WHAT SURVIVES A CODING-AGENT HANDOFF? DECISION RETENTION UNDER COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

During a coding session, a user and an agent establish constraints, reject approaches, and agree on next steps. A successor may need these decisions to continue the work. Replaying the conversation consumes input capacity and may exceed the context window. Summaries offer a shorter alternative. We ask how much of the recorded decision information a successor can still recover from them. We compare conversation replay, freeform summaries, and structured handoff packets using the same questions about prior decisions. Successor answers are assessed against references anchored to session quotations. Across 42 handoff points in private and publicly available coding conversations, successors recover more recorded decision information from replay than from structured packets. The replay advantage persists in sensitivity checks. Packet structure offers no consistent advantage over freeform summaries. Increasing the requested packet length also leaves a replay gap for the tested generator. In short behavioral probes, replay's higher retention is not accompanied by fewer observed breaches than intact summaries. These results support preserving access to the original record when recovering decisions matters. The best transfer policy for subsequent task execution remains unresolved.

1 INTRODUCTION

During a coding session, a user may prohibit a dependency, reject an approach, confirm a repair, or choose the next experiment. The repository records code, but need not explain these decisions. Supplying prior context can reduce rediscovery effort in coding-agent takeovers (KC & Budathoki, 2026). A successor may therefore need conversation history as well as the code and latest request.

Replaying that history has an input cost and may exceed the context window. Context-management studies compare summarization with selective omission to reduce coding-agent costs (Lindenbauer et al., 2025). A freeform summary condenses the conversation; a structured handoff packet organizes it into fields such as constraints, completed work, and next steps. Either may omit decisions. Synthetic agent experiments show that compaction can remove constraints that were followed while visible (Chen, 2026).

Our question is how much recorded decision information a successor can recover from a shorter representation. Task completion alone can miss the cost of reacquiring compressed-away state in controlled agent experiments (Liu, 2026). We examine decision recovery in human-guided coding sessions. An omitted decision may not matter for a particular continuation; an agent may also fail to follow a decision that remains available. We therefore evaluate recovery separately from subsequent behavior.

We measure *decision retention* by asking the same questions about recorded decisions under each context representation. Reference answers are anchored to quotations from the earlier conversation. We check mechanically that those quotations occur in the record. Successors receive replay, a freeform summary, or a structured packet. A judge grades their answers without seeing the representation label. Packets with removed constraints and reversed decisions provide a corruption control.

We evaluate private coding sessions and SWE-chat (Baumann et al., 2026), a public dataset of real interactions between developers and coding agents. The central finding is that the tested summaries lose recoverable decisions, even when organized into explicit fields or given a larger requested budget.

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.

The loss varies by decision type: packets retain some rejected options and next steps better than freeform summaries, but retain less completed-state information.

The study contributes a decision-level evaluation and a paired comparison of context representations. It also tests two boundaries of the finding. Short behavioral probes ask whether replay's retention advantage accompanies fewer breaches; it does not. Beyond the context window, a policy comparison asks which parts of history remain accessible after selection. Public quizzes, result rows, and an offline analysis make the publicly released portion of the evidence inspectable. Together, these components separate what a handoff preserves from what a successor subsequently does with it.

2 Related Work

Takeover and session transfer. Handoff Debt (KC & Budathoki, 2026) compares repository-only, raw-trace, and note-based takeovers in interrupted agent-only SWE-bench runs. Its evidence on the effort of recovering prior work motivates a decision-level question: what information remains available to the successor? Our study uses human-guided sessions and measures constraints, rejected approaches, completed work, and next steps. It compares summary structure at a common prompted budget and varies both artifact authors and evaluators. SWE-Together (Wu et al., 2026) evaluates agents using recorded sessions and a user simulator. Our decision-retention measure assesses the transferred content; task-level evaluation is still needed.

Compaction and memory. Governance Decay (Chen, 2026) reports safety-constraint violations on a synthetic benchmark: 0% with full context, 30% overall after compaction, and up to 59% for individual models. A vendor guide discusses paraphrased-path risks (Morph, 2026); a user report describes context loss (jameslowman, 2025). We examine recovery of decisions from recorded coding sessions and test one concrete summary-generation procedure, including its response to a larger output budget. Memory systems (Packer et al., 2024; Chhikara et al., 2025; Zhou et al., 2025) and benchmarks (Maharana et al., 2024; Hu et al., 2026; Xu et al., 2026) study information access across long interactions. Our overflow experiment connects that problem to handoff by comparing recent history, merged summaries, and retrieval under a common context ceiling. Because these policies use different queries and realized lengths, the comparison measures complete context-selection policies.

Interaction and supervision. Tool-agent-user benchmarks (Yao et al., 2024; Barres et al., 2025) use simulated interactions, whose reliability can affect evaluation conclusions (Seshadri et al., 2026). Recorded human–agent sessions anchor decisions in actual development work. They also introduce selection and privacy constraints. Manager-agent systems address the coordination of ongoing work (Masters et al., 2025). Human attention is also a concern in research on fan-out, supervisory control, and levels and limits of automation (Olsen & Wood, 2004; Sheridan, 1992; Parasuraman et al., 2000; Bainbridge, 1983). These concerns motivate transfers that require limited operator attention. The present study addresses the narrower information-recovery problem that such transfers create.

3 Measuring Decision Retention

Suppose a user has ruled out a dependency: can a successor recover that constraint from its context? Whether it follows the constraint during later work is a separate test. To measure recovery, we fix a handoff point, vary the supplied context, and check answers to shared questions against the earlier conversation.

Choosing a handoff point. A successor's takeover of ongoing, human-guided coding work is *session adoption*. Eligible sessions have at least three real user prompts, one file-mutating action, and twenty minutes of wall-clock span.[1] We place the handoff point, or *cut*, at 60% of the session's user prompts. The preceding conversation, the *prefix*, supplies the transfer history and evidence for our questions.

[1] In the census, 39/236 (16.5%) of one researcher's substantial local sessions qualify, compared with 0/176 from an automated pipeline. Of the 5,851 public SWE-chat sessions in our census (Baumann et al., 2026), 2,281 (39.0%) qualify. These are workflow-specific eligibility counts, not successful-handoff rates.

Varying the successor's context. We compare keeping the history with compressing it. **Replay (A)** supplies the rendered prefix, with tool outputs capped at 400 characters and a 300K-character tail window. **Freeform summary (B)** condenses it for continuation with a 550-word prompt budget. **Structured packet (C)** uses the same prompt budget. Its fields cover goal, state, completed work, next actions, constraints, rejected options with reasons, acceptance criteria, and key artifacts. The stated caps limit replay. Summary budgets specify requested lengths. The **corrupted packet (Cx)** checks whether the evaluation detects damaged decisions. It deterministically replaces C's next actions with rejected options and empties the rejection and constraint lists. The letters label experimental conditions.

Grounding questions in the history. Questions ask about abandoned approaches and reasons (*rejected*), user-imposed rules (*constraint*), completed work (*state*), and agreed next steps (*next*). A generator reads the rendered prefix and produces each question, a concise reference answer, and a verbatim evidence quote. For example, "What execution-control constraints did the user impose on the repair work?" has the reference answer "work directly in the current session; no background tasks or subagents," anchored to the user's instruction. The verifier locates each quote by substring matching after normalizing case and whitespace. Quotes must contain at least 15 normalized characters. We retain a cut only if it yields at least five verified questions. At least two must concern rejections, constraints, or next steps, the categories affected by corruption. Matching verifies text presence. Question adequacy and quote interpretation remain labeling judgments.

Comparing answers. A successor answers the shared questions from one condition's context, with "unknown" allowed. A judge scores answers 1, 0.5, or 0 against their anchored references without seeing the condition label. The average score measures *decision retention*. The original protocol averages admitted graded items (*question-pooled retention*). The subsequent analysis also averages within each cut (*cut-level retention*). For a paired comparison, we first average grades within each cut and condition, then subtract the two condition scores for that cut. The reported mean gives each cut equal weight, even when quizzes differ in size. Bootstrap intervals resample cuts; sign tests count positive and negative differences, excluding ties. These analyses keep a cut's questions together rather than treating them as independent observations. Reusing questions with another successor or judge adds no independent cuts. Pooled $\pm$ values are descriptive 95% normal-approximation half-widths: 1.96 times the item-level standard error. They are not the uncertainty estimates used for the paired comparisons.

The score estimates recovery of the admitted, quote-anchored decisions from each supplied context. It does not estimate the fraction of every decision in a session that survives. Replay is an evidence-access reference: it contains the same rendered record used to construct the questions. Comparing shorter representations with it measures changes in recoverable decision information, including the effect of reduced context size. It does not isolate format at matched realized lengths. Figure 1 shows this shared source.

Checking sensitivity to decision loss. If the questions detect the affected decisions, damaged packets should score below intact ones. They do: retention is 0.640 versus 0.813 on private cuts and 0.689 versus 0.815 across 32 public cuts. Rejected-item retention falls to 0.182 in the private control. An earlier measure compared file/command overlap between proposed next actions and the actual continuation. Orientation actions dominated, and corruption did not change that score. The questions instead target the affected constraints and choices. The next section specifies the session samples and model comparisons for this evaluation.

4 Experimental Design

Corpora and sampling. The primary comparison contains ten private cuts from bilingual research-engineering sessions (7–332 user prompts) and 32 public SWE-chat cuts (3–523 prompts). Sessions are selected at evenly spaced ranks after sorting eligible candidates by prompt count. The public sample comprises ten initial and 22 extension cuts. Appendix B records candidate exclusions and admission counts. We also evaluate 12 local Codex command-line interface (CLI) sessions (3–735 user messages). The record reports the same rendering caps and admission gates. All 12 are admitted, yielding 93 questions.

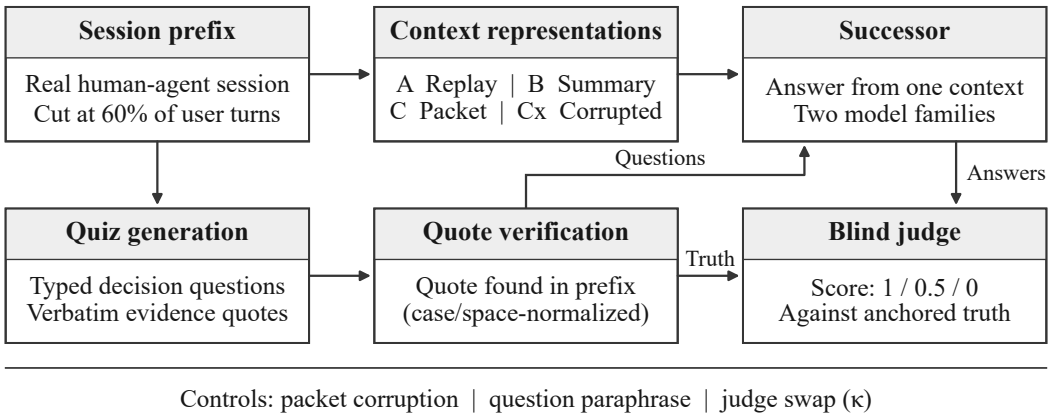

Figure 1: Decision-retention evaluation. A: transcript replay; B: freeform summary; C: structured packet; Cx: corrupted packet. These labels identify experimental conditions. Questions are generated from the session prefix, and their evidence quotes are mechanically verified against it. Successors answer from one condition's context. A blind judge grades against anchored reference answers. Corruption, paraphrase, and judge-swap controls assess complementary sources of measurement error.

Comparisons and model roles. The primary analysis combines the ten private cuts and 32 public cuts. A separate 20-cut bank contains the private ten and initial public ten. Context artifacts are freshly generated and saved for each cut. GPT-5.6 authors these artifacts. GPT-5.6 (schema-constrained CLI) and Claude Opus 5 (headless, tools disabled) each answer and judge the same bank, forming a 2×2 matrix. Re-authoring the artifacts with Opus 5 tests sensitivity to the artifact author. These repeated evaluations are sensitivity checks on overlapping sessions. Table 1 distinguishes their sample sizes from the primary 42-cut comparison.

The remaining experiments test budget, behavior, and transfer beyond the replay window. The budget sweep adds 1100- and 2200-word packet prompts on the ten private cuts. Behavioral probes use the session's actual next user requests. The overflow comparison uses five histories at 403K–973K characters. An extension tier adds three at 357K–397K. Quizzes are quote-verified within each window, and all policies share a 300K-character context ceiling. The protocol document records contracts v1–v6 as frozen before the corresponding experiments, including pre-data amendments and the subsequently invalidated control.

Item audit and wording control. An agent audit of 30 items marks 28 fully clean and two with minor defects; none is marked invalid or mistyped. The sample and checks appear in Appendix B. This is not independent human validation. For the wording control, GPT-5.6-Sol is instructed to rephrase each of the 151 questions in the 20-cut bank while preserving its decision target and identifier. All 151 strings change. Contexts, reference answers, and evidence quotes remain fixed. Replay and packet conditions are then answered and graded with the same procedure. Target preservation is an instruction to the generator, without an independent human equivalence audit.

5 Results

5.1 How much decision information survives the handoff?

The primary comparison covers 42 cuts. Replay exceeds structured packets by 0.178 in mean cut-level retention (bootstrap 95% CI $[0.140, 0.218]$). The sign test gives $38+/1-$ and $p = 1.5 \times 10^{-10}$. Against freeform summaries, the mean advantage is 0.129 ($[0.093, 0.168]$, $p = 7.9 \times 10^{-9}$). Intact packets exceed their corrupted versions by 0.141 ($[0.101, 0.183]$). Figure 2 shows the individual replay-minus-packet differences, including three ties and one negative difference.

Across the 32 public cuts and 246 questions, pooled replay retention is 0.994 ± 0.007, compared with 0.815 ± 0.041 for packets. The gap of 0.179 exceeds the pre-specified public replication criterion

Table 1: Evaluation samples. Questions are retention items per condition, not behavioral audit denominators. Rows overlap: the public and matrix samples are subsets of the primary comparison, and the sweep reuses its private cuts. The matrix regenerates artifacts on its 20 cuts. Counts must not be added.

Evaluation	Cuts	Questions
Primary within-window comparison	42	321
Public subset (SWE-chat)	32	246
Successor–judge matrix subset	20	151
Private budget-sweep subset	10	75
Codex-format comparison	12	93
Overflow: longer / near-window tiers	5 / 3	60 / 36

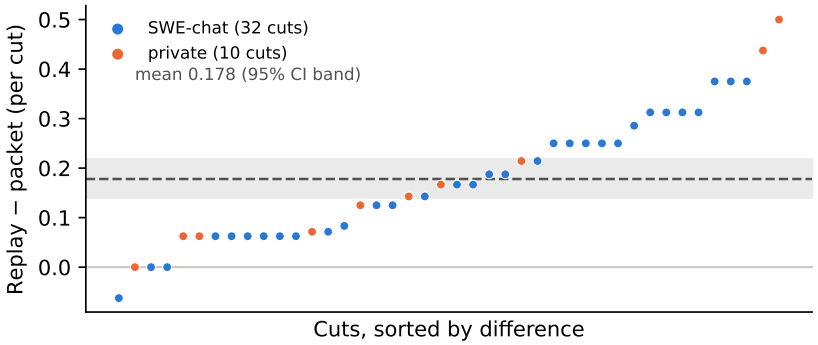

Figure 2: Replay-minus-packet differences for all 42 within-window cuts, sorted by magnitude. The difference is positive in 38 of 39 non-tied cuts. Cuts from both corpora occur throughout the range. The shaded band shows the bootstrap 95% CI around the mean difference of 0.178.

of > 0.05. A later offline reanalysis of these released rows gives a cut-averaged difference of 0.178 (bootstrap 95% CI $[0.137, 0.219]$), with 29 positive differences, one negative difference, and two ties. This checks the publicly inspectable portion of the historical finding. It uses the same observations and does not reconstruct the ten withheld private cuts. Appendix C specifies the reanalysis.

Successor and judge sensitivity. On the 20-cut persisted bank, replay has the highest score in all four successor–judge cells (Table 2). The cut-level replay–packet gap ranges from 0.086 to 0.173. Opus-as-judge cells assign higher scores and narrower replay gaps, showing that the magnitude depends on the evaluator. The primary 42-cut analysis and this matrix use overlapping sessions but separate experimental runs.

Question wording. On the 151 rephrased items, replay scores 0.970 and packets 0.798. Both scores decline, but their gap of 0.172 matches the original GPT-5.6 successor–judge cell to three decimal places. The observed replay advantage therefore persists under this wording intervention.

Artifact author. The recorded re-authoring comparison with Opus 5 holds the cuts, frozen prompts, and prompt budget fixed. With GPT-5.6 as successor and judge, packet retention rises from 0.811 to 0.871. Replay remains ahead for both successors ($\mathrm{A} - \mathrm{C}_{\mathrm{opus}} = 0.113$ and 0.119). Packet-minus-summary differences are $-0.010/-0.020$. Corruption-control separations are 0.142/0.136. These historical results suggest sensitivity to the artifact author without closing the replay gap. The underlying re-authoring rows are absent from the release, limiting independent verification.

Session ecosystem. The Codex-CLI comparison covers 12 sessions and 93 questions per condition. Reported retention is 0.984 ± 0.018 for replay, 0.823 for summaries, 0.790 for packets, and 0.581 for corrupted packets. The replay–packet difference is 0.194, the intact–corrupted difference is 0.210,

Table 2: Replay has the highest retention in every successor×judge cell. A: replay; B: summary; C: packet; Cx: corrupted packet. Results use the 20-cut artifact bank (151 questions per arm per cell). The primary 42-cut statistics appear in the text. The corruption control (Cx) separates in every cell. Cross-family judge agreement is $\kappa = 0.755$ on GPT-5.6 answers and 0.695 on Opus answers (n=604 pairs each).

Successor × Judge	A	B	C	Cx	A−C	A−C (cut)
GPT-5.6 × GPT-5.6	**0.983**	0.828	0.811	0.659	0.172	0.173
GPT-5.6 × Opus 5	**0.997**	0.854	0.871	0.719	0.126	0.125
Opus 5 × GPT-5.6	**0.990**	0.858	0.831	0.689	0.159	0.157
Opus 5 × Opus 5	**1.000**	0.894	0.914	0.755	0.086	0.086

Table 3: Decision retention by type on the 32 public cuts. Each row reports question-pooled means for the same items under replay (A), freeform summary (B), structured packet (C), and corrupted packet (Cx). Counts reflect the admitted quizzes. These descriptive subgroups share cuts.

Decision type	Questions	A	B	C	Cx
Constraint	61	1.000	0.902	0.869	0.656
Completed state	117	0.987	0.953	0.786	0.769
Rejected option	41	1.000	0.805	0.854	0.573
Next step	27	1.000	0.667	0.759	0.593

and the packet–summary difference is -0.032. The reported ordering extends to this rollout format. Only aggregate scores are released, without the private per-item results.

5.2 Does packet structure improve on a freeform summary?

The equal-prompt-budget comparison tests the value of the chosen field structure. Packet-minus-summary differences range from -0.027 to $+0.020$ across the four matrix cells, changing sign with the judge. On the public corpus, packet retention is 0.815 versus 0.884 for the freeform summary. Across all 42 cuts, the mean paired difference is -0.049 (bootstrap 95% CI $[-0.080, -0.017]$). The paired sign test gives $p = 0.052$. The mean reflects the magnitudes of paired differences. The sign test reflects the balance of wins and losses.

Table 3 shows where the public-corpus aggregate difference arises. Packets outperform freeform summaries on rejected options and next steps, while freeform summaries retain more constraints and completed-state information. State accounts for 117 of the 246 questions, so its larger loss weighs heavily in the aggregate. This is the mixture in the admitted quizzes, not an estimate of how often each decision type occurs in coding work. The admission gate deliberately requires items sensitive to corruption.

The same breakdown helps interpret the corruption control. Scores fall most for constraints, rejected options, and next steps, the categories targeted by the intervention. The state score changes much less, but is not invariant. This pattern supports sensitivity to the damaged content. It does not establish complete quiz coverage or a causal account of the summary trade-off. In particular, the schema may change which information the generator includes, but field allocation was not independently manipulated. The category means are descriptive; shared cuts and unequal category sizes preclude treating them as four independent replications.

5.3 Does a larger packet budget close the gap?

The budget sweep holds the ten private cuts, generator, and schema fixed and changes the packet's prompted word budget. Budgets of 550, 1100, and 2200 words yield retention 0.813, 0.840, and 0.813 on the same 75 questions (Figure 3). At four times the initial budget, the gap to replay's 0.993 remains 0.180. Additional allowed output length does not close the gap for this procedure, and the observed response is non-monotonic.

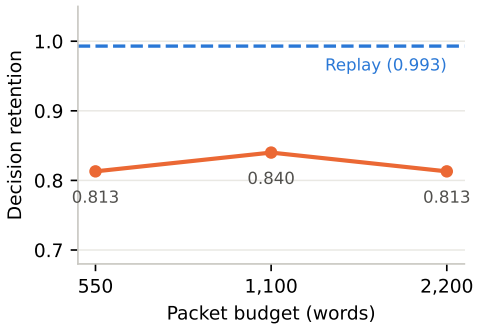

Figure 3: Increasing the prompted packet budget from 550 to 2200 words leaves a 0.180 gap to replay (10 private cuts, 75 questions per arm).

The cause of the loss remains open. A vendor guide's discussion of paraphrased-path risks (Morph, 2026) and synthetic evidence of constraint loss (Chen, 2026) motivate examining decision selection and restatement. The sweep tests one procedure's response to generation instructions. It does not isolate a mechanism or verify matched realized artifact lengths.

5.4 Does higher retention imply fewer behavioral breaches?

The behavioral probe asks whether higher-retention conditions also show fewer observed breaches. Under each context condition, the successor responds to the actual next user request. An auditor checks the proposed response against recorded decisions. The proposed actions are not executed, and task completion is not measured. The auditor marks next-step items and untouched topics as not applicable and excludes them from the breach denominator. Each condition therefore has its own denominator.

Corrupted packets have the highest observed breach rate, 16/90 (0.178), and produce five re-proposals among 14 audited rejected items. Intact conditions produce none among their 12–16 audited rejected items. Overall breach rates are 10/92 (0.109) for replay, 7/97 (0.072) for freeform summaries, and 7/94 (0.074) for structured packets. Thus, the predicted increase in breaches for intact summaries relative to replay is not observed.

The protocol-recorded three-turn extension uses ten private cuts. Before each next real user request, the successor receives excerpts of earlier requests and responses. These retain up to 800 characters per earlier request and 1,500 per response. Breach rates are 0.122 for replay, 0.103 for intact packets, and 0.139 for corrupted packets. The corrupted condition produces no re-proposals among six audited rejected items in this extension. Corruption has the highest overall rate in both probes, but its single-turn re-proposal pattern does not recur in the three-turn sample. These are descriptive, conditional rates with small, differing audit denominators and a limited continuation horizon.

The two tests ask different conditional questions. A quiz explicitly requests a recorded decision. A response audit asks whether a proposed continuation breaches a decision judged applicable to that response. A decision can be recoverable without being applied, or omitted from a summary without becoming relevant to the next request. Excluding untouched topics also changes which opportunities enter each condition's breach rate. The observed reversal therefore cannot show that summaries protect against breaches, or that the conditions are behaviorally equivalent. Establishing either would require common opportunities to apply decisions and executed continuations.

5.5 Which policies retain decisions beyond the replay window?

Policies and information access. The overflow experiment compares context-allocation policies under a common 300K-character ceiling. It uses the five longer and three near-window histories described above. In this regime, tool outputs are rendered with a 4,000-character cap rather than the within-window cap of 400. Tail replay keeps the last 300K characters.

Table 4: Retention beyond the replay window. The first two columns separate the overflow tiers; the final column pools their 96 questions. Question-conditioned retrieval uses evaluation-item queries. Blind retrieval uses the working context.

Policy	Longer histories (5 cuts, 60 Q)	Near window (3 cuts, 36 Q)	All eight (96 Q)
Tail replay	0.717	0.903	0.786
Merged summary	0.500	0.667	0.563
Blind retrieval	0.575	0.875	0.688
Question-conditioned retrieval	**0.817**	**0.917**	**0.854**

The merged-summary policy divides history into four portions and requests about 250 words per portion. It merges the summaries into about 550 words. The released code supplies only the first 200K characters of each portion. For histories longer than 800K, this would omit input before summarization. That limit qualifies the comparison. The private summary inputs are unavailable, so historical execution of this truncation cannot be verified.

Question-blind retrieval uses lexical matching (BM25). Its working query contains the final two user requests, clipped to 4,000 characters. It selects the top 40 chunks without seeing evaluation questions. Question-conditioned retrieval takes the union of the top eight chunks for each quiz item. Chunks contain 2,000 characters before position labels and are combined in transcript order. The blind query is available at adoption time. The conditioned policy can instead use what the evaluation asks for. Realized context lengths also differ. Merged summaries use 1.9K–5.3K characters, compared with 300K for tail replay. Blind retrieval uses 76.2K–80.6K and conditioned retrieval 66.5K–129.0K. The comparison changes context length and selection rules as well as queries.

Observed policy differences. Table 4 reports each tier separately and the pooled 96 questions. Blind retrieval scores 0.688, exceeding merged summaries by 0.125. Across sessions it has six wins, one tie, and one loss. Inspection of the losing case suggests a stale working query, an explanation that remains to be tested. Blind retrieval's pooled retention is below tail replay's 0.786. Conditioned retrieval scores 0.854 and exceeds the merged summary on all eight sessions. The mean paired difference is 0.292 (bootstrap 95% CI $[0.203, 0.391]$; sign-test $p = 0.008$). Against tail replay, the difference is 0.068, with CI $[-0.005, 0.141]$. This interval includes zero.

The tier breakdown shows a larger tail–conditioned gap on longer histories than near the window boundary. Recent history retains much of the evaluated information in the latter tier. Figure 4 in the appendix plots these tiers and the mismatched-summary control. Given the access and length differences above, the results compare complete policies rather than isolate the effect of retrieval itself.

Failed and replacement controls. The overflow comparison includes a mismatched summary from another session as its validity control. An attempted field-level control appended "reversal" lines containing rejected-item ground truths. That construction leaked the answers and scored 1.000 on targeted items. It cannot establish sensitivity to corruption. The failed control and its outcome are retained in the record rather than interpreted as a valid counterpart to the within-window corruption experiment.

5.6 What is the input-cost trade-off?

On the ten private cuts, packet answer calls average approximately 17.8K input tokens, compared with 64.5K for replay. About 16K tokens per call are fixed CLI overhead. After subtracting that overhead, the content-token ratio is about 0.04; the call-total ratio is 0.28. These counts concern answer calls and exclude generation and indexing costs. In the five longer overflow sessions, question-conditioned retrieval averages about 48.6K input tokens per answer call, compared with 17.7K for merged summaries and 111.1K for tail replay. Compression substantially reduces answer input. The retention comparisons show how much information remains accessible under each policy. These data do not estimate end-to-end operational cost or the cost of completing a task.

6 Implications for Handoff Evaluation

Evaluate what a summary must preserve. The decision-type breakdown shows why an evaluation should specify the decisions the successor needs and report their recovery separately. The quote-anchored procedure compares a candidate representation with its source record. The corruption control checks sensitivity to selected damage; the paraphrase control checks dependence on question wording; the successor–judge matrix checks evaluator sensitivity. These controls address different concerns and do not replace validation of which decisions matter for the task.

Separate archival access from working context. The original record keeps omitted decisions available for inspection; a summary can organize the immediate task with less input. Combining a recent tail with retrieval of older evidence is thus a candidate to test, not an evaluated recommendation. Its retrieval query must be available at takeover time; evaluation-question access would change the problem. The input-cost results also exclude generation and indexing, which such a comparison must include.

Use controls that distinguish explanations. Three follow-up tests target specific ambiguities. First, measure realized artifact lengths and compare authors and schemas at matched lengths. This would separate failure to use a requested budget from differences in content selection. Second, verify that summaries receive all quiz anchors, then match retrieval length and chunk allocation while varying the query. This would separate source coverage and context volume from query relevance. Third, pair recovery questions with common opportunities to apply the same decisions in executed continuations. This would test whether access predicts compliance and task progress. These proposed experiments remain necessary to distinguish the competing explanations.

7 Limitations

Measurement. Questions and replay derive from the same rendered prefix. The evaluation does not cover information excluded by the rendering caps. Mechanical quote matching does not validate every interpretation of a quote. The 30-item agent audit and paraphrase control address some labeling and wording concerns. Neither provides independent human validation or establishes that the quiz covers all consequential decisions. Behavioral evidence covers one and three subsequent turns, with no executed task-completion comparison.

Evaluators and sampling. Judge swaps alter both gap magnitudes and the packet–summary ordering. Cross-family agreement is $\kappa = 0.70$–0.76, leaving evaluator choice a source of uncertainty. Selection by eligibility and prompt-count rank can still affect the results. Private sessions come from one operator. Public SWE-chat sessions come from developers who opted into logging. Results from the three corpora and two CLI ecosystems should be interpreted within those populations.

Policy and evidence coverage. The packet experiments test one schema, two artifact authors, and a limited budget sweep. Overflow results cover eight sessions, one lexical retriever, and one summarize-then-merge procedure; their questions favor early and middle history. Generalization to other compressors, retrievers, and decision distributions remains untested. Verification also differs across results: public primary rows can be reanalyzed, whereas private results are available only as aggregates. The re-authoring result files and the calculation supporting Codex-format uncertainty are absent. These evidence gaps limit independent checking of the corresponding sensitivity claims.

8 Conclusion

A handoff can retain a readable account of the work while losing specific recorded decisions. The tested summaries show this loss, with trade-offs across decision types that their aggregate scores conceal. Quote-anchored questions make the loss inspectable, and the public-row reanalysis makes part of the comparison directly reproducible. Short behavioral probes do not establish a corresponding benefit for replay during continued work. Preserving access to the source record is therefore a bounded implication for decision recovery; choosing a deployment policy still requires evidence about actions and total cost.

AI USE STATEMENT

Large language models were used as part of the experimental evaluation to generate and compare experimental outputs across different LLMs. Generative AI tools were additionally used only for language polishing, grammar correction, and improving the clarity and readability of the manuscript. They were not used to generate the scientific content, design the methodology, interpret the results, or determine the conclusions. The authors reviewed all AI-assisted edits and take full responsibility for the final manuscript.

REFERENCES

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. doi: 10.1016/0005-1098(83)90046-8.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. doi: 10.48550/arXiv.2506.07982.

Joachim Baumann, Vishakh Padmakumar, Xiang Li, John Yang, Diyi Yang, and Sanmi Koyejo. SWE-chat: Coding agent interactions from real users in the wild. *arXiv preprint arXiv:2604.20779*, 2026. doi: 10.48550/arXiv.2604.20779.

Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. *arXiv preprint arXiv:2606.22528*, 2026. doi: 10.48550/arXiv.2606.22528.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI*, 2025. doi: 10.3233/FAIA251160. arXiv:2504.19413.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257v4*, 2026. doi: 10.48550/arXiv.2507.05257. URL https://arxiv.org/abs/2507.05257v4. Version 4, June 28, 2026; first posted July 2025.

jameslowman. [Bug] auto-compact causes context loss and degraded performance. https://github.com/anthropics/claude-code/issues/13112, 2025. Claude Code GitHub issue 13112, opened December 5, 2025. Accessed September 26, 2026.

Dipesh KC and Anjila Budathoki. Handoff debt: The rediscovery cost when coding agents take over interrupted tasks. *arXiv preprint arXiv:2606.02875*, 2026. doi: 10.48550/arXiv.2606.02875.

Tobias Lindenbauer, Igor Slinko, Ludwig Felder, Egor Bogomolov, and Yaroslav Zharov. The complexity trap: Simple observation masking is as efficient as LLM summarization for agent context management. *arXiv preprint arXiv:2508.21433*, 2025. doi: 10.48550/arXiv.2508.21433.

Shuyu Liu. What does context compression cost an agent? interaction costs unrevealed by task-completion metrics. *arXiv preprint arXiv:2608.16370*, 2026. doi: 10.48550/arXiv.2608.16370.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024. doi: 10.48550/arXiv.2402.17753.

Charlie Masters, Advaith Vellanki, Jiangbo Shangguan, Bart Kultys, Jonathan Gilmore, Alastair Moore, and Stefano V. Albrecht. Orchestrating human-AI teams: The manager agent as a unifying research challenge. In *Proceedings of the 7th International Conference on Distributed Artificial Intelligence (DAI)*, 2025. doi: 10.1145/3772429.3772439.

Morph. Context compaction: The technical guide to deletion-based context management. https://www.morphllm.com/context-compaction, 2026. Technical guide. Accessed September 26, 2026.

Dan R. Olsen, Jr. and Stephen Bart Wood. Fan-out: Measuring human control of multiple robots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 231–238, 2004. doi: 10.1145/985692.985722.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560v2*, 2024. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560v2. Version 2, February 12, 2024; first posted October 2023.

Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, 2000. doi: 10.1109/3468.844354.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations. *arXiv preprint arXiv:2601.17087*, 2026. doi: 10.48550/arXiv.2601.17087.

Thomas B. Sheridan. *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, 1992.

Yifan Wu, Zhuokai Zhao, Songlin Li, Ho Hin Lee, Jiacheng Zhu, Shirley Wu, Tianhe Yu, Serena Li, Lizhu Zhang, Xiangjun Fan, and Shengzhi Li. SWE-Together: Evaluating coding agents in interactive user sessions. *arXiv preprint arXiv:2606.29957*, 2026. doi: 10.48550/arXiv.2606.29957.

Wujiang Xu, Yu Wang, Kai Mei, Kaiqu Liang, Zhenting Wang, Mingyu Jin, Han Zhang, Shi-Xiong Zhang, Wenyue Hua, Sambit Sahu, and Dimitris N. Metaxas. MemGym: a long-horizon memory environment for LLM agents. *arXiv preprint arXiv:2605.20833*, 2026. doi: 10.48550/arXiv.2605.20833.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. doi: 10.48550/arXiv.2406.12045.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841v2*, 2025. doi: 10.48550/arXiv.2506.15841.

A Protocol and Artifact Provenance

One workstation's session store contained 3,311 auto-compaction events across 1,810 sessions. These counts describe only that local store.

The protocol record includes pre-data amendments to the corruption-sensitive admission rule and to the fixed-budget regime after the size census. It also retains the invalidated control construction and its reported outcome. These records document the experiment history. The release lacks independent timestamps establishing when each contract was frozen. It also omits private inputs and result rows needed to reconstruct that history.

Two provenance gaps concern sampling and model identification. The saved selection summary reports 23 additional public candidates. The released preparation code targets 24. Resolving this discrepancy requires the original candidate and skip logs. The Claude drivers request the mutable model alias `opus`. The manuscript retains the recorded Opus 5 designation, but the supplied result rows do not independently establish the resolved model version for each call.

Reported end-to-end model expenditure is approximately 44M input tokens. This includes a discarded round caused by a headless-output parsing defect. The revised parser handles three JSON shapes emitted by the successor family. This historical total differs from the answer-call costs. Editorial checks use saved experimental artifacts and make no model calls.

B Selection and Item-Audit Details

Private selection excludes earlier instrument-probe sessions, files larger than 12MB, and sessions modified within the preceding two hours. This leaves 24 fresh candidates. Twelve are selected. One yields no verified questions and another only state questions, leaving ten cuts.

The public selection record lists 12 initial and 23 additional candidates. The second batch excludes earlier selections. Admission removes two initial candidates and one additional candidate. One has fewer than five verified questions; two have fewer than two corruption-sensitive items. The public batches therefore contain ten and 22 cuts.

Item audit. The saved 30-item audit sample uses seed 20260729 and contains 14 state, seven constraint, five rejected-option, and four next-step items. The Fable agent reads each item and performs three full-quote spot checks. It marks 28 items fully clean. Two have minor defects: an answer clause extending beyond its quote and an ambiguous no-op next step. None is marked invalid or mistyped. This is an agent audit, not independent human validation.

C Public-Row Reanalysis

The later reanalysis validates 128 complete cut–condition rows across 32 public cuts (246 questions per condition). Checks cover grade values, question types, cut identifiers, pooled means and descriptive half-widths, and the historical source hashes of both primary result files.

The paired check averages replay-minus-packet differences in cut-level means. Cuts, sorted by session identifier, are resampled with replacement 10,000 times using seed 20260729. Percentile endpoints use sorted bootstrap means at zero-based indices 250 and 9750. This later analysis is neither preregistered nor an independent replication. Table 3 pools grades by decision type without category-level inferential tests.

Run `python3 code/public_retention_reanalysis.py` from the supplement directory. It requires only the Python standard library and writes derived JSON with input hashes and settings. It makes no model calls or data changes. Absent prefix and artifact texts prevent checking original quote occurrence and generated lengths. Withheld private results cannot be reconstructed.

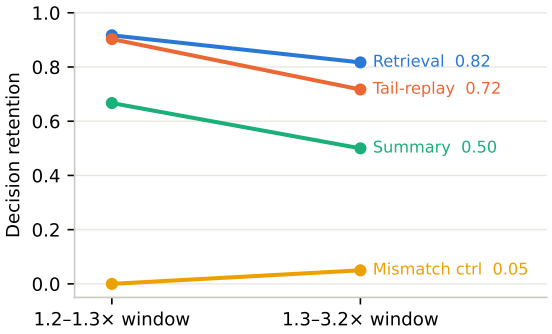

Figure 4: Question-conditioned retrieval preserves more decisions than merged summaries in both overflow tiers; tail replay is closest to retrieval near the window boundary (5 + 3 sessions, 96 questions per arm). Blind retrieval is reported separately in the text.

D Overflow Visualization and Supporting Tool

The supporting tool indexes two providers' sessions, tags eligibility, supports resumption, and counts compactions. Separate metadata records ownership, permissions, acceptance criteria, and escalation policy; an inbox collects operator requests. Tool availability is not evidence of better adoption outcomes.