Encevra Research Round 51

Title: Valid Evidence, Wrong Verdict: Transfer Tests on Code Review and Raw Table Claims, with Independent Recovery

Round code: XFR-51

Creator: Wilkins, Benjamin. Encevra. ORCID: 0009-0005-6532-3200

Date: November 8, 2026

Status: All planned experimental calls attempted once. One reader output was empty and was not repeated. The first synthesis call returned no text; a second synthesis call completed and both are disclosed. One authored label was found wrong after the live reads and both scorings are reported. Prepared for review, not published by this run. DOI pending reservation.

Abstract

The Theory of Collaborative Intelligence roadmap sets Round 51 as a third, clearly different task: not another document and clause packet. Propositions that fail transfer cannot enter the broader theory unchanged. This round moves the evidence, agreement and rewrite checks of Rounds 40 to 50 onto two non-document tasks whose gold does not depend on an author's reading: code review against a written specification, with gold fixed by executing a hidden test suite, and numerical claims against raw tables, with gold fixed by recomputation. A third track uses 22 real SEC credit agreement exhibits, none used before, with one numeric covenant claim each.

Structurally, the contract-era limits transfer exactly. An exact evidence-span gate admits all 64 of 64 wrong status mutations of the authored gold ledgers, and a run-time check that executes the code on the specification's own examples admits all 4 subtle code defects as PASS. Across all 81 truth, baseline, rewrite and adjudicator states, preserving the baseline loses 12 correct decisive repair states; adding recovery through an independent reader restores four of them and adds eight unsafe states, every one of which requires the rewrite and the independent reader to share the same wrong decisive verdict.

Live, the failure classes built from contract work did not reproduce. On the 32 authored code and table items, Witnesses B, C and D made no decisive error relative to execution or recomputation, caught every subtle defect and every table trap, and held every genuinely open dependency. The validator changed none of 53 natural baseline ledgers, so preservation and recovery had nothing to act on in live data. Two concrete findings remain. First, Witness B read one item labeled HOLD as FAIL and gave a counterexample: the unshown helper receives a floating point value, and two valid inputs with different required answers produce the same value, so no helper can satisfy the specification. A deterministic check confirms it. The authored label, the two-world construction, the blind label auditor and two other readers were wrong; a single dissenting reader was right, and agreement routing would have held its correct answer. Second, Witness C read four SEC compliance claims as FAIL in blind reads but kept HOLD when validating another reader's HOLD ledger for the same items, so a validation call is not an independent read. The wording of those claims makes FAIL defensible, so those four reads are reported as contestable rather than unsafe in any settled sense.

1. Archive check and continuity

The staging registry was read before any live call. Its newest entry is Round 50, ORT-50, so this is Round 51. The audit covered all 44 accessible entries (Rounds 3 and 8 to 50); Rounds 1, 2 and 4 to 7 are not in the accessible archive, so this is not a certification about unavailable history. No accessible round used executable code under review, hidden test suites, raw numerical tables, or covenant threshold claims. Round 16 proposed contrastive scenario pairs and Round 26 characterized observationally identical worlds; here a two-world construction only certifies authored HOLD gold and is not the tested mechanism. Rounds 35, 42 and 50 measured joint error on contract packets; Rounds 40, 43 and 50 tested rewriting validators from seeded ledgers. This round uses natural

baselines and adds a recovery policy that no prior round tested. All 22 SEC accessions and filers are new against the recorded prior set through Round 50. The continuity audit is `anti-repetition-verification.md`.

Round 50 handed off five items. This round takes item 1 (the planned different task, with baseline trust and recovery as testable conditions) and item 3 (route changed statuses to a separate reader with raw evidence). Item 2 is avoided by construction, since code and table gold do not depend on wording. Item 4, a direct replication of the Round 49 near misses, is not attempted and remains open. Item 5 belongs to Round 52.

2. Fixed plan and hypotheses

The plan was fixed before any live call (`run-plan-r51.md`). A verdict is unsafe when it is PASS or FAIL and differs from gold, and a false hold is HOLD on a PASS or FAIL item.

ID	Fixed statement
H51.1	Structural: in both non-document domains an exact evidence-span gate admits every wrong status mutation, and a visible-example execution gate admits every subtle code defect as PASS
H51.2	In each non-document domain at least one live decisive wrong read cites only exact packet evidence
H51.3	In each non-document domain at least one reader returns a decisive verdict on an open-dependency HOLD item
H51.4	Agreement routing of Witnesses C and D yields no more unsafe outcomes than the better single reader in each domain, with a false hold cost
H51.5	With natural baselines, preservation loses at least one correct decisive repair, and independent recovery restores at least one without admitting a wrong decisive change
H51.6	Structural: under recovery, every unsafe state that preservation would have blocked has the rewrite and the independent reader sharing the same wrong decisive verdict

Related work

Code generation benchmarks score programs by executing unit tests [S99], which makes executable gold possible but also shows that a passing test set is only as strong as its coverage. Claim verification against evidence has standard formulations for text [S100] and for tables [S101]. Self-correction without external feedback can fail to improve, or degrade, answers [S102], which motivates separating a validator's rewrite from an independent read. Dependent failures among separately developed programs [S98] motivate measuring, rather than assuming, the independence of readers.

3. Materials and gold

Each code family has a written specification with three shown examples and four implementations: P correct; S a subtle defect that passes every shown example; O a defect that fails a shown example; H correct-looking code that imports an unshown helper. Each table family has raw rows, a computation note and four claims: P true; S false but true under one specific misreading; O false against a single cell; H undetermined because rows are declared missing. The packet convention is an open record: anything not shown is unknown.

Family	Domain	Task	S variant built around
X01	Code review	Whole-number percentage share (`percent_share`)	exact halves rounded to even by the built-in rounding

Family	Domain	Task	S variant built around
X02	Code review	Run counting (`count_runs`)	the empty string
X03	Code review	Median of a list (`median_of`)	even-length lists
X04	Code review	Fixed-size chunking (`chunk`)	a final partial chunk
T01	Raw tables	Greenhouse trial heights	rows the note excludes
T02	Raw tables	Clinic waiting load	mean of weekly ratios instead of a ratio of totals
T03	Raw tables	Branch library late returns	a rate exactly at the threshold
T04	Raw tables	Building energy meters	a row recorded in MWh

Gold for P, S and O code items is fixed by running a hidden suite (7 to 10 cases per family); every P passes it and every S and O fails it. Gold for table claims is recomputed from the rows; each S claim is verified to be true under its named misreading. Gold HOLD for H items was fixed by a two-world construction: one helper or one completion of the missing rows makes the claim true, another makes it false. For tables the completions are concrete rows. For code, the two worlds were checked against the hidden suite only, which section 8 shows was insufficient for one family.

The SEC track uses 22 real credit agreement exhibits from EDGAR full-text search, fetched with a declared User-Agent, with whitespace normalization only. Each window ends with one covenant sentence that sets a single ratio threshold. Claims state the correct threshold and direction (6, PASS), a shifted number (5, FAIL), a reversed maximum or minimum (5, FAIL), or that the window establishes actual compliance at the latest test date (6, HOLD). Windows can contain other numbers as distractors.

Item	Filer	Exhibit	Filing date	Accession	Claim type	Source
S01	Necessity Retail REIT, Inc.	EX-10.1	2023-05-10	0001568162-23-000016	correct threshold	exhibit
S02	O REILLY AUTOMOTIVE INC	EX-10.1	2025-04-01	0001104659-25-030592	reversed direction	exhibit
S03	GENTEX CORP	EX-10.14	2023-02-22	0000355811-23-000011	correct threshold	exhibit
S04	Global Medical REIT Inc.	EX-10.1	2025-10-08	0001104659-25-097935	correct threshold	exhibit
S05	TUPPERWARE BRANDS CORP	EX-10.1	2023-05-08	0001008654-23-000035	correct threshold	exhibit
S06	BWX Technologies, Inc.	EX-10.1	2026-11-03	0001104659-26-121983	shifted number	exhibit
S07	VIAD CORP	EX-10.1	2024-08-07	0000950170-24-092988	correct threshold	exhibit
S08	Noble Corp plc	EX-10.1	2026-06-01	0001895262-26-000129	actual compliance	exhibit
S09	Neutron Holdings, Inc.	EX-10.1	2026-07-06	0001628280-26-047289	shifted number	exhibit
S10	Leonardo DRS, Inc.	EX-10.1	2026-01-28	0001833756-26-000007	reversed direction	exhibit
S11	Sleep Number Corp	EX-10.4	2023-05-09	0000827187-23-000045	actual compliance	exhibit

Item	Filer	Exhibit	Filing date	Accession	Claim type	Source
S12	Principal Credit Real Estate Income Trust	EX-10.2	2025-08-11	0002026448-25-000067	reversed direction	exhibit
S13	COFFEE HOLDING CO INC	EX-10.28	2024-02-09	0001493152-24-005726	shifted number	exhibit
S14	Ares Commercial Real Estate Corp	EX-10.4	2024-08-06	0001628280-24-034916	reversed direction	exhibit
S15	BGSF, INC.	EX-10	2024-11-07	0001474903-24-000089	reversed direction	exhibit
S16	Cars.com Inc.	EX-10.1	2024-05-09	0000950170-24-056894	shifted number	exhibit
S17	Knowles Corp	EX-10.1	2023-02-09	0001587523-23-000019	actual compliance	exhibit
S18	Jamf Holding Corp.	EX-10.1	2025-05-27	0001628280-25-027919	actual compliance	exhibit
S19	AvePoint, Inc.	EX-10.1	2026-11-05	0001437749-26-035271	correct threshold	exhibit
S20	TransUnion	EX-10.1	2024-04-25	0001552033-24-000040	actual compliance	exhibit
S21	Evercore Inc.	EX-10.4	2025-11-05	0001360901-25-000025	actual compliance	exhibit
S22	PayPal Holdings, Inc.	EX-10.1	2023-06-13	0001193125-23-165556	shifted number	exhibit

4. Readers, ledgers and policies

Witness A blind-labeled every item once and flagged wording that admits more than one label; it is a validity audit, not gold. Witnesses B, C and D each read every item in a fresh call with no ledger and no other reader, returning a verdict and one to three exact evidence strings. Witness D's own read then became a natural starting ledger. In a separate fresh call, Witness C validated and could rewrite that ledger. Three policies are replayed on the recorded outputs: accept the rewrite; preserve the baseline, routing any status change to HOLD; and preserve with recovery, where a changed status survives only if Witness B's separate blind read gives the same decisive verdict. These are call-output replays, not a live end-to-end orchestration comparison, and they are not matched call baselines. Item order was shuffled with seed 51.

5. Deterministic checks

Structural findings

• Exact evidence-span validity admits 64 of 64 wrong status mutations of the 32 authored gold ledgers, in both non-document domains. Finding a real line of code or a real table row does not certify the verdict drawn from it.

• Executing each implementation on the specification's own examples admits all 4 S implementations as PASS and routes all 4 O implementations. A cheap executable check behaves like a quote check: it confirms what was shown, not what was specified.

• All four table S claims flip from false to true under their named misreading, so each is a near miss of one specific error.

• Across 81 states:

Policy	Unsafe states of 81	Lost correct decisive repair states
Accept the rewrite	36	0
Preserve the baseline	12	12
Preserve with independent recovery	20	8

Preservation blocks unsafe changes but loses every state where the rewrite is a correct decisive repair of a wrong baseline. Recovery restores four of those twelve states, exactly where the independent reader agrees with the correct repair, and adds eight unsafe states, each of which has the rewrite and the independent reader sharing the same wrong decisive verdict. The trade is set by joint error between rewriter and adjudicator, not by either alone. The twelve unsafe states common to all three policies are unchanged wrong baselines, which no change-based gate can see.

6. Empirical reader results

Witness	Domain	Correct (fixed labels)	Unsafe	False holds	Correct (corrected gold)	Usable
Witness B	Code review	15	1	0	16	16 of 16
Witness B	Raw tables	16	0	0	16	16 of 16
Witness B	SEC covenant claims	22	0	0	22	22 of 22
Witness B	All	53	1	0	54	54 of 54
Witness C	Code review	16	0	0	15	16 of 16
Witness C	Raw tables	16	0	0	16	16 of 16
Witness C	SEC covenant claims	18	4	0	18	22 of 22
Witness C	All	50	4	0	49	54 of 54
Witness D	Code review	15	0	0	15	15 of 16
Witness D	Raw tables	16	0	0	16	16 of 16
Witness D	SEC covenant claims	22	0	0	22	22 of 22
Witness D	All	53	0	0	53	53 of 54

Fixed labels are the labels set before any live call. Corrected gold changes one item, X01H, from HOLD to FAIL (section 8). Witness D's one empty output fell on X01H and is excluded, not scored as HOLD.

On the 32 authored code and table items, no reader made a decisive error against execution or recomputation. Every reader returned FAIL on all four subtle code defects that pass the shown examples, rejected all four table traps (for example, computing 1850 over 60 visits as 30.83 minutes rather than averaging weekly ratios), and returned HOLD on every genuinely open table dependency and on the three code dependencies that are genuinely open. Decisive reads differing from the fixed labels with exact evidence: Witness B on X01H (correct under corrected gold) and Witness C on S08, S11, S18 and S20. In the SEC track all three readers were correct on every threshold, shifted number and reversed direction claim; only the compliance wording produced disagreement.

Pair	Domain	Complete pairs	Routed correct	Routed unsafe	Routed false holds	Disagreements
Witnesses C and D	Code review	15	15	0	0	0
Witnesses C and D	Raw tables	16	16	0	0	0
Witnesses C and D	SEC covenant claims	22	22	0	0	4
Witnesses B and C	Code review	16	16	0	0	1
Witnesses B and C	Raw tables	16	16	0	0	0
Witnesses B and C	SEC covenant claims	22	22	0	0	4
Witnesses B and D	Code review	15	15	0	0	0
Witnesses B and D	Raw tables	16	16	0	0	0
Witnesses B and D	SEC covenant claims	22	22	0	0	0

No pair shared a wrong decisive verdict on any item. Routing therefore never released an unsafe outcome, but it also had no unsafe single reads to stop in the non-document domains. Under corrected gold, the Witness B and Witness C pair on code routes X01H to HOLD: agreement routing turns the one correct dissent into a false hold.

7. Natural baselines and rewrite policies

Domain	Usable rewrites	Status changes	Baseline correct	Accept	Span gate	Preserve	Recover
Code review	15	0	15	15	15	15	15
Raw tables	16	0	16	16	16	16	16
SEC covenant claims	22	0	22	22	22	22	22
All	53	0	53	53	53	53	53

Witness C changed the status of none of the 53 usable baseline ledgers, and every baseline was already correct under fixed labels. Accept, span gate, preserve and recover are identical on live data, so this round provides no live evidence about preservation loss or recovery benefit. The structural table in section 5 is the only evidence on those policies here.

One live pattern does bear on independence. On S08, S11, S18 and S20, Witness C's blind read was FAIL. When the same witness validated Witness D's HOLD ledger for those items, it kept HOLD in all four, with reasons such as the covenant "does not provide actual reported metrics demonstrating compliance". The validator's output depended on the proposed ledger, so a validation call cannot be counted as a second independent read.

8. A label error found by one reader

Item X01H asks whether this implementation meets the percentage specification for every valid input:

```
from shares import round_half_up
def percent_share(part, whole):
```

```
    return round_half_up(100 * part / whole)
```

The fixed label was HOLD, because the helper is not shown, and the two-world construction found one helper that passes the hidden suite and one that fails it. Witness B returned FAIL and gave a counterexample. With whole equal to 8 times 10 to the 18th power, part equal to 10 to the 18th requires 13, and part one less requires 12. Both divisions round to the same floating point value, 12.5, before the helper sees them, so any helper must return the same answer for both, and at least one is wrong. A deterministic check in the archive confirms the collision (true). The defect is in the call, not in the unshown helper, so FAIL is established from the packet. The two-world check was sound only on the hidden suite, not on every input in the stated domain.

Witness A's blind audit, Witness C and the authored label all said HOLD; Witness D's output for this item was empty. The other three code H items pass the helper an exact integer, string or list, and their passing world computes the same result as the correct implementation, so their HOLD labels stand. The fixed label is not overwritten: both scorings are reported, and the rule for which counts was not fixed in advance.

9. Findings versus interpretations

Empirical findings

• On 32 authored code and table items, Witnesses B, C and D made no decisive error against execution or recomputation, under either scoring.

• Witness B identified a wrong authored HOLD label with a correct counterexample; Witness A's blind label audit agreed with all 54 fixed labels, including the wrong one.

• Witness C read four of six SEC compliance claims as FAIL against fixed HOLD labels, with exact quotes; Witness A had flagged that wording as admitting FAIL.

• Witness C changed none of 53 natural baselines and, on four items, kept a HOLD baseline that contradicted its own blind read.

• One reader output and the first synthesis call were empty; neither was scored as a verdict.

Interpretation

The structural limits of span, example execution and preservation are properties of the gates, so they transfer to any domain by construction. The live failure classes from contract rounds, open record reads and near miss traps, did not appear on these non-document items, so this round cannot say those failures transfer; it can say they are not universal across tasks for these readers. The items were too easy to separate policies: ceiling performance is a design limitation, not evidence that routing or recovery is unnecessary. The X01H case is a concrete example of the theory's central caution about agreement: unanimity among the label author, the auditor and two readers was wrong, and the useful signal was a dissent with a checkable counterexample. A routing rule that only lets a second read stop a release cannot use that signal; a rule that lets a checkable counterexample override labels could, but this round did not test one.

10. Hypothesis outcomes

ID	Outcome	Basis
H51.1	Supported	64 of 64 wrong mutations pass the span gate; 4 of 4 subtle defects pass example execution
H51.2	Not supported	No decisive wrong read in tables; the only decisive code read differing from labels is correct under corrected gold

ID	Outcome	Basis
H51.3	Not supported	No decisive verdict on a genuinely open dependency in either non-document domain; X01H was not genuinely open
H51.4	Inconclusive	Routing released no unsafe outcome, but single readers had none to stop in code or tables; the false hold cost appeared only under corrected gold (one item)
H51.5	Inconclusive	Zero of 53 rewrites changed status, so neither preservation loss nor recovery could occur live
H51.6	Supported	Recovery adds 8 unsafe states over preservation, all with a shared wrong decisive verdict, and restores 4 of 12 lost repair states

11. Synthesis and analysis audit

Witness A reviewed a summary of the completed analysis, after the fact, as commentary rather than gold. The first call returned no text and is preserved verbatim; a second call with a larger output allowance and an instruction to answer directly completed. Its assessment: "Structural: 54 items (16 code, 16 table, 22 SEC), labels fixed before calls, 64 mutations and 81 constructed recovery states. The policy counts (unsafe 36/12/20, lost repairs 0/12/8) describe constructed states, not live reads. Empirical: fixed labels give B 53/54, C 50/54, D 53/53 with X01H empty; corrected gold gives B 54/54, C 49/54; all pairs route 100 percent; rewrites changed 0 of 53, so accept, preserve and recover are identical by construction and the live data say nothing about policy differences. Interpretation and overclaims: X01H gold was changed after seeing B's read; the collision check and x01h_no_helper_can_pass support FAIL, but that conflicts with code_two_world_ok, which asserts a passing world existed for every H item. The five live unsafe reads (B on X01H, C on S08, S11, S18, S20) are FAIL given to HOLD items; FAIL is the conservative direction and the SEC wording admits FAIL, so these are label and wording defects, not reader failures, and the fixed label unsafe counts overstate risk. C kept HOLD when validating D ledgers on items it independently read FAIL, so validation calls anchor on the ledger and pair agreement is not independent confirmation. Blind review agreed 54/54 on labels including the later overturned X01H; agreement did not establish truth."

Its caveats were: (1) D's perfect score excludes its one empty call, which fell on the disputed item X01H; a missing read is not a correct read. (2) B, C and D may share a model and prompts; separate calls are not independent evidence and the pair tables cannot be read as replication. (3) Both X01H scorings are shown, but the rule for which one counts was not fixed in advance; report both, not the more favorable one. (4) Zero rewrite changes mean recover and preserve margins (lost repairs 8 and 12) rest entirely on constructed states and are untested live. The assessment's description of the B, C and D unsafe reads as conservative is not adopted, since FAIL is a decisive verdict under the convention. Its remark that witnesses may share underlying systems is not something the published record can settle; the witnesses are separate configurations, and independence is in any case measured here, not assumed. Its first next check is answered in section 8.

12. Completion disclosure and limitations

Every planned experimental call was attempted once: one blind label audit, 162 independent reads and 53 validations (one per usable baseline). One independent read (Witness D on X01H) was empty and not repeated. No call was stopped by the spend guard and no call errored. The first synthesis call was empty; the second synthesis call used a larger output allowance and one added sentence, disclosed here. No packet, claim or label was edited after outcomes were observed; the X01H correction is reported beside the fixed label.

Limitations: the authored pool is small (32 items) and proved too easy for these readers; four policies coincide live; the SEC compliance claims were worded so that FAIL is defensible, a flaw Witness A flagged before scoring; code

gold relies on hidden suites, which section 8 shows can miss a whole-domain defect; numeric representation at extreme magnitudes is a residual source of disagreement for other items, although all readers judged those items alike; one reader output is missing on the one disputed item. Results are about these readers, prompts and items, not about any reader class.

13. Next checks for the theory program

1. Round 52 must compare matched call baselines, and should use items hard enough that single readers make decisive errors in each domain; this round shows that ceiling items cannot rank routing, preservation or recovery.

2. Reword the SEC compliance claims so that HOLD is the only defensible reading, and keep wording flagged by the blind audit out of scored sets.

3. Test a dissent rule in which a decisive minority verdict with an executable counterexample is checked rather than vetoed.

4. Measure validator anchoring directly by seeding wrong decisive ledgers on items where the validator's blind read is known.

5. The direct replication of the Round 49 near misses remains open.

14. Reproducibility

The archive includes the fixed plan, continuity audit, deviation record, material builder with all specifications, implementations, hidden suites, tables and two-world completions, the SEC track with exact windows, accessions and links, source verification, the live runner, verbatim transcripts, the blind audit and synthesis, replay analysis, the exact verifier, the chart script and this manuscript with its PDF builder. The verifier makes no network calls and checks 0 conditions. The private execution configuration is not included and is not required to replay the recorded evidence.

15. Verified references

• [S98] John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. IEEE Transactions on Software Engineering SE-12(1), 96-109, 1986. doi:10.1109/TSE.1986.6312924.

• [S99] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda and 50 further authors. Evaluating Large Language Models Trained on Code. arXiv:2107.03374, first posted July 7, 2021. https://arxiv.org/abs/2107.03374

• [S100] James Thorne, Andreas Vlachos, Christos Christodoulopoulos and Arpit Mittal. FEVER: a large-scale dataset for Fact Extraction and VERification. arXiv:1803.05355, first posted March 14, 2018. https://arxiv.org/abs/1803.05355

• [S101] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou and William Yang Wang. TabFact: A Large-scale Dataset for Table-based Fact Verification. arXiv:1909.02164, first posted September 5, 2019. https://arxiv.org/abs/1909.02164

• [S102] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798, first posted October 3, 2023. https://arxiv.org/abs/2310.01798

• [S30] U.S. Securities and Exchange Commission. EDGAR Archives. https://www.sec.gov/Archives/edgar/data/. The 22 exhibits in section 3 were opened and verified for this run.