

Enceavra Research Round 42: Role or Witness? Measured Joint Misreading of Undetermined Facts by One Witness in Two Roles and by Two Witnesses in One Role

Date: 2026-10-31

Research lab: Enceavra Research

Author: Ben Wilkins, Enceavra (ORCID 0009-0005-6532-3200)

Short code: RDX-42 (Role Decorrelation)

Primary tracks: Collaborative track ("When does dividing reasoning across specialists help, and when does it fragment evidence?"), Track 5 (Calibration, Abstention, and Release Control)

Position in the program: second round of the Round 41 to 50 "Theory v1" block. The roadmap item for Round 42 is measured decorrelation: joint misreading rates across witness pairs and prompts (proposition P4).

Carry-forward boundary: Round 41 (HPS-41) found that the only authored error was unanimous (F07, a missing delivery fact) and that the one composition that held it used the same witness in an extraction role. Its handoff asked whether one witness in two roles has lower joint error than two witnesses in one role. Round 42 answers that question on a pool built by construction. It does not rerun the Round 41 selection study or the Round 40 topology comparison.

Abstract

Proposition P4 says that decorrelation, not headcount, is the safety budget of a multi witness gate, and Round 41 added that role may be a lever for it. Round 42 measures that lever directly. Six authored contract families are each written as two contrast variants that differ by exactly one record. In the U variant the packet holds a record that suggests the decisive fact (a draft, a board authorization, a broker email, a wire instruction, a recital pointing to a missing annex) but nothing that establishes it, and the label is HOLD. In the K variant the establishing record is added and the label is PASS or FAIL. Witnesses B, C, and D read every item twice: once giving a direct verdict and once filling an evidence ledger that deterministic code resolves and that fails closed. The same six readers read 20 new public SEC EDGAR exhibits, balanced 10 CERTIFY and 10 HOLD. Pairings were fixed before any call.

We report three strictly separated evidence categories.

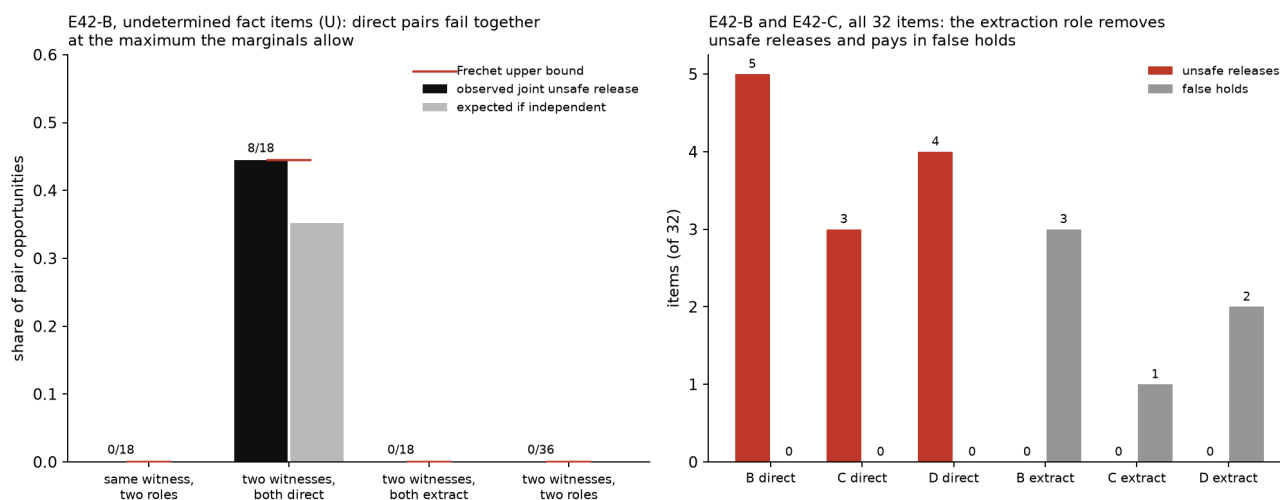
Deterministic (E42-A, no model calls). Every contrast pair differs by exactly one record with the audit question unchanged; a gold evidence ledger reproduces all 12 labels through the same resolver the witnesses face; all 20 exhibit windows contain their anchors and no exhibit or company repeats an earlier Enceavra track. The Frechet Hoeffding bounds show that the joint unsafe release of any pair lies between $\max(0, p + q - 1)$ and $\min(p, q)$ of its two marginal rates, so marginal accuracy cannot certify a pair, and a pair that contains a reader with zero unsafe releases is certified at zero by its marginal alone. Exact binomial limits show that zero events in 18 pair opportunities still allow a true rate of 0.153.

Live authored (E42-B, 6 families, 12 contrast items, 72 calls). On the six U items the direct readers released an unsafe verdict 5, 2, and 4 times (Witnesses B, C, D), and every one of those releases was

FAIL: the reader treated the missing record as proof that the event did not happen. The three extraction readers held all six U items. Two different witnesses in the direct role released the same wrong verdict together on 8 of 18 pair opportunities, which is the largest value their marginal rates allow (Frechet upper bound 8; independence would predict 6.33). The same witness in two roles did so on 0 of 18. On the K items every direct reader was correct, and the extraction role paid 4 reader false holds (Witness B twice, Witness D twice, Witness C none).

Live real exhibits (E42-C, 20 SEC EDGAR exhibits, 120 calls). On a balanced real track the panel was near ceiling. One direct unsafe release occurred (Z16, an amendment effective "subject to the delivery" of executed signatures, certified by Witness C directly and held by its extraction). One exhibit drew two extraction false holds (Z19, where Witnesses B and C extracted a counterpart sufficiency clause as a condition). No pair group in any role released a joint unsafe verdict.

Interpretation, kept separate. On this pool, one witness in two roles never failed jointly, while two witnesses in the direct role failed together as much as arithmetic permits. The safety gain, however, comes from the extraction role's marginal, not from a measured decorrelation: the fail closed resolver never released a wrong verdict, so every pair containing it is bounded at zero. What role does change is the kind of error. A witness errs unsafely as a direct reader and conservatively as an extractor, so its errors across roles are negatively correlated (error phi -0.14, -0.06, -0.10), while the errors of different witnesses within a role are positively correlated (direct 0.45 to 0.88). We revise P4 accordingly.



1. Archive continuity and anti-repetition verification

Before designing any Round 42 experiment we rebuilt the archive ledger from the live research registry (prior-rounds-ledger.json, 35 indexed rounds). A keyword scan of registry titles, questions, and open questions for the Round 42 constructs (role decorrelation, same witness, two roles, role pair, extraction role, contrast, joint error, joint misreading, double fault, label contrast, undetermined fact, Frechet, decorrelation) returned hits only for "extraction role", "joint error", and "undetermined fact" in the Round 41 handoff, which this round answers, and one hit for "decorrelation" in the Round 37 open question,

which asked whether deliberately decorrelated witnesses produce informative split votes. A second scan over registry summaries, findings, and hypotheses also found Round 35's pairwise error correlation, which entered a quorum model as an assumed parameter rather than a measured pair rate, and Round 26's contrastive sufficiency probe, a probe for missing decision conditions rather than a contrast pair design. We then compared the closest rounds by hand. Round 38 described its witness families as "role decorrelated", but there the role was an evidentiary lens (precautionary or presumptive) assigned to different witnesses, and the measure was split votes; Round 39 balanced those lenses two and two. Neither put the same witness in a direct role and an extraction role, and neither measured joint error per pair against bounds.

Prior round	Focus	How Round 42 differs
HPS-41 (Round 41)	Whether disagreement mining builds a sound hard pool	Round 42 builds by construction and measures joint error per fixed reader pair; no selection step is used
NCT-40 (Round 40)	Six composition topologies compared on outcome counts	Round 42 measures the dependence between readers that those topologies rely on; topologies are reported only for continuity
CEV-39 (Round 39)	Evidentiary lens balancing on conflicting records	No records conflict here; the decisive record is either absent or present
CIV-38 (Round 38)	Lens instructions assigned to different witnesses to produce split votes	Role here is a task (direct verdict or evidence ledger) given to the same witness; the measure is joint error per pair
Round 35	Correlated semantic failures in grounded quorums	Round 35 assumed a pairwise error correlation in a quorum model and counted unanimous wrong verdicts; Round 42 measures joint error per fixed pair against Frechet bounds and independence, and adds role as a factor

No packet, exhibit, or prompt task is reused as a test of the same question. The direct, exhibit, and extraction prompts are the Round 41 prompts, unchanged, so that results are comparable. Eleven exhibits come from a new EDGAR full text search; nine come from the Round 39 candidate pool and had never been put to a witness.

2. Where this round sits in the program

The roadmap names Rounds 41 to 50 "Theory v1". Round 40 proposed P1 to P5; Round 41 added P6 (disagreement is a biased detector) and revised P4 to read "decorrelation is the safety budget, and role is a lever", with a removal test: the joint error of the same witness in two roles against two witnesses in one role. Round 42 runs that removal test.

3. Design

3.1 Readers, pairs, and scoring

A reader is a witness in a role. **BD**, **CD**, **DD** are Witnesses B, C, and D giving a direct verdict. **BX**, **CX**, **DX** are the same witnesses filling an evidence ledger (conditions, verbatim quotes, and a status of MET,

UNMET, or UNDETERMINED for each), which deterministic code resolves: any unverifiable quote or any UNDETERMINED condition gives HOLD. For exhibits the extraction lists conditions to effectiveness and an "effective on its date" quote, as in Rounds 40 and 41.

A pair releases only when both readers give the same determinate verdict; otherwise it holds. A **joint unsafe release** is both readers releasing the same wrong verdict. A **joint error** is both readers being wrong in any way, including a false hold. Four pair groups were fixed before any call:

- role pairs: BD+BX, CD+CX, DD+DX (one witness, two roles);
- cross witness direct: BD+CD, BD+DD, CD+DD;
- cross witness extract: BX+CX, BX+DX, CX+DX;
- cross witness cross role: the six pairs of different witnesses in different roles.

Each reader decision is a correct release, an unsafe release (a determinate verdict that differs from truth, including any release when truth is HOLD), a correct hold, or a false hold.

3.2 Contrast variants instead of twins

The Round 41 handoff asked for held out twins. Twins guard against selection on noise, and this round selects nothing: every item and every pairing is fixed in advance and every item is reported. We therefore spent the item budget on contrast variants in the sense of Gardner et al. [S40]: each family is one rule program rendered once without the establishing record (U, label HOLD) and once with it (K, label PASS or FAIL). The build asserts that K equals U plus exactly one record and that the audit question is identical. This gives each family a built in check that a reader who holds on U is not simply holding on everything.

3.3 Families

Family	Decisive fact	What the U packet holds instead	K label
R01	Lender received a signed compliance certificate	An internal memo with draft ratios and an unsigned draft certificate	PASS
R02	A non renewal notice was received in time	Board minutes authorizing management to send it	PASS
R03	Lender received an insurance certificate before funding	A broker email saying the certificate "will follow"	PASS
R04	The agency acceptance notice is dated on or before the deadline	A draft press release that "expects" acceptance	FAIL (notice dated after the deadline)
R05	The counterparty signed a prior written consent	An assignment reciting that consent was obtained, "see Annex 1", with the annex not attached	PASS
R06	The cure payment was received within 10 business days of notice	A wire initiation confirmation from the payer's bank	FAIL (received after the cure period)

The HOLD label on U items rests on a stated convention, unchanged from Rounds 37 to 41: an audit packet is the evidence put before the gate, not a closed register of everything that happened, so a fact that the packet neither establishes nor excludes is undetermined. Under the opposite convention, a

closed record, FAIL would be the correct answer on every U item. Section 9 treats this as the central threat to validity.

3.4 Witness budget

Witnesses B, C, and D each read every item in both roles. Witness A was reserved for the synthesis review at maximum reasoning; the run budget did not allow it in both roles across both tracks. One Witness A direct call on R01U was made while sizing the run; it is kept in the transcripts and excluded from every metric. Every call used temperature 0 and was made once.

4. Related work

Software reliability met this question before language models did. Knight and Leveson found that independently written program versions fail together more often than independence predicts [S43]. Eckhardt and Lee showed that coincident failure follows from variation in input difficulty even when development is independent [S44], and Littlewood and Miller showed that forcing diversity of method, rather than of team, can reduce coincident failure and in principle beat independence [S45]. Round 42 is a direct analogue: the same witness under a different method (role) against different witnesses under the same method.

For language models, Kim et al. measured error correlation across many language models and found substantial agreement on errors, with larger and more accurate models showing highly correlated errors [S41]. Goel et al. found that model similarity is rising with capability and that model judges favor similar models, which undermines oversight by panels of similar models [S42]. Juries of diverse models [S26] and self consistency [S09] assume some independence that these results put in doubt. Classifier ensemble work offers pairwise measures of diversity, including the correlation of errors and the double fault rate, and warns that diversity measures relate loosely to ensemble accuracy [S46]. Chain of Verification separates answering from verifying within one model [S48], which is the closest prior design to a role split; it measures factual precision, not joint release. For zero event counts we use the exact binomial limit and the rule of three [S47]. Contrast sets come from Gardner et al. [S40]; abstention follows the survey of Wen et al. [S24].

5. Hypotheses (stated before the live run)

- **H42.1.** On U items, role pairs have lower joint unsafe release than cross witness pairs in the same role.
- **H42.2.** The extraction role holds more often than the direct role on U items (the F07 pattern from Round 41 replicates).
- **H42.3.** The extraction role pays a measurable false hold cost on K items.
- **H42.4.** The real exhibit track shows the same ordering of pair groups.
- **H42.5 (structural).** Pools of this size cannot distinguish small differences in joint unsafe rate.

6. Evidence category 1: deterministic diagnostic (E42-A)

6.1 Certificates

Certificate	Result
S1 contrast: K equals U plus one record, same question, U is HOLD, K determinate	6 of 6 families pass
S2 gold resolver: a gold ledger with verbatim quotes reproduces each label through the witness resolver	12 of 12 items pass
S3 exhibits: anchors present, track balanced, no exhibit or company from earlier Encevera tracks	20 of 20 pass; 10 CERTIFY, 10 HOLD
S4 baselines	always hold: 0 unsafe, 6 authored and 10 exhibit false holds; always certify on exhibits: 10 unsafe; always PASS on authored items: 8 unsafe

6.2 Bounds

For two readers with marginal unsafe rates p and q on the same items, the joint unsafe rate is bounded by $\max(0, p + q - 1)$ and $\min(p, q)$. Independence gives $p \cdot q$. Two consequences follow without any data. First, two readers with good marginals can still fail together at their full marginal rate, so a pair is not certified by its members' accuracy. Second, if one member has a zero marginal, the pair is bounded at zero whatever the dependence, so a zero joint rate for such a pair says nothing about decorrelation.

p	q	lower	independent	upper
0.1	0.1	0	0.01	0.1
0.2	0.3	0	0.06	0.2
0.3	0.5	0	0.15	0.3
0.5	0.5	0	0.25	0.5

6.3 Resolution

With zero events in n opportunities, the exact one sided 95 percent upper limit is 0.221 at $n = 12$, 0.139 at 20, 0.080 at 36, and 0.049 at 60; the rule of three gives nearly the same values. A one sided exact test needs 4 events against 0 at $n = 12$ and 5 events at every size from 20 to 300. Pair opportunities that share an item are not independent, so these limits are optimistic for pair counts.

7. Evidence category 2: live authored contrast families (E42-B)

7.1 Protocol

Each of the 12 items went to Witnesses B, C, and D in both roles (72 calls, 0 missing). Prompts are the Round 41 prompts, unchanged. Labels were computed by the family rule functions before any call.

7.2 Reader results

Reader	U: unsafe / held (of 6)	K: correct / false hold (of 6)	Correct on both variants (of 6 families)
BD	5 / 1	6 / 0	1
CD	2 / 4	6 / 0	4
DD	4 / 2	6 / 0	2
BX	0 / 6	4 / 2	4
CX	0 / 6	6 / 0	6
DX	0 / 6	4 / 2	4

All eleven unsafe direct releases on U items were FAIL. R05U (consent recited in a missing annex) was held by all six readers. R03U (broker email) and R04U (draft press release) were released as FAIL by all three direct readers.

7.3 Pair results on U items

Pair group	Opportunities	Joint unsafe	Expected if independent	Frechet upper
Same witness, two roles	18	0	0	0
Two witnesses, both direct	18	8	6.33	8
Two witnesses, both extract	18	0	0	0
Two witnesses, two roles	36	0	0	0

Per direct pair, joint unsafe equalled the smaller marginal in every case: BD+CD 2 (upper 2), BD+DD 4 (upper 4), CD+DD 2 (upper 2). Whenever the more careful of two direct readers released a wrong FAIL, the other did too.

On K items, the pair rule converted extraction holds into pair false holds: role pairs 4 of 18, cross witness extract pairs 7 of 18 (one joint error, R02K, held by both BX and DX), cross witness direct pairs 0 of 18.

7.4 Topologies, for continuity with Round 40

Topology over the six readers	Correct release	Unsafe	Correct hold	False hold
Direct majority (B, C, D)	6	4	2	0
Direct unanimous	6	2	4	0
Extraction unanimous	3	0	6	3
Five of six readers agree	5	0	6	1

7.5 Failure walk-through: R03U

The construction loan agreement says the lender "is obligated to fund an Advance only if, on or before the funding date, Lender has received a certificate of builder's risk insurance that names Lender as loss payee." The packet holds one other record, a broker email: the policy "has been bound with Lender as

loss payee. The certificate will follow under separate cover." The question: the lender funded Advance No. 3 on August 5, 2026; was the insurance condition satisfied?

All three direct readers answered FAIL. Witness C's direct rationale reads: "The agreement strictly requires actual receipt of an insurance certificate naming Lender as loss payee prior to funding. The record shows only an email stating that coverage was bound and that the certificate would follow later." That reasoning is correct about what the packet shows and wrong about what it proves: the packet does not say the certificate was not received before August 5. The same three witnesses, asked only to fill the ledger, each marked the receipt condition UNDETERMINED, and the resolver held. In the K variant, a certificate stamped received on August 3 is added; all six readers released PASS.

7.6 Extraction false holds on K

On R02K, BX and DX both marked the delivery method condition ("by hand or by national courier") UNDETERMINED: the record is a courier proof of delivery, and the packet does not say the courier was national. That hold is defensible and points to a gap in the authored K packet rather than a reading error; we keep the authored label and disclose it. On R01K, DX marked "reduced margin covers all of June 2026" UNDETERMINED although receipt on May 14 settles it. On R06K, BX joined two records into one evidence quote, the span check rejected it, and the resolver failed closed, the same formatting cause as the Round 41 F03E hold; the Round 41 method change allowing separate quotes was not yet implemented. None of these holds is unsafe; each is a lost release.

8. Evidence category 3: real SEC EDGAR exhibits (E42-c)

8.1 Track

Twenty exhibits filed with the SEC between January 2025 and October 2026, fixed with labels before any call and run in an order that alternates CERTIFY and HOLD. The label rule is unchanged from Rounds 39 to 41: CERTIFY if the exhibit text alone establishes that the instrument's principal operative effect is in effect; HOLD if that effect depends on any condition the exhibit does not itself establish. Receipt or delivery of counterparts, fees, officer certificates, and the closing of another transaction are such conditions; ongoing covenants are not. Windows hold the opening 1,800 characters of the exhibit and 4,600 characters around the effectiveness passage.

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Link
Z01	Certara, Inc.	EX-10.1	2025-04-14	0001827090-25-000021	effective on its face	CERTIFY	exhibit
Z02	MOLINA HEALTHCARE, INC.	EX-10.1	2026-02-06	0001437749-26-003384	as of trap	HOLD	exhibit
Z03	HERTZ GLOBAL HOLDINGS, INC	EX-10.1	2025-03-24	0001657853-25-000031	effective on its face, distractor	CERTIFY	exhibit
Z04	Nine Energy Service, Inc.	EX-10.3	2026-03-10	0001213900-26-025721	internal conflict	HOLD	exhibit
Z05	LIPELLA PHARMACEUTICALS INC.	EX-10.1	2025-02-24	0001753926-25-000276	effective on its face	CERTIFY	exhibit

ID	Filer	Exhibit	Filed	Accession	Stratum	Label	Link
Z06	TXNM ENERGY INC	EX-10.1	2025-05-27	0001108426-25-000056	proviso trap	HOLD	exhibit
Z07	RF Acquisition Corp II	EX-10.1	2025-10-02	0001829126-25-007836	effective on its face, distractor	CERTIFY	exhibit
Z08	Lazydays Holdings, Inc.	EX-10.1	2025-08-04	0001641172-25-021971	receipt or concurrency	HOLD	exhibit
Z09	SONIDA SENIOR LIVING, INC.	EX-10.2	2026-03-11	0001193125-26-102271	effective on its face	CERTIFY	exhibit
Z10	DigitalOcean Holdings, Inc.	EX-10.1	2026-05-05	0001582961-26-000045	as of trap	HOLD	exhibit
Z11	AMERICOLD REALTY TRUST	EX-10.1	2026-08-31	0001193125-26-377096	effective on its face	CERTIFY	exhibit
Z12	Compass Diversified Holdings	EX-10.2	2025-12-19	0001345126-25-000089	receipt or concurrency	HOLD	exhibit
Z13	Lodging Fund REIT III, Inc.	EX-10.1	2025-01-03	0001558370-25-000020	effective on its face	CERTIFY	exhibit
Z14	CORE MOLDING TECHNOLOGIES INC	EX-10.1	2026-02-17	0001026655-26-000003	as of trap	HOLD	exhibit
Z15	AST SpaceMobile, Inc.	EX-10.1	2025-02-07	0001493152-25-005306	effective on its face	CERTIFY	exhibit
Z16	LAGO Evergreen Credit	EX-10.1	2025-12-17	0001193125-25-322306	receipt or concurrency	HOLD	exhibit
Z17	TELEDYNE TECHNOLOGIES INC	EX-10.1	2025-12-18	0001094285-25-000145	effective on its face, distractor	CERTIFY	exhibit
Z18	NABORS INDUSTRIES LTD	EX-10.2	2025-09-02	0001104659-25-086363	receipt or concurrency	HOLD	exhibit
Z19	Ramaco Resources, Inc.	EX-10.1	2025-07-31	0001104659-25-072699	effective on its face	CERTIFY	exhibit
Z20	T1 Energy Inc.	EX-10.1	2025-03-14	0001213900-25-024191	conditions article	HOLD	exhibit

8.2 Results

Reader	Correct release	Unsafe	Correct hold	False hold
BD	10	0	10	0
CD	10	1	9	0
DD	10	0	10	0
BX	9	0	10	1
CX	9	0	10	1
DX	10	0	10	0

Joint unsafe release was 0 in every pair group (60 opportunities each for the three pair groups of three, 120 for cross role pairs). The one joint error was BX+CX on Z19, a false hold.

8.3 Failure walk-throughs: Z16 and Z19

Z16 (LAGO Evergreen Credit, accession 0001193125-25-322306) reads: "This Amendment shall become effective as of the date hereof, subject to the delivery of this Amendment, executed and delivered by duly authorized officers of the parties hereto". Witness C, reading directly, certified on the ground that no conditions "outside the executed instrument" remain. Under the label rule, delivery of executed signatures is a condition a filed exhibit does not establish. Witness C's own extraction listed that clause as a condition and the resolver held. This is the execution only stratum that split the panel in Round 41.

Z19 (Ramaco Resources, accession 0001104659-25-072699) says "This Agreement shall be effective as of the date first written above" and, separately, that the exchange of a fully executed agreement by electronic transmission "shall be sufficient to bind the parties". Witnesses B and C, as extractors, listed the sufficiency clause as a condition; the resolver holds on any verified condition, so both held. All three direct readers certified correctly.

9. Findings and interpretations

9.1 Measured findings

- **F42.1 (deterministic, E42-A).** All contrast, gold resolver, and exhibit certificates pass. Joint unsafe release of a pair is bounded by the smaller marginal; with zero events in 18 opportunities the exact 95 percent upper limit is 0.153.
- **F42.2 (live, E42-B).** On U items, direct readers released unsafe verdicts 11 times in 18 reader decisions, all FAIL; extraction readers released none.
- **F42.3 (live, E42-B).** Two witnesses in the direct role released the same wrong verdict on 8 of 18 U pair opportunities, equal to the Frechet upper bound (independence 6.33). One witness in two roles did so on 0 of 18.
- **F42.4 (live, E42-B).** On K items the direct role was correct 18 of 18; the extraction role gave 4 false holds in 18, and the pair rule turned them into 4 of 18 role pair false holds and 7 of 18 cross witness extraction pair false holds.
- **F42.5 (live, E42-C).** On a balanced real track, one unsafe direct release (Z16) and two extraction false holds (Z19) occurred; no pair group released a joint unsafe verdict.
- **F42.6 (live, pooled over 32 items).** Joint error, counting false holds, was 0 of 96 for role pairs, 8 of 96 for cross witness direct pairs, and 2 of 96 for cross witness extraction pairs. Error correlation within role pairs was negative for every witness (-0.14, -0.06, -0.10) and positive for every cross witness direct pair (0.45, 0.88, 0.53).

9.2 Interpretation

The answer to the Round 41 handoff is yes on this pool, with a qualification that matters for design. The role pairs' zero is produced by the extraction marginal, which was zero everywhere; a pair of two different witnesses in different roles has the same zero. On this evidence, role works less like a source of independent opinions and more like a change in failure mode: the direct role fails by releasing (here, by reading absence as negation), and the extraction role with a fail closed resolver fails by holding. Because the two failure modes point in opposite directions, the same witness's errors across roles are negatively correlated, which is the property P4 asks for, obtained by design rather than by hoping two models think differently.

Cross witness pairs within the direct role were at the maximum dependence their marginals allowed. Adding a second direct witness bought nothing on U items: every wrong FAIL from the more careful witness was matched by the other. This is consistent with a shared reading convention rather than idiosyncratic noise, as in Round 41's F07, but in the opposite direction (FAIL there was PASS).

9.3 Hypothesis status

Hypothesis	Status
H42.1 role pairs below cross witness same role pairs	Supported against direct pairs (0 of 18 against 8 of 18); not distinguishable from extraction pairs (0 against 0); the mechanism is the extraction marginal, not measured independence
H42.2 extraction holds more on U	Supported (18 of 18 held against 7 of 18)
H42.3 extraction false hold cost on K	Supported (4 of 18 reader false holds; 0 for Witness C)
H42.4 exhibits show the same ordering	Not testable: no joint unsafe release in any group
H42.5 structural resolution	Supported: a zero in 18 opportunities allows a true rate up to 0.153

10. Toward Theory v1

- **P4 revised again. Decorrelate by role, then verify by structure.** The safety of a pair depends on whether its members fail in the same direction. Different witnesses in the same role failed together at the Frechet maximum; the same witness in a direct role and an extraction role made opposite kinds of error. Evidence: F42.3, F42.6. Removal test: an extraction role with a nonzero unsafe marginal whose unsafe releases coincide with the direct role's on the same items.
- **P7 (new, proposed). Absence is not negation.** Direct readers resolve a missing establishing record as proof that the event did not happen; a ledger with an explicit UNDETERMINED status does not. Evidence: F42.2, R03U. Removal test: a direct prompt that states whether the packet is a closed record, with the FAIL releases on U items unchanged.

11. Method changes for later rounds

1. Report joint error per fixed pair with its Frechet bounds and independence expectation, not only topology outcomes.
 2. Treat a zero joint rate for a pair that contains a zero marginal reader as a marginal result, not as evidence of decorrelation.
 3. Every evaluation pool states its record closure convention to witnesses or tests both conventions.
 4. Count pair opportunities and items separately; limits on pair counts are optimistic when pairs share items.
-

12. Synthesis review (interpretation, not evidence)

After all results were recorded, Witness A at maximum reasoning received the structural and live results and was asked to separate what the data show from what they suggest. Its review is archived verbatim (`synthesis-review.json`). It agreed that the role pair zero "is produced entirely by the extraction marginal being zero", that direct readers "appear to share a single failure mechanism, treating absence of a record as proof of nonoccurrence", and that the exhibit set was too easy to discriminate configurations. It flagged as a threat that "under the opposite convention the direct FAIL verdicts would be correct and the headline difference would largely vanish", and proposed a direct role variant told explicitly that the packet is not a closed record. That proposal is adopted in Section 14. The review is a second reading, not a source of any number in this paper.

13. Limitations and threats to validity

- The authored pool is small: 6 families, 12 items, 18 pair opportunities per group on U items. Opportunities share items and are not independent samples.
 - The HOLD label on U items is a convention. A reader that treats the packet as a closed record answers FAIL, and under that convention the direct readers' eleven unsafe releases would be correct. The direct prompt says to return HOLD "only if no determinate PASS or FAIL verdict is supportable from the packet", which does not settle the convention.
 - The extraction marginal was zero, so the design could not measure dependence inside the extraction role or between roles on unsafe releases.
 - Witness A did not read items in this round, so the four witness quorum of earlier rounds is not reproduced.
 - Families and labels were written by the research team; no external reviewer checked labels before the run. One K label (R02K) depends on reading a courier proof of delivery as delivery by national courier, which the packet does not state. Every call was made once at temperature 0 .
 - The real exhibit track was near ceiling, so it rules out gross regression but does not test the pair ordering.
-

14. Next check

Run the direct role twice on new undetermined fact contrast families, once with the packet declared an open record and once declared closed, alongside extraction readers designed to have a nonzero unsafe marginal (decisive records present but easy to misread). Measure whether the direct FAIL releases follow the convention or persist, and whether the same witness's unsafe releases coincide across roles once both roles can err. This also serves the roadmap's Round 43 item, flag only validators, by giving them items whose correct flags are known.

15. References

- [S09] Xuezhi Wang et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171, 2022.
- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S26] Pat Verga, Sebastian Hofstatter, et al. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796, 2024.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. <https://efits.sec.gov/LATEST/search-index> and <https://www.sec.gov/Archives/edgar/data/>.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.
- [S41] Elliot Kim, Avi Garg, Kenny Peng, Nikhil Garg. Correlated Errors in Large Language Models. Proceedings of the 42nd International Conference on Machine Learning (ICML), PMLR 267, 2025. arXiv:2506.07962.
- [S42] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, et al. Great Models Think Alike and this Undermines AI Oversight. arXiv:2502.04313, 2025.
- [S43] John C. Knight, Nancy G. Leveson. An Experimental Evaluation of the Assumption of Independence in Multiversion Programming. IEEE Transactions on Software Engineering 12(1), 1986. doi:10.1109/TSE.1986.6312924.
- [S44] Dave E. Eckhardt, Larry D. Lee. A Theoretical Basis for the Analysis of Multiversion Software Subject to Coincident Errors. IEEE Transactions on Software Engineering 11(12), 1985. doi:10.1109/TSE.1985.231895.
- [S45] Bev Littlewood, Douglas R. Miller. Conceptual Modeling of Coincident Failures in Multiversion Software. IEEE Transactions on Software Engineering 15(12), 1989. doi:10.1109/32.58771.
- [S46] Ludmila I. Kuncheva, Christopher J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51, 2003. doi:10.1023/A:1022859003006.

- [S47] James A. Hanley, Abby Lippman-Hand. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA 249(13), 1983. doi:10.1001/jama.1983.03330370053031.
- [S48] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, Jason Weston. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495, 2023.