A naive rule that reads a strict

local input-output interface as root independence is wrong in **50%** of the paired worlds. This is an impossibility result for the declared diagnostic, not a population estimate.

We propose an **Interventional Lineage Test (ILT)**: maintain an explicit candidate set; choose safe, decision-relevant interventions that separate candidates; require an intervention receipt proving what actually changed; update only after fidelity checks; and stop at the finest decision-relevant equivalence class the admissible actions can identify. Content perturbations test behavioral coupling. Path isolation, root replacement, or another root-addressing action is required to test provenance. Even verified root rotation establishes a new post-intervention trust boundary; it does not retroactively prove historical independence.

1. Research question

Primary question

Under a fixed verification budget, which controlled interventions can distinguish independent evidence roots from shared-root facades in a collaborative reasoning system, and what evidence justifies an independence claim rather than an explicit unresolved-equivalence finding?

Decision target

The system must choose among three operational conclusions:

1. **Dependency detected:** at least one channel's output behavior or evidence lineage depends on the other channel or on a shared root.
2. **Independence established within a declared scope:** the admissible intervention family separates the relevant dependency alternatives and the observed, fidelity-validated trace eliminates them.
3. **Unresolved equivalence:** two or more decision-relevant structures predict the same outcomes under every admissible intervention; the system must not convert an untestable distinction into a clean bill of health.

The third conclusion is not indecision. It is the identifiable result.

2. Why this is the next round

The Lovable archive was inspected before topic selection. At archive commit 7e3a99e454a140d1206c308f1b9bda2 the published sequence ended at Round 28, so the correct next number is Round 29, not the older requested number 11.

The carry-forward is cumulative:

- **Round 9** proposed learning reusable procedures only from verified experience. Round 29 does not study procedural learning; it studies what an intervention can verify before an experience is trusted.
- **Round 10** introduced structured evidence exchange, challenge, selective revision, and matched-cost evaluation. Round 29 keeps those principles but does not repeat them. It isolates the upstream question of whether apparently separate evidence paths are actually distinguishable.
- **Rounds 24 and 25** made source authority, lineage, and omission part of the evidence object. Their implication is that agreement is not two votes when both votes inherit the same root.
- **Round 26** separated a good-looking observation from observation sufficiency. Round 29 changes the environment rather than asking a passive observer to infer what it cannot see.
- **Round 27** separated verification allocation from adjudication and showed why a dependent check can consume budget without adding independent evidence. Round 29 asks which verification actions are genuinely separating.
- **Round 28** separated target confidence, structural ambiguity, and model adequacy. Its constructed hidden-common-origin alias survived passive diagnostics. Its explicit handoff was to

test source isolation, evidence withholding, authority rotation, delayed re-observation, or challenge queries.

The novelty is therefore not “agents should exchange evidence” or “agents should revise selectively.” It is the testable distinction between **observing more outputs** and **changing an evidence-generating path in a way that can discriminate provenance hypotheses**.

3. Scope and exclusions

In scope

- Two or more evidence channels contributing to one decision.
- Hidden common origins, cross-channel leakage, cached evidence, and facade independence.
- Safe task-content probes, path isolation, source withholding, and root rotation.
- Fixed-cost intervention selection and decision-relevant stopping.
- Intervention-target uncertainty and off-target effects.
- Honest certification relative to an explicitly declared candidate family.

Out of scope

- Training a new foundation model.
- Claiming that a machine system literally implements a brain mechanism.
- General causal discovery over arbitrary unbounded programs.
- Proving historical provenance from a newly independent configuration.
- Treating an intervention response as evidence when the action itself is not verified.
- Reporting proposed experiments as completed experiments.

4. Definitions that prevent category errors

4.1 Evidence root

An **evidence root** is the earliest decision-relevant upstream origin that can transmit the same error, manipulation, or unsupported assumption into multiple downstream channels. A root may be a document, database, retrieval index, model-generated summary, analyst, cached artifact, or shared preprocessing step.

4.2 Behavioral dependency

Channel B is behaviorally dependent on channel A, relative to intervention family I, when at least one valid intervention on A changes B’s observable response while relevant background conditions are held fixed. This is a claim about the response surface, not necessarily about ultimate provenance.

4.3 Root-provenance independence

Two channels are root-independent only if their relevant upstream error pathways do not share a root within the declared boundary. Local responsiveness does not prove this. Two wrappers can respond independently to prompts while drawing decisive content from the same database or cached synthesis.

4.4 Interventional equivalence

Candidate structures are interventionally equivalent under I when every action in I produces the same observable distribution or deterministic trace in each structure. No policy restricted to I can distinguish those structures, regardless of how many times it repeats the actions.

4.5 Action fidelity

Action fidelity is the degree to which the actual intervention matches the intended target, intensity, isolation boundary, and timing. It is not safe to infer structure from an intended action when the action may have missed its target, changed additional variables, leaked through a cache, or altered the task itself.

4.6 Certificate scope

An intervention can support a certificate only relative to:

- a candidate model family;
- an admissible intervention family;
- a declared observation interface;
- validated action fidelity;
- and an explicit decision predicate.

“No dependency found” without those qualifiers is not a certificate.

5. Literature map: what is established, contested, and missing

5.1 Position A: interventions can break observational equivalence

Hauser and Buhlmann characterize how intervention targets refine observational Markov-equivalence classes. Hyttinen, Eberhardt, and Hoyer formulate experiment selection through separating systems and cut-coverings. Eberhardt, Glymour, and Scheines derive worst-case experiment bounds under strong assumptions. He and Geng propose batch and sequential designs using minimax and maximum-entropy criteria. Together these works support a precise point: **active changes can distinguish structures that passive observations cannot, but only when the changes separate the candidate structures under the assumed model** [S01-S04].

5.2 Position B: context and latent structure can be modeled, but assumptions remain decisive

Joint Causal Inference represents multiple contexts and interventions in one causal framework, including settings where intervention type or target is not fully known. Work on intervention design with latent variables shows that useful experiment design is still possible with hidden structure and explicit costs. This broadens the positive case, but it does not remove assumptions about causal semantics, faithfulness, allowable interventions, or the model family [S05-S06].

5.3 Position C: the intervention itself may be uncertain

Varici et al. and Hagele et al. treat intervention targets as objects of inference rather than trusted metadata. Choo et al. analyze off-target interventions in which an attempted action may affect a random subset. Yang et al. show that latent confounding can prevent unique target identification, making a candidate superset the appropriate output. These results directly motivate an intervention receipt and a fidelity gate: an unverified manipulation cannot safely update the structural ledger as though it were exact [S07-S10].

5.4 Position D: invariance and negative controls are useful diagnostics, not universal provenance proofs

Invariant prediction uses stability across environments to support causal conclusions under explicit invariance assumptions. Negative controls can expose confounding and bias when the control’s structural conditions hold. Both are valuable, especially when direct intervention is unsafe, but a stable or clean result can still be non-separating for the provenance question [S11-S12].

5.5 Position E: active certificates inherit their version-space boundary

A September 2026 preprint proposes actively selecting interventions that separate observationally equivalent causal models and then certifying agreement within a current version space. Its own bypass analysis makes the central limitation visible: if the true mechanism lies outside that space, a certificate can be void. Random validation interventions refuted many void certificates in the reported setting, but not all. This is promising recent evidence, not a settled guarantee [S13].

5.6 Agent benchmarks are becoming interactive, but do not settle provenance certification

CausalGame evaluates LLM agents in interactive causal games with confounding, selection, and measurement challenges. ACED-Bench evaluates evidence acquisition and action in hidden probabilistic graphical models with a controlled intervention interface and evidence ledger. They provide useful test scaffolds and metrics for experiment choice. Neither result establishes that a real multi-agent evidence path has independent roots; ACED-Bench also excludes latent confounding and model mis-specification in its current formulation [S14-S15].

5.7 Human evidence offers a design analogy, not a machine guarantee

Human studies show that confidence and information seeking can dissociate, and that people can select interventions that balance expected information with inferential complexity. Round 29 uses this only as a functional analogy: a verifier should be able to trigger a targeted “information-seeking mode” instead of treating subjective certainty as sufficient. It does not claim that the proposed protocol copies a brain or inherits human optimality [S16-S17].

6. Claim-evidence map

Claim	Status	Direct support	What would falsify or narrow it
Passive agreement cannot generally identify a hidden common root.	Established in the Encevera constructed archive; generally consistent with observational equivalence.	Round 28 [A01]; interventional-equivalence literature [S01-S03].	A passive observable proven to differ for every relevant common-root alternative in the declared family.
A separating intervention can refine a candidate structural class.	Established under model assumptions.	[S01-S06].	A candidate pair that the claimed intervention predicts differently but that remains identical under a faithful implementation.
Intended targets cannot be assumed to equal actual targets.	Established concern; methods exist for several formal settings.	[S07-S10].	Direct, independently verified execution semantics that rule out target and spillover uncertainty for the tested action.
Three content probes identify every behavior in the finite Round 29 model.	Exact result for this diagnostic only.	Exhaustive E29-A diagnostic and eight passing tests.	A behavior pair in the enumerated family with the same four-point response signature.
Those same probes identify root provenance in none of the paired worlds.	Exact impossibility result for this diagnostic only.	Provenance twins have identical traces under all admissible content probes.	A content-only trace that differs between either member of a paired provenance twin.

Claim	Status	Direct support	What would falsify or narrow it
A verified root-addressing intervention is necessary for a root-independence claim when provenance twins are content-interventionally equivalent.	Logical consequence within the declared candidate family; Encevera hypothesis for practice.	E29-A equivalence classes plus [S07-S10, S13].	A non-root-addressing observable that validly separates every decision-relevant provenance twin.
Decision-relevant stopping can reduce intervention cost without forcing complete graph recovery.	Supported in formal experiment-selection work; not yet tested in Encevera's live setting.	[S02-S04, S13].	Matched-budget experiments in which stopping at a predicate-equivalence class increases target decision error or costs more.
Intervention receipts will reduce false-clean provenance judgments in deployed agent workflows.	Open Encevera hypothesis.	Motivated by unknown/off-target-intervention literature [S07-S10].	A preregistered matched-cost evaluation showing no reduction, or a regression in task utility that dominates any gain.

7. Proposed mechanism: the Interventional Lineage Test

ILT is not a universal causal-discovery algorithm. It is a decision protocol for testing a narrow evidence-lineage predicate under a fixed budget.

7.1 Inputs

For each decision, ILT requires:

1. **Decision predicate:** for example, "Are channels A and B sufficiently root-independent to count as two corroborating sources?"
2. **Candidate set:** a bounded set of plausible dependency structures, including at least one hidden-common-root or facade alternative.
3. **Admissible actions:** safe interventions the system is actually authorized and able to perform.
4. **Observation interface:** outputs, source accesses, hashes, timing, cache status, and execution telemetry available after each action.
5. **Cost envelope:** tokens, model calls, tool calls, external requests, elapsed-time ceiling, and any human-review cost.
6. **Loss function:** the relative cost of a false independence claim, a false dependency claim, an abstention, and delayed action.

7.2 Intervention hierarchy

ILT distinguishes actions by what they can test.

Level	Intervention	Example	What it may identify	What it cannot establish alone
0	Passive repeat	Ask both agents again with the same evidence	Stochastic consistency	Structural independence
1	Content perturbation	Change only source A's claim-relevant fact	Behavioral coupling through the exposed interface	Ultimate root provenance

Level	Intervention	Example	What it may identify	What it cannot establish alone
2	Path isolation	Block A-to-B messages and shared caches; rerun B	Dependence on the blocked path if fidelity is verified	Absence of an unblocked common root
3	Root replacement	Replace A's upstream root with an independently administered source	Prospective response to a new root and a new trust boundary	Historical independence before replacement
4	Root intervention plus control	Root replacement together with canaries, negative controls, and telemetry	Stronger evidence that the intended boundary changed	Truth outside the declared observation and model family

7.3 The intervention receipt

Every structural update is conditioned on a machine-readable receipt:

```
intervention_id
decision_id
intended_target
actual_target_evidence
isolation_boundary
pre_state_hash
post_state_hash
cache_state
content_delta
allowed_spillovers
observed_spillovers
start_time / end_time
executor_identity
telemetry_links
fidelity_status: valid | uncertain | invalid
```

An invalid action is not counted as a negative result. An uncertain action preserves, rather than eliminates, candidates that would be ruled out only by exact targeting.

7.4 Selection rule

At each step, ILT groups surviving candidates by the outcome each admissible action predicts. It selects a safe action that maximizes expected reduction in **decision-relevant** ambiguity per unit cost, subject to a worst-case guardrail. If no action can separate candidates with different decision labels, ILT stops and returns unresolved equivalence. If all surviving candidates share the same label, it may stop without recovering the complete graph.

A practical score is:

$\text{score}(\text{action}) = \text{expected decision-entropy reduction} / \text{verified total cost}$

with a constraint that the largest possible remaining opposite-decision class must not exceed a pre-registered threshold. The score is only a ranking heuristic; the certificate comes from the surviving set and validated trace.

7.5 Update and stop rules

After an action:

1. validate the receipt and classify fidelity;
2. compare the observation with each candidate's predicted outcome set;
3. retain candidates consistent with both observation and fidelity uncertainty;
4. run a model-adequacy check or a reserved validation probe;
5. stop when the decision predicate is identified, the budget is exhausted, or the remaining candidates are equivalent under every remaining action.

The output must name the surviving equivalence class and the intervention family. It must not render "all surviving candidates are clean" as "the world is clean" when the true world could lie outside the family.

8. Exact diagnostic E29-A

8.1 Purpose

E29-A tests a minimal claim before a costly live benchmark: can rich content interventions distinguish observable behavior while still leaving root provenance fundamentally unidentified?

8.2 Model

There are two binary source facts, x_A and x_B , and two channel outputs, y_A and y_B . Each output uses one of the six monotone Boolean functions of the two facts:

ZERO, AND, A, B, OR, ONE.

This yields 36 ordered behavior pairs. Every behavior is instantiated with two structural implementations:

- **SEPARATE_ROOTS**
- **SHARED_ROOT_FACADE**

The paired implementations expose the same input-output function by construction. There are therefore 72 structural worlds.

The baseline input is $(0, 0)$. The admissible content interventions are:

- **A_ONLY** = $(1, 0)$
- **B_ONLY** = $(0, 1)$
- **JOINT** = $(1, 1)$

The experiment exhaustively evaluates six policies at budgets 0 through 3:

- passive repeat;
- one-sided A;
- reciprocal A then B;
- full factorial A, B, then joint;
- greedy expected-entropy reduction;
- minimax worst-class splitting.

All 72 worlds, six policies, and four budgets produce **1,728 policy-world-budget traces**. There is no sampling error because enumeration is complete. The uniform paired construction is a diagnostic design, not a prevalence model.

8.3 Reproducibility

The bundle includes:

- `round29_interventional_lineage.py`;
- `test_round29_interventional_lineage.py`;
- all world traces;
- policy aggregates;
- full-factorial equivalence classes;
- an exact summary;
- source-inspection ledger;
- SHA-256 manifest.

The eight unit tests verify model cardinalities, unique behavior signatures, complete full-factorial behavior identification, provenance-twin equivalence, non-identification of provenance under every content policy, the residual alias after reciprocal probes, the informationlessness of passive repeats, and deterministic artifact generation.

8.4 Results

Policy and budget	Exact behavior identified	Coupling predicate decided	Strict-local-interface predicate decided	Root provenance identified	Mean probes
Passive repeat, budget 3	2.78%	2.78%	30.56%	0%	3.00
One-sided A, budget 1	2.78%	36.11%	83.33%	0%	1.00
Reciprocal A/B, budget 2	44.44%	66.67%	100%	0%	2.00
Full factorial, budget 3	100%	100%	100%	0%	3.00
Greedy entropy, budget 3	100%	100%	100%	0%	2.50
Minimax split, budget 3	100%	100%	100%	0%	2.50

Four facts matter:

1. Repetition adds no information in this deterministic model.
2. Reciprocal single-factor probes resolve the strict-local-interface predicate but leave interaction aliases; the joint probe is needed to identify every monotone behavior.
3. Greedy and minimax policies identify all behaviors while stopping early in some worlds, averaging 2.5 probes.
4. No content-probe policy identifies provenance. Full behavioral identification and zero provenance identification coexist.

8.5 Worked example: the perfect facade

Consider the local behavior:

- channel A outputs x_A ;
- channel B outputs x_B .

The complete trace is:

Action	Input (x_A, x_B)	Output (y_A, y_B)
Baseline	(0, 0)	(0, 0)
A only	(1, 0)	(1, 0)
B only	(0, 1)	(0, 1)
Joint	(1, 1)	(1, 1)

This is the cleanest possible local response surface. Yet both structural implementations generate it:

- in one world, the two channels really have separate roots;
- in the other, a shared upstream root feeds two wrappers that expose locally controlled outputs.

No additional content probe within the declared binary interface can separate the pair. If a verifier equates strict local response with root independence, it labels both worlds independent and is wrong in one of the two - an exact **50% false-root-independence rate within this deliberately balanced pair**. The correct output is: behavioral locality identified; provenance unresolved.

Now rotate channel B to a genuinely independent, independently administered root and verify the isolation receipt. The new run can establish a prospective independence boundary for the rotated configuration. It still cannot prove that the earlier B was historically independent.

9. Brain-inspired hypothesis, kept testable

The useful analogy is experimental blocking, not literal neural copying. In human active causal-structure tasks, participants can intervene on or block elements and appear to trade expected information against inferential complexity [S17]. Enceavra's engineering hypothesis is correspondingly modest:

A verifier with a small repertoire of targeted “block, replace, and re-observe” operations will achieve lower decision-relevant ambiguity per unit cost than one that merely repeats observations or introspects confidence.

This is falsifiable in the E29-B design below. A failure would reject the engineering hypothesis; it would not imply anything broad about human cognition or brains.

10. Alternative explanations and rival mechanisms

An apparent intervention effect does not uniquely imply the intended dependency. ILT must consider at least these alternatives:

- **Task redefinition:** the perturbation changed the question, not only the evidence path.
- **Cache invalidation:** output changed because a cache was refreshed, not because the target path was causal.
- **Prompt sensitivity:** a model reacted to surface wording rather than source content.
- **Global context leakage:** the “isolated” agent received the intervention through shared conversation state, retrieval, or logs.
- **Off-target action:** another source, tool, or preprocessing stage changed.
- **Challenge awareness:** the system detected that it was being tested and altered behavior.
- **Stochastic variation:** the observed difference lies inside normal sampling variability.

- **Adjudicator dependence:** the judge shares the same root or model family as the channels being evaluated.
- **Version-space omission:** the true mechanism is not represented among candidates.
- **Equifinality:** multiple structures genuinely predict the same admissible trace.

The protocol's value is not that it makes these disappear. It forces them into the candidate set, receipt, or stated limitation.

11. Falsifiable hypotheses

H29.1 - Content-separation hypothesis

Within a correctly specified finite response-function family, an adaptive separating policy will decide the target behavioral-dependency predicate more often than passive repetition at equal total cost.

Reject if: the paired, matched-cost lower confidence bound on improvement is at or below zero, or the gain vanishes under held-out mechanisms.

H29.2 - Provenance non-identification hypothesis

When separate-root and shared-root implementations are interventionally equivalent under the admissible content interface, no content-only policy will outperform the prior on root provenance.

Reject if: a leakage-free content-only trace contains reproducible information that separates a paired provenance twin.

H29.3 - Receipt hypothesis

Requiring verified action receipts will reduce false-clean dependency judgments under unknown and off-target interventions compared with trusting intended targets.

Reject if: the preregistered false-clean rate does not fall at matched coverage and cost, or if overall decision loss increases beyond the tolerance margin.

H29.4 - Root-rotation hypothesis

Verified root rotation plus path isolation will reduce prospective shared-failure correlation and false corroboration more than content perturbation alone.

Reject if: the rotated channel retains matched-root failure correlation or does not improve dependency detection under an independently audited receipt.

H29.5 - Predicate-stopping hypothesis

Stopping when all surviving candidates share the same decision label will preserve target decision accuracy while using fewer interventions than full graph identification.

Reject if: target decision error exceeds the preregistered non-inferiority margin or total measured cost fails to decrease.

H29.6 - Misspecification hypothesis

Certificates formed without reserved out-of-version-space validation will exhibit a higher false-clean rate on omitted mechanisms than certificates with a reserved validation action.

Reject if: the reserved action produces no improvement across preregistered omitted-mechanism strata at matched cost.

12. Proposed live experiment E29-B

No E29-B results are reported in this round. The following is a preregistration-ready design.

12.1 Dataset and hidden structure

Construct 720 paired decision cases across six equally sized strata:

1. genuinely separate roots;
2. explicit shared root;
3. shared cached summary behind separate wrappers;
4. cross-channel message leakage;
5. off-target or partially applied intervention;
6. mechanism omitted from the initial candidate family.

Half of the cases should be controlled synthetic or sandbox tasks with an oracle dependency graph. The other half should be replayable, permissioned workflow traces whose roots and isolation boundaries can be independently audited. Split mechanism templates, not just surface prompts, between development and test.

12.2 Arms

Evaluate all arms on the same paired cases:

Arm	Policy
P0	Passive repeat and confidence report
P1	Round 28-style passive structural/model-adequacy audit only
P2	Random valid content probes
P3	Fixed reciprocal content probes
P4	ILT selection with candidate set, fidelity gate, and receipts
P5	ILT plus verified path isolation/root rotation when content probes are non-separating
Oracle	Uses the hidden dependency graph; evaluation ceiling, not a deployable method

12.3 Controlled total compute and cost

Each non-oracle arm receives the same envelope per case:

- at most three intervention slots;
- the same model family and decoding policy;
- 18,000 generated tokens, counted across reasoning and agent messages;
- six external source/tool transactions;
- one adjudication call;
- a fixed 120-second wall-clock ceiling;
- identical source bytes at the start of the case.

If root rotation costs more than a content probe, the cost ledger charges the additional transactions and tokens against the same envelope. Unused budget is recorded and not silently reassigned. Report both nominal slots and measured resource use.

12.4 Baselines and ablations

Preregister these P4/P5 ablations:

- remove the joint probe;
- remove the receipt and trust intended targets;
- remove path isolation;
- replace content canaries with metadata-only canaries;
- force a binary verdict instead of allowing unresolved equivalence;
- retain only the single best candidate rather than the version space;
- omit the hidden-cache and off-target families from the candidate set;
- use a judge sharing the tested agents' model family versus an independent audited judge.

12.5 Primary metrics

- **False-root-independence rate:** proportion of dependent cases certified independent.
- **Decision-relevant dependency accuracy:** correct dependent/independent/unresolved outcome under the preregistered loss matrix.
- **Dependency-detection AUROC and AUPRC:** when the protocol emits a score.
- **Selective risk and coverage:** error among non-abstained cases across coverage levels.
- **Correct-abstention rate:** unresolved output when opposite-decision candidates remain interventionally equivalent.
- **Action-fidelity violation rate:** intended actions whose receipts are uncertain or invalid.
- **Harmful revision rate:** initially correct answers made incorrect after an intervention or dependency judgment.
- **Answer utility:** task correctness, calibration, and citation/source correctness before and after the protocol.
- **Cost:** generated tokens, tool transactions, intervention count, latency, and human-review minutes.

12.6 Regression and safety metrics

Report worst-stratum results, not only averages. Specifically track:

- clean-case false alarms;
- high-stakes-case false-clean rate;
- independent-minority-source suppression;
- unnecessary root rotations;
- source exposure or privacy violations caused by probing;
- cache and state contamination across cases;
- judge-channel error correlation;
- robustness under paraphrase and reordered interventions;
- performance on a sealed omitted-mechanism challenge set.

Any arm that improves dependency detection while materially increasing harmful revision, privacy breach, or clean-case false alarms fails the safety criterion.

12.7 Analysis plan

Use paired comparisons by case and cluster uncertainty by mechanism template. Predeclare the loss weights and a non-inferiority margin for answer utility. Report exact denominators, confidence intervals, and multiplicity-adjusted secondary analyses. Do not pool natural and sandbox strata until heterogeneity is reported. Release all valid and invalid receipts so failures of execution can be distinguished from failures of inference.

13. Expected failure modes

1. **Facade completeness:** a shared-root system perfectly emulates every admissible content response.
2. **Off-target interventions:** the intended path is not the path that changes.
3. **Isolation leakage:** shared memory, retrieval, or monitoring transmits the manipulated evidence.
4. **Intervention-induced target shift:** the task after intervention is not the original decision problem.
5. **Adaptive evasion:** an agent behaves differently only when a probe is detected.
6. **Overfitted probe library:** interventions separate authored training mechanisms but fail on novel ones.
7. **Stale or hidden caches:** an old root survives a nominal replacement.
8. **Correlated adjudication:** the verifier repeats the same upstream error.
9. **Certificate laundering:** a within-family result is presented without its family boundary.
10. **Unsafe probing:** a theoretically informative action violates authorization, privacy, or operational constraints.

Mitigations include sealed canaries, randomized timing, independent telemetry, cache-state hashes, alternative judges, reserved misspecification probes, and a hard rule that unsafe actions are inadmissible rather than merely expensive.

14. Limitations

- E29-A is a deliberately small deterministic model. Its functions are monotone and binary; real systems are stochastic, stateful, strategic, and high-dimensional.
- The 50% false-independence figure comes from equal pairing of two provenance variants. It is not an estimate of production prevalence.
- Exact behavior identification relies on a complete four-point binary truth table. Real interventions may be infeasible, unsafe, noisy, or nonstationary.
- The protocol cannot guarantee that its version space contains the true mechanism.
- An intervention receipt may itself be incomplete or falsified unless telemetry and administration are independently trustworthy.
- Root rotation can create a new boundary but cannot reconstruct the past.
- The literature mostly studies formal causal graphs, biological or simulated systems, or simplified agent environments. Transfer to collaborative reasoning systems requires direct evaluation.
- No new Encevera live benchmark, language-model benchmark, production study, or Noetis answer evaluation was run.

15. Practical conclusions

1. **Do not count channels; count independently tested roots.** Two outputs are not two corroborations when they may share an origin.
2. **Match the intervention to the claim.** Content perturbations test behavioral coupling. Provenance claims require a root-addressing action or an unresolved result.
3. **Treat action fidelity as evidence.** An intended isolation without a receipt is not an isolation result.
4. **Preserve the version space.** A single favored explanation hides equivalence and misspecification risk.
5. **Stop at the decision predicate.** Complete graph recovery is unnecessary when every survivor implies the same operational action.
6. **Reserve budget for model challenge.** A clean within-family certificate needs an out-of-family stress test.
7. **Name the temporal scope.** A new independent source establishes a prospective boundary, not a historical fact.

16. Next-round handoff

Round 30 should execute E29-B in shadow mode on a pre-registered sandbox and a permissioned replay set. The priority is not a larger synthetic model. It is an end-to-end audit of whether one real evidence workflow can:

- address and change a root rather than merely a prompt;
- produce an independently verifiable intervention receipt;
- prevent leakage through shared memory and caches;
- preserve a candidate set under fidelity uncertainty;
- return unresolved equivalence without being forced into a binary verdict;
- and improve false-root-independence at matched total compute without increasing harmful revision.

If no workflow can satisfy those prerequisites, Round 30 should report that infrastructure gap rather than simulate a deployment claim.

17. Sources and inspection record

The accompanying `source_inspection_ledger.csv` records publication dates, exact URLs, material inspected, supported claims, and limitations. The entries below are the sources actually inspected for this round.

Enceavra archive

- **[A01] Enceavra Research Round 28.** “Confidence About What? Separating Target Probability, Structural Ambiguity, and Model Adequacy.” Full canonical manuscript inspected at archive commit 7e3a99e454a140d1206c308f1b9bda2daf3625ef.
- **[A02] Enceavra Research Round 27.** “Which Check Should Run Next? Model-Set-Safe Verification Under a Fixed Budget.” Full canonical manuscript inspected at the same commit.
- **[A03] Enceavra Research Round 10.** “Evidence exchange, selective revision, and a testable path toward brain-inspired intelligence.” Full route manuscript inspected at the same commit.
- **[A04] Enceavra Research Round 9.** “A research agenda for learning reusable procedures from verified experience.” Full route manuscript inspected at the same commit.
- The Research overview, rounds index, publications route, Round 26, Round 25, and Round 24 were also inspected to confirm numbering, navigation, and overlap boundaries.

Primary and authoritative literature

- **[S01] Hauser, A., & Buhlmann, P. (2012).** “Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs.” *Journal of Machine Learning Research*, 13. Full paper. Inspected: definitions, characterization theorem, examples, assumptions, and evaluation.
- **[S02] Hyttinen, A., Eberhardt, F., & Hoyer, P. O. (2013).** “Experiment Selection for Causal Discovery.” *Journal of Machine Learning Research*, 14. Full paper. Inspected: separating systems, cut-coverings, constructions, and assumptions.
- **[S03] Eberhardt, F., Glymour, C., & Scheines, R. (2005).** “On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among N Variables.” UAI 2005. Caltech repository record. Inspected: abstract, accepted-paper metadata, bounds, and stated assumptions.
- **[S04] He, Y., & Geng, Z. (2008).** “Active Learning of Causal Networks with Intervention Experiments and Optimal Designs.” *Journal of Machine Learning Research*, 9. Official article page. Inspected: abstract and batch/sequential minimax and maximum-entropy designs.
- **[S05] Mooij, J. M., Magliacane, S., & Claassen, T. (2020).** “Joint Causal Inference from Multiple Contexts.” *Journal of Machine Learning Research*, 21. Full paper. Inspected: framework,

context variables, experimentation, assumptions, and evaluation.

- **[S06] Addanki, R., Kasiviswanathan, S., McGregor, A., & Musco, C. (2020).** “Efficient Intervention Design for Causal Discovery with Latents.” ICML 2020. Official PMLR page. Inspected: abstract, objectives, latent-variable setting, and cost models.
- **[S07] Varici, B., Shanmugam, K., Sattigeri, P., & Tajer, A. (2022).** “Intervention Target Estimation in the Presence of Latent Variables.” UAI 2022. Official PMLR page. Inspected: target-estimation scope, assumptions, and evaluation.
- **[S08] Hagele, A., Rothfuss, J., Llorente, I., Heinze-Deml, C., & Krause, A. (2023).** “Ba-CaDI: Bayesian Causal Discovery with Unknown Interventions.” AISTATS 2023. Official PMLR page. Inspected: unknown-target formulation, method scope, and experiments.
- **[S09] Choo, D., Shiragur, K., Uhler, C., & Bhattacharyya, A. (2024).** “Causal Discovery under Off-Target Interventions.” AISTATS 2024. Official PMLR page. Inspected: stochastic off-target model, verification and search questions, and limitations.
- **[S10] Yang, K. D., Katcoff, A., & Uhler, C. (2024).** “Learning Unknown Intervention Targets in Structural Causal Models from Heterogeneous Data.” AISTATS 2024. Official PMLR page. Inspected: sufficient conditions, latent-confounder boundary, and candidate-set result.
- **[S11] Peters, J., Buhlmann, P., & Meinshausen, N. (2016).** “Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals.” *Journal of the Royal Statistical Society: Series B*, 78(5). Publisher record. Inspected: abstract, method summary, and publication record.
- **[S12] Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010).** “Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies.” *Epidemiology*, 21(3). PubMed record. Inspected: bibliographic record, abstract, and full-text repository result.
- **[S13] “Certifying Interventional Agreement Among Observationally Equivalent Causal Models” (2026-09-26).** Recent version-1 preprint. Full HTML. Inspected: algorithm, Boolean and continuous experiments, bypass failure, discussion, and limitations. Treated as provisional.
- **[S14] “CausalGame: Benchmarking Causal Thinking of LLM Agents in Games” (2026-07-05).** ICML 2026 oral / preprint. Full HTML. Inspected: task design, results, trajectory failures, and limitations.
- **[S15] “ACED-Bench: Evaluating How LLM Agents Acquire and Act on Causal Evidence” (2026-10-09).** Recent version-1 preprint. Full HTML. Inspected: hidden-PGM construction, intervention interface, evidence-ledger analysis, findings, and limitations. Treated as provisional.
- **[S16] Van Marcke, H., & Desender, K. (2025).** “Context-dependent Role of Confidence in Information-seeking.” PubMed record. Inspected: bibliographic record and abstract for two preregistered human experiments.
- **[S17] Gong, T., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. (2023).** “Active Causal Structure Learning in Continuous Time.” *Cognitive Psychology*, 140. Full paper. Inspected: methods, intervention/blocking task, results, computational-cost model, and discussion.

18. Reproducibility statement

The numerical claims in Section 8 are generated by the included standard-library Python program. The complete state space and all policy traces are exported; no random seed is needed. Re-run:

```
python3 round29_interventional_lineage.py --output-dir diagnostic
python3 -m unittest -v test_round29_interventional_lineage.py
```

The SHA256SUMS file binds the paper, source ledger, code, tests, and generated diagnostic artifacts. The PDF is a rendered publication copy of this manuscript. Differences caused solely by PDF metadata are not treated as changes to the analytic result.