Encevra Research Round 52

Title: The Second Reader or the Second Call? Matched Call Baselines for Agreement Routing, Validation and Recovery

Round code: MCB-52

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

Date: November 9, 2026

Status: All planned experimental calls attempted once. Twenty one of 797 experimental outputs were unusable (one error, twelve empty, eight incomplete) and were not repeated. The remaining phases were reordered once for the internal spend guard before any of their outputs existed; no call was stopped by the guard. Published November 9, 2026. DOI 10.5281/zenodo.24081926.

Abstract

The Theory of Collaborative Intelligence roadmap sets Round 52 as matched call baselines for the propositions that survived Rounds 40 to 51, so that an improvement counts only against a fair resource comparison. This round asks a narrow version of that question: when a second or third model call improves a verdict, is it because a different reader looked, or only because another call was made? Every composition is compared with baselines that use the same number of calls: the same reader resampled, the same reader validating its own read, and one call with room to argue each verdict.

The materials were chosen so that single readers make decisive errors. Track R replays all 36 Round 49 packets verbatim, with the Round 49 prompt, opaque identifiers and labels; it is also the direct replication of the Round 49 near misses that Rounds 50 and 51 left open. Track N adds 16 authored raw table claims (shipments, grant expenses, inverter output, rain gauges) with thresholds within 2 percent of the correct total and declared missing rows. Track S adds 24 real SEC merger agreement exhibits, none used before, each with one termination fee claim.

Structurally, over all 81 states of gold and three verdicts, unanimity has 4 unsafe states, preservation with independent recovery 20 and majority 28, each set strictly inside the next, and every unsafe recovery state needs two reads that share the same wrong decisive verdict.

Live, the replication holds exactly: Witness D's 36 verdicts match Round 49 on 36 of 36, including a wrong FAIL on all 9 near miss items, where nothing in the packet proves or excludes the missing qualifying record. A second call from the same reader changed nothing: two resamples at a nonzero temperature and a self validation gave the same wrong FAIL on 9 of 9, and self validation changed none of 75 reads. At two calls, routing Witnesses C and D released 6 unsafe verdicts against 5 for Witness C routed with its own resample, so a second reader did not beat a second call from the better of the two. Witness C as validator of Witness D's read changed 2 of 7 wrong decisive reads to HOLD and turned one correct HOLD into PASS, against no change at all for self validation. Recovery released 5 unsafe verdicts against 0 for three reader unanimity, because a wrong baseline that the validator kept was released whatever the adjudicator said. The largest single gain came from one call: Witness D with room to argue each verdict moved 5 near miss FAILs to HOLD and released 2 unsafe verdicts on 69 usable items (9 if every unusable output had been unsafe). Witness B alone released none. Tracks N and S produced no decisive error from any reader, so all separation comes from track R.

1. Archive check and continuity

The staging registry was read before any live call. Its newest entry is Round 51, XFR-51, so this is Round 52. The audit covered all 45 accessible entries (Rounds 3 and 8 to 51); Rounds 1, 2 and 4 to 7 are not in the accessible archive, so this is not a certification about unavailable history. Round 35 compared homogeneous and

heterogeneous quorums at matched compute on truncated evidence; Round 40 compared composition topologies without fixing the call count; Round 42 compared one witness in two roles with two witnesses in one role at temperature zero. No accessible round compared a second reader with the same reader resampled at the same call count, or with one call given room to argue. All 24 SEC accessions and filers are new against the recorded prior set through Round 51. The continuity audit is `anti-repetition-verification.md`.

Round 51 handed off five items. This round takes item 1 (matched call baselines on items hard enough that single readers err), item 2 (SEC HOLD claims worded so that HOLD is the only defensible reading: the claim is that the fee has been paid, a world fact, not that the window establishes something) and item 5 (the direct Round 49 near miss replication). Items 3 (a dissent rule with executable counterexamples) and 4 (seeded wrong ledgers to measure validator anchoring) are not attempted and remain open.

2. Fixed plan and hypotheses

The plan was fixed before any live call (`run-plan-r52.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.

ID	Fixed statement
H52.1	Structural: the unsafe states of three reader unanimity are a strict subset of those of recovery, which are a strict subset of those of majority, and every unsafe recovery state has two reads sharing the same wrong decisive verdict
H52.2	Replication: with the identical Round 49 prompt and packets, Witness D's single read is decisive on at least 5 of the 9 Round 49 near miss items
H52.3	At two calls, routing Witnesses C and D has fewer unsafe outcomes than routing either reader with its own resample, pooled over all items
H52.4	Routing C and D does not dominate routing the better of the two readers (by correct count) with its own resample: it pays at least as many false holds
H52.5	One Witness D call with twice the output allowance and a counter case step has no more unsafe outcomes than Witness D's two call self routing
H52.6	At three calls, recovery (D read, C validation, B adjudicator) has no more unsafe outcomes than three reader unanimity and fewer false holds
H52.7	Cross validation changes more of Witness D's wrong decisive reads to gold or HOLD than self validation
H52.8	Manipulation check: each of Witnesses B, C and D makes at least one decisive error on tracks R or N, and at least one reader is decisive on at least one SEC payment claim

Related work

Sampling several reasoning paths from one model and voting improves accuracy on many tasks [S103], and accuracy can keep rising with the number of sampled agents [S104], although more calls can also lower accuracy on hard queries [S105]. Debate among several model instances improves some factual and reasoning tasks [S106], but comparisons of debate strategies [S109] and of single agents with strong prompts against multi agent discussion [S110] show that much of the gain can be had without a second agent. Spending inference compute adaptively can beat larger models [S108], which makes the unit of comparison matter. Evaluations of agents that ignore cost overstate the value of complex designs, and cost controlled baselines change rankings [S107]. Self correction without external feedback often fails to improve answers [S102]. This round applies these cautions to the propositions of Rounds 40 to 51 by fixing the call count and testing the same reader as the baseline.

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 (36 of 36). The nine families state one requirement with a universal part and an aggregate threshold: five contract families (joint venture funding confirmations, pipeline right of way grants, noteholder consents, lender estoppels, offtaker acknowledgments) and four quality release record families (aircraft inspections, food lot tests, calibration certificates, bridge load tests). Each family has four variants: P (two qualifying records meet the threshold, PASS), S (the second record is a near miss that fails the universal part, 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 Round 49 blind review disputed M02C and M03C and flagged the lender definition in M04; the sensitivity subset excludes those six items.

Track N (16 items, raw tables). Each family is a 20 row table that requires a filtered sum with a unit conversion and an exclusion rule. P and F claims are about the rows shown and sit between the correct total and the total a reader gets by ignoring the exclusion rule, so that one misreading flips each decisive verdict. HU and HO claims are about the full record when two rows are declared missing with unknown, non negative values; the shown rows alone make HU look false and HO look true, and two concrete completions make each claim true in one world and false in the other.

Family	Table	Correct total	Total if the exclusion is ignored	P threshold (below)	F threshold (exceeds)	H threshold
N1	Warehouse outbound shipments (kg and lb, cancelled excluded)	5101.6 kg	5235.6 kg	5141.8	5195.4	5127.1
N2	Research grant expenses (USD and EUR, disallowed excluded)	5861.3 USD	6009.3 USD	5905.7	5964.9	5890.6
N3	Solar plant inverter output (kWh and MWh, estimated excluded)	7405.0 kWh	7578.0 kWh	7456.9	7526.1	7442.0
N4	Rain gauge network (mm and inches, faulty excluded)	3970.6 mm	4065.6 mm	3999.1	4037.1	3990.5

Track S (24 items, real exhibits). Candidates came from EDGAR full text search for termination fee definitions in merger agreements filed as EX-2.1 from 2023 to 2026, fetched with a declared User-Agent, normalized for whitespace only, and excluded when the accession or filer appears in any prior round. A definition was eligible when it starts a defined term, states one dollar amount and has no proviso. Twenty four were drawn with seed 52. Each window ends with the definition sentence. Claims state the amount (8, PASS), the amount with two adjacent digits transposed (8, FAIL; where no transposition keeps a nonzero leading digit, the leading digit is raised by one), or that the fee has been paid (8, HOLD).

Item	Filer	Exhibit	Filing date	Accession	Fee	Claim type	Source
S01	ATLAS TECHNICAL CONSULTANTS, INC.	EX-2.1	2023-02-01	0001213900-23-006542	Parent Termination Fee	fee paid	exhibit
S02	Diversey Holdings, Ltd.	EX-2.1	2023-03-08	0001104659-23-030166	Company Termination Fee	transposed amount	exhibit
S03	MOMENTIVE GLOBAL INC.	EX-2.1	2023-03-14	0001193125-23-069366	Company Termination Fee	stated amount	exhibit
S04	CVENT HOLDING CORP.	EX-2.1	2023-03-14	0001193125-23-070287	Company Termination Fee	fee paid	exhibit
S05	Embark Technology, Inc.	EX-2.1	2023-05-25	0001104659-23-064691	Company Termination Fee	fee paid	exhibit
S06	Ellington Financial Inc.	EX-2.1	2023-07-03	0001104659-23-077395	Company Termination Fee	transposed amount	exhibit
S07	Heritage-Crystal Clean, Inc.	EX-2.1	2023-07-20	0001193125-23-190211	Parent Termination Fee	stated amount	exhibit
S08	AVANTAX, INC.	EX-2.1	2023-09-11	0001387131-23-011013	Company Termination Fee	fee paid	exhibit
S09	BSQUARE CORP /WA	EX-2.1	2023-10-11	0001437749-23-028060	Company Termination Fee	transposed amount	exhibit
S10	Chord Energy Corp	EX-2.1	2024-02-26	0001193125-24-046215	Parent Termination Fee	transposed amount	exhibit
S11	Apartment Income REIT Corp.	EX-2.1	2024-04-09	0001193125-24-089819	Company Termination Fee	stated amount	exhibit
S12	ENCORE WIRE CORP	EX-2.1	2024-04-15	0001193125-24-095443	Company Termination Fee	stated amount	exhibit
S13	DecisionPoint Systems, Inc.	EX-2.1	2024-05-01	0001213900-24-038523	Parent Termination Fee	transposed amount	exhibit
S14	Squarespace, Inc.	EX-2.1	2024-05-13	0001496963-24-000038	Company Termination Fee	stated amount	exhibit
S15	Aaron's Company, Inc.	EX-2.1	2024-06-17	0001193125-24-162017	Company Termination Fee	fee paid	exhibit
S16	INSTRUCTURE HOLDINGS, INC.	EX-2.1	2024-07-25	0001193125-24-184954	Company Termination Fee	stated amount	exhibit

Item	Filer	Exhibit	Filing date	Accession	Fee	Claim type	Source
S17	PetIQ, Inc.	EX-2.1	2024-08-07	0001104659-24-086551	Company Termination Fee	transposed amount	exhibit
S18	Hall of Fame Resort & Entertainment Co	EX-2.1	2025-05-08	0001213900-25-041134	Company Termination Fee	fee paid	exhibit
S19	E2open Parent Holdings, Inc.	EX-2.1	2025-05-27	0001193125-25-126316	Company Termination Fee	fee paid	exhibit
S20	Couchbase, Inc.	EX-2.1	2025-06-20	0001193125-25-143735	Parent Termination Fee	transposed amount	exhibit
S21	RAYONIER INC	EX-2.1	2025-10-14	0001193125-25-237976	Company Termination Fee	fee paid	exhibit
S22	Eventbrite, Inc.	EX-2.1	2025-12-03	0001193125-25-305559	Company Termination Fee	stated amount	exhibit
S23	SILICON LABORATORIES INC.	EX-2.1	2026-02-04	0001193125-26-036712	Company Termination Fee	transposed amount	exhibit
S24	SAFETY INSURANCE GROUP INC	EX-2.1	2026-07-24	0001104659-26-086488	Company Termination Fee	stated amount	exhibit

4. Calls and arms at matched call counts

Every reader call uses the Round 49 reviewer prompt, so the Witness D read is a direct replication of the Round 49 no memory arm, and the procedure card arm reuses the nine Round 49 cards verbatim. Witnesses B, C and D each read every item once at temperature zero. Witnesses C and D each read every item twice more at temperature 0.7. Witness D made one deliberate call per item with twice the output allowance and an instruction to state the strongest case for PASS, FAIL and HOLD before deciding. Witness C validated each Witness D read, and Witness D validated its own read, shown in the same words as another reviewer's verdict. Witness A audited the labels of tracks N and S blind (40 of 40 agree, no flags).

Arms are replayed from these recorded calls. Routing releases a decisive verdict only when every routed read gives it, otherwise HOLD. Majority releases the verdict that more than half of three reads give. Recovery keeps the baseline when the validator keeps it, and lets a changed verdict through only when the independent adjudicator gives the same verdict. Calls are matched by count only: reader capability and resource per call differ between Witnesses B, C and D, so the same reader repeated is the fair baseline for any second call.

5. Structural results (E52-A)

Over all 81 combinations of gold and three verdicts, unanimity is unsafe in 4 states (all three reads share the same wrong decisive verdict), recovery in 20 and majority in 28; unanimity's states are a strict subset of recovery's and recovery's of majority's. Every unsafe recovery state has two reads that share the wrong decisive verdict: either the validator kept a wrong baseline, or it changed the verdict and the adjudicator agreed with the wrong change. Recovery therefore cannot be safer than unanimity, and is safer than majority only in the 8 states where the baseline and the adjudicator agree on a wrong verdict and the validator dissents. Two reader routing is unsafe in 12 of the 81 states and pays false holds in 42.

All 16 table claims recompute to their gold; all 8 P and F claims flip under the exclusion misreading; all 8 HU and HO claims have one true and one false completion. All 24 SEC windows contain their definition sentence and stated amount, and no transposed amount appears in its window. All track R prompts are identical to Round 49.

6. The Round 49 near misses, replicated (E52-B)

Witness D's 36 single read verdicts match its Round 49 verdicts on 36 of 36. On the 9 near miss items it again answered FAIL on 9 of 9. With the Round 49 procedure card it answered HOLD on 4 of 9 (Round 49: 5 of 9), and 30 of its 36 card verdicts match Round 49; two card outputs were empty. Witness C's single reads match Round 49 on 34 of 36 (two outputs incomplete) and were FAIL on 6 of the 7 usable near miss items. Witness B answered HOLD on 9 of 9.

Call	Near miss FAIL (wrong)	Near miss HOLD (correct)	Unusable
Witness D read	9	0	0
Witness D resample 1	9	0	0
Witness D resample 2	9	0	0
Witness D validates its own read	9	0	0
Witness C read	6	1	2
Witness C validates Witness D read	5	2	2
Witness D with procedure card	5	4	0
Witness D, room to argue each verdict	2	5	2
Witness B read	0	9	0

7. Matched call comparisons (E52-B and E52-C)

All 76 items, with the sensitivity subset (70 items without the six Round 49 flagged items) in the last column:

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe, sensitivity subset
Witness B read	1	73 of 75	0	2	1	0
Witness C read	1	67 of 74	6	1	2	5
Witness D read	1	66 of 76	9	1	0	8
Witness D, one call with room to argue each verdict	1	67 of 69	2	0	7	2
Witness D with the Round 49 procedure card (track R only)	1	28 of 34	5	1	2	4
Route Witness D with its own resample	2	66 of 76	9	1	0	8
Route Witness C with its own resample	2	68 of 74	5	1	2	4

Arm	Calls	Correct	Unsafe	False holds	Unusable	Unsafe, sensitivity subset
Route Witnesses C and D	2	67 of 74	6	1	2	5
Route Witnesses B and D	2	73 of 75	0	2	1	0
Route Witnesses B and C	2	71 of 73	0	2	3	0
Witness D validates its own read	2	65 of 75	9	1	1	8
Witness C validates Witness D read	2	67 of 74	6	1	2	5
Witness D three reads, majority	3	64 of 74	9	1	2	8
Witness D three reads, all agree	3	64 of 74	9	1	2	8
Witness C three reads, majority	3	66 of 72	5	1	4	5
Witness C three reads, all agree	3	67 of 72	4	1	4	4
Witnesses B, C and D, majority	3	66 of 73	6	1	3	5
Recovery: D read, C validation, B adjudicator	3	67 of 73	5	1	3	4
Witnesses B, C and D, all agree	3	71 of 73	0	2	3	0

By track, every single read and every arm was correct on all 16 table claims and all 24 SEC claims, apart from unusable outputs and one validation (section 8). No reader was decisive on any SEC payment claim and no reader made a decisive error on the near threshold tables, so every unsafe verdict in the table above falls on track R except that one validation, and every unsafe single read by Witnesses C and D falls on a near miss item.

Paired on items where both arms are usable, routing C and D was worse than routing C with its own resample on one item (RN01S, where Witness C's resample said HOLD) and better than routing D with its own resample on one item (RN04S, where Witness C's read said HOLD); the two near miss items where Witness C's read was unusable drop out of both comparisons. Witness D's deliberate call was better than Witness D's two call self routing on 5 items and worse on none (exact sign test, $p = 0.0625$, descriptive). Recovery was worse than unanimity on 4 items and better on none ($p = 0.125$) and identical to majority on every usable item. The joint same wrong decisive counts explain these results: Witness D and its own resample shared a wrong verdict on 9 items, Witnesses C and D on 6, Witness C and its own resample on 5, and Witness B with either C or D on none.

8. Validation: a second reader versus the same reader (E52-C)

Witness D validating its own read changed none of 75 usable reads, including all 9 wrong near miss FAILs. Witness C validating Witness D's read changed 3 of 76: two wrong FAILs became HOLD (RM01S and RN04S),

and one correct HOLD became PASS (N4HU). On N4HU the validator returned verdict PASS with changed false, while its rationale says that the two missing readings could raise the total above the threshold, so the packet neither establishes nor refutes the claim. The recorded verdict is scored. On the other five near miss items where Witness D was wrong, Witness C kept the wrong FAIL as validator; in its own blind read it had given the same FAIL on four of them, so validation added a different reader but not a different reading.

9. One near miss, walked through

Packet P16 (Round 49 item M03S) requires consents from holders of at least 66 percent of principal, each delivered before the Expiration Time, 5:00 p.m. on August 24. Ashgrove (41 percent) consented on August 19; Benton (35 percent) consented at 6:40 p.m. on August 24. The question asks about the Expiration Time "as finally in effect". Nothing in the packet establishes or excludes an extension, so the open record label is HOLD.

Witness D answered FAIL: "Benton's 35 percent consent was received at 6:40 p.m. on August 24, 2026, after the 5:00 p.m. Expiration Time, so it does not qualify." Its two resamples, its self validation, its procedure card read, Witness C's read, both of Witness C's resamples and Witness C's validation all answered FAIL. Every FAIL quoted exact packet text. Two calls answered HOLD: Witness B ("The packet establishes the stated Expiration Time but does not establish whether it was later extended") and Witness D's one deliberate call, whose case for HOLD was "The packet does not establish or exclude an extension of the Expiration Time, and an extension past 6:40 p.m. would make Benton's consent qualifying and change the result." Agreement among nine calls did not make the FAIL right, and routing any two of them released it.

10. Findings versus interpretations

Structural findings

• Unanimity, recovery and majority have 4, 20 and 28 unsafe states of 81, nested strictly; every unsafe recovery state needs two reads sharing the wrong verdict.

• All track R prompts are identical to Round 49; all table and SEC gold certificates pass.

Empirical findings

• Witness D reproduced its Round 49 verdicts on 36 of 36 items, including FAIL on 9 of 9 near misses.

• A second or third call from Witness D (resampling or self validation) changed no near miss verdict, and self validation changed none of 75 reads.

• At two calls, routing C and D released 6 unsafe verdicts, routing C with its own resample 5 and routing D with its own resample 9.

• Witness C as validator moved 2 of 7 wrong Witness D reads to HOLD and damaged one correct HOLD.

• Recovery released 5 unsafe verdicts against 0 for three reader unanimity, with one false hold against two.

• One Witness D call with room to argue each verdict released 2 unsafe verdicts on 69 usable items, against 9 for its two call self routing; 7 of its outputs were empty, two of them on near miss items.

• Witness B alone released no unsafe verdict, with 2 false holds; no reader made a decisive error on tracks N or S.

Interpretation

On these items the second call was worth what the second reader brought, and only when that reader read differently: a resample or a self check from the same reader reproduced its error, and adding a reader that shared the same error (Witnesses C and D) bought less than resampling the better of the two. Matched call baselines

therefore change how the surviving propositions should be stated. Agreement routing is a property of the joint error between reads, not of the number of reads; validation by another reader is weakly useful and can damage a correct read; recovery inherits any wrong baseline that its validator keeps. The strongest one call arm (room to argue each verdict) moved more near miss errors than any two or three call arm built from the same reader, which suggests that what helped was a different reading procedure, not a different reader or more calls. Because tracks N and S were at ceiling, none of this is shown outside the Round 49 near miss structure, and because readers differ in capability and resource per call, no ranking across witnesses is a cost comparison.

11. Hypothesis outcomes

ID	Outcome	Basis
H52.1	Supported	Unsafe states 4, 20 and 28 of 81, strictly nested; every unsafe recovery state has two reads sharing the wrong verdict
H52.2	Supported	Witness D decisive (FAIL) on 9 of 9 near miss items, matching Round 49 on 36 of 36
H52.3	Rejected	Routing C and D released 6 unsafe verdicts, routing C with its own resample 5; routing D with its own resample released 9 (7 on the 74 items usable for both)
H52.4	Supported	Routing C and D does not dominate routing C with its own resample (6 against 5 unsafe, 1 against 1 false hold)
H52.5	Supported	2 unsafe on 69 usable items against 9; at most 9 if every unusable output were unsafe
H52.6	Rejected	Recovery released 5 unsafe verdicts against 0 for unanimity, with 1 false hold against 2
H52.7	Supported	Cross validation moved 2 of 7 wrong decisive reads to HOLD; self validation moved 0 of 9; cross validation also damaged one correct HOLD
H52.8	Rejected	Witnesses C and D erred decisively on track R, but Witness B made no decisive error, no reader erred on track N, and no reader was decisive on an SEC payment claim

12. 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: "Joint same wrong CD 6 and DD 9 show C errors nest inside D errors and D resamples reproduce them, so separate calls were not independent; every D only arm equals D alone, and every B containing arm equals B alone minus missing items." It judged H52.2 and H52.4 supported, H52.3 and H52.8 failed and H52.6 failed in its safety half, and described H52.5 and H52.7 as directional only ($p = 0.0625$; two fixes against one damage). Its caveats were: (1) matched call counts are not matched cost, since readers differ, the deliberate arm had twice the output allowance, resamples ran at 0.7 and validation prompts carry a prior read; (2) unusable outputs shrink denominators, and the deliberate arm lost seven track R items; (3) composite arms inherit the missingness of every member call; (4) agreement is not truth, and Witness D matching Round 49 on 36 of 36 shows determinism, not corroboration. Caveat 2 is answered by the bound reported for H52.5. Its remark that track R labels were unaudited is not adopted: they were audited blind in Round 49, where the review agreed with 34 of 36 open labels and the six flagged items form the sensitivity subset, which changes no outcome. Its suggestion to report resource use per arm is noted as a limitation; this record matches call counts and does not publish resource

use.

13. Completion disclosure and limitations

Every planned experimental call was attempted once: one blind label audit, 228 single reads, 304 resampled reads, 76 deliberate reads, 36 procedure card reads and 152 validations, 797 in all, plus one synthesis call. Twenty one outputs were unusable: one error (read_B on S23), twelve empty (two resamples, seven deliberate calls, two card calls, one self validation) and eight whose JSON ended before the object closed (two Witness C reads, two Witness C validations, four Witness C resamples). None was repeated. The remaining phases were reordered once, during the second Witness D resample, so that the validations ran before the last Witness C resample; this was decided before any of their outputs existed. No call was stopped by the spend guard. No packet, claim or label was edited after outcomes were observed. The deviation record is `deviations-r52.md`.

Limitations: all separation comes from 36 replicated items and mostly from 9 near miss items, so differences of one to five items are descriptive; tracks N and S were at ceiling despite near threshold margins, missing rows and real exhibits, so the round says nothing about matched baselines on those materials except that no arm was needed; reader capability and resource per call are not matched across witnesses; the deliberate arm also had twice the output allowance and more empty outputs; temperature 0.7 is one setting; track R labels come from Round 49; results are about these readers, prompts and items, not about any reader class.

14. Next checks for the theory program

1. Round 53 should state candidate propositions in domain general language that already include the matched call result: a second read helps in proportion to how differently it reads, not because it is a second call.

2. Test a different reading procedure (room to argue each verdict) against a different reader at matched calls and matched output allowance, on held out near miss structures outside Round 49.

3. Build non-contract items that produce single reader errors; near threshold arithmetic and declared missing rows did not.

4. The Round 51 dissent rule with executable counterexamples and the seeded validator anchoring test remain open.

15. Reproducibility

The archive includes the fixed plan, continuity audit, deviation record, the Round 49 packet builder and the Round 49 reference calls used for the identity checks, the procedure cards, the material builder with all tables, completions and certificates, the SEC track with exact windows, accessions and links, source and bibliography verification, the live runner with every prompt, verbatim transcripts of all 798 calls, the blind audit and synthesis, replay analysis, the exact verifier, the chart script and this manuscript with its PDF builder. The verifier makes no network calls and checks 395 conditions. The private execution configuration is not included and is not required to replay the recorded evidence.

16. 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

• [S104] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu and Deheng Ye. More Agents Is All You Need. arXiv:2402.05120, first posted February 3, 2024. https://arxiv.org/abs/2402.05120

• [S105] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia and James Zou. Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems. arXiv:2403.02419, first posted March 4, 2024. https://arxiv.org/abs/2403.02419

• [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

• [S107] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir and Arvind Narayanan. AI Agents That Matter. arXiv:2407.01502, first posted July 1, 2024. https://arxiv.org/abs/2407.01502

• [S108] Charlie Snell, Jaehoon Lee, Kelvin Xu and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314, first posted August 6, 2024. https://arxiv.org/abs/2408.03314

• [S109] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett and Arnu Pretorius. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. arXiv:2311.17371, first posted November 29, 2023. https://arxiv.org/abs/2311.17371

• [S110] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong and Yangqiu Song. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key?. arXiv:2402.18272, first posted February 28, 2024. https://arxiv.org/abs/2402.18272

• [S30] U.S. Securities and Exchange Commission. EDGAR Archives. https://www.sec.gov/Archives/edgar/data/. The 24 exhibits in section 3 were opened and verified for this run.