Encevra Research Round 44: This Record Fails, or No Record Exists? Record Level Status Ledgers, Completeness Records, and Independently Reviewed HOLD Labels

Date: 2026-11-02
Research lab: Encevra Research
Author: Ben Wilkins, Encevra (ORCID 0009-0005-6532-3200)
Short code: `RLS-44` (Record Level Status)
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: fourth round of the Round 41 to 50 "Theory v1" block. The roadmap item for Round 44 is abstention at scale with externally reviewed HOLD labels (proposition P1, structural abstention).
Carry-forward boundary: Round 43 (`RCV-43`) found that a declared open record stops readers treating a missing record as a negative but does not stop them treating a defective record as one, and that condition ledgers made this error on 14 of 18 near miss items. Its handoff asked for a ledger with a status per record, a resolver that reads "unmet" only on a closed record or a declared complete record set, new near miss kinds, and harder seeded validator defects, and noted that the near miss labels themselves needed outside review. Round 44 builds and tests that ledger, adds a completeness record variant, and puts every authored label through a blind review before any test call. It does not rerun the Round 43 convention study, the Round 42 role pair design, or the Round 41 selection study.

Abstract

A record that fails a contract requirement shows that this record does not satisfy it. It does not show that no satisfying record exists. Round 43 found that evidence readers draw the second conclusion: asked whether a condition was MET, UNMET, or UNDETERMINED, they marked it UNMET on 14 of 18 items where the packet held a defective record, and under an open record that released an unsafe FAIL. Round 44 changes the question the reader answers. The record level ledger asks, for each packet record, whether that record SATISFIES the requirement, FAILS it, or is IRRELEVANT, and asks separately for any words in the packet that exclude every other qualifying record. Deterministic code then decides: a condition is met if a record satisfies it, unmet only if an exclusion is quoted, and otherwise undetermined, which holds under an open record and fails under a closed one.

Eight new authored families (a cure payment USD 100,000 short, an extension notice received before its window, an insurance certificate naming the owner only as certificate holder, an unsigned execution version, a letter of credit that expires too early, a lien waiver covering too short a period, an opinion from in house counsel, and a PDF image of a share certificate where the original was required) are each written in three variants: K, where the final record satisfies the requirement; T, where it is a near miss; and TC, which adds a completeness record stating that the near miss record is the only one. Every

HOLD label carries a two completion certificate, and all 24 packets were reviewed blind under both conventions by a witness outside the test panel before any test call.

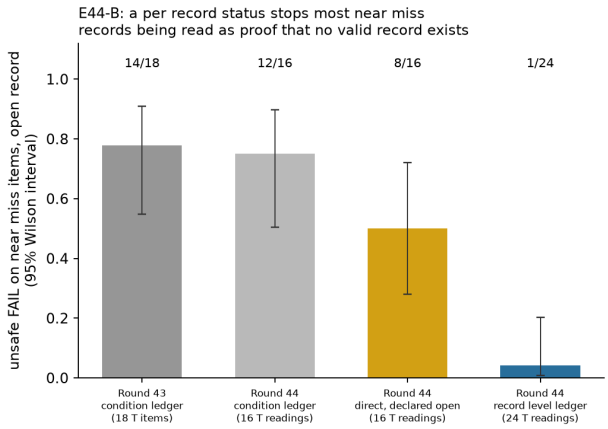
We report three strictly separated evidence categories.

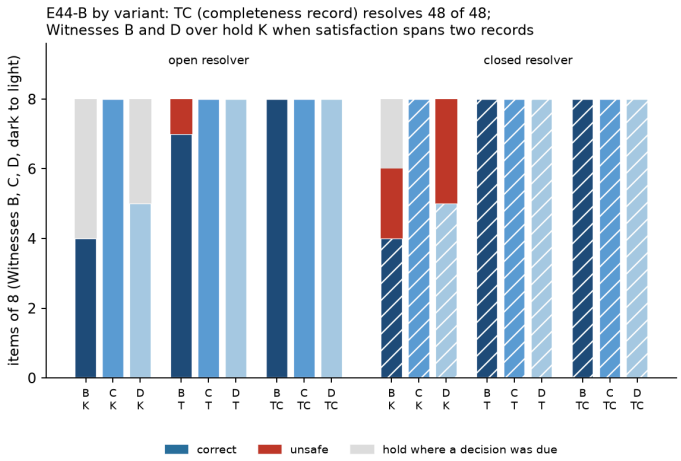
Deterministic (`E44-A`, no model calls). All eight certificates pass. K, T, and TC differ only in the records they are built to differ in; exactly the 8 T items admit consistent completions with different verdicts, so exactly they are HOLD under an open record; gold record level ledgers reproduce all 48 labels; and under the record level resolver no ledger can release FAIL on a T item without quoting an exclusion, while the condition ledger releases FAIL from the near miss record alone. All 24 new SEC exhibits overlap none of the 1,100 accessions and 833 filer names used before. The pre registered rewrite gate is monotone: it can never produce a release that the original ledger did not already produce.

Live authored (`E44-B`, 8 families, 24 items, 128 completed calls). The blind review agreed with all 48 authored labels and reported packet defects in 6 of 8 families. With the record level ledger and the open resolver, Witnesses B, C, and D held 23 of 24 near miss T items and released an unsafe FAIL on 1. On the same 8 T items, the Round 41 to 43 condition ledger released an unsafe FAIL on 12 of 16 (Witness B 4 of 8, Witness C 8 of 8), and declared open direct readers released one on 8 of 16 (4 of 8 each, on the same four items). All 24 TC ledgers and all 16 TC direct readings resolved to the correct FAIL, and under the closed resolver all 48 T and TC ledgers were correct. The weakness appeared on K: Witness C resolved all 8 K items correctly, but Witness B held 4 of 8 and Witness D 3 of 8, and under the closed resolver they released 2 and 3 unsafe FAILs, mostly because the requirement was met by two records together and the schema asks whether one record meets it "by itself". The one T error is a scope error in the new exclusion field: Witness B quoted "Additional insured: none listed." from the defective certificate as if it excluded every other certificate.

Real exhibits (`E44-C`, 24 SEC EDGAR exhibits). The track, gold ledgers, and seeded ledgers with harder defects (span valid "near miss evidence" marked SATISFIED, and spurious unestablished conditions) are fixed and certified. Witness C validated all 48 ledgers and Witness B the 24 seeded ledgers, each in a flag only and a rewriting arm (144 calls). No arm made an unsafe release. Both validators flagged all 24 seeded ledgers; the flag only arm therefore caught every near miss evidence seed but, by design, left every spurious condition seed as a false hold (12 of 24 per witness). The rewriting arm repaired all 24 seeded ledgers for both witnesses and changed no gold ledger's outcome. The pre registered gate accepted every hold creating repair and routed the 12 hold to release repairs per witness to review. These phases were first stopped by the run budget and were rerun later with the same prompts and settings (Section 3.5).

Interpretation, kept separate. Moving the status from the condition to the record, and giving exclusion its own field, did what Round 43 predicted on near miss items for all three witnesses tested, while the unchanged condition ledger repeated the Round 43 error on the new families, and the completeness variant shows the same readers will return FAIL when the packet actually supports it. The schema has two new weak points: joint satisfaction (a requirement met only by two records read together) and exclusion scope (a negative statement inside one record read as a statement about all records). The blind review mattered most where we did not expect it. It confirmed every label, but the defects it raised fell mostly on K packets, where our own PASS labels rested on inferences (a certificate date read as a delivery date, a waiver's presence read as its receipt), and those were the same K items on which Witnesses B and D held.





1. Archive continuity and anti-repetition verification

Before designing any Round 44 experiment we rebuilt the archive ledger from the live research registry (`prior-rounds-ledger.json`, `37` indexed rounds, `173` registry hypotheses, `67` registry experiments). A keyword scan of titles, questions, summaries, findings, open questions, hypotheses, and experiments for the Round 44 constructs returned:

- "record level" and "satisfies, fails" only in Round 43, whose method change and open question this round carries out;
- "externally reviewed" and "reviewed hold" only in the Round 40 roadmap, whose Round 44 item this is;
- "completion" in Rounds 16 and 17. Round 16 asked readers to produce a validated pair of evidence compatible scenarios with different decisions before committing. Round 44 uses the same logical idea on the label side, to certify which authored labels are HOLD; no reader is asked for scenarios. Round 17 concerned compressed handoffs;
- "independent review" in Rounds 25 and 30, about silent authorities and a readiness gate; neither reviewed benchmark labels;
- "near miss" only in Round 43, whose six near miss kinds do not recur here.

"Completeness record", "exclusion", "two completion", "label review", "blind review", and "gated rewrite" returned no hits. The full table is in `anti-repetition-verification.md`.

Prior round	Focus	How Round 44 differs
`RCV-43` (Round 43)	Declared conventions; condition ledgers read near miss records as negation; flag only against rewriting validators	Status per record and an explicit exclusion field; a completeness record variant; new families and near miss kinds; a fixed rewrite gate tested on harder seeded defects
`RDX-42` (Round 42)	Absence read as negation under an unstated convention	Not rerun
`NCT-40` (Round 40)	Six topologies; the roadmap	Round 44 is the roadmap item; labels reviewed blind by a witness outside the test panel

Prior round	Focus	How Round 44 differs
`Round 16`	Reader produced scenario pairs before commitment	Label side completion certificates only

The condition ledger prompt and the open declaration are the Round 43 prompts, unchanged, and serve only as baselines. All 24 exhibits come from new EDGAR full text searches.

---

2. Where this round sits in the program

Round 40 proposed P1 to P5; Round 41 added P6; Round 42 proposed P7 (absence is not negation); Round 43 revised P4, P5, and P7. Round 43 left three removal tests: for P7, a record level schema under which near miss T items hold; for P4, the same schema with cross role coincidence on T items falling; for P5, a rewriting validator that releases wrongly at a rate a hold only gate would not catch. The roadmap item for Round 44 is P1 at scale with HOLD labels that someone other than the authors has checked.

---

3. Design

3.1 Families and variants

Each family is one rule program rendered three times. All variants share the contract, the first record, and the audit question.

Family	Requirement	K: final record	T: near miss final record	TC: added completeness record
`N01`	Cure by paying at least the overdue interest (USD 612,500) within 10 Business Days of notice	Wire of USD 612,500 on June 10	Wire of USD 512,500 on June 10	Agent's officer: the June 10 wire is the only payment received June 1 to 30
`N02`	Extension notice received 270 to 180 days before the Term ends	Notice stamped received May 15	Notice stamped received February 10, before the window opened	Landlord certificate: exactly one extension communication since January 1
`N03`	Certificate of insurance naming Owner as additional insured, delivered before work	Certificate naming Owner as additional insured	Certificate naming Owner only as certificate holder; "Additional insured: none listed."	Contractor's risk manager: that certificate is the only evidence of insurance delivered before September 30
`N04`	Outside Date extended only by an amendment executed by both parties before October 1	Amendment No. 2, signed by both	Execution version with blank signature blocks	Joint letter: no amendment was executed by both parties before October 1
`N05`	Standby letter of credit of at least USD 250,000 not expiring before May 30, 2027	Expiry June 30, 2027	Expiry December 31, 2026	Issuing bank: this is the only letter of credit for the account of Customer in favor of Supplier

Family	Requirement	K: final record	T: near miss final record	TC: added completeness record
`N06`	Lender received a final unconditional lien waiver through Substantial Completion (June 30)	Waiver through June 30	Waiver through May 31	Lender's administrator: the waivers received consist solely of the July 8 waiver
`N07`	Trustee received an opinion of independent outside counsel before executing	Opinion of outside law firm	Opinion of the Company's General Counsel	Trustee's receipt: these documents and no others were received before execution
`N08`	Original share certificate with stock power delivered within 5 Business Days	Custody receipt for the original certificate	Custody receipt for a PDF image by email	Custody log: one entry, the PDF image; no other items received

Labels: K is PASS under both conventions. T is HOLD under an open record and FAIL under a closed record. TC is FAIL under both, because the completeness record excludes any other qualifying record. Every packet was shown under an opaque identifier (`P01` to `P24`, fixed by a seeded shuffle) so that no identifier revealed its variant.

3.2 Two completion label certificates

A HOLD label is a claim that the packet is consistent with both answers. For each item we enumerate the completions the packet allows: the recorded facts alone, and, where no record excludes it, the recorded facts plus one further qualifying event. The family's rule function computes the verdict on each completion. An item is certified HOLD under an open record exactly when its completions give both PASS and FAIL, and certified determinate when every completion gives the same verdict. A completeness record removes the second completion, so TC is determinate even under an open record.

3.3 The record level ledger and resolver

Readers fill one ledger per item. For every requirement they quote the contract, list each packet record that bears on it with a verbatim quote and a status (SATISFIES if that record by itself meets every stated part of the requirement, FAILS if it is the right kind of record but does not, IRRELEVANT otherwise), and fill an `exclusion_quote` only with words in the packet that rule out any other qualifying record or event. No convention is declared to the reader. The resolver decides:

1. any quote that is not a verbatim span of the packet (normalized, at least 15 characters) gives HOLD;
2. a condition is MET if some record marked SATISFIES has a verified quote;
3. otherwise it is UNMET only if a verified exclusion quote is present;
4. otherwise it is UNDETERMINED: HOLD under the open resolver, UNMET under the closed resolver.

A record marked FAILS never makes a condition UNMET by itself. The baseline is the Round 41 to 43 condition ledger (one status per condition) with its unchanged resolver.

3.4 Blind label review

Before any test call, Witness A at maximum reasoning reviewed each family's three packets in a shuffled order, without the authored labels, and returned an open and a closed verdict for each packet together with any packet defects and whether the label was contestable. The adjudication rule was

fixed in advance: a label is confirmed when both verdicts agree with it; if the reviewer reports a packet defect, the packet is fixed and reviewed again when the run budget allows, and otherwise the packet and label are kept unchanged, the family is marked contested, and every authored result is reported both on all families and on the uncontested families only. Witness A was not a test reader in this round. It is an independent model reviewer, not a human legal expert, and we use the word independent in that sense only.

3.5 Run order, budget, and completion

The phases were fixed before any call: review, then the test phases. The review used a much larger share of the run budget per call than planned. After the review and before any test call, the test phases were reordered and narrowed to fit the remainder (record level ledgers with Witnesses B and C on all 24 items; the condition ledger baseline on the 8 T items; Witness C validators on the 24 exhibits; direct open declared readers on T and TC; Witness D record level ledgers; Witness B validators), and a re-review of defective packets, Witness D in the baseline and direct phases, direct reading of K items, and the synthesis call were dropped. This revision is recorded in `run-plan-r44.md`. No test outcome had been observed when it was made.

A budget rule fixed before the run skips every later call once the run budget is reached. It was reached during the condition ledger baseline after 3 of 16 calls (`N01T` Witnesses B and C, `N02T` Witness B), and every later call was skipped: the other 13 baseline calls, all validator calls on the exhibits, all direct reader calls, all Witness D calls, and the synthesis review. The 13 stopped calls remain in the transcripts with their stopped status.

Later the same day, after the first results had been seen, the stopped phases were completed in the revised order with the same prompts, the same harness, temperature `0`, and the same witnesses: the 13 remaining baseline calls, the Witness C validators (96 calls), the direct readers (32), the Witness D record level ledgers (24), the Witness B validators (48), and the synthesis review. No completed call was repeated, and no item, prompt, scoring rule, or phase was changed. The phases dropped in the revision before any test call (a re-review of defective packets, Witness D in the baseline and direct phases, direct reading of K items) were not added. In total 274 calls completed: 8 review, 72 record level, 16 baseline, 32 direct, 144 validator, and 2 synthesis calls (the first synthesis reply was cut off by its response length limit and was repeated once with a larger limit). Because the completed phases ran after the first results were known, they are a completion of the fixed design, not an independent replication.

---

4. Related work

Unanswerable questions as a test of whether a reader knows what it does not know go back to SQuAD 2.0 [S54]; selective prediction formalizes the trade of coverage for risk [S55], and selective question answering studies it under shift [S56]. Kirichenko et al. find that reasoning models often fail to abstain on unanswerable questions [S60], and Wen et al. survey abstention [S24]. Contract reading benchmarks such as ContractNLI include a "not mentioned" class alongside entailment and contradiction [S53], and LegalBench collects legal reasoning tasks [S59]; Round 44 differs in deriving the abstention label from the record's completeness rather than from the text's silence. The open and

closed record conventions are the open and closed world assumptions of deductive databases [S49], and the closed reading is negation as failure [S50]. Northcutt et al. showed that label errors in common test sets can change which model ranks first [S57], which is why our HOLD labels carry certificates and a blind review. Using a model as a reviewer follows the judge literature [S58], whose known biases are one reason we report the review as independent rather than external. Contrast variants follow Gardner et al. [S40]; zero event limits use the exact binomial limit [S47].

---

5. Hypotheses

These hypotheses restate the removal tests published by Round 43 and the roadmap item. They were written into this manuscript after the design was fixed and are not a separate preregistration.

- **H44.1 (P7).** With a record level ledger, near miss T items hold under the open resolver: the unsafe FAIL rate on T falls well below the condition ledger's.
- **H44.2.** The record level ledger stays determinate where the record supports a decision: K resolves PASS, and TC resolves FAIL under both resolvers.
- **H44.3.** Declared open direct readers err on T items and not on TC items.
- **H44.4 (P4).** Under the record level ledger, unsafe releases on T items do not coincide across witnesses.
- **H44.5 (P1).** The authored HOLD labels survive a blind independent review under both conventions.
- **H44.6 (P5).** A gated rewriting validator makes no unsafe release on exhibit ledgers with harder seeded defects.

---

6. Evidence category 1: deterministic diagnostic (`E44-A`)

Certificate	Result
S1 contrast: K and T differ only in the final record; TC is T plus exactly one completeness record	8 of 8 families
S2 two completion certificates: HOLD under an open record exactly when completions give both verdicts; K and TC invariant	24 of 24 items; exactly the 8 T items certified HOLD
S3 resolver contrast by enumeration over the decisive record (status times exclusion present) on the 8 T items	record level: FAIL only with an exclusion quote (16 of 48 enumerated ledgers, all with one; FAILS or IRRELEVANT without one gives HOLD in 16 of 16); condition ledger: UNMET quoting the near miss record gives FAIL in 8 of 8. Under both schemas a SATISFIES or MET status on the near miss record releases PASS, so both still depend on the reader classifying the record
S4 gold record level ledgers reproduce every label under both resolvers	48 of 48

Certificate	Result
S5 exhibit track: new, balanced, stratified	24 of 24 new accessions and filer names against 1,321 prior records (1,100 accessions, 833 filer names); 24 distinct filers; 12 HOLD, 12 CERTIFY; 8 conditions listed, 4 receipt or delivery only, 6 effective on face, 6 effective on face with a distractor
S6 exhibit ledgers	gold 24 of 24 correct; seeded ledgers give 12 unsafe releases and 12 false holds; 84 of 84 quotes are verbatim spans
S7 rewrite gate monotonicity over every pair of original and rewritten outcomes	the gate releases only when both the original and the rewrite release; it never creates a release
S8 resolution	zero events allow a true rate of 0.312 at n = 8, 0.171 at 16, 0.117 at 24, and 0.061 at 48 (exact one sided 95 percent limit)

S3 carries a limit that S4 does not show. The resolver checks that an exclusion quote is a verbatim span; it cannot check what the quote excludes. A negative statement that describes one record ("none listed" on a certificate) passes the span check as readily as a statement that no other record exists. Section 7.5 shows a live case.

---

7. Evidence category 2: live authored families (`E44-B`)

7.1 Blind label review

Family	Open and closed verdicts agree with labels	Defects reported	Reviewer called the label contestable
`N01` cure payment	6 of 6	none	no
`N02` extension notice	6 of 6	waiver or estoppel argument from a Landlord reminder; completeness certificate from an interested party	yes
`N03` insurance certificate	6 of 6	delivery date inferred from the certificate date (K, TC)	yes
`N04` outside date amendment	6 of 6	title "Amendment No. 2" implies an unseen Amendment No. 1	no
`N05` letter of credit	6 of 6	an issuing bank can speak only to its own letters of credit, so TC could be read as HOLD	yes
`N06` lien waiver	6 of 6	no record that Lender received the waiver (K)	yes
`N07` opinion of counsel	6 of 6	no receipt record for the opinion (K); "outside counsel to the Company" against "independent" (K)	yes
`N08` share certificate	6 of 6	none	no

The reviewer agreed with all 48 authored labels. Under the fixed rule, the six families with reported defects (`N02` to `N07`) are contested; `N01` and `N08` are uncontested. A re-review was not run (Section 3.5), so no packet was changed. Of the eleven defect notes, five concern K packets, three concern TC

packets, and three concern T packets. The reviewer's own open and closed verdicts matched every T label; its T notes concern how the packet could be argued, not the verdict.

7.2 Record level ledgers

Resolver	Variant (label)	Witness B correct / held instead of a verdict / unsafe	Witness C correct / held instead of a verdict / unsafe	Witness D correct / held instead of a verdict / unsafe
Open	K (PASS)	4 / 4 / 0	8 / 0 / 0	5 / 3 / 0
Open	T (HOLD)	7 / not applicable / 1	8 / not applicable / 0	8 / not applicable / 0
Open	TC (FAIL)	8 / 0 / 0	8 / 0 / 0	8 / 0 / 0
Closed	K (PASS)	4 / 2 / 2	8 / 0 / 0	5 / 0 / 3
Closed	T (FAIL)	8 / 0 / 0	8 / 0 / 0	8 / 0 / 0
Closed	TC (FAIL)	8 / 0 / 0	8 / 0 / 0	8 / 0 / 0

Across the three witnesses under the open resolver: T unsafe 1 of 24; TC correct 24 of 24; K correct 17 of 24 with no unsafe release. On the uncontested families only (`N01`, `N08`): T 6 of 6 held, TC 6 of 6 correct, K 4 of 6 correct (Witness B held `N08K`, Witness D held `N01K`).

How readers classified the decisive record (Witnesses B, C, and D): on T items, 15 of 24 ledgers marked it FAILS and nothing else, 8 marked it FAILS under one condition and SATISFIES under another (for example, a wire that satisfies the timing condition and fails the amount condition), and 1 marked it FAILS and added an exclusion quote. On TC items, all 24 ledgers quoted the completeness record as the exclusion. On K items, 16 of 24 marked the final record SATISFIES only, 7 marked it FAILS under at least one condition, and 1 marked it IRRELEVANT under one condition.

7.3 Condition ledger baseline and direct readers on the same T items

Reader on the 8 T items (open label HOLD)	Witness B unsafe FAIL	Witness C unsafe FAIL	Both unsafe on the same item
Record level ledger, open resolver	1 of 8	0 of 8	0
Condition ledger (Rounds 41 to 43), open resolver	4 of 8	8 of 8	4
Direct reader, open declaration	4 of 8	4 of 8	4

Witness D record level ledgers released no unsafe FAIL on T (0 of 8). The condition ledger released FAIL on 12 of 16 T readings by marking the requirement UNMET on the strength of the near miss record alone, the error Round 43 measured on 14 of 18 T items in other families; Witness B's other 4 baseline readings held (2 undetermined, 2 with an unverified evidence quote). Under the closed resolver the condition ledger was correct on 14 of 16 T readings and held on 2. The declared open direct readers released FAIL on `N02T`, `N03T`, `N05T`, and `N06T`, both witnesses on the same four items, and held on the other four. On TC both direct readers returned the correct FAIL on 16 of 16. Direct reading of K items was dropped before any test call, so the direct readers' K accuracy is not measured in this round.

7.4 Where the record level ledger erred on K

All four Witness B holds on K have a recorded cause. On `N03K` and `N06K` Witness B marked every record FAILS for a requirement that two records meet together: the certificate dated September 2 and the site log showing work began September 8 together establish delivery before work, and the waiver and the completion record together establish coverage through Substantial Completion. Its reasons read "does not by itself establish when work started" and "does not by itself establish that this was the date of Substantial Completion". The schema's definition of SATISFIES ("that record by itself meets every stated part of the requirement") asked for exactly that. Under the open resolver these were holds; under the closed resolver, where undetermined reads as unmet, they became unsafe FAILs. On `N07K` the condition quote joined two sentences that are not adjacent in the contract, and on `N08K` it was the two word phrase "Pledged Shares", below the minimum quote length; both failed the span check and held.

Witness D held `N01K`, `N03K`, and `N07K` for the same reason. In each ledger it marked the decisive records FAILS with a "by itself" reason: the wire "does not by itself state the notice date", the certificate and the site log each do not by themselves show delivery before work, and the opinion and the execution record each do not show receipt before execution. Under the closed resolver all three became unsafe FAILs. No Witness D ledger failed the span check.

Five of these items (`N03K`, `N06K`, `N07K` for Witness B; `N03K`, `N07K` for Witness D) are K packets on which the blind reviewer had reported a defect, and on each the defect is the same gap: the packet shows the certificate date but not the delivery date, the waiver but not its receipt, the opinion's date but not its receipt. Witness C resolved all of them PASS.

7.5 Failure walk-through: `N03T`

The construction contract says: "Before starting work at the site, Contractor shall deliver to Owner a certificate of commercial general liability insurance that names Owner as an additional insured. A certificate that names Owner only as certificate holder does not satisfy this Section." The packet holds the site log ("Mobilization and start of excavation on September 8, 2026.") and one certificate dated September 2, delivered to Owner: "Certificate holder: Ferncastle Medical Properties LLC. Additional insured: none listed." The question is whether Contractor satisfied the Section before starting work. Under an open record the answer is HOLD: this certificate fails, and nothing in the packet says it was the only one. Under a closed record, or with the risk manager's letter that TC adds, it is FAIL.

Witness C's ledger marked the certificate FAILS, left the exclusion field empty, and the resolver held. Witness B's ledger also marked the certificate FAILS, with the reason "The timely commercial general liability certificate names Owner only as certificate holder and lists no additional insured", which is right. It then filled the exclusion field with "Additional insured: none listed." That line is a verbatim span, so it passed the span check, and the resolver read the condition as UNMET and released FAIL. The line describes this certificate. It says nothing about whether another certificate was delivered. The error moved from the status to the exclusion field, and the span check cannot see it.

---

8. Evidence category 3: real SEC exhibits (`E44-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. The label rule is unchanged from Rounds 39 to 43: 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. Windows hold the opening 1,800 characters of the exhibit and 4,600 characters around the effectiveness passage.

The ledgers use a record level form: each condition to effectiveness is quoted with a status, SATISFIED or NOT ESTABLISHED, and an evidence quote. The **gold ledger** resolves to the label. The **seeded ledger** carries one defect that is harder than Round 43's omissions. For a HOLD exhibit, the condition is kept but marked SATISFIED, with a verbatim evidence quote that is a near miss for satisfaction: the text of the requirement itself, an execution phrase, a representation, an acknowledgment, a recital, a signature block, or a clause on signature mechanics. For a CERTIFY exhibit, a spurious condition marked NOT ESTABLISHED is added from a nearby clause that is not a condition to effectiveness. Every seeded quote passes the span check.

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Seeded defect	Link
`W01`	SIGMATRON INTERNATIONAL INC	EX-10.29	2025-08-27	`0000915358-25-000018`	conditions listed	HOLD	SATISFIED on near miss evidence (requirement text)	exhibit
`W02`	Cineverse Corp.	EX-10.1	2025-01-06	`0000950170-25-001415`	effective on its face	CERTIFY	spurious condition	exhibit
`W03`	Shake Shack Inc.	EX-10.1	2025-07-31	`0001620533-25-000065`	conditions listed	HOLD	SATISFIED on near miss evidence (requirement text)	exhibit
`W04`	Pacira BioSciences, Inc.	EX-10.37	2025-02-27	`0001396814-25-000041`	effective on its face	CERTIFY	spurious condition	exhibit
`W05`	Kinsale Capital Group, Inc.	EX-10.1	2025-12-11	`0001669162-25-000064`	conditions listed	HOLD	SATISFIED on near miss evidence (execution phrase)	exhibit
`W06`	Mineralys Therapeutics, Inc.	EX-10.1	2026-08-11	`0001933414-26-000114`	effective on its face	CERTIFY	spurious condition	exhibit
`W07`	WHIRLPOOL CORP /DE/	EX-10.2	2026-05-07	`0000106640-26-000036`	conditions listed	HOLD	SATISFIED on near miss evidence (representation)	exhibit
`W08`	Lantheus Holdings, Inc.	EX-10.1	2025-08-06	`0000950170-25-103544`	effective on its face	CERTIFY	spurious condition	exhibit
`W09`	RPM INTERNATIONAL INC/DE/	EX-10.1	2026-04-08	`0001193125-26-147191`	conditions listed	HOLD	SATISFIED on near miss evidence (requirement text)	exhibit
`W10`	HENRY SCHEIN INC	EX-10.1	2025-05-02	`0001193125-25-111677`	effective on its face	CERTIFY	spurious condition	exhibit
`W11`	Karyopharm Therapeutics Inc.	EX-10.6	2025-10-14	`0001193125-25-239015`	conditions listed	HOLD	SATISFIED on near miss evidence (signature mechanics)	exhibit

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Seeded defect	Link
W12	Sezzle Inc.	EX-10.1	2025-08-04	0001662991-25-000209	effective on its face	CERTIFY	spurious condition	exhibit
W13	Vera Bradley, Inc.	EX-10.16	2025-03-28	0001495320-25-000019	conditions listed	HOLD	SATISFIED on near miss evidence (acknowledgment)	exhibit
W14	STAAR SURGICAL CO	EX-2.1	2025-12-09	0001193125-25-311982	effective on its face, distractor	CERTIFY	spurious condition	exhibit
W15	C. H. ROBINSON WORLDWIDE, INC.	EX-10.1	2025-08-12	0001043277-25-000044	conditions listed	HOLD	SATISFIED on near miss evidence (requirement text)	exhibit
W16	Blink Charging Co.	EX-2.1	2025-04-09	0001641172-25-003426	effective on its face, distractor	CERTIFY	spurious condition	exhibit
W17	ONEOK INC /NEW/	EX-10.1	2025-02-05	0001213900-25-010560	receipt or delivery only	HOLD	SATISFIED on near miss evidence (recital)	exhibit
W18	NorthView Acquisition Corp	EX-2.1	2025-02-19	0001213900-25-014987	effective on its face, distractor	CERTIFY	spurious condition	exhibit
W19	Prologis, Inc.	EX-10.2	2026-10-28	0001193125-26-436160	receipt or delivery only	HOLD	SATISFIED on near miss evidence (representation)	exhibit
W20	Mativ Holdings, Inc.	EX-10.1	2026-02-11	0001000623-26-000009	effective on its face, distractor	CERTIFY	spurious condition	exhibit
W21	Mobile Infrastructure Corp	EX-10.1	2026-09-23	0001493152-26-043917	receipt or delivery only	HOLD	SATISFIED on near miss evidence (signature block)	exhibit
W22	Titan Acquisition Corp.	EX-2.1	2026-06-12	0001829126-26-006454	effective on its face, distractor	CERTIFY	spurious condition	exhibit
W23	INTEGRA LIFESCIENCES HOLDINGS CORP	EX-10.2	2026-10-19	0001193125-26-425271	receipt or delivery only	HOLD	SATISFIED on near miss evidence (execution phrase)	exhibit
W24	Korsana Biosciences, Inc.	EX-2.2	2026-09-11	0001193125-26-389318	effective on its face, distractor	CERTIFY	spurious condition	exhibit

8.2 Validator results

Scoring is fixed in `analyze_results.py`. Flag only arm: any flag, or an unparsed reply, gives HOLD; otherwise the original ledger is resolved. Rewriting arm: the rewritten ledger is resolved by the record level exhibit resolver with every quote checked against the window, and the pre registered gate (S7) is applied to the original and rewritten outcomes. Outcomes are scored against the exhibit label: unsafe release (CERTIFY on a HOLD exhibit), false hold (HOLD on a CERTIFY exhibit), or correct.

Witness, arm, ledger	n	Correct	Unsafe release	False hold	Gate: routed to review
Witness C, flag only, gold	24	24	0	0	not applicable
Witness C, flag only, seeded	24	12	0	12	not applicable
Witness C, rewrite, gold	24	24	0	0	0
Witness C, rewrite, seeded	24	24	0	0	12
Witness B, flag only, seeded	24	12	0	12	not applicable
Witness B, rewrite, seeded	24	24	0	0	12

Both validators flagged all 24 seeded ledgers. Witness C flagged none of the 12 gold CERTIFY ledgers and 8 of the 12 gold HOLD ledgers; those 8 flags could not change the outcome, which was already HOLD. In the flag only arm every near miss evidence seed on a HOLD exhibit was caught (12 of 12 per witness, so 0 unsafe releases where the seeded ledger alone would release 12), and every spurious condition seed on a CERTIFY exhibit stayed a false hold, because a flag can only hold. In the rewriting arm both witnesses removed the false SATISFIED status from all 12 near miss seeds and the spurious condition from all 12 CERTIFY seeds, and no rewrite of a gold ledger changed its outcome. Under the gate, the 12 near miss repairs (release to hold) were accepted and the 12 spurious condition repairs (hold to release) were routed to review, as the gate requires; none of the 144 calls produced an unsafe release with or without the gate. Zero unsafe releases in 48 Witness C rewrites allow a true rate up to 0.061 (S8).

8.3 Seeded defect walk-through: `W21`

`W21` (Mobile Infrastructure Corp, accession `0001493152-26-043917`) reads: "Except as provided in Section 6, this Agreement shall become effective when it shall have been executed by Administrative Agent and when Administrative Agent shall have received counterparts hereof that, when taken together, bear the signatures of each of the other parties hereto." The gold ledger records that condition as NOT ESTABLISHED, since a filed exhibit does not show the agent's receipt of counterparts, and resolves HOLD. The seeded ledger marks the same condition SATISFIED with the evidence quote "IN WITNESS WHEREOF, the parties hereto have caused this Agreement to be duly executed as of the day and year first above written". That sentence is in the exhibit and passes the span check. It is the signature block's opening formula, which appears whether or not counterparts were ever received. The resolver releases CERTIFY. A validator must read the window to see that the evidence does not meet the condition; no structural check can.

---

9. Findings and interpretations

9.1 Measured findings

- **F44.1 (deterministic, E44-A).** All eight certificates pass. Exactly the 8 T items are HOLD under an open record by two completion certificate. Under the record level resolver no T ledger can release FAIL without a verified exclusion quote; under the condition ledger a single quote of the near miss record suffices.
- **F44.2 (live, E44-B).** The blind review agreed with 48 of 48 authored labels and reported defects in 6 of 8 families; 5 of 11 defect notes concern K packets, and the reviewer's own verdicts matched every T label.
- **F44.3 (live, E44-B).** Under the open resolver, record level ledgers held 23 of 24 T readings (Witnesses B, C, D) and released an unsafe FAIL on 1 (Witness B, `N03T`, a record scoped negative quoted as an exclusion). On the same 8 T items the condition ledger released an unsafe FAIL on 12 of 16 readings (Witness B 4, Witness C 8), close to Round 43's 14 of 18 on other families, and declared open direct readers on 8 of 16 (4 per witness, on the same four items).
- **F44.4 (live, E44-B).** Record level ledgers resolved all 24 TC items to FAIL under both resolvers and all 24 T items to FAIL under the closed resolver. Declared open direct readers returned FAIL on all 16 TC readings.
- **F44.5 (live, E44-B).** On K, Witness C was correct on 8 of 8 under both resolvers. Witness B held 4 of 8 and Witness D 3 of 8 under the open resolver, and they released 2 and 3 unsafe FAILs under the closed resolver, mostly on requirements met only by two records together; 5 of their 7 K holds fell on K packets the reviewer had flagged.
- **F44.6 (live, E44-B).** Under the record level ledger no two witnesses shared an unsafe release on T (1, 0, and 0 of 8; error correlation undefined). Under the condition ledger Witnesses B and C were both unsafe on 4 of 8 T items, and as direct readers on the same 4 of 8.
- **F44.7 (live, E44-C).** Across 144 validator calls on the 24 exhibits, no arm and no witness made an unsafe release. The flag only arm caught all 24 near miss evidence seeds across the two witnesses and kept all 24 spurious condition seeds as false holds. The rewriting arm repaired all 48 seeded ledgers and changed no gold ledger's outcome; the gate routed the 24 hold to release repairs to review.

9.2 Interpretation

The near miss result supports Round 43's diagnosis. Asked whether a condition is met, readers turn a defective record into an unmet condition. Asked what each record shows and, separately, whether anything excludes other records, the same kind of reader states the narrower conclusion and lets the resolver decide. The TC variant matters here: the readers did not hold because they had become cautious. Given a record that actually excludes the alternative, all 16 ledgers quoted it and the resolver returned FAIL. The improvement is consistent with a change in what the schema lets a reader say, not a general move toward abstention. The same-item baseline now supports this: on the same 8 T items the condition ledger released FAIL on 12 of 16 readings and the record level ledger on 1 of 24. Declared open direct readers fell in between (8 of 16) and erred on the same four items, so declaring the convention helps a direct reader only partly. With eight families and decisions that share families, this is still directional evidence, not a measured effect size. Part of the gain is structural by design: the record

level resolver cannot release FAIL on a T item without an exclusion quote (S3), so the measured question is whether readers supply false exclusions, and only one did.

The validator result is narrower than it looks. On near miss evidence both arms were safe, and flagging alone was enough. On spurious conditions a flag cannot help, and only the rewriting arm repaired them; the gate then sends every such repair to review because it turns a hold into a release. On these seeds the gate cost 24 reviews and prevented no error, since no rewrite was wrong. Whether the gate earns its cost depends on rewrites that are wrong, which this track did not produce.

The schema moved two errors rather than removing all of them. The first is joint satisfaction: a requirement met by two records read together has no record that satisfies it alone, and a careful reader who follows the definition reports every record as failing. Under an open record that yields a hold. Under a closed record it releases a wrong FAIL, so the closed resolver needs record sets, not single records. The second is exclusion scope: the resolver trusts any verbatim exclusion quote, and a negative line inside one record looks like an exclusion. Both are cases where the fixed rule, not the reader, is wrong in a predictable way, which is the roadmap's next question (P2).

The review result is the one we did not plan for. We expected the contested labels to be the T labels, which encode the open convention. The reviewer accepted every T label and instead found that several K packets reach PASS only by inference: a certificate dated before work is taken as delivered before work; a waiver in the packet is taken as received by the lender. Witness B's and Witness D's holds on K line up with those inferences, and under a strict open reading they are defensible. So the labels most in need of outside review were the PASS labels, and a reader that holds on them may be closer to right than our scoring credits. We keep the authored labels, as the fixed rule requires, and report every K result on uncontested families as well.

9.3 Hypothesis status

Hypothesis	Status
H44.1 record level ledger holds near miss T items	Supported (1 of 24 unsafe across three witnesses, against 12 of 16 for the condition ledger on the same items)
H44.2 determinate where the record supports a decision	Supported for TC (24 of 24, both resolvers) and for T under the closed resolver (24 of 24); partly supported for K (17 of 24 open; Witnesses B and D 5 unsafe under closed)
H44.3 direct readers err on T and not TC	Supported (8 of 16 unsafe on T, 0 of 16 on TC)
H44.4 no cross witness coincidence on T	Consistent under the record level ledger (no shared unsafe release across three witnesses) but not testable as a correlation with one error; under the condition ledger and direct reading, errors coincided on 4 of 8 items
H44.5 labels survive blind review	Supported for verdicts (48 of 48); defects reported in 6 of 8 families, mostly on K packets
H44.6 gated rewriting validator safe on harder seeded defects	Supported on this track (0 unsafe releases in 72 rewrites, with or without the gate); the gate routed 24 correct repairs to review and caught no wrong rewrite, because none occurred

10. Toward Theory v1

- **P7 revised again. A failed record is not a failed condition.** A condition is unmet only on an exclusion or a closed convention; a record level schema lets readers say so. Evidence: F44.1, F44.3, F44.4. Remaining threats: exclusion scope (`N03T`) and joint satisfaction (`N01K`, `N03K`, `N06K`, `N07K`). Removal test: a schema with record sets and a scoped exclusion field (what is excluded, by whom, over what period) on which T unsafe and K false holds both stay near zero across four witnesses.
- **P1 extended. Abstention labels are claims and need certificates.** A HOLD label should carry a two completion certificate, and every label should pass a review blind to it. Evidence: S2, F44.2. Removal test: a pool on which reviewed and unreviewed labels give the same reader rankings.
- **P2, carried to Round 45.** The record level resolver's fixed rules are wrong in two predictable places (record sets, exclusion scope), and on `N03K`, `N06K`, and `N07K` the witnesses that followed the rule erred while the witness that read the records together was right. This is the roadmap's Round 45 question.
- **P5 refined. A validator that can only flag cannot repair a spurious condition.** On harder seeded defects both validator designs were safe; rewriting repaired what flagging could only hold, and the monotone gate (S7) routed every hold to release repair to review. Evidence: F44.7. Removal test: a track where rewrites are sometimes wrong, on which the gate's review queue catches no more errors than it costs.

---

11. Method changes for later rounds

1. Record level ledgers allow a record set: a list of records that together satisfy one requirement.
2. The exclusion field states its scope (which records or events it excludes, who makes the statement, and for what period), and the resolver treats an exclusion as valid only if its scope covers the requirement.
3. K packets include receipt and delivery records wherever a requirement turns on receipt or delivery, so PASS labels do not rest on inference.
4. Every authored label carries a completion certificate and a blind review result; the review is planned per call within the run budget before the round.

---

12. Synthesis review

After the completed analysis, Witness A at maximum reasoning received the recorded structural and live results and was asked for interpretation, threats to validity, and a next check (`synthesis_r44.py`, verbatim in `synthesis-review.json`). Its output is interpretation, not evidence. In summary:

- It agreed that the near miss error largely disappears under the record level format, and cautioned that part of the gain is structural, because the resolver makes HOLD the default when no exclusion is quoted.

- It read the TC result as showing that the completeness record is the operative closing signal, and noted that uniform FAIL on TC cannot separate correct closure from a tendency to read any added record as closing.
- It judged that rewriting dominates flagging on these seeds because a flag cannot remove a spurious condition, and that the gate's value is unmeasured because no rewrite was wrong.
- It proposed a TC control whose added record does not assert completeness, and a second reviewer that judges each exclusion quote for meaning rather than span presence.

Two of its points are corrected here from the recorded data. The closed resolver FAILs on K by Witnesses B and D come from undetermined conditions read as unmet, not from exclusion quotes. Direct readers' errors did coincide: Witnesses B and C erred on the same four T items (Section 7.3). The validator intervals in `live_results.json` are two sided Wilson intervals, while S8 gives exact one sided limits; the paper quotes S8.

---

13. Limitations and threats to validity

- The run budget stopped the round early, and the stopped phases were completed later the same day, after the first results had been seen, with unchanged prompts, settings, and scoring. They complete the fixed design but are not blind to the earlier results. The phases dropped before any test call (re-review, Witness D in the baseline and direct phases, direct K reading) remain unrun.
- Three witnesses wrote record level ledgers and two ran the baseline and direct readers. With one record level error, a cross witness correlation could not be estimated.
- The exhibit validators were never wrong in a way the gate could catch, and the seeded defects follow two known kinds, so the validators may partly be detecting the seeding pattern.
- The review used one model reviewer, not a human legal expert, and found defects it could not have resolved without the re-review that the budget dropped. Six of eight families are contested under the fixed rule; the uncontested subset has 6 items.
- The authored labels for `N03K`, `N06K`, and `N07K` rest on inferences the reviewer flagged; a strict open reading could label them HOLD, which would turn Witness B's and Witness D's open resolver holds into correct holds.
- The pool is small: 8 families, 24 items, 8 items per variant per witness. Decisions that share a family are not independent. Every call was made once at temperature `0`.
- Families, labels, gold ledgers, and seeded ledgers were written by the research team.

---

14. Next check

The roadmap's Round 45 item asks where fixed resolver rules are wrong and witnesses are right (P2). Round 44 supplies two concrete cases. Rerun the eight families (and the eight Round 43 near miss families with new surface forms) with record sets and a scoped exclusion field, add K packets that need two records and T packets whose decisive record carries a record scoped negative, and score both the

record level resolver and a witness's direct reading against labels that are certified by completions and reviewed blind with a re-review budgeted in advance. Add a TC control whose added record does not assert completeness, and an exhibit validator track on which rewrites can be wrong, so the gate's review cost can be weighed against errors it catches.

---

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.
- [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.
- [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.
- [S53] Yuta Koreeda, Christopher D. Manning. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Findings of EMNLP 2021. arXiv:2110.01799.
- [S54] Pranav Rajpurkar, Robin Jia, Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. ACL 2018. arXiv:1806.03822.
- [S55] Yonatan Geifman, Ran El-Yaniv. Selective Classification for Deep Neural Networks. NeurIPS 2017. arXiv:1705.08500.
- [S56] Amita Kamath, Robin Jia, Percy Liang. Selective Question Answering under Domain Shift. ACL 2020. arXiv:2006.09462.
- [S57] Curtis G. Northcutt, Anish Athalye, Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. NeurIPS 2021 Datasets and Benchmarks Track. arXiv:2103.14749.
- [S58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.
- [S59] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2308.11462.
- [S60] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, Samuel J. Bell. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. arXiv:2506.09038, 2025.