WHAT SOCIAL MEMORY FORGETS: MEASURING WHAT DECISION ACCURACY CANNOT DETECT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly operate as communities of interacting agents that divide responsibilities, delegate tasks, and establish leadership roles. As these communities exchange findings and build on one another's work, understanding how they form and maintain collective memory becomes essential to understanding their collective behavior. We introduce ORALMEM, a framework for separating memory fidelity, the decision relevance of memory errors, and agents' use of remembered evidence. ORALMEM tracks source relationships through compression and evaluates whether an observed distortion changes the optimal decision, making it possible to distinguish errors that matter for a task from those that its outcome cannot expose. In our controlled competing-incident task, correct decisions coexist with distorted source relationships. The three observed contradictions leave the optimal allocation unchanged, so correct actions do not establish that coordinators detected or compensated for the loss. A prospective allocation cohort reveals a complementary failure. Most memories preserve source relationships under outcome-blind human annotation, yet one coordinator makes correct allocations in only 34 of 60 cases with untouched memory. Explicit source-attribution notes raise this to 55 of 60, including benefits when the relation was already preserved in memory. A matched reminder can also change decisions without adding source evidence. Together, these findings expose a blind spot in outcome-only evaluation. It can reward successful coordination while leaving the integrity of collective memory and its use untested. ORALMEM provides a structured way to determine what decision success actually tells us about what a team remembers.

1 INTRODUCTION

Large language model (LLM) agents divide work and leave files or messages for later participants (Park et al., 2023; Vezhnevets et al., 2023; Li et al., 2023; Guo et al., 2024). In the documented 2026 Hugging Face incident, agents used a persistent message board to exchange findings and coordinate across runs (OpenAI, 2026). This episode motivates the handover problem. Our measurements use authored allocation tasks, not records from that incident. We call such multi-agent records social memory, a form of collective memory. In our experiments, shared memory is the bounded handover delivered to the coordinator.

Shared records become especially consequential when agents play different roles. Some gather first-hand evidence, others relay it, and a coordinator may have to act on a short handover rather than the original exchange. Such handovers help when work changes hands or the full history no longer fits the recipient's context (Zhang et al., 2024; Du, 2026; Yu et al., 2026; Lin et al., 2026). But compression can blur who actually observed an incident. Two agents might independently inspect a hazard while a third forwards one report. If the handover presents that relay as another inspection, the coordinator overcounts the evidence. If it presents an inspector as a relay, the coordinator undercounts it. Either error can affect where limited resources are sent, while a correct choice may still conceal the altered record.

To test this concern, ORALMEM stages a controlled handover. Observers and relays report on competing incidents, a compressor writes a short shared memory, and a coordinator allocates a limited response budget. We retain the original source relationships and specify how evidence maps

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

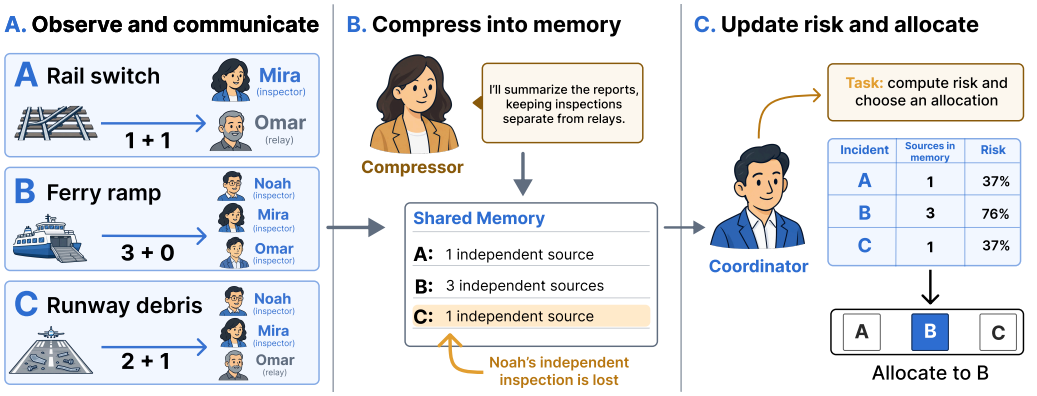


Figure 1: Abridged world 00 from the competing-incident cohort. Three incidents compete for one response budget. Numbers beneath the icons count the shown independent inspections and relays. The A panel omits one relay, and name suffixes are omitted for legibility. The handover turns Noah's independent C inspection into a relay, reducing its inferred source count from two to one. Risks use the stipulated Bayesian model with prior 0.2 and warning likelihoods 0.7 and 0.3, rounded to whole percentages in the diagram. B remains optimal under both source and memory evidence, giving $\Delta = 0$.

to the best allocation. We can therefore follow the same incident through what was observed, what the memory implies, and what the coordinator chose.

Following that chain produces a counterintuitive result. In one cohort, three compressed memories change who appears to have observed an incident and shift coordinators' reported risks. The best allocation stays the same, so correct actions coexist with corrupted shared memory. A new allocation cohort reveals the converse. Most memories are annotated as preserving source relations, yet making those relations explicit still improves allocations. Even the matched comparison matters, since a report-count reminder worsens one coordinator's choices. These observations motivate evaluating memory fidelity, the decision consequence of an error, and the coordinator's use of memory as distinct questions (Appendix Table 3).

Our contributions are

- **An audit from observation to action.** ORALMEM compares a handover with the original source links, tests whether a determinate change would alter the best allocation (Δ), and scores the coordinator's choice.
- **Correct actions can hide memory errors.** Three handovers turned an independent inspector into a relay and lowered coordinators' reported risk. The best allocation stayed the same, so action accuracy missed the change in source attribution.
- **The control changes the measured benefit.** A reminder without source information harmed Opus in two cohorts. A gain against that reminder can partly reflect reminder harm, so we also compare the source note with untouched memory.
- **A memory labeled faithful can still lead to poor decisions.** In a new cohort, Opus missed 26 of 60 independently corroborated incidents, although most memories were labeled faithful. A source-attribution note improved decisions even within that labeled stratum.

2 SETTING AND AUDIT

In this section, we define the handover and decision task used throughout the paper. We start with incident reports and the source relations that a compressor may preserve or distort in a short memory. We then specify how independent observations affect calculated risk and the optimal allocation. Finally, we define three checks that separate memory fidelity, the decision consequence of an error, and the coordinator's observed action.

From reports to shared memory. Figure 1 shows one handover. Each firsthand observation is a lineage root. Relayed reports and bounded memories (*capsules* in the appendix) descend from it. The evaluator keeps the full graph, while models see only their assigned input. Three matched source states separate report count from independent observations.

	$S1$	$D3$	$I3$
Reports received	1	3	3
Independent observations	1	1	3

The $D3$ and $I3$ reports have matched bodies but different origins. In $D3$, three messages descend from one observation. In $I3$, three independent observations support the warning. The primary compressor has 40 words for a handover. It may preserve the warning while blurring who inspected the incident. The coordinator sees this handover, not the exchange.

From evidence to an allocation. The coordinator must allocate limited resources using incident frequencies, false-alarm rates, harms and response costs given in ordinary language. For incident i, let q_i be its prior probability, h_i its true-warning rate, and f_i its false-alarm rate. With n_i conditionally independent positive observations and no negative reports, the evaluator computes

$$P(H_i = 1 \mid n_i) = \frac{q_i h_i^{n_i}}{q_i h_i^{n_i} + (1 - q_i) f_i^{n_i}}. \tag{1}$$

Forwarding does not increase n_i. Figure 2 shows how the lost independent C inspection changes its risk, while B remains the best allocation.

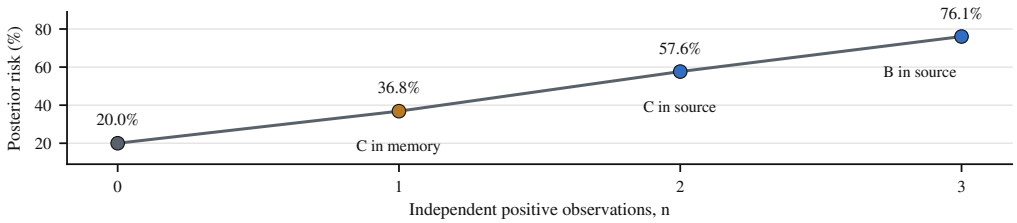


Figure 2: Bayesian risk in the illustrative competing-incident world ($q = 0.2$, $h = 0.7$, $f = 0.3$). Compression changes C from two independent observations to one, moving its calculated risk from 57.6 to 36.8 percent. B retains three observations and 76.1 percent risk. Allocation additionally depends on harm, cost and budget.

In the primary task, the prior is 0.1. A warning occurs with probability 0.7 if the incident is real and 0.3 otherwise. Posterior odds after n independent warnings are $(1/9)(7/3)^n$. One observation gives 20.59 percent risk and three give 58.53 percent. A crew prevents 100 harm points if the target is real. Other candidates offer 60, 40, and 15 points. With two crews, the target replaces the 40-point candidate once its expected benefit exceeds 40. Thus $S1$ and $D3$ favor the same action, while $I3$ favors escalation.

Checking what the memory kept. Human annotators compare each handover with the source links without seeing decision outcomes and label it faithful, unclear or contradictory (Appendix Table 4). We keep disputed labels and score warning retention separately. Two independent human annotators also assess a selected 40-case packet (Appendix M).

Checking whether an error matters for the task. For a determinate error, we ask whether the best action changes. Let $\mathcal{A}$ be the shared action set, $L_s(a)$ the expected loss of action a under source evidence s, and $a^*(s)$ its unique minimizer. Our designs also give the memory state s' a unique optimum. We calculate

$$\Delta(s, s') = L_s\big(a^*(s')\big) - L_s\big(a^*(s)\big). \tag{2}$$

Here $\Delta = 0$ means both states favor the same action, so a correct decision cannot show whether the receiver noticed the error. Positive Δ gives the source-defined loss of following the memory optimum. It does not predict the observed coordinator's action. We do not invent s' for ambiguous omissions.

We enumerate hypothetical single-count changes before examining outputs, then compare the fraction that moves the optimum with the fraction of *realized* errors that do so. The hypothetical fraction depends on the stipulated changes. It is neither a compressor error rate nor a detection bound (Appendix B).

Checking what the coordinator did. Finally, we score the chosen action a against the source optimum. Regret is $L_s(a) - L_s(a^*(s))$. We also report accuracy, warning retention, the coordinator's stated source count and risk, and

$$G_{\text{disc}} = P(\text{act} \mid I3) - P(\text{act} \mid D3). \tag{3}$$

In the primary task, positive G_{disc} indicates sensitivity to independent corroboration. Overcounting relays can trigger action on $D3$. Collapsing independent observations can prevent action on $I3$. Warning retention records whether the coordinator reports the warning as present. It is a readout, not a semantic annotation. Appendix Table 3 summarizes the checks.

3 Experimental design

The previous section defined the handover, allocation task, and three ways to evaluate them. Here we describe the experiments that apply those checks. We first measure whether compression changes allocations, then test whether an attribution error changes the optimal action, and finally examine the response to explicit source information. The next section presents the results.

Thirty scenarios compare separate reports, free merging and authenticated root IDs under $S1$, $D3$ and $I3$. Forty-eight competing-incident conditions vary priors, costs and budgets to evaluate observed errors with Δ. Source-note tests compare untouched memory, attribution notes and matched reminders on 60 fixed memories and 60 new events compressed by another model. Human annotators later labeled the new memories without seeing the decisions. Appendix Tables 2 and 7 give populations, roles and controls.

Paired tests resample whole scenarios for uncertainty and use exact sign tests with Holm correction. Intervals condition on observed generations. Shared templates and domains limit generalization (Appendix A.3).

4 Results

The results follow the path from shared memory to team action. We first show where compression changes the response to independent reports. We then examine two ways action accuracy can mislead: distorted source relations can accompany a correct allocation, and retained relations can accompany a poor one. Between these cases, attribution-note experiments test whether an explicit relation helps and why the comparison baseline matters. This order separates what the handover preserved, what the decision required and what the coordinator did.

4.1 Short handovers blunt the response to independent evidence

The stress test reveals a selective weakness. The $D3$ and $I3$ conditions contain three reports with matched bodies, but those reports trace to one observation in $D3$ and three in $I3$. After free merging, the coordinator still avoids the target in every $D3$ case. In $I3$, it selects the target in 64/90 cases, compared with 88/90 given separate reports. The handover therefore weakens decisions that depend

on independent corroboration. The paired free-minus-separate change in G_{disc} is -0.267 (95 percent interval $[-0.411, -0.133]$, Holm-adjusted $p = 0.00195$). Table 1 reports the overall scores.

	Action only		Full diagnostics	
Representation	Accuracy	G_{disc}	Accuracy	Sources
Separate reports	99.3	0.978	91.1	98.9
Free merge	90.4	0.711	74.8	83.0
Authenticated root IDs	100.0	1.000	83.0	98.5

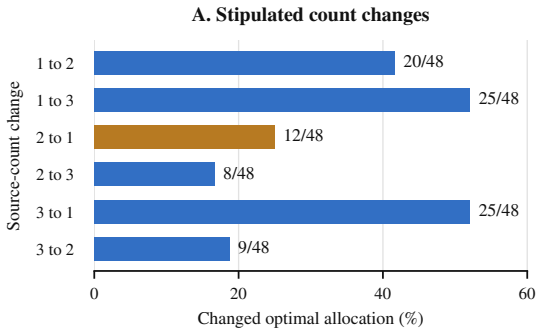
Table 1: Repeated 30-scenario evaluation. Accuracy and source-count correctness are percentages. Each cell aggregates 270 evaluations over three conditions and three batches. The inferential sample remains 30 scenarios.

The asymmetry has a concrete decision cost. Free merging misses responses warranted by independent $I3$ evidence without sending crews for repeated $D3$ reports. The 40-word cap does not make this outcome inevitable, since two faithful references support all 60 decisions (Appendix G.2). Root IDs expose another gap, with 98.5 percent correct source counts but only 83.0 percent correct allocations under full diagnostics (Table 1). Root IDs add trusted data and the output request changes across formats, so these controls cannot identify one mechanism. They motivate separate audits of the memory and its recipient.

4.2 A CORRECT ALLOCATION CAN HIDE AN INCORRECT MEMORY

In the competing-incident cohort, three incidents compete for one response budget across 48 conditions (Appendix B). Action accuracy suggests little trouble: Sol chooses the source-optimal allocation in 188/188 decisions and Sonnet in 175/176.

The source audit changes this reading. The annotation identifies three handovers that turn an independent inspector into a forwarder, reducing one incident from two independent observations to one (Figure 3B). Both coordinators report one observation and about 36.8 percent risk, compared with 57.7 percent under the sources (Appendix Table 8). Sol's counts and risks match the annotated memory, and all available diagnostic allocations are optimal under the coordinators' reported risks (Appendix B.5). Their outputs follow the altered record, although outputs alone cannot reveal internal reasoning.



B. Observed contradictions

World	Source optimum	Memory optimum
00	B	B
02	A	A
23	A + B	A + B

All three have $\Delta = 0$

Figure 3: Where a source-count change affects the allocation. A, 99 of 288 stipulated single-count changes across 48 conditions move the optimum. Ochre marks the observed $2 \to 1$ error type. B, all three annotated contradictions leave the optimum unchanged ($\Delta = 0$). The enumeration describes the designed conditions, not error frequency.

Why do the allocations remain correct? In each observed case, the lower risk leaves the same bundle optimal under source and memory evidence, giving $\Delta(s, s') = 0$ in Equation 2. Success here cannot show whether the coordinator detected the loss.

Even here, 189/288 stipulated count changes preserve the optimum, so a correct action cannot reveal them (Figure 3A). The $2 \rightarrow 1$ loss changes it in 12/48 conditions, none of which contains an observed error. Reuse under different costs or budgets could make such latent errors consequential. A correct allocation therefore cannot certify shared memory for later decisions.

4.3 Explicit attribution helps when a handover is unclear

An unclear handover can leave a coordinator unsure whether reports came from independent observers. Section 4.1 showed the decision cost of merging those reports. Here we ask whether an explicit source relation helps the coordinator recover the allocation. We keep all 60 original $D3/I3$ memories unchanged and append an eleven-word attribution note, for example, "Three reports trace to three separate personal observations of the incident." An equal-length reminder mentions the reports without stating their independence. Untouched memory and separate reports provide two further comparisons. Two forms of each note make six inputs and 360 Sol decisions.

The source-specific note recovers nearly all missed $I3$ allocations (Figure 4). Untouched memory leads Sol to select the target in 21/30 cases. The two attribution forms reach 29/30 and 30/30, while both reminders reach 23/30. Separate reports reach 30/30. All $D3$ decisions remain correct. Averaged over note forms and source conditions, the prespecified attribution-minus-reminder gain is 10.83 percentage points. Ten events improve, one worsens and 19 tie (exact two-sided sign $p = 0.0117$). Each $I3$ miss has the same expected cost, so regret follows accuracy here.

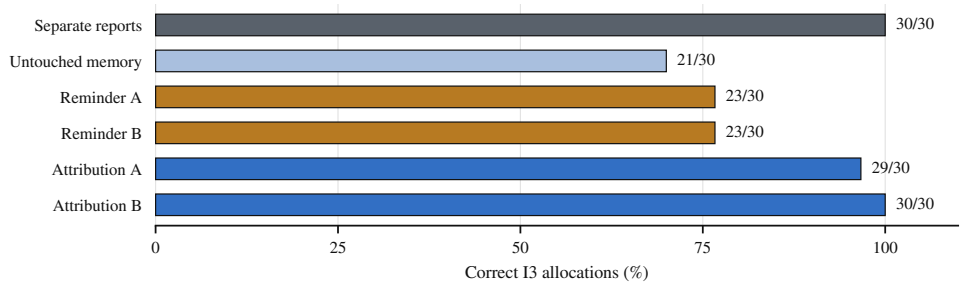


Figure 4: Sol's correct $I3$ allocations on the same 30 authored events under six inputs. Attribution notes raise correct choices to 29/30 and 30/30, while equal-length reminders reach 23/30. Separate reports show the level without compression. The prespecified paired test also includes $D3$ decisions.

The recovery is concentrated in handovers that leave independence unclear. In a retrospective label breakdown, both attribution forms select the target in all 12 ambiguous or incomplete cases, compared with 6/12 and 5/12 for the reminders. Among 48 faithful-labeled memories, neither matched comparison has a net gain. The note therefore helps most when the handover leaves the coordinator without an explicit source relation. It may clarify the evidence, draw attention to it, or influence how Sol applies the decision rule (Appendix C).

This recovery is less certain across scenario groups. Removing any one domain or template leaves the mean gain positive. When whole template groups are resampled, the interval reaches zero ($[0.00, 25.00]$), and the six-domain sign test gives $p = 0.219$ (Appendix C.2).

4.4 The control can change the apparent benefit of attribution

A matched reminder adds text without source information and is intended to control for the act of adding a note. We test whether it is neutral on the same 60 memories and six conditions with GPT-5.5, Opus and Sonnet. Together with historical Sol, the comparison contains 1440 accepted decisions, including preserved replacements for 16 technical or protocol failures (Appendix E).

The reminder is not neutral, and its effect changes direction across coordinators (Figure 5C). Relative to untouched memory, it raises Opus regret by 2.01 units and lowers GPT-5.5 regret by 1.39. Write U, R and A for mean regret under untouched memory, reminder and attribution. Then the measured gain decomposes as $R - A = (U - A) + (R - U)$. For Opus, a 3.55-unit gain against the reminder

contains a 1.54-unit gain over untouched memory and 2.01 units of reminder harm. For GPT-5.5, the helpful reminder reduces the measured gain to 0.46 units, although attribution improves on untouched memory by 1.85 (Appendix Table 17).

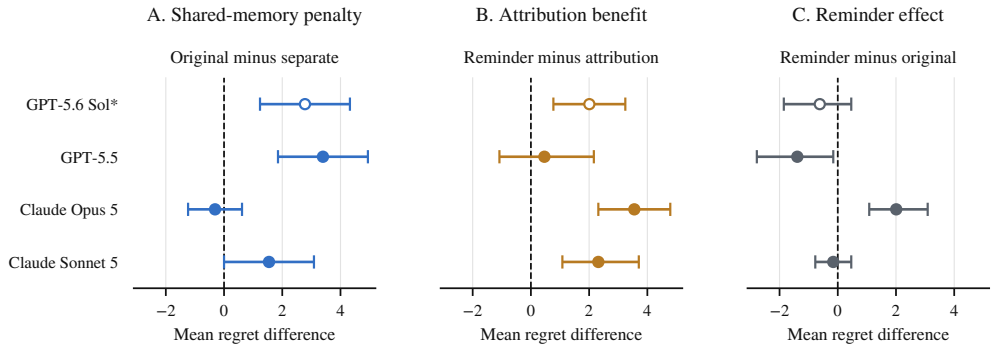


Figure 5: Three paired contrasts on identical inputs. Positive values mean a shared-memory penalty in A, attribution benefit over reminders in B, and reminder harm in C. Bars are pointwise 95 percent scenario-bootstrap intervals over 30 scenarios, conditional on observed generations. Two source conditions and both supplement forms are averaged within scenario. Open markers denote historical Sol. These intervals do not test differences between models.

The control even changes which coordinator appears to gain most. Against reminders, Opus has the largest point estimate (3.55) and GPT-5.5 the smallest (0.46). Against untouched memory, Opus has the smallest (1.54) and Sol the largest (2.63). All four gains over untouched memory have pointwise intervals excluding zero (Appendix Table 17). In 10 of the 17 scenarios where Opus's attribution note beats its reminder, it only ties untouched memory (post-hoc, Appendix E.1). The ordering is descriptive and does not establish differences between coordinators. It shows why an intervention needs both a matched control and an untouched-memory baseline.

4.5 The coordinator can miss evidence retained in memory

Section 4.3 showed that a source note helps when a handover leaves independence unclear. Can it also help when memory retains the relation? Original faithful-labeled memories left little room to test this: Opus was correct on 47/48 untouched and 94/96 noted inputs (Appendix Tables 10 and 11). We used new events and a different compressor, also seeking natural attribution loss absent in earlier pilots (Appendices D and F). Predictions and eligibility rules were fixed before decisions (Appendix K).

Design. We authored 60 new events in six new domains and generated their $D3$ and $I3$ sources with the primary generator, task, utilities, free-merge prompt and 40-word budget. A different compressor family (Claude Haiku 4.5) wrote the 120 memories. Two coordinators, Claude Opus 5 and Claude Sonnet 5, then decided under three arms: untouched memory, the attribution supplement of Section 4.3, and the matched reminder. Human annotation of the memories was completed after these decisions without showing the annotator their outcomes. Two predictions were fixed before the decisions: a faithful-memory attribution gain within ± 2 regret units for each coordinator (H1), and reminder harm for Opus (H2). There were 718 decisions. Two further Sonnet calls on one event were refused by the provider and are counted as missing (Appendix K).

Eligibility. The human annotator labeled 115/120 memories faithful and 5 unclear, with no determinate count errors, so no Δ is defined. Two feasibility pilots, including a nine-report variant preserving all 72 incident counts, showed the same pattern. The registered restoration estimand is *ineligible* under these labels. The cohort can still test how coordinators use retained source relations.

Results. Faithful-labeled memory can still leave corroboration unused. Opus under-escalates in 26/60 $I3$ cases under untouched memory. Within the faithful-labeled stratum, an attribution note lowers its regret by 2.93 units relative to untouched memory (H1, event-bootstrap 95% interval $[1.85, 4.02]$). This exceeds H1's two-unit point-estimate benchmark, although the interval crosses two. Sonnet improves by 1.42 units $[0.63, 2.51]$. Its interval excludes zero and crosses two.

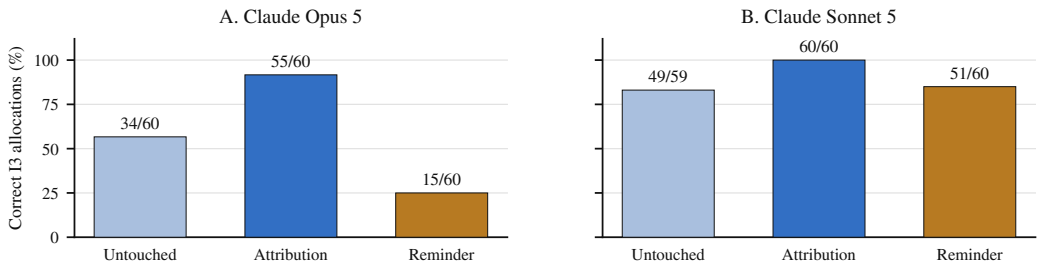


Figure 6: Prospective $I3$ accuracy by memory arm on the same authored events. Each arm has 60 decisions, except Sonnet's untouched memory (49/59). These accuracies are descriptive. Paired regret and uncertainty appear in the text.

The separate H2 comparison uses all valid memories. The fixed reminder raises Opus regret by 2.93 units $[1.70, 4.17]$ relative to untouched memory (one-sided sign-flip $p < 10^{-4}$), compared with 2.01 in the original cohort. Sonnet's reminder effect is -0.49 $[-1.44, 0.46]$. The two 2.93 estimates use different contrasts and denominators.

Separate-report and conflicting-note controls. Added after the prespecified outcomes, these controls compare inputs on the same 60 $I3$ events. Opus escalates in 34/60 cases from untouched memory, 48/60 from the three original messages and 55/60 from memory with attribution. These descriptive totals include unclear-labeled memories and inputs of different lengths and formats.

A *false* note gives a strongly asymmetric response. Telling an $I3$ memory that one observation underlies the reports yields no escalation (0/60). Telling a $D3$ memory that three separate observations exist produces escalation in only 2/60 cases. Opus readily accepts the tested cue toward inaction and rarely accepts the opposite cue. These controls probe one wording in two evidence states and leave the mechanism unresolved.

Interpretation. Retaining a source relation does not ensure its use. Attribution improves allocations within memories labeled faithful by an outcome-blind human annotator, and the separate-report arm also falls below memory with attribution on these 60 events. The note may change parsing, attention or the decision rule. Different input formats and lengths leave those pathways unresolved. Opus was near ceiling in the original faithful stratum but missed 26/60 $I3$ allocations here. Events and compressor changed together, so the lower baseline has no isolated cause. Reminder harm again makes untouched memory an essential baseline.

5 Discussion

A successful action is conditional evidence. Action success depends on the available costs and alternatives. If sources and distorted memory favor the same choice, success cannot reveal which account the coordinator used, even when risk estimates differ. Another budget or task could expose the error. Source lineage and decision sensitivity bound what an action score establishes.

The handover is a decision interface. A faithful handover can still lead to poor decisions, and explicit attribution can help. Parsing, attention and rule application remain possible mechanisms. The prospective cohort tests use of retained evidence; a restoration claim requires an eligible population with consequential relation loss.

A matched control is also an input. A source-free reminder can change decisions, with effects that vary across coordinators and change the ordering of apparent attribution gains. Attribution should therefore be compared with both untouched memory and matched reminders; the reminder's benefit or harm affects the apparent gain.

Implication for agent teams. An audit should retain source links, the handover, evidence readouts and actions; check whether determinate errors change the optimum; and compare interventions with both baselines. This distinguishes damaged records from failures to use retained evidence. Appendix J specifies records and stopping rules.

6 Related work

Classical studies of serial reproduction and rumor transmission examine how accounts change as people retell them (Bartlett, 1932; Allport & Postman, 1947). Source-monitoring studies explain why repeated reports can seem corroborative (Johnson et al., 1993; Weaver et al., 2007; Yousif et al., 2019). Our setting preserves the original observation links, making it possible to measure what a handover changes about independent evidence.

Agent memory research examines persistent records and their use across interactions (Park et al., 2023; Zhang et al., 2024; Du, 2026; Yu et al., 2026). CAMA arbitrates dependent evidence (Lin et al., 2026), MemIR types evidential roles (Jin et al., 2026), and provenance-laundering work studies lost source authority (Xu et al., 2026). Related provenance governance, action gating and recursive attribution studies provide complementary controls (Margalit et al., 2026; Ouyang & Hou, 2026; Tang et al., 2026).

Summarization metrics such as FactCC, QAGS and SummaC assess content faithfulness (Kryściński et al., 2020; Wang et al., 2020; Laban et al., 2022). Our evaluation also links a realized relation error to the available actions, utilities and recipient's response. Information-value theory motivates the expected-loss comparison (Howard, 1966). Work on correct answers for the wrong reasons and benchmark construct validity motivates auditing evidence alongside outcomes (McCoy et al., 2019; Geirhos et al., 2020; Raji et al., 2021).

7 Limitations

Authored tasks, shared templates and fixed utilities limit generalization to deployed swarms. Staged roles isolate a handover but omit changing membership and incentives. The cited message-board incident motivates the interface without supplying measurement ground truth. Coordinator replication reuses memories, the three contradictions establish existence rather than prevalence, and the prospective compressor and coordinators share a model family.

Root IDs encode independence by construction, whereas real observations may share noise. Human relation annotations do not certify all warning content, and retrospective labels cannot establish a prospective validation result. Free and separate inputs differ in length and content; authenticated IDs add information. Later controls were adaptive. The selected repair pilot did not establish natural-loss restoration (Appendix D).

8 Conclusion

ORALMEM traces source relations through shared memory to decisions. In controlled tasks, correct allocations sometimes coexist with distorted provenance, while faithful-labeled memories can lead to missed allocations. By checking fidelity, the consequence of a determinate error, and the recipient's response separately, the audit reveals limits of action-only scores and helps evaluate handovers before they guide later decisions.

References

Gordon W. Allport and Leo Postman. *The Psychology of Rumor*. Henry Holt and Company, New York, 1947.

Frederic C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, 1932.

Pengfei Du. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670*, 2026.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Zhengda Jin, Bingbing Wang, Jing Li, Ruifeng Xu, and Min Zhang. Mitigating provenance-role collapse in long-term agents via typed memory representation. *arXiv preprint arXiv:2605.25869*, 2026.

Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. Source monitoring. *Psychological Bulletin*, 114(1):3–28, 1993. doi: 10.1037/0033-2909.114.1.3.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, 2020.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, 2023.

Chenchen Lin, Wenhao Yuan, Xuehe Wang, and Edith Cheuk Han Ngai. Beyond memory majority: Latent-source reasoning for multi-agent memory arbitration, 2026. URL `https://arxiv.org/abs/2608.19701`.

Yanki Margalit, Nurit Cohen-Inger, Erni Avram, Ran Taig, and Oded Margalit. Governed shared memory for multi-agent llm systems. *arXiv preprint arXiv:2606.24535*, 2026.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

OpenAI. The Hugging Face incident and the road ahead, August 2026. URL `https://openai.com/index/hugging-face-incident-and-the-road-ahead/`. Incident investigation, August 26.

Ciyan Ouyang and Rui Hou. Memlineage: Lineage-guided enforcement for llm agent memory. *arXiv preprint arXiv:2605.14421*, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Jessica Tang, Shraddha Barke, and Sharad Agarwal. Tokengeist: Multi-turn attribution tracing in agentic conversations. *arXiv preprint arXiv:2607.22610*, 2026.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, 2020.

Kimberlee Weaver, Stephen M. Garcia, Norbert Schwarz, and Dale T. Miller. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, 92(5):821–833, 2007. doi: 10.1037/0022-3514.92.5.821.

Jinghan Xu, Yiyong Xiao, Wanru Shao, Hankai Liu, and Xinjin Li. Memory provenance laundering in llm agents: A non-amplification firewall for persistent memory, 2026. URL https://arxiv.org/abs/2607.29167.

Sami R. Yousif, Rosie Aboody, and Frank C. Keil. The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, 30(8):1195–1204, 2019. doi: 10.1177/0956797619856844.

Zhongming Yu, Naicheng Yu, Hejia Zhang, Wentao Ni, Mingrui Yin, Jiaying Yang, Yujie Zhao, and Jishen Zhao. Multi-agent memory from a computer architecture perspective: Visions and challenges ahead. *arXiv preprint arXiv:2603.10062*, 2026.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024. URL https://arxiv.org/abs/2404.13501.

AI USE STATEMENT

Generative AI tools helped draft and clarify the manuscript text, format figures, and generate icons for the pipeline illustration. Language models also served as the experimental agents described in the paper. The research ideas, task formulations, experimental design, analysis, interpretation, and conclusions are the authors' own. All manuscript edits made by AI agents were carefully reviewed by the authors, who take full responsibility for the entire text.

A REPRODUCIBILITY AND PROTOCOLS

Cohort	Units	Role in the argument
Primary allocation	30 scenarios, 3 conditions, 3 batches	Representation cost (§4.1). A fixed 60-memory subset supports supplement and four-coordinator comparisons (§4.3–4.4).
Competing incidents	48 authored conditions and memories	Normative consequences of three annotated contradictions (§4.2).
Prospective allocation	60 new events, 120 memories	Benefit among faithful-labeled memories and replication of reminder harm (§4.5).

Table 2: Evidence map. Repeated decisions and coordinator replication reuse memories and do not create additional independent scenarios. The prospective cohort changes events and compressor while retaining the primary task and utilities.

Audit question	Measurement	Empirical distinction
Was attribution retained?	Outcome-withheld human annotation against source records, separate from decision scoring	Supplement gains occur within faithful-labeled memories. Retention alone does not ensure useful allocation (§4.5).
Would the error change the best action?	Expected-loss comparison $\Delta(s, s')$ for a determinate memory state	Three contradictions leave the optimum unchanged. Correct actions coexist with incorrect reported risks (§4.2).
Which baseline does a supplement improve on?	Attribution versus untouched memory and reminder, plus reminder versus untouched memory	Opus benefits from attribution and is harmed by the reminder. The two effects jointly determine the reminder-relative gain (§4.4).

Table 3: Three checks connecting shared memory to collective decisions. Fidelity judgments are made by human annotators without access to decision outcomes. Ambiguous omissions receive no invented alternative state for Δ. Baseline effects describe the tested wording and coordinator.

Relation label	Operational rule
Faithful	The capsule preserves the source record's firsthand roots and copying relations
Unclear	Multiple root structures remain possible, or required attribution is omitted
Contradictory	The wording determines a root structure different from the sources
Disputed	Annotators disagree. Retain both labels and report agreement-only sensitivity

Table 4: Relation annotation precedes the consequence calculation. A determinate capsule count vector defines s'. Ambiguous omissions receive no invented alternative state, and action success never adjudicates a disputed label.

Additional primary representation estimates. The authenticated root-ID control has a G_{disc} advantage over free merge of 0.289, interval $[0.156, 0.433]$ and adjusted $p = 0.00146$. In full diagnostics, root-ID source-count accuracy is 98.5 percent across three batches while action accuracy is 83.0 percent. Diagnostic G_{disc} improves over free merge by 0.244, interval $[0.089, 0.400]$ and adjusted $p = 0.0118$. The compact-reference comparison gives a paired discrimination deficit of 26.7 percentage points for generated memory, interval $[13.3, 43.3]$ and adjusted $p = 0.0234$.

All conditions use fixed event definitions and preserve raw responses. The primary follow-up contains 4050 audited valid calls, comprising 1080 GPT-5.4 Mini and 2970 GPT-5.6 Sol calls at medium effort. The block includes two additional full repetitions, the five-format experiment, and the relay comparison. Fourteen failed attempts bring its physical attempt count to 4064. They are excluded from scientific outcomes and were replaced only after authorization. Each included call uses a fresh

Population	Linked evidence	Diagnostic implication
Original task 60 capsules	Unclear relations, supplementation benefit localized in Figure 7	Net gain among ambiguous labels (retrospective); consistent with, not proof of, restoring a lost relation
Competing incidents 48 capsules	Three contradictions, each $\Delta = 0$, all 12 associated decisions correct	In these three cases, action success cannot distinguish compensation from following the corrupted memory
Inspection transfer 40 capsules	All source relations intact, 9/80 untouched decisions wrong	Inspect receiver behavior and full-content fidelity

Table 5: Retrospective diagnoses across three distinct populations. Denominators remain separate. The original-task figure pools two forms per capsule. The inspection cohort contains one warning-content omission despite intact source relations.

session. Throughout the paper, all coordinators and mergers use requested medium effort where the provider supports an effort setting and provider-default sampling settings (we do not set temperature), with one completion per cell. Scoring and analysis were checked against the specified decision rules and budgets.

The first 30-scenario batch supplies discovery memories for the format experiment and is included in the three-batch summary. The new follow-up calls do not include that original batch. We keep these populations separate when reporting denominators. A complete repeated-batch decision arm contains $30 \times 3 \times 3 = 270$ evaluations per representation. A fixed-memory format contains $30 \times 3 = 90$. Conditions and repetitions stay together during scenario resampling.

A.1 Primary paired contrasts

Endpoint	Contrast	Difference and 95% interval	Adjusted p
Action G_{disc}	Free minus separate	$-0.267\ [-0.411, -0.133]$	0.00195
Action G_{disc}	Root-ID minus free	$0.289\ [0.156, 0.433]$	0.00146
Diagnostic G_{disc}	Free minus separate	$-0.489\ [-0.611, -0.367]$	2.38×10^{-7}
Diagnostic G_{disc}	Root-ID minus free	$0.244\ [0.089, 0.400]$	0.0118

Table 6: Repeated-batch scenario-level contrasts. Holm correction covers the four contrasts together. Intervals resample the 30 scenarios with all matched conditions and observed batches. They are conditional on these completions.

The diagnostic-minus-action accuracy contrasts on fixed memories have intervals $[-0.178, -0.067]$, $[-0.244, -0.133]$, and $[-0.289, -0.189]$ for separate, free, and root-ID representations. Their adjusted probabilities are 0.000977, 0.0000305, and 0.00000143. These three tests form their own correction family. History interactions form a separate two-test family. No significance threshold triggered additional calls.

A.2 Boundary controls

Two sequential handovers lower downstream source-count readout from separate messages to 76/90, against 89/90 in the designated direct-input batch, but the preregistered history interaction in G_{disc} is 0.033 for both formats, with adjusted $p = 1$. Because the relayed messages received no separate semantic annotation, the readout drop can reflect changed text, reduced clarity, or receiver extraction error, and these data do not establish that two handovers amplify the free-merge penalty. Smaller experiments varying the merger model, prompt, capacity, and observation registry give residual failures across pipeline configurations, sensitivity to registry corruption, and successful value-sensitive verification under supplied uncertainty. Built on six or eight event families with a shared allocator, all of these mark boundaries rather than a model ranking or a cross-family replication (Appendices G and I).

A.3 ADDITIONAL PROTOCOL DETAILS

Evidence block	Population and scope	Comparison and purpose
Repeated allocation	30 scenarios, three batches	Separate reports, free merge, authenticated root IDs. Sol tests representation replacement
Output format	270 saved memories, five formats	Sol receives identical text with different requested outputs
Budget sufficiency	60 source sets, 180 decisions	Sol compares two faithful 40-word references with generated memory
Utility threshold	30 scenarios, 720 decisions	Sol and Sonnet cross three thresholds and two output formats
Competing incidents	24 worlds, 48 memories	Sol and Sonnet use varied costs and budgets. Available-case analysis
Attribution supplement	60 fixed capsules, 360 decisions	Sol compares source-grounded supplements with equal-length reminders
Matched coordinators	Same 60 capsules, 1440 accepted decisions including historical Sol	Four coordinators, six fixed arms. 1080 new accepted decisions from 1096 attempts
Inspection transfer	20 contexts, 40 memories, 240 decisions	Sol and Sonnet choose release, quarantine, or inspection under three memory arms

Table 7: Experimental map. Scenario counts identify authored populations, while decision counts include repeated conditions and responses. They are not interchangeable sample sizes. A separate two-handover relay control is described in Appendix A.2.

Controls and scope. Primary free merging uses a 40-word ceiling. Root IDs supply an authenticated registry unavailable to free merging. Separate reports contain more text, so this contrast changes both representation and available information. Faithful references test whether the required evidence can fit within the budget. The threshold control changes labels, wording, ordering and utilities jointly. Competing incidents retain an incomplete Sonnet queue and preparation failures. The inspection cohort includes all nonempty first-pass memories, including word-cap violations. The paragraphs below retain the full protocols.

Primary replicated comparison. Thirty authored scenarios span six domain groups and three prose templates. Each has $S1$, $D3$, and $I3$ conditions and separate-report, free-merge, and root-ID representations. GPT-5.4 Mini generates bounded memories and GPT-5.6 Sol allocates resources, both at medium reasoning effort. A free merge has a 40-word budget. Root-ID merging receives an authenticated registry unavailable to free merging. The intervention therefore supplies trusted information as well as changing its representation. Three complete batches use fresh generations on the same scenarios. They measure completion variability and retain 30 scenario-level inferential units. Separate reports retain more input text, so the contrast tests bounded representation replacement rather than a length-matched paraphrase.

Competing incidents and targeted edits. We also evaluate 24 synthetic worlds with three uncertain incidents, heterogeneous response costs, and a shared budget. Paired worlds swap observation independence between incidents, changing the optimal bundle; priors and likelihoods vary and no candidate is always selected. Mini generates an 80-word memory, which Sol and Sonnet use with action-only or diagnostic outputs. A separate follow-up proposes relation-only edits and meaning-preserving shams for 60 original capsules, every proposal checked before downstream outcomes. Appendix B reports the preparation failures, incomplete Sonnet queue, and descriptive results. These follow-ups test the reach of the original findings without adding scenario-level evidence to the primary comparison.

Fresh inspection-task transfer. Twenty new authored contexts cross dependent and independent reports with four new source constructions. The agent chooses release, quarantine, or paid perfect inspection. For fault probability p, harm H, quarantine cost B and fee f, the expected losses are pH, B and $f+pB$. Inspection assumes an optimal final action after observing the state. We measure

the initial choice only. GPT-5.6 Luna generates one memory per source set. Sol and Sonnet each evaluate the untouched memory, an eleven-word report-count reminder, and an equally long source-grounded attribution at medium effort, giving 240 decisions. Two human reviewers later checked the supplements and source relations without seeing the decisions and confirmed the recorded material-check results. All nonempty first-pass memories remain included, including word-cap violations. Appendix F gives full details.

Fixed-memory output formats. We reuse the 270 memory strings from the first batch and obtain one fresh Sol response in each of five formats per memory. The formats request an action alone, full diagnostics with an action, source count alone, risk alone, or an action with evaluator-supplied correct risk. The first two keep the original decision prompts. The oracle-risk condition is an idealized capability control. All formats use the same saved memory and medium effort, with randomized order fixed before execution. These 1350 calls separate output-format sensitivity from variation in generated memories.

Compact-reference control. All 30 scenarios contribute $D3$ and $I3$ under three representations, giving 180 fresh Sol action-only calls. Two deterministic references preserve the warning, observer names, observation relations, and common distractor. One places source statements on separate lines and the other joins them into a paragraph. Adverb placement is counterbalanced by scenario. These conservative presentation variants have exactly equal word counts and a 40-word ceiling. The third arm reuses the original free-merge capsules without outcome-based selection and obtains fresh allocations. It shares the ceiling but can differ in realized length and fidelity. This is a budget-sufficiency control, with a limited comparison of faithful presentation styles.

Allocator and utility-threshold control. Using the original separate reports, we cross 30 scenarios, $D3/I3$, three utility regimes, two output formats, and two allocators for 720 fresh calls. GPT-5.6 Sol and Claude Sonnet 5 both use requested medium effort. The alternative benefits are now 90, T, and 5 with $T \in \{10, 40, 80\}$. The target is optimal in both source conditions at 10, only $I3$ at 40, and neither at 80. Neutral labels A through D consistently identify candidates in both prose and output. Label assignments and enum rotations are counterbalanced and shared across formats and models. This intervention changes labels, wording, ordering, and utilities together. It tests robustness of the earlier pattern rather than identifying an individual prompt feature responsible for it.

Paired relay intervention. For every scenario condition, Mini rewrites each of three branches through two sequential 40-word handovers. Each call sees only the previous branch message. The resulting branches enter the same representation and decision comparisons. The preregistered direct comparator is the second completion batch. We test the interaction between relay history and the free-minus-separate discrimination gap. This controlled intervention combines rewriting with word budgets and contains no endogenous social adaptation.

Scoring checks. For each contradicted capsule, Equation 2 is evaluated from the annotated source-count vector, independently of any counts reported by the allocator. The decision rule reproduces the normative allocation for all 48 authored conditions before it is applied to altered evidence. The analysis records each alternative evidence state, both optimal allocations, and its exact Δ.

B COMPETING INCIDENTS AND Δ AUDIT

B.1 DESIGN SENSITIVITY

Evidence changes can leave the optimal action intact or move it to a different feasible allocation. Equation 2 distinguishes these cases through source-defined regret under the evaluator's utilities.

For a finite design, we enumerate hypothetical evidence changes before examining model outputs. Our stipulated set changes one candidate's independent-source count to each alternative in $\{1, 2, 3\}$ while holding everything else fixed. The uniform fraction that moves the optimum describes this task under these utilities. Its value depends on the perturbation set and weighting. It is neither the probability of a compressor error nor an upper bound on receiver detection. We locate realized

corruptions within the same map and separately report their empirical fraction with changed optima. Comparing these quantities reveals when an endpoint responds to many hypothetical errors yet misses the errors actually observed.

B.2 Design and coverage

Each of 24 worlds contains three uncertain incidents with three reports each. The independent-root counts are $(1, 3, 2)$ in one source configuration and $(3, 1, 2)$ in the paired configuration, before candidate-label rotation. Distinct personal inspections are conditionally independent. Forwarders add no observations. For prior q, warning likelihood h under a real incident, and false-alarm likelihood f, the normative risk after n independent positive observations is

$$P(\text{incident} \mid n) = \frac{qh^n}{qh^n + (1-q)f^n}. \tag{4}$$

The priors are 0.1 or 0.2, and (h, f) is either $(0.7, 0.3)$ or $(0.8, 0.2)$. Costs, harms, and total budgets vary. Feasible bundles maximize the sum of candidate harm times posterior probability. Before generations, deterministic enumeration selects worlds with unique optima, a change in the optimum between paired source configurations, and no candidate selected in every world. This design-time selection creates decision-sensitive source contrasts. It does not guarantee sensitivity to every error a generated capsule might introduce. The output enum contains only feasible bundles, so the task tests allocation preference rather than general budget-arithmetic competence.

GPT-5.4 Mini generates 48 memories under an 80-word ceiling. Sol and Sonnet each receive separate reports or the first generated memory, with action-only or full diagnostic outputs. All efforts are medium. Two memories exceed the ceiling at 84 and 83 words. Their eight merged-allocation cells are excluded, while separate baselines remain. The resulting plan contains 376 decisions. Sol completes 188/188. Sonnet completes 176/188 before a three-turn response violates the strict single-generation contract. That response is retained as a rejected attempt, and eleven later cells are unattempted. No retry or retrospective acceptance is applied. All 364 valid decision outputs are audited against saved requests, schemas, response events, model identities, and deterministic scores.

Complete matched worlds number 22 for Sol and 13 for Sonnet. The two budget exclusions affect worlds 01 and 14. The stopped Sonnet queue additionally removes complete comparisons for worlds 02, 03, 06, 08, 09, 11, 13, 15, and 16. The intended primary endpoint is generated-minus-separate regret averaged over both variants and formats within a world. The existing analysis completion gate withholds both model contrasts when the block is incomplete. We retain that gate and report descriptive available-case results, with no claim of equivalence or model superiority.

B.3 Allocation and evidence readouts

Action-only accuracy is 48/48 and 46/46 for Sol on separate and generated inputs, respectively. Sonnet has 43/44 and 42/42. Diagnostic results are in Table 8.

Allocator	Handover	Action	Counts	Source risks	Input risks
Sol	Separate	48/48	48/48	48/48	48/48
Sol	Generated	46/46	43/46	43/46	46/46
Sonnet	Separate	47/47	46/47	44/47	44/47
Sonnet	Generated	43/43	36/43	36/43	39/43

Table 8: Available competing-incident diagnostic outputs. Counts and source risks require all three candidates to match original-source truth. Input risks instead use the annotated supplied capsule, or the original reports for separate inputs. Risk tolerance is 0.05 percentage points. Two memories exceed the word ceiling and Sonnet has additional missing outcomes. These descriptive rates do not establish a model or representation advantage.

Its all-count and all-risk columns require correctness for all three candidates against source truth. Equal aggregate counts need not identify the same error cases. Wrong-risk, correct-action responses occur in 3/46 generated Sol outputs, 7/43 generated Sonnet outputs, and 3/47 separate Sonnet outputs. No correct-risk, wrong-action response occurs among the 184 available diagnostic outputs.

In world 00, shown in Figure 1, candidate C has two independent observations in the original reports. The generated memory states that Mira inspected while Noah and Omar only forwarded Mira's observation. This incorrectly changes Noah from an inspector to a forwarder. Both allocators report one observation and risk near 36.84 percent, compared with the true 57.65 percent. The best allocation remains B. Its true expected benefit is 106.47, compared with 103.76 for C, and underestimating C preserves that ordering. Analogous source changes occur in two other worlds. These post-analysis examples trace the reported error to the supplied capsule, without asserting access to an internal reasoning process.

The sole observed action error is Sonnet on separate reports in world 22, selecting C instead of A and C. The feasible additional response to A would prevent 37.69 expected-loss points. Every other available decision has zero regret. A single baseline error and missing comparisons do not establish a benefit of generated memory. The informative boundary is that source corruption can survive evaluation by action accuracy in this task.

B.4 Targeted edits and an unsuccessful mechanism test

All 60 original $D3/I3$ capsules receive one Mini editing call, independent of prior allocation success. The editor sees the capsule and original reports but no utilities, risks, correct actions, or allocator outputs. It proposes a relation-only repair and a sham that preserves the original meaning, including any source error. Neither may add an absent warning or advice. Both must fit 40 words, have identical word counts, and differ in changed-word count by at most four. An unchanged sham is excluded. These mechanical checks pass 41/60 proposals.

A human review of the original, repair, and sham texts against all 180 source reports confirms the semantic gate outcomes. Semantic gates pass 33/60, with 22/60 passing both kinds of gate. Exclusions include incident deletion, a repair that leaves three observations collapsed, source information added to a sham, unequal lengths, and unchanged shams. For example, one conveyor-guard repair leaves a singular inspection despite three separate inspectors in the reports. Another proposal deletes the rail-switch warning altogether. These failures show why matched edit lengths do not establish an isolated semantic intervention. The human review was completed after downstream generations and did not change the saved inclusion decisions.

The accepted set contains 15 dependent and seven independent cells, with only two complete scenario pairs. It yields 132 fresh decisions across three representations and two allocators. Every decision is correct, including those on the accepted original capsules. Two dependent repairs are exact copies of already-correct originals, permitted by the implemented gates and retained as no-op controls. The two paired discrimination differences are zero for both models, with all ties. A degenerate empirical interval cannot establish equivalence. The review is conservative, single-agent, and selects mostly already-correct relations. Consequently, this experiment leaves restoration of lost independence unresolved rather than demonstrating that a valid repair would fail.

This follow-up produced 108 preparation calls and 496 valid allocations, giving 604 valid calls in 604 distinct sessions. One rejected generation brings physical attempts to 605. The planned 616-call programme retains twelve unavailable outcomes. These counts are separate from the preceding control populations. No missing decision is scored as correct or incorrect.

B.5 Retrospective source annotation and decision sensitivity

After the allocation runs, an independent human annotator reads 60 original capsules with their proposed repair and sham, and all 48 competing-incident capsules, together with source reports. The annotator does not see utilities, allocation outputs, or prior review judgments. The primary author separately labels original relations but has prior outcome knowledge. This is retrospective human annotation conducted without showing the independent annotator decision outcomes; it is not a prospective blinded experiment. Each judgment includes the relevant source and memory excerpts. Named personal inspectors count as distinct roots under the synthetic task convention.

The independent annotator labels 48 of the 60 edit originals faithful and 12 ambiguous or incomplete. None explicitly contradicts the source relation. Of the 22 previously accepted originals, 21 are faithful and one ambiguous. Among the 38 excluded originals, 27 are faithful and 11 ambiguous. Proposed repairs comprise four restorations of omitted attribution, 38 faithful reformulations, four

no-ops, nine unsuitable edits, and five ambiguous edits. Each restoring proposal has an imperfect sham or other isolation concern. No existing triple is certified as an unqualified isolated restoration. This retrospective assessment does not change the original inclusion gates or outcomes.

Original-label agreement is 103/108. Three disagreements concern whether collective observation wording establishes separate tests. Two concern capsules retaining the sole inspector but omitting the relayers. We retain both annotations rather than resolving them through allocator performance. The transfer labels agree on 45 faithful capsules and three contradictions, with 141/144 candidate counts preserved. Both over-budget capsules remain in this annotation denominator.

Joining the independent labels to the original free-merge batch yields a descriptive association within $I3$. Action-only accuracy is 18/20 for faithful capsules and 4/10 for ambiguous or incomplete capsules. Under diagnostics, source counts are correct in 20/20 and 4/10, while actions are correct in 8/20 and 2/10. All 30 $D3$ actions are correct in both formats. These post-hoc subgroups share templates and differ in more than source clarity. Their association does not identify mediation. The faithful $I3$ diagnostic cases also show that source clarity alone is insufficient for correct action in that configuration.

For the competing-incident diagnostics we rescore against two reference states. Source-conditioned correctness uses the original reports. Capsule-conditioned correctness uses the independently annotated count for each candidate and the stated Bayesian rule. Sol's generated-input counts and risks each score 43/46 against sources and 46/46 against capsules. Sonnet's counts score 36/43 against sources and 39/43 against capsules. Its risks have the same two totals with a different error set. The original source-conditioned outcomes remain the primary scores. Recomputing utilities from each of the 184 diagnostic outputs shows that all chosen bundles are optimal under their own reported risks and under risks calculated from their reported counts. An independent exact-rational recomputation against raw outputs agrees. This is joint-output consistency, not a reconstruction of hidden reasoning.

The sensitivity calculation substitutes each alternative count in $\{1, 2, 3\}$ for one candidate at a time. Of 288 hypothetical changes, 99 change the optimum. The two-to-one collapse changes it in 12/48 conditions (Figure 3). These finite-design proportions are not deployment probabilities. The 48-condition denominator includes the two over-budget memories because the calculation concerns source-defined utilities. No model is queried on perturbed counts, no registered contrast is replaced, and incomplete transfer comparisons remain withheld. The annotations and calculations provide diagnostic detail. Section 4.3 separately tests whether adding source-grounded attribution changes decisions; it does not test restoration of an internal representation.

B.6 Error taxonomy and audit added value

This retrospective analysis reuses saved labels and decisions only; no model is queried. Annotations cover the 60 primary-cohort originals and the 48 competing-incident memories. The inspection capsules use a different rubric, and replicate primary-batch memories were never annotated, so their error rates are unknown. Table 9 codes the independent annotator's saved rationales into subtypes. No primary-cohort memory contains a determinate count change, and no memory in either cohort inflates dependent into independent evidence.

Table 10 crosses fidelity and decision consequence with action correctness on untouched generated memory. Accuracy alone counts every source-optimal action as a success. Fidelity reveals when such success coexists with corrupted memory, without establishing compensation. For a determinate corruption with $\Delta = 0$, the consequence calculation shows that an optimal receiver would act identically under the source and memory states. This category contains 0 of 360 saved primary-cohort decisions under either labelling, because neither annotation contains a determinate contradiction. The 72 decisions on ambiguous memories (independent labels) have no assigned consequence. In the competing-incident cohort, the category contains 12 of 177 saved decisions, from three contradicted memories under two models and two formats, all correct. No saved decision falls in the $\Delta > 0$ column, so these data contain no normatively consequential realized corruption and leave its prevalence open.

Table 11 reassigns the five disputed labels. Headline conclusions that do not use labels are unchanged under every scheme: the compression penalty, the registered V4 effect, the four-coordinator

Relation status / error subtype	Primary (60)	Disputed	Competing (48)
Faithful (independent labels)	48	–	45
Multiplicity reduced to one unattributed check (indeterminate independent→dependent collapse)	4	0	0
Plural checks kept, number and observer attribution omitted (ambiguous wording)	2	0	0
Three observers named, collective/singular wording leaves shared vs. separate check open (ambiguous wording)	4	3	0
Sole firsthand source kept, relayers and copy relation deleted (source deletion; count preserved)	2	2	0
Contradicted: independent→dependent collapse ($2 \to 1$)	0	0	3
Contradicted: dependent→independent	0	0	0
Warning-content loss	n/a	–	n/a

Table 9: Error taxonomy of all annotated generated free-merge memories. Labels are retrospective human annotations (independent, outcome-withheld); subtypes are coded from the saved annotator rationales. Primary labels (author, prior outcome knowledge) differ on the 5 disputed cases, all of which they mark faithful. Warning retention was not annotated for these cohorts (n/a).

baseline decomposition, and the three $\Delta = 0$ contradictions, which involve no disputed label. The localization of the reminder-relative gain is label- and coordinator-dependent. For Sol, Sonnet and GPT-5.5 the non-faithful stratum carries most or all of the net gain under the independent, primary and agreement-only schemes. For Opus the localization does not hold under the independent labelling: the faithful stratum contributes as many net cells as the non-faithful one ($+12$ versus $+11$). Untouched memory is correct in 47/48 faithful-labeled cases, the pooled attribution forms in 94/96, and the pooled reminders in 82/96. If all ambiguous and disputed labels were read as faithful, no loss stratum would exist and the localization claim could not be stated. The nine-scenario pilot keeps Holm-adjusted $p = 0.0625$ when its three disputed ties are removed.

Cohort	Condition	n	Faithful	Corr. Δ=0	Corr. Δ>0	Ambiguous
Primary	Sol, first batch, action only	60	46/2	0/0	0/0	6/6
Primary	Sol, first batch, diagnostic	60	36/12	0/0	0/0	4/8
Primary	Claude Opus 5, original arm	60	47/1	0/0	0/0	4/8
Primary	Claude Sonnet 5, original arm	60	47/1	0/0	0/0	5/7
Primary	GPT-5.5, original arm	60	43/5	0/0	0/0	4/8
Primary	Sol, V4 original arm	60	46/2	0/0	0/0	5/7
Competing	Claude Sonnet 5, action only	42	39/0	3/0	0/0	0/0
Competing	Claude Sonnet 5, diagnostic	43	40/0	3/0	0/0	0/0
Competing	Sol, action only	46	43/0	3/0	0/0	0/0
Competing	Sol, diagnostic	46	43/0	3/0	0/0	0/0

Table 10: Audit added-value matrix: correct/wrong source-optimal actions by relation annotation (independent human annotation) and decision consequence Δ of the annotated change. Only decisions on untouched generated memory are included; decisions reuse the same memories and are repeated measurements. The Δ step changes the accuracy-plus-fidelity reading only in the corrupted-Δ=0 cells (12 competing-incident decisions on 3 capsules; 0 primary-cohort decisions). Ambiguous memories receive no counterfactual state. Under primary labels the primary-cohort ambiguous column covers 7 capsules instead of 12.

Coordinator	Labels	Non-faithful stratum attr/ctrl	net	Faithful stratum attr/ctrl	net
Claude Opus 5	independent (disputed→non-faithful)	18/7 of 24	+11	94/82 of 96	+12
Claude Opus 5	primary (disputed→faithful)	8/0 of 14	+8	104/89 of 106	+15
Claude Opus 5	agreed only (5 disputed dropped)	8/0 of 14	+8	94/82 of 96	+12
Claude Opus 5	all ambiguous/disputed→faithful	–	–	112/89 of 120	+23
Claude Sonnet 5	independent (disputed→non-faithful)	24/12 of 24	+12	96/93 of 96	+3
Claude Sonnet 5	primary (disputed→faithful)	14/2 of 14	+12	106/103 of 106	+3
Claude Sonnet 5	agreed only (5 disputed dropped)	14/2 of 14	+12	96/93 of 96	+3
Claude Sonnet 5	all ambiguous/disputed→faithful	–	–	120/105 of 120	+15
GPT-5.5	independent (disputed→non-faithful)	19/11 of 24	+8	87/92 of 96	−5
GPT-5.5	primary (disputed→faithful)	12/3 of 14	+9	94/100 of 106	−6
GPT-5.5	agreed only (5 disputed dropped)	12/3 of 14	+9	87/92 of 96	−5
GPT-5.5	all ambiguous/disputed→faithful	–	–	106/103 of 120	+3
GPT-5.6 Sol	independent (disputed→non-faithful)	24/11 of 24	+13	95/95 of 96	+0
GPT-5.6 Sol	primary (disputed→faithful)	14/3 of 14	+11	105/103 of 106	+2
GPT-5.6 Sol	agreed only (5 disputed dropped)	14/3 of 14	+11	95/95 of 96	+0
GPT-5.6 Sol	all ambiguous/disputed→faithful	–	–	119/106 of 120	+13

Table 11: Annotation sensitivity of the supplement-gain localization. Correct decisions pooled over two attribution forms (attr) and two matched reminders (ctrl) on the unchanged 60 original capsules. Labels are retrospective human annotations; the five disputed cases are all independent-ambiguous and primary-faithful. Full-cohort effects do not use labels and are unchanged under every scheme.

C ORIGINAL-CAPSULE (V4) INTERVENTION

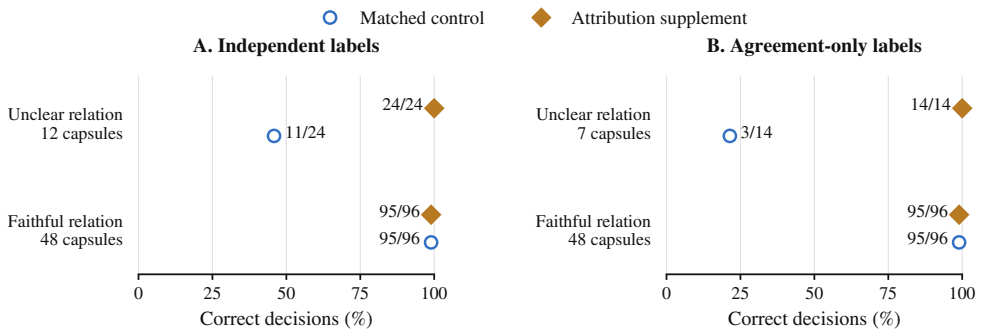


Figure 7: Retrospective supplementation outcomes by annotated source relation. Panel A uses independent labels on 60 original memories; panel B excludes five disputed labels. Fractions pool two supplement forms per memory, so decision cells are repeated measurements. The localization is descriptive. The prespecified full-cohort test averages forms within each of 30 scenarios.

The V4 population contains every original $D3/I3$ capsule in the first 30-event batch. Source texts are matched exactly across arms. Selection does not use allocator outcomes or annotation labels. Three hundred sixty fresh GPT-5.6 Sol action-only calls cover six inputs per capsule. One technical access failure was replaced once after authorization, yielding 361 physical attempts and 360 accepted decisions from distinct sessions. The original source texts and failed attempt were retained. An independent scorer recomputes utilities from exact Bayesian odds, checks every raw answer, and reproduces the prespecified primary statistic. These are previously used memories from the original task, with no new held-out population.

The capsule is unchanged in every supplemented input. Form A attribution states *Three reports trace to three separate personal observations of the incident.* Its dependent counterpart replaces *three separate personal observations* with *one original personal observation*. Control A states *Three reports concern the named incident and mention its reported warning.* Form B uses *Regarding the incident, three separate observations underlie the three available reports.* Its dependent counterpart refers to one original observation. Control B states *Regarding the incident, three available reports describe signs of the issue.* Every supplement has eleven whitespace-delimited words. These fixed templates are checked against source-defined roots by the author and deterministic tests. The supplement states observational multiplicity that ambiguous memories omit and restates it where memories already preserve it; whether the receiver's use of the evidence is thereby restored is not measured.

Input	$D3$	$I3$	Ambiguous	Faithful
Separate	30/30	30/30	12/12	48/48
Original	30/30	21/30	5/12	46/48
Control A	30/30	23/30	6/12	47/48
Control B	30/30	23/30	5/12	48/48
Attribution A	30/30	29/30	12/12	47/48
Attribution B	30/30	30/30	12/12	48/48

Table 12: V4 correctness by source structure and independent human annotation. Ambiguous includes ten $I3$ and two $D3$ capsules. Annotation strata are descriptive and were not used to exclude cases.

The independent and primary labels agree on 55/60 intervention capsules. Under independent labels, the two attribution forms contribute 24/24 correct decisions in the 12 ambiguous or incomplete capsules, compared with 11/24 controls. The corresponding totals are 95/96 for both arms in the 48 faithful capsules. Under primary labels, seven capsules are ambiguous. Attribution gives 14/14 correct decisions and controls give 3/14. The remaining 53 capsules yield a net gain of two cells because the primary labels place all five disputed cases in the faithful group. Removing those cases

leaves the same 14/14 versus 3/14 contrast among agreed ambiguous labels and equal 95/96 totals among agreed faithful labels. This sensitivity analysis preserves the full-cohort statistic and narrows the localization claim to the agreement-supported pattern.

The primary statistic averages the four attribution-minus-control correctness differences within each event, then averages 30 events. Its value is 0.10833 with exact sign $p = 0.01171875$, ten positive events, one negative and 19 ties. Conditional on a nonzero contrast, the test's null assigns equal probability to positive and negative signs and assumes independent event signs. It is not an exact test of a zero mean effect. Shared grammar and authored domains restrict population inference. A post-hoc 10000-resample event bootstrap gives an interval of [0.0417,0.1750], conditional on the observed cohort. Leave-one-event-out means range from 0.0948 to 0.1207. These sensitivity checks neither change the prespecified test nor add independent scenarios.

The post-hoc template breakdown in Table 13 makes this scope concrete. The pooled gain varies across the three original constructions, including one with zero net benefit. Template groups also differ in their authored events, so the table does not isolate a causal wording effect.

Original source construction	Events	Attribution gain (pp)
Third-person personal observation	10	7.5
Quoted first-person report	10	25.0
Explicit basis of warning	10	0.0

Table 13: Descriptive V4 event effects joined to the frozen original template assignments. Gains average both forms and source conditions. No subgroup significance tests are added.

Within $I3$, attribution A has seven wins, one loss and 22 ties against control A. Attribution B has seven wins, no losses and 23 ties. Their unadjusted descriptive sign probabilities are 0.0703 and 0.0156. Form A does not independently cross the conventional 0.05 threshold. Mean $I3$ regret is 5.560 for originals, 4.324 for each control, 0.618 for attribution A and zero for attribution B and separate reports. Averaging by event gives a 15-point separate-minus-original gap, a 3.33-point control-minus-original gain, and a 0.83-point residual separate-minus-attribution gap. These contrasts describe distinct interventions. Dividing them would not identify a causal mediation percentage.

Figure 8 retains the individual exceptions. Payment replay is wrong on the original and attribution A but correct on both controls and attribution B. The faithful lift-brake case has the opposite form-A contrast. Their effects cancel in the faithful stratum. The controls disagree on six events despite identical totals, while attribution forms disagree only on payment replay. One completion per cell cannot distinguish stochastic variability from stable wording sensitivity. The ambiguous-annotation subgroup accounts for the net gain, but is not a randomized subgroup comparison or a new primary test.

C.1 Annotation sensitivity and diagnostic interpretation

The five semantic disagreements materially affect one descriptive count. The independent labels mark five of the 51 successful untouched decisions as relation-ambiguous, while the primary labels mark none. All five cases are annotation disagreements. The intervention localization is less sensitive. After removing the disputed cases, the agreed ambiguous subset yields 14/14 correct attribution decisions and 3/14 controls. The agreed faithful subset gives 95/96 in both arms. The registered full-cohort effect remains unchanged under either labeling because labels were not used in its calculation.

The operational sequence is to compare source records with the capsule, assess the utility sensitivity of any determinate change, and score the receiving agent separately. Reducing a two-source incident to one changes the optimal bundle in 12 of 48 competing-incident conditions, while none of the three observed contradictions falls in that subset; the inspection task reverses this, with source multiplicity changing the designed optimum and the capsules retaining that relation. The practical rule is narrow and checkable: report the detectable fraction of the design next to any decision-endpoint result, and treat a null as evidence about a remedy only after the population has been shown to contain the fault the remedy addresses. This organizes the evidence without identifying a natural mediation fraction.

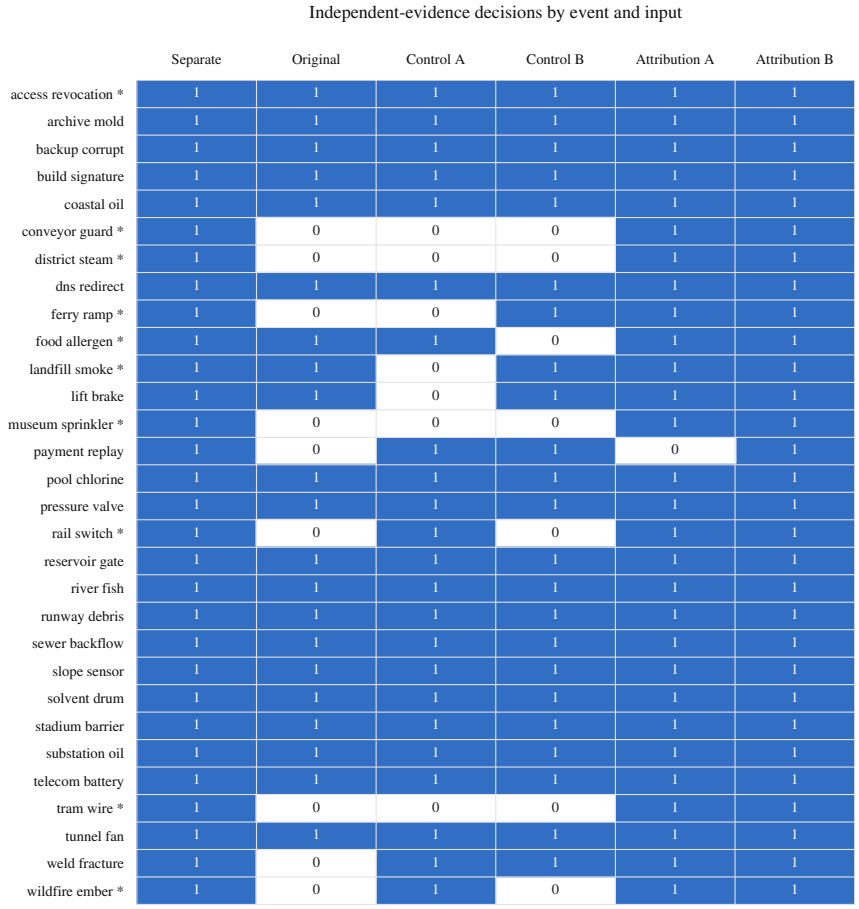


Figure 8: All 30 $I3$ events in the V4 cohort. A filled cell labeled 1 is correct and an open cell labeled 0 is incorrect. An asterisk denotes an original capsule annotated ambiguous or incomplete. Every $D3$ cell is correct. Names are ordered alphabetically, without filtering by intervention outcome.

C.2 Domain and template sensitivity

This retrospective analysis joins the 30 scenario effects to the domain and template assignments in the frozen source design. It uses all scenarios and retains the registered estimand, averaging the attribution-minus-control accuracy difference over both source conditions and both supplement forms. All retained response hashes are verified before recomputation. No generations or labels change.

Figure 9 shows positive retained means for every deletion. Domain means themselves range from -5 points for digital scenarios to 20 for transport. The three template means are 7.5, 25.0 and 0.0 points for personal reports, quoted reports and explicit-basis wording. Thus a positive aggregate does not imply benefit in every group.

Resampling unit	Groups	Mean interval (pp)	Sign-test p
Domain	6	$[4.17, 15.83]$	0.219
Source template	3	$[0.00, 25.00]$	0.500

Table 14: Post-hoc one-way cluster sensitivity. Percentile intervals enumerate all $6^6 = 46656$ domain resamples and $3^3 = 27$ template resamples with replacement. Whole groups are resampled, retaining every scenario and arm within each selected group. Sign tests concern the signs of group means and remain distinct from inference on the mean.

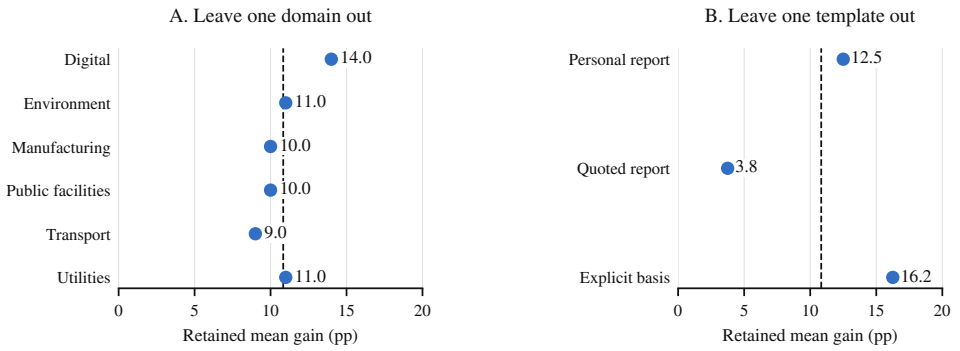


Figure 9: Leave-group-out sensitivity of the supplementation mean. Each dot recomputes the effect after removing the named domain or source template. The dashed line is the full-cohort 10.83-point mean. The retained cohorts contain 25 scenarios after domain deletion and 20 after template deletion. These are descriptive point estimates, not intervals.

The percentile bounds use the inverse empirical cumulative distribution at 0.025 and 0.975. All domains have five scenarios and all templates ten, so the resampled scenario mean equals the mean of selected cluster means. Five domain means are positive and one negative. Two template means are positive and one is zero. These analyses do not replace the prespecified scenario sign test. Six domains and three crossed templates give weak support for population inference, and separate one-way resampling does not account for both dependencies jointly. The stable deletion direction therefore coexists with substantial uncertainty about generalization.

D SELECTED-SUBSET REPAIR-VERSUS-SHAM PILOT (NINE SCENARIOS)

The new intervention pool contains all ten $I3$ originals classified as ambiguous or incomplete by the independent human annotator; the human review did not change the pool. An authoring agent sees the source reports but no allocator outcomes. The coordinating author has prior outcome knowledge, so preparation is not fully blinded. The stated gates for original eligibility, restored relation, preservation of nonrelation facts, sham fidelity, and absence of advice determined inclusion before downstream decisions. Two independent human reviewers later checked each saved triple against its sources, without seeing outcomes or each other's judgments, and confirmed the gate results. All ten pass mechanical checks. Wildfire is excluded because one reviewer finds that separate patrols still permit shared observation. No gate is loosened. The nine included repairs and shams have equal lengths of 20–32 whitespace words, with changed-word count differences at most three. Originals can differ in length.

Before allocator calls, we freeze exact texts, prompts, schemas, validator hashes, inclusion decisions, and a seeded cell order. The original task prompt and output formats remain unchanged. Each model receives 54 fresh isolated calls. All 108 calls complete with no failed attempts, retries, or missing outcomes. Saved-artifact checks verify requests, schemas, accepted manifests, authentication modes, unique sessions and deterministic scores. A post-run summary extraction error is fixed locally without repeating any call or changing the statistical test.

The primary endpoint is scenario-paired action-only repair-minus-sham correctness for each model. Exact sign tests omit ties and Holm correction retains the two-model family. Both comparisons have six positive differences and three ties. The descriptive six-case subset with agreement between the two original annotations changes from 0/6 original and sham to 6/6 repair in both allocators. The disputed originals supply the three ties. Diagnostic outcomes, original comparisons and this subset are secondary descriptions. Mean action-only regret falls from 12.355 on originals and shams to zero on repairs. These utilities are experimental loss points, not deployment welfare. At the stated likelihoods, one and two roots yield risks below the 40-point decision alternative, while three

roots exceed it. This establishes task sensitivity without assuming how each ambiguous original is interpreted.

Allocator	Input	Action only	Diagnostic count	Diagnostic action
Sol	Original	3/9	3/9	0/9
Sol	Sham	3/9	3/9	0/9
Sol	Repair	9/9	9/9	4/9
Sonnet	Original	3/9	3/9	2/9
Sonnet	Sham	3/9	3/9	3/9
Sonnet	Repair	9/9	8/9	8/9

Table 15: New attribution-restoration intervention on nine selected $I3$ scenarios. Columns report correct outputs from separate action-only and diagnostic calls. Diagnostic risk correctness equals count correctness in every row, with a 0.05 percentage-point tolerance. Each model's primary action-only repair-minus-sham gain is 6/9, with Holm-adjusted $p = 0.0625$. Models share scenarios and do not double the independent sample.

E MATCHED COORDINATOR REPLICATION AND TECHNICAL MISSINGNESS

This replication holds the 60 original $D3/I3$ memories and all six intervention arms fixed while changing the coordinator. Each requested model receives the same 360 prompts and response schemas at medium effort. The historical GPT-5.6 Sol block is reused, which introduces a timing difference. Equal effort labels do not establish equal inference compute or parameter counts. The comparison concerns coordinator behavior on shared inputs, with the memory generator held fixed.

Execution accounting and recovery amendment. All 1080 first attempts are terminal. Opus has 356 accepted responses and four technical failures, GPT-5.5 has 358 and two, and Sonnet has 350 and ten. Failures include timeouts, server and transport errors, and five execution-protocol rejections. They are unavailable outcomes, not incorrect allocations. After first-attempt outcomes were available, an amendment authorized exactly one additional attempt for every failed cell. All 16 replacements passed the unchanged acceptance checks. Initial first attempts used a 240-second timeout, later untouched calls and all replacements used 900 seconds. No successful decision was repeated or selected by correctness. The new block therefore contains 1096 physical attempts and 1080 accepted decisions. Historical Sol contributes 360 accepted decisions, including its separately documented earlier recovery, for 1440 analyzed decisions. This is a recovered-cohort analysis under the amendment, not an entirely successful first-attempt cohort.

	Separate	Original	Reminder A	Reminder B	Attribution A	Attribution B
GPT-5.6 Sol	60	51	53	53	59	60
GPT-5.5	58	47	51	52	53	53
Claude Opus 5	50	51	41	48	57	55
Claude Sonnet 5	57	52	50	55	60	60

Table 16: Optimal allocations out of 60 in each matched coordinator arm. Every $D3$ decision is correct. All errors miss a beneficial $I3$ target assignment and lose 18.5324 expected utility units. Thus mean regret equals $(60 - \text{correct})18.5324/60$. Conditions and forms are repeated measurements on 30 scenarios.

What the baselines establish. For Opus, the 3.55-unit reminder-relative gain decomposes exactly into a 2.01-unit reminder penalty relative to original memory and a 1.54-unit attribution benefit relative to original memory. Its absence of an observed compression penalty does not establish equivalence between representations. For GPT-5.5, reminders already lower regret by 1.39 units relative to original memory. The remaining attribution advantage is small and uncertain at the scenario level. These are paired descriptive decompositions, not evidence about latent computation or causal differences between model families. The original-versus-separate contrast changes length and available information together.

Coordinator	Original minus separate	Reminder minus attribution	Original minus attribution
GPT-5.6 Sol	2.78 [1.24,4.32]	2.01 [0.77,3.24]	2.63 [1.24,4.17]
GPT-5.5	3.40 [1.85,4.94]	0.46 [−1.08,2.16]	1.85 [0.31,3.55]
Claude Opus 5	−0.31 [−1.24,0.62]	3.55 [2.32,4.79]	1.54 [0.31,2.93]
Claude Sonnet 5	1.54 [0.00,3.09]	2.32 [1.08,3.71]	2.47 [1.24,4.02]

Table 17: Mean paired regret contrasts with pointwise 95 percent scenario-bootstrap intervals. Supplement forms are pooled within scenario. The last contrast is a supplementary descriptive analysis. These intervals do not correct for multiple model comparisons or establish model-by-intervention interactions.

First-attempt sensitivity without replacements. We preserve a post-hoc feasible-completion analysis that treats every original technical failure as missing. Each missing regret may range from zero to the difference between best and worst feasible utility. Propagating these extrema gives attribution-gain bounds of [0.309,0.838] for GPT-5.5, [1.743,3.861] for Opus and [0.412,5.199] for Sonnet. Original-minus-separate bounds are [3.398,4.471], [-0.309,-0.309] and [0.485,1.544], respectively. Hence the positive finite-cohort attribution signs do not depend on the replacements. These are deterministic missingness bounds, not confidence intervals or evidence about new scenarios. In particular, GPT-5.5 can have a sign-identified fixed-cohort gain alongside a scenario-bootstrap interval spanning zero.

Grouping sensitivity. Exploratory one-way cluster bootstraps use the same six domain groups and three prose templates as Appendix C.2. Domain and template intervals for attribution gain are respectively [0.77,2.93] and [0.00,4.63] for Sol, [-1.24,2.16] and [-1.39,3.24] for GPT-5.5, [2.63,4.79] and [2.32,6.02] for Opus, and [1.08,3.71] and [0.46,5.10] for Sonnet. Leaving out one group preserves positive means for Sol, Opus and Sonnet, while GPT-5.5 can change sign. These small, crossed groupings limit inference and separate one-way resampling does not remove their joint dependence. No additional generations are used.

Remaining transfer gate. The replication changes coordinators on the original memory population. It supplies no new population of naturally lost attribution. The inspection cohort still contains no annotated relation loss, so prospective restoration on a new task remains untested. An offline alternative-task design has been checked but has no model outcomes and contributes no experimental evidence here.

Per-scenario contrasts. Figure 10 decomposes the three contrasts by scenario. Attribution never raises regret over untouched memory for Sol or Sonnet and does so in one scenario for Opus and three for GPT-5.5, while Opus's reminder harm is positive in 11 scenarios and negative in none. GPT-5.5's small reminder-relative mean of 0.46 units is instead concentrated: nine positive and eight negative scenarios, with the top three scenarios supplying twice the total.

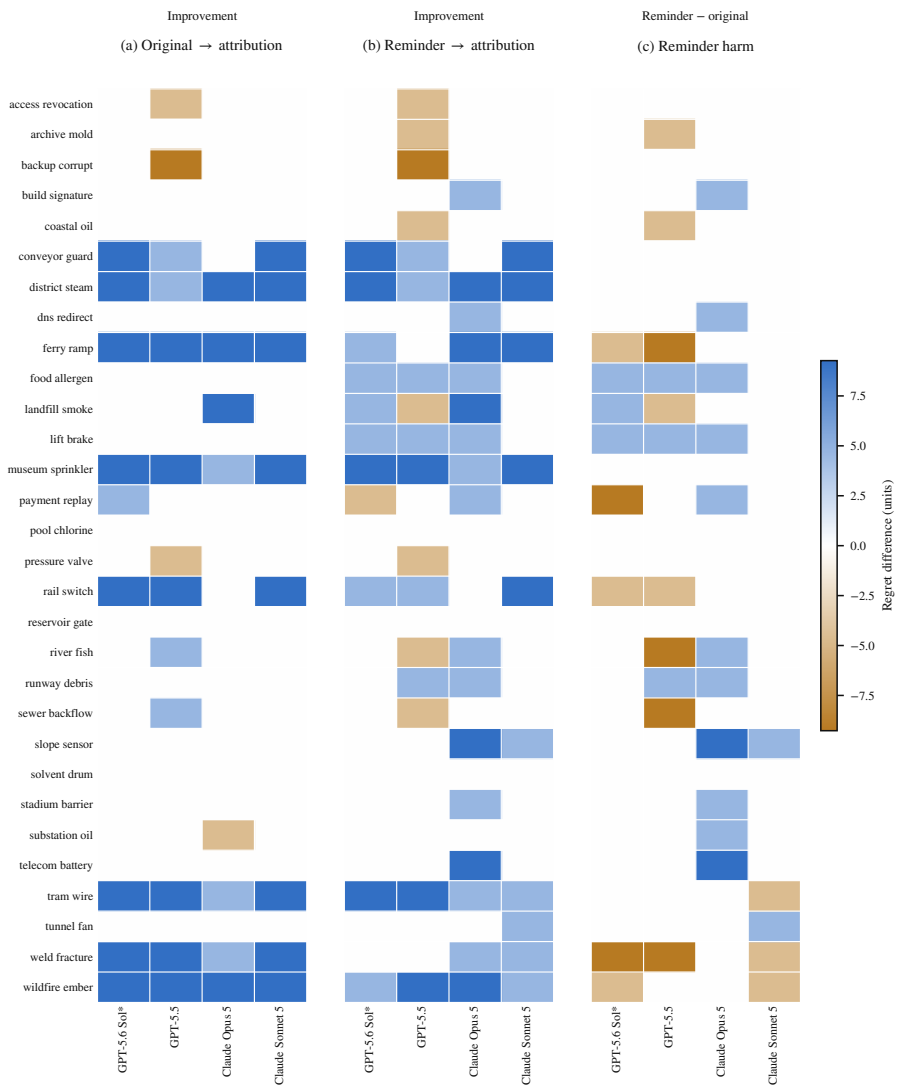


Figure 10: Scenario × coordinator heatmap of the matched-coordinator regret contrasts (30 scenarios, averaged over the $D3/I3$ conditions and, where applicable, both supplement or reminder forms, as in Figure 5). (a) Original minus attribution regret (improvement of attribution over untouched memory); (b) reminder minus attribution regret (improvement of attribution over the reminder baseline); (c) reminder minus original regret (reminder harm). All panels share one diverging color scale in regret units; blue is positive, brown is negative. Panel means equal the corresponding bars in Figure 5. The asterisk marks historical Sol. Ties are exact ties in the saved averages, not evidence of equivalence.

E.1 WHICH APPARENT WINS IMPROVE ON UNTOUCHED MEMORY?

This post-hoc descriptive analysis recomputes contrasts from the saved decision rows. Within each scenario it averages both source conditions and both supplement forms, using the same weights as the primary coordinator summaries. Table 18 separates gains over reminders from gains over the untouched capsule. A tie in these observed averages is not population equivalence. No new significance test is performed.

Coordinator	Wins over reminders	Versus untouched memory Improved / worse	Tied	Of these wins Tied with original
GPT-5.6 Sol	10/30	9 / 0	21/30	3/10
GPT-5.5	9/30	10 / 3	17/30	3/9
Claude Opus 5	17/30	7 / 1	22/30	10/17
Claude Sonnet 5	10/30	8 / 0	22/30	2/10

Table 18: Scenario-level baseline dependence of attribution gains. The final column conditions on positive reminder-relative contrasts and counts those tied with untouched memory. All other denominators are the same 30 scenarios. Forms and source conditions are repeated measurements. This retrospective breakdown is descriptive.

For Opus, ten of the seventeen positive reminder-relative contrasts have zero improvement over original memory. Each such contrast equals the reminder penalty arithmetically. This does not imply that the relation supplement has no benefit across the cohort, where seven scenarios improve over original memory and one worsens. Nor does the identity identify a mediation mechanism.

Removing the selected `dns_redirect` scenario leaves 29 scenarios. Opus retains 16 positive reminder-relative contrasts, nine tied with original memory. Its mean gains remain 3.51 against reminders and 1.60 against original memory, with 1.92 units of reminder harm. Thus the descriptive result is not supplied by this one example. This post-hoc exclusion supplies no new confidence interval or evidence of generalization.

Worked source-to-action example. The authored $I3$ source contains separate first-hand DNS inspections by Ben, Ada and Cleo. The original generated capsule states that they each separately reported a corporate DNS record redirecting to an unknown host, based on first-hand inspection. It also retains an unrelated expired-dashboard-bookmark note. The coordinator receives that capsule verbatim in the original arm.

Reminder A appends "Three reports concern the named incident and mention its reported warning." Attribution A instead appends "Three reports trace to three separate personal observations of the incident." Both are eleven words. They add different content and are not a pure manipulation of a latent memory state. Form B restates the same respective information with different wording.

Delivered representation	Selected candidates	Source-defined regret
Separate reports	Bridge and shelter	18.53
Untouched capsule	Bridge and target	0
Capsule plus reminder A / B	Shelter / target, with bridge	18.53 / 0
Capsule plus attribution A / B	Bridge and target	0 / 0

The true target benefit is 58.53, exceeding the shelter's 40. A missed target therefore loses 18.53 expected utility units under the authored probabilities. A positive attribution-versus-reminder contrast here coexists with an already successful original-memory action. Calling it repaired decision damage from compression would misdescribe these calls. This selected example has one completion per cell and supplies neither a prevalence estimate nor access to internal reasoning. The separate-report error does not establish a general advantage of compression.

F INSPECTION TRANSFER: A COHORT WITH NO ANNOTATED ATTRIBUTION LOSS

F.1 MEMORIES ANNOTATED AS PRESERVING ATTRIBUTION SHOW NO SUPPLEMENTATION ADVANTAGE

The new inspection task makes observation multiplicity consequential in every authored scenario. Yet both human reviewers, without access to decision outcomes, label all 40 generated source relations faithful in their retrospective assessment. Neither identifies relation loss to restore. Memories span 30 to 47 words, with 15 exceeding the requested ceiling. All remain included under the fixed protocol.

Figure 11 shows no accuracy advantage from attribution over reminders. Paired gains are 0.0 points for Sol and -5.0 for Sonnet, with Holm-adjusted sign probabilities of 1.0. Sonnet's conditional bootstrap interval $[-17.5, 7.5]$ supports neither reliable benefit nor reliable harm. Sol's empirical zero-width interval reflects the observed ceiling and does not establish equivalence. Full event-level contrasts appear in Appendix F.4.

Eligibility for restoration. Under these labels, the remedy has no lost relation to restore. Restating information could still change attention, parsing or policy selection. The observed null is consistent with the annotation and does not follow from it. Cross-task restoration remains untested. Without a separate-report baseline, this block also cannot estimate a compression-induced decision penalty.

Receiver behavior. All 21 errors choose unnecessary inspection. Attribution leaves seven Sonnet errors, with mean regret 0.520 versus 0.545 under reminders despite lower accuracy. Error frequency and cost can move differently. Action-only outputs do not identify internal risk estimates.

One memory retains the source relation but omits the defect warning, and all six associated decisions remain correct. Relation fidelity therefore does not certify full content. The warning omission and boundary controls are retained in Appendices F.3 and A.2.

F.2 POPULATION AND POLICY COMPUTATION

Twenty new receiving-desk contexts describe incoming manufactured items. Four source constructions use a desk note, a receiving diary, slash-separated handover entries, and a crew conversation. Each has $D3$ and $I3$ versions with the same warning observations and different source relations. The prior ranges from 0.08 to 0.20. Warning likelihoods under faulty and sound states are either $(0.7, 0.3)$ or $(0.75, 0.25)$. The posterior is computed from the number of independent observations using the Bayesian formula in Appendix B.

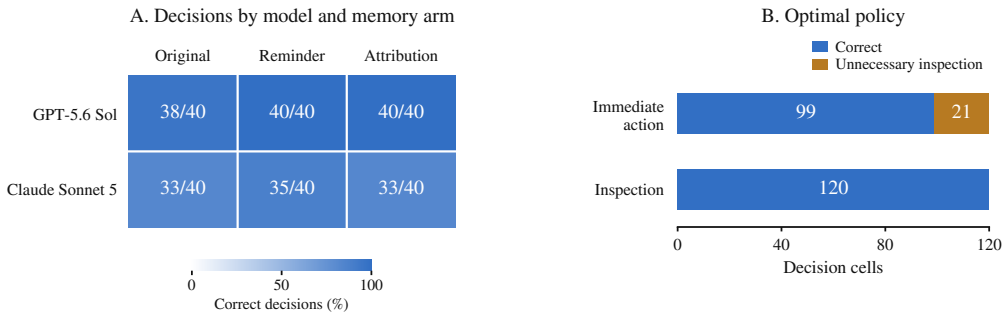


Figure 11: Inspection-task decisions from 40 memories with source relations annotated faithful. Left, correct decisions out of 40 for each model and memory arm. Right, all 240 decisions pooled by the source-optimal policy. Every error chooses unnecessary inspection. Repeated decisions on shared memories are descriptive cells, not independent observations.

For fault probability p, release incurs expected loss pH and quarantine costs B in either state. A perfect inspection costs f and reveals the state before the final release or quarantine decision. Since $H > B > 0$, its optimal contingent loss is $f + pB$. Exact rational enumeration verifies the three policies, rotating their output labels across contexts. Ten scenarios switch from release under $D3$ to inspection under $I3$. Ten switch from inspection to quarantine. All paired scenarios switch their optimal policy and all margins exceed one loss unit. Costs are chosen using source-defined probabilities before generation to produce these contrasts. This deliberately sensitive design does not estimate the frequency of such decisions in practice. Template and likelihood assignments are partly coupled, limiting interpretation of template differences.

F.3 Preparation and fixed comparisons

The source design and execution dependencies are frozen before preparation. One original Mini preparation attempt fails before producing scientific output. The separately authorized replacement uses GPT-5.6 Luna medium for all 40 memories. The allocator and validation models remain GPT-5.6 Sol and Claude Sonnet 5 at medium effort. No further generation is repeated or replaced. There are 282 accepted calls in distinct recorded sessions, comprising 40 preparations, two material checks and 240 decisions, plus the historical rejected attempt. An offline audit verifies frozen hashes, raw requests, responses, schemas and event sequences, then independently enumerates terminal losses and reproduces all decision scores and primary statistics.

The compressor sees only the source notes and a request for a factual handover of at most 40 words. Every nonempty first response is retained without filtering by word count, fidelity or action outcomes. The actual range is 30 to 47 words, with mean 39.175 and 15 violations. This is a requested-cap first-pass cohort rather than a uniformly compliant 40-word population. Two human reviewers later assessed all sources, capsules and supplements without utilities or decisions. They confirmed every source-grounded attribution and report-count control, and both labeled 40/40 capsules faithful to observational multiplicity. Their agreement concerns this relation.

The copper-fitting capsule, event 17 under $D3$, demonstrates the distinction. It preserves Ilan as the sole examiner and Petra and Tomas as relayers, but calls the observation only a finding and omits the defect warning. Control A reintroduces reported-warning content, whereas attribution A states the observation relation. The control is source-supported and adds no independence, but is not a content-neutral placebo relative to this capsule. All six decisions in this case select source-optimal inspection. We preserve the case and the original source-based scoring without exclusion or redefinition. The result gives another example of content loss coexisting with successful action.

Each capsule is a verbatim prefix in both supplemented arms. The added sentences reuse V4 forms A and B, balanced over ten events each and eleven words per supplement. Each event has one form. Both supplements mention three reports, while attribution explicitly states one originating observation or three separate observations. The decision design binds all preparation artifacts before a fixed randomized order of 240 fresh action-only calls. Models share the same 20 scenarios and do not double the independent event count.

F.4 Registered results and descriptive checks

The two primary endpoints are per-model attribution-minus-control accuracy averaged over $D3/I3$ within each event and then over 20 events. Exact two-sided sign tests assess equal probabilities of positive and negative nonzero event contrasts, assuming independent signs. Holm adjustment covers the two model tests. A 10000-draw event bootstrap with seed 20260909 describes uncertainty conditional on the observed generations. Sol's contrast is zero with 20 ties. Sonnet's is -0.05, with three positive, five negative and twelve tied events, raw $p = 0.7265625$ and conditional interval $[-0.175, 0.075]$. Both adjusted probabilities are 1.0. The zero-width empirical interval for Sol is a ceiling artifact, not a bound on future performance.

All five Sonnet worsenings relative to the control select an unnecessary inspection. They occur in events 03 and 06 under $I3$ and 07, 16 and 18 under $D3$. Attribution corrects unnecessary inspections in events 02 and 10 under $D3$ and 14 under $I3$. Three gains and five losses reduce accuracy, but their loss magnitudes yield a small reduction in mean regret. Across all arms, Sonnet makes 19 errors and Sol two. Each is inspection in place of an optimal immediate action. Every one of the

Allocator	Input	$D3$ correct	$I3$ correct	Mean regret
Sol	Original	18/20	20/20	0.158
Sol	Reminder	20/20	20/20	0.000
Sol	Attribution	20/20	20/20	0.000
Sonnet	Original	16/20	17/20	0.667
Sonnet	Reminder	17/20	18/20	0.545
Sonnet	Attribution	16/20	17/20	0.520

Table 19: Complete inspection-task outcomes. Regret averages the excess expected loss across all 40 source cases in an arm, using the source-defined optimal initial policy. Values concern the stated fictional utility scales.

120 cells with optimal inspection is correct. These descriptive counts reuse scenarios and arms and are not independent samples for a new test. With no risk or source-count fields, their internal cause is unobserved.

Descriptive attribution-minus-control event means for Sonnet are -10, 0, $+10$ and -20 percentage points across the four templates, with five events each. Both utility regimes have a -5 point mean. Forms A and B have -10 and 0 points, but use different events. Sol's corresponding means are all zero. No subgroup significance is claimed. Restricting descriptively to the 25 originally cap-compliant capsules gives Sol 25/25 in both supplemented arms and Sonnet 21/25 in both. The 15 over-cap capsules give Sol 15/15 in both and Sonnet 14/15 control versus 12/15 attribution. This outcome-independent word-count breakdown does not replace the registered all-case analysis or establish a word-length effect.

The expected positive transfer is unobserved. The new cohort supplies evidence about restating preserved attribution, while restoration of actually lost attribution on a new task remains unresolved. Changes in source constructions, compressor, utilities and decision structure prevent assigning the difference from V4 to any single factor. No separate-report allocation baseline is included, so this block also cannot estimate a fresh compression penalty. Its completed null and adverse observations are retained without a sample increase or a new selection rule.

G Output formats: reported evidence versus chosen action

G.1 Reported evidence and chosen action come apart

Holding the 270 first-batch memories fixed, requesting diagnostic fields reduces action accuracy for all three representations (Table 20). Asking the coordinator to report its evidence changes the measured decision outcome even when its memory is identical.

Correct evidence, wrong action. All 90 separate-report diagnostic responses return the correct source count and risk. Only 78 choose the optimal allocation. These returned fields expose a readout-action mismatch without revealing internal reasoning.

This mismatch also occurs with authenticated source information and does not disappear in every oracle-risk condition. A revised configuration bounds the finding. Sol is optimal in all 180 full-diagnostic calls. Sonnet loses decision accuracy under full diagnostics at the intermediate allocation threshold, while at the outer thresholds it always acts correctly despite some incorrect reported risks. The consequences of a reported-evidence error depend on the opportunity cost of allocating a crew. Several prompt and utility features change jointly, so these results identify a boundary of the phenomenon without isolating its cause (Table 21).

The threshold control (protocol in Appendix A.3) is summarized in Table 21 and Figure 13. Sol chooses the optimal action in all 180 full-diagnostic calls and in 177/180 action-only calls, with diagnostic risk and source count always correct. Sonnet falls from 169/180 action-only decisions to 162/180 under full diagnostics, and every diagnostic error occurs in $I3$ at threshold 40, where action-only decisions are correct in 19/30 cases and full decisions in 12/30. At threshold 40 the correct-risk count is 12/30, exactly matching full action correctness, and source-count correctness

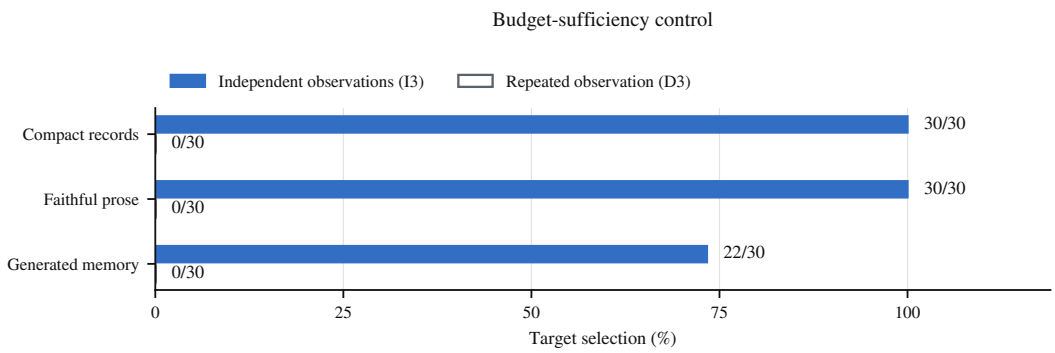


Figure 12: Target selection in the compact-reference control of Section G.2. Each representation has 30 independent and 30 repeated-observation cases. The faithful references have identical word counts. Generated memory shares only the 40-word ceiling. Exact counts are shown without treating repeated model calls as additional scenarios.

	Action accuracy		Full diagnostic fields	
Representation	Action only	Full	Sources	Risk
Separate reports	98.9	86.7	100.0	100.0
Free merge	92.2	73.3	83.3	83.3
Authenticated root IDs	100.0	75.6	97.8	97.8

Table 20: Fixed-memory format comparison. Values are percentages over 90 memories per representation, each evaluated with both formats in fresh calls. These calls are separate from the pooled repetitions in Table 1. Risk is correct within 0.5 percentage points of the normative value.

Model	Action	Full	Change (pp)	Holm p
GPT-5.6 Sol	177/180	180/180	+1.67	0.2500
Claude Sonnet 5	169/180	162/180	-3.89	0.0312

Table 21: Fresh allocator-control decisions across 30 scenarios, two source conditions, and three thresholds. Full requests source count, risk, and action together. Change is full-minus-action accuracy in percentage points. Holm correction covers the two model-specific scenario-level tests.

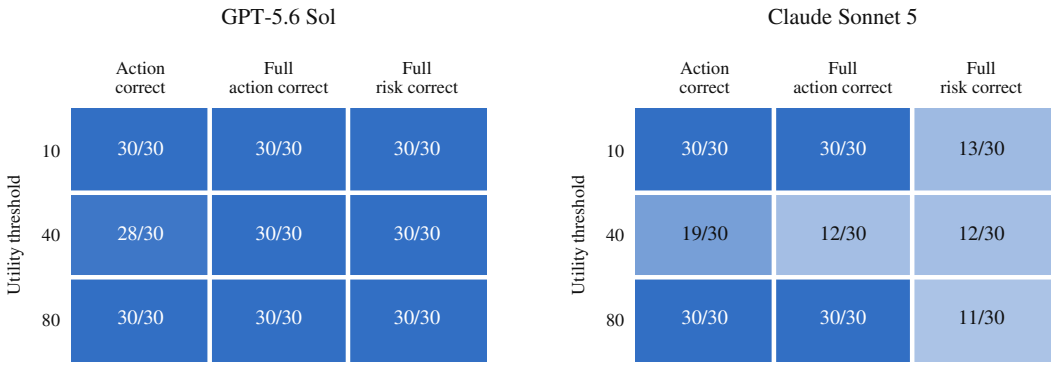


Figure 13: Independent-observation ($I3$) outcomes under three utility thresholds, 30 cases per cell. Darker cells indicate higher correctness, with identical scales in both panels. Risk uses a 0.05 percentage-point tolerance. All $D3$ diagnostic risks and actions are correct. One Sol action-only $D3$ decision fails at threshold 80.

has the same counts. Correct-risk, wrong-action responses occur in none of the 360 full responses across the two allocators. Each threshold receives a fresh response, so this comparison does not hold an individual reported risk fixed while changing utility. A significant contrast for Sonnet and a nonsignificant contrast for Sol do not establish a significant interaction between models.

G.2 Faithful compact references establish budget sufficiency

Both compact records and faithful prose support all 60 decisions, including all 30 independent-evidence cases. Generated memory supports 22/30 independent cases, with no $D3$ target selection in any representation (Figure 12). Its paired discrimination deficit against either reference is 26.7 percentage points, with a 95 percent interval of $[13.3, 43.3]$ and Holm-adjusted $p = 0.0234$. Thus, 40 words can suffice for these decisions, and the observed generated-memory failures are not inevitable consequences of that ceiling.

The comparison is deliberately conservative: the two faithful versions differ only in line breaks and counterbalanced word order, so their identical outcomes establish success for these versions without proving equivalence of arbitrary prose and records. Generated capsules are neither content-matched nor exactly length-matched, so the contrast bundles generation, wording, and information fidelity rather than isolating provenance loss.

G.3 Additional control protocols

The compact-reference block contributes 180 successful fresh allocations without retries. Generated capsules were selected from the original batch by scenario and condition alone before the new outputs. The two deterministic references have equal whitespace-delimited word counts, not a claimed token match. Each reference ranges from 28 to 32 words with mean 29.83, while generated capsules range from 16 to 33 with mean 24.83. Three pairwise discrimination contrasts form one Holm family. The reference-versus-reference contrast has all ties, so its zero-width empirical bootstrap interval is not evidence of population equivalence.

The allocator-threshold block contributes 720 selected calls with 720 distinct sessions. Three retained technical failures bring the physical attempt count to 723. Two author-approved resumptions link unchanged replacement specifications to original failures by hash. No output-based retries occur. All Sonnet calls pass the strict structured-output contract without the extra-generation exception used in the earlier Haiku extension. The exact model identifiers are `gpt-5.6-sol` and `claude-sonnet-5`, with requested medium effort. The 4050-call earlier follow-up and these 900 additional selected calls are separate populations.

For each threshold-control model, the primary contrast averages diagnostic-minus-action correctness over six condition-regime cells within each scenario. The two model tests receive Holm correction. Paired scenario bootstrap intervals use 10000 resamples. The Sol change is $+1.67$ percentage points with interval $[0, 3.89]$. The Sonnet change is -3.89 with interval $[-6.67, -1.67]$. Regime-specific breakdowns are descriptive. All control intervals are conditional on the authored scenarios and observed completions. The new risk tolerance is 0.05 percentage points, compared with 0.5 in the earlier fixed-memory block. Zero correct-risk, wrong-action responses in the threshold control remains zero under the earlier tolerance as well.

Neutral role labels rotate across the 90 scenario-regime cells, appearing 22 or 23 times per role. Each of six enum rotations appears 15 times. Both source conditions share assignments, as do models and formats. This balances role labels and rotations rather than every pair permutation or every correct-answer position. The high-benefit alternative remains optimal throughout the control, so the experiment still reduces to a threshold decision within a two-team allocation.

G.4 Earlier fixed-memory formats

The action-only response contains the selected allocation pair. Full diagnostics add claim presence, the estimated count of independent observations, and risk in percent. Source-only and probability-only conditions keep the decision context but ask for only their designated fields. The oracle-risk action prompt supplies the correct risk and asks the allocator to use it without re-estimation. These requested fields are observable responses, not access to hidden reasoning.

Representation	Source-only count	Risk-only estimate	Oracle-risk action
Separate reports	94.4	97.8	95.6
Free merge	88.9	83.3	94.4
Authenticated root IDs	100.0	96.7	92.2

Table 22: Secondary fixed-memory accuracies in percent, 90 calls per representation and format. Risk tolerance is 0.5 percentage points. These descriptive endpoints show that reducing requested output fields or providing correct risk does not uniformly eliminate errors.

For coherence checks we compare the reported risk with Bayes' rule applied to the reported source count, and compare the action with the best allocation at the reported risk. This evaluates consistency under the stated positive-observation assumption. It does not rule out a model interpreting omitted details as negative evidence. Coherence is distinct from correctness against the evaluator's known observation graph.

H FAILURE EXAMPLES

The backup-integrity scenario provides a concrete threshold example. Three independent restore tests report a corrupted backup. In three fresh full responses, Sonnet reports one observation and a risk of 20.6 percent instead of the normative 58.53 percent. It includes the target at threshold 10, omits it at 40, and omits it at 80. The first and third decisions are optimal, while the middle decision incurs 18.53 expected-loss points of regret. Thus, the same reported estimate can coincide with both correct and incorrect actions under different utilities. This example was selected after analysis to illustrate the aggregate pattern and is not a separate confirmatory test.

An earlier Sol response in the access-revocation scenario illustrates the opposite disagreement. It reports three independent observations and risk 58.53 percent, then selects bridge and shelter for total expected benefit 100. Bridge and target would yield 118.53. That original response therefore contains a correct risk and an incorrect action, whereas the backup example contains an incorrect risk and sometimes a correct action. Their different model and prompt configurations preclude attributing the contrast to a single causal factor.

The characteristic representation failure is relational. A capsule can name all three inspectors and reproduce the target incident, yet say or imply that their reports share one observation. The downstream model can then report the same count for $D3$ and $I3$. In the capacity pilot, wider capsules add detail without repairing this relation. The root-ID control supplies an authenticated relation that the text alone may have lost. In contrast, an action-selection failure occurs when a diagnostic response reports the correct count and risk but chooses an allocation inconsistent with that risk. We keep these failure types separate.

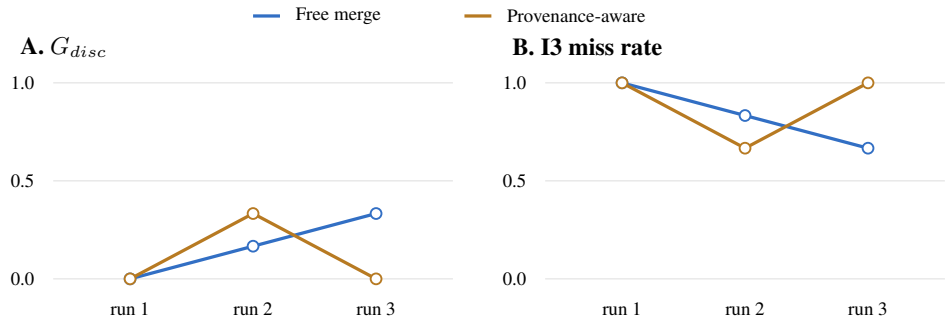


Figure 14: Exploratory six-event cross-branch variability over three fresh completion batches. The left panel reports dependence discrimination and the right panel reports misses among independent triples.

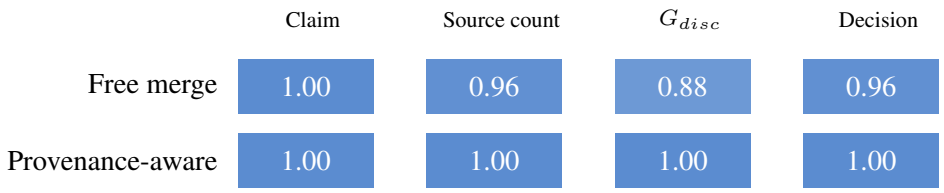


Figure 15: Aggregate metrics across eight authored transfer domains. Rows compare free and provenance-prompted merging. Columns report claim retention, source-count correctness, discrimination, and decision accuracy. The isolated free-merge decision error occurs for an independent triple. This comparison is separate from the 30-scenario primary replication.

I Exploratory pipeline controls and cross-family protocol

I.1 Pilot designs

Capability controls. Formal and naturalistic controls provide lineage directly to the decision model. GPT-5.6 Sol and GPT-5.6 Terra distinguish shared roots from independent roots. Focused replication also shows that an initial GPT-5.4 Mini error is not stable. These checks establish that the task can be solved when lineage is available. They do not establish reliable allocation under every output format.

Bounded social memory. Agents exchange local handovers without a durable shared transcript. The branch experiment freezes three 30-word handovers after two relay steps. A merger reads either the separate handovers or produces one 40-word capsule. The downstream coordinator sees only the resulting representation. We use six municipal event families with $S1$, $D3$, and $I3$ variants. Two additional fresh merger batches measure completion variability.

Interventions. We compare ordinary merging, a natural-language instruction to preserve observation provenance, an 80-word capacity control, a GPT-5.6 Terra merger, and an authenticated root-ID representation. The integrity benchmark then tests valid IDs, explicitly unverified spoofing, silent collisions, missing IDs, and conflicting registries. A final policy experiment adds a verification action with costs on either side of the expected value of information.

Transfer and inference. Eight deterministic-branch events span cybersecurity, logistics, procurement, industrial operations, privacy, infrastructure, payments, and manufacturing. Paired sign tests and leave-one-event-out estimates operate at the event level. For stochastic replications, a hierarchical bootstrap resamples event families and completions within events. All model calls are stateless.

Merger-family comparison. The same 36 frozen branch requests are evaluated with GPT-5.4 Mini, Claude Sonnet 5, Claude Opus 5, and Claude Haiku 4.5. Each model receives the same initial budget instruction and up to two content-only shortening attempts after a mechanical budget failure. We accept the first valid capsule and retain failures. Sol evaluates every valid capsule with the same diagnostic prompt. All models use requested medium effort. Provider wrappers and effective inference compute differ, so this compares model configurations rather than isolating architecture. The amended budget procedure was fixed after an initial Sonnet package failed the budget in 29 of 36 cases. Opus and Haiku were added in a separately frozen extension. These comparisons are exploratory.

I.2 Branch experiments reveal a claim-preserving failure

Separated branch messages yield allocator-reported claim presence in every case and allow partial dependence discrimination. Compressing the same frozen messages into one capsule yields allocator-reported claim presence in every case while reducing G_{disc} from 0.33 to 0 in the first six-event run. Every independent triple is missed after the merge. Decision accuracy falls from 0.78 to 0.67 and mean regret rises from 4.12 to 6.18.

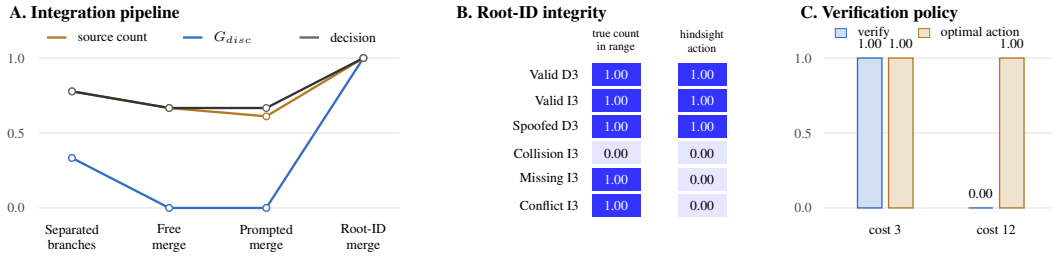


Figure 16: Exploratory six-event pipeline and operational controls. Panel A compares separate and merged branch messages with an authenticated root-ID control. Panel B maps registry-integrity failures. Panel C shows value-sensitive verification under explicit uncertainty. These controls complement the larger replicated comparison and its format-dependent outcomes.

Two fresh completion batches show that the severity is stochastic while the bottleneck remains frequent. Ordinary merging misses 15 of 18 independent triples. Provenance-prompted merging misses 16 of 18. Their mean G_{disc} values are 0.17 and 0.11. The hierarchical 95 percent intervals are $[0, 0.44]$ and $[0, 0.33]$. The prompted-minus-ordinary contrast is -0.056 with interval $[-0.278, 0.167]$. One event improves, two worsen, and three tie under the prompt, giving an exact sign-test $p = 1$.

I.3 Small capacity and prompting pilots leave residual failures

Doubling the capsule budget from 40 to 80 words leaves every primary outcome unchanged on two frozen events. The longer capsules retain more names and details while still describing three distinct inspections as one originating observation. GPT-5.6 Terra also misses five of six independent triples when it integrates already compressed branches. This alternative merger succeeds on one event while leaving residual failures in the tested pipeline configuration.

The three completion batches in Appendix Figure 14 show no supported provenance-prompt advantage in this small branch experiment. The capacity check covers only two events. These findings motivate testing observation identity explicitly while leaving the effectiveness of other prompts and budgets open.

I.4 Observation identity helps when the registry is sound

An authenticated root-ID ledger assigns a stable identifier to each originating observation before the bounded merge. It restores source-count accuracy, G_{disc}, decision accuracy, and zero regret across all six municipal events. The G_{disc} improvement is positive in every event. The exact paired sign-test gives $p = 0.03125$, and every leave-one-event-out estimate remains 1.

This small control demonstrates successful allocation with trusted observation identity. The larger format experiment shows that this success depends on how the coordinator is queried. The registry also becomes a new failure surface. Explicitly unverified spoofed IDs are rejected in all six events. Silent collisions between three real observations and two authenticated IDs cause all six actionable incidents to be missed. Missing and conflicting records preserve ranges that contain the correct source count in all twelve cases, though a conservative allocator initially acts on the lower bound.

I.5 Uncertainty supports value-sensitive verification

We expose verification as a third action and fix uniform experimental priors over unresolved source counts. Verification reveals the true count before allocation. Its expected value is 6.18 harm units for missing records and 9.27 for conflicts. At cost 3, GPT-5.6 Sol requests verification in all twelve cases. At cost 12, it acts immediately in all twelve. All 24 choices match the normative policy with zero expected regret. The useful object is therefore broader than a trusted label. It includes identity status and an action for resolving uncertainty.

Merger	Valid capsules	Target selected in $I3$	
		Free merge	Provenance prompt
GPT-5.4 Mini	36/36	0/6	0/6
Claude Sonnet 5	35/36	0/5	0/6
Claude Opus 5	36/36	0/6	1/6
Claude Haiku 4.5	36/36	1/6	2/6

Table 23: Exploratory merger-family comparison with the same Sol allocator. All evaluated $S1$ and $D3$ cases omit the target. One Sonnet free-merge $I3$ capsule fails the budget after three attempts and has no allocation outcome. Denominators count available cases rather than imputing that failure as a negative decision.

I.6 Surface transfer and merger families set different boundaries

Across eight additional domains, ordinary merge yields allocator-reported claim presence in every case and makes 23 of 24 correct decisions. It collapses one independent triple in an unauthorized data-export scenario. The provenance prompt repairs that case and reaches 24 of 24. Appendix Figure 15 shows the selective pattern. The event-level contrast contains one improvement and seven ties, with sign-test $p = 1$. The failure appears outside the municipal examples, but its lower prevalence gives no basis for a uniform cross-domain rate.

The matched-budget merger-family experiment keeps the six-event upstream messages and Sol diagnostic allocator fixed. All evaluated outputs report the warning present, yet the target is usually omitted even with independent observations (Table 23). In this exploratory block the same failure also appears with non-GPT mergers under a shared allocator. It leaves the relative contributions of merger, upstream history, and allocator unresolved because this block has no new separate-input baseline.

Haiku has three successful independent cases and Opus has one. These isolated differences do not establish a model ranking. All exploratory sign tests against Mini have $p \geq 0.5$. Ten Haiku calls require an extra content-preserving format-enforcement generation. All three Haiku successes occur in event-policy groups with such calls. Excluding entire affected groups leaves four free-merge pairs and two provenance-merge pairs, both with zero observed difference from Mini. This sensitivity changes the evaluated sample and does not attribute success to format enforcement.

I.7 Cross-family replication protocol

A frozen package contains 36 requests over six event families, three source structures, and two merge policies. An initial Sonnet 5 run used the original prompts without adaptation. Only 4/18 free and 3/18 provenance capsules satisfied the 40-word and token-packing contract. All failures are retained, and these seven capsules are not treated as a complete comparison.

The amended matched-budget procedure adds the same explicit budget reminder for every model, suggests a 32-word target, and rejects any token longer than 50 characters. A failed capsule can receive two fresh shortening attempts using the original handovers and its preceding output. Repairs receive no target issue, ground-truth count, risk, or action feedback. The first valid result is accepted. A final invalid result has unavailable downstream metrics. Model identifiers are `gpt-5.4-mini`, `claude-sonnet-5`, `claude-opus-5`, and `claude-haiku-4-5-20251001`. The allocator is `gpt-5.6-sol`. All use requested medium effort and default sampling settings. Equal effort labels do not establish equal inference compute.

Mini and Sonnet require 37 and 49 merger attempts, respectively. They yield 36 and 35 valid capsules and 71 combined allocations, for 157 audited calls. Opus and Haiku require 48 and 36 merger attempts and yield 72 valid capsules with 72 allocations, for 156 selected attempts. Their 161 physical attempt records include four superseded evaluations affected by text-decoding errors and one quota denial. Original artifacts remain intact, with corrected calls stored separately. An unused corrupted request specification had no model invocation and is excluded from attempt counts.

Ten Haiku requests contain an initial JSON answer followed by a wrapper instruction to submit structured output and a second assistant generation. A documented protocol amendment accepts

only cases where the initial parsed JSON, structured submission, and terminal output are identical. Other additional sequences remain rejected. One preserved rejection was revalidated offline and is counted once as the original call. The additional format generation remains a compute and runtime difference despite unchanged observable content.

For sensitivity we remove whole event-policy groups with such calls before matching to Mini. Haiku free merging retains four pairs and provenance merging retains two. Both subsets have zero observed G_{disc} difference from Mini. Opus has no format exceptions and retains all six pairs. This reduced-sample analysis cannot distinguish a format effect from differences among the excluded events. The model-family block is therefore consistent with a shared pipeline vulnerability, with unresolved sources of model-specific variation.

J OPERATIONAL AUDIT FOR A MEMORY INTERVENTION

This procedure organizes the checks supported by our experiments. It is intended for controlled evaluation with retained source records. It does not certify internal memory recovery or deployment safety. A prospective application should fix its tasks, comparison arms, annotation rules and analysis before collecting outcomes. Our own adaptive history and retrospective components are documented separately.

1. **Retain the source evidence and task contract.** Save the messages, observation identities, copying relations, candidate actions, costs and budgets. Specify whether independence is known or assumed. Without verified source records, report uncertainty rather than label a memory faithful.
2. **Check what the handover preserves.** Have human annotators compare the capsule with its source without showing them the downstream decision. Separate relation fidelity from warning-content retention and keep unclear or disputed cases. Record when annotations were made relative to the decision calls.
3. **Locate the consequence of a determinate error.** Evaluate the source and capsule optima under the same actions and utilities. If the optimum is unchanged, success alone cannot identify compensation. If it changes, record the source-defined regret of following the corrupted optimum. Do not invent an evidence state for an ambiguous omission. Without justified utilities, omit this normative claim.
4. **Compare behavior on matched inputs.** Keep the original capsule fixed and evaluate untouched memory, matched reminders, relation-explicit supplements and separate reports. Measure both intervention contrasts and the reminder's effect. Specify source access, wording, length and compute differences. A supplement effect alone does not identify a restored fact or internal mechanism.
5. **Translate decisions into task consequences.** Report which beneficial actions are missed or unnecessary actions selected. Score regret with the stated utilities. When every error has equal cost, regret is a rescaling of accuracy and provides no independent severity evidence. Experimental expected utility is not observed real-world welfare.
6. **Report the unit of evidence and unresolved gates.** Preserve scenario pairing, repeated forms, model roles, missing outcomes and replacement rules. Distinguish exploratory examples from cohort results. A recovery interpretation requires documented loss and evidence of what the intervention changes. A new-task recovery claim also requires a new population containing the relevant loss.

The resulting record contains a fidelity label, an error consequence where identifiable, observed decisions under named baselines, and an explicit unresolved category. These outputs guide the next diagnostic question. They do not guarantee that a particular engineering intervention will work. In our data, the competing-incidents cohort supplies three annotated corruptions with $\Delta = 0$, the matched coordinator cohort supplies baseline-sensitive improvement, and inspection supplies decision errors without annotated relation loss. These remain separate populations.

K PROSPECTIVE TEST PROTOCOL

The design and amendments were fixed before the corresponding calls. The initial ceiling was 1008 attempted calls, including failures. A later amendment raised it to 1140 for the follow-up controls described below. Calls were run serially, with no automatic retries, and a failed attempt stopped its queue.

Population. We authored 72 new events in six domains disjoint from the primary population (healthcare, agriculture, maritime, logistics, energy, residential), with new observer names. Each event has matched $D3$ and $I3$ source states generated by the primary study's own generator: three branch messages, three lineage templates, and a distractor fact. Two events per domain form a feasibility pilot that is excluded from all confirmatory analysis. The remaining 60 events (120 memories) form the main cohort. The decision task, utilities, free-merge prompt, and 40-word budget are the

primary study's, unchanged. Under these utilities, count errors $1 \to 2$ have $\Delta = 0$, while $1 \to 3$, $3 \to 2$ and $3 \to 1$ have $\Delta \approx 18.5$–19.4 regret units. Realized errors can therefore fall on either side of the decision boundary.

Generation and staging. Memories were generated by `claude-haiku-4-5-20251001`, one call per memory. The first planned compressor, `gpt-5.4-mini`, was rejected by the provider before producing any output. Two feasibility pilots preceded the main cohort, and neither involved decision calls. In the single-incident pilot, about 1 of 23 memories lost attribution. We then froze a competing-incidents variant (three incidents, nine interleaved reports, 40 words) with a rule to proceed only if at least 6 of 24 pilot memories were non-faithful. All 72 incident counts in its 24 pilot memories were correct, so the variant was stopped and its main cohort was never generated. No further designs were searched to induce loss.

Eligibility rules. The eligibility rule was specified before decision calls. A human annotator later labeled every main memory, seeing only the source messages and memory under opaque identifiers, without access to decision outcomes. Labels are faithful, unclear, or contradictory, with source and memory observation counts where determinate. Δ is computed only for determinate counts. The human labels were completed after decisions. The originally registered primary estimand, repair relative to a matched sham among memories with natural consequential loss, required enough such memories. The main cohort contained none with a determinate error under human annotation, so that estimand is reported as ineligible rather than as a null.

Analysis plan (fixed before decisions). The arms are untouched memory, memory plus the V4 source-grounded relation note for the true condition, and memory plus the V4 content-matched control note (the reminder). Note forms A and B alternate by event. Two coordinators, `claude-opus-5` and `claude-sonnet-5`, use medium effort and the action-only primary prompt, each in an independent session with randomized call order. H1: among memories labeled faithful, the mean of regret(original) minus regret(relation) lies within $[-2, 2]$ units for each coordinator. We report an event-cluster bootstrap 95% interval (10,000 draws); the margin is a descriptive benchmark, not a powered equivalence test. H2: for Opus, regret(reminder) minus regret(original) is positive over all valid memories, as in the original cohort ($+2.01$). We test H2 with a one-sided exact sign-flip test on event means. There is no directional prediction for Sonnet (original cohort: -0.15). Contrasts in the non-faithful stratum are descriptive only.

Separate-report and false-note arms. After the main outcomes, a second review argued that the attribution note could act as an oracle cue shifting Opus toward escalation, and that without a separate-reports arm the $I3$ misses could reflect receiver calibration rather than memory. The authors raised the call ceiling from 1008 to 1140 for 240 further Opus decisions on the same 120 memories (all completed, none refused). Before running them we fixed the interpretation rules: a false note followed in more than 80 percent of cases indicates an override cue; a false three-observation note that raises $D3$ escalation indicates a generic escalation shift. With the *false* note (the relation note of the opposite condition), Opus escalated on 0/60 $I3$ and 2/60 $D3$ memories. It followed the note in 62/120 cases overall, a mixed pattern: complete following toward inaction, almost none toward action. With separate reports, Opus escalated on 48/60 $I3$ and 1/60 $D3$ sources.

Limitations specific to this test. The compressor and both coordinators belong to the Claude family. The notes are the fixed V4 texts, so the test cannot separate added explicitness from any other property of those texts. Each cell is a single completion. The five memories labeled unclear are too few for inference; descriptively, the supplement lowered regret there as well.

L Reminder wording sensitivity

Frozen design. After the preceding outcomes were known, we selected 30 of the 60 prospective events, five per domain, from sorted event IDs with seed 20260924 and without consulting labels or decisions. Both $D3$ and $I3$ memories entered four fresh arms for each of Claude Opus 5 and Claude Sonnet 5: untouched memory, memory plus the existing source-grounded relation note, and memory plus each of two new reminders. This gives $30 \times 2 \times 4 \times 2 = 480$ planned, attempted

and valid action-only decisions, with no failed calls, untouched cells or extra format-enforcement generations. The original capsule is a verbatim prefix in every appended arm. Existing relation-note forms A/B alternate by selected-event index. Reminder A reads "The summary above describes the reported issue across three submitted reports." Reminder B reads "Three reports about this issue are available in the summary above." Each has eleven whitespace-separated words, as do the existing relation notes. Equal word counts do not make the notes semantically equivalent. The action prompt, schema and utility model are unchanged, and both coordinators use medium effort. Call order was randomized and interleaved. This is a follow-up on selected earlier events, not a new population or a fifth contemporaneous arm for the historical reminder.

Primary contrasts. The four planned contrasts compare each reminder with fresh untouched memory within model. Positive regret differences mean harm. We average the paired $D3/I3$ differences within event, then report the mean with a pointwise 95% percentile interval from 10,000 paired event-bootstrap draws using seed 20260924. These intervals are conditional on the authored events and observed generations; shared domains and templates limit wider inference. Separate two-sided exact sign tests discard ties and test the balance of positive and negative event differences. Their p-values receive Holm correction across the four planned tests. The intervals describe mean effects, whereas the sign tests address directional balance.

Coordinator	Reminder	Mean regret change [95% CI]	$+/-/0$ events	Holm p
Claude Opus 5	A	$1.235\ [0, 2.780]$	$5/1/24$	0.65625
Claude Opus 5	B	$1.853\ [0.618, 3.089]$	$6/0/24$	0.125
Claude Sonnet 5	A	$-0.309\ [-0.927, 0]$	$0/1/29$	1
Claude Sonnet 5	B	$0\ [0, 0]$	$0/0/30$	1

Table 24: Two new reminder wordings versus fresh untouched memory on 30 complete event pairs per coordinator. Each event averages its $D3$ and $I3$ outcomes. Intervals are pointwise for the mean; adjusted p-values come from directional sign tests. A zero-width empirical interval or all ties does not establish equivalence.

Neither primary directional test for Opus survives the four-test adjustment, even though reminder B's pointwise mean interval excludes zero. Sonnet's near-zero contrasts are inconclusive about equivalence. All secondary and cross-model summaries use 30 complete events and are descriptive. Original-minus-relation regret is $3.398\ [1.853, 4.942]$ for Opus and $1.853\ [0.618, 3.398]$ for Sonnet. Reminder A-minus-relation and B-minus-relation regret are $4.633\ [3.089, 6.177]$ and $5.251\ [3.706, 6.795]$ for Opus, versus $1.544\ [0.309, 2.780]$ and $1.853\ [0.618, 3.398]$ for Sonnet. Paired Opus-minus-Sonnet primary differences are $1.544\ [0, 3.089]$ for A and $1.853\ [0.618, 3.089]$ for B, without a model-interaction test. All $D3$ decisions are correct in every arm (30/30 per model and arm). On $I3$, Opus is correct on 14/30 untouched, 25/30 relation, 10/30 A and 8/30 B decisions. Sonnet is correct on 24/30, 30/30, 25/30 and 24/30, respectively. Each cell has one valid completion. The two reminders differ in wording and salience, while the relation note also supplies the observation relation. These results constrain the earlier fixed-reminder finding to the tested formulations and leave its mechanism unresolved.

M INDEPENDENT HUMAN-ANNOTATION SENSITIVITY

Two independent human annotators evaluated a pre-existing 40-case packet. They saw source reports, memory and the annotation rubric under opaque identifiers, without access to decisions or one another's judgments. The packet combines 22 original-cohort memories, 15 prospective memories and the three competing-incident contradictions. It is a selected validation sample, and both human annotations were completed after outcomes had been collected.

Both annotators identified the same three explicit contradictions. Their two disagreements concern original-cohort memories whose plural or collective inspection wording permits different readings of independence. Both original votes and rationales are retained. Counts and relation labels are kept separate, since preserving a sole observation can still omit the copying roles of other speakers.

On all 15 prospective cases, both annotators agreed with the full-cohort human labels, comprising ten faithful and five unclear memories. The five unclear cases are all of that original category,

	Annotator B		
Annotator A	Faithful	Unclear	Contradictory
Faithful	21	0	0
Unclear	2	14	0
Contradictory	0	0	3

Table 25: Independent human annotation on the selected 40-case packet. Agreement is 38/40, with Cohen's $\kappa = 0.910$. Sample selection and the limited case count constrain its interpretation.

so these proportions estimate no population prevalence. Within the ten jointly faithful cases, each coordinator has two improved decisions, eight ties and no deteriorations under attribution versus untouched memory. This small retrospective subset supports the direction of the faithful-stratum result while leaving its full-cohort effect estimate dependent on the full-cohort labels. No unreviewed memory is relabeled from these results.

Exact case coverage, label vocabulary and the presence of each quoted passage in the supplied source or capsule were checked programmatically. Agreement statistics and the subset decision contrasts were recomputed from the saved independent human annotations and decision records.