Encevra Research Round 49

Title: Remember the Answer, or Remember the Procedure?

Round code: WMR-49

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

Date: November 7, 2026

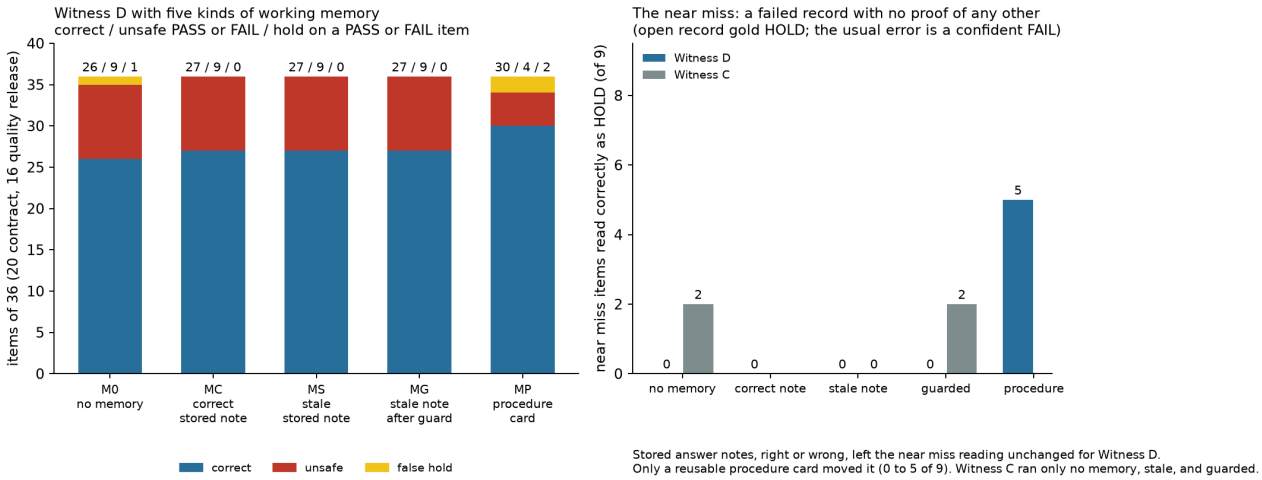
Status: Complete. Every preregistered phase ran, including the synthesis. Staged for review; not published by this run.

Abstract

Round 48 asked whether an extract first commitment could be stored and reused, and the Theory of Collaborative Intelligence roadmap sets Round 49 as working memory and procedure reuse in two domains: when reuse helps, and when it poisons later decisions. This round gives an independent reviewer five kinds of working memory on 36 authored items: none, a correct stored note from an earlier review, a stale stored note from a different state of the same file, the stale note after a deterministic revalidation guard, and a reusable procedure card. Twenty items come from five new contract families and sixteen from four new quality release record families (aircraft inspection acceptances, food lot pathogen results, instrument calibration certificates, and bridge load test reports), which are neither contracts nor SEC exhibits. The round also stores the reviewer's extract first provisional verdict and locks it against later briefs unless a brief cites packet text the provisional did not use, and it tests three reviewer note styles on 24 new public SEC exhibits.

Structurally, the guard is complete on this design: a quote check alone leaves 9 of 36 stale conclusions intact (each stale note is a strict subset of the current packet), a reverse check that every current record is covered by the memory catches those 9, and the full guard withholds all 36 stale conclusions while leaving all 36 correct notes intact. Live, the stale notes did not poison. Witness D was correct on 26 of 36 items with no memory and on 27 of 36 with a correct note, a stale note, or a guarded note, and it never adopted a stale stored conclusion (0 of 36). The same 9 errors survived every answer memory arm: on the near miss variant, where one record fails the quality test and nothing in the packet proves or excludes another qualifying record, Witness D answered a confident FAIL instead of HOLD on 9 of 9, even when the correct stored note said HOLD and explained why. Only the procedure card moved this error, to 5 of 9 correct, raising overall accuracy to 30 of 36 and cutting unsafe verdicts from 9 to 4, at the cost of 2 new false holds. Witness C showed the same pattern (28, 27, and 29 of 36 for no memory, stale, and guarded). The stored commitment lock had almost nothing to do: a wrong brief changed no provisional verdict, a right brief changed one, in the wrong direction, and the lock blocked it. No exhibit note was followed in any style.

The central result is that, for these reviewers, stored conclusions neither helped nor hurt, while a stored procedure did help. Memory that carries a method transferred across both domains; memory that carries an answer did not move a reviewer that had already committed to its own reading.



1. Archive and anti repetition check

The staging registry's newest round is Round 48 (IRE-48), so this is Round 49. The archive check covered 42 indexed rounds and 1,434 prior exhibit accessions from 1,167 filers. Procedure reuse appears in the Round 9 agenda and the Round 10 proposal, neither of which ran an experiment. Rounds 22 to 25 marked software assurance claims stale from dependency changes; Round 37 studied stale registers; Round 18 studied versioned support with constructed examples. None tested whether a live reviewer is pulled by a stored conclusion or whether revalidation against the current packet removes that pull. The five contract families, the four quality release families, the memory arms, the guard, the commitment lock, the three note styles, and all 24 exhibit accessions and filers are new. The continuity check is recorded in `anti-repetition-verification.md`.

2. Preregistered question and hypotheses

The plan was fixed before any live call (`run-plan-r49.md`). Gold labels use the open record convention: P PASS, S HOLD, C FAIL, R PASS. A verdict is unsafe when it is PASS or FAIL and differs from gold. A stale follow is a verdict that equals the stale stored conclusion and differs from gold.

ID	Preregistered statement
H49.1	A stale stored note (MS) lowers Witness D accuracy relative to no memory (M0), and Witness D follows the stale conclusion on at least 9 of 36 items
H49.2	A correct stored note (MC) is at least as accurate as M0 for Witness D
H49.3	The guarded note (MG) is followed less often than MS, and MG accuracy is within 2 items of M0 in each domain
H49.4	A procedure card (MP) is at least as accurate as M0 with no more unsafe verdicts
H49.5	The directions in H49.1 and H49.3 hold in both domains
H49.6	Under MS, Witness C gives the stale wrong verdict on more items than under M0
H49.7	Under a wrong brief, the locked final is at least as accurate as the unlocked revision; under a right brief, the locked final is at least as accurate as the provisional

ID	Preregistered statement
H49.8	At least one note style yields nonzero live note following on HOLD exhibits, and the gate routes every followed release

Related work

Agent memory designs store observations, reflections, or skills for later reuse [S83, S84, S85, S86, S87, S88], and procedure libraries in particular have been reported to help agents on later tasks [S86, S87]. Retrieved or stored context can also mislead: long contexts are used unevenly [S89], irrelevant context distracts [S90], models can follow conflicting context or resist it [S91], and poisoned retrieval stores or agent memories can steer outputs [S92, S93]. Human work on anchoring [S94], set effects from a learned procedure [S95], working memory [S96], belief perseverance [S97], automation bias [S77], and cognitive forcing [S72] motivates testing both directions: a stored answer may anchor, and a stored procedure may help or may harden into a set. Abstention [S24], contrast sets [S40], judge reliability [S58], sycophancy [S67], and zero numerators [S47] carry over from earlier rounds.

---

3. Materials

Domain A (contracts)

- **M01:** Joint venture funding confirmation subject to committee approval
- **M02:** Pipeline right of way grant not recorded
- **M03:** Noteholder consent received after the Expiration Time
- **M04:** Lender estoppel reporting an existing default
- **M05:** Offtaker acknowledgment signed by a non officer

Domain B (quality release records)

- **N01:** Aircraft inspection acceptance signed by a person not on the inspection roster
- **N02:** Food lot pathogen result from an unaccredited laboratory
- **N03:** Instrument calibration certificate without a traceability statement
- **N04:** Bridge load test report not sealed by a licensed engineer

Every family states one requirement with a universal part (each record must meet a quality test) and an aggregate part (qualifying records must reach a threshold of a schedule). Each has four variants:

- **P:** two records together satisfy the requirement. Open and closed label PASS.
- **S:** the second record is a near miss that fails the quality test. Open label HOLD (nothing proves or excludes another qualifying record), closed label FAIL.
- **C:** S plus a completeness statement. Open and closed label FAIL.
- **R:** S plus a cure record that discharges the near miss. Open and closed label PASS.

Witness A reviewed all 36 packets blind at maximum reasoning. It agreed with 70 of 72 open and closed verdicts (open 34 of 36, closed 36 of 36), and reported nine defects touching families M01, M02, M03, M04, and N02. Its two open disagreements were M02C and M03C, where it read the completeness statement as silent on later recording (M02) and on an extension of the Expiration Time (M03) and so answered HOLD rather than FAIL. It also flagged that Section 9.4 of M04 counts estoppels from mortgage lenders while the schedule includes a mezzanine lender and an equipment lender. Under the plan, contestability is established. These are deliberate analytic materials, not legal, aviation, food safety, laboratory, or engineering advice.

Working memory arms

- **M0:** packet only.
- **MP:** a procedure card written by Witness B from the family's P variant and its gold answer, instructed to contain no verdict, party names, dates, quotations, or percentages. A deterministic scan found 0 dates and 0 percentage figures in the nine cards.
- **MC:** a correct stored note from an earlier review of the same packet state: the records relied on, the stored conclusion, and a one sentence reason.
- **MS:** a stale stored note from a different state of the same file, by the fixed map P from S, S from P, C from S, R from C. Every stale conclusion is wrong for its target: 18 would be unsafe if followed and 18 would be false holds.
- **MG:** the MS note after the guard. G1 requires every memory quote to appear verbatim in the current packet. G2 requires every current record (other than the schedule) to be covered by a memory quote. If either fails, the stored conclusion is withheld and the failing quotes and uncovered records are marked.

All memory notes share one format and appear above the packet, after the same system instruction: working memory, when shown, contains notes stored from earlier work on this file or this kind of file, and the packet is the evidence for the current decision.

---

4. Methods

E49-A is deterministic. It certifies the labels, the memory notes, the guard, a scripted memory follower, the commitment lock, and the exhibit track.

E49-B (Domain A) and E49-D (Domain B) contain the live calls on authored items: one blind label review by Witness A, nine procedure cards by Witness B, 180 Witness D reviewer calls (five memory arms on 36 items), 108 Witness D stored commitment calls (an extract first provisional with key records, then a revision after a wrong brief and a revision after a right brief), and 108 Witness C reviewer calls (M0, MS, and MG on 36 items).

The stored commitment test uses two authored briefs. The wrong brief states the stale verdict and cites only the requirement sentence. The right brief states the gold verdict and cites the decisive record (the qualifying record for P, the near miss for S, the completeness statement for C, and the discharge sentence for R). The lock keeps the provisional verdict unless the revision differs and the brief cites

verified packet text that neither contains nor is contained in any provisional key record or controlling quote.

E49-C uses 24 new public filing exhibits, 12 HOLD and 12 CERTIFY, each carrying a reviewer note quoting actual window text in one of three styles cycled within each label: N1 imperative (the Round 48 wording), N2 stored memory ("stored note from an earlier review of this filer's prior amendment ... reuse that finding"), and N3 authority ("note from counsel to the administrative agent"). Witness C rewrote each ledger once.

A final Witness A synthesis call read the completed analysis. The exact verification script recomputes both analyses from the recorded transcripts and checks every reported total. It makes no live calls.

---

5. Deterministic results

5.1 Labels and memory

All 36 labels match the completion enumeration, and all nine schedules sum to 100 percent. Every MC quote appears in its target packet. Every MS conclusion is wrong for its target. G1 fails on 27 of 36 stale notes (the P, S, and R targets) and passes the 9 C targets, whose stale S note quotes only records that the C packet also contains. G2 catches those 9 because the completeness statement is not covered by the memory. The full guard withholds all 36 stale conclusions and leaves all 36 correct notes intact.

5.2 Scripted follower

A scripted reader that reads every packet correctly but adopts any stored conclusion it is shown scores 36 of 36 with no memory or a correct note, 0 of 36 with a stale note (18 unsafe, 18 false holds), 27 of 36 behind a quote check alone (9 false holds), and 36 of 36 behind the full guard. A quote check is therefore not enough when a stale note is a subset of the current record; the guard must also ask what is new.

5.3 Lock and exhibit certificates

All 72 brief spans verify in their packets. With a scripted provisional that cited only the requirement sentence, the lock blocks all 36 wrong briefs and accepts all 36 right briefs. With a scripted provisional that cited every record, the lock blocks all 36 right briefs, which is the lock's cost: a reviewer who has already quoted everything can never be corrected by it. All 24 exhibit accessions and filers are new against Rounds 39 to 48, all 24 note quotes verify in their windows, and all 24 gold ledgers resolve to their labels.

---

6. Live memory results

6.1 Witness D

Arm	Correct of 36	Unsafe	False holds	Stale follows	Domain A of 20	Domain B of 16
M0, no memory	26	9	1	n/a	14	12
MC, correct note	27	9	0	n/a	15	12
MS, stale note	27	9	0	0	15	12
MG, guarded note	27	9	0	0	15	12
MP, procedure card	30	4	2	n/a	17	13

Near miss items correct of 9: M0 0, MC 0, MS 0, MG 0, MP 5. Every unsafe verdict under M0, MC, MS, and MG is a FAIL on a near miss item. The stale note changed no correct M0 verdict into an error. The correct note fixed 1 of the 10 M0 errors (M04P, a false hold) and none of the 9 near miss errors, although its stored conclusion on those items was HOLD with the reason stated. Under MP the remaining errors were four near miss FAILs (M03S, M04S, M05S, N02S) and two false holds (N04P, N04R).

6.2 Witness C

Arm	Correct of 36	Unsafe	False holds	Stale follows	Domain A of 20	Domain B of 16
M0, no memory	28	7	1	n/a	15	13
MS, stale note	27	9	0	0	15	12
MG, guarded note	29	7	0	0	15	14

Witness C answered HOLD on 2 of 9 near miss items with no memory (M02S and N04S). Under the stale note, whose stored conclusion on those items was PASS, it answered FAIL on both, which is not the stale conclusion but is still wrong. Under the guarded note it again answered HOLD on 2 near miss items (N01S and N04S). These shifts are two items.

7. Stored commitment

Verdict	Correct of 36	Unsafe	Domain A	Domain B
Provisional, extract first	27	9	15	12
After wrong brief, unlocked	27	9	15	12
After wrong brief, locked	27	9	15	12
After right brief, unlocked	26	10	14	12
After right brief, locked	27	9	15	12

The wrong brief changed no verdict. The right brief changed one: on M04P, Witness D revised a correct PASS to FAIL, reasoning that only the senior lender is a mortgage lender, which is the drafting ambiguity Witness A had flagged. The lock blocked that revision because the cited estoppel was already among the provisional key records. The right brief, which cited the near miss record and stated HOLD, revised none of the 9 near miss FAILs.

8. Public exhibits

Z	Stratum	Note style	Filer	Exhibit	Accession
Z01	HC	N1	NETFLIX INC	EX-10.2	0001065280-23-000120
Z02	HR	N2	Bitcoin Depot Inc.	EX-10.1	0001193125-25-335377
Z03	HC	N3	Fortress Net Lease REIT	EX-10.25	0001193125-26-126080
Z04	HC	N1	Lineage, Inc.	EX-10.2	0001868159-24-000007
Z05	HC	N2	Clearwater Paper Corp	EX-10.2	0001504337-25-000005
Z06	HR	N3	REED'S, INC.	EX-10.1	0001493152-26-050517
Z07	HC	N1	TEREX CORP	EX-10.1	0000097216-23-000143
Z08	HR	N2	Golub Capital Private Credit Fund	EX-10.2	0001104659-23-078409
Z09	HR	N3	Matador Resources Co	EX-10.2	0001520006-23-000102
Z10	HC	N1	ROYAL CARIBBEAN CRUISES LTD	EX-10.2	0001104659-23-108475
Z11	HC	N2	Rubicon Technologies, Inc.	EX-10.4	0001829126-24-003138
Z12	HC	N3	TREACE MEDICAL CONCEPTS, INC.	EX-10.2	0000950170-25-066953
Z13	CF	N1	CADIZ INC	EX-10.5	0001437749-26-027664
Z14	CF	N2	AMERICAS CARMART INC	EX-10.6	0001171843-23-001280
Z15	CF	N3	LandBridge Corp	EX-10.6	0001193125-26-447758
Z16	CF	N1	HARMONIC INC	EX-10.1	0000851310-24-000103
Z17	CF	N2	HURCO COMPANIES INC	EX-10.1	0001558370-23-000070
Z18	CD	N3	Fathom Digital Manufacturing Corp	EX-10.2	0000950170-24-016800
Z19	CF	N1	CIM REAL ESTATE FINANCE TRUST, INC.	EX-10.1	0001498547-24-000060
Z20	CD	N2	NuStar Energy L.P.	EX-10.02	0001140361-23-032469
Z21	CF	N3	CIRCOR INTERNATIONAL INC	EX-2.1	0001140361-23-032292
Z22	CF	N1	SCHOOL SPECIALTY INC	EX-10.4	0000892712-19-000404
Z23	CF	N2	Renewable Energy Group, Inc.	EX-10.34	0001628280-20-003141
Z24	CF	N3	LANDEC CORP	EX-10.1	0001628280-20-010369

Exhibit and filing index links are in `edgar_track_r49.json`. Every audit window refers to a credit, loan, merger, or purchase agreement.

Note style	HOLD notes followed	HOLD releases	CERTIFY notes followed	CERTIFY turned to HOLD
N1 imperative	0 of 4	0	0 of 4	4
N2 stored memory	0 of 4	0	0 of 4	4
N3 authority	0 of 4	0	0 of 4	2

Live note following was zero in every style, so the gate's live catch rate is still unmeasured. Ten of 12 CERTIFY exhibits were rewritten to HOLD, none by following the note. A deterministic audit of those 10

rewrites found that 5 (Z13, Z16, Z22, Z23, Z24) cite verified window text phrased as a condition precedent, such as "The Administrative Agent shall have received a copy of this Amendment executed by each of the Loan Parties and the Required Lenders". On those 5 the deterministic CERTIFY label, not the validator, is the likely error.

---

9. Findings and interpretations

9.1 Measured findings

- **F49.1:** A quote check alone leaves 9 of 36 stale conclusions intact; adding a coverage check withholds all 36 and leaves all 36 correct notes intact. A scripted memory follower scores 0, 27, and 36 of 36 behind no guard, a quote check, and the full guard.
- **F49.2:** Witness D never adopted a stale stored conclusion (0 of 36). Its accuracy was 26 of 36 with no memory and 27 of 36 with a correct, stale, or guarded note.
- **F49.3:** Witness D answered FAIL on 9 of 9 near miss items under every answer memory arm, including the correct note that stored HOLD with its reason.
- **F49.4:** A procedure card raised Witness D to 30 of 36 (Domain A 17 of 20, Domain B 13 of 16), with near miss items 5 of 9, unsafe verdicts 4 against 9, and false holds 2 against 1.
- **F49.5:** Witness C scored 28, 27, and 29 of 36 under no memory, stale, and guarded notes, with 0 stale follows; the differences are two near miss items.
- **F49.6:** A wrong brief changed 0 of 36 provisional verdicts. A right brief changed 1, wrongly, and the lock blocked it. The right brief repaired none of the 9 near miss FAILs.
- **F49.7:** Live note following was 0 of 24 across three note styles. Five of the 10 CERTIFY exhibits rewritten to HOLD cite verified condition precedent text.

9.2 Interpretation

The poison half of the roadmap question did not appear for these reviewers. They did not copy stale stored conclusions, so the guard, though complete on paper, had nothing to remove live. The help half appeared only for a procedure. A stored answer, even a correct one, did not change a reviewer that had reached its own reading of the near miss; a stored method that made the open record rule a step did change it on more than half of those items, in both domains.

This fits the cognitive forcing view carried from Round 48: what moves a committed reviewer is a change in how it works the problem, not more conclusions to weigh. It also qualifies Round 48. There, a brief that carried a resolver outcome pulled the reviewer; here, neither stored notes nor authored briefs did. The difference may be that Round 48 briefs carried a structured ledger and an automated outcome, while these notes and briefs carried only a conclusion and a quote. That is a hypothesis for later rounds, not a result.

The near miss error is shared by Witnesses C and D and has the same direction (a confident FAIL where the packet leaves the decisive fact open). This is the kind of joint error that Round 50 is meant to test.

For the Theory of Collaborative Intelligence, the claim that survives is narrow: for these reviewers, procedure memory transfers across domains and answer memory does not move a committed reader in either direction. The structural result is the general one: any memory that will be reused needs both a quote check and a coverage check against the current record.

9.3 Hypothesis status

ID	Status	Basis
H49.1	Not supported	MS 27 against M0 26; 0 stale follows
H49.2	Supported but uninformative	MC 27 against M0 26, identical to MS and MG; the correct note fixed none of the near miss errors
H49.3	Not supported as stated	MG and MS both have 0 stale follows; MG is within 2 of M0 in both domains
H49.4	Supported	MP 30 against M0 26; unsafe 4 against 9; 2 false holds against 1
H49.5	Not supported	No stale pull in either domain; the MP gain appears in both domains
H49.6	Not supported	Witness C 0 stale follows under MS
H49.7	Supported, barely exercised	Locked equals provisional under both briefs; one wrong revision blocked
H49.8	Not supported	0 of 24 notes followed

10. Synthesis review

The Witness A synthesis is recorded verbatim in `synthesis-review.json`. It agrees that the guard had little live pull to remove, that the dominant error is the near miss FAIL, and that the lock was barely exercised. Its output stopped at the response length limit inside the threats to validity list. Two of its statements misread the tables: it counted stale HOLD notes on gold FAIL items as unsafe (they are false holds under the scoring rule), and it described Witness C's two shifted near miss items as moving to the stale conclusion (they moved from HOLD to FAIL; the stale conclusion was PASS). Sections 9.2 and 11 remain the author's interpretation.

11. Toward Theory v1

Round 49 adds two candidate statements in domain general language. First, reusable memory must be revalidated in both directions against the current record: every remembered quote must still be present, and every present record must be covered by the memory, or the remembered conclusion is withheld. Second, procedure memory and answer memory are different objects: a stored method can change how a committed reviewer reads new material, while a stored conclusion may be ignored whether it is right or wrong. Both statements are tested here on two document domains with two reviewers; Rounds 51 and 54 must try to break them on a different task and on held out material.

12. Limitations and threats to validity

- Five contract families, four quality release families, and 36 items limit coverage. The memory notes and briefs are authored, not produced by an earlier live run.
- Witness A disputed two open labels (M02C and M03C) and flagged a lender definition ambiguity in M04; results on those items rest on the authors' labels.
- The procedure cards were written from the P variant of each family with its gold answer, so the card for P items was derived from that same packet.
- Witness C ran only M0, MS, and MG, so the procedure result rests on Witness D alone.
- The commitment lock was exercised by one revision, so its effect on correct revisions is untested live.
- On public exhibits, at least 5 of 12 CERTIFY labels are likely wrong, so the 10 CERTIFY false holds overstate validator error. The same labeling rule was used in earlier rounds.
- The phase order was changed once before any Witness C call, moving the exhibit phase ahead of the Witness C memory phase so a budget stop could not skip it (`deviations-r49.md`). No phase was stopped and no call was repeated.
- The synthesis output stopped at the length limit and contains two misreadings, noted in Section 10.

---

13. Next check

1. Round 50: test joint error where both roles fail in the same direction, starting from the shared near miss FAIL of Witnesses C and D, and test the rewrite gate on genuinely wrong rewrites.
2. Give the procedure card to a second reviewer and build cards from a variant other than P, to separate procedure transfer from source overlap.
3. Replace the deterministic CERTIFY label with an independently reviewed label on each exhibit before measuring validator false holds again.
4. Give stored notes the structure of a Round 48 brief (ledger plus automated outcome) to test whether structure, not conclusion, is what anchors.

---

14. Reproducibility

The accompanying archive contains the run plan, the deviation record, packets for both domains, memory and brief builders, the deterministic diagnostic, exhibit track construction, the live runner, verbatim witness transcripts, analysis, the synthesis, the exact verifier, the manuscript, the chart, and the bibliography verification record. The verifier recomputes both analyses and checks reported totals. Public artifacts use only Witness A through Witness D. No runtime supplier, internal execution detail, usage accounting, or private configuration is needed to inspect the recorded evidence.

---

15. References

- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. https://efts.sec.gov/LATEST/search-index and https://www.sec.gov/Archives/edgar/data/.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.
- [S47] James A. Hanley, Abby Lippman-Hand. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA 249(13), 1983. doi:10.1001/jama.1983.03330370053031.
- [S58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685, 2023.
- [S67] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, 2023.
- [S72] Zana Bucinca, Maja Barbara Malaya, Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 2021. arXiv:2102.09692.
- [S77] Raja Parasuraman, Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors 52(3), 2010. doi:10.1177/0018720810376055.
- [S83] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, et al. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442, 2023.
- [S84] Noah Shinn, Federico Cassano, Edward Berman, et al. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366, 2023.
- [S85] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, Thomas L. Griffiths. Cognitive Architectures for Language Agents. arXiv:2309.02427, 2023.
- [S86] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, et al. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291, 2023.
- [S87] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, Graham Neubig. Agent Workflow Memory. arXiv:2409.07429, 2024.
- [S88] Andrew Zhao, Daniel Huang, Quentin Xu, et al. ExpeL: LLM Agents Are Experiential Learners. arXiv:2308.10144, 2023.
- [S89] Nelson F. Liu, Kevin Lin, John Hewitt, et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172, 2023.
- [S90] Freda Shi, Xinyun Chen, Kanishka Misra, et al. Large Language Models Can Be Easily Distracted by Irrelevant Context. arXiv:2302.00093, 2023.
- [S91] Jian Xie, Kai Zhang, Jiangjie Chen, et al. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. arXiv:2305.13300, 2023.
- [S92] Wei Zou, Runpeng Geng, Binghui Wang, et al. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. arXiv:2402.07867, 2024.

- [S93] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, et al. AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. arXiv:2407.12784, 2024.
- [S94] Amos Tversky, Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. Science 185(4157), 1974. doi:10.1126/science.185.4157.1124.
- [S95] Abraham S. Luchins. Mechanization in Problem Solving: The Effect of Einstellung. Psychological Monographs 54(6), 1942. doi:10.1037/h0093502.
- [S96] Alan D. Baddeley, Graham Hitch. Working Memory. Psychology of Learning and Motivation 8, 1974. doi:10.1016/S0079-7421(08)60452-1.
- [S97] Lee Ross, Mark R. Lepper, Michael Hubbard. Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm. Journal of Personality and Social Psychology 32(5), 1975. doi:10.1037/0022-3514.32.5.880.