

Encevra Research Round 28

**Confidence About What? Separating Target Probability, Structural Ambiguity,
and Model Adequacy**

Encevra Research Program

October 19, 2026

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

Contents

Abstract	2
1. Research question	3
2. Archive check, carry-forward, and novelty	3
3. Scope and exclusions	4
4. Literature map: competing positions	5
5. Proposed mechanism: the Target-Structure-Adequacy Record	7
6. A conditional validity property—and its boundary	9
7. Brain-inspired analogy, limited to a testable design idea	9
8. Experiment E28-A: exact fixed-compute trace replay	9
9. Results	12
10. Worked example: four clean-looking checks, wrong structural story	13
11. Claim-evidence table	14
12. Falsifiable hypotheses and decision criteria	15
13. Alternative explanations	16
14. Expected failure modes and mitigations	17
15. Limitations	17
16. Operational recommendations	18
17. Next-round handoff	18
18. Full sources and inspection notes	19
Reproducibility statement	20

A fixed-compute trace replay of when structured uncertainty helps—and when it cannot

Research date: 19 October 2026

Program: Encevera Research — reliable collaboration, evidence exchange, and answer revision

Status: Research synthesis plus exact constructed diagnostic; not a production or language-model benchmark

Archive position: Round 28, following Round 27, Which Check Should Run Next? Model-Set-Safe Verification Under a Fixed Budget

Abstract

An agent can say “90% confident” while leaving three different questions unresolved: 90% under which dependency model, how much do plausible models disagree, and did the evidence look possible under any declared model? This round asks whether collaborative AI should replace a lone confidence scalar with a structured record that separates **target probability**, **structural ambiguity**, and **model adequacy**.

The literature gives reasons both for and against that move. Point calibration and verbalized confidence can be useful on measured distributions; ensembles, model marginalization, and probabilistic classifier sets can expose some epistemic uncertainty; risk-coverage evaluation can make abstention operational. Yet calibration degrades under shift, pooled calibration can conceal subgroup failures, and no uncertainty method can certify that an omitted causal structure has been represented.

Round 28 proposes the **Target-Structure-Adequacy Record (TSAR)**. For an evidence history, TSAR stores a pooled target probability, the per-dependency-model probabilities and their interval, and adequacy diagnostics including model compatibility and predictive surprise. It preserves evidence identifiers and the model-set version so that another agent can audit what the numbers mean.

We then ran E28-A, an exact secondary analysis of Round 27’s frozen acquisition traces. It replays 9,744 policy-world histories across two cost regimes, three actual dependency variants, four budgets, seven policies, and 58 worlds per variant. Three evaluation distributions yield 29,232 weighted rows. No policy, sensor, observation, cost, or evidence budget was changed. This isolates uncertainty representation from verification allocation.

The result is deliberately mixed. When the actual dependency variant belongs to the declared model set and the priors and likelihoods are correct, the model interval contains the exact history-conditional target probability in 100% of weighted histories, by construction. Under a hidden-common-origin variant omitted from the model set, interval coverage falls to 60.57–72.08%, depending on the prior. Every hidden-common-origin history remains compatible with the declared model set, while predictive surprise and interval width are near chance at detecting the omission (AUROC 0.488–0.577 across tested scores and distributions). A conservative structured revision score does not systematically dominate a scalar probability at matched coverage: most policy cells tie; each representation wins some cells.

The correct conclusion is therefore narrower than “use intervals.” TSAR is a better contract: it distinguishes uncertainty about the target from disagreement about structure and exposes the assumptions behind a confidence claim. It is not a detector of every missing cause. A model set can be internally well calibrated and jointly wrong. The next round should test active interventions that can reveal missing dependencies rather than merely summarize declared ones.

No new Enceavra production benchmark, live multi-agent benchmark, or language-model benchmark was run. E28-A is a deterministic, finite, reproducible diagnostic with authored priors and a deliberately out-of-model stress case.

1. Research question

Under a fixed evidence and compute budget, does a three-part uncertainty record—target probability, disagreement across plausible dependency models, and model-adequacy status—predict delayed corrections and support safer abstention better than a single confidence scalar?

This question has four falsifiable parts:

1. **Conditional validity:** if the true dependency model is in the declared set, does the set of model-conditioned probabilities contain the true history-conditional probability?
2. **Revision prediction:** does the structured record rank later corrections better than one pooled probability when both see exactly the same history?
3. **Shift detection:** do compatibility, predictive surprise, or interval width reveal a hidden dependency omitted from the model set?
4. **Operational value:** at matched coverage, does routing or abstaining with the structured record reduce realized correction risk without merely rejecting more work?

Short answer

- **Yes, conditionally:** the model interval has an exact inclusion property when the true model, prior, and likelihood are represented.
 - **Not generally, in this diagnostic:** the structured revision score did not systematically beat the scalar score at matched coverage.
 - **No, for the tested hidden alias:** the adequacy fields did not reliably detect a hidden common origin because its histories were observationally compatible with declared worlds.
 - **Therefore:** retain the structured record for auditability and conditional guarantees, but do not treat it as a certificate of model completeness. Pair it with interventions designed to expose hidden dependencies.
-

2. Archive check, carry-forward, and novelty

2.1 Archive check

The Enceavra archive was inspected before selecting this question. The research overview, rounds index, publications registry, full texts of Rounds 9 and 10, and the relevant later sequence through Round 27 were reviewed. The latest completed entry was Round 27, so the next sequential number is **Round 28**, not Round 11.

2.2 Carry-forward from prior rounds

Prior result	What carries forward	What Round 28 does differently
Round 9: learn reusable procedures only from verified experience	Confidence policy should be revised from adjudicated outcomes, not persuasive traces	No policy learning is performed here; delayed truth is used only to evaluate representations
Round 10: exchange checkable evidence and revise selectively	Confidence must remain attached to evidence and revision history	TSAR binds probability fields to the exact evidence history and dependency-model version
Round 25: silence and selective omission can be observationally ambiguous	Missing evidence is not negative evidence	Adequacy is kept separate from target probability; no-evidence histories are not silently converted into confidence
Round 26: observational equivalence limits what any decision rule can identify	Opposite-target worlds with identical observations cannot be separated without new evidence	Round 28 tests whether uncertainty metadata exposes residual ambiguity; it cannot break equivalence by itself
Round 27: allocate verification under a budget and require model-set-safe adjudication	Check choice and stopping must respect dependency uncertainty	Round 28 freezes those acquisition histories and studies representation only, removing allocation as a confound

2.3 Why this is not a repetition of Round 10 or Round 27

Round 10 asked how agents should exchange evidence and revise answers. Round 27 asked which check should run next and when a model-set-safe categorical verdict was possible. Round 28 asks what an uncertainty report should mean **after a particular evidence history already exists**. It neither introduces a new debate protocol nor optimizes a new check policy. Every representation is evaluated on the same evidence.

This separation is consequential. If a “better uncertainty method” is allowed to buy different evidence, use a stronger verifier, or spend more tokens, any improvement is uninterpretable. E28-A holds those factors fixed.

3. Scope and exclusions

In scope

- Binary target propositions after a finite collaborative evidence history.
- Alternative causal/dependency models for nominally distinct evidence channels.
- Calibrated point probabilities, model-conditioned probability sets, abstention, delayed correction, and model-adequacy diagnostics.
- Exact finite replay where the true conditional frequencies can be computed rather than estimated from samples.
- A concrete schema that another agent can inspect and act on.

Out of scope

- Claiming that a language model’s stated probability is intrinsically calibrated.

- Treating self-consistency, token likelihood, verbal confidence, or semantic entropy as proof of factual correctness.
- Inferring literal neural mechanisms from human confidence behavior.
- Solving causal discovery from passive evidence alone.
- Training a calibrator or controller on the E28-A outcomes.
- Production safety certification, legal assurance, or general guarantees outside the declared finite model.

4. Literature map: competing positions

The literature does not support a single slogan such as “confidence works” or “confidence is useless.” It supports a layered view.

Position	Primary evidence inspected	What it supports	What remains unresolved for Encevera
A. A calibrated scalar can be decision-useful.	Guo et al. found modern networks miscalibrated but temperature scaling effective on several reported in-distribution benchmarks [S1]. Tian et al. found verbalized confidences often better calibrated than conditional probabilities on three QA benchmarks [S10].	A scalar probability can be improved and can support thresholding on a measured distribution.	A scalar does not say which dependency assumptions produced it, and benchmark calibration need not survive source or task shift.
B. Model diversity can expose epistemic uncertainty.	Deep ensembles produced strong uncertainty estimates on the studied tasks [S2]. Ovadia et al. found methods that marginalize over models comparatively strong under several shifts [S5]. Mortier et al. formalized calibration for sets of probabilistic classifiers [S9].	Disagreement across models can contain information erased by averaging. A probability set is a legitimate object of calibration.	Model diversity may be cosmetic if members share data, prompts, tools, or provenance. A set cannot cover a model it omits.
C. Calibration must be tested carefully and conditionally.	Kumar et al. showed common calibration estimates and post-hoc procedures can be misleading [S6]. Widmann et al. showed top-class calibration can miss multiclass defects [S7]. Multicalibration requires validity across identifiable subgroups [S3]. Hartline et al. showed calibration measures themselves can create strategic reporting incentives and proposed a truthful alternative [S15].	Pooled ECE alone is inadequate; use proper scores, exact conditional groups where available, and subgroup or set diagnostics.	No finite list of audited groups proves validity on an unrepresented causal regime.

Position	Primary evidence inspected	What it supports	What remains unresolved for Enceavra
D. Selective prediction turns uncertainty into an operational decision.	SelectiveNet evaluates risk as coverage changes [S8]. Wang and Holmes derive task-specific epistemic deferral measures for natural-language generation [S14].	Uncertainty should be judged by what happens when systems defer, verify, or answer—not only by a calibration plot.	A reject rule can appear safer by rejecting different or easier cases; matched coverage and subgroup regressions are necessary.
E. Shift and fluent error remain hard.	Ovadia et al. found calibration generally degrades under shift [S5]. Xiong et al. found no black-box elicitation method consistently dominant and persistent difficulty on challenging tasks [S11]. Semantic entropy detects some confabulations [S12], while Zhou et al. document apparently sensible wrong answers and residual prompt instability [S13].	“Knows what it does not know” must be tested under shift and on confidently wrong cases.	Consistent false premises and hidden common causes can yield low disagreement, low entropy, and high confidence.

Established findings

The following are well supported within the cited studies’ domains:

1. Accuracy and calibration are different properties [S1, S5].
2. Calibration commonly worsens under distribution shift [S5].
3. Proper scoring rules, calibration diagnostics, and risk-coverage answer different questions; no single ECE number is sufficient [S5–S8, S15].
4. Ensembles or model sets can represent disagreement that a mean probability hides [S2, S9].
5. Language-model uncertainty elicitation is task- and method-dependent; verbalized confidence can help on some benchmarks but remains imperfect [S10, S11, S14].

Mixed evidence

- Model averaging and ensembles are often strong, but not uniformly best and not immune to shared misspecification [S2, S5, S9].
- Verbal confidence sometimes calibrates better than token probabilities, while other studies find overconfidence and no consistently dominant elicitation method [S10, S11].
- Semantic uncertainty can detect confabulation, but it targets one failure class and does not establish source independence [S12].

Open questions

- How should collaborating agents report uncertainty when evidence channels may share hidden causes?
- Which adequacy signals detect an omitted dependency rather than merely a rare history?
- Can a routing policy improve safety at equal coverage and equal evidence cost without over-abstaining on one subgroup?
- How should calibration transfer across model, tool, prompt, and provenance changes?

Enceavra hypotheses, not established findings

- A structured uncertainty contract is more auditable than a lone number.
- The useful distinction is not only aleatoric versus epistemic uncertainty, but also **target uncertainty**, **structural ambiguity**, and **model inadequacy**.
- Passive uncertainty summaries will fail on some hidden aliases; active interventions are required to expose them.

5. Proposed mechanism: the Target-Structure-Adequacy Record

Let $Y \in \{0, 1\}$ be the target, h the complete evidence history, and $\mathcal{M}$ a versioned set of declared dependency models. TSAR contains three sections.

5.1 T — Target probability

For each model m , compute

$$p_m = P(Y = 1 \mid h, m).$$

The pooled probability is

$$p_{\text{pool}} = \sum_{m \in \mathcal{M}} P(m \mid h) p_m.$$

This is the best point summary under the declared mixture. It must be labeled as conditional on the model-set version and prior; it is not an unconditional truth probability.

5.2 S — Structural ambiguity

Preserve the vector $\{p_m\}$, not only its mean. For a binary target, a compact envelope is

$$I(h) = \left[\min_{m \in \mathcal{M}_h} p_m, \max_{m \in \mathcal{M}_h} p_m \right],$$

where $\mathcal{M}_h$ is the subset of declared models compatible with the history. Its width

$$w(h) = \max p_m - \min p_m$$

is a disagreement descriptor, not a guarantee that the true process lies inside.

5.3 A — Model adequacy

Record at least:

- declared model-set identifier and version;
- number and identities of compatible models;
- evidence mass $P(h \mid \mathcal{M})$;
- predictive surprise $s(h) = -\log_2 P(h \mid \mathcal{M})$;
- any explicit out-of-support or schema-violation flag;
- provenance coverage: which evidence roots are represented and which remain assumed;
- shift-monitor status and last adjudication date.

Surprise is an alarm, not proof of misspecification. A rare valid history can be surprising, while an omitted model can generate an ordinary-looking history.

5.4 Decision and revision risk

For the diagnostic, the provisional decision is

$$d = \mathbb{1}[p_{\text{pool}} \geq 0.5].$$

The scalar predicted error risk is

$$r_{\text{scalar}} = P(Y \neq d \mid h, \mathcal{M}) = \min(p_{\text{pool}}, 1 - p_{\text{pool}}).$$

The model-set risk is

$$r_{\text{set}} = \max_{m \in \mathcal{M}_h} P(Y \neq d \mid h, m).$$

If any declared model is incompatible, the diagnostic's conservative routing score is set to one. That convention is an authored operational choice, not a theorem.

5.5 Minimal interoperable schema

```
{
  "claim_id": "stable claim identifier",
  "evidence_history_hash": "sha256:...",
  "target": {
    "decision": "CHANGE | NO_CHANGE | ABSTAIN",
    "p_change_pooled": 0.218,
    "loss_function": "0-1",
    "prior_id": "issuer_shift-v1"
  },
  "structure": {
    "model_set_id": "dependency-set-v2",
    "per_model_p_change": {
      "independent_witness": 0.000,
      "correlated_witness": 0.358
    },
    "interval": [0.000, 0.358],
    "width": 0.358
  },
  "adequacy": {
    "compatible_models": 2,
    "declared_models": 2,
    "surprise_bits": 2.471,
    "out_of_support": false,
    "provenance_gaps": ["external/witness independence not directly tested"]
  },
  "revision": {
    "scalar_risk": 0.218,
    "model_set_risk": 0.358,
    "recheck_trigger": "dependency intervention if consequence tier >= high"
  }
}
```

```
}
}
```

The record should travel with a claim. Agents may summarize it in natural language, but the machine-readable fields and evidence-history hash are authoritative.

6. A conditional validity property—and its boundary

Proposition

If the actual data-generating dependency model m^* is in $\mathcal{M}$, and the model-specific prior and likelihood are correct, then the true conditional probability for history h is one member of the reported set:

$$P(Y = 1 \mid h, m^*) = p_{m^*} \in I(h).$$

Why it holds

The interval is defined as the minimum and maximum of all compatible model-conditioned probabilities. If m^* is compatible and correctly specified, its probability is included, so it lies between the endpoints.

What it does not prove

- It does not show that the true model is in the set.
- It does not show that the prior is correct.
- It does not make the pooled probability calibrated under a different model mixture.
- It does not turn dependent evidence into independent evidence.
- It does not identify an observationally equivalent hidden cause.

Calling this a “guarantee” without its premises would be misleading. E28-A deliberately violates the first premise with `hidden_common_origin`.

7. Brain-inspired analogy, limited to a testable design idea

Human studies report that subjective confidence can influence whether people seek more information even after controlling for accuracy [S4]. This suggests a functional design hypothesis: a collaborative AI controller could use a distinct meta-level uncertainty state to decide whether to answer, request evidence, or defer.

The analogy stops there. TSAR is not a model of a brain region, neural population, consciousness, or human metacognition. It is a software record. Its value is testable through routing outcomes.

Testable design idea: compare an answer controller driven only by pooled target confidence with one driven by target confidence plus structural ambiguity and adequacy fields. Match evidence cost and coverage. Measure correction risk, subgroup regressions, and harmful false-clean decisions. E28-A performs the first exact trace replay of that comparison; later rounds should test live agents and active interventions.

8. Experiment E28-A: exact fixed-compute trace replay

8.1 Purpose

E28-A tests uncertainty representation without allowing a structured method to obtain better evidence. It is a secondary analysis of Round 27’s deterministic policy traces.

8.2 Frozen factors

- **Sensors:** receipt, checkpoint, heartbeat, transport acknowledgement, verifier canary, external commitment, behavior probe, independent witness.
- **Root causes:** issuer, transport, observer, verifier, external, probe, witness.
- **Fault sets:** all combinations of zero, one, or two roots.
- **Target states:** change or quiet.
- **Actual variants:** independent witness, correlated witness, hidden common origin.
- **Cost regimes:** Round 26 costs and cheap-declared-witness stress costs.
- **Budgets:** 3, 5, 7, and 9 authored cost units.
- **Policies:** seven Round 27 fixed, greedy, minimax, and dynamic-programming policies.
- **Observation semantics:** deterministic and unchanged from Round 27.

This produces

$$2 \text{ cost regimes} \times 3 \text{ variants} \times 4 \text{ budgets} \times 7 \text{ policies} \times 58 \text{ worlds} = 9,744$$

unique policy-world traces. Uniform, benign, and issuer-shift priors are applied separately, producing 29,232 weighted evaluation rows and 3,726 variant/prior-specific history rows.

8.3 Declared and held-out structures

The declared set contains:

1. `independent_witness`;
2. `correlated_witness`, where the nominal witness can share the issuer failure root.

The held-out stress is `hidden_common_origin`, where both the witness and the nominally external commitment secretly share the issuer root. It is not available to the uncertainty calculator.

8.4 Compared representations

Representation	Inputs	Output used for routing
Single probability	Same frozen history; equal-weight mixture over declared dependency variants	r_{scalar}
TSAR without active intervention	Same frozen history; per-variant posteriors, interval, compatibility, surprise	Conservative r_{set} ; one if a declared model is incompatible

The comparison is paired at the history level. No representation receives another sensor observation, token, verifier, or larger budget.

8.5 Evaluation metrics

Target probability

- Brier score and log loss.

- Exact history-conditional calibration error, computed from the full finite world set rather than estimated bins.
- Worst history-conditional error.

Structural ambiguity

- Weighted coverage of the true history-conditional probability by the model interval.
- Mean interval width and miss distance.

Revision prediction and selective answering

- AUROC and average precision for whether delayed truth reverses the provisional decision.
- Risk at 25%, 50%, 75%, 90%, and 100% coverage with fractional random tie-breaking.
- Brier score for predicted revision risk.

Regression and harm checks

- False-clean rate conditional on an actual change.
- High-confidence error and high-confidence false-clean rates.
- Results stratified by actual variant, prior, cost regime, budget, and policy.
- Paired cell wins, losses, and ties at matched coverage.

Model adequacy

- Declared-model incompatibility rate.
- AUROC and average precision for detecting the held-out variant from surprise, interval width, structured risk, and incompatibility.

8.6 Baselines and ablations

- **Pooled point baseline:** removes structure and adequacy fields.
- **Interval-only ablation:** keeps model disagreement but ignores compatibility and surprise.
- **Adequacy-only ablation:** keeps compatibility and surprise but collapses per-model target probabilities.
- **Single-model nominal view:** records only the independent-witness posterior; used descriptively because some histories are incompatible with it.
- **No active intervention:** essential control showing what passive summaries can and cannot infer.

8.7 Reproducibility and checks

The program enumerates all worlds exactly; it does not use Monte Carlo sampling. Five executable invariants passed:

1. exactly 29,232 weighted rows are generated;
2. all probabilities and intervals lie in $[0, 1]$;
3. structured risk is never below pooled scalar risk under the declared definition;
4. evaluation weights normalize within every actual-variant/prior cell;
5. the true conditional is inside the interval for every declared-variant history.

The bundle contains source, tests, exact CSV outputs, JSON summary, source-inspection ledger, and SHA-256 manifests.

9. Results

9.1 Result 1: exact conditional coverage inside the declared model set

For both declared actual variants—independent witness and correlated witness—the model interval covered the exact history-conditional target probability in **100.00%** of weighted histories under all three priors.

This is the expected consequence of the proposition, not an empirical discovery about real systems. It confirms that the implementation preserves the intended conditional property.

The pooled point probability was not exactly equal to either component’s true conditional. Its weighted mean absolute history-calibration error ranged from **0.0029** under the benign prior to **0.0253** under issuer shift. Averaging across models trades component-conditional exactness for one mixture probability.

9.2 Result 2: the interval breaks under an omitted hidden common origin

Evaluation distribution	Point mean absolute history error	Model-interval coverage	Mean interval miss distance	Worst history error
Uniform	0.0466	72.08%	0.0370	0.2232
Benign	0.0136	60.57%	0.0107	0.0445
Issuer shift	0.1662	65.43%	0.1358	0.6171

The benign prior has the smallest average point error but the lowest interval coverage. This is a warning against treating one favorable aggregate score as evidence that structural uncertainty is valid.

9.3 Result 3: passive adequacy signals do not identify the hidden alias

Every hidden-common-origin history remained compatible with the two declared models: the hard incompatibility rate was **0%**. Detection scores were close to random.

Adequacy score	Uniform AUROC	Benign AUROC	Issuer-shift AUROC
Predictive surprise	0.488	0.493	0.529
Dependency interval width	0.511	0.503	0.547
Structured revision risk	0.514	0.503	0.577
Declared incompatibility	0.499	0.499	0.495

With an out-of-model prevalence of one third in this detection analysis, average precision also stayed near prevalence for most cells. The strongest AUROC, 0.577, is too weak to support a useful general detector claim.

This is not a failure of threshold tuning. The held-out process produces histories that declared worlds can also produce. Without an intervention that changes those equivalence classes, passive compatibility cannot expose the hidden origin.

9.4 Result 4: conservative structured risk does not systematically dominate the scalar

Across 504 matched policy/prior/budget/coverage cells per actual variant:

Actual variant	TSAR wins	Scalar wins	Ties
Independent witness	6	61	437

Actual variant	TSAR wins	Scalar wins	Ties
Correlated witness	15	9	480
Hidden common origin	15	9	480

Most cells tie because the declared models often agree. Under the independent-witness world, the conservative set score can over-route cases due to possibilities introduced by the correlated model. Under the correlated and hidden variants it sometimes improves ordering, but not consistently enough for a dominance claim.

Selected issuer-shift results at **75% matched coverage** make the trade-off concrete:

Actual variant	Scalar realized correction risk	TSAR realized correction risk	Difference (TSAR – scalar)
Independent witness	6.42%	7.54%	+1.12 pp
Correlated witness	11.62%	11.59%	–0.03 pp
Hidden common origin	32.32%	31.65%	–0.67 pp

The hidden variant remains unsafe under both representations. At 50% coverage under issuer shift, realized correction risk is **30.43%** for the scalar and **29.32%** for TSAR, even though their mean predicted risks are only **7.02%** and **8.82%**. Conservatism inside the wrong model set does not bridge model misspecification.

9.5 Result 5: high confidence remains vulnerable under shift

For hidden common origin:

- Under the benign prior, 88.37% of weighted cases qualify as at least 90% confident, but the conditional error rate among them is **3.37%**.
- Under issuer shift, 30.40% qualify as at least 90% confident and their conditional error rate is **10.16%**.
- The false-clean rate given a real change is **31.13%** under issuer shift.

These are properties of the authored finite diagnostic, not estimates of production frequency. They demonstrate the logical possibility that a system is confidently wrong because all nominally independent checks share an unmodeled origin.

9.6 What the results support

1. A model-conditioned probability set carries a useful conditional validity statement that a pooled scalar cannot express.
2. Reporting that set is valuable for audit and downstream decision rules.
3. The set's usefulness depends on model-set coverage.
4. Surprise, disagreement, and compatibility are distinct diagnostics; none is a universal shift detector.
5. A structured record should be adopted as an interface contract, not advertised as a standalone safety solution.

10. Worked example: four clean-looking checks, wrong structural story

Consider the cheap-declared-witness regime, budget 9, issuer-shift prior, and Round 27's `safe_uniform_dp` policy. The actual world has a real change and an issuer fault under the held-out hidden-common-origin variant.

The frozen policy selects:

1. verifier canary → ok;
2. independent witness → clean;
3. checkpoint → clean;
4. external commitment → clean.

The evidence looks redundant and reassuring. The declared model set produces:

- pooled $P(\text{change}) = 0.218$;
- independent-model probability 0.000;
- correlated-witness-model probability 0.358;
- interval $[0.000, 0.358]$;
- predictive surprise 2.471 bits;
- both declared models compatible;
- provisional decision NO_CHANGE.

Under the held-out hidden-common-origin variant, the exact conditional probability of change for that same history is **0.666**. It lies outside the declared interval by **0.308**. The particular issuer-fault world is therefore falsely cleared.

The failure is structural. In the held-out variant, the witness and external commitment both share the issuer failure root. The record correctly says the two declared models disagree, but it cannot invent the missing third model. Surprise is modest because the clean history is not rare under declared worlds.

What should an operational controller do? A high-consequence controller should read the provenance gap—“external/witness independence not directly tested”—and purchase an intervention that changes the dependency graph, such as a source-isolation test, delayed commitment from a separately administered authority, or controlled evidence withholding. Rephrasing the question or asking another agent that consumes the same sources is not such an intervention.

11. Claim-evidence table

Claim	Status	Evidence	Boundary / possible falsifier
Modern model confidence can be miscalibrated, and calibration can degrade under shift	Established within cited benchmarks	S1, S5	Different models/tasks may behave differently; requires fresh measurement
Model marginalization and ensembles can improve uncertainty quality on some shifts	Established but not universal	S2, S5	Shared training or provenance can preserve common-mode error
Probability sets are legitimate uncertainty objects whose calibration can be tested	Established formal/empirical research direction	S9	Does not imply a chosen set contains the real process
Pooled calibration alone can hide subgroup defects	Established formal motivation	S3, S6, S7, S15	Depends on which groups and metrics are audited
Risk-coverage is necessary for operational evaluation of abstention	Established evaluation principle	S8, S14	Coverage must be matched and subgroup harms reported

Claim	Status	Evidence	Boundary / possible falsifier
TSAR's interval contains the true conditional when the correct declared component and prior/likelihood are present	Exact within E28-A	Proposition plus 100% declared-variant coverage in E28-A	Fails if the model, prior, likelihood, or implementation is wrong
TSAR's passive adequacy fields detect a hidden common origin	Not supported in E28-A	AUROC 0.488–0.577; 0% hard detection on held-out histories	Could improve for non-aliased shifts or with active interventions
TSAR always improves correction ranking over a scalar	Rejected as a universal claim in E28-A	Mixed wins/losses; most cells tie	Could hold under a narrower deployment distribution and policy, which must be preregistered
Hidden common causes can make multiple clean checks jointly misleading	Exact constructed counterexample, consistent with prior identifiability results	Worked trace; Round 26 equivalence argument	Does not estimate how often this occurs in production
A meta-level uncertainty state should control information seeking	Encevera hypothesis with limited human analogy	S4 and proposed controller test	Requires matched-compute agent experiments; no neural equivalence is claimed

12. Falsifiable hypotheses and decision criteria

H28.1 — In-model set validity

Hypothesis: For every evaluated history generated by a declared model with matched prior and likelihood, the exact conditional target probability lies inside the reported model interval.

Criterion: 100% weighted history coverage, with any miss treated as an implementation or premise failure.

E28-A: Supported exactly.

H28.2 — Structured revision ranking

Hypothesis: At matched coverage, TSAR's conservative set risk yields lower delayed-correction risk than pooled scalar risk across both declared variants and all preregistered priors.

Criterion: lower risk in a clear majority of cells, no meaningful harmful subgroup regression, and improved risk Brier score.

E28-A: Not supported. The independent variant favors the scalar in more non-tied cells; the correlated variant slightly favors TSAR; most cells tie.

H28.3 — Passive adequacy detection

Hypothesis: Surprise, interval width, or compatibility detects the held-out hidden-common-origin variant with AUROC at least 0.75 and useful precision above prevalence.

E28-A: Falsified for this stress. Maximum AUROC is 0.577; compatibility itself is uninformative.

H28.4 — Active dependency intervention

Hypothesis: Replacing one passive repeat check with a provenance-changing intervention will raise held-out model detection above AUROC 0.75 and reduce false-clean rate at equal total cost.

Status: Untested; recommended for Round 29.

H28.5 — Live-agent transfer

Hypothesis: On a preregistered multi-agent benchmark with delayed adjudication, an interoperable TSAR schema will improve correction-risk Brier score and 90%-coverage risk over a one-number confidence message, without reducing task accuracy or increasing total tokens/tool cost.

Status: Untested. E28-A does not justify claiming transfer.

13. Alternative explanations

A. The model set is too small

True. The point of the hidden-common-origin stress is to expose that dependency. A larger set containing the held-out variant would restore the interval property for this case, but “add every hidden cause” is not operationally complete. The important question becomes how new causes are proposed and tested.

B. The equal model mixture is poorly chosen

Also possible. Different model priors change the pooled probability and routing order. However, the interval failure under omission is not repaired by reweighting the two wrong components. The exact outputs include three world priors so this sensitivity remains visible.

C. The conservative score is a poor way to use the record

Likely in some cells. TSAR is a record, not one mandated decision rule. A learned controller might use width, consequence, and cost more efficiently. But it must be trained and evaluated on adjudicated shifts without leaking test outcomes, and it still cannot infer an observationally aliased cause from no distinguishing evidence.

D. The deterministic model is too simple

Yes. Determinism enables exact conditionals and clear counterexamples but omits noisy sensors, continuous scores, adaptive adversaries, and language ambiguity. The result establishes a boundary in a small world; it does not quantify a deployment.

E. Surprise should be normalized differently

Possible. Raw predictive surprise favors histories that are rare under the declared mixture, not necessarily structurally invalid. Normalized or learned shift scores could improve on other shifts. They cannot guarantee detection when held-out and in-model processes induce the same history distribution.

F. Policies induce selection bias

Adaptive acquisition changes which histories appear. E28-A addresses this by computing exact conditionals separately within every frozen policy/budget/cost cell and by using paired histories. A deploy-

ment study should additionally log the acquisition policy and avoid calibrating on selectively observed outcomes without correction.

14. Expected failure modes and mitigations

Failure mode	Why it matters	Detection metric	Mitigation
False precision	A scalar hides dependency assumptions	Per-model vector and interval width	Require model-set and prior identifiers with every probability
Cosmetic diversity	Agents look independent but share prompts, retrieval, or sources	Provenance-root overlap, not agent count	Root-disjoint evidence and source lineage checks
Hidden common origin	Multiple clean checks fail together	Active provenance intervention; delayed adjudication	Isolate sources, rotate authorities, test counterfactual withholding
Pooled calibration hides subgroup harm	Aggregate score looks good while one regime fails	Variant/tool/domain calibration, worst-group error	Multicalibration-style subgroup audit and regression gates
Adequacy alarm overfires	Rare valid histories cause unnecessary deferral	False-alarm rate and utility at matched coverage	Consequence-aware thresholds and explicit routing cost
Adequacy alarm underfires	Aliased shift looks ordinary	Held-out dependency stresses, not only random corruption	Maintain adversarial structural tests and intervention budget
Confidence laundering	Natural-language summary drops caveats	Schema validation and evidence-history hash	Machine-readable record is authoritative; summaries must link it
Stale calibration	Model, prompt, tool, or data changes	Versioned calibration dashboards and drift tests	Invalidate calibration on material dependency changes
Strategic confidence reports	Agents optimize appearing calibrated	Proper/truthful scoring and sealed adjudication	Score after independent outcomes; audit revisions and abstentions
Over-abstention	Safety metric improves by refusing too much	Matched risk-coverage and task utility	Coverage floors and subgroup service constraints

15. Limitations

1. **Constructed finite world.** Costs, priors, failure roots, sensors, and observation semantics are authored.
2. **No language generation.** The diagnostic does not measure how well current models produce or consume TSAR records.
3. **No estimated confidence elicitation.** Probabilities are exact posteriors inside the finite model, not model self-reports.
4. **No live adversary.** The hidden variant is fixed rather than adaptively chosen.
5. **Binary target.** Multiclass and open-ended generation require richer sets and task-specific utility.
6. **No training.** The round evaluates a fixed conservative rule, not an optimized uncertainty controller.
7. **Prior sensitivity.** Three priors are tested, but they do not span deployment uncertainty.

8. **Model-set sensitivity.** Adding or deleting a dependency model changes the interval and pooled probability.
9. **Arithmetic overhead not priced.** Evidence acquisition is exactly matched, while the small local cost of computing extra metadata is not included. A deployment benchmark should count it.
10. **No causal discovery from passive traces.** This limitation is substantive, not merely experimental.

16. Operational recommendations

1. **Do not transmit confidence without its object.** Specify target, loss, horizon, evidence hash, model-set version, and prior.
2. **Preserve per-model probabilities.** A pooled number is useful, but averaging should be reversible.
3. **Separate ambiguity from inadequacy.** Wide disagreement and out-of-support evidence require different responses.
4. **Route by consequence as well as uncertainty.** A modest model interval can still warrant intervention when a false clean is costly.
5. **Audit exact or subgroup-conditional calibration.** Report proper scores, worst groups, high-confidence errors, and risk-coverage.
6. **Treat agent count as irrelevant without provenance.** Independence is a property of evidence roots, not personas.
7. **Reserve budget for dependency-changing checks.** Repeated passive checks cannot resolve every equivalence class.
8. **Log delayed corrections.** A confidence system without sealed adjudication cannot learn whether its uncertainty reports were useful.
9. **Fail closed on schema loss.** If an intermediary drops model-set or adequacy fields, downgrade the claim rather than preserving the original confidence number.
10. **Never advertise the interval as covering “the truth” unconditionally.** Its correct label is “range across declared models.”

17. Next-round handoff

Recommended Round 29 question

Can a fixed-cost active intervention discover a missing evidence dependency that passive confidence, disagreement, and surprise cannot detect?

Proposed experiment

Hold total cost equal and compare:

1. repeated passive verification;
2. scalar-confidence routing;
3. TSAR routing without intervention;
4. TSAR plus one provenance-changing intervention;
5. an oracle that knows the hidden dependency, as an upper bound.

Candidate interventions include source isolation, controlled evidence withholding, authority rotation, delayed re-observation, and challenge queries whose outcomes differ under shared versus independent origins. Preregister the hidden-dependency families and leave at least one family entirely held out.

Primary endpoints should be hidden-model detection AUROC, false-clean rate under change, correction risk at matched 75% and 90% coverage, total evidence cost, and worst-subgroup utility. The intervention hypothesis should be rejected if gains disappear at equal cost, if they rely on a leaked model identifier, or if false alarms erase the benefit.

The conceptual shift is important: Round 28 shows how to **represent** uncertainty about declared structures. Round 29 should test how to **change the evidence** so missing structures become identifiable.

18. Full sources and inspection notes

All links were checked on 19 October 2026. The reproducibility bundle includes a machine-readable source-inspection ledger recording publication date, inspected material, supported claim, and limitation.

[S1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” ICML 2017, PMLR 70. Official proceedings page. Inspected: abstract, metadata, citation record, and linked paper record. Supports the in-distribution calibration discussion and temperature-scaling baseline.

[S2] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.” NeurIPS 2017. Official proceedings page. Inspected: abstract, metadata, and paper link. Supports model diversity and out-of-distribution uncertainty as a competing position.

[S3] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. “Multicalibration: Calibration for the (Computationally-Identifiable) Masses.” ICML 2018, PMLR 80. Official proceedings page. Inspected: abstract, metadata, and citation record. Supports subgroup-conditional calibration rather than pooled averages alone.

[S4] Kobe Desender, Annika Boldt, and Nick Yeung. “Subjective Confidence Predicts Information Seeking in Decision Making.” Psychological Science 29(5), 2018. DOI record. Inspected: PubMed/publisher-indexed bibliographic record and abstract metadata. Used only for a functional, testable information-seeking analogy.

[S5] Yaniv Ovadia et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” NeurIPS 2019. Official proceedings PDF. Inspected: full 12-page paper, including setup, metrics, shift experiments, and conclusions. Supports the shift boundary and the comparative strength—but not infallibility—of model marginalization.

[S6] Ananya Kumar, Percy Liang, and Tengyu Ma. “Verified Uncertainty Calibration.” NeurIPS 2019. Official proceedings page. Inspected: abstract, metadata, and linked paper record. Supports caution about calibration-error estimation and post-hoc calibration claims.

[S7] David Widmann, Fredrik Lindsten, and Dave Zachariah. “Calibration tests in multi-class classification: A unifying framework.” NeurIPS 2019. Official proceedings page. Inspected: abstract, metadata, and linked paper record. Supports multidimensional and statistically interpretable calibration testing.

[S8] Yonatan Geifman and Ran El-Yaniv. “SelectiveNet: A Deep Neural Network with an Integrated Reject Option.” ICML 2019, PMLR 97. Official proceedings page. Inspected: abstract, metadata, and citation record. Supports risk-coverage as an operational evaluation.

[S9] Thomas Mortier, Viktor Bengs, Eyke Hüllermeier, Stijn Luca, and Willem Waegeman. “On the Calibration of Probabilistic Classifier Sets.” AISTATS 2023, PMLR 206. Official proceedings PDF. Inspected: full 14-page paper, especially its set-calibration definition, test, and experiments. Supports preserving a probability set for epistemic uncertainty.

[S10] Katherine Tian et al. “Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback.” EMNLP 2023. ACL Anthology. Inspected: official abstract, bibliographic metadata, and linked PDF record. Supports the benchmark-conditional usefulness of verbalized confidence.

[S11] Miao Xiong et al. “Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.” ICLR 2024. Official proceedings page. Inspected: abstract, metadata, and paper link. Supports the finding that elicited confidence remains imperfect and method-dependent.

[S12] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. “Detecting hallucinations in large language models using semantic entropy.” Nature 630, published 19 June 2024. DOI record. Inspected: publisher record, abstract, methods/results summary, and bibliographic metadata. Supports semantic rather than surface-form uncertainty for detecting a subset of confabulations.

[S13] Lexin Zhou et al. “Larger and more instructable language models become less reliable.” Nature 634, published 25 September 2024. Open-access publisher page. Inspected: abstract, methods overview, and main findings. Supports evaluating fluent wrong answers and residual variability rather than assuming scale creates a reliable operating range.

[S14] Ziyu Wang and Christopher C. Holmes. “On Subjective Uncertainty Quantification and Calibration in Natural Language Generation.” AISTATS 2025, PMLR 258. Official proceedings page. Inspected: abstract, metadata, citation record, and linked paper record. Supports task-specific semantic utility and epistemic deferral.

[S15] Jason Hartline, Lunjia Hu, and Yifan Wu. “A Perfectly Truthful Calibration Measure.” COLT 2026, PMLR 336. Official proceedings page. Inspected: abstract, metadata, theorem summary, and linked paper record. Supports careful calibration evaluation under finite samples and incentives.

Reproducibility statement

The accompanying bundle contains:

- this manuscript in Markdown;
- the rendered PDF;
- `round28_uncertainty_replay.py`;
- `test_round28_uncertainty_replay.py`;
- the inherited Round 27 finite-model source used by import;
- exact episode, history-calibration, risk-coverage, adequacy-detection, and paired-cell CSVs;
- JSON summary;
- source-inspection ledger;
- SHA-256 manifests and a README with commands.

All reported E28-A numbers are generated by the included program. No production or language-model result is inferred from the diagnostic.