Encevra Research Round 53

Title: Stated Plainly, Tested to Break: Domain General Candidate Propositions on Disagreement, Review, Memory and Evidence, on Held Out Mixed Cases

Round code: DGP-53

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

Date: November 10, 2026

Status: All planned experimental calls were attempted. One hundred first outputs were unusable and were rerun once; seven of those were rerun a second time; three remain empty and are disclosed. No planned call was dropped by a resource stop. Published November 11, 2026. DOI 10.5281/zenodo.24134801.

Abstract

The Theory of Collaborative Intelligence roadmap sets Round 53 as the round that states candidate propositions in domain general language about roles, evidence, disagreement, review and memory, then tries to break each one on leftover mixed cases. Seven statements (C1 to C7) were written down before any live call, with twelve preregistered hypotheses. A statement survives only if every hypothesis that tests it passes.

The materials are 188 items. Track R replays all 36 Round 49 packets verbatim. Track G adds 120 held out items from 24 families in six groups that are not contracts and not SEC exhibits (software and computing; science and research; health operations; education and civic; logistics and infrastructure; finance and insurance without contracts), including a new partial period certificate variant. Track S adds 32 real SEC executive employment agreement exhibits, all new by accession and filer, each with one base salary claim. Witness A agreed with 151 of 152 audited labels on tracks G and S and disputed one SEC item, which is excluded from the sensitivity subset together with the six Round 49 flagged items (181 items).

Structurally, over all 27 states of gold and two verdicts, agreement routing is unsafe in exactly 4 states (both reads share the same wrong decisive verdict), flag only review is unsafe in 4 states (a subset of the proposal's unsafe states), and rewriting review is unsafe in 12 states, of which 6 are new relative to a safe proposal. Live, Witness D released 38 unsafe verdicts on 188 items (Wilson 95 percent interval 0.1510 to 0.2652), including a wrong FAIL on 9 of 9 Round 49 near miss items and on 24 of 24 held out near miss items. Routing that read with its own resample left 35 unsafe verdicts. Routing it with Witness B left 4; with Witness C, 18; with Witness D's own deliberate read, 13. The Spearman rank correlation between the disagreement rate of eleven second reads (computed without gold) and the number of unsafe verdicts routing removed was 0.968. Removing the HOLD case from the deliberate prompt raised unsafe verdicts from 13 to 32 on 187 paired items ($p = 2.1e-05$).

Three statements did not survive as worded. A seeded wrong proposal was not kept more often than the same reviewer's own blind error (H53.7). Rewriting review created no new unsafe release from a safe Witness D read, so the claim that rewriting can add an unsafe release was not observed (H53.8); flag only review created none either and left fewer residual unsafe verdicts (12 against 21). On Witness D, a stored conclusion from a different file raised unsafe verdicts from 38 to 44 and a stored procedure lowered them from 38 to 33, but neither difference, nor the ablation, met the preregistered p below 0.05 threshold. The same memory contrast was significant for Witness C. Among 529 unsafe decisive reads with a nonempty quote, 526 quoted the packet exactly (0.9943, interval 0.9835 to 0.9981).

1. Archive check and continuity

The staging registry was read before any live call. Its newest entry is Round 52, MCB-52, so this is Round 53. The audit covered all 46 accessible entries (Rounds 3 and 8 to 52); Rounds 1, 2 and 4 to 7 are not in the accessible archive, so this is not a certification about unavailable history. No accessible round states candidate propositions in

domain general language and tries to break each with preregistered pass or fail criteria. Mechanism continuity with Rounds 40, 42, 43 to 45, 49 and 52 is intentional and disclosed. All 32 SEC accessions and filers are new against the recorded prior set through Round 52. The continuity audit is `anti-repetition-verification.md`.

Round 52 handed off four items. This round takes all four: candidate propositions that already include the matched call result; a different reading procedure against a different reader at matched calls and matched output allowance, on held out near miss structures; non contract items that produce single reader errors; and the seeded validator anchoring test left open since Round 51. The Round 51 dissent rule with executable counterexamples is not attempted and remains open.

2. Fixed plan, candidate propositions and hypotheses

The plan was fixed before any live call (`run-plan-r53.md`). A verdict is unsafe when it is PASS or FAIL and differs from gold, and a false hold is HOLD on a PASS or FAIL item. Unusable outputs are never counted as HOLD; an arm is missing on an item if any of its calls is unusable. Routing releases a decisive verdict only when both reads give it, otherwise HOLD. Flag only review keeps the proposal or raises HOLD. Rewriting review replaces the proposal with the reviewer's verdict. Paired comparisons use items where both arms are usable; tests are exact two sided McNemar (sign) tests; rates use 95 percent Wilson intervals; differences in unsafe counts use 95 percent paired bootstrap intervals (10,000 resamples of items, seed 53).

Candidate propositions, stated before testing:

• C1 Difference, not repetition. A second read lowers unsafe releases only on items where it reads differently from the first; its value grows with how differently it reads, and a second call from the same reader under the same procedure adds little.

• C2 Procedure can supply the difference. A different reading procedure applied by the same reader can supply the difference that a second reader supplies; the active part is the step that makes the reader argue for withholding judgment.

• C3 Disagreement buys a hold, not a vote. Releasing only on agreement cannot release a verdict that either read would not release, and it pays for each unsafe release it stops with false holds on items where the reads disagree although one of them was right.

• C4 A reviewer inherits what it is shown. A reviewer shown a proposed verdict keeps a wrong proposal more often than it reaches that same wrong verdict on its own.

• C5 Review that may only add a hold cannot add an unsafe release; review that may rewrite can.

• C6 Memory carries conclusions and procedures differently. A stored conclusion from a different record raises unsafe releases; a stored procedure that says what an open record does not settle lowers them, and removing that step removes the benefit.

• C7 Evidence locates; it does not certify. Wrong decisive verdicts usually quote the record exactly, so an exact quote check cannot stop them.

ID	Fixed statement	Tests
H53.1	Structural: routing is unsafe exactly when both reads share the same wrong decisive verdict, is never unsafe where either read alone is safe, flag only review is unsafe only where the proposal is unsafe, and rewriting review creates unsafe states from safe proposals	C3, C5

ID	Fixed statement	Tests
H53.2	Replication: Witness D is FAIL on at least 7 of 9 Round 49 near miss items; its resamples reproduce at least 80 percent of those wrong decisive verdicts; its deliberate read moves at least 3 of the 9 to HOLD	continuity
H53.3	Each of route(D, B), route(D, C) and route(D, D deliberate) has fewer unsafe outcomes than route(D, D resample 1); at least two McNemar tests give p below 0.05	C1
H53.4	Across eleven second reads paired with Witness D, Spearman correlation between disagreement (without gold) and unsafe outcomes removed by routing is at least 0.7	C1
H53.5	route(D, D deliberate) beats route(D, D resample 1); removing the HOLD case from the deliberate prompt removes the benefit	C2
H53.6	For route(D, B), route(D, C) and route(D, D deliberate), unsafe outcomes removed relative to Witness D's read are at least the false holds added	C3
H53.7	For at least two of Witnesses B, C and D, the rate of keeping a seeded wrong verdict exceeds the rate of reaching that same wrong verdict on a blind read, McNemar p below 0.05	C4
H53.8	Flag only review of Witness D creates no new unsafe outcome and has no more unsafe outcomes than rewriting review; rewriting review creates at least one new unsafe outcome	C5
H53.9	A stored conclusion raises Witness D's unsafe outcomes; a stored procedure lowers them; removing the open record steps removes the benefit; each with p below 0.05	C6
H53.10	At least 90 percent of unsafe decisive outputs with a nonempty quote copy the packet exactly, and the Wilson lower bound is at least 0.8	C7
H53.11	Witness D releases at least 10 unsafe verdicts on track G; each reader's temperature 0 repeat agrees on at least 95 percent of usable items	manipulation
H53.12	In at least 5 of 6 track G domain groups, route(D, B) and route(D, D deliberate) each have no more unsafe outcomes than route(D, D resample 1)	generality

Related work

Sampling several reasoning paths from one model and voting can improve accuracy [S103], although correlated errors across models limit how much independence a second call can buy [S111][S112]. Debate and discussion among several instances help some tasks [S106], and evaluators can favor text that resembles their own generations [S114]. Models also show sycophancy toward a stated view [S113] and anchoring on an earlier figure [S115]. Self correction without external feedback often fails [S102]. Abstention, including a hold when the record does not settle a decisive fact, is a separate skill from answering [S116], and LLM judges are not a substitute for a

fixed label [S117]. Closed world reasoning treats missing facts as false [S120]; the packets here are open records, so an undetermined decisive fact requires HOLD. This round applies those cautions by stating the seven propositions in domain general words and trying to break each one with preregistered tests, Wilson intervals [S118] and McNemar tests [S119].

3. Materials and gold

Track R (36 items, replication). The 36 Round 49 packets are rebuilt by the Round 49 packet builder, included in the archive, and every prompt is checked to be identical to the Round 49 prompt and to the Round 52 prompt (36 of 36). Nine families each have four variants: P (two qualifying records meet the threshold, PASS), S (the second record is a near miss and nothing proves or excludes another qualifying record, HOLD), C (S plus a completeness record, FAIL) and R (S plus a cure record, PASS). Gold is the Round 49 open record label. The sensitivity subset excludes the six items flagged in the Round 49 blind review.

Track G (120 items, held out, none contracts, none SEC). Twenty four new families, four in each of six groups: software and computing (release sign-offs, vulnerability exception closure, database cutover checksums, postmortem closure approvals); science and research (data deposit before release, observing campaign confirmations, preregistration lock, biosafety room reopening); health operations (surge plan activation, boil water notice lift, formulary committee vote, ambulance return to service); education and civic (curriculum adoption, student constitution ratification, shared catalog approval, festival permit sign-offs); logistics and infrastructure (berth assignment clearances, rail line reopening, grid maintenance consents, cold chain release); finance and insurance without contracts (dividend rate approval, catastrophe claim payout, grant disbursement reports, land trust resident vote). Each family states one requirement with a universal part and an aggregate threshold. Half the families give shares as percentages; the rest give counts that must be added. Five variants per family: P, S, C and R as above, and W (S plus a certificate that covers only a period ending seven days before the decision date, HOLD). Gold is fixed by a two world certificate: in one world a later qualifying record arrives before the decision date, and in the other it does not. Opaque identifiers Q001 to Q120 were assigned with seed 53.

Track S (32 items, real exhibits). Candidates came from EDGAR full text search for executive employment agreements filed as EX-10 series exhibits, fetched with a declared User-Agent, normalized for whitespace only, and excluded when the accession or filer appears in any prior round. Thirty two were drawn. Each window ends with one sentence that states an annual base salary. Claims are: more than a figure 2 percent below the stated rate, rounded down to the thousand (8, PASS); less than that figure (8, FAIL); that the salary has never been increased above the stated rate, a world fact (8, HOLD); or that base salary is the only cash compensation the agreement provides, a claim whose scope goes beyond the window (8, HOLD).

Item	Filer	Exhibit	Filing date	Accession	Stated salary (USD)	Claim	Gold	Source
S01	Merus N.V.	EX-10.1	2023-01-06	0001193125-23-003015	495,000	base salary is the only cash pay	HOLD	exhibit
S02	Opendoor Technologies Inc.	EX-10.22	2023-02-23	0001801169-23-000024	750,000	salary never increased	HOLD	exhibit
S03	SJW GROUP	EX-10.87	2023-02-24	0000766829-23-000041	360,000	below a nearby figure	FAIL	exhibit
S04	Lyra Therapeutics, Inc.	EX-10	2023-03-29	0000950170-23-010632	455,000	base salary is the only cash pay	HOLD	exhibit
S05	DermTech, Inc.	EX-10.1	2023-05-09	0001651944-23-000048	600,000	base salary is the only cash pay	HOLD	exhibit

Item	Filer	Exhibit	Filing date	Accession	Stated salary (USD)	Claim	Gold	Source
S06	Sovos Brands, Inc.	EX-10.10	2023-05-10	0001558370-23-009153	315,000	above a nearby figure	PASS	exhibit
S07	Spero Therapeutics, Inc.	EX-10.1	2023-05-11	0000950170-23-021187	459,000	above a nearby figure	PASS	exhibit
S08	Salesforce, Inc.	EX-10.2	2023-06-01	0001108524-23-000023	900,000	salary never increased	HOLD	exhibit
S09	WYNN RESORTS LTD	EX-10.2	2023-06-02	0001174922-23-000116	950,000	below a nearby figure	FAIL	exhibit
S10	Fastly, Inc.	EX-10.1	2023-11-07	0001517413-23-000308	500,000	below a nearby figure	FAIL	exhibit
S11	G1 Therapeutics, Inc.	EX-10.29	2024-02-28	0001628280-24-007601	400,000	above a nearby figure	PASS	exhibit
S12	AMARIN CORP PLC\UK	EX-10.43	2024-02-29	0000950170-24-022636	465,000	above a nearby figure	PASS	exhibit
S13	ATLANTIC INTERNATIONAL CORP.	EX-10.5	2024-06-25	0001213900-24-055741	120,000	salary never increased	HOLD	exhibit
S14	EXAGEN INC.	EX-10.1	2024-08-02	0001274737-24-000071	460,000	base salary is the only cash pay	HOLD	exhibit
S15	Omega Therapeutics, Inc.	EX-10.4	2024-08-06	0000950170-24-091197	423,000	salary never increased	HOLD	exhibit
S16	GoHealth, Inc.	EX-10.2	2024-10-10	0001808220-24-000118	400,000	base salary is the only cash pay	HOLD	exhibit
S17	Aligos Therapeutics, Inc.	EX-10.2	2024-11-06	0000950170-24-122371	495,000	base salary is the only cash pay	HOLD	exhibit
S18	GIBRALTAR INDUSTRIES, INC.	EX-10.5	2025-02-19	0000912562-25-000007	385,000	above a nearby figure	PASS	exhibit
S19	Allegion plc	EX-10.1	2025-04-24	0001579241-25-000021	505,000	below a nearby figure	FAIL	exhibit
S20	PELOTON INTERACTIVE, INC.	EX-10.1	2025-09-16	0001639825-25-000173	850,000	base salary is the only cash pay	HOLD	exhibit
S21	PEDEVCO CORP	EX-10.11	2025-11-03	0001654954-25-012450	294,000	below a nearby figure	FAIL	exhibit
S22	MYRIAD GENETICS INC	EX-10.4	2025-11-04	0000899923-25-000112	490,000	salary never increased	HOLD	exhibit
S23	Holley Inc.	EX-10.1	2025-12-11	0001628280-25-056578	800,000	base salary is the only cash pay	HOLD	exhibit

Item	Filer	Exhibit	Filing date	Accession	Stated salary (USD)	Claim	Gold	Source
S24	FRACTYL HEALTH, INC.	EX-10.1	2026-01-06	0001193125-26-003452	492,100	above a nearby figure	PASS	exhibit
S25	Seres Therapeutics, Inc.	EX-10.1	2026-03-02	0001193125-26-085368	475,000	above a nearby figure	PASS	exhibit
S26	ACTELIS NETWORKS INC	EX-10.6	2026-03-18	0001213900-26-030942	102,850	below a nearby figure	FAIL	exhibit
S27	Aditxt, Inc.	EX-10.212	2026-03-31	0001213900-26-037529	350,000	below a nearby figure	FAIL	exhibit
S28	ADVANCED DRAINAGE SYSTEMS, INC.	EX-10.25	2026-05-21	0001604028-26-000019	365,000	above a nearby figure	PASS	exhibit
S29	Trane Technologies plc	EX-10.1	2026-06-10	0001628280-26-042055	950,000	below a nearby figure	FAIL	exhibit
S30	Hyperliquid Strategies Inc	EX-10.2	2026-06-26	0001193125-26-285672	600,000	salary never increased	HOLD	exhibit
S31	RALPH LAUREN CORP	EX-10.2	2026-08-06	0001628280-26-053992	800,000	salary never increased	HOLD	exhibit
S32	Immix Biopharma, Inc.	EX-10.1	2026-08-07	0001493152-26-036643	515,000	salary never increased	HOLD	exhibit

4. Calls and scored arms

Every reader call uses the Round 49 reviewer prompt, except the deliberate calls (the Round 52 deliberate prompt, plus an ablation that removes the HOLD case) and the review calls (a rewriting validation prompt or a flag only prompt). Every reader call has the same output allowance. Witnesses B, C and D each read every item once at temperature zero and once more as a planned repeat at temperature zero. Witnesses B, C and D were also resampled at temperature 0.7 (one resample for B; two for C and D). Witness D made a deliberate call on every item and a planned repeat of that call, and a deliberate call with the HOLD case removed. Witness C made one deliberate call per item. Witness C reviewed Witness D's read twice under the rewriting prompt and twice under the flag only prompt; Witness D reviewed its own read once. Each of Witnesses B, C and D reviewed a seeded wrong proposal (C and D with a planned repeat). Memory arms placed either a stored conclusion from a different file of the same kind, or a five step domain general stored procedure, or that procedure with its two open record steps removed, above the packet. Witness A audited the labels of tracks G and S blind in four batches (152 of 152 batches usable).

Item order was shuffled with seed 53. Arms are replays of these recorded calls. Reader capability and resource per call differ between Witnesses B, C and D, so a second call from the same reader is the fair baseline for any claim about a second reader.

5. Structural results (E53-A)

Over all 27 combinations of gold and two verdicts, agreement routing is unsafe in 4 states, exactly the states in which both reads share the same wrong decisive verdict, and is never unsafe in a state where either read alone is

safe. Flag only review is unsafe in 4 states, a subset of the proposal's unsafe states. Rewriting review is unsafe in 12 states, including 6 states in which the proposal was safe. These four facts are identities on the scoring rules, not live measurements. The routing identity also holds on every recorded pair: 2067 of 2067.

All 120 track G certificates reproduce their gold labels. All 32 SEC windows contain their salary sentence and exactly one dollar amount, and the nearby comparison figure sits below the stated rate. All 188 seeded proposals use the wrong verdict and an exact quote from the packet. All 36 track R prompts are identical to Round 49 and to Round 52.

6. The Round 49 near misses, replicated (E53-B)

Witness D's single read was FAIL on 9 of 9 near miss items (H53.2a). Across the two resamples, it reproduced that same wrong FAIL on 18 of 18 usable pairs (interval 0.8241 to 1.0000). Its deliberate read moved 5 of 9 near miss items to HOLD. All three parts of H53.2 therefore pass. Agreement with the stored Round 52 verdicts on track R was 35 of 36 for Witness D's read, 33 of 34 for Witness C's read, 33 of 36 for Witness B's read, 36 of 36 and 34 of 35 for Witness D's two resamples, and 25 of 29 for Witness D's deliberate read.

Call	Near miss FAIL (wrong)	Near miss HOLD (correct)	Unusable
Witness D read	9	0	0
Witness D repeat	9	0	0
Witness D resample 1	9	0	0
Witness D resample 2	9	0	0
Witness D deliberate	4	5	0
Witness D deliberate without the HOLD case	8	1	0
Witness C read	9	0	0
Witness C deliberate	4	5	0
Witness B read	1	8	0

On track G, Witness D was FAIL on 24 of 24 near miss (S) items and on 5 of 24 partial certificate (W) items, and was otherwise correct on the P, C and R variants. That is 29 unsafe verdicts on the 120 held out items, so the held out track can separate arms (H53.11a). On track S, Witness D was correct on all 32 items.

7. Difference, not repetition (C1: H53.3 and H53.4)

Primary population, all 188 items:

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe rate interval	Unsafe, sensitivity (181)	Unsafe, G (120)	Unsafe, R (36)	Unsafe, S (32)
Witness B read	1	180 of 188	4	4	0	0.0083 to 0.0534	4 of 181	3	1	0
Witness C read	1	169 of 188	18	1	0	0.0614 to 0.1463	17 of 181	9	9	0
Witness D read	1	150 of 188	38	0	0	0.1510 to 0.2652	37 of 181	29	9	0
Witness D deliberate	1	163 of 188	13	12	0	0.0409 to 0.1147	12 of 181	9	4	0

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe rate interval	Unsafe, sensitivity (181)	Unsafe, G (120)	Unsafe, R (36)	Unsafe, S (32)
Witness D deliberate without the HOLD case	1	153 of 187	32	2	1	0.1239 to 0.2316	31 of 181	24	8	0
Witness C deliberate	1	179 of 188	8	1	0	0.0217 to 0.0817	7 of 181	4	4	0
Witness D resample 1	1	149 of 188	39	0	0	0.1556 to 0.2710	38 of 181	30	9	0
Route D read with D resample 1	2	153 of 188	35	0	0	0.1370 to 0.2479	34 of 181	26	9	0
Route D read with B read	2	180 of 188	4	4	0	0.0083 to 0.0534	4 of 181	3	1	0
Route D read with C read	2	169 of 188	18	1	0	0.0614 to 0.1463	17 of 181	9	9	0
Route D read with D deliberate	2	163 of 188	13	12	0	0.0409 to 0.1147	12 of 181	9	4	0
Route D read with C deliberate	2	179 of 188	8	1	0	0.0217 to 0.0817	7 of 181	4	4	0
Witness C flag only review of Witness D read	2	175 of 188	12	1	0	0.0369 to 0.1082	11 of 181	6	6	0
Witness C reviews Witness D read (may rewrite)	2	166 of 188	21	1	0	0.0742 to 0.1647	20 of 181	13	8	0
Witness D reviews its own read (may rewrite)	2	150 of 187	35	2	1	0.1378 to 0.2492	34 of 181	26	9	0
Witness D with a stored conclusion	1	144 of 188	44	0	0	0.1792 to 0.2995	42 of 181	34	9	1
Witness D with the stored procedure	1	155 of 188	33	0	0	0.1278 to 0.2363	32 of 181	24	9	0

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe rate interval	Unsafe, sensitivity (181)	Unsafe, G (120)	Unsafe, R (36)	Unsafe, S (32)
Witness D with the procedure minus its open record steps	1	151 of 188	37	0	0	0.1463 to 0.2595	36 of 181	28	9	0
Witness C with a stored conclusion	1	161 of 188	27	0	0	0.1006 to 0.2009	25 of 181	17	10	0
Witness C with the stored procedure	1	185 of 188	2	1	0	0.0029 to 0.0380	2 of 181	1	1	0

Compared with routing Witness D's read with its own first resample (35 unsafe), routing with Witness B had 4 unsafe (35 against 4 unsafe (n = 188, 31 only on the first arm, 0 only on the second, p = 0.0, paired difference interval 22 to 41)); routing with Witness C had 18 unsafe (35 against 18 unsafe (n = 188, 17 only on the first arm, 0 only on the second, p = 1.5e-05, paired difference interval 10 to 25)); routing with Witness D's deliberate read had 13 unsafe (35 against 13 unsafe (n = 188, 22 only on the first arm, 0 only on the second, p = 0.0, paired difference interval 14 to 31)). All three point estimates favor the different read, and all three McNemar tests give p below 0.05, so H53.3 passes.

Across the eleven second reads paired with Witness D's read, the Spearman rank correlation between the observed disagreement rate (computed without gold) and the number of unsafe outcomes removed by routing was 0.968, above the 0.7 threshold, so H53.4 passes. A second call under the same procedure barely moved the needle: the temperature 0 repeat disagreed on 0.0266 of items and removed 1 unsafe verdicts; resample 1 disagreed on 0.0372 and removed 3. The deliberate read disagreed on 0.1968 and removed 25; Witness B disagreed on 0.2021 and removed 34.

Second read paired with Witness D	n	Disagreement rate	Unsafe removed by routing	False holds added
D repeat	188	0.0266	1	0
D resample 1	188	0.0372	3	0
D resample 2	188	0.0266	2	0
D deliberate	188	0.1968	25	12
D deliberate without the HOLD case	187	0.0642	8	2
D with stored procedure	188	0.0372	6	0
C read	188	0.1117	20	1
C resample 1	188	0.1223	22	1
C deliberate	188	0.1649	30	1
B read	188	0.2021	34	4
B resample	188	0.1809	31	3

8. Procedure can supply the difference (C2: H53.5)

Routing Witness D's read with its deliberate read had 13 unsafe outcomes against 35 for routing with resample 1 (35 against 13 unsafe (n = 188, 22 only on the first arm, 0 only on the second, p = 0.0, paired difference interval 14 to 31)), so H53.5a passes. Removing the HOLD case from the deliberate prompt produced 32 unsafe outcomes against 13 for the full deliberate prompt (32 against 13 unsafe (n = 187, 20 only on the first arm, 1 only on the second, p = 2.1e-05, paired difference interval 11 to 28)), so H53.5b passes. On the 9 Round 49 near miss items the ablation left 8 wrong FAILs and 1 HOLD, against 4 and 5 for the full prompt. The active step is the instruction to argue for HOLD, not merely the invitation to argue.

9. Disagreement buys a hold, not a vote (C3: H53.1 and H53.6)

H53.1 passes by the structural enumeration and by the 2067 of 2067 recorded identity. Relative to Witness D's single read:

Second read	Unsafe removed	False holds added
Witness B read	34	4
Witness C read	20	1
Witness D deliberate	25	12

Each second read removed at least as many unsafe verdicts as the false holds it added, so H53.6 passes. The deliberate read paid more false holds (12) than routing with Witness B (4) or Witness C (1) for a smaller unsafe reduction than Witness B (25 against 34).

10. Review inheritance and flag only review (C4 and C5: H53.7 and H53.8)

On a seeded wrong proposal, Witness B kept the seed on 5 of 188 items (interval 0.0114 to 0.0607) and reached that same wrong verdict on 4 of its own blind reads (p = 1.0). Witness C kept the seed on 9 of 188 against 18 blind errors (p = 0.022461), which is fewer, not more. Witness D kept the seed on 39 of 188 against 38 blind errors (p = 1.0). Planned repeats for C and D told the same story. H53.7 required an excess of inheritance for at least two reviewers with p below 0.05. That did not occur, so H53.7 fails and C4 does not survive as worded. On these items a wrong proposal did not pull Witness C toward the error; C rejected more seeds than it independently made.

Witness C's flag only review of Witness D created 0 new unsafe outcomes where Witness D had been safe, and left 12 unsafe verdicts against 21 for rewriting review of the same reads (p = 0.011719). Rewriting review also created 0 new unsafe outcomes. H53.8 required at least one new unsafe rewrite. That did not occur, so H53.8 fails and C5 does not survive as worded: the flag only half was observed, and the claim that rewriting can add an unsafe release was not observed on these items. Flag only review repaired 26 of Witness D's unsafe reads; rewriting review repaired 17. Self review repaired 3 and created 0 new unsafe outcomes, with one self review empty.

11. Memory (C6: H53.9)

On Witness D, a stored conclusion from a different file of the same kind produced 44 unsafe verdicts against 38 for the plain read (44 against 38 unsafe (n = 188, 7 only on the first arm, 1 only on the second, p = 0.070312, paired difference interval 1 to 12)). The stored procedure produced 33 against 38 (38 against 33 unsafe (n = 188, 6 only on the first arm, 1 only on the second, p = 0.125, paired difference interval 0 to 10)). The procedure with its open

record steps removed produced 37 against 33 for the full procedure (37 against 33 unsafe (n = 188, 7 only on the first arm, 3 only on the second, p = 0.34375, paired difference interval -2 to 10)). None of the three primary tests met p below 0.05, so H53.9 fails as preregistered and C6 does not survive as worded for the primary reader. The planned D repeats were similar for the stored conclusion (41 against 38, p = 0.375) and stronger for the stored procedure (38 against 32, p = 0.03125).

As a second reader check, Witness C with the stored conclusion produced 27 unsafe verdicts against 18 for its plain read (p = 0.011719, interval 3 to 16). Witness C with the stored procedure produced 2 against 18 (p = 3.1e-05, interval 9 to 24). Those checks are not a substitute for the failed primary tests.

12. Evidence locates; it does not certify (C7: H53.10)

Among 529 unsafe decisive outputs of single read calls (reads, repeats, resamples, deliberate reads and memory calls) that returned a nonempty quote, 526 quoted the packet exactly after whitespace and quotation mark normalization (0.9943, Wilson interval 0.9835 to 0.9981). Both the 0.90 rate and the 0.80 lower bound thresholds are met, so H53.10 passes and C7 survives.

13. Stability and domain generality (H53.11 and H53.12)

Temperature 0 repeats agreed with the first read on 183 of 188 (0.9393 to 0.9886) items for Witness B, 180 of 188 (0.9183 to 0.9783) for Witness C, and 183 of 188 (0.9393 to 0.9886) for Witness D. Each rate is at least 0.95, so H53.11b passes. Witness D released 29 unsafe verdicts on track G, so H53.11a passes.

In each of the six track G groups (20 items each), both route(D, B) and route(D, D deliberate) had no more unsafe outcomes than route(D, D resample 1):

Group	Route D with D resample 1	Route D with B	Route D with D deliberate	Witness D read
software	5	0	2	5
science	4	0	2	4
health	4	1	2	4
civic	4	1	2	5
logistics	5	1	0	6
finance	4	0	1	5

H53.12 therefore passes. These are 20 item slices, so the per group counts are descriptive.

14. One held out near miss, walked through

Packet G16S (Harbor Lights Festival permit, City of Bayview). Subsection (d) allows the permit only when the Special Events Office has received neighborhood sign-offs that together represent at least 75 percent of adjacent households, and each sign-off attaches a signed noise plan. Old Pier Association (46 percent) attached a signed plan; Seawall Terrace Association (34 percent) attached an unsigned plan; Marina Heights Association (20 percent) does not appear. Nothing in the packet establishes or excludes a later signed Seawall plan before the Permit Date, July 6, 2026, so the open record label is HOLD.

Witness D answered FAIL: qualifying coverage is only 46 percent, below 75 percent. Its first resample gave the same FAIL and the same arithmetic. Witness C also answered FAIL. Two calls answered HOLD: Witness B (a later qualifying Seawall submission would raise the total to 80 percent and is neither shown nor excluded) and Witness D's deliberate read, whose case for HOLD was that the packet neither establishes nor excludes a signed

replacement from Seawall by July 6, and that this fact determines whether the qualifying total is 46 or 80 percent. Agreement between Witnesses C and D would have released the wrong FAIL. Routing Witness D with Witness B, or with its own deliberate read, would have held.

15. Findings versus interpretations

Structural findings

• Routing is unsafe in 4 of 27 states, exactly when both reads share a wrong decisive verdict, and is never worse than either read; flag only review is unsafe in 4 states; rewriting review is unsafe in 12 states, including 6 new ones. The routing identity holds on 2067 of 2067 recorded pairs.

• All track R prompts match Round 49 and Round 52; all 120 track G certificates and all 32 SEC windows check.

Empirical findings

• Witness D released 38 unsafe verdicts on 188 items (interval 0.1510 to 0.2652), including 9 of 9 Round 49 near misses and 24 of 24 held out near misses, and 29 unsafe verdicts on track G. Witness C released 18; Witness B released 4. No reader released an unsafe verdict on the 32 SEC items except Witness D with a stored conclusion (1 of 32).

• Routing Witness D with its own resample left 35 unsafe verdicts; with Witness B, 4; with Witness C, 18; with its deliberate read, 13. Spearman rho between disagreement and unsafe verdicts removed was 0.968.

• Removing the HOLD case from the deliberate prompt raised unsafe verdicts from 13 to 32 on 187 items (p = 2.1e-05).

• Routing with B, C or the deliberate read each removed at least as many unsafe verdicts as the false holds it added (34 against 4; 20 against 1; 25 against 12).

• Seeded wrong proposals were not kept more often than the same reviewer's blind error for two of three reviewers. Witness C kept 9 seeds against 18 blind errors.

• Flag only review created 0 new unsafe releases and left 12 unsafe verdicts; rewriting review created 0 new unsafe releases and left 21.

• On Witness D, stale memory produced 44 unsafe verdicts against 38 (p = 0.070312) and the stored procedure produced 33 against 38 (p = 0.125). On Witness C the same contrasts were 27 against 18 (p = 0.011719) and 2 against 18 (p = 3.1e-05).

• 526 of 529 unsafe decisive quotes copied the packet exactly (0.9943, interval 0.9835 to 0.9981).

• Temperature 0 repeats agreed on 183 of 188, 180 of 188, and 183 of 188 items. Both different-read routes beat same-reader resampling in all 6 track G groups.

Interpretation

On these items a second read helped when it read differently, and almost not at all when it was the same reader under the same procedure. A different procedure (argue each verdict, including HOLD) supplied a large part of that difference for Witness D, and removing the HOLD case took the benefit away. Agreement routing could not release a verdict that either read would not release, and the unsafe verdicts it stopped were at least the false holds it added, though the deliberate read paid a higher hold cost than routing with Witness B. Those four points are the surviving content of C1, C2, C3 and C7.

C4, C5 and C6 do not survive as worded. A wrong proposal did not produce the predicted inheritance; if anything Witness C used the seed as something to reject. Rewriting review did not create a new unsafe release here, so the

second clause of C5 is unsupported on this pool even though flag only review behaved as predicted. Memory effects ran in the predicted direction for Witness D but missed the preregistered significance threshold; they were clear for Witness C. Because readers differ in capability, no ranking across witnesses is a resource comparison, and because track G is 24 families with five variants each, treating 188 items as independent overstates precision.

16. Hypothesis outcomes

ID	Outcome	Basis
H53.1	Supported	Enumeration of 27 states and 2067 of 2067 recorded routing identities
H53.2	Supported	FAIL on 9 of 9 near misses; resamples reproduced 18 of 18; deliberate HOLD on 5 of 9
H53.3	Supported	Route with B, C and D deliberate: 4, 18 and 13 unsafe against 35 for D resample; all three p below 0.05
H53.4	Supported	Spearman rho = 0.968 across 11 second reads
H53.5	Supported	Both parts: deliberate routing beat resample routing (p = 0.0); HOLD ablation raised unsafe verdicts (p = 2.1e-05)
H53.6	Supported	Unsafe removed at least matched false holds added for B, C and D deliberate
H53.7	Rejected	No reviewer showed a significant excess of keeping the seeded wrong verdict over its own blind error; Witness C kept fewer
H53.8	Rejected	Flag only review created 0 new unsafe verdicts and had fewer residual unsafe verdicts than rewriting, but rewriting also created 0 new unsafe verdicts
H53.9	Rejected	On Witness D, stale p = 0.070312, procedure p = 0.125, ablation p = 0.34375, all above 0.05
H53.10	Supported	526 of 529 exact quotes, interval 0.9835 to 0.9981
H53.11	Supported	Both parts: 29 unsafe on track G; repeats at or above 95 percent
H53.12	Supported	Both different-read routes had no more unsafe outcomes than resample routing in all 6 groups

C1, C2, C3 and C7 survive this round. C4, C5 and C6 do not survive as worded.

17. Synthesis and analysis audit

Witness A reviewed a summary of the completed analysis, after the fact, as commentary rather than gold. Its assessment, in part: "Structural facts hold by construction: route unsafe only when both calls share a wrong label, route never worse than either call, flag unsafe a subset, rewrite opens six new unsafe states; H53.1 only confirms records obey the rule. Empirical: reads D 38, C 18, B 4 unsafe; D deliberate 13 unsafe plus 12 false holds; every two call route containing D read equals its stronger member exactly, and D read with D resample removes 3 of 38 while sharing 35 wrong. C with the stored procedure reaches 2 unsafe; D moves 38 to 33. H53.7, H53.8 and H53.9 fail, so C4, C5 and C6 fall. Overreach: C1 credits a second call, but matched two call arms differ only in which reader is present and the second call never alters the stronger reader, so the evidence shows reader capability, not routing; same reader resamples are not independent evidence. C2 credits deliberation, yet without the HOLD

option unsafe returns to 32, so the gain is mostly HOLD, paid for with 12 false holds. H53.4's Spearman over 11 overlapping rows all paired to one D read is descriptive, not causal, and ignores false holds. H53.12 rests on 0 to 6 events per 20 item group. Round 52 agreement and 183/188 stability show repeatability, not truth."

Its caveats were: (1) G is 24 families times 5 variants; CIs and paired tests treat 188 items as independent and overstate precision. S is 32/32 in every arm and discriminates nothing. (2) S13 gold HOLD is disputed as FAIL; if the audit is right every arm hides one false hold on S. 36 labels appear unaudited. (3) The procedure effect is reader specific (C 18 to 2, D 38 to 33, open record ablation not significant); if authored alongside the G families it may encode their construction. (4) Anchoring evidence conflicts: seeded wrong proposals cut C's errors (9 vs 18) yet C rewriting D's read gave 21 unsafe against 18 blind; the seed arm cannot separate anchoring from attention drawn to the disputed clause. (5) 100 unusable first outputs were rerun and 3 empties dropped without stated criteria; if unusability tracks difficulty, rerun arms are favoured.

Caveat 2 is answered in part: track R labels were audited blind in Round 49, and the six flagged items plus the one disputed SEC item (S13) form the sensitivity subset (181 items), which does not change any hypothesis outcome reported above. The remark that 36 labels appear unaudited is not adopted for that reason. Caveat 5 is answered by the deviation record: the rerun rule was unusable output (error, empty, timeout, or a JSON object that did not parse to a verdict), applied after the planned arms finished and disclosed before scoring. Caveats 1, 3 and 4 are adopted as limitations. Its suggested next checks are taken up in section 19.

18. Completion disclosure and limitations

Every planned experimental call was attempted: 4 blind label batches and 32 reader or reviewer phases of 188 items, 6016 experimental calls, plus one synthesis call (6021 recorded calls in the public transcripts). One hundred first outputs were unusable and were rerun once so that each planned arm would be complete; seven of those were rerun a second time. Three outputs remain empty after both reruns: the repeated deliberate read on G16S, the HOLD ablation on RM04P, and Witness D's self review on RM04P. Unusable outputs are never scored as HOLD. No packet, claim, label, hypothesis or scoring rule was edited after outcomes were seen. No planned call was dropped by a resource stop. The deviation record is `deviations-r53.md`.

Limitations: track G is 24 families with five variants, so item level intervals treat dependent variants as independent and overstate precision; track S was at ceiling for every reader without memory, so it does not separate those arms; reader capability and resource per call are not matched across witnesses; temperature 0.7 is one setting; track R labels come from Round 49; the stored procedure is one domain general text and may fit the construction of the authored families more neatly than a procedure written without sight of them would; results are about these readers, prompts and items, not about any reader class.

19. Next checks for the theory program

1. Round 54 should replicate the surviving statements (C1, C2, C3, C7) on new packets and surface forms in the standing domains, and should report family clustered intervals as well as item level ones.

2. Restate C4, C5 and C6 more narrowly, or drop them: inheritance was not observed; rewriting added no new unsafe release here; memory effects were reader specific and, for Witness D, below the preregistered threshold.

3. Test the stored procedure on families authored without sight of its wording, and include Witness B as well as C and D.

4. The Round 51 dissent rule with executable counterexamples remains open.

20. Reproducibility

The archive includes the fixed plan, continuity audit, deviation record, the Round 49 packet builder and the Round 49 and Round 52 reference calls used for identity checks, the held out family builder, the material builder with

certificates, the SEC track with exact windows, accessions and links, source and bibliography verification, the live runner with every prompt, verbatim transcripts of all recorded calls, the blind audit and synthesis, replay analysis, the exact verifier and this manuscript with its PDF builder. The verifier makes no network calls. The private execution configuration is not included and is not required to replay the recorded evidence.

21. Verified references

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

• [S103] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171, first posted March 21, 2022. https://arxiv.org/abs/2203.11171

• [S106] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum and Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325, first posted May 23, 2023. https://arxiv.org/abs/2305.14325

• [S111] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge and Jonas Geiping. Great Models Think Alike and this Undermines AI Oversight. arXiv:2502.04313, first posted February 6, 2025. https://arxiv.org/abs/2502.04313

• [S112] Elliot Kim, Avi Garg, Kenny Peng and Nikhil Garg. Correlated Errors in Large Language Models. arXiv:2506.07962, first posted June 9, 2025. https://arxiv.org/abs/2506.07962

• [S113] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang and Ethan Perez. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, first posted October 20, 2023. https://arxiv.org/abs/2310.13548

• [S114] Arjun Panickssery, Samuel R. Bowman and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076, first posted April 15, 2024. https://arxiv.org/abs/2404.13076

• [S115] Jiaxu Lou and Yifan Sun. Anchoring Bias in Large Language Models: An Experimental Study. arXiv:2412.06593, first posted December 9, 2024. https://arxiv.org/abs/2412.06593

• [S116] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe and Lucy Lu Wang. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, first posted July 25, 2024. https://arxiv.org/abs/2407.18418

• [S117] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685, first posted June 9, 2023. https://arxiv.org/abs/2306.05685

• [S118] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association, 1927. https://doi.org/10.1080/01621459.1927.10502953

• [S119] Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. Psychometrika, 1947. https://doi.org/10.1007/BF02295996

• [S120] Raymond Reiter. On Closed World Data Bases. Logic and Data Bases, 1978. https://doi.org/10.1007/978-1-4684-3384-5_3

• [S30] U.S. Securities and Exchange Commission. EDGAR Archives. https://www.sec.gov/Archives/edgar/data/. The 32 exhibits in section 3 were opened and verified for this run.