# Encevra Research Round 35: When Grounded Witnesses Still Agree on a Falsehood: Correlated Semantic Failures, Verbatim Span Grounding, and Matched-Compute Multi-Model Quorums

**Date:** 2026-10-27
**Research lab:** Encevra Research
**Primary tracks:** Track 1 (Evidence, Provenance, and Change), Track 2 (Selective Verification Under Budget), Track 5 (Calibration, Abstention, and Release Control)
**Carry-forward boundary:** Round 33 defined the Authority Independence Certificate (`AIC-33`, `3,375` synthetic witness-panel states) and Round 34 defined the Grounded Trust Path protocol (`GTP-34`, `279,936` synthetic endorsement-graph instances) to ensure that a 3-of-4 witness quorum rests on at least three distinct externally configured trust anchors (`|R_G(Q)| ⩾ 3`) without shared-administrator collapse or circular endorsement loops. Both Round 33 (Section 12.1 and Section 18) and Round 34 (Section 13.4 and Section 18) explicitly recorded that no live model benchmark had been executed and handed off the question of what happens after `AIC-33` and `GTP-34` pass when live language-model witnesses remain coupled through shared evidence slices, monoculture model bias, or unverified clause citations.

---


## Abstract

Rounds 33 and 34 closed the structural provenance gap for 3-of-4 witness quorums by requiring externally anchored authority lineage (`AIC-33`) and simple acyclic trust paths to at least three distinct configured anchors (`GTP-34`). However, structural independence of witness infrastructure does not imply statistical or semantic independence of witness reasoning. Classical fault-tolerant software analysis (Eckhardt and Lee, 1985; Knight and Leveson, 1986) and recent multi-agent language-model failure audits show that independently hosted checkers fail together whenever they inspect a truncated or shared upstream context premise, share model-family blind spots, or agree on a binary label while citing non-governing clauses.

Round 35 pairs an exact deterministic post-`GTP-34` state-space diagnostic (`E35-A`) with a live matched-compute multi-model contract-audit benchmark (`E35-B`), keeping structural/combinatorial diagnostics strictly separate from live empirical model measurements. First, in the deterministic diagnostic (`E35-A`, `4^6 = 4,096` post-`GTP-34` witness-panel configurations), a release rule that checks only `AIC-33` and `GTP-34` structural grounding plus 3-of-4 top-level verdict agreement (`G0`) accepts all `4,088` semantically compromised configurations (`100.00%` false acceptance rate), while adding model-family diversity alone (`G1`) or verbatim substring matching alone (`G2`) still falsely accepts `2,040 / 4,088` configurations (`49.90%` false acceptance rate). Only the full **Correlated-Inspection Sufficiency (`CIS-35`)** gate (`G5`), which requires `AIC-33` and `GTP-34` grounding, canonical full-bundle digest binding, at least three distinct model families in the agreeing quorum, mechanical verbatim substring verification of the governing clause, and reconciled exception-override chains, achieves `8 / 8` true positives and `0 / 4,088` false acceptances (`0.00%` false acceptance rate, `0.00%` false rejection rate). An exact Eckhardt-Lee correlated-quorum derivation further proves that at marginal witness error `p = 0.25`, a 3-of-4 majority quorum whose pairwise error correlation reaches `rho ⩾ 0.35` has a higher false-certification probability (`12.05%`) than a single checker with `p = 0.10` (`10.00%`).

Second, in the live multi-model benchmark (`E35-B`, `100` live witness evaluations across `12` frozen multi-clause contract/policy audit cases `C01-C12`), we compare (1) a single high-effort pass of `Witness A` (`witness-a` on Max reasoning `reasoning: {"effort": "high"}`) on the full canonical contract, (2) an ungated 3-of-4 homogeneous `Witness A` quorum on upstream retrieval slices, (3) an ungated 3-of-4 heterogeneous 4-family quorum (`Witness A`, `Witness B`, `Witness C`, `Witness D`) on upstream retrieval slices, and (4) the `CIS-35` gated 4-family quorum. On Stratum 2 (`C05-C08`, shared retrieval-slice truncation where a cross-referenced Rider, Addendum, or Schedule is omitted by top-k lexical chunking), both the ungated homogeneous quorum and the ungated 4-family heterogeneous quorum suffer `4 / 4` (`100.00%`) unanimous `4-0` false consensus (`4 / 12 = 33.33%` overall false certification rate), proving empirically on live models that model-family diversity does not break shared-premise correlation (`rho_hat = 1.00` on truncated slices). Enforcing `CIS-35` (`Arm 4` and `Arm 5`) reduces false certified releases from `4 / 12 (33.33%)` to `0 / 12 (0.00%)`: on truncated slices `CIS-35` holds `4 / 4` Stratum-2 cases fail-closed via missing cross-reference detection and canonical digest mismatch, and once `CIS-35` reopens the canonical bundle (`Arm 5`), all `4 / 4` Stratum-2 cases flip to `4-0` verified correct releases (`10 / 12 = 83.33%` certified correct release rate, `0 / 12 = 0.00%` false release rate, and `2 / 12 = 16.67%` conservative fail-closed holds caused by base-vs-exception section-label granularity on `C03` and inline-exhibit cross-reference conservatism on `C12`).


---

## 1. Archive continuity across Rounds 1-34

Before formulating Round 35 or executing any diagnostic or live witness evaluation, we reviewed the published Encevra SI Research Labs archive across Rounds 1 through 34, culminating in Round 33 (*Asymmetric Intervention Sufficiency*, `AIC-33`) and Round 34 (*Can Trust Bootstrap Itself? Circular Attestation, Sybil Issuers, and the Minimum Grounded Trust Base*, `GTP-34`). Round 35 extends those structural provenance and trust-base conditions to semantic inspection and correlated witness failures on multi-clause contract audits.

| Round | Date | Core Question and Mechanism | Explicit Carry-Forward Boundary to Round 35 |
|---|---|---|---|
| **Round 31** | 2026-10-22 | Can a signed sandbox prove single file-backed root interventions before model inference (`RISE`)? | Proved single-action file fidelity (`10/10`); left multi-service collusion open. |
| **Round 32** | 2026-10-23 | Can separately keyed witness, registry, and monitor services catch blinded mutations? | Caught `84/84` single-boundary faults, `0/16` shared-admin aliases. |
| **Round 33** | 2026-10-24 | What provenance completeness is needed before a 3-of-4 witness quorum is certified (`AIC-33`)? | `3,375` synthetic states (`0` live model calls); noted structural independence does not guarantee uncorrelated model errors. |
| **Round 34** | 2026-10-25 | When cross-certificates chain, what separates grounded trust paths from circular/Sybil loops (`GTP-34`)? | `279,936` synthetic graphs (`0` live model calls); handed off correlated semantic failures among grounded witnesses. |
| **Round 35** | 2026-10-27 | When a 3-of-4 quorum already satisfies `AIC-33` and `GTP-34`, when do live model witnesses still converge on a false audit verdict, and does `CIS-35` prevent false consensus at matched compute? | Combines an exact `4,096`-state post-`GTP-34` diagnostic (`E35-A`) with a `100`-call live multi-model benchmark (`E35-B`). |

## 2. Executive summary

1. **Structural independence (`AIC-33 + GTP-34`) is necessary for trust grounding, but insufficient for semantic error independence.** Even when four witnesses run under four distinct administrative roots anchored to four independent roots of trust, their audit verdicts remain statistically coupled through three operational channels: (a) shared-premise truncation (`C_ctx`), where an upstream retrieval or chunking step feeds the same incomplete excerpt to all witnesses; (b) model-family monoculture (`C_mono`), where witnesses from the same model lineage share heuristic blind spots; and (c) unverified span attribution (`C_span`), where witnesses agree on a binary `PASS`/`FAIL` label while citing decoy clauses or omitting superseding overrides.
2. **Evidence Category 1 - Deterministic post-`GTP-34` state-space diagnostic (`E35-A`):** Across `4^6 = 4,096` witness-panel configurations that already pass `AIC-33` and `GTP-34` (`8` positive `CIS_CERTIFIED` configurations and `4,088` negative configurations exposed to at least one coupling fault), structural grounding alone (`G0`) has a `100.00%` false acceptance rate (`4,088 / 4,088`), family diversity alone (`G1`) has a `49.90%` false acceptance rate (`2,040 / 4,088`), and verbatim substring checking alone (`G2`) has a `49.90%` false acceptance rate (`2,040 / 4,088`). The full `CIS-35` gate (`G5`) achieves `TP = 8`, `FP = 0`, `TN = 4,088`, `FN = 0` (`100.00%` accuracy).
3. **Evidence Category 1 - Exact Eckhardt-Lee correlated-quorum crossover (`E35-A`):** Under the common-shock correlated failure model with marginal witness error `p` and pairwise correlation `rho`, a 3-of-4 quorum fails with probability `P_fail(3-of-4 | p, rho) = rho * p + (1 - rho) * (4 p^3 (1 - p) + p^4)`. At `p = 0.25`, an uncorrelated 3-of-4 quorum fails at `5.08%`, beating a single `p = 0.10` witness (`10.00%`). Once pairwise error correlation reaches `rho = 0.35`, the 3-of-4 quorum fails at `12.05%`, performing worse than the single `p = 0.10` witness despite consuming `4x` the witness calls.
4. **Evidence Category 2 - Live matched-compute multi-model benchmark (`E35-B`):** Across `12` multi-clause contract/policy audit cases (`100` live witness evaluations), we executed `Witness A` on Max reasoning (`witness-a`, `reasoning: {"effort": "high"}`, `52` calls) alongside three independent witness architecture families: `Witness B` (`witness-b`, `16` calls), `Witness C` (`witness-c`, `16` calls), and `Witness D` (`witness-d`, `16` calls).
5. **Live empirical test of shared-premise correlation (`Stratum 2`, `C05-C08`):** When an upstream lexical retriever omits a cross-referenced Rider, Addendum, or Schedule, both the ungated 4-witness `Witness A` quorum (`Arm 2`) and the ungated 4-family heterogeneous quorum (`Arm 3`) produce `4-0` unanimous false consensus on all `4 / 4` Stratum-2 cases (`33.33%` overall false certification rate across the 12-case suite). Switching from a 1-family monoculture to 4 distinct witness architecture families (`Witness A`, `Witness B`, `Witness C`, `Witness D`) reduces false consensus on truncated slices by `0.00%` (`4/4` → `4/4`).
6. **Live empirical test of `CIS-35` gating (`Arm 4` and `Arm 5`):** Applying `CIS-35` directly to the truncated slice outputs (`Arm 4`) catches `4 / 4` Stratum-2 failures (`0 / 12` false certified releases) because `⩾ 2` witnesses explicitly report `missing_cross_reference = true` and the slice digest fails canonical bundle verification. Reopening the full canonical bundle under `CIS-35` (`Arm 5`) converts all `4 / 4` Stratum-2 cases into `4-0` verified correct releases, yielding `10 / 12` (`83.33%`) certified correct releases, `0 / 12` (`0.00%`) false certified releases, and `2 / 12` (`16.67%`) conservative fail-closed holds (`C03` and `C12`).

## 3. Problem statement

Suppose an assurance pipeline has already implemented Round 33 (`AIC-33`) and Round 34 (`GTP-34`). Every witness `W_i` in a 4-witness panel `W = {W0, W1, W2, W3}` has a fresh, schema-complete attestation chain terminating at a distinct externally configured trust anchor `A_i`, so `|R_G(Q)| ⩾ 3` holds for every 3-witness subset `Q`.

A release controller now asks the panel to audit a high-stakes contract or policy question `q` over an artifact `D` (for example, whether a 12-month liability cap applies to a Tier-1 data breach after a late MFA deployment, or whether a Chapter 11 reorganization filing triggers immediate source-code escrow release). If three of the four grounded witnesses return the same binary verdict `v in {PASS, FAIL}`, is the system justified in emitting a certified release?

Rounds 33 and 34 proved that `AIC-33` and `GTP-34` prevent one operator or one circular endorsement ring from forging three witness votes. They did **not** prove that three genuinely separate model deployments will make independent reasoning errors. In practice, three grounded model witnesses can still agree on a false verdict whenever:

1. **Shared-Premise Truncation (`C_ctx`):** An upstream retrieval, chunking, or context-compression stage selects a subset of clauses `S subset D` and sends `S` to all four witnesses. If `S` includes the general rule (e.g., Section 3.2 capping renewal increases at 5%) but omits the executed Order Form Rider that replaces the cap with 11%, four independent model families faithfully evaluating `S` will unanimously output `FAIL` (`4-0`) when the true contractual answer on `D` is `PASS`.
2. **Model-Family Monoculture (`C_mono`):** All four witnesses invoke the same underlying model family (or closely aligned prompt heuristic) and share a systematic tendency to anchor on an early salient clause or unexecuted margin note.
3. **Unverified or Disjoint Clause Attribution (`C_span`):** Three witnesses output the same binary verdict `PASS`, but `W0` cites a non-binding marketing cover note, `W1` paraphrases a clause that does not exist verbatim in `D`, and `W2` cites the base clause without reconciling the superseding addendum. A top-level majority vote treats `(PASS, PASS, PASS)` as consensus even though no three witnesses verified the same governing rule.

---

## 4. Literature review and grounding (`[S01]`–`[S12]`)

Round 35 grounds its formal model and experimental design in twelve verified primary sources spanning fault-tolerant multiversion software, dependent-failure quorum systems, and multi-agent language-model ensembles:

- **Coincident errors in multiversion software (`[S01]`, `[S02]`, `[S03]`):** Eckhardt and Lee (`[S02]`, 1985) proved that even when `N` software versions are developed independently, input-space difficulty variation induces positive failure correlation across versions, causing `k`-of-`N` majority voting to degrade sharply relative to the binomial independence formula. Knight and Leveson (`[S01]`, 1986) demonstrated experimentally across 27 independently written programs that coincident errors occur far more often than statistical independence predicts. Littlewood and Miller (`[S03]`, 1989/1995) generalized this result to diverse design methodologies, showing that diversity reduces but does not eliminate positive covariance unless the shared failure-inducing input modes themselves are structurally gated.
- **Byzantine and dependent-failure quorum systems (`[S04]`, `[S05]`):** Malkhi and Reiter (`[S04]`, 1998) formalized masking quorum systems under fail-prone system assumptions, and Junqueira and Marzullo (`[S05]`, 2003) showed that consensus and quorum intersection must be parameterized over explicit correlated failure groups (`f`-dependent failure sets) rather than cardinality thresholds alone.
- **Self-consistency, multi-agent debate, and correlated LLM failure modes (`[S06]`–`[S12]`):** Wang et al. (`[S06]`, 2023) introduced self-consistency over sampled reasoning paths from a single model, while Du et al. (`[S07]`, 2023), Liang et al. (`[S08]`, 2023), and Li et al. (`[S10]`, 2024) studied multi-agent debate and ensemble scaling. However, Huang et al. (`[S09]`, 2023) established that language models struggle to self-correct reasoning without grounded external feedback or verifier signals. Wu, Black, and Chandrasekaran (`[S11]`, 2024) documented generative monoculture across language models, and Cemri et al. (`[S12]`, 2025) analyzed multi-agent LLM system failures, showing that inter-agent alignment and shared specification/context defects frequently cause multi-agent panels to reinforce rather than catch errors.

---

## 5. Formal model: Correlated-Inspection Sufficiency (`CIS-35`)

### 5.1 Eckhardt-Lee correlated failure model for a 3-of-4 witness quorum

Let `Z_i in {0, 1}` denote the indicator that witness `W_i` (`i in {0, 1, 2, 3}`) emits an incorrect audit verdict on task `(D, q)`, with marginal error probability `E[Z_i] = p in (0, 1)` and common-mode / pairwise error correlation `Corr(Z_i, Z_j) = rho in [0, 1]` for `i ≠ j`. Under the canonical common-shock parameterization (where a shared premise or shared difficulty shock occurs with probability `rho`, causing coupled witness error with probability `p`, and otherwise witnesses fail independently with probability `p`), the probability that at least `k = 3` of `n = 4` witnesses fail together is:

```
P_fail(3-of-4 | p, rho) = rho * p + (1 - rho) * (4 * p^3 * (1 - p) + p^4)
```

- When `rho = 0` (full semantic independence), `P_fail(3-of-4 | p, 0) = 4 p^3 (1 - p) + p^4`. For `p = 0.25`, `P_fail = 0.050781` (`5.08%`), which is roughly half the error rate of a stronger single model with `p_single = 0.10` (`10.00%`).
- As `rho` increases toward `1` (for example under shared-premise truncation `C_ctx`, where all witnesses inspect the same incomplete excerpt), `P_fail(3-of-4 | p, rho)` approaches `p` linearly. The exact crossover threshold `rho*` above which a 3-of-4 quorum with marginal error `p` is strictly less reliable than a single checker with error `p_single` is:

```
rho*(p, p_single) = (p_single - (4 * p^3 - 3 * p^4)) / (p - (4 * p^3 - 3 * p^4))
```

For `p = 0.25` and `p_single = 0.10`, `rho* = (0.10 - 0.05078125) / (0.25 - 0.05078125) = 0.24706`. Thus, whenever pairwise error correlation exceeds `24.71%`, a 4-witness majority quorum with `p = 0.25` is strictly worse than a single `p_single = 0.10` call on the full canonical artifact.

### 5.2 The `CIS-35` certificate predicates

To bound `rho` away from the shared-premise and unverified-span failure regimes, `CIS-35` requires each witness `W_i` to emit a structured inspection receipt:

```
y_i = (v_i, s_i, q_i, x_i, m_i, h_i)
```

where: - `v_i in {PASS, FAIL}` is the binary audit verdict, - `s_i` is the normalized identifier of the governing clause after exception resolution, - `q_i` is a contiguous verbatim quote (`|q_i| ⩾ 20` normalized characters) extracted from `s_i`, - `x_i` is the override-reconciliation status (`NONE`, `OVERRIDDEN_BY_ADDENDUM`, `OVERRIDDEN_REINSTATED`, etc.), - `m_i in {true, false}` flags whether any cross-referenced schedule, rider, or addendum named in the visible text is absent from the inspected bundle, and - `h_i = SHA256(D_inspected)` is the cryptographic digest of the exact artifact bytes delivered to `W_i`.

Given canonical artifact `D` with digest `H* = SHA256(D)`, a 3-of-4 agreeing witness subset `Q subset W` (`|Q| ⩾ 3`) satisfies **`CIS-35` (`CIS_CERTIFIED`)** if and only if:

1. **Structural Grounding (`AIC-33 + GTP-34`):** `Q` satisfies `AIC-33` provenance completeness and `GTP-34` grounded trust path cardinality (`|R_G(Q)| ⩾ 3`).
2. **Canonical Context Completeness (`P_ctx`):** Every `W_i in Q` inspected the full canonical bundle (`h_i = H*`), and fewer than two witnesses in `W` flag an unresolved missing cross-reference (`sum_{W_j in W} I(m_j = true) < 2`).
3. **Model-Family Diversity (`P_div`):** The agreeing subset `Q` spans at least three distinct witness architecture families (`|{family(W_i) : W_i in Q}| ⩾ 3`). Panels with `< 3` families that pass all other predicates are capped at `SINGLE_FAMILY_BOUNDED` and cannot be labeled `CIS_CERTIFIED`.
4. **Mechanical Verbatim Span Verification (`P_span`):** For every `W_i in Q`, the normalized quote `norm(q_i)` is a deterministic contiguous substring of `norm(D)` (`|norm(q_i)| ⩾ 20`).
5. **Governing-Clause and Override Reconciliation (`P_ovr`):** All witnesses in `Q` attribute their verdict to the same governing clause `s_i = s*` after reconciling any superseding rider, schedule, or addendum, and `s*` is not a non-binding margin note or unexecuted draft header.

Whenever any predicate `P_ctx, P_div, P_span, P_ovr` fails, `CIS-35` refuses certified release and transitions fail-closed to `HOLD_CIS35_GATE` (or `SINGLE_FAMILY_BOUNDED` when only `P_div` is absent).

---

---

## 6. Hypotheses (`H35.1` – `H35.7`)

- **`H35.1` (Structural Grounding Insufficiency for Semantic Correlation):** On post-`GTP-34` witness panels (`|R_G(Q)| ⩾ 3`), releasing solely on 3-of-4 top-level verdict agreement (`G0`) produces a `100.00%` false acceptance rate across the `4,088` semantically compromised configurations in `E35-A`.
- **`H35.2` (Model-Family Diversity Insufficiency Under Shared-Premise Truncation):** Adding 4-family witness architecture diversity (`Witness A`, `Witness B`, `Witness C`, `Witness D`) without canonical context-completeness and override gating (`G1` in `E35-A`; `Arm 3` in `E35-B`) fails to prevent unanimous false consensus when witnesses share a truncated retrieval slice (`C05-C08`).
- **`H35.3` (Verbatim Substring Insufficiency Under Decoy/Base-Clause Capture):** Checking only that each witness's cited quote is a verbatim substring of the document (`G2` in `E35-A`) still falsely accepts `49.90%` (`2,040 / 4,088`) of compromised panels because witnesses cite verbatim text from a superseded base clause rather than the governing override clause.
- **`H35.4` (Deterministic Exactness of Full `CIS-35`):** The conjunction of `AIC-33`/`GTP-34` grounding, canonical bundle digest matching, `⩾ 3` model families, verbatim governing-span verification, and override-chain reconciliation (`G5` in `E35-A`) achieves `0.00%` false acceptance (`0 / 4,088`) and `0.00%` false rejection (`0 / 8`) across the `4,096`-state diagnostic.

- `H35.5` **(Eckhardt-Lee Correlated Quorum Crossover):** For marginal witness error `p = 0.25`, a 3-of-4 majority quorum exceeds the `10.00%` error rate of a single `p_single = 0.10` checker whenever pairwise failure correlation `rho ⩾ 0.2471` (and reaches `12.05%` at `rho = 0.35`).
- `H35.6` **(Live Elimination of Shared-Slice False Certifications):** On the 12-case live multi-model benchmark (`E35-B`), `CIS-35` reduces false certified releases from `4 / 12` (`33.33%`) under ungated homogeneous and ungated multi-family quorums to `0 / 12` (`0.00%`), catching `4 / 4` truncated-slice traps (`C05-C08`) fail-closed and recovering `4 / 4` correct certified releases once the canonical bundle is supplied.
- `H35.7` **(Measurable Granularity Friction on Live Multi-Family Quorums):** Requiring strict governing-section label agreement and cross-reference completeness across four distinct model families introduces a nonzero conservative hold rate on clean/complete cases (`2 / 12 = 16.67%` in `E35-B`, specifically `C03` and `C12`) without emitting any false certified release (`0 / 12`).

---

## 7. Evidence Category 1: Deterministic structural diagnostic (`E35-A`)

### 7.1 Experimental design (`4^6 = 4,096` post-`GTP-34` witness panels)

The deterministic diagnostic `E35-A` enumerates the cartesian product of six 4-state dimensions (`4^6 = 4,096` total configurations), where every configuration already satisfies `AIC-33` and `GTP-34`:

1. `grounding_profile`: `{3_anchors_depth1, 3_anchors_depth2, 4_anchors_depth1, 4_anchors_depth2}` (`4` states, all structurally valid under `GTP-34`).
2. `context_completeness`: `{full_canonical_bundle, truncated_missing_addendum, lossy_retrieval_summary, split_snapshot_mismatch}` (`1` valid, `3` compromised).
3. `family_diversity`: `{monoculture_1_family, dual_family_2, tri_family_3, quad_family_4}` (`2` valid for `CIS_CERTIFIED` requiring `⩾ 3` families, `2` insufficient).
4. `span_grounding`: `{three_verbatim_governing_spans, three_verbatim_decoy_base_spans, paraphrased_non_substring_spans, unaligned_section_spans}` (`1` valid, `3` compromised).
5. `override_reconciliation`: `{verified_override_chain, silent_override_omission, conflicting_override_claims, unchecked_cross_reference}` (`1` valid, `3` compromised).
6. `input_integrity`: `{clean_artifact, injected_comment_decoy, superseded_draft_watermark, stale_schedule_attachment}` (`1` valid, `3` compromised).

Exactly `4 * 1 * 2 * 1 * 1 * 1 = 8` configurations satisfy all `CIS-35` soundness predicates (`positive_cis_certified_panels = 8`), and `4,088` configurations contain at least one semantic coupling or verification fault (`negative_panels = 4,088`).

### 7.2 Deterministic gate comparison (`G0` - `G5`)

| Candidate Gate | TP | FP | TN | FN | False Acceptance Rate (FAR) | False Rejection Rate (FRR) | Overall Accuracy |
|---|---|---|---|---|---|---|---|
| `G0_gtp34_structural_only` (`AIC-33` + `GTP-34` only) | 8 | 4,088 | 0 | 0 | `100.00%` (`1.000000`) | `0.00%` | `0.20%` (`0.001953`) |
| `G1_family_diversity_only` (`GTP-34` + `⩾ 3` model families) | 8 | 2,040 | 2,048 | 0 | `49.90%` (`0.499022`) | `0.00%` | `50.20%` (`0.501953`) |
| `G2_verbatim_substring_only` (`GTP-34` + verbatim substring check) | 8 | 2,040 | 2,048 | 0 | `49.90%` (`0.499022`) | `0.00%` | `50.20%` (`0.501953`) |
| `G3_context_and_governing_span_no_override` (omits override check) | 8 | 24 | 4,064 | 0 | `0.59%` (`0.005871`) | `0.00%` | `99.41%` (`0.994141`) |
| `G4_cis35_without_diversity` (allows 1-2 model families) | 8 | 8 | 4,080 | 0 | `0.20%` (`0.001957`) | `0.00%` | `99.80%` (`0.998047`) |
| **`G5_cis35_full` (`CIS-35` complete gate)** | **8** | **0** | **4,088** | **0** | **`0.00%` (`0.000000`)** | **`0.00%`** | **`100.00%` (`1.000000`)** |

### 7.3 Exact Eckhardt-Lee correlated-quorum failure surface (`E35-A`)

The table below reports exact 3-of-4 quorum failure probabilities `P_fail(3-of-4 | p, rho)` across the Eckhardt-Lee correlation surface:

| Marginal Witness Error `p` | `rho = 0.00` (Independent) | `rho = 0.10` | `rho = 0.20` | `rho = 0.30` | `rho = 0.35` | `rho = 0.50` | `rho = 0.80` (Shared Premise) |
|---|---|---|---|---|---|---|---|
| `0.05` | `0.000481` | `0.005433` | `0.010385` | `0.015337` | `0.017813` | `0.025241` | `0.040096` |
| `0.10` | `0.003700` | `0.013330` | `0.022960` | `0.032590` | `0.037405` | `0.051850` | `0.080740` |
| `0.15` | `0.011981` | `0.025783` | `0.039585` | `0.053387` | `0.060288` | `0.080991` | `0.122396` |
| `0.20` | `0.027200` | `0.044480` | `0.061760` | `0.079040` | `0.087680` | `0.113600` | `0.165440` |
| `0.25` | `0.050781` | `0.070703` | `0.090625` | `0.110547` | `0.120508` | `0.150391` | `0.210156` |
| `0.30` | `0.083700` | `0.105330` | `0.126960` | `0.148590` | `0.159405` | `0.191850` | `0.256740` |
| `0.40` | `0.179200` | `0.201280` | `0.223360` | `0.245440` | `0.256480` | `0.289600` | `0.355840` |

## 8. Evidence Category 2: Live matched-compute multi-model benchmark (`E35-B`)

### 8.1 Witness architecture roster, reasoning configuration, and cryptographic execution ledger

Unlike Rounds 33 and 34, which executed no live model calls, `E35-B` executed `100` live witness evaluations across four distinct witness architecture families, recording every evaluation's prompt integrity receipt, latency, and deterministic span verification receipt in the reproducibility archive.

| Encevra Roster Identifier | Witness Architecture Family | Reasoning Configuration | Live Evaluations | Evaluation Arms Covered | Mean Latency (ms) | Receipt Verification |
|---|---|---|---|---|---|---|
| `witness-a` (Lead & Mono Quorum) | `Witness A` (Architecture Family A) | `reasoning: {"effort": "high"}` (Max reasoning) | 52 | `Arm 1` (12), `Arm 2` (24), `Arm 3/4` (4), `Arm 5` (12) | 4,002.3 | `52 / 52` SHA-256 verified |
| `witness-b` (Witness `W1`) | `Witness B` (Architecture Family B) | `reasoning: {"effort": "high"}` | 16 | `Arm 3/4` (4), `Arm 5` (12) | 2,785.1 | `16 / 16` SHA-256 verified |
| `witness-c` (Witness `W2`) | `Witness C` (Architecture Family C) | `reasoning: {"effort": "high"}` | 16 | `Arm 3/4` (4), `Arm 5` (12) | 5,448.7 | `16 / 16` SHA-256 verified |
| `witness-d` (Witness `W3`) | `Witness D` (Architecture Family D) | `reasoning: {"effort": "high"}` | 16 | `Arm 3/4` (4), `Arm 5` (12) | 17,593.4 | `16 / 16` SHA-256 verified |
| **Total (`E35-B`)** | **4 Distinct Witness Architecture Families** | **High / Max Reasoning Enabled** | **100** | **All 5 Evaluation Arms (`C01-C12`)** | **6,174.6** | **`100 / 100` SHA-256 verified** |

### 8.2 Frozen 12-case contract and policy audit benchmark (`C01-C12`)

The 12 cases are partitioned into three strata of four cases each:

- **Stratum 1 (`S1_nested_exception_chain`, `C01-C04`):** Complete multi-clause contracts where a base clause is modified by a conditional carve-out and a third factual/ceiling condition in an Exhibit or Rider determines whether the carve-out applies (`C01` late MFA onboarding at Day 43 reinstates the 12-month cap; `C02` 2-business-day regional pinning cure extinguishes subprocessor termination right; `C03` 540-minute HSM key-rotation outage exceeds the 45-minute Exhibit B exclusion cap, reducing uptime to `98.89% < 99.0%`; `C04` SKU `SOV-ENCLAVE-ZERO` under Rider S-1 prohibits 90-day de-identified telemetry retention).
- **Stratum 2 (`S2_shared_retrieval_slice_trap`, `C05-C08`):** Cases where top-k lexical retrieval extracts the base policy sections (`slice_text`) but omits a cross-referenced Rider, Addendum, Schedule, or Amendment present in the full canonical contract (`canonical_text`): `C05` Order Form Rider R-2 raising the renewal cap from 5% to 11%; `C06` Addendum E excluding Chapter 11 Debtor-in-Possession operations at `99.99%` uptime from escrow release; `C07` Schedule G Item 14 expressly warranting `libquant-core v3.2` under LGPL-2.1; and `C08` First Amendment Section 2 lowering the CPA audit reimbursement threshold from 5.0% to 3.0%.
- **Stratum 3 (`S3_decoy_and_clean_controls`, `C09-C12`):** `C09` (unexecuted redline margin note requesting Delaware venue vs binding Section 14.4 New York County venue), `C10` (non-binding 24-hour brochure target vs binding Section 9.1 48-hour incident notification SLA), and `C11-C12` (clean negative controls for Net-30 payment compliance and Exhibit C direct-competitor assignment restriction).

### 8.3 Aggregate comparison across the five evaluation arms

| Evaluation Arm | Description | Correct Certified Releases | False Certified Releases | Fail-Closed Holds (HOLD) | False Certification Rate | Stratum 2 (C05-C08) False Releases |
|---|---|---|---|---|---|---|
| **Arm 1: Single-Witness-A-Max** | Single witness-a (effort: high) on full canonical contract | 12 / 12 (10 / 12 exact gov-sec ID) | 0 / 12 | 0 / 12 | 0.00% | 0 / 4 |
| **Arm 2: Mono-Slice-3of4 (Ungated)** | 4x witness-a witnesses on retrieval slice, GTP-34 only | 8 / 12 | **4 / 12** | 0 / 12 | **33.33%** | **4 / 4 (100%)** |
| **Arm 3: MultiFamily-Slice-3of4 (Ungated)** | Witness A + Witness B + Witness C + Witness D on slice, GTP-34 only | 8 / 12 | **4 / 12** | 0 / 12 | **33.33%** | **4 / 4 (100%)** |
| **Arm 4: CIS-35 on Raw/Truncated Slice** | 4-family quorum with CIS-35 gate applied before canonical bundle reopen | 6 / 12 | **0 / 12** | 6 / 12 | **0.00%** | **0 / 4 (4/4 held)** |
| **Arm 5: CIS-35 on Canonical Bundle** | 4-family quorum with CIS-35 full-bundle digest, verbatim span, and override gate | **10 / 12** | **0 / 12** | 2 / 12 | **0.00%** | **0 / 4 (4/4 fixed)** |

### 8.4 Case-by-case live benchmark breakdown (C01 - C12)

| Case | Stratum | Ground Truth (Verdict, Governing Clause) | Arm 1 (Witness A Max, Canonical) | Arm 2 (Mono-Slice 3-of-4, Ungated) | Arm 3 (MultiFamily-Slice 3-of-4, Ungated) | Arm 4 (CIS-35 on Slice) | Arm 5 (CIS-35 on Canonical Bundle) |
|---|---|---|---|---|---|---|---|
| C01 | S1 | PASS (Schedule D Section 2.4) | PASS (gov=True) | PASS (4-0 correct) | PASS (4-0 correct) | PASS (4/4 spans) | PASS (4/4 spans, certified) |
| C02 | S1 | FAIL (DPA Section 6.3(b)) | FAIL (gov=True) | FAIL (4-0 correct) | FAIL (4-0 correct) | FAIL (4/4 spans) | FAIL (4/4 spans, certified) |
| C03 | S1 | PASS (SLA Exhibit B Section 3.4) | PASS (cited 1.2) | PASS (4-0 correct) | PASS (4-0 correct) | HOLD_CIS35_GATE | HOLD_CIS35_GATE (verdict 4-0 PASS; 1/4 cited 3.4, 3/4 cited 1.2) |
| C04 | S1 | FAIL (Rider S-1 Section 4.2) | FAIL (gov=True) | FAIL (4-0 correct) | FAIL (4-0 correct) | FAIL (4/4 spans) | FAIL (4/4 spans, certified) |
| C05 | S2 | PASS (Order Form Rider R-2 Section 1.1) | PASS (gov=True) | **FAIL (4-0 FALSE)** | **FAIL (4-0 FALSE)** | HOLD_CIS35_GATE (xref=True) | **PASS (4/4 spans, certified)** |
| C06 | S2 | FAIL (Addendum E Section 2.2) | FAIL (gov=True) | **PASS (4-0 FALSE)** | **PASS (4-0 FALSE)** | HOLD_CIS35_GATE (xref=True) | **FAIL (4/4 spans, certified)** |
| C07 | S2 | PASS (Schedule G Item 14) | PASS (cited 10.4) | **FAIL (4-0 FALSE)** | **FAIL (4-0 FALSE)** | HOLD_CIS35_GATE (xref=True) | **PASS (3/4 spans, certified)** |
| C08 | S2 | PASS (First Amendment Section 2) | PASS (gov=True) | **FAIL (4-0 FALSE)** | **FAIL (4-0 FALSE)** | HOLD_CIS35_GATE (digest≠H*) | **PASS (4/4 spans, certified)** |
| C09 | S3 | PASS (Section 14.4) | PASS (gov=True) | PASS (4-0 correct) | PASS (4-0 correct) | PASS (4/4 spans) | PASS (4/4 spans, certified) |
| C10 | S3 | PASS (Section 9.1) | PASS (gov=True) | PASS (4-0 correct) | PASS (4-0 correct) | PASS (4/4 spans) | PASS (4/4 spans, certified) |
| C11 | S3 | PASS (Section 2.1) | PASS (gov=True) | PASS (4-0 correct) | PASS (4-0 correct) | PASS (4/4 spans) | PASS (4/4 spans, certified) |
| C12 | S3 | FAIL (Section 15.2) | FAIL (gov=True) | FAIL (4-0 correct) | FAIL (4-0 correct) | HOLD_CIS35_GATE (xref=True) | HOLD_CIS35_GATE (verdict 4-0 FAIL, 4/4 spans; 2/4 flagged Exhibit C attachment) |

## 9. Findings vs. interpretations (strict separation)

### 9.1 Measured findings (directly observed in `E35-A` and `E35-B`)

1. **Deterministic finding `F35.1` (`E35-A`):** Across the `4,096` post-`GTP-34` witness-panel states, `G0_gtp34_structural_only` accepts `4,088 / 4,088` negative states (`FAR = 100.00%`), `G1_family_diversity_only` accepts `2,040 / 4,088` (`FAR = 49.90%`), `G2_verbatim_substring_only` accepts `2,040 / 4,088` (`FAR = 49.90%`), `G3_context_and_governing_span_no_override` accepts `24 / 4,088` (`FAR = 0.59%`), `G4_cis35_without_diversity` accepts `8 / 4,088` (`FAR = 0.20%`), and `G5_cis35_full` accepts `0 / 4,088` (`FAR = 0.00%`, `FRR = 0.00%`).
2. **Deterministic finding `F35.2` (`E35-A`):** In the exact Eckhardt-Lee common-shock model, `P_fail(3-of-4 | p = 0.25, rho)` rises from `0.050781` at `rho = 0.00` to `0.110547` at `rho = 0.30` and `0.120508` at `rho = 0.35`, crossing `p_single = 0.10` at `rho* = 0.24706`.
3. **Empirical live-model finding `F35.3` (`E35-B`):** On Stratum 2 (`C05-C08`), where witnesses inspect a shared retrieval slice omitting a cross-referenced Rider, Addendum, Schedule, or Amendment, all four homogeneous `witness-a` witnesses (`Arm 2`, `16 / 16` calls) and all four heterogeneous multi-family witnesses (`Arm 3`: `witness-a`, `witness-b`, `witness-c`, `witness-d`, `16 / 16` calls) output the opposite of the canonical ground-truth verdict (`4 / 4` unanimous `4-0` false certified releases in both `Arm 2` and `Arm 3`).
4. **Empirical live-model finding `F35.4` (`E35-B`):** Under `CIS-35` on the canonical contract bundle (`Arm 5`), false certified releases drop from `4 / 12` (`33.33%`) to `0 / 12` (`0.00%`), while all four Stratum-2 cases (`C05-C08`) achieve `4-0` correct verdict consensus and pass `CIS-35` with `4 / 4` (`C05, C06, C08`) and `3 / 4` (`C07`) verified governing-clause spans across distinct model families.
5. **Empirical live-model finding `F35.5` (`E35-B`):** `CIS-35` on the canonical bundle (`Arm 5`) holds `2 / 12` (`16.67%`) cases fail-closed (`HOLD_CIS35_GATE`) despite `4-0` correct top-level verdict agreement:
6. On `C03`, all four families output `PASS` (`4-0`) and provide verbatim quotes from the contract, but only `W1_B` (plus `W1, W2, W3` in the homogeneous `Witness A` arm) labels `SLA Exhibit B Section 3.4` as `governing_section`, whereas `W0_A`, `W2_C`, and `W3_D` apply `Section 3.4` in their reasoning chain and label the credit-entitlement clause `SLA Section 1.2` as `governing_section`.
7. On `C12`, all four families output `FAIL` (`4-0`) and cite `Section 15.2` with verbatim quotes (`4 / 4`), but `W2_C` and `W3_D` set `missing_cross_reference = true` because `Exhibit C (Named Direct Competitors)` is summarized inline in the Transaction Record rather than attached as a separate exhibit block.

### 9.2 Interpretations and architectural implications

1. **Why model diversity cannot substitute for canonical input-digest and cross-reference verification:** Because instruction-following frontier models are trained to reason strictly from the supplied prompt context, feeding four distinct model families the same truncated RAG excerpt (`C05-C08`) turns instruction fidelity into a `100%` common-mode failure multiplier. `AIC-33` and `GTP-34` certify *who* ran the witness; `CIS-35` certifies *what exact byte digest and clause chain* the witness inspected.
2. **Why `C03` and `C12` demonstrate a real calibration trade-off in structured span verification:** Requiring witnesses to emit a single scalar `governing_section` string creates false-hold friction when a legal conclusion depends on a two-clause composition (an entitlement clause `SLA Section 1.2` plus an exclusion ceiling `SLA Exhibit B Section 3.4`). Splitting the witness schema into `entitlement_section` and `decisive_override_section` is the natural next refinement to eliminate `C03`-style label splits without relaxing override verification.

## 10. Scope limitations and threats to validity

1. **Sample size and evaluation scope:** `E35-B` evaluated `12` contract and policy audit cases across `100` live witness evaluations (`5` evaluation arms). While the `4 / 4` (`16 / 16` witness calls) failure correlation on Stratum 2 is sharp, larger multi-hundred-document legal corpora will exhibit broader paraphrase and cross-reference topologies.
2. **Single-pass non-adversarial witness prompting:** Witnesses in `E35-B` ran as zero-shot structured JSON auditors without multi-turn cross-examination. Multi-turn debate can either repair `C03`-style section-label splits or amplify sycophantic conformity (`[S09]`, `[S12]`).
3. **Orchestration and verification scope:** Encevra builds and evaluates the verification, evidence-grounding, and quorum release-control architecture (`AIC-33`, `GTP-34`, `CIS-35`) rather than training foundation model weights; all live measurements in `E35-B` evaluate witness outputs under deterministic orchestration and verification gates.

## 11. Next check (Round 36 handoff)

Round 36 should test **Compositional Clause-Chain Schemas (`CCS-36`)** on a live matched-compute multi-model benchmark: replacing the single `governing_section` field in `CIS-35` with an ordered `(base_entitlement_section, decisive_override_section, inline_fact_sufficiency)` tuple to test whether the `2 / 12` (`16.67%`) conservative fail-closed holds observed on `C03` and `C12` can be reduced to `0 / 12` while preserving `0.00%` false certified releases on truncated and decoy-injected contracts.

---

## 12. References (`[S01]` - `[S12]`)


- **[S01]** John C. Knight and Nancy G. Leveson (1986). "An Experimental Evaluation of the Assumption of Independence in Multi-Version Programming." *IEEE Transactions on Software Engineering*, SE-12(1):96-109. https://doi.org/10.1109/TSE.1986.6312924
- **[S02]** Dave E. Eckhardt and Larry D. Lee (1985). "A Theoretical Basis for the Analysis of Multiversion Software Subject to Coincident Errors." *IEEE Transactions on Software Engineering*, SE-11(12):1511-1517. https://doi.org/10.1109/TSE.1985.231895
- **[S03]** Bev Littlewood and Douglas R. Miller (1995). "A Conceptual Model of Multi-Version Software." *Twenty-Fifth International Symposium on Fault-Tolerant Computing (FTCS-25 Highlights)*, pp. 150-159. https://doi.org/10.1109/FTCSH.1995.532633
- **[S04]** Dahlia Malkhi and Michael Reiter (1998). "Byzantine Quorum Systems." *Distributed Computing*, 11(4):203-213. https://doi.org/10.1007/s004460050050
- **[S05]** Flavio P. Junqueira and Keith Marzullo (2003). "Synchronous Consensus for Dependent Process Failures." *Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS 2003)*, pp. 274-283. https://doi.org/10.1109/ICDCS.2003.1203476
- **[S06]** Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *International Conference on Learning Representations (ICLR 2023)*. https://arxiv.org/abs/2203.11171
- **[S07]** Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch (2023). "Improving Factuality and Reasoning in Language Models through Multiagent Debate." *arXiv preprint arXiv:2305.14325*. https://arxiv.org/abs/2305.14325
- **[S08]** Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu (2023). "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate." *arXiv preprint arXiv:2305.19118*. https://arxiv.org/abs/2305.19118
- **[S09]** Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou (2023). "Large Language Models Cannot Self-Correct Reasoning Yet." *arXiv preprint arXiv:2310.01798*. https://arxiv.org/abs/2310.01798
- **[S10]** Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye (2024). "More Agents Is All You Need." *Transactions on Machine Learning Research (TMLR 2024)*. https://arxiv.org/abs/2402.05120
- **[S11]** Fan Wu, Emily Black, and Varun Chandrasekaran (2024). "Generative Monoculture in Large Language Models." *arXiv preprint arXiv:2407.02209*. https://arxiv.org/abs/2407.02209
- **[S12]** Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica (2025). "Why Do Multi-Agent LLM Systems Fail?" *arXiv preprint arXiv:2503.13657*. https://arxiv.org/abs/2503.13657