Encevra Research Round 54

Title: The Same Statements, New Cases: Held Out Replication of Domain General Propositions on New Packets and New Surface Forms

Round code: HRP-54

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

Date: November 11, 2026

Status: All planned experimental calls were attempted. One hundred and five first outputs were unusable and were rerun once; one of those was rerun a second time and then parsed; none remain unusable. No planned call was dropped by a resource stop. Published November 13, 2026. DOI 10.5281/zenodo.24149599.

Abstract

The Theory of Collaborative Intelligence roadmap sets Round 54 as the last experimental round before Theory v1 (Round 55, a multi-day round). It is a held out replication of the seven Round 53 candidate propositions (C1 to C7), worded identically, on new packets and new surface forms. A statement survives this round only if every hypothesis that tests it passes. A statement is recommended for Theory v1 only if it survives this round and also survived Round 53. If it failed in Round 53 and fails again, it is not recommended as worded. If the two rounds disagree, the statement is deferred rather than forced into v1.

The materials are 152 items, all new. Track H adds 120 held out items from 24 families in six domain groups that were not used in Round 53 (agriculture and food safety; environmental monitoring; aviation and airport operations; museums, archives and cultural heritage; telecommunications and broadcast; housing, building inspection and occupancy), with the same five variants and two world certificates as Round 53, presented in three surface forms that Round 53 did not use (a records table, a dated operations log, and correspondence letters; 40 items each). Track S adds 32 real SEC notes or indenture windows, all new by accession and filer against the recorded prior set through Round 53, each with one fixed annual interest-rate claim. There is no replay of Round 49 or Round 53 packets. Witness A agreed with 148 of 152 audited labels and disputed four items, which are excluded from the sensitivity subset (148 items). The sensitivity subset does not change any hypothesis outcome reported below.

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. The routing identity holds on 1,672 of 1,672 recorded pairs. Live, Witness D released 35 unsafe verdicts on 152 items (Wilson 95 percent interval 0.1705 to 0.3033), all on track H, including a wrong FAIL on 22 of 24 held out near miss items and on 12 of 24 partial period certificates. Routing that read with its own resample left 29 unsafe verdicts. Routing it with Witness B left 3; with Witness C, 8; with Witness D's own deliberate read, 14. The Spearman rank correlation between the disagreement rate of eleven second reads (computed without gold) and the number of unsafe verdicts routing removed was 0.9495. Removing the HOLD case from the deliberate prompt raised unsafe verdicts from 15 to 26 ($p = 0.003418$).

C1, C2, C3 and C7 survive this round, as they did in Round 53, and are recommended for Theory v1. C4, C5 and C6 do not survive as worded, as they did not in Round 53, and are not recommended as worded. No statement is deferred by disagreement between the two rounds. A seeded wrong proposal was not kept more often than the same reviewer's own blind error (H54.7). 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 (H54.8); flag only review created none either and left fewer residual unsafe verdicts (5 against 11). On Witness D, a stored conclusion from a different file did not raise unsafe verdicts (33 against 35, $p = 0.753906$); a stored procedure did lower them (26 against 35, $p = 0.022461$); removing the open record steps did not remove that benefit at the preregistered threshold (28 against 26, $p = 0.753906$). Among 395 unsafe decisive reads with a nonempty quote, 393 quoted the

packet exactly (0.9949, interval 0.9817 to 0.9986). Temperature 0 repeats agreed on 134 of 152, 146 of 152 and 136 of 152 items for Witnesses B, C and D; only Witness C met the preregistered 95 percent stability threshold.

1. Archive check and continuity

The staging registry was read before any live call. Its newest published entry is Round 53, DGP-53, DOI 10.5281/zenodo.24134801, so this is Round 54. Mechanism continuity with Round 53 is intentional and disclosed: the seven candidate propositions are worded identically, the reader prompt, deliberate prompt, review prompts, scoring rules and two world certificate are the same, and the stored procedure is the same text. What is new is the packets, the domain groups, the surface forms, and the SEC instrument type (notes and indentures with a fixed annual interest-rate claim, not executive employment agreements). All 32 SEC accessions and filers are new against the recorded prior set through Round 53. The continuity audit is `anti-repetition-verification.md`.

Round 53 handed off a clear v1 rule: retain only what survives both this held out replication and Round 53. This round takes that rule as fixed before testing. Before the materials were frozen, one Track S filer was replaced because an affiliate name had appeared in a prior round; that substitution is in the continuity audit and in the deviation record, and it was made before any live call.

2. Fixed plan, candidate propositions and hypotheses

The plan was fixed before any live call (`run-plan-r54.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 54). The primary population is all 152 items. Surface populations are the 40 table items, the 40 log items and the 40 letter items, each judged separately where a hypothesis names them. Sensitivity excludes any H or S item whose label Witness A disputes.

Candidate propositions, stated before testing, worded identically to Round 53:

• 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
H54.1	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; the routing identity holds on every recorded pair	C3, C5
H54.2	On the 24 new track H near miss items, Witness D is FAIL on at least 18 of 24; its resamples reproduce at least 80 percent of its wrong decisive verdicts; its deliberate read moves at least 8 of the 24 to HOLD	continuity
H54.3	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
H54.4	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
H54.5	route(D, D deliberate) beats route(D, D resample 1); removing the HOLD case from the deliberate prompt removes the benefit	C2
H54.6	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
H54.7	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
H54.8	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
H54.9	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
H54.10	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
H54.11	Witness D releases at least 10 unsafe verdicts on track H; each reader's temperature 0 repeat agrees on at least 95 percent of usable items	manipulation
H54.12	In at least 5 of 6 track H domain groups, route(D, B) and route(D, D deliberate) each have no more unsafe outcomes than route(D, D resample 1)	generality

ID	Fixed statement	Tests
H54.13	On at least two of the three surfaces (table, log, letter), route(D, B) and route(D, D deliberate) each have no more unsafe outcomes than route(D, D resample 1)	surface

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]. Models also show anchoring on an earlier figure [S115]. Abstention, including a hold when the record does not settle a decisive fact, is a separate skill from answering [S116]. Held out evaluation across domains is a standard precaution against fitting a method to one distribution [S121], and tables are a surface on which factual claims are often verified [S122]. This round applies those cautions by taking the seven propositions in the same domain general words as Round 53 and trying to break each one on new packets and new surfaces, with preregistered tests, Wilson intervals [S118], McNemar tests [S119] and paired bootstrap intervals [S123].

3. Materials and gold

Track H (120 items) is 24 new authored families, four in each of six groups that were not used in Round 53, none of them contracts and none of them SEC exhibits. Each family states one requirement with a universal part and an aggregate threshold. Five variants per family, gold by the same two world certificate as Round 53: P (two qualifying records reach the threshold, PASS), S (the second record is a near miss and nothing establishes or excludes a later qualifying record, HOLD), C (S plus a certificate through the decision date that no other record was received, FAIL), W (S plus a certificate that covers only a period ending seven days before the decision date, HOLD), R (S plus a replacement qualifying record, PASS). Surface forms are new relative to the Round 53 labeled document and record packets: eight families as a records table, eight as a dated operations log, and eight as correspondence letters (40 items each). Opaque identifiers Q001 to Q120 were assigned by seed 54.

Track S (32 items) is 32 notes or indenture exhibits (28 EX-4 series, and four 424B pricing documents where a clean EX-4 window was not available), all new by accession and filer against every prior recorded exhibit through Round 53. Each window contains one sentence that states a fixed annual interest rate. Claims: more than a figure below the stated rate (GT, PASS, 8), less than that figure (LT, FAIL, 8), the interest rate has never been amended above the stated rate (NOCHG, a world fact, HOLD, 8), interest is the only payment the issuer is required to make under the instrument (ONLY, scope beyond the window, HOLD, 8). Gold totals: 56 PASS, 64 HOLD, 32 FAIL (track H 48, 48, 24; track S 8, 8, 16).

There is no replay of Round 49 or Round 53 packets. Every item is new. All 120 track H certificates reproduce their gold labels. All 32 SEC windows contain their rate sentence, and the nearby comparison figure sits below the stated rate. All 152 seeded proposals use the wrong verdict and an exact quote from the packet.

ID	Filer	Exhibit	Filed	Accession	Stated rate	Claim	Gold	Source
S01	FIRST BUSEY CORP /NV/	EX-4.2	2022-06-03	0001104659-22-068088	5.000%	above a nearby figure	PASS	exhibit
S02	PECO ENERGY CO	EX-4.1	2022-08-23	0000078100-22-000007	4.375%	interest is the only required payment	HOLD	exhibit

ID	Filer	Exhibit	Filed	Accession	Stated rate	Claim	Gold	Source
S03	INTEL CORP	EX-4.1	2023-02-10	0001193125-23-032608	4.875%	rate never amended above the stated rate	HOLD	exhibit
S04	PEPSICO INC	424B2	2023-02-14	0001104659-23-021900	4.550%	below a nearby figure	FAIL	exhibit
S05	Cheniere Energy Partners, L.P.	424B3	2022-06-03	0001193125-22-167333	3.25%	interest is the only required payment	HOLD	exhibit
S06	SEMPRA ENERGY	EX-4.2	2023-02-28	0001032208-23-000008	5.75%	below a nearby figure	FAIL	exhibit
S07	PPL Corp	EX-4.(B)	2023-03-02	0000922224-23-000018	5.00%	above a nearby figure	PASS	exhibit
S08	IDAHO POWER CO	424B5	2023-03-09	0001193125-23-066303	5.50%	interest is the only required payment	HOLD	exhibit
S09	VIRGINIA ELECTRIC & POWER CO	EX-4.3	2023-03-30	0001193125-23-084700	5.45%	rate never amended above the stated rate	HOLD	exhibit
S10	WarnerMedia Holdings, Inc.	424B3	2023-03-31	0001193125-23-088017	3.428%	below a nearby figure	FAIL	exhibit
S11	PORTLAND GENERAL ELECTRIC CO /OR/	EX-4.1	2023-08-29	0000784977-23-000126	5.44%	rate never amended above the stated rate	HOLD	exhibit
S12	ATMOS ENERGY CORP	EX-4.1	2023-10-10	0001193125-23-253163	5.900%	above a nearby figure	PASS	exhibit
S13	Energy Transfer LP	EX-4.2	2023-10-13	0001193125-23-256129	6.050%	rate never amended above the stated rate	HOLD	exhibit
S14	ELI LILLY & Co	EX-4.3	2024-02-09	0001193125-24-029551	4.500%	above a nearby figure	PASS	exhibit
S15	Keurig Dr Pepper Inc.	EX-4.2	2024-03-07	0001193125-24-062259	5.100%	below a nearby figure	FAIL	exhibit
S16	Ingersoll Rand Inc.	EX-4.1	2024-05-10	0001140361-24-025480	5.197%	above a nearby figure	PASS	exhibit
S17	NOKIA CORP	424B3	2024-08-21	0001193125-24-204391	2.125%	interest is the only required payment	HOLD	exhibit
S18	SOUTHERN CO	EX-4.F3	2024-10-31	0000092122-24-000089	4.63%	rate never amended above the stated rate	HOLD	exhibit

ID	Filer	Exhibit	Filed	Accession	Stated rate	Claim	Gold	Source
S19	EASTERN GAS TRANS MISSION & STORAGE, INC.	EX-4.2	2024-12-10	0001193125-24-274355	5.020%	interest is the only required payment	HOLD	exhibit
S20	EASTERN ENERGY GAS HOLDINGS, LLC	EX-4.2	2025-01-15	0001193125-25-007024	5.800%	above a nearby figure	PASS	exhibit
S21	AFFILIATED MANAGERS GROUP, INC.	EX-4.15	2025-02-14	0001004434-25-000010	5.875%	rate never amended above the stated rate	HOLD	exhibit
S22	Targa Resources Corp.	EX-4.2	2025-02-27	0001193125-25-039300	5.550%	above a nearby figure	PASS	exhibit
S23	Motorola Solutions, Inc.	EX-4.2	2025-06-16	0001193125-25-141449	4.850%	interest is the only required payment	HOLD	exhibit
S24	AMAZON COM INC	EX-4.1	2025-11-20	0001104659-25-114647	3.900%	rate never amended above the stated rate	HOLD	exhibit
S25	JOHNSON & JOHNSON	EX-4.B	2026-02-11	0000200406-26-000016	1.150%	rate never amended above the stated rate	HOLD	exhibit
S26	NASDAQ, INC.	EX-4.21	2026-02-12	0001628280-26-007703	4.500%	below a nearby figure	FAIL	exhibit
S27	ARRAY DIGITAL INF RASTRUCTU RE, INC.	EX-4.7	2026-02-20	0000821130-26-000012	6.25%	below a nearby figure	FAIL	exhibit
S28	REALTY INCOME CORP	EX-4.96	2026-02-25	0000726728-26-000011	5.000%	above a nearby figure	PASS	exhibit
S29	CLOROX CO /DE/	EX-4.2	2026-05-11	0001193125-26-216807	4.700%	below a nearby figure	FAIL	exhibit
S30	STATE STREET CORP	EX-4.1	2026-07-23	0001193125-26-314120	4.701%	below a nearby figure	FAIL	exhibit
S31	Aon plc	EX-4.2	2026-09-17	0001193125-26-394533	5.350%	interest is the only required payment	HOLD	exhibit
S32	MICROSOFT CORP	EX-4.1	2026-11-04	0001193125-26-447173	4.850%	interest is the only required payment	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, subject to the disclosed mid-run adjustment in `deviations-r54.md`. 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 H and S blind in four batches (4 of 4 batches usable after a disclosed rerun of the first batch).

Item order was shuffled with seed 54. 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 (E54-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: 1,672 of 1,672 (H54.1).

6. The near miss pattern, on new items (E54-B)

Witness D's single read was FAIL on 22 of 24 new track H near miss items (H54.2a; the threshold was at least 18 of 24). The two items it did not fail were H10S, which it correctly held (this family is one of the four Witness A disputes), and H23S, which it also correctly held. Across the two resamples, it reproduced its own wrong decisive verdict on 58 of 70 usable (item, resample) pairs (0.8286, interval 0.7238 to 0.8991), which meets the preregistered point estimate of at least 80 percent. The interval includes values below 0.8; that is a limitation of the precision of this check, not a change to the pass rule, which was the point estimate. Its deliberate read moved 10 of the 24 near miss items to HOLD, meeting the threshold of at least 8. All three parts of H54.2 therefore pass.

Call	Near miss FAIL (wrong)	Near miss HOLD (correct)	Unusable
Witness D read	22	2	0
Witness D repeat	22	2	0
Witness D resample 1	21	3	0
Witness D resample 2	23	1	0
Witness D deliberate	14	10	0
Witness D deliberate, repeat	15	9	0
Witness D deliberate without the HOLD case	21	3	0
Witness C read	9	15	0
Witness C deliberate	3	21	0
Witness B read	3	21	0

On track H, Witness D was FAIL on 22 of 24 near miss (S) items and on 12 of 24 partial certificate (W) items, and was FAIL on 1 of 24 replacement (R) items. It released 8 false holds on the 24 passing (P) items, 1 on the closed certificate (C) items, and 1 on the replacement items. That is 35 unsafe verdicts on the 120 held out items (interval 0.2178 to 0.3784), so the held out track can separate arms (H54.11a). Twenty-three of the 24 families have at least one unsafe Witness D verdict (12 families have two, 11 have one, and the disputed H10 family has none). Item level intervals treat those dependent variants as independent and overstate precision. On track S, Witness D, Witness C and Witness B were each correct on all 32 items, and no scored arm released an unsafe verdict on track S.

7. Difference, not repetition (C1: H54.3 and H54.4)

Primary population, all 152 items:

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe rate interval	Unsafe, sensitivity (148)	Unsafe, H (120)	Unsafe, S (32)
Witness B read	1	133 of 152	3	16	0	0.0067 to 0.0564	3 of 148	3	0
Witness C read	1	139 of 152	9	4	0	0.0315 to 0.1087	9 of 148	9	0
Witness D read	1	107 of 152	35	10	0	0.1705 to 0.3033	35 of 148	35	0
Witness D deliberate	1	112 of 152	15	25	0	0.0607 to 0.1564	15 of 148	15	0
Witness D deliberate without the HOLD case	1	109 of 152	26	17	0	0.1195 to 0.2388	26 of 148	26	0
Witness C deliberate	1	143 of 152	3	6	0	0.0067 to 0.0564	3 of 148	3	0
Witness D resample 1	1	111 of 152	33	8	0	0.1590 to 0.2892	33 of 148	33	0
Route D read with D resample 1	2	110 of 152	29	13	0	0.1363 to 0.2606	29 of 148	29	0
Route D read with B read	2	128 of 152	3	21	0	0.0067 to 0.0564	3 of 148	3	0
Route D read with C read	2	133 of 152	8	11	0	0.0269 to 0.1004	8 of 148	8	0
Route D read with D deliberate	2	112 of 152	14	26	0	0.0557 to 0.1487	14 of 148	14	0
Route D read with C deliberate	2	137 of 152	3	12	0	0.0067 to 0.0564	3 of 148	3	0

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe rate interval	Unsafe, sensitivity (148)	Unsafe, H (120)	Unsafe, S (32)
Witness C flag only review of Witness D read	2	134 of 152	5	13	0	0.0141 to 0.0747	5 of 148	5	0
Witness C reviews Witness D read (may rewrite)	2	131 of 152	11	10	0	0.0409 to 0.1249	11 of 148	11	0
Witness D reviews its own read (may rewrite)	2	110 of 152	29	13	0	0.1363 to 0.2606	29 of 148	29	0
Witness D with a stored conclusion	1	110 of 152	33	9	0	0.1590 to 0.2892	33 of 148	33	0
Witness D with the stored procedure	1	113 of 152	26	13	0	0.1195 to 0.2388	26 of 148	26	0
Witness D with the procedure minus its open record steps	1	114 of 152	28	10	0	0.1306 to 0.2534	28 of 148	28	0
Witness C with a stored conclusion	1	131 of 152	17	4	0	0.0710 to 0.1718	17 of 148	17	0
Witness C with the stored procedure	1	146 of 152	0	6	0	0.0000 to 0.0246	0 of 148	0	0

Routing Witness D with its own resample left 29 unsafe verdicts on 152 items (interval 0.1363 to 0.2606). Routing it with Witness B left 3 (interval 0.0067 to 0.0564; McNemar p = 0.0; bootstrap difference interval 17 to 35). Routing it with Witness C left 8 (interval 0.0269 to 0.1004; p = 0.000006; interval 13 to 30). Routing it with Witness D's deliberate read left 14 (interval 0.0557 to 0.1487; p = 0.000275; interval 8 to 23). All three point estimates favor the different read, and all three McNemar tests are below 0.05, so H54.3 passes.

Across the eleven second reads paired with Witness D, the Spearman rank correlation between the observed disagreement rate (computed without gold) and the number of unsafe verdicts routing removed was 0.9495, above the preregistered 0.7, so H54.4 passes. The rows, in brief: D repeat, disagreement 0.1053, 7 unsafe removed; D resample 1, 0.0921, 6; D resample 2, 0.0855, 6; D deliberate, 0.2500, 21; D deliberate without the HOLD case, 0.1118, 10; D with the stored procedure, 0.1316, 11; C read, 0.2237, 27; C resample 1, 0.2105, 26; C deliberate, 0.2500, 32; B read, 0.3026, 32; B resample, 0.3026, 33. A second call from the same reader under the same procedure (the repeat and the two resamples) removes little; a different reader or a different procedure removes more, in proportion to how often it disagrees.

C1 therefore survives this round. It also survived Round 53, so it is recommended for Theory v1.

8. Procedure can supply the difference (C2: H54.5)

Route(D, D deliberate) left 14 unsafe verdicts against 29 for route(D, D resample 1), p = 0.000275 (H54.5a). Witness D's deliberate read without the HOLD case left 26 unsafe verdicts against 15 for the intact deliberate prompt, p = 0.003418 (H54.5b). Both parts pass. The gain is concentrated on arguing for a hold: when the prompt no longer asks for the strongest case for HOLD, the near miss FAIL count rises from 14 to 21 of 24. C2 therefore survives this round. It also survived Round 53, so it is recommended for Theory v1.

9. Disagreement buys a hold, not a vote (C3: H54.1 and H54.6)

The structural half is H54.1, which passes: routing cannot release a verdict that either read would not release, and the identity holds on every recorded pair. The cost half is H54.6. Relative to Witness D's single read (35 unsafe, 10 false holds), route(D, B) removed 32 unsafe verdicts and added 11 false holds; route(D, C) removed 27 and added 1; route(D, D deliberate) removed 21 and added 16. In all three cases the unsafe verdicts removed are at least the false holds added, so H54.6 passes. C3 therefore survives this round. It also survived Round 53, so it is recommended for Theory v1. The deliberate route pays a larger false hold bill (16) than the B or C routes; that is the price of the hold, not a vote, and it is disclosed rather than netted away.

10. Review inheritance and flag only review (C4 and C5: H54.7 and H54.8)

No reviewer kept a seeded wrong verdict more often than it reached that same wrong verdict on its own blind read, at the preregistered threshold. Witness B kept 2 seeds and made 3 matching blind errors (p = 1.0). Witness C kept 3 and made 9 (p = 0.070312). Witness D kept 29 and made 35 (p = 0.070312). Planned repeats of C and D agree in direction. H54.7 fails, as H53.7 failed. C4 does not survive as worded, failed in both rounds, and is not recommended for Theory v1.

Flag only review by Witness C of Witness D's read created 0 new unsafe verdicts and left 5 residual unsafe verdicts, having fixed 30. Rewriting review by Witness C created 0 new unsafe verdicts and left 11, having fixed 24. Flag only therefore created no new unsafe release and had fewer residual unsafe verdicts than rewriting, but rewriting also created no new unsafe release, so the claim that rewriting can add an unsafe release was not observed on this reviewer. H54.8 fails, as H53.8 failed. (Witness D's self review did create 2 new unsafe verdicts and left 29, but the preregistered test is Witness C's rewriting review.) C5 does not survive as worded, failed in both rounds, and is not recommended for Theory v1. The narrower observation that flag only review added no new unsafe release holds in both rounds and is available for a restatement, if Theory v1 wants one, with a new test.

11. Memory (C6: H54.9)

On Witness D, a stored conclusion from a different file produced 33 unsafe verdicts against 35 for the blind read (p = 0.753906), so H54.9a fails, and the point estimate is in the opposite direction from the statement. A stored procedure produced 26 against 35 (p = 0.022461), so H54.9b passes; the planned repeat was 21 against 35 (p = 0.000519). Removing the open record steps produced 28 against 26 (p = 0.753906), so H54.9c fails. On Witness C, reported as a second reader check, the stale contrast was 17 against 9 (p = 0.007812) and the procedure contrast was 0 against 9 (p = 0.003906). Because parts a and c fail on Witness D, H54.9 as a whole does not pass and C6 does not survive as worded. It also failed in Round 53 (where even the procedure contrast on Witness D was below the threshold). It is not recommended for Theory v1 as worded. The procedure contrast is now significant for Witness D as well as for Witness C; that is a narrower empirical finding, not a survival of the three-part statement.

12. Evidence locates; it does not certify (C7: H54.10)

Among 395 unsafe decisive outputs of single read calls (read, repeat, resample, deliberate, memory) with a nonempty quote, 393 copied the packet exactly after whitespace and quotation mark normalization (0.9949, interval 0.9817 to 0.9986). Both the rate and the Wilson lower bound meet the preregistered thresholds (0.9 and 0.8). H54.10 passes. C7 therefore survives this round. It also survived Round 53, so it is recommended for Theory v1.

13. Stability, domain generality and surface form (H54.11, H54.12, H54.13)

H54.11a passes: Witness D released 35 unsafe verdicts on track H, above the threshold of 10. H54.11b fails: temperature 0 repeats agreed with the first read on 134 of 152 items for Witness B (0.8816), 146 of 152 for Witness C (0.9605), and 136 of 152 for Witness D (0.8947). Only Witness C met the preregistered 95 percent threshold. In Round 53 all three readers met it. This is a stability check, not a test of C1 to C7, and it is a limitation for any claim that a single temperature 0 call is a stable property of the reader on these new surfaces.

H54.12 passes: in all 6 track H domain groups, route(D, B) and route(D, D deliberate) each had no more unsafe outcomes than route(D, D resample 1). The group counts of unsafe verdicts, in the order resample, route with B, route with deliberate, and Witness D's read, are: agriculture 5, 3, 3, 6; aviation 4, 0, 2, 4; cultural 4, 0, 2, 5; environment 5, 0, 3, 7; housing 4, 0, 2, 6; telecom 7, 0, 2, 7.

H54.13 passes: on all three surfaces, both different-read routes had no more unsafe outcomes than resample routing (the threshold was at least two of three). Unsafe counts, in the same order, on 40 items each: table 9, 2, 6, and 11 for Witness D's read; log 10, 1, 4, and 12; letter 10, 0, 4, and 12. H54.13 does not by itself decide whether C1 survives; it is the preregistered check that the C1 pattern is not tied to one surface form.

14. One held out near miss, walked through

Family H02 is a Crowmarsh Dairy Board herd movement rule, presented as an operations log. Clause 5 allows shipment of herd lot 18 only when the Board Veterinarian has received tuberculin test reports from barns that together house at least 80 percent of the milking herd, and each such report records a clear result for every animal listed. An inconclusive result is not qualifying. The shipment date is May 11, 2026. North Barn houses 51 percent, South Barn 34 percent, and Calf Barn 15 percent. On the near miss variant (H02S, gold HOLD), the log shows a clear North Barn report on April 20 and an inconclusive South Barn report on April 28, and it is silent about any later report. Qualifying coverage on the open record is 51 percent, but nothing establishes or excludes a later qualifying South Barn or Calf Barn report before the shipment date, so the decisive fact is undetermined.

Witness D's read was FAIL. It quoted the 80 percent rule, counted only the North Barn report as qualifying, and treated the absence of a later report as the absence of a later qualifying fact. Both of its resamples were also FAIL, with the same closed treatment of the open record. Its deliberate read was HOLD: North Barn and South Barn together house 85 percent, so a later qualifying South Barn report would satisfy the clause, and the packet does not exclude one before May 11. Witness C was also HOLD, for the same reason. Witness B was FAIL, and added that even a qualifying Calf Barn report would reach only 66 percent, which is true but does not close the South Barn possibility. The closed world FAIL is the pattern H54.2 records on 22 of 24 new near misses; the deliberate prompt's request to argue for HOLD is the step that moves this item, and 9 others, off that FAIL.

15. Findings versus interpretations

Structural findings

• Over all 27 states, routing is unsafe in 4 states, flag only review in 4, and rewriting review in 12, of which 6 are new. The routing identity holds on 1,672 of 1,672 recorded pairs.

• All 120 track H certificates and all 32 SEC windows check. All 152 seeded proposals are wrong decisive and quote the packet exactly. All 32 SEC accessions and filers are new against the recorded prior set through Round 53.

Empirical findings

• Witness D released 35 unsafe verdicts on 152 items (interval 0.1705 to 0.3033), including 22 of 24 held out near misses and 12 of 24 partial certificates, and 35 unsafe verdicts on track H. Witness C released 9; Witness B released 3. No scored arm released an unsafe verdict on the 32 SEC items.

• Routing Witness D with its own resample left 29 unsafe verdicts; routing with Witness B left 3; with Witness C, 8; with Witness D's deliberate read, 14. Spearman rho between disagreement and unsafe verdicts removed was 0.9495.

• Removing the HOLD case from the deliberate prompt raised unsafe verdicts from 15 to 26 ($p = 0.003418$).

• A seeded wrong proposal was not kept more often than the same reviewer's own blind error for any of the three reviewers.

• Flag only review created 0 new unsafe releases and left 5; rewriting review created 0 new unsafe releases and left 11.

• On Witness D, a stored conclusion produced 33 unsafe verdicts against 35 ($p = 0.753906$), and a stored procedure produced 26 against 35 ($p = 0.022461$); the ablation was 28 against 26 ($p = 0.753906$). On Witness C the same contrasts were 17 against 9 and 0 against 9, both below 0.05.

• 393 of 395 unsafe decisive quotes copied the packet exactly (0.9949, interval 0.9817 to 0.9986).

• Temperature 0 repeats agreed on 134 of 152, 146 of 152 and 136 of 152 items for Witnesses B, C and D. The C1 pattern held in all 6 domain groups and on all 3 surfaces.

• Witness A agreed with 148 of 152 labels and disputed four items (H10P, H10R, H10C and S19). The 148-item sensitivity subset does not change any hypothesis outcome.

Interpretation

C1, C2, C3 and C7 are recommended for Theory v1 because they survived this held out replication and survived Round 53. C4, C5 and C6 are not recommended as worded, because they failed in both rounds. That is a recommendation about these statements as worded, on these readers, prompts and items, not a claim that no narrower restatement could be tested later. Track S was at ceiling for every scored arm, so it does not separate arms and is not evidence that every statement generalizes to notes and indentures; it is evidence that these particular interest-rate claims were easy for these readers. Item level intervals on track H overstate precision because variants of one family are not independent. The H54.2 resample rate meets its point-estimate rule, but its interval includes values below 0.8. Matched call counts are not matched reader capability. A temperature 0 call was less stable on these materials than the Round 53 threshold assumed.

16. Hypothesis outcomes

ID	Outcome	Short result
H54.1	Supported	Enumeration holds, and the routing identity holds on 1,672 of 1,672 recorded pairs

ID	Outcome	Short result
H54.2	Supported	FAIL on 22 of 24 near misses; resamples reproduced 58 of 70 (interval 0.7238 to 0.8991); deliberate HOLD on 10 of 24
H54.3	Supported	3, 8 and 14 unsafe against 29; all three p below 0.05
H54.4	Supported	rho = 0.9495
H54.5	Supported	Both parts: deliberate routing beat resample routing (p = 0.000275); HOLD ablation raised unsafe verdicts (p = 0.003418)
H54.6	Supported	Unsafe removed at least matched false holds added for B (32 against 11), C (27 against 1) and D deliberate (21 against 16)
H54.7	Rejected	No reviewer showed a significant excess of keeping the seeded wrong verdict over its own blind error
H54.8	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
H54.9	Rejected	On Witness D, stale p = 0.753906, procedure p = 0.022461, ablation p = 0.753906; parts a and c fail
H54.10	Supported	393 of 395 exact quotes, interval 0.9817 to 0.9986
H54.11	Mixed	Part a supported (35 unsafe on track H); part b rejected (repeats 134 of 152, 146 of 152 and 136 of 152)
H54.12	Supported	Both different-read routes had no more unsafe outcomes than resample routing in all 6 groups
H54.13	Supported	Both different-read routes had no more unsafe outcomes than resample routing on all 3 surfaces

C1, C2, C3 and C7 survive this round and also survived Round 53, and are recommended for Theory v1. C4, C5 and C6 do not survive as worded in either round, and are not recommended as worded. No statement is deferred.

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: 152 new items; the 120 H items are 24 families whose gold label is fixed by variant letter (P and R PASS, S and W HOLD, C FAIL); 32 S items; certificates and the 1672 of 1672 enumeration check; route and flag relations hold by construction. Empirical: D read 35 unsafe, C 9, B 3, all on track H; S is at ceiling in every arm and discriminates nothing. Route D with B yields 3 unsafe and 21 false holds, the same unsafe count as B alone with 5 more false holds and fewer correct (128 versus 133), so H54.3 and H54.6 measure the second reader's capability rather than a second opinion effect; matched call counts are not a fair comparison without a stronger reader twice baseline. Temperature 0 repeats disagree on 6 to 12 percent of items (H54.11b fails), so single call differences such as H54.9b (11 versus 2 discordant) and the procedure repeat (26 versus 21) sit near run noise. Interpretation: C1, C2, C3 and C7 survive under the stated rule, but C1 and C3 are supported only as stronger reader second claims, and C6 was tested on D only; C's stored conclusion arm (9 to 17 unsafe) shows the

stale effect is reader specific, so neither acceptance nor rejection generalizes across readers."

Its caveats were: (1) Item level intervals and McNemar tests treat 120 H items as independent; they are 24 families; the H10 family is flagged as ambiguous, and the sensitivity set drops only the 4 disputed items rather than the 11 flagged items. (2) The completion allowance was raised mid-run for Witnesses C and D, and 105 unusable calls were rerun until they parsed, which conditions on format success. (3) For H54.8, a McNemar p of 0.07 sits alongside a difference interval of 12 fewer to 1 fewer residual unsafe verdicts for flag only review, and "the fail call depends on which criterion was preregistered." (4) H54.2 passes on a point estimate whose interval includes values below 0.8, and H54.4's Spearman uses 11 rows that include near duplicate arms. (5) Audit agreement of 148 of 152 is one author and one auditor and is not evidence of label truth; the H10 rule wording is ambiguous across all five variants.

Caveat 3 is answered on the preregistered rule: H54.8 fails because Witness C's rewriting review created 0 new unsafe verdicts, which is the third required clause; flag only review created 0 and left 5 against 11, so the first two clauses hold. The McNemar p of 0.07 is not what decides that failure. Caveat 2 is answered in part by the deviation record: the rerun rule was unusable output (empty, or a JSON object that did not parse to a verdict), applied after the planned arms finished and disclosed before scoring; the mid-run change to the allowed completion length is also disclosed there. The caution that this conditions on format success is adopted as a limitation. The remark that C6 was tested on Witness D only is not adopted: Witness C's memory arms were run and reported as a second reader check, and the preregistered pass rule is on Witness D. Caveats 1, 4 and 5, the family independence point in caveat 1, and the interpretation that route(D, B) matches Witness B's own unsafe count (so C1 and C3 should not be read as a pure second opinion effect, and matched call counts are not matched reader capability) 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 152 items, 4,864 experimental reader calls and 4 audit batches (4,868 recorded calls in the working public transcript), plus one synthesis call after analysis. One hundred and five first outputs were unusable and were rerun once so that each planned arm would be complete; one of those (Witness D's deliberate read on H07P) was rerun a second time and then parsed. None remain unusable. After Witness D reads that had been unusable became usable, the validation and flag only arms were run for the first time on those 23 items. 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-r54.md`.

Limitations: track H 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 scored arm, so it does not separate those arms; reader capability and resource per call are not matched across witnesses; temperature 0.7 is one setting; 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; temperature 0 repeats were less stable here than in Round 53; the H54.2 resample interval includes values below the 80 percent point-estimate rule; results are about these readers, prompts and items, not about any reader class.

19. Next checks for the theory program

1. Round 55, the multi-day Theory v1 round, should admit C1, C2, C3 and C7 as worded, on the strength of survival in both Round 53 and this held out replication, and should not admit C4, C5 and C6 as worded.

2. If C4, C5 or C6 are restated more narrowly, they need new preregistered tests. Inheritance was not observed in either round. Rewriting added no new unsafe release by Witness C in either round; flag only added none either. The stored procedure contrast is now significant for Witness D as well as Witness C, but the stale and ablation parts of C6 failed again.

3. Theory v1 should report family clustered intervals as well as item level ones, and should treat the 24 families, not the 120 variants, as the independent units. Track H is 24 families, and 23 of them contribute at least one unsafe Witness D verdict. A sensitivity that drops the whole H10 family, not only the four disputed items, is a useful extra check.

4. Track S has now been at ceiling on employment agreements (Round 53) and on notes and indentures (this round). A harder instrument, or a claim that is not a single quoted rate or salary, is needed if Theory v1 wants a real SEC track to separate arms.

5. Temperature 0 stability should not be assumed at 95 percent on new surfaces; only one of three readers met that threshold here. Key arms should be repeated and between-run variance reported before a p below 0.05 is treated as settled.

6. Equal call baselines (a stronger reader twice, and a deliberate second call from the same reader) should sit beside route(D, B) and route(D, C), so that a second opinion is not confused with a stronger reader. Memory arms should include Witness B as well as C and D.

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

20. Reproducibility

The archive includes the fixed plan, continuity audit, deviation record, 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

• [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

• [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

• [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

• [S121] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine and Chelsea Finn. WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, first posted December 14, 2020. https://arxiv.org/abs/2012.07421

• [S122] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou and William Yang Wang. TabFact: A Large-scale Dataset for Table-based Fact Verification. arXiv:1909.02164, first posted September 5, 2019. https://arxiv.org/abs/1909.02164

• [S123] B. Efron. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 1979. https://doi.org/10.1214/aos/1176344552

• [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.