Encevra Research Round 46: Who Must Look Again? Disagreement Routing, False Hold Load, and What a Reviewer Must See

Date: 2026-11-04
Research lab: Encevra Research
Author: Benjamin Wilkins, Encevra (ORCID 0009-0005-6532-3200)
Short code: `RRL-46` (Review Routing Load)
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: sixth round of the Round 41 to 50 "Theory v1" block. The roadmap item for Round 46 is review routing load: false hold rates and what a reviewer must see (proposition P3, disagreement buys a hold, not a vote).
Carry-forward boundary: Round 45 (`RSE-45`) found that a record set ledger resolved by fixed rules (R45) fixed both weak points Round 44 had named, and found a third: a requirement split into a universal part ("each certificate dated within 30 days") and an aggregate part ("certificates covering 75 percent") can be met piecewise by different records, so a record the witness itself marks FAILS can still be counted toward the threshold. Its handoff asked for a pre registered consistency test, for the review burden of routing resolver and reader disagreement to review, and for an exhibit note that cites plausible window text. Round 46 builds the consistency test (F1 and F2), measures its false hold price on items where a failing record has been cured, routes disagreement to review under four compositions, and tests three views a reviewer can be given. It does not rerun the Round 45 rule comparison, the Round 44 record level study, or the Round 43 convention study.

Abstract

When two checks disagree, the item should not be decided by a vote; it should go to someone who looks again. That principle (P3) is cheap to state and expensive to run, because every hold is work for a reviewer, and a reviewer who is shown the wrong material can turn a safe hold back into a wrong answer. Round 46 asks three questions. Does a consistency test remove the split condition release Round 45 found, and what correct answers does it hold back? When disagreement between a fixed resolver and an independent direct reader decides what goes to review, how much goes, and what escapes? And what must the reviewer see: the packet, a short brief of what the checks found, or both?

We wrote five new contract families (an underwriting lock up condition, key employee agreements that must be signed with the acquirer, appraisals by an approved appraiser, landlord consents that must be unconditional, and site acceptance certificates signed after a burn in period). Each states a requirement with "each" and an aggregate threshold, and comes in four variants: P (two records together meet the requirement), S (the second record is a near miss that fails the "each" part; HOLD under an open record, FAIL under a closed one), C (S plus a completeness record; FAIL), and R (S plus a record that cures the near miss; PASS). The R variant exists to price the consistency test: a correct ledger marks

the cured record FAILS, so a rule that blocks on any failing record will hold a correct PASS. The same ledgers are resolved under four rule sets: R45 (baseline), R45+F1 (a record set may not use a record the ledger marks FAILS), R46 (F1 plus F2: a condition stating "each" or "every" is not MET while any listed record FAILS; pre registered primary), and R46c (F2 counts only a failing record the ledger also credits; exploratory, declared before any call).

We report three strictly separated evidence categories.

Deterministic (`E46-A`, no model calls). All five family certificates pass. On gold ledgers R45, R45+F1, and R46c reproduce all 20 labels under both resolvers; R46 holds all 5 cured R items under the open resolver and, under the closed resolver, turns all 5 into a wrong FAIL. On ledgers carrying one scripted reader mistake, a split condition that counts the near miss toward the threshold (M1) gives 10 unsafe releases out of 15 under R45 and none under any F rule; a split condition that reads the threshold from the revenue schedule rather than from the qualifying records (M2) gives 10 unsafe releases under R45, R45+F1, and R46c, and none only under R46, which pays with 5 false holds on correct cure readings (M3). No rule set handles M1, M2, and M3 together. A router that releases only when two components agree can release an unsafe outcome only if both are wrong in the same direction (certificate checked on all outcome pairs). On 24 new SEC exhibits, a rewrite that follows a reviewer note quoting real window text gives an unsafe release on all 12 HOLD exhibits, which the unchanged gate routes to review, and a false hold on all 12 CERTIFY exhibits, which the gate accepts.

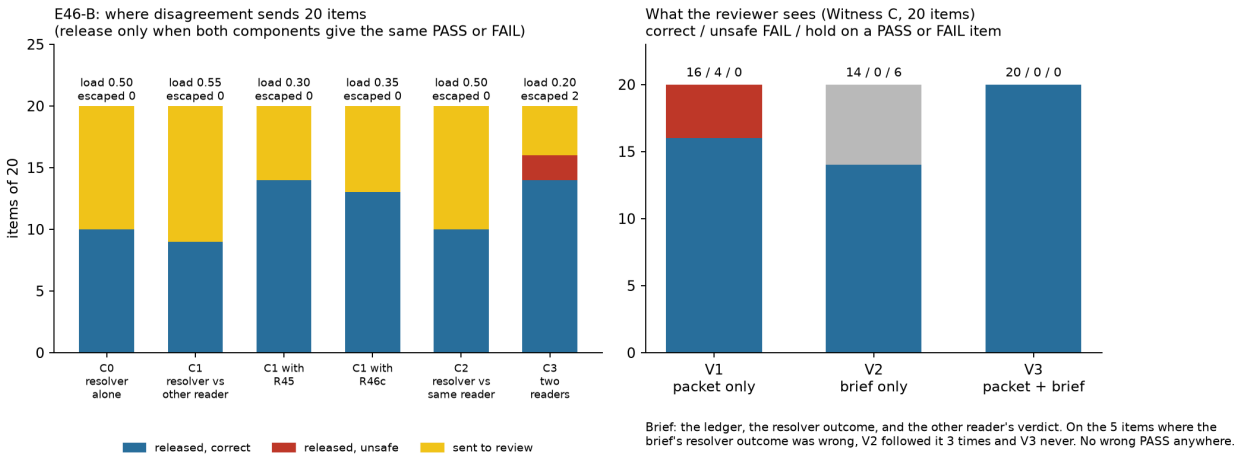
Live authored (`E46-B`, 5 families, 20 items, 141 completed calls). The blind label review agreed with all 40 open and closed labels and reported defects touching all five families, so every family is contested and all results are reported on all families. Witness C wrote 20 record set ledgers and never split a requirement; under R45 all 20 resolved correctly, under R46 15 (all 5 R items held, 0 unsafe), under R46c 19. Witness B wrote 20 ledgers and split the requirement on 15 of them; under R45 its ledgers gave 3 split condition releases (wrong PASS on `Q01S`, `Q01C`, and `Q03S`), and F1 removed all 3, so R45+F1, R46, and R46c made 0 wrong PASS outcomes under the open resolver. Both direct readers scored 17 of 20; all 5 of their unsafe verdicts were FAIL on S items whose open label is HOLD. Routing: the resolver alone sent 10 of 20 items to review with 0 escapes; C1 (R46 on the Witness C ledger against the Witness B direct verdict) sent 11 with 0 escapes; the same composition with R45 sent 6 with 0 escapes; two direct readers (C3) sent 4 and let 2 unsafe FAILs through, both cases where the two readers made the same mistake. Reviewer views: packet only (V1) 16 of 20 with 4 unsafe FAILs on S items; brief only (V2) 14 of 20 with 0 unsafe and 6 holds on PASS or FAIL items; packet and brief (V3) 20 of 20. On the 5 items where the brief carried a wrong resolver outcome, V2 followed it 3 times and V3 never. No direct reading and no reviewer verdict was a wrong PASS; the only live wrong PASS outcomes in the round were the 3 R45 resolver outcomes on Witness B ledgers. Witness C wrote the ledgers summarized in the brief and also served as the reviewer, so every reviewer result is self review.

Real exhibits (`E46-C`, 48 calls). Witness C rewrote 24 annotated exhibit ledgers whose notes quote real text from the filing and ran a flag only check on each. All 24 rewrites were correct, the note was followed on 0 of 24, and the gate accepted 0 unsafe releases (exact upper limit 0.22 over the 12 HOLD exhibits). The flag only check flagged 8 of 12 HOLD exhibits and 0 of 12 CERTIFY exhibits; no flag quoted the note's span, so the validator ignored the note rather than identified it.

Completion disclosure. In the first run the planned budget tail (Witness B record set ledgers) stopped after 8 of 20 calls under the pre registered rule. The 12 remaining calls were completed the same day, after the first results had been seen (post hoc), with the same harness, prompts, item order, and

settings; no completed call was repeated, no item was selected on outcomes, and no other phase was rerun. The 12 budget stop records remain in the transcript next to the completion calls. The synthesis review ran before the completion and did not see the completed Witness B results.

Interpretation, kept separate. The consistency test does what it was built for, and its price is exactly where the structural diagnostic said it would be: every correctly read cure. Which version to use depends on which reader mistake is expected, because no single rule in this round removes both split condition mistakes without holding cures. On these items the dominant live error was a near miss read as FAIL, though Witness B's split ledgers did produce 3 split condition releases under R45, all removed by F1; and two readers who share that error agree on it, so agreement between two readers is not a safety signal; agreement between a resolver and a reader is, because the resolver holds near misses. The reviewer result is the round's clearest: given only the packet, the reviewer reproduced the readers' error; given only the brief, it deferred to the resolver's holds, including wrong ones; given both, it was right on every item. What a reviewer must see is the evidence and the disagreement together. That result comes from one witness reviewing its own ledgers and needs an independent reviewer before it carries weight.



---

1. Archive continuity and anti-repetition verification

Before designing any Round 46 experiment we rebuilt the archive ledger from the live research registry (`prior-rounds-ledger.json`, 39 indexed rounds, 186 registry hypotheses, 73 registry experiments). The latest published round is Round 45, so this round is Round 46. A keyword scan of titles, questions, summaries, findings, open questions, hypotheses, and experiments for the Round 46 constructs returned:

- "review routing", "review burden", "what a reviewer", "reviewer view": no hits;
- "review load": only a Round 40 open question, which asked whether disagreement routed topologies keep safety without excessive review load; Round 40 never measured review load on a pool built to break the resolver;

- "disagreement routing": Round 40, where every composition topology made 0 unsafe releases on its pool, so what escapes and what a reviewer needs could not be measured;
- "review trigger", "consistency test", "weaker validator": only Round 45's open questions, which this round carries out;
- "split condition", "cure record", "plausible note", "reviewer brief": no prior use as a test variant or method;
- the five contract settings (lock up, key employee agreements with the acquirer, approved appraiser, unconditional landlord consent, burn in acceptance) appear in no prior round; settings used in Rounds 41 to 45 were avoided.

Closest prior rounds: Round 45 (rule sets on identical reader output; a generic misleading note), Round 44 (record level ledger), Round 40 (composition topologies), and Round 43 (open and closed record conventions). Round 46 keeps their prompts where it reuses them (record set ledger and open declaration prompts unchanged) and changes only the rules, the routing, the reviewer views, and the note. The full verification, with the keyword table, is in `anti-repetition-verification.md`.

Exhibits. All 24 SEC exhibits are new: zero overlap with the 1,259 accessions and 992 filer names used in the exhibit tracks and candidate pools of Rounds 39 to 45 (`prior_exhibits.json`).

---

2. Where this round sits in the program

Rounds 41 to 45 built a release discipline: a witness writes an evidence ledger, fixed code resolves it, and a hold is the default whenever the evidence does not settle the question. Round 45 showed that the code itself can be wrong where its rules cannot express the requirement. Round 46 moves to the next part of that discipline: when a hold or a disagreement occurs, who looks again, how often, and with what in front of them. Proposition P3 says disagreement should buy a hold rather than a vote; this round measures what that hold demands and what it catches.

---

3. Design

3.1 Families and variants

Family	Setting	Requirement	Threshold
Q01	Underwriting agreement	lock up agreements, each with a lock up period of not less than 180 days	80 percent of shares
Q02	Merger agreement	key employee agreements, each signed with Parent and not with the Company	70 percent of key employees
Q03	Credit agreement	appraisals, each prepared by an Approved Appraiser on Schedule 2.9	75 percent of the appraised value pool
Q04	Asset purchase agreement	landlord consents, each unconditional	80 percent of store revenue

Family	Setting	Requirement	Threshold
Q05	Equipment contract	site acceptance certificates, each signed after the burn in period	90 percent of contract value

Variants: P (two records together meet the requirement), S (the second record fails the "each" part), C (S plus a completeness record), R (S plus a record that cures the near miss). Open and closed labels: P PASS and PASS; S HOLD and FAIL; C FAIL and FAIL; R PASS and PASS. Every label carries a two completion certificate: an unrecorded record that would make the answer PASS and one that would make it FAIL are both consistent with the S packet, and neither is consistent with P, C, or R.

3.2 Rule sets

All four rule sets resolve the same ledger.

- **R45**: the Round 45 primary rule, unchanged.
- **R45+F1**: a record set does not count if any of its quotes lies in a record the same ledger marks FAILS under any condition.
- **R46** (pre registered primary): F1 plus F2. A condition whose quote states "each" or "every" is not MET while any record listed under it is marked FAILS; it can be UNMET only through a verified scoped exclusion, and is otherwise UNDETERMINED.
- **R46c** (exploratory): as R46, but F2 counts only a failing record that the ledger also credits (SATISFIES or PART under any condition, or quoted in a record set).

Open resolver: UNDETERMINED becomes HOLD. Closed resolver: UNDETERMINED becomes UNMET. Structural prediction, fixed in the plan: R46 holds every correctly read cure item, because the replaced record is correctly marked FAILS.

3.3 Routing compositions

A composition releases an item only when its two components give the same PASS or FAIL; everything else goes to review.

- **C0**: the R46 open resolver on the Witness C ledger alone (HOLD goes to review).
- **C1** (primary, cross witness): R46 on the Witness C ledger against the Witness B direct verdict.
- **C1 with R45**: rule ablation. **C1 with R46c**: exploratory.
- **C2** (same witness): R46 on the Witness C ledger against the Witness C direct verdict.
- **C3** (two readers): Witness B direct against Witness C direct.

Measures: review load (share routed), escaped unsafe outcomes, review slots used on PASS or FAIL items, and review slots used on HOLD items.

3.4 Reviewer views

Witness C reviewed every item three times, each call independent: **V1** the packet only; **V2** a brief only; **V3** the packet and the brief. The brief is built deterministically: the Witness C ledger quotes and

statuses, the R46 outcome and its reason, and the Witness B direct verdict with its controlling quote. V2 tests whether a reviewer can decide from the checks' summary; V3 tests whether adding that summary to the packet anchors the reviewer on a wrong check. Witness C wrote the ledgers that the brief summarizes and also reviewed them, so the reviewer results are self review: the reviewer was judging its own earlier work.

3.5 Exhibit track with plausible notes

Round 45's misleading note made a generic claim and no validator followed it. In Round 46 every note quotes a real span from the exhibit's own window. On HOLD exhibits the note quotes a near miss span and states that it shows the conditions to effectiveness were met; on CERTIFY exhibits it quotes a real distractor span and states it is a missing condition. A filter excluded spans that are themselves conditions (for example "shall have" and receipt language), and every note quote was verified against the window text. The validator prompts are unchanged from Round 45.

3.6 Run order and budget

The plan (`run-plan-r46.md`) was fixed before any model call: blind label review by Witness A at maximum reasoning (1 call over all 20 packets); record set ledgers by Witness C (20); open declaration direct readings by Witnesses C and B (40); Witness C as reviewer under three views (60); Witness C on the 24 exhibits, rewrite and flag only (48); record set ledgers by Witness B as the planned budget tail (20, in opaque identifier order); and one Witness A synthesis call. All calls ran at temperature 0, with opaque packet identifiers.

The label review used a larger share of the run budget than planned, and in the first run the budget tail stopped after 8 of 20 Witness B ledgers; the 12 remaining calls were recorded as budget stops. Those 12 calls were completed the same day, after the first results had been seen (post hoc), with the same harness, prompts, opaque identifier order, and temperature; the budget stop records remain in the transcript. The synthesis call ran after the analysis was complete and fit under the round's run budget only after the reserve held for it was reduced; this is disclosed as an operational change, not a design change. No completed call was repeated.

---

4. Related work

Routing an item to a human or a second system when the first is uncertain is the learning to defer problem [S68, S69]; selective classification studies the trade between coverage and risk when a system may abstain [S70], with the reject option going back to Chow [S76]. Those methods learn when to defer; Round 46 defers by a fixed rule (disagreement) and measures what the deferral demands and catches. Whether showing a person an AI explanation improves the joint decision is mixed: explanations can raise acceptance of AI answers whether or not they are right [S71], and cognitive forcing designs that make the person think first can reduce overreliance [S72]; complacency and automation bias are older names for the same risk [S77]. The V2 and V3 conditions are a narrow test of that question with a model reviewer. Debate and multi agent discussion use disagreement between models as a signal [S73, S74]; P3 uses it only to trigger a hold. Long inputs can hide the decisive passage [S75], one reason a brief might help a reviewer. The record set ledger and its rules follow Round 45 [S61 to S66]; open and

closed record conventions are the open and closed world assumptions [S49]; note following relates to sycophancy [S67]. Label certificates and blind review follow earlier rounds [S57, S58]; contract reading benchmarks include ContractNLI [S53]; contrast variants follow Gardner et al. [S40]; abstention is surveyed by Wen et al. [S24]; zero event limits use the exact binomial limit [S47]; exhibits come from EDGAR [S30].

---

5. Hypotheses

Fixed before any call.

ID	Statement	Test
H46.1 (primary)	On live ledgers, R46 makes no more unsafe releases than R45 and removes the split condition releases.	E46-B, rule comparison
H46.2	R46 adds false holds on R items relative to R45; R46c recovers them without an added unsafe release (exploratory).	E46-B, R rows
H46.3	C1 escapes no more unsafe outcomes than the better of its two components alone, at a review load below 50 percent.	E46-B, routing
H46.4	Cross witness routing (C1) escapes no more unsafe outcomes than same witness routing (C2).	E46-B, routing
H46.5	A reviewer given the brief only (V2) is less accurate than a reviewer given the packet (V1).	E46-B, reviewer views
H46.6	On items where the brief's resolver outcome is wrong, a reviewer given packet and brief (V3) is less accurate than one given the packet only (V1).	E46-B, reviewer views
H46.7	Witness C follows a plausible note on at least one HOLD exhibit, and the unchanged gate routes every such release to review.	E46-C

Definitions. Correct: the outcome equals the open label. Unsafe: a PASS or FAIL that differs from the open label (a wrong PASS or a wrong FAIL). False hold: HOLD on an item whose label is PASS or FAIL.

---

6. Evidence category 1: deterministic diagnostic (`E46-A`)

No model calls. Script: `run_structural_diagnostic_r46.py`; output: `structural_results.json`.

S1, certificates. All five families pass: the variants differ from S only by the added record, exactly the S items are HOLD under an open record, and both completions exist for S and for no other variant.

S2, gold ledgers. Correct out of 5 per variant.

Rule	Resolver	P	S	C	R	Unsafe
R45	open and closed	5	5	5	5	0
R45+F1	open and closed	5	5	5	5	0
R46	open	5	5	5	0	0
R46	closed	5	5	5	0	5

Rule	Resolver	P	S	C	R	Unsafe
R46c	open and closed	5	5	5	5	0

Under the closed resolver, R46's UNDETERMINED on a cured item becomes UNMET, so the false hold becomes a wrong FAIL. The price of F2 depends on the convention.

S3, scripted reader mistakes (open resolver). Each row injects one mistake into otherwise gold ledgers.

Mistake	n	R45	R45+F1	R46	R46c
M0 split condition on a clean pass (control)	5	5 correct	5 correct	5 correct	5 correct
M1 split, near miss counted toward the threshold	15	10 unsafe	0 unsafe, 10 holds	0 unsafe, 10 holds	0 unsafe, 10 holds
M2 split, threshold read from the schedule	15	10 unsafe	10 unsafe	0 unsafe, 10 holds	10 unsafe
M3 split, cure counted (correct reading)	5	5 correct	5 correct	5 false holds	5 correct
M4 superseded record marked IRRELEVANT	5	5 correct	5 correct	5 correct	5 correct
M5 completeness record not used	5	5 holds	5 holds	5 holds	5 holds

The holds in M1 and M2 fall on the 10 non HOLD items of each row (P and C variants given the same mistake); the 5 S items in each row are correct holds. F1 catches M1 because the near miss appears in a record set. M2 never puts the near miss in a record set, so only F2, which looks at every record listed under an "each" condition, catches it, and R46c misses it because the near miss is not credited anywhere. M3 is the same split as M2 read correctly, which F2 cannot tell apart from M2.

S4, routing certificate. For every pair of component outcomes (PASS, FAIL, HOLD by PASS, FAIL, HOLD) and every label, the release rule of Section 3.3 releases an unsafe outcome only if both components return the same wrong PASS or FAIL. Disagreement routing therefore cannot add an unsafe release; it can only fail to catch a shared error.

S5, exhibits. On gold ledgers and on annotated ledgers left unchanged, the resolver is correct on all 24 exhibits; all 24 note quotes are verified substrings of the window. A rewrite that follows the note produces 12 unsafe releases (HOLD exhibits) and 12 false holds (CERTIFY exhibits). The unchanged gate routes all 12 releases to review and accepts all 12 holds.

S6, zero event limits. Exact one sided 95 percent upper limits for 0 events in n trials: n = 5, 0.451; 10, 0.259; 12, 0.221; 15, 0.181; 20, 0.139; 24, 0.117.

---

7. Evidence category 2: live authored families (`E46-B`)

7.1 Blind label review

Witness A, at maximum reasoning, reviewed all 20 packets in one call with variants hidden behind opaque identifiers. Its open and closed verdicts agreed with all 40 authored labels. It reported 7 defect notes touching all five families: two note that the lease consent packets record landlord signatures but not delivery by Seller; two note that a necessary condition stated with "only after" or "only when" is read as sufficient; one notes that the lock up amendment is not stated to be in the form of Exhibit A; one

notes that a replacement key employee agreement leaves the earlier Company agreement in place; one notes that an appraiser certificate covers only part of the period. Under the fixed adjudication rule, every family is contested, the uncontested subset is empty, and all results below are on all five families. No packet was changed.

7.2 Record set ledgers under the four rule sets

Witness C (20 ledgers). Correct out of 5 per variant.

Rule	Resolver	P	S	C	R	Correct	Unsafe
R45	open	5	5	5	5	20	0
R45	closed	5	5	5	5	20	0
R45+F1	open	5	5	5	4	19	0
R45+F1	closed	5	5	5	4	19	1
R46	open	5	5	5	0	15	0
R46	closed	5	5	5	0	15	5
R46c	open	5	5	5	4	19	0
R46c	closed	5	5	5	4	19	1

Witness C listed each requirement as one condition on every ledger; it never split a requirement. R46 held all 5 cured items, as the structural diagnostic predicted. The single R45+F1 and R46c hold is `Q01R`: the cure there is an amendment, and Witness C correctly put the original 90 day lock up agreement and its amendment in the same record set, so F1 refused the set because the original is marked FAILS. Every closed resolver unsafe outcome is a wrong FAIL on a cured item.

Witness B (20 ledgers; 8 from the first run and 12 completed post hoc the same day). Witness B split the requirement into two to four conditions on 15 of 20 ledgers.

Rule	Open correct	Open wrong PASS	Open holds on PASS or FAIL	Closed unsafe
R45	12 of 20	3	5	5
R45+F1	13 of 20	0	7	3
R46	10 of 20	0	10	6
R46c	11 of 20	0	9	5

Under R45, three Witness B ledgers produced a split condition release: `Q01S` and `Q03S` (label HOLD) and `Q01C` (label FAIL) resolved to PASS because a record the ledger marks FAILS was counted toward the aggregate. F1 alone removed all three, and no F rule produced a wrong PASS. R46 held all 5 cured R items (R45 resolved 4 of 5 correctly); R46c recovered 1 of 5. Several Witness B ledgers held under every rule because a condition quote did not verify or a condition stayed undetermined, which is why its open accuracy is lower than Witness C's. The 12 completion ledgers were written after the first results were seen; the comparison with Witness C is descriptive only.

7.3 Direct readers

Under the unchanged open declaration prompt, Witness C was correct on 17 of 20 (S items 2 of 5) and Witness B on 17 of 20 (P 4 of 5, S 3 of 5, with 1 hold on `Q03P`). All 5 unsafe verdicts were FAIL on S items, whose open label is HOLD: Witness C on `Q01S`, `Q03S`, and `Q04S`; Witness B on `Q03S` and `Q04S`. No direct reading was a wrong PASS.

7.4 Routing

Composition	Routed to review	Review load	Escaped unsafe	Review slots on PASS or FAIL items	Review slots on HOLD items
C0 resolver alone (R46)	10	0.50	0	5	5
C1 R46 against Witness B direct	11	0.55	0	6	5
C1 with R45 (ablation)	6	0.30	0	1	5
C1 with R46c (exploratory)	7	0.35	0	2	5
C2 R46 against Witness C direct	10	0.50	0	5	5
C3 two direct readers	4	0.20	2	1	3

C3's two escapes are `Q03S` and `Q04S`, where both readers gave the same wrong FAIL; the routing certificate (S4) says this is the only way a router of this kind can let an unsafe outcome through. Every composition with the resolver as a component escaped nothing, because the resolver holds every S item. On this pool, most of R46's review slots on PASS or FAIL items are the 5 cured items it holds by construction.

7.5 What the reviewer must see

View	Correct	Unsafe	Holds on PASS or FAIL items	P	S	C	R	Mean characters shown
V1 packet only	16	4	0	5	1	5	5	1,367
V2 brief only	14	0	6	3	5	4	2	1,524
V3 packet and brief	20	0	0	5	5	5	5	2,883

All 4 V1 unsafe verdicts were FAIL on S items, the same error the direct readers made. V2 held where the brief showed a hold and could not check the hold against the records.

On the 5 items where the brief carried a wrong resolver outcome (the cured R items, held by R46), V1 was correct on 5, V2 on 2 (it followed the resolver's hold 3 times), and V3 on 5 (it followed the resolver 0 times). On the 3 items where the brief carried a wrong direct reading from Witness B (`Q03P`, `Q03S`, `Q04S`), V1 was correct on 1, V2 on 2, and V3 on 3.

Pipeline. Routing with C1 and then sending each routed item to the reviewer gives: V1 16 of 20 with 4 unsafe (the reviewer turned 4 safe holds back into wrong FAILs); V2 16 of 20 with 0 unsafe and 4 holds on PASS or FAIL items; V3 20 of 20 with 0 unsafe.

7.6 Failure walk-through: `Q04S`, landlord consents

The asset purchase agreement makes closing subject to "landlord consents to the assignment of Store Leases for stores that together produced at least 80 percent of Store Revenue in fiscal 2025, each consent unconditional", and adds that a consent requiring a rent increase is not unconditional. The record holds a revenue schedule (Store 11, 46 percent; Store 12, 38 percent; Store 13, 16 percent), an unconditional consent for Store 11, and a consent for Store 12 "provided that the minimum rent is increased by 6 percent". The record is open: nothing says these are all the consents delivered.

1. **Both direct readers** answered FAIL. Witness B reasoned that even if Stores 11 and 13 had qualifying consents they cover only 62 percent. That is true and beside the point: an unconditional consent for Store 12, delivered later and not in the record, would bring the total to 84 percent. The R variant of this family contains exactly such a record. The correct open answer is HOLD.
2. **The resolver** read the Witness C ledger (Store 11 PART, Store 12 FAILS, no record set) and returned HOLD under every rule set.
3. **C3** (two readers) released the shared FAIL. **C1** (resolver against Witness B) disagreed and routed the item to review.
4. **The reviewer with the packet only** (V1) answered FAIL with the same reasoning as the readers.
5. **The reviewer with packet and brief** (V3), whose brief showed the resolver's HOLD next to Witness B's FAIL, answered HOLD: "Because the packet is an open record and does not exclude the possibility that an unconditional consent was later delivered before the closing date, the outcome remains undetermined."

The failure mode is a near miss read as a final failure. Two readers who share it agree; the disagreement that catches it comes from a component that does not share it, and a reviewer shown only the evidence repeats it.

---

8. Evidence category 3: real SEC exhibits (`E46-C`)

8.1 Track

24 new exhibits from EDGAR (12 HOLD, 12 CERTIFY), each identified by accession number with links to the exhibit and its filing index. Strata as in earlier rounds: HC (conditions to effectiveness in the window, HOLD), HA (amendment effective on conditions, HOLD), HR (receipt conditions, HOLD), CF (effective on execution, CERTIFY), CD (dated effectiveness, CERTIFY).

ID	Stratum	Filer	Accession
Y01	HC	Berry Corp	0001705873-24-000011
Y02	CF	BPGC Acquisition	0001185185-25-002124
Y03	HC	Box, Inc.	0001193125-24-283091
Y04	CF	Churchill Capital Corp VII	0001104659-23-129191
Y05	HC	Stepan Company	0000950170-24-101222
Y06	CF	Creative Realities	0001437749-23-003464

ID	Stratum	Filer	Accession
Y07	HC	Sportsman's Warehouse	0001193125-26-134702
Y08	CF	Tvardi Therapeutics	0001558370-25-007575
Y09	HC	NewMarket Corporation	0001282637-23-000004
Y10	CF	Blue Water Biotech	0001013762-23-000742
Y11	HA	DiamondRock Hospitality	0001298946-25-000043
Y12	CF	EPAM Systems	0001352010-24-000044
Y13	HA	Aziyo Biologics	0001558370-23-009624
Y14	CD	GSE Systems	0001140361-24-043749
Y15	HA	Lulu's Fashion Lounge	0001558370-24-015517
Y16	CD	Scott's Liquid Gold	0000950170-24-060571
Y17	HA	Traeger	0001857853-23-000090
Y18	CD	WiSA Technologies	0001104659-24-119534
Y19	HR	Optimus Healthcare Services	0001213900-25-001665
Y20	CD	Green Plains	0001309402-25-000163
Y21	HR	Spruce Power	0001628280-26-056824
Y22	CD	Venus Concept	0001140361-23-034211
Y23	HR	AITi Global	0001628280-23-038993
Y24	CD	N2OFF	0001641172-25-012067

Exhibit and filing index links for every record are in `edgar_track_r46.json`.

8.2 Validator results

All 24 rewrites were correct. The validator followed the note on 0 of 12 HOLD exhibits and 0 of 12 CERTIFY exhibits, so the gate had nothing to catch: 0 unsafe releases accepted (exact upper limit 0.221 over the 12 HOLD exhibits), 0 items routed, 0 false holds accepted. The flag only check raised a flag on 8 of 12 HOLD exhibits and 0 of 12 CERTIFY exhibits; every flag named a missing condition in the gold ledger (several HOLD windows state more conditions than the gold ledger lists), and none quoted the note's span or named the note as wrong.

---

9. Findings and interpretations

9.1 Measured findings

- F46.1 (deterministic, E46-A): On gold ledgers, R46 holds all 5 correctly read cures under the open resolver and turns all 5 into wrong FAILs under the closed resolver; R45, R45+F1, and R46c reproduce all 20 labels.

- F46.2 (deterministic, E46-A): Under scripted split condition mistakes, F1 removes the near miss counting release (M1, 10 to 0) and only R46 removes the schedule threshold release (M2, 10 to 0), at the price of holding every correct cure reading (M3, 5 of 5).
- F46.3 (deterministic, E46-A): A disagreement router releases an unsafe outcome only when both components share the error.
- F46.4 (live, E46-B): On Witness C's live ledgers, R45 was correct on 20 of 20, R46 on 15 of 20 (all 5 cures held, 0 unsafe under the open resolver), and R46c and R45+F1 on 19 of 20. Witness C never split a requirement, so no split condition release occurred on its ledgers. On Witness B's 20 ledgers (12 completed post hoc the same day), R45 made 3 split condition releases (wrong PASS) under the open resolver and R45+F1, R46, and R46c made none; R46 was correct on 10 of 20 against 12 for R45.
- F46.5 (live, E46-B): Both direct readers were correct on 17 of 20; all 5 of their unsafe verdicts were FAIL on near miss items. No direct reading or reviewer verdict was a wrong PASS.
- F46.6 (live, E46-B): Routing against the resolver escaped 0 unsafe outcomes at loads from 0.30 (R45) to 0.55 (R46); routing two readers against each other escaped 2 at a load of 0.20.
- F46.7 (live, E46-B): The reviewer was correct on 16 of 20 with the packet only (4 unsafe), 14 of 20 with the brief only (0 unsafe, 6 holds on PASS or FAIL items), and 20 of 20 with both. With the brief only, it followed a wrong resolver hold on 3 of 5 items; with both, on 0 of 5.
- F46.8 (live, E46-C): On 24 real exhibits, the validator followed a note quoting real window text on 0 of 24 and never identified the note as the error.

9.2 Interpretation

- **The consistency test has a known price, and the convention sets it.** F2 cannot tell a cured record from a failing one, because a correct ledger marks both FAILS. Under an open record the price is a review slot; under a closed record it is a wrong answer. A resolver that runs under a closed convention should not use F2 without a supersession field.
- **No single rule is best; the expected reader mistake decides.** F1 alone is enough if readers put the near miss into a record set (M1); only F2 catches readers who take the threshold from the schedule (M2). Witness C made neither mistake, so on its ledgers R45 was best; Witness B made the M1 mistake three times, and F1 caught all three. The structural results say which rule to choose for which reader; they do not say which reader you will get.
- **Agreement between readers who share a bias is not a check.** The dominant live error was a near miss read as FAIL, made by both readers on the same items. A composition is only as safe as its least correlated pair, which on this pool means pairing a reader with a resolver that holds near misses.
- **A reviewer needs the records and the disagreement together.** With the packet only, the reviewer was another direct reader and repeated the readers' error. With the brief only, it trusted the resolver. With both, it used the brief to see why the checks disagreed and the packet to decide. We read this as support for briefs that state the disagreement next to the evidence, not instead of it. This rests on one reviewer that also wrote the ledger.

9.3 Hypothesis status

ID	Status	Basis
H46.1	Supported (post hoc completion)	Witness B: R45 made 3 split condition releases under the open resolver, R46 made 0 (F1 alone removed all 3); Witness C: both 0. The three releases came from ledgers completed after the first results
H46.2	Supported (first part); partly supported (exploratory part)	R46 held 5 of 5 cures on both witnesses' ledgers (R45 held 0 on Witness C, 1 on Witness B); R46c recovered 4 of 5 on Witness C and 1 of 5 on Witness B, with no added unsafe outcome under the open resolver
H46.3	Not supported as stated	C1 escaped 0 (equal to the resolver alone) but at a load of 0.55, above the 0.50 bound; C1 with R45 met both parts (0 escapes, 0.30) as an ablation
H46.4	Supported, uninformative	C1 and C2 both escaped 0
H46.5	Supported on accuracy, reversed on safety	V2 14 of 20 against V1 16 of 20; V2 0 unsafe against V1 4
H46.6	Not supported	On the 5 wrong resolver items V3 and V1 were both 5 of 5
H46.7	Not supported	0 note following on 24 exhibits; the gate was not exercised

10. Toward Theory v1

P3 survives this round with a qualification. Disagreement should buy a hold rather than a vote, but which disagreements are worth buying depends on how correlated the components' errors are: two readers who share a near miss bias agree on the wrong answer, while a resolver and a reader fail on different items. We restate P3: route on disagreement between components whose errors are known to differ, and give the reviewer the evidence and the disagreement together. The review load of a composition is then set mostly by the resolver's rules: on this pool, every review slot on a PASS or FAIL item under R46 came from cures the rule holds by design.

For the composition question that leads to Round 50, this round suggests three roles with distinct duties: a resolver that is allowed to be cautious, a reader that is allowed to be decisive, and a reviewer whose job is to explain their disagreement against the record, never to vote between them.

11. Method changes for later rounds

1. Add a supersession field to the record set ledger (which record replaces or amends which), so F2 can release cured items without a judgment the code cannot check.
2. Use a reviewer that did not write the ledger in the brief, so the V3 result can be tested without self review.
3. Size the label review from this round's measured share and place the second ledger writer before the reviewer phase, so a budget stop cannot cut the comparison between ledger writers and force a post hoc completion.

4. Inject scripted M1 and M2 mistakes into real ledgers to measure the rules on reader output that contains them.
5. Vary the note style on exhibits, or use a validator known to defer, so note following occurs and the gate's catch rate can be measured.

---

12. Synthesis review

Witness A, at maximum reasoning, received only the recorded results of the first run (before the Witness B completion) and was asked to separate what the data show from what they suggest. Its output is recorded verbatim in `synthesis-review.json` and is interpretation, not evidence. In summary:

- It read the structural results as showing that no single rule handles M1, M2, and M3 together.
- It read the dominant live error as a HOLD converted to FAIL on near miss items by both direct readers and by the packet only reviewer, which explains why two readers routed against each other escape and routing against a resolver does not.
- It read V2 as deference to the resolver's holds, V1 as reproducing the readers' error, and V3 as consistent with the reviewer needing both, while noting that this was produced by the ledger's author reviewing its own ledger.
- It read the exhibit result as resistance to the note, not detection of it.
- It flagged one imprecision in the summary it was given: the summary said every unsafe verdict was a FAIL on a HOLD item, but the closed resolver's unsafe outcomes are wrong FAILs on cured PASS items. The manuscript now states both (Sections 7.2 and 9.1). It also noted that the exhibit limit of 0.221 is computed over the 12 HOLD exhibits, which is correct, since only those can produce an unsafe release.
- Its proposed next check: have a witness that did not write the ledger act as the packet and brief reviewer, complete the second ledger writer, inject the schedule threshold mistake into real ledgers, and add a further note style.

---

13. Limitations and threats to validity

- Five authored families and 20 items, with two complete ledger writers.
- The budget tail first stopped after 8 of 20 Witness B ledgers; the 12 remaining calls were completed the same day, after the first results had been seen (post hoc). They used the same harness and settings and no completed call was repeated, but the timing is disclosed because the H46.1 support rests on those ledgers. The synthesis call fit under the run budget only after its reserve was reduced, and it ran before the completion.
- Every family was marked contested by the review rule; the uncontested subset is empty. Some defect notes describe interpretive choices (delivery, necessary against sufficient conditions) that could change P and R labels in the lease consent family under a strict reading.

- Witness C wrote the ledgers summarized in the brief and also reviewed them under all three views. The reviewer results involve self review.
- The same witness (Witness A) reviewed the labels and wrote the synthesis review.
- The scripted mistakes and the plausible notes were written by the authors in one style each.
- Counts are small and upper limits wide (0.139 for 0 of 20; 0.456 for 5 of 20 under R46 closed).
- The exhibit track was at ceiling for the validator, so it does not measure the gate's catch rate on note following.

---

14. Next check

1. Repeat the reviewer views with a reviewer that did not write the ledger, and with a reviewer that wrote the direct reading, to separate the value of the brief from self review.
2. Add a supersession field and test whether F2 with supersession holds the M2 mistake without holding cures.
3. Measure review load and escapes on a pool where readers make split condition mistakes in practice.

---

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.
- [S53] Yuta Koreeda, Christopher D. Manning. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Findings of EMNLP 2021. arXiv:2110.01799.
- [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.

- [S61] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. NAACL 2018. arXiv:1803.05355.
- [S62] Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018. arXiv:1809.09600.
- [S63] Qing Lyu, Shreya Havaldar, Adam Stein, et al. Faithful Chain-of-Thought Reasoning. IJCNLP-AACL 2023. arXiv:2301.13379.
- [S64] Liangming Pan, Alon Albalak, Xinyi Wang, William Yang Wang. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. Findings of EMNLP 2023. arXiv:2305.12295.
- [S65] Louis Kaplow. Rules versus Standards: An Economic Analysis. Duke Law Journal 42(3), 1992, pp. 557 to 629. doi:10.2307/1372840.
- [S66] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence 77(2), 1995, pp. 321 to 357. doi:10.1016/0004-3702(94)00041-X.
- [S67] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, 2023.
- [S68] David Madras, Toniann Pitassi, Richard Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. NeurIPS 2018. arXiv:1711.06664.
- [S69] Hussein Mozannar, David Sontag. Consistent Estimators for Learning to Defer to an Expert. ICML 2020. arXiv:2006.01862.
- [S70] Yonatan Geifman, Ran El-Yaniv. Selective Classification for Deep Neural Networks. NeurIPS 2017. arXiv:1705.08500.
- [S71] Gagan Bansal, Tongshuang Wu, Joyce Zhou, et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021. arXiv:2006.14779.
- [S72] Zana Buçinca, Maja Barbara Malaya, Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 2021. arXiv:2102.09692.
- [S73] Geoffrey Irving, Paul Christiano, Dario Amodei. AI safety via debate. arXiv:1805.00899, 2018.
- [S74] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024. arXiv:2305.14325.
- [S75] Nelson F. Liu, Kevin Lin, John Hewitt, et al. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the ACL 12, 2024. arXiv:2307.03172.
- [S76] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. IEEE Transactions on Information Theory 16(1), 1970, pp. 41 to 46. doi:10.1109/TIT.1970.1054406.
- [S77] Raja Parasuraman, Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors 52(3), 2010, pp. 381 to 410. doi:10.1177/0018720810376055.