RECURSIVELY UPDATED SUMMARIES CAN AMPLIFY A REMEMBERED BETRAYAL'S EFFECT ON A QWEN3-8B AGENT'S STATED WILLINGNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

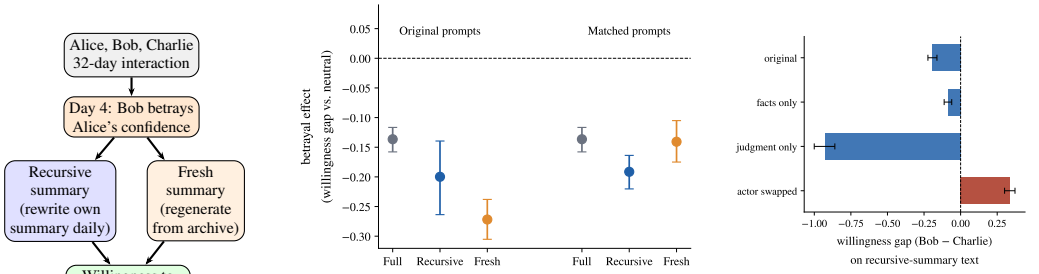
Long-horizon language-model agents compress their interaction history into summaries. We ask whether a single remembered relational event changes an agent's later willingness to rely on the offending party, and whether two common summarization procedures preserve or distort that change. In a scripted three-agent simulation on Qwen3-8B, with a scenario family ($n = 30$ lexical variants of one script) as the unit of inference and a single-candidate digit-logit rating as the instrument, a deliberate betrayal on day 4 lowers Alice's willingness toward Bob by 0.46 on a 7-point scale; the gap between Bob and Charlie, the recipient of the disclosure, falls by 0.136 relative to neutral control (95% CI $[-0.158, -0.117]$; 30/30 families). An initial comparison suggested that both a recursively updated summary and a summary regenerated from the archive amplified this effect. A content audit revealed an instruction asymmetry between the summarizer prompts; in a preregistered replication with matched prompts the regenerated-summary result disappears on the checkpoint average, while a smaller recursive-summary amplification remains (-0.055, CI $[-0.075, -0.037]$, 27/30 families negative; the originally secondary contrast). A preregistered sentence-swap study shows that the verbatim wording of the retained event sentence does not by itself produce this difference; the responsible feature of the surrounding text is not isolated, so we report a difference between recursively updated and regenerated memory states, not an isolated effect of recursion. Minimal edits to the memory text show the effect is bound to whichever name occupies the offender role and depends on the confidentiality-request clause; a per-candidate decomposition shows that compact "judgment" sentences widen the gap mainly by raising willingness toward the comparator, a caution for gap-based instruments. On the original corpus, explicit recall of the event stays intact. The instrument could not be validated on other model families in our calibration protocol, so all findings are limited to one model and lexical variants of one scenario.

1 INTRODUCTION

Large language model (LLM) agents deployed over long horizons cannot retain every interaction verbatim: practical deployments compress memory into summaries, either recursively updated or regenerated from an archive. A basic open question for such agents is whether the *behavioral* influence of a single relational event – for instance, a collaborator's deliberate betrayal of a confidence – survives this compression, and if so, whether it survives *faithfully*, attenuated as detail is lost, or is instead distorted. Figure 1 summarizes the design and the answer.

We study this question with a controlled three-agent simulation (Alice, Bob, Charlie) in which a scripted critical event occurs on day 4 of a 32-day history, under one of four conditions, followed by scripted filler. In the betrayal condition Bob deliberately tells Charlie a confidence that Alice had asked him to keep. At each of four checkpoints Alice's memory is the full history, a recursively updated summary, or a summary regenerated fresh from the archive. We measure her willingness to rely on Bob, and on Charlie as a reference, with a single-candidate digit-logit rating and separately probe explicit recall, appraisal, and stated belief. The unit of inference throughout is the scenario family: 30 lexical variants of one script on Qwen3-8B.

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.



(a) Setup.

(b) Only recursive amplification survives on the checkpoint average.

(c) Edits to the event sentence in the recursive summary.

Figure 1: (a) A scripted relational event is compressed by two procedures. (b) Betrayal effect on willingness (family-level, 95% CIs, $n = 30$): with the two summarization prompts matched and averaged over checkpoints, the fresh-summary amplification disappears while the recursive-summary amplification remains. (c) Edits to the event sentence inside the generated recursive summaries ($n = 26$): the effect follows the name in the offender role, and deleting the confidentiality-request clause removes much of it.

Deliberate betrayal lowers Alice's willingness toward Bob, and also toward Charlie, who received the disclosure; the preregistered gap between the two is -0.136 on a 7-point scale, negative in every family. An initial comparison suggested this effect was *amplified* by either kind of summary. Auditing the summaries, we found the two summarizer prompts differ in an instruction unrelated to compression, and in a preregistered replication with matched prompts the regenerated-summary amplification disappears on the checkpoint average while a smaller recursive-summary amplification remains. That replication also showed that the matched recursive prompt makes the summarizer copy the event sentence verbatim, so a further preregistered study swapped the sentence between summary types: verbatim wording alone does not produce the difference.

Two further questions follow. Does the effect reflect forgetting? Explicit recall of who did what stays near ceiling, and stated appraisal and belief do not track the amplification. And which content in the memory carries the effect? Minimal edits to the event sentence show that it follows whichever name occupies the offender role and depends on the confidentiality-request clause. They also show that compact "judgment" sentences widen the Bob−Charlie gap mainly by raising willingness toward Charlie, because they lift the penalty the original sentence imposes on the recipient of the disclosure. The instrument could not be validated on three other model families in our calibration protocol, so we draw no cross-model conclusion.

Contributions.

- A family-level ($n = 30$) finding on Qwen3-8B that a remembered deliberate betrayal lowers stated willingness toward both parties to the disclosure, the offender most.
- A matched-prompt replication of the initial compression-amplification finding: the regenerated-summary difference disappeared under matched prompts; a recursive-summary amplification survives; and a sentence-swap study shows that verbatim wording of the retained event sentence does not by itself produce it.
- Minimal-edit evidence that the effect is bound to the offender role and to the confidentiality-request clause, together with a per-candidate decomposition showing that "judgment-only" rewrites act mainly on the comparator, a check we recommend for any gap-based instrument.
- A dissociation on the original corpus (recall intact, self-report untracked), audited linear probing (one circular target excluded), and a full record of the cross-model calibration attempts, distinguishing measurement failure from non-replication.

2 Related Work

Agent memory systems. Generative agents (Park et al., 2023) made persistent memory central to long-horizon agent behavior; MemGPT (Packer et al., 2023) and MemoryBank (Zhong et al.,

2023) page or compress memory once it exceeds a context window, and running summaries of the kind studied here are a standard component of such systems. Wang et al. (2023) show recursive summarization is effective for long-term dialogue memory, and LoCoMo (Maharana et al., 2024) benchmarks long-term conversational memory; both evaluate *recall*. We ask a different question: with recall intact, does the form of the summary change what the agent is willing to do?

Iterated rewriting. A recursive summary repeatedly conditions the model on its own prior output. Transmission-chain studies find that LLM retellings favor social and negative content (Acerbi & Stubbersfield, 2023), that iterated LLM outputs drift toward attractors (Perez et al., 2024), and that iterated generation amplifies model biases (Ren et al., 2024); training-time self-consumption degrades models in a related way (Shumailov et al., 2023; Alemohammad et al., 2023). Abstractive summarizers also hallucinate content (Maynez et al., 2020), which we observe here as fabricated filler events (Appendix F). None of this work measures a downstream behavioral consequence of a remembered relational event.

Social reasoning and instruments. TrustLLM (Huang et al., 2024) evaluates trustworthiness of static text; Xie et al. (2024) study LLM trust-game behavior as a disposition rather than as a consequence of a remembered event; Gandhi et al. (2023) probe theory-of-mind with free-text or comparative responses. Logit-scored ratings are sensitive to label and surface-form biases (Zhao et al., 2021; Holtzman et al., 2021) and to irrelevant context (Shi et al., 2023), which motivates our calibration gates and our single-candidate, forced-choice design after an A/B design failed in piloting (Appendix A).

Probing and input edits. Linear probing (Alain & Bengio, 2016) tests linear decodability with known pitfalls (Belinkov, 2022): decodability is not causal use. Our text interventions follow the counterfactual and contrast-set tradition of editing an input minimally (Kaushik et al., 2020; Gardner et al., 2020), applied to a memory text rather than a classification input; they license claims about which input content the behavior depends on, not about internal mechanism (Vig et al., 2020).

3 METHOD

3.1 SIMULATION AND SCENARIO DESIGN

Three simulated agents (Alice, Bob, Charlie) interact over a 32-day timeline. On day 4 a critical event occurs, drawn from one of four frozen templates (deliberate betrayal, accidental disclosure, positive protection, neutral control; full text in Appendix L); every other day is one identical routine-filler sentence. A *scenario family* is one topic noun phrase substituted into these templates; all other text is shared. The betrayal event reads: "On day 4, Bob deliberately told Charlie about {topic}, even though Alice had asked Bob to keep it private. Bob chose to share it on purpose, and Charlie now knows about {topic}." The 30 families are therefore lexical variants of a single scripted scenario, and $n = 30$ supports robustness across topics, not across scenario structures. Alice's memory at each of four checkpoints (days 4, 8, 16, 32) takes one of three forms: the full history (mean 948 tokens at day 32); a recursively updated summary, rewritten each day from the previous summary and that day's event; or a summary regenerated fresh from the whole archive at each checkpoint. Summaries are capped at 1,024 tokens. Summaries are generated by the same Qwen3-8B model with greedy decoding; the summarizer prompts are given verbatim in Appendix L. The main text analyzes betrayal vs. neutral control; the other two conditions are reported in Appendix C. A qualitative audit of the templates (Appendix B) found no apology, remorse, repair, or mitigation wording in any condition.

3.2 WILLINGNESS INSTRUMENT AND ESTIMANDS

Alice is asked, in a single prompt, how willing she is (1–7) to extend one of 12 kinds of reliance to one named candidate (Bob or Charlie, never both), with forward and reverse anchors and two paraphrases per template: 96 prompts per (family, condition, memory, checkpoint) cell. Each prompt is rendered with the model's chat template, thinking mode disabled, and scored in one forward pass by reading the logits at the answer position restricted to the seven validated single-token digits; the *score* is the expected digit under the renormalized digit probabilities, reversed for reverse anchors. No text is generated. The instrument passed a frozen 12-gate calibration battery on Qwen3-8B (Appendix A).

Two quantities are used throughout. The **willingness gap** $g = \bar{w}_{\text{Bob}} - \bar{w}_{\text{Charlie}}$ is the difference in mean score between the two candidates in one cell. The **betrayal effect** $\Delta = g_{\text{betrayal}} - g_{\text{neutral}}$ compares that gap with the neutral-control gap at the same memory condition and checkpoint, averaged over the four checkpoints unless a checkpoint is named. **H1** is Δ at full history. **H2** is the change in Δ from full history to each summary condition (fresh−full and recursive−full); the original preregistration named fresh−full primary and recursive−full secondary (Section 4.2).

Statistics. The scenario family is the sole independent unit; templates, paraphrases, anchors, checkpoints, and repeated passes are never treated as independent samples. Every test reports a family-cluster bootstrap 95% CI (5,000 resamples), a family-level sign-flip Monte Carlo permutation p (10^6 draws; reported as $p < 10^{-6}$, never $p = 0$), Cohen's d_z, leave-one-family-out sign stability, and Holm correction where several comparisons are made. Family counts by direction use a preregistered tie band of ± 0.10; because that band is a large fraction of the effects reported here, counts are descriptive and inference rests on the CIs and permutation tests. Because families differ only by a topic phrase and decoding is deterministic, between-family standard deviations are small (0.06 for H1), so d_z values are large; raw means on the 7-point scale are always reported alongside.

3.3 Memory-state probes

On the persisted memory texts (read-only), each cell was also scored for cued component recall, four-alternative recognition, actor–target binding, free recall, and five-item appraisal and belief batteries. All but free recall use the digit-logit mechanism above; free recall generates a short JSON answer that is scored deterministically against ground truth. Calibration: a synthetic known-answer case plus all 30 families' neutral states (9/9 calibration gates passed).

3.4 Memory-text interventions

Deterministic edits replace *only* the day-4 event sentence in a frozen betrayal memory text, leaving every other character byte-identical. Six conditions (Table 2; sentences in Appendix L) are applied to the full-history text and to the actual recursive summary: *original* (identity), *facts only* (evaluation clause deleted), *judgment only* (episodic detail deleted, one evaluative sentence kept), *neutralized* and *reversed* (evaluation clause replaced by a neutral or permissive one), and *actor swap* (Bob and Charlie exchange roles). A second set swaps the event sentence *between* the two summary types and varies the judgment sentence's vocabulary (Section 4.5). Every text is audited for leakage and structure before inference, and each identity condition is checked against the frozen measurement of the same text.

3.5 Representation probing

Final-layer hidden states at the last input token of one fixed willingness prompt per cell (1,440 cells; embedding and middle layers in the supplementary audit) are probed with linear models, with standardization and PCA (100 components) fit within each training fold of a family-grouped 5-fold cross-validation. Labels come from the separate probe passes above. Section 5 summarizes a forensic audit of this pipeline; Appendix J gives it in full.

3.6 Compute and provenance

All inference ran on H100 GPUs under Slurm. Every stage's plan, config, code, and (for interventions) input texts were version-controlled before its inference began; available SHA-256 manifests and run summaries are in the supplementary material (Appendix K).

4 Results

4.1 H1: deliberate betrayal reduces stated willingness

At full history, deliberate betrayal reduced Alice's willingness gap toward Bob relative to neutral control: $\Delta = -0.136$ on the 7-point scale, 95% CI $[-0.158, -0.117]$, $d_z = -2.29$, negative in 30/30 families (21 beyond the ± 0.10 band), sign-stable under every leave-one-family-out exclusion,

$p < 10^{-6}$ (Figure 2a). The gap is a relative measure: betrayal lowers willingness toward Bob by 0.46 and toward Charlie, who received the disclosure, by 0.33 (both 30/30 families; Appendix C). The question for the rest of the section is what happens to this effect when the history is compressed into a summary.

4.2 H2: AN INITIAL AMPLIFICATION FINDING AND ITS PREREGISTERED CORRECTION

Initial observation. Under the original summarizer prompts the betrayal effect grew as memory was compressed: -0.136 (full history), -0.200 (recursive summary), -0.272 (fresh summary); fresh$-$full -0.135, 95% CI $[-0.174, -0.096]$ (Table 1, first row; Figure 2b). Before treating this as a finding about compression we audited the summaries. The two prompts differ in an instruction unrelated to compression: the fresh prompt asks the model to preserve "all specific facts you can", the recursive prompt only asks for brevity, and at least one of five evaluative-significance words occurs in every family's fresh summaries (29/30 at day 32) and in no recursive summary (Appendix E).

Matched-prompt replication. We regenerated both summary types for all 30 families with prompts that are the union of the two originals' instructions, so nothing is added that was not already in one of them (Appendix L), and reran the identical instrument; full history is reused from the original run. Estimands, statistics, and pass/fail criteria were frozen and committed before inference. Two caveats: the plan declared fresh$-$full and recursive$-$full co-equal, whereas the original preregistration named fresh$-$full primary and recursive$-$full secondary; and both prompts and the generation ceiling changed at once, so the design tests whether amplification survives matched instructions without isolating which instruction was responsible. Under matched prompts the originally primary contrast is null and the originally secondary one survives (Table 1).

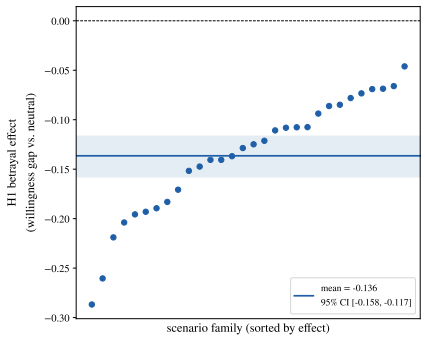
Table 1: H2: change in the betrayal effect Δ from full history to each summary type (7-point scale; $n = 30$ families; family-cluster bootstrap 95% CI; sign-flip permutation p). amp/att: families with contrast < -0.10 / $> +0.10$; neg: families with contrast < 0.

Contrast	Prompts	Mean	95% CI	d_z	p	amp/att/neg
fresh − full	original	-0.135	$[-0.174, -0.096]$	-1.23	$< 10^{-6}$	20/0/28
fresh − full	matched	-0.004	$[-0.039, 0.037]$	-0.04	0.84	5/3/19
recursive − full	original	-0.063	$[-0.126, -0.003]$	-0.37	0.049	11/6/19
recursive − full	matched	-0.055	$[-0.075, -0.037]$	-1.02	2×10^{-6}	5/0/27

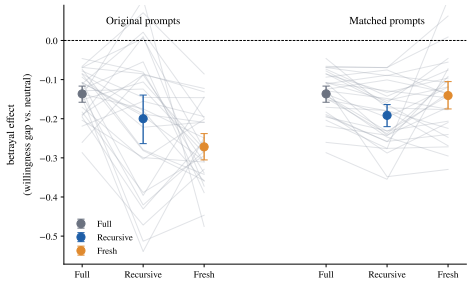
The surviving recursive$-$full effect is -0.055 (about 40% of H1; 25 of 30 families lie inside the ± 0.10 band); its CI excludes zero, 27/30 families are negative, and its sign survives every leave-one-family-out exclusion. It is present at every checkpoint and largest at day 4. The direct recursive$-$fresh difference is concentrated at day 4; the two summary types are similar at days 8, 16, and 32, and the checkpoint-averaged fresh$-$full null combines attenuation at day 4 ($+0.128$) with amplification at day 32 (-0.086) (Appendix F). The matched prompts did not make the two summary types alike (Table 3): matched recursive summaries are three times longer than matched fresh ones (258 vs. 86 tokens at day 32), fabricate filler events in 13/30 families, drop the event in 4/30 by day 32, and, because of the fact-preservation instruction, copy the day-4 template sentence verbatim in 26/30 families, whereas both fresh prompts always paraphrase it. The effect survives excluding the fabricating families (-0.044, $n = 17$) and is numerically larger within them (-0.069). **For the checkpoint-averaged H2 estimand, amplification is observed only in the recursive-summary condition; that condition also differs from the fresh condition in length, verbatim retention, and drift, and Section 4.5 tests the verbatim difference directly.**

4.3 EXPLICIT RECALL STAYS INTACT; SELF-REPORT DOES NOT TRACK THE EFFECT

A natural explanation for any change under compression is that the agent forgets the event. On the original summaries, explicit actor and intent recall stayed at 0.90–1.00 across all memory conditions and checkpoints. Because the probes read the same text that contains the event, recall here measures whether the event is accessible in the memory text; this rules out forgetting as the explanation in families where the event survives summarization. Stated belief (*willingness to depend*) correlates with H1 itself ($r = 0.450$, raw $p = 0.0125$) but not with the original fresh$-$full contrast ($r = 0.047$,



(a) H1: betrayal effect per family, full history.



(b) H2 under original vs. matched prompts; thin lines are families.

Figure 2: H1 and H2 with family-cluster bootstrap 95% CIs (30 families). Under matched prompts, averaged over checkpoints, the fresh-summary point returns to the full-history level while the recursive-summary point does not.

$p = 0.804$); perceived harmfulness correlates with that contrast ($r = -0.477$, raw $p = 0.0078$) but not after Holm correction across the eight tests ($p \approx 0.062$), and its level falls under compression (4.59–4.75 vs. 5.03) while the effect grew (Appendix C). With $n = 30$ a small relationship cannot be excluded; the claim is that none was detected. These probes were run on the original corpus, not on the matched one whose recursive contrast is now the effect of interest.

4.4 Which sentence carries the effect

Separately from why the two summary types differ, which sentence in the memory does the behavior depend on? We ran the six interventions of Section 3.4 on the full-history text (30 families) and on the actual matched recursive summaries at day 32, replacing the day-4 sentence in place; in four families the summarizer had dropped it, so no edit is definable and they were excluded before inference ($n = 26$). Both runs were preregistered, and in both the identity condition reproduces the earlier measurement of the same text to floating-point precision (1.9×10^{-16}, 2.6×10^{-16}). Table 2 and Figure 3 give the results.

Table 2: Interventions, paired vs. *original* within family. Gap g is the Bob−Charlie willingness gap in the betrayal state at day 32 (7-point scale; not the betrayal−neutral contrast Δ). Δg = condition − original; positive = weaker betrayal effect. Families: stronger / within ± 0.10 / weaker-or-redirected. Recursive base $n = 26$, full-history base $n = 30$. All Holm $p < 5 \times 10^{-6}$ except full-history *facts only* (raw $p = 0.00094$); every comparison leave-one-family-out sign-stable. Per-candidate changes in Appendix G.

	Recursive-summary text				Full-history text		
Condition	g	Δg [95% CI]	d_z	families	g	Δg [95% CI]	d_z
original	-0.192	–	–	–	-0.113	–	–
facts only	-0.087	$+0.105\ [0.084, 0.128]$	1.82	0/14/12	-0.093	$+0.020\ [0.010, 0.031]$	0.67
judgment only	$\mathbf{-0.929}$	$-0.737\ [-0.810, -0.664]$	-3.74	26/0/0	-0.564	$-0.451\ [-0.513, -0.386]$	-2.49
neutralized	-0.039	$+0.153\ [0.129, 0.178]$	2.36	0/5/21	-0.052	$+0.061\ [0.050, 0.072]$	1.96
reversed	-0.028	$+0.164\ [0.137, 0.190]$	2.33	0/5/21	-0.041	$+0.072\ [0.061, 0.083]$	2.20
actor swap	$\mathbf{+0.335}$	$+0.527\ [0.470, 0.589]$	3.37	0/0/26	$+0.148$	$+0.261\ [0.235, 0.288]$	3.46

Three things follow. First, the effect is bound to the offender role: *actor swap* redirects it to whichever name occupies that role in every family on both bases. Second, the confidentiality-request clause ("even though Alice had asked Bob to keep it private") matters: deleting or replacing it (*facts only*, *neutralized*, *reversed*) removes 55–85% of the effect on the recursive text and 18–64% on the full history, and the neutral and permissive replacements differ by about 0.01, so what the clause is replaced with matters less than that the original is gone; the intent markers ("deliberately", "on purpose") are retained in all three. Third, *judgment only* produces the largest gap on both bases, but its per-candidate decomposition tempers the headline: on recursive text Bob falls by 0.34 while Charlie *rises* by 0.39, because the sentence no longer names Charlie as the recipient, and the sentence

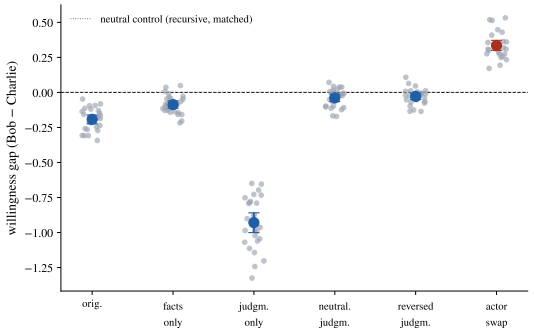


Figure 3: Interventions on the actual recursive summaries: all 156 family-level gaps (26 families × 6 conditions), family-cluster bootstrap 95% CIs.

contains the word *trust*, which the instrument also asks about. Section 4.5 separates these. Every intervention moves the gap further on the 258-token summary than on the 948-token history. These edits license claims about input content, not internal mechanism.

4.5 Swapping the sentence between summary types

The matched recursive prompt copies the day-4 template sentence verbatim, while both fresh prompts paraphrase it, so the recursive/fresh difference might be the wording of the retained sentence. A preregistered study (Appendix H) swapped the sentences: the verbatim template sentence was spliced into each fresh summary in place of its paraphrase ($n = 30$), and each family's own paraphrase into its recursive summary in place of the verbatim sentence ($n = 26$), everything else byte-identical; the identity conditions reproduce the earlier measurements exactly. Figure 4a: the verbatim sentence in a fresh summary produces no detectable change (-0.010, 95% CI $[-0.054, 0.035]$, $p = 0.68$), and the paraphrase in a recursive summary makes the gap *larger*, not smaller (-0.177, $[-0.232, -0.125]$, $d_z = -1.25$, 18/26 families beyond the band, the change almost entirely in Bob's rating). Verbatim retention of the template sentence therefore does not by itself produce the recursive/fresh difference, although wording clearly matters inside recursive text. The swap does not locate the difference: at day 32 the verbatim sentence yields the same betrayal effect in both summary types (-0.201 fresh, -0.193 recursive, relative to matched neutral gaps; Appendix H).

The same study added two judgment-sentence controls (Figure 4b). A judgment sentence with no trust or reliance vocabulary ("Bob did something concerning {topic} that seriously wronged Alice") still widens the gap on both bases (-0.496 recursive, -0.362 full history; every family), so the instrument's construct term is not required. But the per-candidate decomposition qualifies the *judgment only* result: every judgment variant, with or without Charlie named in it, widens the gap mainly by *raising* willingness toward Charlie ($+0.29$ to $+0.52$) while lowering Bob's by 0.00 to 0.34. The original sentence penalizes Charlie as the recipient of a deliberate disclosure; a judgment sentence lifts that penalty. Bob's own additional drop under a compact judgment is real but modest, and we do not claim that a compact judgment carries the effect. For any instrument built on a gap between two candidates, this decomposition is the check that distinguishes a change in the target from a change in the comparator.

5 Representation Probing

If behavior can differ from self-report, is the relevant information still present in the model's activations? We asked whether the probe labels of Section 3.3 are linearly decodable from Qwen3-8B's final-layer residual stream with families held out (Section 3.5). A forensic audit of the pipeline (Appendix J) changed which claims are defensible: it excluded the continuous willingness score as circular (its label is computed from logits that are a near-affine function of the activation being probed); replaced a cell-level permutation null with a family-block null, and, for two targets whose labels follow the same cell pattern in every family, with a within-family null; and added majority-class and random-feature baselines. Five distinct targets survive (Table 9): actor–target binding

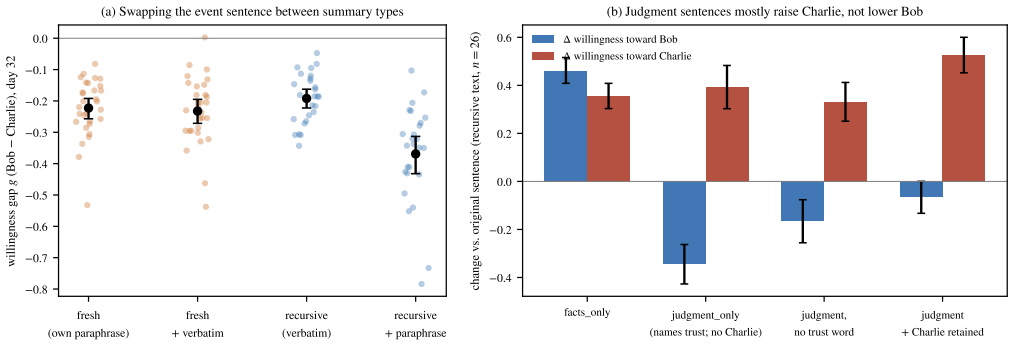


Figure 4: (a) Day-32 willingness gap under the four sentence-swap conditions (points are families; family-cluster bootstrap 95% CIs). (b) Change in willingness toward each candidate, relative to the original sentence, for the evaluation-clause deletion and three judgment sentences on recursive text.

(0.969 accuracy vs. 0.724 majority baseline; the cued actor-recall label is identical to it in all 1,440 cells and is not counted separately), perceived harmfulness and stated belief (R^2 0.959 and 0.783), and event condition and intentionality (0.981 and 0.990 accuracy), each with a valid permutation $p < 0.002$ at the 500-draw resolution used. Two caveats bound the result. The words defining the condition (*deliberately* vs. *accidentally*) are present in the scored context, so the last two targets show that lexical information is linearly preserved, not that the model infers the event abstractly; and because a held-out family differs from the training families only by a topic phrase, the held-out accuracy tests generalization across topics, not across scenarios. Decodability shows information is linearly present, not that the model uses it.

6 Cross-Model Measurement Attempts

Does the phenomenon hold beyond Qwen3-8B? We attempted to extend the Qwen3-8B instrument to Mistral-7B-Instruct-v0.3, Phi-4-mini-instruct, and Yi-1.5-9B-Chat, in three attempts spanning two instrument designs (Appendix I). The first attempt reused the primary instrument unmodified: all three non-Qwen models failed the same forward/reverse anchor-agreement gate (differences 0.972, 0.539, 0.374 against a tolerance of 0.35). The second design, a continuous reservation-threshold instrument scored by full-sequence log-likelihood, failed in its first version on all models including Qwen3-8B because its two continuation strings were not token-length matched; after that implementation defect was repaired and the strings token-matched, the second version still failed a construct-validity gate on Qwen3-8B itself (no family's cooperation curve crossed 50% within the incentive grid), which triggered the preregistered stop rule. No betrayal-condition data was generated for any model, so these attempts yield no evidence for or against the phenomenon on other architectures: a calibration failure is the absence of a valid measurement, not a non-replication. The instrument was developed and validated on Qwen3-8B.

7 Limitations

One model, one scenario. All findings are on Qwen3-8B, the model on which the instrument was validated (Section 6). The 30 families are topic variants of one scripted scenario with identical filler on every non-event day; a summarizer facing heterogeneous logs may behave differently, and the large d_z values reflect how alike the families are.

Small, relative effects. The surviving recursive−full effect is 0.055 on a 7-point scale, and the willingness gap penalizes both parties to the disclosure (Appendix C). The accidental-disclosure condition shows a positive trend, reported in Appendix C. Template word counts differ across conditions (32/36/25/19), which is not ruled out as a contributor for the protection and neutral comparisons.

Operative variable. The sentence-swap study shows that verbatim wording of the retained event sentence does not by itself produce the recursive/fresh difference; it does not identify what in the surrounding recursive text is responsible (length, position, fabricated content, or the update history).

Checkpoint effects are non-monotonic and are reported in Appendix F. Both prompts were changed at once in the replication, so it does not isolate which instruction was responsible for the original fresh-summary difference, and no experiment manipulates update history independently of the resulting text.

Scope of the probes and interventions. The recall, appraisal and belief probes were run on the original corpus, not on the matched one; on text that contains the event, near-ceiling recall indicates that the event is accessible in the text. The interventions edit one sentence; 2/30 recursive summaries contain another sentence referring to the event, left unchanged. Reversed judgment attenuated rather than reversed the effect.

No internal mechanism claim. The interventions establish input-level dependence; probing shows information is linearly decodable across topic variants, not that the model uses it, and we performed no activation-level intervention. All measures are ratings or forced-choice scores; nothing is claimed about felt experience.

8 Discussion and Conclusion

What a builder of agent memory should take from this study is narrow but concrete. In this setting, a running summary that is rewritten at every step widened the gap between an agent's willingness toward the offender and toward the recipient of the disclosure by a consistent amount, on the checkpoint average, that a summary regenerated from the archive did not, and verbatim retention of the event sentence does not by itself account for it. We therefore interpret this result as a difference between recursively updated and freshly regenerated memory states, not as evidence that recursion itself is the isolated causal variable. The recursively produced text is also longer, fabricates filler events in nearly half of the families from the first update, and occasionally loses the event altogether; regenerating from an immutable archive avoided all of these here. Explicit recall was probed on the original corpus only.

Three checks proved more informative than the headline contrasts and are worth reusing. Matching the two summarizer prompts removed a significant-looking effect that would otherwise have been reported as a property of compression. Splicing the event sentence between summary types tested wording directly. Decomposing a gap-based measure into its two candidates showed that the largest intervention effects came from the comparator, not the target. The decisive follow-ups are equally concrete: manipulate update history and summary length independently, re-run the recall and belief probes on the matched corpus, and calibrate the instrument on a second model family, which our calibration attempts did not achieve.

Ethics Statement

This work uses entirely synthetic, scripted simulations of interactions between three fictional named agents (Alice, Bob, Charlie); no human subjects, personal data, or real individuals' information is involved at any stage. All models used (Qwen3-8B, Mistral-7B-Instruct-v0.3, Phi-4-mini-instruct, Yi-1.5-9B-Chat) are publicly released, ungated checkpoints used under their open licenses (Apache-2.0 for Qwen3-8B, Mistral, and Yi-1.5-9B; MIT for Phi-4-mini), used only for research measurement and never fine-tuned, redistributed, or deployed. No paid or access-restricted APIs were used. We see no direct risk of harm from this work. The closest adjacent risk is that findings about relational-memory sensitivity in LLM agents could inform the design of agents that manipulate users' trust; the present work is diagnostic, measuring an existing behavioral tendency, and the interventions studied operate on the agent's own memory content, not on user-facing outputs. Negative and null results are reported in full.

Reproducibility Statement

Model identifiers and exact Hugging Face snapshot revisions, package versions, and decoding seeds are recorded in Appendix K. The supplementary run summaries record available model/revision, seed, configuration hashes, GPU type, timing, and output counts. Frozen plans, configuration files, analysis code, and available hash manifests are included in the anonymized supplementary material.

All prompts, the incentive grid (cross-model attempts), scoring rules, calibration thresholds, and statistical procedures were version-controlled before the corresponding inference began, at every stage of this project. Scenario family ($n = 30$) is the independent unit in every statistical procedure in this paper; no analysis treats templates, paraphrases, anchors, or checkpoints as independent samples. Monte Carlo results are reported as "no permutation among N draws was as extreme as the observed statistic; Monte Carlo $p < 1/N$" wherever zero more-extreme permutations were found, never as $p = 0$. Code and configuration for reproducing every reported number are included in the supplementary material (anonymized).

AI-Assistance Statement

We used generative AI tools to assist with implementation, experiment execution, diagnostic analyses, figure generation, literature discovery, and manuscript drafting. A human researcher formulated the research questions, determined the experimental design, reviewed intermediate results, authorized all follow-up experiments, manually reviewed the principal analyses, figures, and reported findings, and is responsible for the submission.

References

Alberto Acerbi and Joseph M. Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120 (44):e2313790120, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*, 2023.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*, 2023.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*, 2020.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of EMNLP*, 2021.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, et al. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of ACL*, 2020.

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:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

Jérémy Perez, Grgur Kovač, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. When llms play the telephone game: Cultural attractors as conceptual tools to evaluate llms in multi-turn settings. *arXiv preprint arXiv:2407.04503*, 2024.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J. Sutherland. Bias amplification in language model evolution: An iterated learning perspective. In *Advances in Neural Information Processing Systems*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of ICML*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*, 2023.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of ICML*, 2021.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

A Instrument Evolution and Calibration Gates

The measurement instrument evolved across generations, each preserved unmodified and hash-verified: an initial simulation and memory-condition pipeline; an A/B forced-choice willingness instrument, which failed calibration due to order and label effects; and the single-candidate, never-comparative digit-rating instrument used for every reported behavioral result. The main study of Section 4 used that instrument on 30 families.

The frozen 12-gate calibration battery: (1) digit labels validated single-token and mutually distinct; (2) deterministic repeat (spread tolerance 0.0); (3) expected scores finite and in $[1, 7]$; (4) synthetic strong-positive minus strong-negative ≥ 1.0 points; (5) synthetic ordering positive $>$ neutral $>$ negative; (6) forward/reverse agreement ≤ 0.35 points; (7–8) neutral candidate balance ≤ 0.35 aggregate / ≤ 0.75 per template; (9) no digit-probability degeneracy; (10) at least 5/7 digits non-negligible; (11) template-removal stability ≤ 0.20 shift; (12) all target families calibrated. Qwen3-8B passed all 12.

B SCENARIO TEXT AUDIT

All 30 families share four frozen condition templates (Appendix L) differing only in a substituted topic phrase (word counts before substitution: 32 betrayal / 36 accidental / 25 protection / 19 neutral). No apology, remorse, repair, mitigation, or transparency wording occurs in any family in either the betrayal or the accidental-disclosure text. Deliberate-intent marking (*deliberately/on purpose* vs. *accidentally/did not mean to*) is the only evaluative-adjacent language distinguishing these two conditions. No relational adjectives or explicit evaluative words (*good/bad/wrong/right*) appear anywhere. Thirteen families show generic workplace-vocabulary overlap with the probe templates (flagged, not exclusionary).

C MEMORY PROBES: FULL STATISTICS

Component recall, recognition, appraisal, and belief were scored at every (family, event condition, memory condition, checkpoint) cell on the persisted memory texts of the original run (no regeneration). Free recall is scored by deterministic keyword matching of the generated JSON description against the ground-truth event record.

Per-candidate H1. The betrayal effect is a relative measure. Averaged over checkpoints at full history, betrayal lowers willingness toward Bob by -0.464 (95% CI $[-0.518, -0.410]$, 30/30 families negative) and toward Charlie, the recipient of the disclosure, by -0.327 ($[-0.388, -0.268]$, 30/30 negative), from neutral baselines of 4.37 (Bob) and 4.38 (Charlie). H1 (-0.136) is the difference. Charlie is therefore not a neutral comparator: both parties to the disclosure are penalized, Bob more.

Relationship tests. Family-level ($n = 30$; 100,000-draw family-level permutation), eight tests in the Holm family. Against H1 (full history, day 32): *willingness to depend* $r = 0.450$, raw $p = 0.0125$; *perceived harmfulness* $r = 0.250$, $p = 0.187$. Against the original fresh−full contrast (fresh summary, day 32): *willingness to depend* $r = 0.047$, $p = 0.804$; *perceived harmfulness* $r = -0.477$, raw $p = 0.0078$. Against the accidental-disclosure gap: *perceived intentionality* $r = -0.246$, $p = 0.192$; *perceived protection* $r = -0.037$, $p = 0.848$; *perceived harmfulness* $r = 0.324$, $p = 0.080$; *expected reliability* $r = -0.184$, $p = 0.330$. After Holm correction across the eight tests the two smallest raw p-values adjust to approximately 0.062 and 0.087; none survives. Actor recall vs. H1 is undefined (zero variance: always correct). Note that the harmfulness correlation with the original fresh−full contrast has the sign a harm-driven account predicts (families rating the event more harmful amplified more); what does not fit such an account is the level: mean perceived harmfulness is *lower* under the two summary conditions (4.59–4.75) than at full history (5.03) while the effect was larger.

The accidental-disclosure condition. The main study found *accidental disclosure* trending positive, similar in sign to *positive protection*; we report the appraisal profile below without a confirmatory test. Day-32 full-history appraisal means by condition (betrayal / accidental / protection / neutral): perceived harmfulness 5.03 / 4.39 / 3.50 / 1.33; perceived intentionality 5.46 / 3.56 / 4.47 / 2.90; perceived protection 1.35 / 3.38 / 5.36 / 3.71; expected reliability 3.18 / 3.51 / 3.83 / 3.23. The model classifies accidental disclosure as non-deliberate in every family, and its appraisal profile sits closer to neutral and protective than to betrayal; none of the four accidental-specific correlations above reaches significance. Template length differs across conditions and is not ruled out as a contributor for the protection/neutral comparison.

D ORIGINAL H2: LEAVE-ONE-TEMPLATE-OUT CHECK

The preregistered plan specified leave-one-template-out for both H1 and H2; the original analysis run computed it only for H1. Run as a read-only addendum on the same frozen data, the original fresh−full contrast remains negative under removal of each of the 12 question templates (range -0.116 to -0.160 against -0.135). This concerns the original, prompt-confounded contrast and predates the audit of Appendix E.

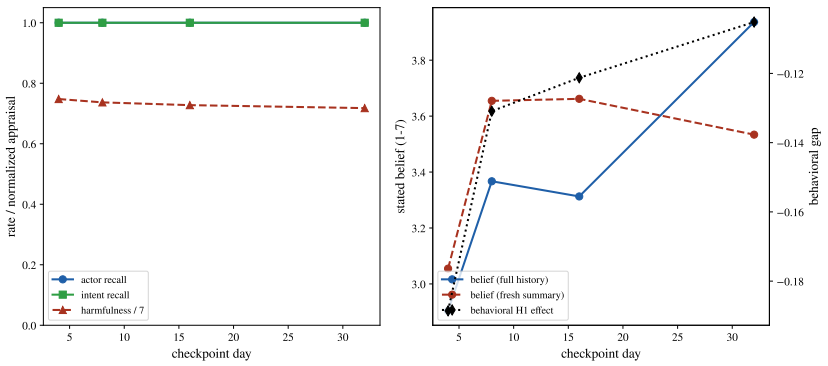


Figure 5: Left: explicit actor and intent recall and perceived harmfulness by checkpoint (full history, betrayal). Right: stated belief (*willingness to depend*) by checkpoint for the full-history and original fresh-summary conditions, against the pooled behavioral H1 effect; the recursive-summary belief series is not plotted. Original (unmatched) corpus.

E Summary-Content Audit

Exploratory, post hoc analysis motivated by inspecting generated summaries. Frozen plan: `reports/summary_asymmetry_v1/analysis_plan.md` (supplementary). Unit: family ($n = 30$); *checkpoints averaged within family*. Reads only the frozen original-run outputs; no new inference. Four deterministic estimands compared fresh- vs. recursive-summary text (betrayal condition): token length, evaluative-marker count (lexicon: *significant, notable, marked, deviation, unremarkable*), marker density, and trailing prominence (position of the last "day 4" reference). Results: token length, fresh 113.3 vs. recursive 190.9 ($d_z = -2.66$, 30/30 families); marker count, fresh 2.47 vs. recursive 0.0 exactly ($d_z = 4.50$, 30/30 families); marker density, same total separation; trailing prominence, null (Holm $p = 0.44$). Permutation: 10^5 draws, four-way Holm, so the two confirmed contrasts have Holm $p < 4 \times 10^{-5}$.

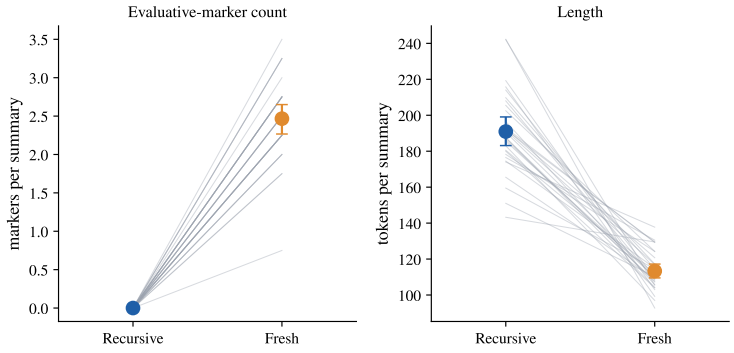


Figure 6: Summary-content audit, original prompts, per family (thin lines) with family-cluster bootstrap 95% CIs: fresh summaries are shorter yet contain evaluative-significance markers in every family; recursive summaries contain none.

F Matched-Prompt Replication

Confirmatory; plan committed before inference (`reports/matched_prompt_v1/analysis_plan.md`). The matched prompts are the union of the two originals' instructions (Appendix L): the recursive prompt gains "preserving all specific facts you can" and the fresh prompt gains "Keep it as short as possible". Scope: 30 families $\times$ {betrayal, neutral} $\times$ {recursive, fresh} $\times$ 4 checkpoints (480 new cells); full history is reused from the original run because it

involves no summarization call. A pre-freeze smoke test found that the matched recursive prompt pushed generation past the existing 512-token generation ceiling, causing mid-sentence truncation that the 1,024-token budget check cannot detect; the ceiling was raised to 1,536 for both prompts and re-verified before freezing. All 480 outputs completed and reconciled.

Estimands. Fresh−full: mean -0.004, 95% CI $[-0.039, 0.037]$, $d_z = -0.04$, $p = 0.84$; 5 amplify / 3 attenuate / 22 tie (band ± 0.10); 19/30 negative. Recursive−full: mean -0.055, 95% CI $[-0.075, -0.037]$, $d_z = -1.02$, $p = 2 \times 10^{-6}$ (2 of 10^6 sign-flip draws as extreme); 5 amplify / 0 attenuate / 25 tie; 27/30 negative; leave-one-family-out sign-consistent in all 30 exclusions. The plan declared the two estimands co-equal after the original preregistration had named fresh−full primary; that revision was frozen before any matched-prompt inference but is post hoc relative to the original plan. Under the original preregistration the primary replication outcome is therefore null and the surviving effect is the originally secondary estimand.

Determinism across runs. Every identity condition in the intervention studies re-scores a text that was scored in an earlier Slurm job; agreement is $\leq 2.6 \times 10^{-16}$ in all cases (Sections 4.4 and 4.5), so reusing the full-history cells from the original run introduces no cross-run drift.

Per-candidate view. Relative to full history, the matched recursive summary raises willingness toward both candidates in the neutral state (Bob $+0.123$, Charlie $+0.120$). In the betrayal state Bob rises less ($+0.027$, CI $[-0.021, 0.077]$) than Charlie ($+0.079$, $[0.033, 0.127]$): the amplification is Bob's rating failing to recover.

By checkpoint (recursive−full, 30 families): day 4 -0.096 $[-0.130, -0.062]$; day 8 -0.026 $[-0.053, -0.001]$; day 16 -0.034 $[-0.065, -0.005]$; day 32 -0.064 $[-0.093, -0.036]$. The effect is already largest at day 4, after four updates. Post hoc, descriptive: fresh−full by checkpoint is $+0.128$ $[0.066, 0.195]$ (day 4), -0.027, -0.033, and -0.086 $[-0.117, -0.057]$ (day 32); the direct matched recursive−fresh contrast is -0.224 $[-0.296, -0.151]$ at day 4 and 0.000 $[-0.060, 0.056]$, -0.001 $[-0.060, 0.058]$, $+0.021$ $[-0.013, 0.055]$ at days 8, 16, 32 (-0.051 $[-0.089, -0.017]$ averaged), so the checkpoint-averaged difference between summary types is concentrated at day 4, where the fresh summary (70 tokens) is shorter and the recursive summary (147 tokens) longer than the 133-token history.

Table 3: Summary text statistics, betrayal condition, 30 families per cell. Tokens: mean at days 4/8/16/32. Markers: mean count of evaluative-significance words at day 32 (families with ≥ 1). Verbatim: families whose summary contains the frozen day-4 sentence verbatim at day 32. Fabricated: families with at least one "A new event occurred on day k" sentence absent from the archive, at day 32 (neutral condition in parentheses). Full history: 133/246/480/948 tokens; verbatim 30/30 at every checkpoint.

Prompts	Summary	Tokens (d4/8/16/32)	Markers d32	Verbatim d32	Fabricated d32
original	recursive	126/235/215/188	0.00 (0)	13/30	2 (3)
original	fresh	108/103/114/128	2.97 (29)	0/30	0 (0)
matched	recursive	147/258/374/258	0.33 (9)	26/30	13 (12)
matched	fresh	70/83/85/86	1.87 (27)	0/30	0 (0)

What the matched prompts did and did not equalize (Table 3). They narrowed the marker difference but reversed the length ordering (matched recursive summaries are three times longer than matched fresh ones), and the recursive prompt's fact-preservation instruction made it copy the day-4 template sentence verbatim (26/30 at day 32, vs. 13/30 under the original recursive prompt and 0/30 for either fresh prompt, which always paraphrase). Section 4.5 tests this last difference directly. Within a condition, family-level amplification correlates with fresh-summary length under the original prompts ($r = -0.51$; longer fresh summaries amplified more) but not with marker count ($r = -0.16$), and with neither under matched prompts ($r = -0.06$, -0.04); the marker audit therefore motivated but does not explain the correction.

Side effects of the recursive prompt. (a) Event loss: the day-4 sentence is present verbatim in 30/30 matched recursive summaries at days 4 and 8, 28/30 at day 16 and 26/30 at day 32 (the four families that lost it are excluded a priori from the recursive-base interventions). (b) Fabrication: 13–14/30 matched recursive summaries contain invented filler events from day 4 onward (12–14/30 in the neutral condition; 2–5 under the original prompts; 0 for fresh), so fabrication appears at the first up-

date rather than accumulating. (c) Truncation in the original corpus: two neutral-condition recursive summaries at day 32 reached the 512-token generation ceiling and end mid-sentence; no betrayal-condition summary did. The original H2 numbers displayed in Section 4.2 include those two cells. (d) Other references: outside the day-4 sentence, 2/30 matched recursive day-32 summaries contain another sentence referring to the event; the interventions leave such sentences unchanged.

Sensitivity to fabrication (post hoc, descriptive). Families whose matched recursive summary fabricates ($n = 13$) show recursive−full -0.069 $[-0.083, -0.053]$, 13/13 negative; the other 17 show -0.044 $[-0.078, -0.017]$, $p = 0.005$; between-group difference -0.025 $[-0.056, 0.012]$. Excluding both fabricating and event-dropping families ($n = 16$): -0.047 $[-0.080, -0.018]$. The effect survives either exclusion and is numerically larger, not smaller, in fabricating families.

F.1 Candidate accounts of the recursive/fresh difference

No experiment here manipulates the number of recursive updates at fixed content. Candidates consistent with the data: (i) verbatim retention of the intent-marked template sentence under the recursive prompt, tested in Section 4.5; (ii) length and position (matched recursive summaries are longest and place the event early; Liu et al., 2023); (iii) iterative self-conditioning, in which each update recommits to the prior rendering of the event (the transmission-chain literature of Section 2 predicts preservation of social gist); (iv) content drift, which the sensitivity analysis above makes an unlikely sole driver. The matched fresh summaries still contain evaluative markers in 27/30 families and do not amplify, so surface evaluative vocabulary is not the operative variable.

G Interventions on the Recursive Summaries

Confirmatory; plan, code, config and all 156 texts hashed before inference (`reports/causal_recursive_v1/`). Base: the matched recursive summary (betrayal, day 32) scored in the replication. The day-4 sentence occurs verbatim exactly once in 26/30 summaries and is substituted in place with prefix and suffix asserted byte-identical; the four families without it are excluded a priori. Audit 156/156 pass; run 156/156 cells; identity-condition fidelity 2.6×10^{-16}; matched neutral recursive day-32 gap 0.0008.

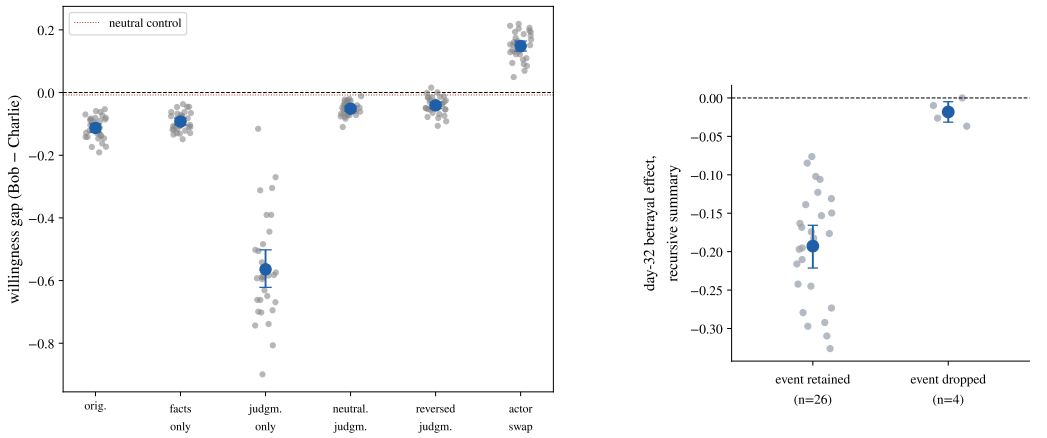
Table 4: Paired comparisons vs. *original* on recursive-summary text ($n = 26$). Δ = condition − original gap; positive Δ = weaker betrayal effect. Family counts: stronger / within ± 0.10 / weaker-or-redirected. All raw $p < 10^{-6}$, Holm $p < 5 \times 10^{-6}$; LOFO 26/26. Bob and Charlie columns are the paired per-candidate changes.

Condition	Δ	95% CI	d_z	families	ΔBob	ΔCharlie
facts only	$+0.105$	$[0.084, 0.128]$	1.82	0/14/12	$+0.46$	$+0.35$
judgment only	-0.737	$[-0.810, -0.664]$	-3.74	26/0/0	-0.34	$+0.39$
neutralized	$+0.153$	$[0.129, 0.178]$	2.36	0/5/21	$+0.50$	$+0.35$
reversed	$+0.164$	$[0.137, 0.190]$	2.33	0/5/21	$+0.68$	$+0.51$
actor swap	$+0.527$	$[0.470, 0.589]$	3.37	0/0/26	$+0.17$	-0.36

The per-candidate columns qualify two headline numbers. Under *judgment only*, Bob falls by 0.34 and Charlie *rises* by 0.39: about half of the gap widening is Charlie being released from the recipient role, since the judgment sentence no longer names him. Under *actor swap*, Charlie-as-offender (3.80) is rated lower than Bob-as-offender was (3.96), while Bob-as-recipient (4.13) remains below Charlie-as-recipient (4.15); the larger swapped gap is this asymmetry, not evidence of anything beyond role binding. On full-history text ($n = 30$) the same pattern holds with smaller magnitudes: *judgment only* Bob -0.17, Charlie $+0.29$; *actor swap* Bob $+0.07$, Charlie -0.19. Every intervention moves the gap further on recursive text than on full history (paired differences 0.08 to 0.29), and is descriptive: the edited sentence is a larger fraction of a 258-token summary than of a 948-token history, so it is not itself evidence about the procedure.

Families whose summary dropped the event. At day 32 the four excluded families' betrayal-vs-neutral gaps on the recursive summary are 0.000, -0.010, -0.037, -0.026 (mean -0.018) against -0.193 for the 26 that retained it (Figure 7b); their recursive−full contrast at day 32 is $+0.071$

(attenuation) against -0.085 for the 26. Averaged over all checkpoints the two groups do not differ (-0.052 vs. -0.055) because the event was still present at days 4 and 8 (and, for two of the four, day 16). Descriptive, $n = 4$.



(a) Full-history text, 30 families × 6 conditions.

(b) Day-32 betrayal effect, event retained ($n = 26$) vs. dropped ($n = 4$).

Figure 7: Family-level observations with family-cluster bootstrap 95% CIs.

H Sentence-Swap and Judgment-Vocabulary Controls

Confirmatory; plan, code, config and all 254 texts hashed before inference (`reports/causal_swap_v1/`); motivated by the verbatim-retention statistics of Table 3. Bases (day 32, betrayal): matched fresh ($n = 30$), matched recursive ($n = 26$, same exclusions as Appendix G), full history ($n = 30$). The paraphrased event span in a fresh summary is located by a rule fixed in the plan (the sentence beginning "On day 4," plus following sentences that name Bob or Charlie; exactly one such span in 30/30 summaries; three summaries carry a later sentence such as "This act of deliberate disclosure …" that names nobody and is left unchanged). Audit: 254/254 pass, no leakage terms; two over-strict audit rules were narrowed after the first build and before hashing (documented in the plan). Run: 254/254 cells (842 s). Identity conditions reproduce the replication and the earlier intervention runs exactly (max $|\Delta| = 0$).

Table 5: Paired comparisons, family level; gap g on the 7-point scale; Holm across six. neg/tie/pos at ± 0.10. Per-candidate columns are the paired changes in willingness toward Bob and Charlie.

Comparison	n	Δg	95% CI	d_z	neg/tie/pos	ΔBob	ΔCharlie
fresh+verbatim − fresh	30	-0.010	$[-0.054, 0.035]$	-0.08	7/17/6	-0.07	-0.06
recursive+paraphrase − recursive	26	-0.177	$[-0.232, -0.125]$	-1.25	18/7/1	-0.18	-0.01
recursive: judgment, no trust word	26	-0.496	$[-0.551, -0.442]$	-3.46	26/0/0	-0.16	$+0.33$
recursive: judgment + Charlie	26	-0.590	$[-0.639, -0.541]$	-4.56	26/0/0	-0.07	$+0.52$
full history: judgment, no trust word	30	-0.362	$[-0.407, -0.315]$	-2.75	30/0/0	-0.08	$+0.29$
full history: judgment + Charlie	30	-0.418	$[-0.465, -0.366]$	-2.96	29/1/0	-0.00	$+0.42$

Raw permutation p: 0.68 for the first row, 2×10^{-6} for the second, $< 10^{-6}$ for the rest; leave-one-family-out sign-stable throughout. Condition-level gaps: fresh -0.223, fresh+verbatim -0.232, recursive -0.192, recursive+paraphrase -0.369; recursive judgment sentences -0.688 (no trust word) and -0.782 (with Charlie), against -0.929 for the original *judgment only*. Neutral day-32 reference gaps: fresh -0.032, recursive $+0.005$. Against the frozen criteria, the verbatim-retention account is not supported (estimand 1 null, estimand 2 opposite in sign to its prediction); the construct-term account for *judgment only* is partial (the effect is reduced by about a third without the word *trust* but remains large); and the comparator account dominates all judgment variants.

I Cross-Model Instrument Attempts: Full Record

I.1 Attempt 1: primary instrument unmodified

Mistral-7B-Instruct-v0.3 and Phi-4-mini-instruct each failed gate 6 (forward/reverse agreement) after one permitted calibration retry each (Mistral diff 0.972, Phi-4-mini 0.539, tolerance 0.35) and passed the other 11 gates. The digit-tokenization validator was corrected for Mistral after it was found to test candidate tokens in isolation rather than at the answer position (an in-situ check, applied to all three models). Yi-1.5-9B-Chat was calibrated once against the same instrument and failed the same gate (diff 0.374), passing the other 11. Qwen3-8B's own gate-6 difference is 0.053.

Table 6: Gate 6 across the three non-Qwen models, Attempt 1.

Model	Diff	Tolerance	Result
Mistral-7B-Instruct-v0.3	0.972	0.35	FAIL
Phi-4-mini-instruct	0.539	0.35	FAIL
Yi-1.5-9B-Chat	0.374	0.35	FAIL

I.2 Attempt 2: reservation-threshold instrument, version 1

A continuous incentive-choice instrument (cooperate-with-actor vs. a neutral outside option; reward varied over an 11-point grid $\{-20, \ldots, 20\}\%$; primary statistic a reservation threshold, with an AUC fallback; full-sequence log-likelihood scoring of complete continuation strings). The continuations "I will work with {name}." (6 tokens on Qwen) and "I will work independently." (5 tokens) were not length-matched, biasing raw sequence likelihoods. All three models, including Qwen3-8B, failed overall calibration (Table 7). An AUC-to-threshold sign inversion in the analysis code was repaired; it fixed the synthetic-ordering direction but not the length confound. The stop rule that later ended the study ("Qwen failing construct validity terminates the cross-model study") was written into the version-2 plan; version 1's Qwen failure was treated as an implementation and design defect and led to the redesign rather than to termination. Version 2 is a redesign of the same instrument; both are reported.

Table 7: Reservation-threshold instrument, version 1 (post-repair). Fractions are the share of test families or templates passing.

Gate	Qwen3-8B	Mistral-7B	Phi-4-mini
Monotonic direction	FAIL (0/5)	FAIL (2/5)	FAIL (0/5)
Name-swap sensitivity	FAIL	FAIL	FAIL
Option-order sensitivity	FAIL	FAIL	FAIL
Template sensitivity	FAIL	PASS	PASS
Raw-vs-normalized agreement	FAIL (20%)	FAIL (50%)	PASS (90%)
Synthetic ordering	PASS	PASS	PASS
Threshold identifiability	FAIL (20%)	PASS (95%)	FAIL (0%)
Overall	**FAIL**	**FAIL**	**FAIL**

I.3 Attempt 3: reservation-threshold instrument, version 2

A tokenizer-level search found a token-matched, name-free pair ("I choose cooperation." / "I choose independence.", 4 tokens on every tokenizer), with actor identity moved into the prompt. This removed the length confound structurally. Qwen3-8B then failed construct validity: synthetic ordering was correct, but no family's cooperation-probability curve crossed 50% within the $\pm 20\%$ grid (Table 8), a dynamic-range limitation. Per the version-2 stop rule, the cross-model study ended; no grid expansion or wording revision was attempted.

Table 8: Reservation-threshold instrument, version 2.

Gate	Qwen3-8B	Mistral-7B	Phi-4-mini
Monotonic direction	FAIL (0/5)	FAIL (0/5)	FAIL (4/5)
No universal choice	PASS	PASS	FAIL (0/5)
Name-swap sensitivity	FAIL	FAIL	PASS
Option-order sensitivity	PASS	FAIL	PASS
Template sensitivity	FAIL	PASS	PASS
Raw-vs-normalized (structural)	PASS	PASS	PASS
Synthetic ordering	PASS	PASS	PASS
Threshold identifiability	FAIL (0%)	PASS (90%)	FAIL (0%)
Overall	**FAIL**	**FAIL**	**FAIL**

J Representation Probing: Forensic Audit and Full Results

Audit items. (1) Activations were extracted in a single forward pass with no generation. (2) Literal lexical equivalents of the target are present in the scored context only for *event condition* and *intentionality*. (3) *behavioral willingness* is circular: its value is computed from logits that are a near-affine function of the activation being probed. (4) Scaling, PCA and hyperparameters are fold-internal. (5) The original analysis shuffled labels at the cell level; a family-block permutation was implemented and found to be a no-op for the two targets whose labels follow the same deterministic cell pattern in every family, so a within-family permutation was used for those two (centering at chance, 0.250/0.501). (6) No cross-family duplicate memory contexts. (7) Majority-class, mean and random-feature baselines added. (8) The *actor target binding* and *cued recall actor* labels are identical in all 1,440 cells and are reported as one target.

Table 9: Final-layer probing, families held out. p: family-block (fb) or within-family (wf) permutation, 500 draws, so the floor is $1/501$. The original protocol used 2,000 draws, whose Holm-adjusted floor across the 21 layer×target tests is 0.0105 for every surviving target.

Target	Score	Baseline	p / status
actor–target binding (= cued actor recall)	0.969 acc	0.724	< 0.002 (fb)
perceived harmfulness	0.959 R^2	-0.010	< 0.002 (fb)
stated belief (willingness to depend)	0.783 R^2	-0.024	< 0.002 (fb)
event condition (4-way)	0.981 acc	0.251	< 0.002 (wf)†
intentionality	0.990 acc	0.476	< 0.002 (wf)†
behavioral willingness	0.936 R^2	-0.020	excluded (circular)

†Condition-defining words are present in the scored context.

The full audit report is `reports/representation_v1/phase7_forensic_audit_report.md` in the supplementary material.

K Reproducibility Details

Model revisions (truncated to 12 characters; full hashes in the committed run configs): `Qwen/Qwen3-8B b968826d9c46`; `mistralai/Mistral-7B-Instruct-v0.3 c170c708c41d`; `microsoft/Phi-4-mini-instruct cfbefacb9925`; `01-ai/Yi-1.5-9B-Chat 1a0fc698cf88`. Packages: `torch==2.6.0+cu124`, `transformers==5.17.0`, `tokenizers==0.23.2`, `huggingface_hub==1.32.0`, Python 3.12.0, CUDA 12.4, NVIDIA H100 80GB. Supplementary run summaries record available model/revision, seed, configuration hashes, GPU type, timing, and output counts. Each stage's frozen plan, calibration tolerances, transformation code, analysis code and, for interventions, input texts were version-controlled before inference; available SHA-256 manifests and the corresponding frozen plans, configs, analyses, and run summaries are included in the anonymized supplementary material. Files referenced by path in this appendix are in the supplementary material.

L EXACT PROMPTS AND TEXTS

Willingness prompt (forward anchor, one of 12 templates): "You are Alice. Here is everything you remember: {memory_context} You must decide whether to assign {name} responsibility for an important upcoming task. Rate your willingness on this scale: 1 = completely unwilling, ..., 7 = completely willing. Respond with exactly one digit from 1 to 7. Do not explain your answer."

Scenario templates. Betrayal: "On day {day}, Bob deliberately told Charlie about {topic}, even though Alice had asked Bob to keep it private. Bob chose to share it on purpose, and Charlie now knows about {topic}." Accidental disclosure: "On day {day}, Bob accidentally let Charlie see a message about {topic}, even though Alice had asked Bob to keep it private. Bob did not mean for this to happen, but Charlie now knows about {topic}." Positive protection: "On day {day}, Charlie asked Bob about {topic}, and Bob deliberately declined to say anything, protecting Alice's privacy. Charlie still does not know about {topic}." Neutral control: "On day {day}, Bob and Alice had an ordinary conversation about their weekend plans. Nothing about {topic} came up." Filler (every other day): "On day {day}, Alice, Bob, and Charlie had a routine, uneventful interaction unrelated to {topic}."

Summarizer prompts. Original recursive: "Update the following summary to include the new event. Keep it as short as possible and under 1024 tokens. PREVIOUS SUMMARY: {previous} NEW EVENT: {event} UPDATED SUMMARY:". Original fresh: "Summarize the following event history in under 1024 tokens, preserving all specific facts you can: {events} SUMMARY:". Matched recursive: "Update the following summary to include the new event. Keep it as short as possible and under 1024 tokens, preserving all specific facts you can. PREVIOUS SUMMARY: {previous} NEW EVENT: {event} UPDATED SUMMARY:". Matched fresh: "Summarize the following event history in under 1024 tokens, preserving all specific facts you can. Keep it as short as possible. {events} SUMMARY:". An over-budget summary is regenerated at most twice with a shortening prompt and then hard-truncated; no reported summary was hard-truncated.

Intervention sentences ({topic} substituted per family). *facts only*: "On day 4, Bob deliberately told Charlie about {topic}. Bob chose to share it on purpose, and Charlie now knows about {topic}." *judgment only*: "On day 4, something happened that damaged Alice's trust in Bob concerning {topic}." *neutralized*: "On day 4, Bob deliberately told Charlie about {topic}, and Alice had mentioned {topic} to Bob on a previous occasion. Bob chose to share it on purpose, and Charlie now knows about {topic}." *reversed*: "On day 4, Bob deliberately told Charlie about {topic}, and Alice had told Bob it was fine to share it. Bob chose to share it on purpose, and Charlie now knows about {topic}." *actor swap*: "On day 4, Charlie deliberately told Bob about {topic}, even though Alice had asked Charlie to keep it private. Charlie chose to share it on purpose, and Bob now knows about {topic}." *judgment, no trust word*: "On day 4, Bob did something concerning {topic} that seriously wronged Alice." *judgment plus Charlie*: "On day 4, something happened that damaged Alice's trust in Bob concerning {topic}, and Charlie now knows about {topic}." A typical fresh-summary paraphrase of the event: "On day 4, Bob intentionally disclosed {topic} to Charlie, despite Alice's request for confidentiality."