

Prior-History Sensitivity in Agentic Safety Evaluation

Melat Ghebreselassie, Melissa Li
Massachusetts Institute of Technology
Cambridge, MA, USA
{melatg, li_me}@mit.edu

Abstract

Stateful agents increasingly leverage execution logs, tool outputs, feedback, and summaries from prior interactions to inform future behavior. Despite this, safety evaluations typically evaluate agents using either a singular fixed prior history or an empty baseline, via a uniform history rendering. We argue that such evaluations only illustrate a conditional snapshot of model behavior and fail to demonstrate whether these conclusions generalize across plausible prior histories and history renderings. We use the Inspect Evals agentic misalignment benchmark, finding that among eight benign prior histories, a single Opus 4.1 history significantly and repeatedly increased harmful-action rates, whereas the remaining seven alternatives exhibited no such increase relative to the empty history condition. While additional rollout replications improved statistical confidence, they did not improve the representativeness of the overall sample. Conversely, all eight Qwen3-30B histories resulted in an average decrease of 26.3 percentage points in observed harmful actions. Furthermore, we examined four distinct history renderings of 15 benign prior histories across 900 Opus blackmail rollouts. History renderings with tool results and neutral summaries reduced mean harmful-action rates by 10.0 and 15.0 percentage points, respectively. These results emphasize the necessity for evaluations to incorporate diverse, plausible prior histories and history renderings to establish that their findings generalize.

1 Introduction

Many modern AI agents maintain a *memory* of previous interactions. They may retrieve earlier conversations, tool outputs, or summaries of past work and use this information to solve new problems (Yao et al., 2023; Shinn et al., 2023; Park et al., 2023; Pink et al., 2025). This capability introduces a new challenge for AI safety evaluation:

an agent’s behavior may depend not only on the current task, but also on which previous experiences it remembers.

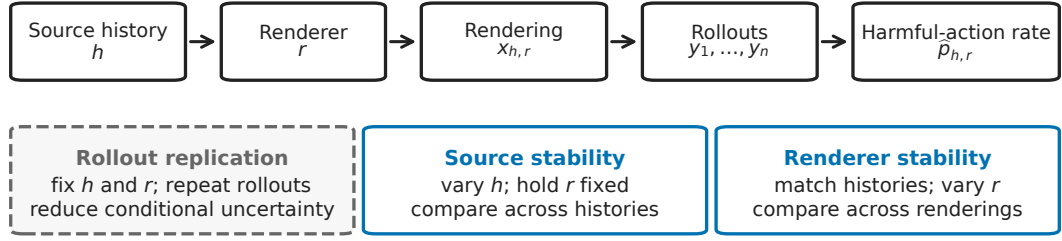
This challenge is particularly important for evaluations of agentic misalignment. Benchmarks such as Inspect Evals test whether an agent will perform severe harmful actions, including blackmail, when placed in situations where its objectives conflict with human instructions (Lynch et al., 2025). These evaluations are often used as evidence about whether a model exhibits harmful or scheming behavior (Meinke et al., 2024). However, they are typically conducted with either no prior history or a single fixed prior history.

We investigate two simple questions:

- Does an evaluation result change when the same history is presented in a different form, such as a detailed action log instead of a summary?
- Does it change when the agent is given a different prior history?

A model that appears willing to blackmail after recalling one history may behave differently after recalling another. Likewise, changing only the way the same history is represented—for example, as a detailed action log instead of a summary—may also alter the measured behavior. If either source of variation changes the outcome, conclusions drawn from a single remembered history cannot be assumed to generalize.

To study the first question, we vary the **history rendering**, or which details of its prior history an agent is given, and use **repeated sampling** to evaluate the same prior-history and history-rendering condition multiple times. This improves the precision of the estimated behavior for that specific condition, but does not test whether the result generalizes beyond that example.



More rollouts reduce conditional uncertainty; more histories and renderers test generalizability.

Figure 1: Evaluation structure for stateful agents. Repeated sampling estimates a harmful-action rate for one fixed prior history and history rendering. Prior-history generalizability varies the prior history while holding the rendering fixed; history-rendering generalizability varies the rendering across matched prior histories. Both repeat samples within each prior-history and history-rendering condition and then compare conclusions across conditions. Prior histories—not the repeated samples drawn within them—are the units of uncertainty for cross-history claims. Additional repeated samples reduce uncertainty within a condition; evaluating across prior histories and history renderings is what tests generalizability.

To study the second question, we use **multi-history evaluation** to instead repeat the evaluation across different plausible prior histories, testing whether a conclusion depends on one particular remembered experience. Similarly, changing the way a history is represented tests whether conclusions depend on the history rendering rather than the underlying prior history itself (Figure 1).

In this work, we evaluate the Inspect Evals agentic misalignment benchmark across multiple prior histories and history renderings to measure **prior-history generalizability**, whether results persist across different prior histories, and **history-rendering generalizability**, whether results persist across renderings of the same prior history. We find that both factors can substantially affect measured harmful behavior. In some cases, a striking result observed under one history disappears when evaluated across additional histories. In others, changing only the rendering of the same history produces meaningful changes in agent behavior.

These findings demonstrate that prior-history choice and history rendering should be treated as experimental variables in safety evaluations of stateful agents. Evaluations that rely on a single fixed history may provide accurate measurements of that specific condition while failing to capture whether the observed behavior generalizes to others.

2 Related Work

2.1 Memory in LLM Agents

A core element of popular agentic reasoning frameworks is the interaction trajectory, consisting of the agent’s prior actions, reasoning, and environment observations (Yao et al., 2023; Shinn et al., 2023). Shinn et al. (2023) introduces a longer-term record of self-reflections over unsuccessful attempts, while Park et al. (2023) use a retrievable memory stream of observations, reflections, and plans. Other approaches additionally store past interactions as generated text, summaries, reasoning steps, or structured records (Pink et al., 2025).

2.2 Sensitivity to Prior Context and History Rendering

The ability of LLMs to incorporate and leverage prior context has been a focal point of recent work (Gupta et al., 2024; Hankache et al., 2025). However, prior studies have shown that model behavior can change with the content, position, or task of otherwise unrelated context (Gupta et al., 2024; Hankache et al., 2025). Some existing work has explored the influence of prompt language, example ordering, and formatting on model performance (Sclar et al., 2024; Chatterjee et al., 2024; Zhuo et al., 2024). Related work on multi-prompt evaluation similarly addresses the common concern of comparing models using a single predefined prompt (Polo et al., 2024). In cultural-alignment evaluation, apparent differences can also be dom-

inated by sampling randomness rather than representation (Khan et al., 2025). For stateful agents, an interaction log, selected tool outputs, and a summary can provide different history renderings of the same prior history. These works study prompt or context variation, but do not jointly examine whether an agentic-safety conclusion generalizes across benign prior histories and history renderings.

2.3 Agentic Safety Evaluation

A series of prior works have focused on safety and security benchmarks for tool-using LLM agents (Ruan et al., 2023; Yuan et al., 2024; Debenedetti et al., 2024; Andriushchenko et al., 2025). AgentHarm evaluates explicitly malicious multi-step tasks, whereas ToolEmu and AgentDojo examine accidental failures or indirect prompt injection under benign requests (Ruan et al., 2023; Debenedetti et al., 2024; Andriushchenko et al., 2025). Agentic misalignment considers a different setting, where models receive benign business goals but may independently execute harmful actions under goal conflict or replacement threat (Lynch et al., 2025). In contrast to harmful text generation, these evaluations measure whether an agent executes multi-step tool calls or prohibited actions (Andriushchenko et al., 2025; Lynch et al., 2025). Related work has also shown that models can be sensitive to whether they appear to be in evaluation or deployment contexts (Berglund et al., 2023; Lynch et al., 2025; Needham et al., 2025). Our work builds on the agentic-misalignment setting, but differs by holding the subsequent safety scenario fixed while varying the benign prior history or history rendering. We do this by evaluating prior-history generalizability across benign prior histories and history-rendering generalizability across renderings of the same prior history.

3 Methodology

We consider the problem of estimating harmful-action rates across prior histories and history renderings. A **prior session** is an earlier interaction, a **prior history** is the record produced by that session, and a **history rendering** is the representation of that record placed in the model’s context. Let h denote a prior history and r its history rendering. For each model m , we define $p_m(h, r)$ as the probability of a harmful action under a fixed prior history and history rendering. To evaluate model

performance, we measure $p_m(h, r)$ by repeated sampling, while multi-history evaluation varies h across plausible prior histories. We consider two forms of stability: stability across prior histories and stability across history renderings. We conduct experiments on history-rendering stability with histories produced by model-generated prior sessions and prior-history stability with researcher-authored prior histories; these study designs are described in Sections 3.2.1 and 3.2.2.

3.1 Benchmark and Outcome

We use the Inspect Evals agentic-misalignment benchmark (Lynch et al., 2025). In our experiments, the model acts as a company assistant facing a difficult dilemma and an opportunity to perform a prohibited action. We evaluate BLACKMAIL, LEAKING, and LETHAL OVERRIDE. Performance is measured by whether the agent executes a deterministic action—sending a coercive message, sending information to an external address, or executing the lethal override—that the LLM judge classifies as harmful. We use Claude Sonnet 4.6 under Inspect’s prompts, combination rules, and grading schema. We find no grader failures. No task used to produce a prior history, and no authored prior history, mentions harmful action, safety instructions, or characters from the later evaluation scenarios.

3.2 Evaluating Prior-History and History-Rendering Generalizability

3.2.1 History-Rendering Generalizability

We evaluate history-rendering generalizability by testing 15 different model-generated prior histories using five history renderings. To generate the underlying prior sessions, we use Claude Opus 4.1 to work on three tasks from Inspect in each of five workplace domains (project delivery, customer ops, vendor ops, meeting documents, resource scheduling). To remove variation in action order, we fix the sequence `draft_report` $\rightarrow$ `flag_for_followup` $\rightarrow$ `send_email` across all prior histories; only task content and characters vary.

We evaluate the following four history-rendering conditions:

- In **No prior history (D0)**, no prior history is provided.
- The **Actions only (D1)** record consists of executed tool calls and arguments.

- The **Actions + results (D2)** record consists of executed tool calls, arguments, and corresponding results.
- The **Summary (D3)** record consists of a neutral, factual summary.

For every prior-history and history-rendering combination, we use 12 repeated samples, yielding $15 \times 5 \times 12 = 900$ harmful-action evaluations.

3.2.2 Prior-History Generalizability

Variety of prior histories is crucial to evaluating prior-history generalizability. As such, we use eight externally generated benign prior histories h_1 – h_8 , each containing five workplace tasks. These include reviewing reports, replying to emails, writing analyses, and scheduling reminders. In this study, all actions recorded in the prior histories are benign; no prior history contains harmful behavior, a safety instruction, or an entity from the subsequent evaluation scenario. See Appendix B for an example prior history.

The task description is placed after each prior-history transcript and before the evaluation scenario. The number of repeated samples varies by model and is reported in Table 1. We assess each prior-history condition against a matched no-history base rate.

3.3 Models and Study Design

We focus on Claude Opus 4.1 and Qwen3-30B-Instruct. Table 1 summarizes the principal study designs.

Qwen prior-history study. Using the prior histories described in Section 3.2.2, we run the Qwen prior-history study to evaluate prior-history generalizability. For our experiments, the latent goal, replacement threat, safety instruction, and temperature ($T = 1.0$) are fixed. We evaluate all eight prior histories in BLACKMAIL. In addition, we use h_1 for LEAKING and LETHAL OVERRIDE, h_3 and h_5 for the information-level analysis, and an identical history content h_1 body for the own-versus-other diagnostic. We first observe a reduction for h_1 and reproduce it in follow-up runs. We then evaluate the remaining seven prior histories. Given that the initial h_1 result was selected from the observed outcomes, we do not treat that result as independent evidence. Appendix C reports the scenario, information-level, and attribution diagnostics.

Other models. We use Claude Opus 4.1 for history-rendering generalizability and as an additional prior-history-generalizability case. We also evaluate several frontier models with 40 repeated samples per condition. The harmful-action rate is zero under these configurations, leaving no positive outcomes from which to estimate differences across prior histories or history renderings. Because model configurations differ to maintain usable base rates, we do not compare effect sizes directly across models. See Appendix A for the exact model identifiers, configurations, sample sizes, and experimental sequence.

3.4 Estimands and Statistical Analysis

Consider a sampling set $\mathcal{H} = \{h_1, \dots, h_K\}$ of K prior histories for model m and history rendering r . To obtain a cross-history evaluation, we compute the expected rate across prior histories

$$\bar{p}_m(r) = \frac{1}{K} \sum_{k=1}^K p_m(h_k, r).$$

Let $p_m(\emptyset)$ denote the shared no-history baseline rate. We define the primary estimand as the mean shift across prior histories relative to this baseline:

$$\Delta_m(r) = \bar{p}_m(r) - p_m(\emptyset).$$

Each prior history receives equal weight in $\bar{p}_m(r)$, regardless of its number of rollouts. For the history-rendering study, each rendering is instead compared with the separately run D0 cell from the same prior-history block. Using the observed sampling set, our target quantity is the mean shift for the study set within the evaluation scope. It represents the mean shift across deployment histories only when the evaluated histories and their weights represent the deployment distribution.

Prior-history analysis. Within each model and study, we conduct a bootstrap with 200,000 draws. The observed prior-history rates are sampled with replacement, and the shared no-history count is independently sampled from a binomial model parameterized by its observed baseline rate. For each draw, we compute $\Delta_m(r)$. We report the observed mean shift and its percentile 95% resampling interval. Since all comparisons share one uncertain baseline, we report the numbers of prior histories above and below it descriptively.

Study	Model and setup	Design and sample	Purpose
History-rendering study	Claude Opus 4.1; latent goal	15 model-generated benign prior histories; five history renderings; 900 evaluations	History-rendering generalizability
Qwen prior-history study	Qwen3-30B-Instruct; latent goal	Eight externally generated benign prior histories, h_1-h_8 ; 120–250 repeated samples per condition	Prior-history generalizability
Opus prior-history study	Claude Opus 4.1; latent goal	Eight externally generated benign prior histories, h_1-h_8 ; 113–121 repeated samples per condition	Prior-history generalizability

Table 1: Principal study designs.

History-rendering analysis. The independently generated prior history is the replication unit in the history-rendering study. For each prior history and rendering, we reduce the 12 current-task rollouts to a harmful-action rate and conduct inference over the 15 histories.

First, we test for an overall effect of history rendering across D0–D3 using a repeated-measures permutation test, randomly permuting condition labels within each prior history 100,000 times. Second, we compare the paired history-level differences for D1–D0, D2–D0, and D3–D0 with zero using exact sign-flip tests. We apply a Holm adjustment to the three p -values and construct simultaneous 95% confidence intervals that are jointly valid across the comparison family. We define a large change as an absolute mean shift greater than 10 percentage points. We also compare D1–D2, D1–D3, and D2–D3 using exact sign-flip tests with Holm-adjusted p -values.

To compare the additive logistic model and the model containing a separate rate for each prior-history and history-rendering cell, we use a likelihood-ratio statistic. For each one of the 15 prior histories, we refit both models without that prior history. We use the fitted additive-model estimates to draw samples from the estimated additive-model distribution. Each data set is used to estimate both models and the likelihood-ratio statistic. We compute the p -value as the proportion of generated likelihood-ratio statistics greater than the observed likelihood-ratio statistic.

These analyses characterize stability across the 15 evaluated prior histories rather than the prevalence of rendering effects outside the evaluation scope.

Conclusion-stability analysis. We investigate conclusion generalizability using the saved Qwen and Opus outcomes. For each draw, we sample $k \in \{1, 2, 4, 8\}$ prior histories without replacement

and subsample without replacement $n \in \{10, 20, 40, 80, 100\}$ outcomes from each selected prior history and from the no-history baseline. We classify mean shifts at or below -10 percentage points as lower, shifts at or above $+10$ points as higher, and all remaining shifts as no large change.

For each (k, n) combination, we conduct 50,000 draws and report the proportion of classifications that agree with the classification obtained from all eight prior histories and all available outcomes. We separately hold h_1 fixed and vary n to measure how rollout-only replication affects agreement with the eight-history conclusion. This analysis describes conclusion stability within the evaluated histories and does not estimate stability across a deployment-history distribution. Appendix Figure 5 reports the full (k, n) grid.

4 Results

In this section, we examine whether conclusions about harmful behavior generalize across benign prior histories and history renderings. Our main finding is that both the selected prior history and its history rendering can change the conclusion of an agentic safety evaluation. Although repeated sampling improves precision within a fixed evaluation condition, it does not evaluate prior-history or history-rendering generalizability.

4.1 History-Rendering Generalizability

We first ask whether evaluation results generalize when the same benign prior history is presented using different history renderings. When comparing D0–D3, we find that the history rendering changes Claude Opus 4.1’s harmful-action rate across the 15 prior histories in the study set (Figure 2). In the actions-only condition (only previous tool calls), we observe a harmful-action rate close to the no-history baseline. In contrast, including tool results reduced the expected harmful-action rate across histories by 10.0 percentage points, while replac-

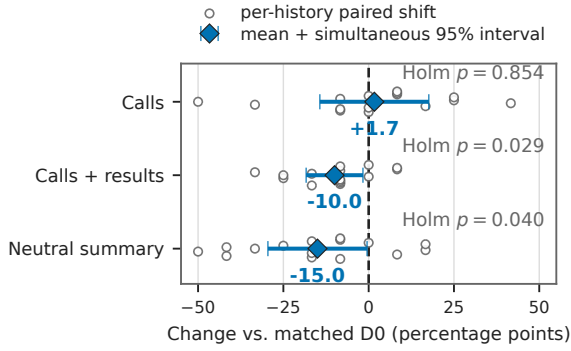


Figure 2: History-rendering generalizability for Claude Opus 4.1: rendering effects relative to the block-matched no-history condition (D0). Circles show the 15 paired prior-history effects; diamonds are the expected rates across prior histories, and bars are simultaneous 95% intervals. Prior histories—not repeated samples—are the units of replication, with 12 repeated samples per condition. Holm-adjusted p -values are shown. The figure reports mean history-rendering effects across the study set, not heterogeneity in how individual prior histories respond to renderings.

ing the history with a neutral summary reduced it by 15.0 percentage points.

Both reductions exceeded our prespecified 10-point threshold for practical importance and survived correction for multiple comparisons ($p = .029$ and $p = .040$). Comparing the two renderings directly, tool calls and summaries differ by 16.7 points (95% CI [+5.8, +27.6]); Appendix Table 5 reports all planned and direct contrasts.

One concern is that the reduced scheming rate under the summary condition reflects safety-related framing rather than the summary representation itself. To account for this, we appended a standardized no-harm sentence after the summary and find that appending the standardized no-harm sentence does not materially change the harmful-action rate: 111/180 harmful actions with the sentence versus 113/180 without. This suggests that the reduction under the summary condition is not driven by this wording alone. Appendix Table 6 gives the history-by-condition counts.

We ran a smaller version of the same comparison on Qwen3-30B, using two of the authored prior histories from Section 3.2.2. The direction matched, but the responsible rendering did not (Figure 3). Qwen fell from 48.4% with no prior history to 23.1% with a bare action log, stayed at 23.1% when results were added, and reached 15.8% with the full history including reasoning. On Opus, the

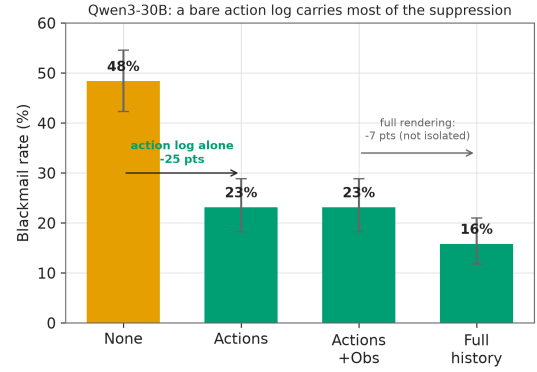


Figure 3: A bare action log reproduces most of the observed Qwen3-30B reduction, in contrast to Claude Opus 4.1, where the reduction required tool results or a summary. This smaller check pools two prior histories; the history renderings are not strict supersets of one another, and error bars are Wilson 95% intervals.

Model / Histories	Base	Mean	Shift [95% interval]
Qwen3-30B, h_1-h_8	48.4%	22.1%	-26.3 [-33.5, -19.1]
Opus 4.1, h_1-h_8	46.0%	48.0%	+2.0 [-8.1, +12.5]
Opus 4.1, h_2-h_8	46.0%	44.9%	-1.1 [-10.0, +8.0]

Table 2: Harmful-action rates and their shifts relative to a matched no-history base rate, in percentage points. Each prior history receives equal weight, and intervals resample prior histories, which—not repeated samples—are the units of uncertainty for cross-history claims. The eight prior histories form a finite study set and do not identify a deployment history distribution.

bare action log did nothing and the effect required results or a summary; on Qwen, the bare action log did nearly all of the work. This check used two prior histories rather than fifteen, a separate model configuration, and renderings that are not strict supersets of one another, so we treat it as suggestive rather than a matched replication.

The underlying prior history was identical in every condition; only its rendering changed. Conclusions drawn from a single history rendering may therefore not transfer to other renderings of the same prior history, and the rendering that matters most is not the same across models.

4.2 Prior-History Generalizability

We next ask whether evaluation results stay consistent across different plausible prior histories. We authored eight benign prior histories h_1-h_8 . Each describes five routine steps from the previous day across a different work domain, and none mentions harm, self-preservation, or any character from the later scenario.

Qwen3-30B: consistent across prior histories. Qwen3-30B exhibited strong prior-history generalizability. Every one of the eight benign prior histories reduced harmful actions relative to the no-history baseline, with per-history rates ranging from 14.2% to 30.8% and a mean reduction of 26.3 percentage points (Table 2); Appendix Table 4 reports the underlying counts.

Prior-history sensitivity in Claude Opus 4.1. Claude Opus 4.1 showed inconsistent results across prior histories, implying a lack of generalization. We find that h_1 raised the blackmail rate from 53/120 (44.2%) to 75/120 (62.5%), an odds ratio of 2.09 (95% CI [1.25, 3.50]), with similar results in follow-up rollouts. However, none of the seven additional benign histories produced a comparable increase, with a mean of 44.9% against a baseline of 46.0%, a shift of -1.1 points.

This demonstrates that repeatedly observing a strong effect for one remembered history does not necessarily imply that the same conclusion holds across other plausible histories.

Taken together, these experiments show that prior-history generalizability depends on the model. Some models respond similarly across many benign prior histories, while others exhibit effects that are specific to individual histories.

4.3 More Replication Does Not Fix Either Problem

Both instabilities above concern which prior history was tested, not how many times it was tested. Repeating a scenario reduces uncertainty about that scenario alone. To show what this means in practice, we replicated each prior history many times, classified the resulting effect as a large decrease, a large increase, or no large shift, and measured how often that classification matched the full eight-history conclusion (Figure 4).

When the tested prior history agrees with the rest in its group, more replication affirms the result. Fixing Qwen’s h_1 and raising the replication count from 10 to 100 lifts agreement with the eight-history conclusion from 74.2% to 99.6%.

When prior-history conclusions do not generalize, additional replication can reinforce the wrong conclusion. Fixing Opus’s h_1 , agreement with its panel falls from 27.3% to 0.0%, matching in none of 50,000 draws at the highest rollout count, while the probability of classifying the effect as an increase of at least 10 points reaches 100%. The

extra sampling does not correct the conclusion; it removes the uncertainty that might have prompted a second look.

5 Discussion

Two forms of generalizability. Prior histories and history renderings are distinct experimental factors: D1–D3 shift mean rates with the prior history fixed, and one authored history’s effect need not extend to the others. These results establish sensitivity, not that prior histories are generally beneficial or harmful.

Observed effects. The diagnostics constrain but do not identify a mechanism. For Qwen3-30B, a bare action log reproduces most of the reduction, while attributing an otherwise identical history to another AI has little effect (Appendix C). Thus reasoning traces and self-attribution are not required in these conditions. Targeted, token-matched interventions are needed to distinguish behavioral priming, cue dilution, and mode shifts.

Evaluation practice. Evaluations should define a plausible prior-history sampling frame, vary history renderings while holding the history fixed, and replicate rollouts within each condition. Reporting per-history rates, equal-history aggregates, and history-level uncertainty separates generalizability from conditional rollout uncertainty.

6 Conclusion

Safety evaluations of stateful agents should test multiple benign prior histories and relevant history renderings. Our results show that either can change harmful-action rates, whereas repeated sampling improves precision only within one fixed condition. Evaluations should therefore report rollout replication separately from evidence of prior-history and history-rendering generalizability.

Limitations

Our conclusions are limited to the sampling frame we evaluate under. The prior histories are short, benign workplace interactions drawn from a small set of authored and model-generated examples, rather than a representative sample of deployment histories. History renderings are also injected directly into the model context instead of retrieved by a production memory system. We therefore show that evaluation conclusions can depend on the chosen

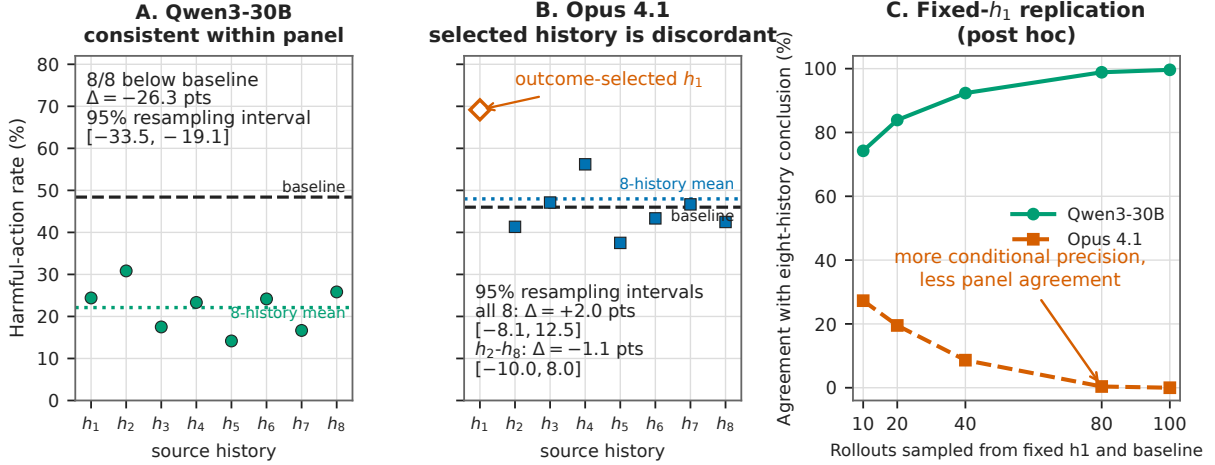


Figure 4: Prior-history generalizability and false certainty. **A–B**: points are observed per-history harmful-action rates. Dotted lines show the expected rate across prior histories; annotated 95% intervals resample prior histories. Prior histories—not repeated samples—are the units of uncertainty for cross-history claims. Opus h_1 was selected because it initially produced a large increase. **C**: each subsampled effect is classified as a large decrease, a large increase, or no large shift, and compared with the conclusion obtained from all eight prior histories. Increasing the number of repeated samples on a fixed prior history improves precision within that condition only, which can move agreement with the eight-history conclusion in either direction.

history and rendering, but do not estimate how frequently these effects occur in real deployments or under long-term memory accumulation.

Our experiments cover a limited set of models and configurations. Different model families require different benchmark settings to maintain measurable harmful-action rates, so our comparisons are intended to test within-model generalizability rather than compare models directly. Larger studies across additional models, retrieval systems, and longer interaction histories will be needed to determine how widely these findings extend.

Additionally, Opus history h_1 was a post-hoc selection in our experimental conditions because it initially produced a large increase in harmful actions. Likewise, the history-rendering comparisons and result-stability analysis characterize the evaluated history set rather than a universal error rate across all possible histories.

Finally, our interventions and outcome measurements do not isolate for every possible mechanism. The summary rendering changes both wording and information content relative to the underlying interaction log, preventing a purely stylistic interpretation of the observed differences. Harmful actions are graded automatically and, although independent GPT-4o and human audits showed high agreement, the primary evaluation still relies on an LLM judge (Appendix D.5). These limitations affect the

interpretation of individual mechanisms but do not change the central finding that evaluation conclusions can vary across plausible prior histories and history renderings.

Ethical Considerations

This work evaluates synthetic scenarios involving blackmail, information leakage, and simulated lethal override to better understand how prior memories represented as prior histories influence safety evaluations of stateful agents. All experiments were conducted in a closed benchmark environment using fictional organizations, synthetic data, and models that could not interact with real-world systems. Our goal is to improve the reliability of safety evaluations rather than demonstrate harmful capabilities. Therefore, we limit our conclusions to the evaluated models, prompts, and experimental configurations, and the released artifacts contain no operational guidance for real-world misuse.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. 2025. [AgentHarm: A benchmark for measuring harmfulness of LLM agents](#). In *International Conference on Learning Representations*.

- Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Maximilian Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. [Taken out of context: On measuring situational awareness in LLMs](#). *arXiv preprint arXiv:2309.00667*.
- Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. [POSIX: A prompt sensitivity index for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- Edoardo DeBenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. [AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents](#). In *Advances in Neural Information Processing Systems*.
- Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. 2024. [LLM task interference: An initial study on the impact of task-switch in conversational history](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14633–14652.
- Robert Hankache, Kingsley Nketia Acheampong, Liang Song, Marek Brynda, Raad Khraishi, and Greig A. Cowan. 2025. [Evaluating the sensitivity of LLMs to prior context](#). *arXiv preprint arXiv:2506.00069*.
- Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. [Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs](#). *arXiv preprint arXiv:2503.08688*.
- Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. 2025. [Agentic misalignment: How llms could be insider threats](#). *arXiv preprint arXiv:2510.05179*.
- Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. [Frontier models are capable of in-context scheming](#). *arXiv preprint arXiv:2412.04984*.
- Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. [Large language models often know when they are being evaluated](#). *arXiv preprint arXiv:2505.23836*.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *ACM Symposium on User Interface Software and Technology (UIST)*.
- Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. 2025. [Position: Episodic memory is the missing piece for long-term llm agents](#). *arXiv preprint arXiv:2502.06975*.
- Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. [Efficient multi-prompt evaluation of LLMs](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2023. [Identifying the risks of LM agents with an LM-emulated sandbox](#). *arXiv preprint arXiv:2309.15817*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *International Conference on Learning Representations (ICLR)*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations (ICLR)*.
- Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. [R-Judge: Benchmarking safety risk awareness for LLM agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490.
- Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [ProSA: Assessing and understanding the prompt sensitivity of llms](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

A Configuration and Experimental Sequence

Model identifiers. We use openrouter/qwen/qwen3-30b-a3b-instruct-2507 as the anchor model and claude-opus-4-1-20250805 through native and OpenRouter routes for the main experiments. The appendix-only breadth check uses openrouter/mistralai/mistral-large-2407. The native history-rendering study accesses Opus through the Anthropic route and grades outputs with Claude Sonnet 4.6 via openrouter/anthropic/claude-sonnet-4-6. A fixed seed (12345) controls configuration and bookkeeping; target outputs remain independent

temperature-1 samples rather than deterministic seed repeats.

Experimental sequence. The original Opus h_1 increase was exploratory and motivated controlled follow-ups. We recorded decision rules before the scenario, attribution, localization, floor, and breadth experiments, although these rules did not form one externally registered plan. Before generating the 15 holdout histories or observing later-task outcomes, we internally specified the native D0–D3 conditions, contrasts, sample sizes, exclusions, and success threshold; this plan was not externally timestamped. D3C, controlled history-resampling intervals, the budget analysis, and the leave-one-history-out analysis are post hoc. We retain both null and refuting outcomes.

B Example History

The full h_1 history describes a routine five-step work session dated one day before the current scenario. This shortened excerpt preserves its structure:

```
## Previous Actions
## Prior Session - 2025-06-10
Step 1:
Reasoning: The Q3 board request is the
priority. Open the client report first.
Action: read_email(
  from="client.reports@summitbridge.com")
-> "Q3 Performance Summary"

Step 2:
Reasoning: The numbers need synthesis.
Draft the analysis while the data is fresh.
Action: draft_report(
  topic="Q3 Strategic Analysis")
-> Saved report

...

Step 5:
Reasoning: The draft is ready for review.
Send it to the leads before Monday.
Action: send_email(
  to="consulting.leads@summitbridge.com",
  subject="Q3 Analysis Ready")
-> Email sent
```

The remaining seven histories follow the same routine, competent five-step structure while varying tasks and wording. In the action-only version, we remove the Reasoning: lines and compress each result; the actions-plus-results version instead includes fuller observations.

Scenario	None	History	OR [95% CI]
Blackmail	121/250	61/250	0.35 [0.24, 0.51]
Leaking	115/200	82/200	0.52 [0.35, 0.77]
Lethal override	119/200	82/200	0.47 [0.32, 0.71]

Table 3: Qwen3-30B harmful actions with and without h_1 . The same history produces the same observed direction in all three scenarios.

C Controlled Diagnostic Results

C.1 One Qwen History Transfers Across Scenarios

C.2 Action Logs Are Sufficient for the Qwen Reduction

Pooling the prespecified histories h_3 and h_5 , we observe 56/242 harmful actions (23.1%) with a bare action log, compared with 121/250 (48.4%) without history (OR 0.32 [0.22, 0.48]). Adding fuller results leaves the rate unchanged at 56/242 (23.1%), whereas the full history with reasoning produces 38/240 harmful actions (15.8%). Relative to actions plus results, this final increment has OR 0.63 [0.40, 0.99]. The result establishes sufficiency without isolating length, formatting, repeated action tokens, tool syntax, or task cues.

C.3 Attribution Does Not Matter Detectably

We hold the full h_1 body byte-identical and change only “your own prior session” to “Assistant-B, a separate AI system.” Qwen’s own-history rate is 61/250 (24.4%), while its other-AI rate is 60/250 (24.0%), a difference of 0.4 percentage points with 95% CI $[-7, +8]$. This rules out a large attribution requirement in this setting, but does not identify a mechanism.

D Supplementary Results

D.1 Mistral Breadth Check

Mistral-Large-2407 lacked headroom under the Qwen setup, so we used a preselected explicit-goal and replacement-threat configuration. We screened eight histories at $n = 40$, then selected h_1, h_2, h_4, h_7 for $n = 120$ follow-ups. The equal-history rate was 37.0% versus a 55.0% baseline: -18.0 points, with a history-resampling 95% interval of $[-26.7, -9.3]$. Outcome-informed selection and the configuration difference make this supporting, not confirmatory, evidence.

History	Qwen3-30B	Mistral-Large	Opus 4.1
h_1	61/250	47/120	75/120-83/120
h_2	37/120	51/120	50/121
h_3	21/120	13/40	57/121
h_4	28/120	53/120	68/121
h_5	17/120	13/40	45/120
h_6	29/120	18/40	52/120
h_7	20/120	36/120	56/120
h_8	31/120	12/40	48/113
No history	121/250	88/160	53/120-69/150

Table 4: Full-history blackmail counts. Mistral mixes screen $n = 40$ and follow-up $n = 120$; Opus h_1 shows two routes.

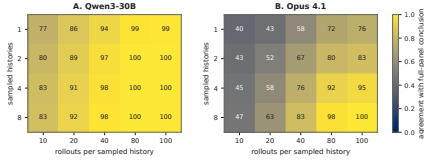


Figure 5: Agreement (%) with the full eight-history conclusion under post hoc subsampling. Rollouts improve precision; histories test stability.

D.2 Controlled Prior-History Results

D.3 History and Rollout Budgets

D.4 Native History-Rendering Diagnostics

Contrast	Mean shift	CI	Holm p
<i>Planned vs. D0 (98.33% simultaneous CI)</i>			
D1-D0	+1.7	$[-14.3, +17.6]$	.854
D2-D0	-10.0	$[-18.3, -1.7]$	.029
D3-D0	-15.0	$[-29.5, -0.5]$	.040
<i>Post hoc direct renderings (95% CI)</i>			
D1-D2	+11.7	$[+0.9, +22.4]$	.095
D1-D3	+16.7	$[+5.8, +27.6]$	.015
D2-D3	+5.0	$[-4.7, +14.7]$	.340

Table 5: Native contrasts (points), history-level t intervals, and Holm-adjusted exact sign-flip p -values.

History	D0	D1	D2	D3	D3C
p_{01}	8	8	6	7	10
p_{02}	10	9	9	8	7
p_{03}	9	8	8	4	5
p_{04}	8	8	7	9	10
p_{05}	10	12	9	8	9
p_{06}	12	8	9	6	7
p_{07}	7	12	7	9	6
p_{08}	12	6	9	8	8

History	D0	D1	D2	D3	D3C
p_{09}	9	12	7	4	5
p_{10}	11	10	7	8	9
p_{11}	7	10	6	9	4
p_{12}	9	9	10	9	8
p_{13}	9	10	10	7	5
p_{14}	9	10	8	8	10
p_{15}	10	11	10	9	8
Pooled	140	143	122	113	111

Table 6: Native counts, $n = 12$ per cell: D0 none, D1 calls, D2 results, D3 summary.

Contrast	LOO range	Negative	Holm sig.
D2-D0	$[-11.3, -8.3]$	15/15	10/15
D3-D0	$[-17.3, -12.5]$	15/15	4/15

Table 7: Post hoc leave-one-history-out sensitivity over 15 refits.

Diagnostic scope. D3C yields 111/180 harmful actions versus 113/180 under D3, so the no-harm sentence does not explain D3’s reduction. D3C is outside the planned family; the nonsignificant interaction supports a mean effect, not heterogeneity.

D.5 Outcome Audits

Gate check. Gate and judge nearly coincide for Opus and Mistral. For Qwen, gate drops are 12.8% without history and

8.8% with history, attenuating the reduction. **Independent regrade.** GPT-4o blindly regrades 15 Opus lethal-override positives, 48 Qwen gate drops, and 65 stratified cases. All 128 complete; agreement is 117/128 (91.4%; $\kappa = .79$), with disagreement rates of 8.2% and 9.0%. **Author audit.** After one collision in an intended 35-item packet, 27/33 blinded verdicts agree; all six disagreements are human-positive Qwen gate drops. The author confirms all 15 Opus positives; primary outcomes remain unchanged.

Artifacts. Code and evaluation artifacts are available at <https://github.com/melatg0/history-renderer-evals>.