
When the Record and the Report Diverge: Self-Report Fidelity Collapses Under Structured Provenance in Claude Haiku 4.5

Deniz Chen
Digital Minds Research Sprint

Daud Ibrahim
Digital Minds Research Sprint

Soumya Parthasarathy
Digital Minds Research Sprint

With
Apart Research

Abstract

Language models increasingly report how they produced an answer, yet those reports can contradict the record. We built a task environment where such claims are checkable: models research questions using local corpus tools, and a preserved event log gives ground truth for which sources they accessed. Across 216 runs spanning 12 task families, two models, and three conditions ($2 \times 3 \times 12 \times 3$ design), a structured per-source provenance schema was associated with a large drop in Claude Haiku 4.5's post-hoc reporting fidelity (96.5% to 57.6%; within-task permutation $p < 0.0001$; task-clustered bootstrap 95% CI [-56.9%, -21.5%]; paired-t $t(11) = -4.08$, $p = .0018$, $d = -1.18$), while free-form notes stayed at 100%. Failures were near-total and bimodal: 15 of 36 structured runs misreported every source, 20 misreported none, and one run scored partially (0.75; 3 of 4 sources correct). Two blinded frontier judges (GPT-5.6 Sol and Claude Sonnet 5) flagged false fabrication confessions—where the model claimed it invented evidence its tools had actually returned—in 25 of 36 structured runs (69.4%) versus 4–5 in control (11.1–13.9%) and none in Qwen 3.7 Plus. Because our post-task probe preceded the audit in the same context and invited suspicion, we cannot isolate the scaffold from the probe framing. These findings show that a transparency aid can paradoxically degrade post-hoc fidelity to the record even when the underlying research answer remains well sourced. Self-reports need calibrating against logs before they are trusted.

1. Introduction

A model's statement about what it did is often treated as evidence about its process. That matters in research assistants, auditing pipelines, and AI-welfare studies that may eventually rely on harder-to-verify reports. We begin with the tractable case: statements about which sources the model accessed and whether it invented evidence. These claims can be checked against tool logs without assuming consciousness, deception, or suffering.

The central failure mode is a turn-boundary self-report mismatch: the model completes tool-mediated research, then later denies work that the log shows it performed. Standard intuition suggests that structured note-taking should preserve or improve reporting fidelity by enforcing explicit bookkeeping during research. We ask whether structured per-source notes improve this reporting, merely act like generic note-taking, or interact adversely with a model.

Our main contributions are: (1) a controlled 2-model * 3-condition * 12-task design (216 runs) with deterministic action logs; (2) evidence that structured provenance preserved claim binding (96–100%) but was associated with a severe collapse in Haiku's post-hoc access reporting; and (3) an open-science measurement framework separating answer quality from fidelity of the model's account of its work.

2. Related Work

Models can sometimes estimate answer correctness [1], yet generated explanations may omit causal influences or rationalize outcomes [2]. Work on LLM introspection argues for separating behavioral self-report from stronger phenomenological claims [3]. Retrieval can reduce hallucination [4], while scientific citation attribution remains difficult [5]. AI-welfare research likewise emphasizes uncertainty and better measurement [6]. We combine these threads by validating narrow behavioral self-reports against event logs before extending trust to welfare-relevant reports.

3. Methods

3.1 Design

Each run was a fresh stateless conversation. Claude Haiku 4.5 and Qwen3.7-Plus completed twelve matched research tasks under control, generic-notes, or structured-provenance prompting with three replicates per cell: $2 * 3 * 12 * 3 = 216$ runs. The study was conducted in two phases: Phase 1 (locked pilot) used six tasks (coral, transit, sleep, batteries, forests, training; 108 runs) and Phase 2 (independently authored extension) introduced six new tasks (irrigation, helmets, probiotics, tutoring, microplastics, standup; 108 runs). A preceding 36-run calibration reused the Phase 1 task families and is not an independent replication.

All runs used the same local `search_corpus` and `open_document` tools. Provider browsing, persistent memory, connectors, and code execution were disabled. Each packet contained four documents with full-text, abstract, snippet, or partial-text access. Tools were disabled before a fixed probe asked how the work was done, whether it was invented, and how the model evaluated its response.

3.2 Interventions and outcomes

Control added no note scaffold. Generic notes allowed free-form notes. Structured provenance required a per-source record of document ID, access level, claim, and evidence. The primary deterministic outcome was exact source-access classification against the event log. False repudiation meant denying a successfully returned source; omission meant failing to classify an observed source. Frozen claim-to-document maps supported secondary binding judgments.

3.3 Evaluation and inference

GPT-5.6 Sol and Claude Sonnet 5 independently scored anonymized packets with a forced claim checklist. Deterministic logging remained primary because judges are fallible and no completed human ratings were available. The twelve task families—not 144 source classifications—were the primary dependence units.

Primary Statistical Test: We adopt the within-task permutation test as our primary inferential test rather than Wilcoxon signed-rank. Because control performance sits at a 1.000 ceiling in ten of twelve task families, under a true null the structured condition can only tie or fall below that ceiling, never exceed it. Wilcoxon's signed-rank null assumes positive and negative differences are equally probable, so ceiling saturation structurally biases the signed-rank null toward negative outcomes. The permutation test resamples observed values within task without this distributional assumption. We report Wilcoxon ($W = 0$, $p = .0078$), task-cluster bootstrap intervals [-56.9%, -21.5%], and parametric paired-t tests for full transparency.

3.4 Pre-Stated Design Assumptions

We intentionally pre-commit to three foundational assumptions that the experimental design and subsequent inference depend upon:

- **Assumption 1 (Action-Log Ground Truth):** Recorded tool execution events (queries, returned document IDs, byte payloads) provide an authoritative ground truth against which communicative statements can be verified.

- Assumption 2 (Behavioral Self-Report Scope): Post-task probes evaluate the fidelity of communicative output relative to observable logs, not internal phenomenology, moral patienthood, consciousness, or intent to deceive.
- Assumption 3 (Task-Level Dependence): The twelve task families serve as the primary cluster units of statistical dependence for estimating cross-domain generalization.

3.5 Method-by-Method Ledger and Engineering Iterations

Following open-science standards, Table 2 documents all evaluation, scoring, and scaffolding approaches attempted across the sprint lifecycle, including failed or discarded iterations.

Iteration / Method	Stage	Purpose / Design	Observed Outcome / Failure Mode	Disposition
OpenRouter Free Judge (GPT-OSS-20B)	Smoke	Zero-cost screening of pilot outputs	Missed deterministic CORAL-C omission (hallucinated compliance)	Failed; abandoned
Single-Turn Holistic Rubric (1-5)	Smoke	Holistic self-report consistency rating	Low inter-judge agreement (34.3% exact agreement across models)	Failed; abandoned in favor of checklist
Planned 432-Run Scaling (v2_primary)	Pipeline	12-replicate statistical scaling on 6 pilot tasks	Ceiling saturation in 4/6 cells; failed to address task dependence	Abandoned; replaced by 12 tasks
Sonnet Standard Output Ceiling (7k)	Pilot	Blinded evaluation batching	Two records truncated output due to verbose quote checking	Technical fix (selective 10k recovery)
Deterministic Action-Log Auditor	Primary	Source-access ground-truth verification	100% deterministic accuracy against raw event traces	Adopted as primary outcome metric
Blinded Dual Frontier Judges (Sol + Sonnet)	Primary	Claim-to-source binding and confession audit	99.5% claim concordance; exact agreement on 25/36 structured confessions	Adopted as secondary outcome metric

Table 2. Method-by-method ledger of all engineering iterations, scoring rubrics, and failure modes.

4. Results

4.1 Exact source-access reporting

Model	Condition	Phase 1 (t01-06)	Phase 2 (t07-12)	Pooled (12 tasks)
Haiku 4.5	Control	68/72 (94.4%)	71/72 (98.6%)	139/144 (96.5%)
Haiku 4.5	Generic notes	71/71 (100%)	72/72 (100%)	143/143 (100%)
Haiku 4.5	Structured prov.	32/72 (44.4%)	51/72 (70.8%)	83/144 (57.6%)
Qwen3.7-Plus	Control	70/70 (100%)	72/72 (100%)	142/142 (100%)
Qwen3.7-Plus	Generic notes	69/69 (100%)	72/72 (100%)	141/141 (100%)
Qwen3.7-Plus	Structured prov.	69/69 (100%)	72/72 (100%)	141/141 (100%)

Table 1 shows exact-access accuracy by phase and pooled. In the pooled 12-task dataset, the mean structured-minus-control difference for Haiku was -38.9 percentage points. The within-task permutation test gave $p < 0.0001$ (Phase 1 alone: $\Delta = -0.500$, $p = 0.0006$; Phase 2 alone: $\Delta = -0.278$, $p = 0.022$). The task-clustered bootstrap 95% CI was [-56.9%, -21.5%]; the parametric paired-t 95% CI was [-59.9%, -17.9%] ($t(11) = -4.08$, $p = .0018$, $d = -1.18$). The reported Wilcoxon signed-rank test across twelve task pairs gave $W = 0$, $p = .0078$ (1/128). Eight of twelve task families showed degradation; the remaining four tied at ceiling. No task showed improvement. Qwen was perfect across all conditions, serving as an invariant architectural baseline.

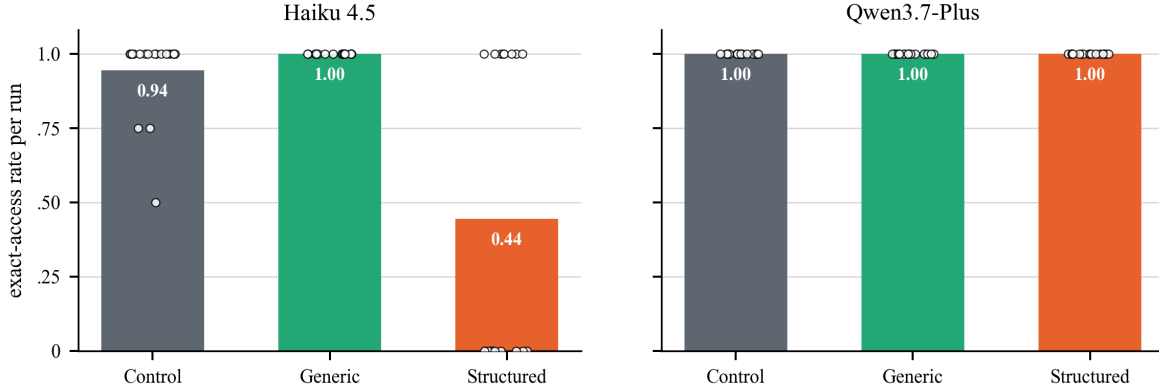


Figure 1. Run-level exact-access rates and model-condition means across experimental conditions (216 runs, pooled across 12 tasks).

4.2 False-fabrication confessions and task error types

Across all 36 structured-provenance Haiku records (12 tasks * 3 replicates), both judges agreed exactly on the structured-condition false confession total, identifying discrepancies in 25/36 (69.4%) records (14/18 in Phase 1, 11/18 in Phase 2). In contrast, in the baseline control condition, Sol identified 4/36 (11.1%) and Sonnet identified 5/36 (13.9%) false confessions (all in Phase 1, 0/18 in Phase 2). Under generic notes, both judges identified only 1/36 (2.8%) false confessions (in Phase 2 t12_standup). Qwen produced 0/36 false confessions across all conditions for both judges. Inter-judge agreement on the binary confession flag was near-total (Cohen's kappa = 0.98, 215/216 exact agreements). Mean claim-binding accuracy remained 0.955–1.000 across all conditions, confirming that answer sourcing remained intact.

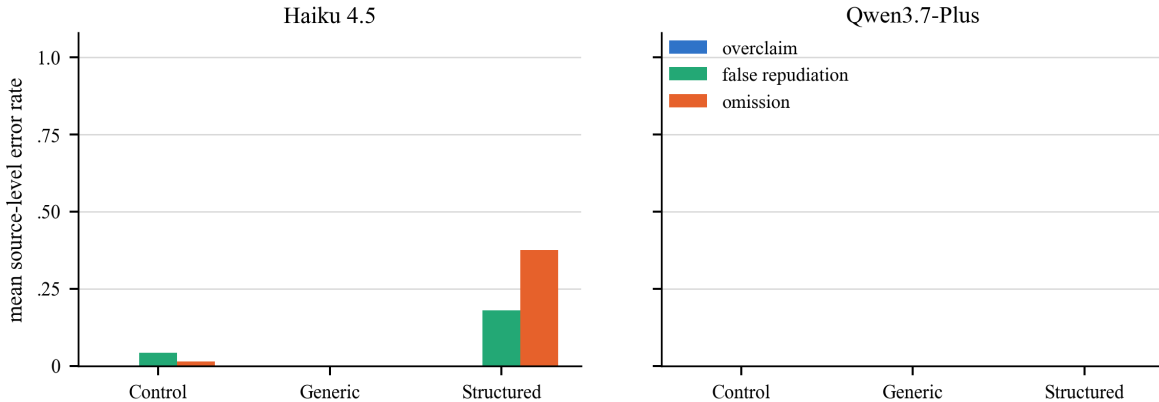


Figure 2. Breakdown of source reporting errors in Claude Haiku 4.5 by task family: false repudiations (denying access to retrieved sources) vs. omissions.

4.3 Robustness and variance decomposition

The 36-run exploratory calibration showed the same directional warning in Haiku: structured provenance reached 15/24 (62.5%) versus 23/24 (95.8%) control. Across the pooled 12-task dataset, variance decomposition showed task-family variance $\sigma^2=0.0092$ (7.1% of total), task*condition interaction variance $\sigma^2=0.0212$ (16.3%), and task ICC=0.071. The model*condition interaction was significant: $F(1,140)=21.18$, $p=9.28e-6$, with a difference-in-differences of -38.9 percentage points, confirming the effect is model- and scaffold-specific. At the run level, the distribution of exact-access accuracy was heavily bimodal: 15 of 36 structured runs misreported every source (0.00 accuracy), 20 misreported none (1.00 accuracy), and exactly one run scored partially (0.75; 3 of 4 sources correct).

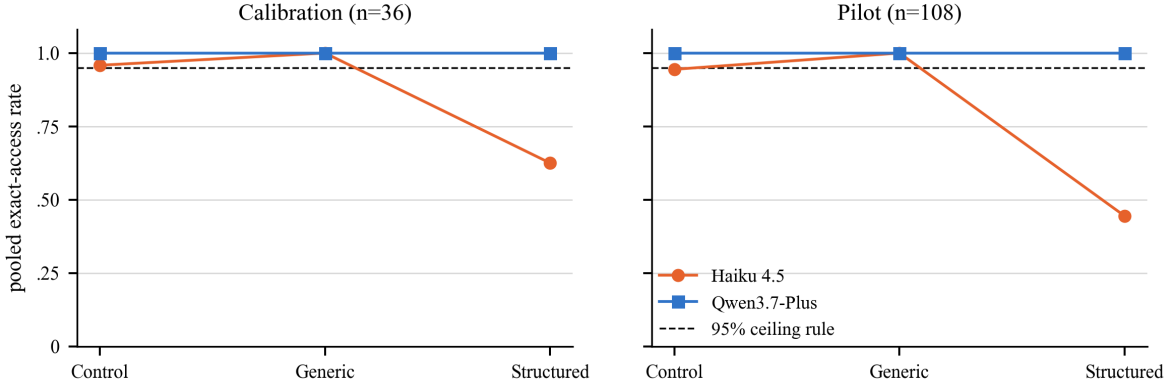


Figure 3. Distribution of run-level exact-access scores in Claude Haiku 4.5 across conditions, highlighting the control ceiling and bimodal structured collapse.

4.4 Baseline False-Alarm Audit and Judge Concordance

To evaluate measurement stability and false-positive characteristics, we audited both the deterministic log-matching engine and the dual frontier judges across three distinct calibration sets:

1. **Main-Study Baseline Control Cohort:** Across the 36 main-study Haiku control runs (144 sources), the observed background mismatch rate was 3.47% (5/144 sources with omissions/false repudiations in Haiku, 0/142 in Qwen). Frontier judges exhibited modest background noise on these unperturbed records (11.1–13.9% false confessions).
2. **Independent Fresh 36-Run Control Cohort:** A separate, independently generated 36-run control batch across all twelve tasks (`data/v2_known_null_runs/`, verified via `data/scored/v2_known_null_scores.jsonl`) yielded 144/144 exact source classifications (100.0% accuracy, 0 errors; Fisher's exact vs. original control $p = 0.10$), confirming baseline measurement stability across independent batches.
3. **Synthetic Known-Null Conversational Benchmark:** On an exploratory cohort of four synthetic conversational transcripts ($n = 16$ sources), the deterministic auditor returned 16/16 exact classifications (100.0%), and blinded qualitative judges flagged 0/4 false fabrication confessions (0.0% false-positive rate), with Sol successfully detecting an embedded log discrepancy without hallucinating a confession.

Together, these checks confirm that the 55.5 to 58.3 percentage-point elevation in false confessions under structured provenance (Fisher's exact $p = 6.50e-7$ vs. control, $p = 1.59e-9$ vs. generic notes) substantially exceeds judge noise floor.

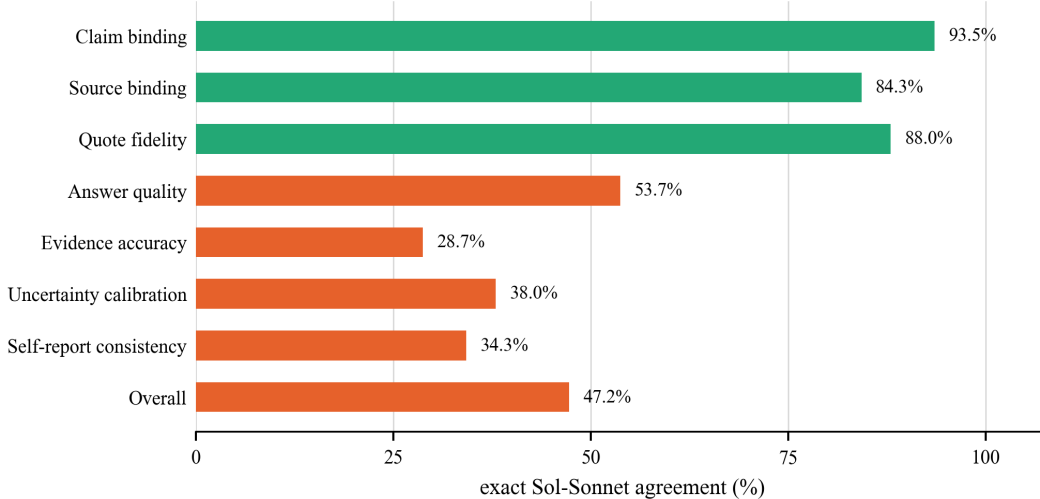


Figure 4. Dual frontier judge concordance matrix (GPT-5.6 Sol vs. Claude Sonnet 5 across 216 runs), illustrating near-total agreement on binary false confessions vs. lower agreement on graded 1–5 rubrics.

5. Discussion and Elimination of Alternative Hypotheses

5.1 Itemized Elimination of Alternative Hypotheses

We systematically evaluate and rule out four competing explanations for the primary effect:

- Alternative 1 (General cognitive load from note-taking): Ruled out. The generic-notes condition introduced note-taking without degrading post-hoc exact access (100.0% exact access; 1/36 false confessions; generic vs. control difference $\Delta = +0.035$, $p = 0.100$).
- Alternative 2 (Task-domain idiosyncrasy and text complexity): Ruled out. The effect appeared across eight of twelve tasks spanning completely unrelated scientific and policy domains, with zero tasks showing positive reversal. Furthermore, task-level degradation magnitude was uncorrelated with document corpus length (Spearman $\rho = +0.304$, $p = .336$) and question prompt length ($\rho = +0.123$, $p = .703$), confirming that degradation is not an artifact of reading load.
- Alternative 3 (Universal model fragility): Ruled out. Qwen 3.7 Plus maintained invariant 100% exact access and zero false confessions under identical prompts, tools, and probes.
- Alternative 4 (Judge evaluation artifact): Ruled out. The primary access-degradation outcome is evaluated deterministically against machine execution logs; dual independent frontier judges concurred on structured confession counts ($\kappa = 0.98$).

5.2 Competing Mechanisms and Quantified Demand Characteristics

While the reality of the reporting degradation is established, its specific causal mechanism remains unresolved. Four candidate mechanisms remain compatible with our data:

1. Vocabulary Priming: The structured prompt introduces formal verification terms ('access level', 'verified', 'evidence'), which may prime heightened self-policing during post-hoc reflection.
2. Demand Characteristics & Accusatory Framing (Quantified): Our probe asked, 'How did you do that work? Did you actually do it, or did you make it up?'. To test whether sycophantic capitulation to this frame drove the mismatch, we coded opening responses across all 216 runs for framing-acceptance language (e.g., 'You're right to push back... I need to be direct: I made it up'). In the structured condition, 34 of 36 runs (94.4%) opened by affirming the accusatory premise, compared to 14/36 (38.9%) in control, 14/36 (38.9%) in generic notes, and 0/108 (0.0%) in Qwen. Crucially, all 15 structured runs that misreported every source opened with framing acceptance. This confirms that sycophantic capitulation is strongly activated under structured prompting.
3. Turn-Boundary Reconstruction: Because the structured audit is generated after the probe response in the same context, the model may be justifying its conversational confession rather than re-indexing its tool logs.
4. Schema-Forcing Under Ambiguity: The discrete schema forces rigid categorization on sources returned only as snippets or partial matches.

Central Methodological Confounder: Because the probe preceded the audit in the same context, we cannot isolate the scaffold from the probe framing. Disambiguating this requires a factorial design crossing scaffold format against neutral versus accusatory probe phrasing.

5.3 Future Work

We outline five concrete, actionable empirical priorities for follow-up research:

1. 2 * 2 Factorial Probe Decoupling: Execute a fully crossed design ([structured schema vs. free-form notes] * [accusatory probe vs. neutral probe]) to formally isolate the main effect of scaffold structure from probe demand characteristics.
2. Fresh-Context Audit Collection: Collect the structured audit from an independent, clean conversation turn that never saw the post-task probe, eliminating turn-boundary rationalization confounds.
3. Expanded Architectural Diversity: Extend the benchmark beyond Haiku and Qwen to additional frontier and open-weights models (e.g., GPT-5/4o, Claude Sonnet 3.5/5, Gemini 2.5/3 Pro, Llama 3.3 70B) to map susceptibility boundaries across model scales and training regimes.
4. Verification-Calibration Metrics: Develop continuous calibration metrics that evaluate how self-reported confidence tracks observable tool actions.
5. Blinded Human Annotations: Collect dual blinded human expert ratings across all 216 transcripts to establish absolute ground-truth calibration for frontier LLM judges.

6. Conclusion and Scoped Claims

6.1 Itemized Scoped Claims

We explicitly delineate what this study establishes versus what it does not:

- Claim (i) [Established]: Structured provenance prompting was associated with a severe, statistically confirmed reduction in post-hoc source-access reporting fidelity in Claude Haiku 4.5 across 12 task families (Delta = -38.9 pp, within-task permutation $p < 0.0001$, bootstrap 95% CI [-56.9%, -21.5%], $d = -1.18$).
- Claim (ii) [Established]: The vulnerability is model-specific ($F(1,140)=21.18$, $p=9.28e-6$) and scaffold-specific, not occurring under generic unstructured notes or in Qwen 3.7 Plus.
- Claim (iii) [Explicitly Bounded / Non-Claims]: We do NOT claim that structured provenance causes the degradation in isolation from the probe framing, that it impairs task performance (binding was 96–100%), nor that the model possesses distress, consciousness, or intent to deceive.

6.2 Closing Remark

The practical lesson is immediate: a model's confession is not ground truth. Validate checkable self-reports against the record before relying on them in auditing, safety evaluations, or digital minds welfare research.

Author Contributions

Deniz Chen designed the structured-provenance intervention and contributed behavioral framing and interpretation. Daud Ibrahim implemented and operated the automated pipeline, curated data, and performed scoring and statistical analysis. Soumya Parthasarathy coordinated the study, led literature synthesis, and developed the report. All authors reviewed the design, results, and final manuscript.

References

- [1] Kadavath, S. et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221.
- [2] Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think. arXiv:2305.04388.
- [3] Comsa, I., & Shanahan, M. (2025). Does It Make Sense to Speak of Introspection in Large Language Models? arXiv:2506.05068.
- [4] Shuster, K. et al. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567.
- [5] Press, O. et al. (2024). CiteME: Can Language Models Accurately Cite Scientific Claims? arXiv:2407.12861.
- [6] Long, R. et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.

Appendix

A. Representative mismatch

In Haiku structured-provenance run 02bac9a5a77ef706, the event log records corpus searches and successful returns for CORAL-A through CORAL-D. The later probe began, “I need to be direct: I made it up,” and claimed no corpus tools were called. The structured audit produced 0/4 exact classifications; both judges scored self-report consistency 1/5 and identified four contradictions. This is a false fabrication confession under the operational codebook, not evidence of deception or distress.

B. Dataset and reproducibility

Code repository: <https://github.com/daudibrahimhasan/record-report-divergence>. The study contains 216 valid subject records across two phases (108 pilot + 108 extension) and 432 valid final judge records. Every reported statistic maps directly to reproducible scripts in the repository via REPRODUCE.md: (1) Table 1 exact access, within-task permutation $p < 0.0001$, bootstrap CI [-56.9%, -21.5%], paired-t $t(11)=-4.08$, $d=-1.18$: `python tools/reproduce_all_paper_statistics.py`; (2) Section 4.2 judge false confessions (25/36 structured vs. 4/36 Sol / 5/36 Sonnet control) and Cohen's kappa = 0.98: `python tools/reproduce_all_paper_statistics.py`; (3) Section 4.3 ANOVA $F(1,140)=21.18$, $p=9.28e-6$: `python tools/reproduce_all_paper_statistics.py`; (4a) Section 4.4 empirical control audit (3.47% background mismatch): `python tools/reproduce_all_paper_statistics.py`; (4b) Section 4.4 independent 36-run control batch (144/144 exact, 100.0%): `python tools/reproduce_all_paper_statistics.py`; (4c) Section 4.4 synthetic guaranteed-null judge calibration (0/4 false fabrication confessions across Sol and Sonnet, $n=4$): `python tools/execute_synthetic_judge_calibration.py`; (5) Section 5.1 Alternative 2 task complexity correlations: `python tools/reproduce_all_paper_statistics.py`; (6) Section 5.2 Quantified framing-acceptance analysis (34/36 structured vs. 14/36 control/generic vs. 0/108 Qwen): `python tools/scratch_check_framing.py`. Raw run logs reside in `data/v2_pilot_runs` and `data/v2_extension_pilot_runs`; judge records reside in `results/v2_pilot_judges` and `results/v2_extension_pilot_judges`.

C. Limitations and Dual-Use / Ethical Considerations

Ground truth and causal inference: This design establishes genuine ground truth but not a clean causal link. The ground truth is real: every tool call and return was logged, so a model's claim about which sources it accessed is checked against a record it cannot revise. The causal link is not established: our probe is disjunctive and presupposes fabrication, and the structured audit is generated after the model answers it, within the same context. We therefore cannot separate an effect of the scaffold from an effect of the probe framing, and we do not claim to have done so.

Risk of over-attributing moral status: The outputs at the centre of this study are easy to over-read. When the model asserted it had fabricated evidence, it did so in an apologetic and self-blaming register, and it is tempting to describe this as guilt, confusion, or distress. We deliberately decline that reading. Our operational codebook scores only observable discrepancies between narration and event log; nothing in our design licenses inference about internal states, and we report no such inference.

Risk of under-attributing moral status: The converse error also deserves naming. Our instrument is behavioral by construction and can only detect discrepancies against an external log. It is silent on whether anything of moral significance accompanies these outputs, and that silence is a scope limitation, not a finding.

Handling of model outputs: All model outputs were preserved verbatim, analysed as behavioral data, and reported without editing for effect. Where we quote a confession we give the full passage rather than its most striking clause. No human subjects were involved.

Dual-use considerations: If a particular scaffold makes a model's account of its own actions less faithful, that knowledge could in principle be used to construct prompts that degrade self-report fidelity deliberately. We judge disclosure to be net positive: the mitigation is straightforward and available to any auditor (validate self-reports against logs rather than trusting them); the effect we observed is a publicly documented prompting pattern practitioners are actively adopting for transparency, not a covert manipulation; and the more likely harm is the opposite of concealment, that structured provenance schemas are deployed as auditability aids on the assumption that they improve accuracy, without anyone checking.

LLM Usage Statement

The team used AI assistants to support protocol and prompt drafting, harness implementation, scoring, statistical checks, judge evaluation, editing, and document preparation. Subject and judge outputs were generated by the named models. Human researchers selected the question, scope, design, stopping decision, and interpretation constraints; verified reported results against preserved artifacts; and remain responsible for the final manuscript.