Encevra Research Round 43: Open Record or Closed Record? Declared Record Conventions, Near Miss Records Read as Negation, and Flag Only Against Rewriting Validators

Date: 2026-11-01
Research lab: Encevra Research
Author: Ben Wilkins, Encevra (ORCID 0009-0005-6532-3200)
Short code: `RCV-43` (Record Convention and Validators)
Primary tracks: Reasoning and Verification ("How does a system know when an answer is ready to release?"), Collaborative track ("When does dividing reasoning across specialists help, and when does it fragment evidence?")
Position in the program: third round of the Round 41 to 50 "Theory v1" block. The roadmap item for Round 43 is flag only validators against rewriting validators (proposition P5, stages that can rewrite evidence add risk).
Carry-forward boundary: Round 42 (`RDX-42`) found that direct readers read a missing record as proof that an event did not happen, proposed P7 (absence is not negation), and named the record convention as its central threat. Its handoff asked for three tests: the direct role under a declared open record and a declared closed record; extraction readers with a nonzero unsafe marginal, to see whether one witness's unsafe releases coincide across roles; and flag only validators on items whose correct flags are known. Round 43 runs those three tests. It does not rerun the Round 42 pair design, the Round 41 selection study, or the Round 40 topology comparison.

Abstract

Round 42 left two readings of the same data. Either direct readers misread absence as negation, or they silently applied a closed record convention under which FAIL is correct. Round 43 separates the two by telling readers which convention applies. Six new authored contract families are each written in three variants that share every record but the last. In U the final record only suggests the decisive fact (a request, a calendar entry, a production report). In K it establishes the fact. In T it is a near miss: a record of the right kind that fails one stated requirement (approval by a lender rather than the agent, notice by email when email is ineffective, consent signed by an excluded signatory, a sale to a subsidiary, reviewed rather than audited statements, a notice citing the wrong agreement number). Under an open record, U and T are HOLD, since the packet neither establishes nor excludes a valid event. Under a closed record, U and T are FAIL. K is the same under both. Witnesses B, C, and D answered every item under a declared open record and a declared closed record, answered U and T items with the undeclared Round 42 prompt, and filled one evidence ledger per item that deterministic code resolves under both conventions. A second experiment puts the roadmap's validator question to 24 new public SEC exhibits, each with a gold evidence ledger and a seeded ledger carrying one known defect, read by a flag only validator and a rewriting validator.

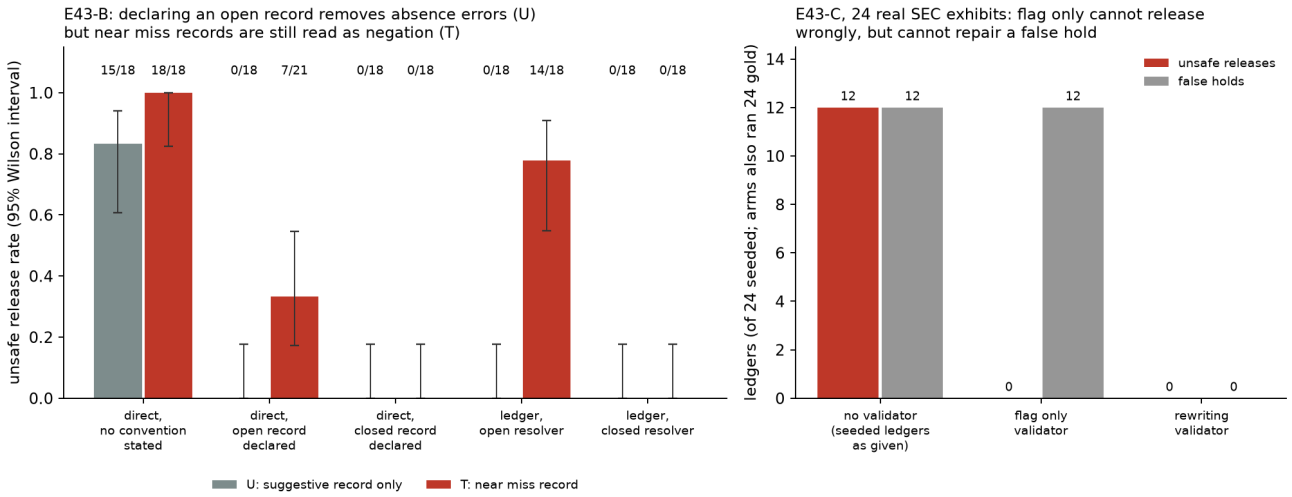
We report three strictly separated evidence categories.

Deterministic (`E43-A`, no model calls). All seven certificates pass. Every family's three variants differ only in the final record; a gold ledger reproduces all 18 labels under both conventions; one extraction ledger yields both conventions' answers through code, while a direct reader needs a separate call per convention; all 24 exhibits are new against the 988 accessions and 721 filers in the exhibit tracks and candidate pools of Rounds 39 to 42. A flag only validator can never create an unsafe release: across all 96 flag outcomes it can only hold or leave the ledger's resolution unchanged. A rewriting validator has no such bound: a rewrite that passes every span check turns each of the 12 gold HOLD ledgers into an unsafe release.

Live authored (`E43-B`, 6 families, 18 items, 201 calls). Without a declared convention, direct readers released FAIL on `33` of `36` U and T decisions (Witness B 11, C 10, D 12 of 12). Declaring an open record removed every absence error: U went from `15` of `18` unsafe to `0` of `18`. It did not remove the near miss errors: T went from `18` of `18` to `6` of `18` (Witness A, three items only: `1` of `3`), and every remaining error sat on two families, `S03T` (consent signed by the property manager, which the lease excludes) and `S06T` (a price notice citing the wrong agreement number), plus Witness C on `S02T`. Declaring a closed record was followed exactly: `36` of `36` correct FAIL, and K was `18` of `18` under both declarations. Extraction readers held all U items under the open resolver but released FAIL on `14` of `18` T items: they marked the requirement UNMET, quoting the defective record. Under the closed resolver extraction was `35` of `36` correct on U and T with no unsafe release. Once both roles could err, the same witness's unsafe releases coincided: `5` of the `6` direct unsafe releases under the open declaration were matched by that witness's own extraction.

Live real exhibits (`E43-C`, 24 SEC EDGAR exhibits, 96 calls, Witness C). The flag only arm made no unsafe release in `48` ledgers, caught all `12` seeded omissions, and left all `12` seeded spurious conditions as false holds, since a flag can stop a release but cannot restore one. The rewriting arm was correct on all `48`: it restored all `12` omitted conditions, removed all `12` spurious ones, and left every gold ledger's resolution unchanged.

Interpretation, kept separate. Declaring the convention is enough to stop readers treating a missing record as a negative, so the Round 42 result was largely a convention effect. It is not enough to stop them treating a defective record as a negative: under an open record, a record that fails a requirement does not exclude a valid one, yet readers in both roles read the near miss as refutation. On these items the extraction role made that error more often than the direct role, and the role pairs that were safe in Round 42 failed together. For validators, the live data favour rewriting, and the structure favours flagging: rewriting repaired everything it was given here, but nothing in its design bounds the harm a wrong rewrite can do, while a flag cannot release anything. We revise P4, P5, and P7 accordingly and propose that rewrites be accepted only when they move a decision toward HOLD.



1. Archive continuity and anti-repetition verification

Before designing any Round 43 experiment we rebuilt the archive ledger from the live research registry (`prior-rounds-ledger.json`, `36` indexed rounds, `168` registry hypotheses). A keyword scan of titles, questions, summaries, findings, open questions, and hypotheses for the Round 43 constructs (record convention, open record, closed record, closed world, closure, flag only, rewrite, validator, seeded ledger, seeded defect, near miss, negation, absence, self correction) returned:

- "closed record", "negation", and "absence" only in Round 42, whose open question this round answers;
- "flag only" and "validator" only in Round 40, whose finding on `X24` (a validator in the role pipeline rewrote a correct ledger by adding five covenants as conditions) and open question ("do flag only validators avoid the errors that rewriting validators introduce?") this round tests, and once in Round 13, where "validator" names an offline trace validator for score records;
- "rewrite" only in Rounds 20 and 23, where it means a privileged operator rewriting a log or a release rewriting historical findings, not a validator rewriting an evidence ledger;
- "closure" in Rounds 16, 20, 22, and 36 to 38, where it means premature answer closure (Round 16), cross document reference closure (Round 36), or register anchored completeness (Rounds 37 and 38). Round 37 is the closest: it gave witnesses a register that anchors which later documents exist. It did not declare whether the packet is a complete record, and it did not contrast an open declaration with a closed one on fixed items.

"Record convention", "open record", "closed world", "seeded ledger", "near miss", and "self correction" returned no hits.

Prior round	Focus	How Round 43 differs
`RDX-42` (Round 42)	One witness in two roles against two witnesses in one role; absence read as negation under an unstated convention	Round 43 states the convention to the reader, tests both conventions on the same items, and adds near miss records so that extraction can err unsafely

Prior round	Focus	How Round 43 differs
`HPS-41` (Round 41)	Disagreement mined pools and twin holdouts	No selection; every item and exhibit fixed in advance
`NCT-40` (Round 40)	Six composition topologies; one rewriting validator error on `X24`	Round 43 compares a flag only and a rewriting validator head to head on ledgers whose correct flags are known by construction
`TDR-37` (Round 37)	Register anchored completeness for later documents	Round 43 does not supply a register; it declares the record open or closed and measures compliance

No packet, exhibit, or prompt task is reused as a test of the same question. The undeclared direct prompt and the extraction prompt are the Round 42 prompts, unchanged; the two declarations are single added paragraphs. All 24 exhibits come from new EDGAR full text searches and are checked against every exhibit track and candidate pool of Rounds 39 to 42.

---

2. Where this round sits in the program

Round 40 proposed P1 to P5 and a path for Rounds 41 to 50; Round 41 added P6; Round 42 revised P4 (decorrelate by role, then verify by structure) and proposed P7 (absence is not negation). Round 43 carries two removal tests from Round 42 (P7: a declared convention with the FAIL releases unchanged; P4: an extraction role with a nonzero unsafe marginal whose errors coincide with the direct role's) and the Round 40 removal test for P5 (validators restricted to flagging).

---

3. Design

3.1 Record conventions

An **open record** is the evidence put before the gate, not a complete register: a fact the packet neither establishes nor excludes is undetermined, and an undetermined decisive fact requires HOLD. A **closed record** is the complete register of every notice, consent, delivery, payment, and approval relating to the matter: an event the packet does not record did not occur. This is the open and closed world distinction of deductive databases [S49], and the closed reading is negation as failure [S50]. Round 42 used the open convention for its labels without telling readers. Round 43 uses three direct framings:

- `D0`: the Round 42 direct prompt, unchanged, no convention stated (U and T items only);
- `DO`: `D0` plus one paragraph declaring an open record;
- `DC`: `D0` plus one paragraph declaring a closed record.

`DO` is scored against open labels and `DC` against closed labels. `D0` is scored against open labels, as in Round 42.

3.2 Families and variants

Each family is one rule program rendered three times. The variants share the contract and every record except the final one, and the audit question is identical. The build asserts both.

Family	Requirement	U: final record	K: final record (label)	T: near miss final record
`S01`	Agent's written approval of a replacement chief executive	Borrower's request to the Agent	Agent's approval letter (PASS)	Approval email from a Lender, not the Agent
`S02`	Objection notice received within 45 days, by hand or national courier; email ineffective	Counsel's calendar entry to send the notice	Courier delivery signed for on May 22, after the window closed May 16 (FAIL)	The notice sent by email on May 12, acknowledged as received
`S03`	Sublease consent signed by a Landlord officer at Vice President or above; property manager excluded	Tenant's email asking the property manager to route the request	Consent signed by a Vice President (PASS)	Consent signed by the Property Manager
`S04`	First Commercial Sale to a third party	Quality release of the first lot	Invoice to an unaffiliated distributor (PASS)	Invoice to a wholly owned subsidiary
`S05`	Guaranty release on audited statements received by April 30	Email that the audit committee will meet	Audited statements logged May 6 (FAIL)	Reviewed, not audited, statements logged on time
`S06`	Price increase notice identifying Supply Agreement SA-2291, at least 60 days ahead	Sales deck announcing new list prices	Notice citing SA-2291 received July 18 (PASS)	Notice citing SA-1874 received July 18

Labels: open U and T HOLD; closed U and T FAIL; K as shown under both. The T variants test what U cannot: a record exists, it is the right kind of record, and it fails. Under the open convention it shows that this record does not satisfy the requirement, not that no satisfying record exists.

3.3 Readers and resolvers

Witnesses B, C, and D answered every item under `D0` and `DC` and every U and T item under `D0`. Each also filled one evidence ledger per item with the Round 41 and 42 extraction prompt (conditions, verbatim quotes, and a status of MET, UNMET, or UNDETERMINED for each). Deterministic code resolves each ledger twice. The **open resolver** is the Round 42 resolver: any unverifiable quote or any UNDETERMINED status gives HOLD. The **closed resolver** is the same except that a verified UNDETERMINED status is read as UNMET. Every call used temperature `0` and was made once.

3.4 Validators on real exhibits

For each of the 24 exhibits we built two ledgers in the Round 40 exhibit format (conditions to the effectiveness of the principal operative effect, each with a verbatim quote, and an "effective on its date" quote). The **gold ledger** resolves to the exhibit's label. The **seeded ledger** carries one known defect. For a HOLD exhibit the conditions are omitted and a fragment of the effectiveness sentence is given as the effectiveness quote, so the seeded ledger releases unsafely. For a CERTIFY exhibit a spurious condition (a quote from the window that is not a condition to effectiveness, such as a merger agreement's own closing conditions or a counterparts clause) is added, so the seeded ledger holds falsely. All seeded spans are verbatim, so every defect passes the span check and only a reader of the window can find it.

The **flag only arm** returns flags and may not change the ledger; any flag, or an unparsed reply, gives HOLD; otherwise the original ledger is resolved. The **rewriting arm** returns a corrected ledger, which is span checked and resolved. Both arms ran on all 48 ledgers with Witness C.

3.5 Run order and budget

The phases ran in a fixed order: declared conventions, extraction, validators with Witness C, undeclared baseline, Witness A on the six T items under `DO`, and a repeat of the validator experiment with Witness B. The order after the first phase was fixed from that phase's measured budget use per call and before any later call. A budget rule fixed before the run skips every later call once the run budget is reached. It stopped the Witness A phase after three of six items (`S01T`, `S02T`, `S03T`; `S04T` to `S06T` not run), skipped the Witness B validator repeat entirely, and stopped the Witness A synthesis review (Section 12). Every skipped call is recorded in the transcripts with its status. No item, reader, or phase was selected on outcomes.

---

4. Related work

The convention question is old in database theory. Reiter distinguished open world query evaluation, where only provable answers count, from closed world evaluation, where failure to prove a positive fact licenses its negation [S49]; Clark gave negation as failure its semantics through the completed database [S50]. An audit packet does not carry its own convention, and Round 42 showed that readers default to the closed one. Round 43 states it. Abstention follows the survey of Wen et al. [S24], and contrast variants follow Gardner et al. [S40].

The validator question belongs to the self correction literature. Huang et al. found that models struggle to correct their own reasoning without external feedback and can degrade after trying [S51]. Kamoi et al. concluded that correction works when feedback is reliable and external, and that the bottleneck is generating accurate feedback [S52]. Chain of Verification separates answering from verifying [S48]. Here the external feedback is partly structural (span checks and a fixed resolver), and the validator is a different witness from the one whose ledger it checks, which is closer to the reliable feedback setting. Software reliability work on coincident failure [S43, S44, S45] and measured error correlation across models [S41, S42] frame the cross role coincidence analysis; diversity measures follow Kuncheva and Whitaker [S46], and zero event limits use the exact binomial limit [S47].

---

5. Hypotheses

These hypotheses restate the removal tests that Rounds 40 and 42 published before this round. They were written into this manuscript after the design was fixed and are not a separate preregistration.

- **H43.1 (P7).** Declaring an open record removes the direct FAIL releases on U and T items.
- **H43.2.** Under a declared closed record, direct readers answer FAIL on U and T items and keep K unchanged.

- **H43.3 (P4).** Once extraction readers can err unsafely, a witness's unsafe releases coincide across its two roles.
- **H43.4 (P5).** A flag only validator creates no unsafe release; a rewriting validator can.
- **H43.5.** A flag only validator pays for its safety in false holds that it cannot repair.

---

6. Evidence category 1: deterministic diagnostic (`E43-A`)

Certificate	Result
S1 contrast: U, K, T share the contract and all records but the final one; same question; labels follow the conventions	6 of 6 families pass
S2 gold ledgers: a verbatim gold ledger reproduces each label through the open and the closed resolver	18 of 18 items, both conventions
S3 one ledger, two conventions: on every recorded extraction ledger the two resolvers agree except where the open resolver holds for an undetermined condition	54 of 54 ledgers; the convention changed the outcome of 23; a direct reader needs 2 calls per item for both conventions, an extraction reader 1
S4 exhibits: anchors present, balanced, new	24 of 24; 12 CERTIFY, 12 HOLD; 0 overlap with the 988 accessions and 721 filers in the tracks and candidate pools of Rounds 39 to 42
S5 exhibit ledgers	gold 24 of 24 correct; seeded give 12 unsafe releases and 12 false holds; all spans verified
S6 validator monotonicity	flag only: 0 unsafe releases possible across 96 enumerated outcomes, and 0 false holds repairable; rewriting: a span valid rewrite exists that turns each of the 12 gold HOLD ledgers into an unsafe release
S7 resolution	zero events allow a true rate of 0.632 at n = 3, 0.221 at 12, 0.153 at 18, 0.080 at 36, and 0.061 at 48 (exact one sided 95 percent limit)

S6 is the structural content of P5. A flag only stage is monotone toward HOLD: whatever it says, the decision is the original ledger's decision or HOLD. A rewriting stage is not monotone: the span check guarantees that its quotes exist, not that its ledger is complete, and deleting a verified condition passes every check.

---

7. Evidence category 2: live authored families (`E43-B`)

7.1 Protocol

108 declared convention calls (18 items, 3 witnesses, 2 declarations), 54 extraction calls, 36 undeclared baseline calls on U and T items, and 3 Witness A calls under `DO` on T items, all completed. Three further Witness A calls were stopped by the budget rule. Labels were computed by the family rule functions before any call.

7.2 Direct readers by framing

Framing (scored against)	U unsafe	T unsafe	U and T unsafe (95% interval)	K correct
`D0` undeclared (open labels)	15 / 18	18 / 18	33 / 36 (0.78 to 0.97)	not run
`DO` open declared (open labels)	0 / 18	6 / 18	6 / 36 (0.08 to 0.32)	18 / 18
`DO`, Witness A, T items only		1 / 3		
`DC` closed declared (closed labels)	0 / 18	0 / 18	0 / 36 (0 to 0.10)	18 / 18

Per witness under `D0`: Witness B 11, C 10, D 12 unsafe of 12. Under `DO`: Witnesses B, C, and D each 2 of 12, all on T. Paired by item and witness, `D0` to `DO` moved 27 decisions from FAIL to HOLD, left 6 at FAIL, and left 3 at HOLD; none moved toward a release. Every unsafe direct release in this round was FAIL; no reader released PASS on any U or T item under any framing.

The six `DO` errors by Witnesses B, C, and D are `S03T` (all three), `S06T` (B and D), and `S02T` (C). Witness A held `S01T` and `S02T` and released FAIL on `S03T`.

7.3 Extraction readers under both resolvers

Resolver	U unsafe	T unsafe	K correct / false hold / unsafe
Open	0 / 18	14 / 18	14 / 4 / 0
Closed	0 / 18 (18 correct FAIL)	0 / 18 (17 correct FAIL, 1 hold)	16 / 2 / 0

Per witness, open resolver, U and T: Witness B 4, C 6, D 4 unsafe of 12. All 14 unsafe extraction releases were FAIL on T items: every T family drew at least one, and `S02T`, `S03T`, and `S05T` drew all three. In each case the extractor marked the requirement UNMET with the near miss record as its evidence quote. The four K false holds are span failures (`S01K` Witness B, `S06K` Witness D) and an UNDETERMINED delivery method on `S02K` (Witnesses B and D), where the courier record does not say that the courier was national, the same gap disclosed for Round 42's `R02K`; we keep the authored label. The resolver records 23 holds for an undetermined condition, 18 resolutions on an unmet condition, 10 on all conditions met, and 3 span failures.

7.4 Cross role and cross witness coincidence, open convention

Pair	Unsafe a / b (of 12)	Joint unsafe	Independence	Frechet upper	Error phi
Witness B, direct and extraction	2 / 4	1	0.67	2	0.16
Witness C, direct and extraction	2 / 6	2	1.00	2	0.45
Witness D, direct and extraction	2 / 4	2	0.67	2	0.63
Witnesses B and C, both direct	2 / 2	1	0.33	2	0.40
Witnesses B and D, both direct	2 / 2	2	0.33	2	1.00
Witnesses C and D, both direct	2 / 2	1	0.33	2	0.40
Witnesses B and C, both extraction	4 / 6	4	2.00	4	0.71
Witnesses B and D, both extraction	4 / 4	3	1.33	4	0.63
Witnesses C and D, both extraction	6 / 4	4	2.00	4	0.71

Summed, the three role pairs failed together 5 times (independence 2.33, upper bound 6); the three cross witness direct pairs 4 times (independence 1.00, upper 6); the three cross witness extraction pairs 11 times (independence 5.33, upper 12). In Round 42 the role pairs' errors were negatively correlated; here every pair's errors are positively correlated.

7.5 Failure walk-through: `S03T`

The lease says: "A consent binds Landlord only if signed by an officer of Landlord at the level of Vice President or above. Neither the property manager nor any building staff may grant consent." It adds that a sublease made without a binding consent is void. The packet holds the sublease, which recites that consent was requested, and one consent: "Calloway Tower Owner LLC consents to the sublease of part of the 14th floor to Pellworth Studio LLC. Signed by Ross Ingle, Property Manager, for Calloway Tower Owner LLC." The question is whether the sublease is supported by a binding consent.

The reader is told the packet is an open record and that an undetermined decisive fact requires HOLD. All four witnesses that read `S03T` under that declaration answered FAIL. Witness A wrote: "The only consent in the packet is signed by Ross Ingle as Property Manager, a signatory Section 12.1 expressly disqualifies, so it does not bind Landlord. With no binding consent supporting it, the sublease falls within Section 12.4 and is void on this record." The first sentence is right. The second moves from "this consent does not bind" to "no binding consent exists", which the declared convention forbids. In the same family, all three witnesses held `S03U`, where the packet holds only the request. The extraction readers marked the consent requirement UNMET, quoting the property manager's signature, and the open resolver released FAIL. In K, a consent signed by a Vice President, all readers released PASS.

`S02T` shows the difference between witnesses. Under `DO` Witness A held: "the only delivery shown is the May 12 email, which Section 11.2 makes ineffective ... The packet neither establishes nor excludes a hand or courier delivery of the Objection Notice received by May 16". Witness C, under the same declaration, answered FAIL because "notices sent by email are not effective", and Witness D's extraction marked the 45 day requirement MET on the strength of the email.

---

8. Evidence category 3: validators on real SEC exhibits (`E43-C`)

8.1 Track

Twenty four exhibits filed with the SEC between January 2025 and October 2026, from 24 different filers, fixed with labels and ledgers before any call and run in an order that alternates HOLD and CERTIFY. The label rule is unchanged from Rounds 39 to 42: CERTIFY if the exhibit text alone establishes that the instrument's principal operative effect is in effect; HOLD if that effect depends on any condition the exhibit does not itself establish. Receipt or delivery of counterparts, fees, officer certificates, and the closing of another transaction are such conditions; ongoing covenants are not. Windows hold the opening 1,800 characters of the exhibit and 4,600 characters around the effectiveness passage. Exhibit types are EX-10 (credit agreement amendments and other material contracts) and EX-2 (merger agreement amendments). Strata: 8 with listed conditions, 4 effective only on receipt or delivery, 6 effective on their face, and 6 effective on their face with a distracting conditional clause nearby.

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Seeded defect	Link
V01	ILLINOIS TOOL WORKS INC	EX-10.A	2025-02-27	0000049826-25-000009	conditions listed	HOLD	omitted conditions	exhibit
V02	Verde Clean Fuels, Inc.	EX-10.2	2025-01-29	0001213900-25-007878	effective on its face	CERTIFY	spurious condition	exhibit
V03	EASTMAN KODAK CO	EX-10.3	2025-05-08	0000950170-25-067006	conditions listed	HOLD	omitted conditions	exhibit
V04	BROOKFIELD REAL ESTATE INCOME TRUST INC.	EX-10.4	2025-11-12	0001713407-25-000134	effective on its face	CERTIFY	spurious condition	exhibit
V05	BREAD FINANCIAL HOLDINGS, INC.	EX-10.1	2026-09-08	0001101215-26-000081	conditions listed	HOLD	omitted conditions	exhibit
V06	HARLEY-DAVIDSON, INC.	EX-2.6	2025-11-06	0001104659-25-107721	effective on its face	CERTIFY	spurious condition	exhibit
V07	TransDigm Group INC	EX-10.1	2025-08-05	0001260221-25-000051	conditions listed	HOLD	omitted conditions	exhibit
V08	Vita Coco Company, Inc.	EX-10.3	2026-10-22	0001482981-26-000185	effective on its face	CERTIFY	spurious condition	exhibit
V09	Celcuity Inc.	EX-10.3	2025-05-15	0001641172-25-011053	conditions listed	HOLD	omitted conditions	exhibit
V10	Mechanics Bancorp	EX-2.1	2025-09-02	0001140361-25-033560	effective on its face	CERTIFY	spurious condition	exhibit
V11	KKR Real Estate Finance Trust Inc.	EX-10.5	2025-04-23	0001631596-25-000029	conditions listed	HOLD	omitted conditions	exhibit
V12	Iron Horse Acquisition II Corp.	EX-2.1	2026-05-18	0001213900-26-058488	effective on its face	CERTIFY	spurious condition	exhibit
V13	AMCON DISTRIBUTING CO	EX-10.1	2025-07-18	0001558370-25-009300	conditions listed	HOLD	omitted conditions	exhibit
V14	Vacasa, Inc.	EX-2.1	2025-03-28	0001140361-25-010956	effective on its face, distractor	CERTIFY	spurious condition	exhibit
V15	Blackstone Secured Lending Fund	EX-10.2	2026-08-06	0001736035-26-000016	conditions listed	HOLD	omitted conditions	exhibit
V16	LPL Financial Holdings Inc.	EX-2.1	2025-11-03	0001628280-25-047934	effective on its face, distractor	CERTIFY	spurious condition	exhibit
V17	BRANDYWINE REALTY TRUST	EX-10.1	2025-11-07	0000790816-25-000046	receipt or delivery only	HOLD	omitted conditions	exhibit
V18	ACELYRIN, Inc.	EX-2.1	2025-04-21	0001193125-25-086274	effective on its face, distractor	CERTIFY	spurious condition	exhibit

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Seeded defect	Link
`V19`	CBIZ, Inc.	EX-10.1	2025-04-25	`0001628280-25-019921`	receipt or delivery only	HOLD	omitted conditions	exhibit
`V20`	Black Stone Minerals, L.P.	EX-10.1	2025-08-27	`0001621434-25-000119`	effective on its face, distractor	CERTIFY	spurious condition	exhibit
`V21`	CAVCO INDUSTRIES, INC.	EX-10.2	2025-08-04	`0000278166-25-000128`	receipt or delivery only	HOLD	omitted conditions	exhibit
`V22`	Fox Corp	EX-10.2	2025-09-10	`0001193125-25-200134`	effective on its face, distractor	CERTIFY	spurious condition	exhibit
`V23`	Longevity Health Holdings, Inc.	EX-2.1	2025-06-25	`0000950170-25-090015`	receipt or delivery only	HOLD	omitted conditions	exhibit
`V24`	Cantor Equity Partners I, Inc.	EX-2.1	2026-03-26	`0001213900-26-034847`	effective on its face, distractor	CERTIFY	spurious condition	exhibit

8.2 Results

Arm	Ledgers	Correct	Unsafe releases	False holds	Repaired	Corrupted
No validator (ledgers as given)	48	24	12	12		
Flag only, Witness C	48	36	0 (95% upper 0.074)	12	12	0
Rewriting, Witness C	48	48	0 (95% upper 0.074)	0	24	0

The flag only arm flagged all 12 seeded HOLD ledgers ("missing condition", 23 flags), all 12 seeded CERTIFY ledgers ("not a condition", 12 flags), 9 of 12 gold HOLD ledgers (15 "missing condition" and 2 "not a condition" flags; harmless, since those ledgers already hold), and none of the 12 gold CERTIFY ledgers. So it raised no flag that would have stopped a correct release. Its 12 false holds are the seeded CERTIFY ledgers, which it diagnosed correctly and could not repair. The rewriting arm returned every gold ledger with an unchanged resolution, restored the missing condition in all 12 seeded HOLD ledgers, and removed the spurious condition from all 12 seeded CERTIFY ledgers. The Witness B repeat of this experiment did not run (Section 3.5).

A post hoc arm, defined after these results were seen and reported only as such, accepts a rewrite when it moves the decision toward HOLD or leaves it unchanged and routes any rewrite that would turn a HOLD into a release to review. Applied to the recorded rewrites, it gives 36 correct, 0 unsafe, 0 false holds, and 12 routed to review (the 12 seeded CERTIFY repairs).

8.3 Failure walk-through: `V17`

`V17` (Brandywine Realty Trust, accession `0000790816-25-000046`) reads: "This Amendment shall become effective upon receipt by the Administrative Agent of counterparts of this Amendment duly executed by the Credit Parties, Lenders constituting Required Lenders and the Administrative Agent." The seeded ledger lists no condition and gives "This Amendment shall become effective" as the effectiveness quote. Every word of that quote is in the exhibit, so it passes the span check, and the

resolver releases CERTIFY on an amendment that is not yet effective on the face of the filing. The flag only validator returned one flag, "missing condition", quoting the receipt clause, and the decision became HOLD. The rewriting validator returned a ledger with that clause as a condition. Both arms were right here. The same span valid truncation is the move S6 shows a rewriting validator could make on any gold HOLD ledger: delete the condition, quote the opening words, and release.

---

9. Findings and interpretations

9.1 Measured findings

- **F43.1 (deterministic, E43-A).** All seven certificates pass. A flag only validator cannot create an unsafe release; a rewriting validator can, on every one of 12 gold HOLD ledgers, without failing a span check.
- **F43.2 (live, E43-B).** Undeclared, direct readers released FAIL on 33 of 36 U and T decisions. Declared open, U errors fell from 15 of 18 to 0 of 18 and T errors from 18 of 18 to 6 of 18; all remaining errors were FAIL on near miss items (`S03T`, `S06T`, `S02T`).
- **F43.3 (live, E43-B).** Declared closed, direct readers were correct on 36 of 36 U and T decisions and 18 of 18 K decisions. K was also 18 of 18 under the open declaration.
- **F43.4 (live, E43-B).** Extraction readers under the open resolver held 18 of 18 U items and released FAIL on 14 of 18 T items, marking the requirement UNMET from the near miss record. Under the closed resolver they were correct on 35 of 36 U and T items with no unsafe release.
- **F43.5 (live, E43-B).** Under the open convention, 5 of 6 direct unsafe releases coincided with the same witness's extraction unsafe release (role pairs joint 5, independence 2.33); error correlation was positive for every role pair (0.16, 0.45, 0.63) and every cross witness pair (0.40 to 1.00).
- **F43.6 (live, E43-C).** On 48 exhibit ledgers, the flag only arm made 0 unsafe releases and 12 false holds and repaired 12; the rewriting arm made 0 unsafe releases and 0 false holds and repaired 24; neither corrupted a correct ledger.

9.2 Interpretation

The convention result answers the Round 42 threat. Most of the Round 42 absence errors were a convention effect: told that the record is open, readers held every U item, and told that it is closed, they failed every U item, both correctly. The reader's default, when nothing is said, is the closed world.

The near miss result shows the default is not only about absence. Under a declared open record, a record that fails a requirement shows that this record does not satisfy it; it does not show that no satisfying record exists. Readers in both roles drew the second conclusion. The extraction ledger made this easier: its schema asks whether a condition is MET, UNMET, or UNDETERMINED, and a defective record looks like evidence that the condition is UNMET. The schema has no place to say "the record shown does not satisfy the condition, and no other record is shown". This is why extraction, which was the safe role in Round 42, was the less safe role here.

Round 42's role decorrelation worked because the two roles erred in opposite directions on absence. On near miss items both roles err in the same direction, and their errors coincide. Role is a lever for decorrelation only when the roles' failure modes differ on the item type at hand.

The validator result is mixed in a useful way. Live, the rewriting validator was strictly better: same safety, no false holds. Structurally, only the flag only validator has a guarantee. One validator witness on 48 ledgers cannot show that rewriting is safe, only that this witness did not err here (95 percent upper limit 0.074 per ledger). The asymmetry suggests a composition that keeps both properties: accept rewrites that add conditions or otherwise move toward HOLD, since they can only cost a hold, and send rewrites that remove conditions to review, since only those can release wrongly.

9.3 Hypothesis status

Hypothesis	Status
H43.1 open declaration removes direct FAIL releases	Supported on U (15 of 18 to 0 of 18); not supported on T (18 of 18 to 6 of 18)
H43.2 closed declaration followed, K unchanged	Supported (36 of 36; K 18 of 18)
H43.3 cross role coincidence once both roles err	Supported (5 of 6 direct unsafe releases coincide; positive error correlation in every role pair)
H43.4 flag only creates no unsafe release, rewriting can	Supported structurally (S6); live, neither arm created one, so the rewriting risk was not observed
H43.5 flag only pays in unrepairable false holds	Supported (12 of 12 seeded CERTIFY ledgers)

---

10. Toward Theory v1

- **P7 revised. Absence is not negation, and a failed record is not proof of absence.** The undeclared default is the closed world. A declared open convention removes absence errors but not near miss errors. Evidence: F43.2, F43.3, `S03T`. Removal test: an extraction schema with a record level status (this record satisfies, fails, or is irrelevant) in which near miss T items hold under the open resolver.
- **P4 revised. Role decorrelates only where the roles' failure modes differ.** On absence the direct role errs by releasing and the extraction role holds; on near miss records both release the same wrong FAIL. Evidence: F42.6 against F43.5. Removal test: the record level schema above, with the cross role coincidence on T items falling to the independence level or below.
- **P5 revised. Rewrites that can release need a gate; rewrites toward HOLD do not.** Flag only stages are safe by construction and cost unrepairable holds; rewriting stages repair more and carry an unbounded structural risk. Evidence: F43.1, F43.6, `X24` (Round 40). Removal test: a rewriting validator that releases wrongly on gold or seeded HOLD ledgers at a rate the post hoc hold only rule would not have caught.

---

11. Method changes for later rounds

1. Every evaluation pool and every direct prompt states its record convention.
2. Extraction ledgers record a status per record as well as per condition, so that "this record fails" is separable from "the condition is unmet".
3. Validator comparisons report gold and seeded ledgers separately and count repairs, created errors, and corruptions, not only accuracy.
4. A validator that may rewrite is gated: rewrites that move a decision from HOLD to a release go to review.

12. Synthesis review

As in earlier rounds, Witness A at maximum reasoning was asked to review all recorded results after the analysis. Two attempts returned no review text, and the budget rule then stopped the call. The request and its status are archived (`synthesis-review.json`). No synthesis review is reported for this round, and no part of this paper depends on one.

13. Limitations and threats to validity

- The T labels under the open convention are the round's most contestable choice. A reader may hold that a consent signed by an excluded signatory (`S03T`) or a notice citing another agreement (`S06T`) establishes the negative in practice, because a second valid record is unlikely. The open declaration says otherwise, and we score against it, but a reviewer applying a likelihood standard could label some T items FAIL.
- The pool is small: 6 families, 18 items, 12 U and T decisions per witness. Decisions that share an item are not independent.
- Witness A read only three T items, all under `DO`; its other planned items were stopped by the budget rule. The four witness quorum of earlier rounds is not reproduced.
- The validator experiment used one witness. The planned Witness B repeat did not run, so validator differences between witnesses are untested.
- Seeded defects are of two kinds (omitted conditions with a truncated effectiveness quote, and one spurious condition). Real validator errors may be subtler, such as a reworded condition or a partly correct ledger.
- Families, labels, and ledgers were written by the research team; no external reviewer checked them before the run. Every call was made once at temperature `0`.

14. Next check

Add a record level status to the extraction ledger (for each cited record: satisfies, fails, or irrelevant to the condition) and resolve "condition unmet" only when the convention is closed or the record set is declared complete. Rerun the six T families with new surface forms and new near miss kinds, Witness A on all items, and a second validator witness on the Round 43 exhibit ledgers with harder seeded defects (reworded conditions, partial omissions). This also serves the roadmap's Round 44 item, abstention at scale with externally reviewed HOLD labels, since the T items need exactly that review.

---

15. References

- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. https://efts.sec.gov/LATEST/search-index and https://www.sec.gov/Archives/edgar/data/.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.
- [S41] Elliot Kim, Avi Garg, Kenny Peng, Nikhil Garg. Correlated Errors in Large Language Models. Proceedings of the 42nd International Conference on Machine Learning (ICML), PMLR 267, 2025. arXiv:2506.07962.
- [S42] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, et al. Great Models Think Alike and this Undermines AI Oversight. arXiv:2502.04313, 2025.
- [S43] John C. Knight, Nancy G. Leveson. An Experimental Evaluation of the Assumption of Independence in Multiversion Programming. IEEE Transactions on Software Engineering 12(1), 1986. doi:10.1109/TSE.1986.6312924.
- [S44] Dave E. Eckhardt, Larry D. Lee. A Theoretical Basis for the Analysis of Multiversion Software Subject to Coincident Errors. IEEE Transactions on Software Engineering 11(12), 1985. doi:10.1109/TSE.1985.231895.
- [S45] Bev Littlewood, Douglas R. Miller. Conceptual Modeling of Coincident Failures in Multiversion Software. IEEE Transactions on Software Engineering 15(12), 1989. doi:10.1109/32.58771.
- [S46] Ludmila I. Kuncheva, Christopher J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51, 2003. doi:10.1023/A:1022859003006.
- [S47] James A. Hanley, Abby Lippman-Hand. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA 249(13), 1983. doi:10.1001/jama.1983.03330370053031.
- [S48] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, Jason Weston. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495, 2023.
- [S49] Raymond Reiter. On Closed World Data Bases. In Hervé Gallaire and Jack Minker (eds.), Logic and Data Bases, Plenum Press, 1978, pp. 55 to 76. doi:10.1007/978-1-4684-3384-5_3.

- [S50] Keith L. Clark. Negation as Failure. In Hervé Gallaire and Jack Minker (eds.), Logic and Data Bases, Plenum Press, 1978, pp. 293 to 322. doi:10.1007/978-1-4684-3384-5_11.
- [S51] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations (ICLR), 2024. arXiv:2310.01798.
- [S52] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, Rui Zhang. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Transactions of the Association for Computational Linguistics 12, 2024, pp. 1417 to 1440. doi:10.1162/tacl_a_00713. arXiv:2406.01297.