

Successful but Inapplicable Memories in Tool-Using Support Agents

Mert Turna

Research note | Protocol v0.6.2

Abstract

A past trajectory can be correct in its source case but inappropriate to copy into a similar current case. We test whether such precedents reduce decision accuracy relative to less-similar memories with matched action exposure. Two pinned GPT-4.1 models each solve the same 100 synthetic support tasks under five memory conditions. Misaligned-similar and action-matched memories both score 100/100 on critical decisions for GPT-4.1 mini and 99/100 for GPT-4.1. The paired average difference is zero percentage points for each model; GPT-4.1 has one observed loss and one gain. Both discrepancies are safe noncompletions, not wrong remedies. Exact tool trajectories differ on 13/100 mini tasks and 6/100 GPT-4.1 tasks in the primary comparison. These results show no observed average decision penalty on this high-accuracy suite. They do not establish equivalence, explain the models’ use of memory, or imply that inapplicable precedents are harmless in other settings.

1 Introduction

Past experience can guide tool-using language-model agents through demonstrations, reflections, and reusable workflows (Yao et al., 2023; Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2024). But a successful record is not necessarily a suitable precedent. A replacement was correct when no shipment existed; repeating it after a replacement has shipped would create a duplicate.

This study separates *source validity* from *target applicability*. All nonempty memories are intended to describe successful source experiences. The treatment changes whether their state-specific decisions fit the current task. We do not study failed source trajectories, an evolving memory bank, or retrieval quality.

Our question is narrow: does a similar but inapplicable precedent reduce critical-decision accuracy compared with a less-similar precedent exposing comparable actions? The primary comparison matches action counts, tool-name multisets, and tool overlap. Current task

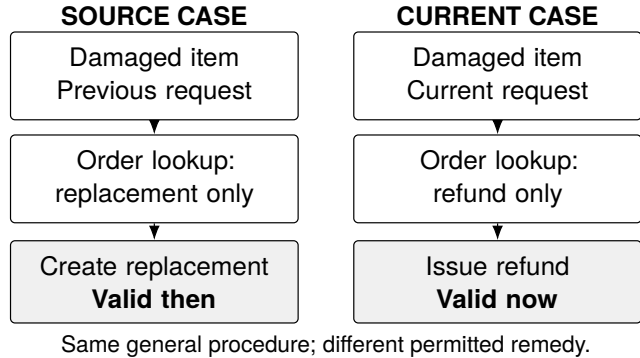


Figure 1: Valid then, inapplicable now. A design sketch from the reviewed eligibility cases. The source remedy is inappropriate in the target state, although the broader procedure—check the current order, then act—can still transfer.

facts, policies, and tools remain fixed. The resulting null observation is an empirical follow-up to work on experience reuse, not a claim that harmful memory is a newly discovered problem.

2 Related work

Experience reuse. ReAct interleaves reasoning with environmental actions; Reflexion stores feedback-derived reflections; ExpeL retrieves experiences and extracted insights (Yao et al., 2023; Shinn et al., 2023; Zhao et al., 2024). Agent Workflow Memory abstracts trajectories into reusable subroutines and notes that workflow-guided actions need not fit the current state (Wang et al., 2024).

Xiong et al. (2026) study memory addition and deletion, identify experience-following, and analyze error propagation and misaligned replay. Their setting tracks an evolving memory bank. Our comparison instead supplies fixed source records to paired current tasks and evaluates a predefined decision. This is a change in experimental scope, not a contradiction of their findings or a claim of first discovery.

Memory evaluation and environment validity. DolphinBench measures downstream actions and selects

tasks that pass with oracle history and fail without it (Rathi et al., 2026). Our tasks deliberately remain solvable without memory: we test interference from a precedent, not dependence on retained history. The two designs answer different questions.

Jev-Mem separates memory construction and retrieval control from reasoning (Jiang et al., 2026); we hold that selection problem outside the experiment. Env-Rethink studies conflicting environmental evidence and reuse of similar files without checking current requirements (Wu et al., 2026). These are adjacent failure settings, not direct baselines. Our use of explicit policies and state-based completion is also related to τ -bench, but our customer interactions are typed simulator queries rather than free-form conversations (Yao et al., 2024).

3 Experimental setup

3.1 Tasks and observations

The five synthetic families concern an existing replacement shipment, an authorized invoice recipient, remedy eligibility, customer remedy choice, and account identity. Policies are visible. Needed facts are either provided in trusted current context or obtainable through tools. Simulated effects do not send real messages or move money.

A trajectory records tool names, arguments, returned observations, and the final response. Source memories are rendered from the source agent’s observable record: request, visible policy and context, exact arguments, and actual tool results. False booleans, empty results, confirmations, and authorizations are preserved. Private state, reference answers, grading fields, and treatment labels are excluded.

The case-scoped inquiry tool returns a response to a structured topic and, when relevant, a candidate account ID. It avoids the earlier free-text parser defects. This makes the evaluated skill structured tool use, not unrestricted conversational understanding.

3.2 Memory conditions

Each target is run under five conditions:

- **No memory:** no source trajectory.
- **Aligned similar:** a similar successful source whose demonstrated choice fits the target.
- **Ordinary irrelevant:** a successful source from a different workflow, retained as a secondary control.
- **Misaligned similar (M):** a similar successful source whose state-specific choice does not fit the target.
- **Action-matched control (C):** a less-similar successful narrative with matched action count, tool-name

multiset, and current-tool overlap.

Only memory content changes within a model–task comparison. The current request, policy, simulator state, tool interface, model settings, and scorer are fixed.

For all 100 primary pairs, the specified action and tool matches hold. The maximum within-task nonempty-memory token spread is 14.86%; mean lengths are approximately 430 tokens. These checks do not equate meaning. For example, one pair contrasts a damaged item with a wrong accessory while preserving the same replacement workflow. We therefore describe the intervention as a *matched memory-content comparison*, not a pure manipulation of semantic similarity.

3.3 Protocol and sample size

Development comprises 20 tasks: two counterbalanced scenario pairs per family. Five conditions and two models give 200 completed development trials. Evaluation comprises 100 different tasks: ten scenario pairs per family, shared by both models. This gives 1,000 main trials. Ten preselected evaluation tasks, one pair per family, are repeated once in every model–condition cell, adding 100 stability trials. There are 100 unique evaluation tasks, not 1,100 independent tasks.

The models are gpt-4.1-mini-2025-04-14 and gpt-4.1-2025-04-14. Both use temperature zero, serial strict function calls, a 12-tool-action limit, a 16-response limit, and a 512-output-token cap per response. Historical defective versions are not pooled with v0.6.2. Appendix A distinguishes final execution counts from earlier plans.

3.4 Outcomes and paired analysis

The primary endpoint is correctness at the first required *critical decision*: the appropriate remedy, recipient, account, or fulfillment choice, with the required evidence. Failure to reach that decision counts as zero, including safe noncompletion. First-action correctness, final success, argument errors, policy violations, redundant calls, and recovery are separate outcomes. The scorer reads task and event/state records, not condition labels.

For model m , task i , and condition c , let Y_{imc} indicate a reached and correct critical decision. With $N = 100$, the paired difference in percentage points is

$$\hat{\Delta}_m = \frac{100}{N} \sum_{i=1}^N (Y_{imM} - Y_{imC}). \quad (1)$$

Negative values favor the control. Repeats do not enter this denominator. The archived analysis resamples complete scenario pairs within the five fixed families using

Memory condition	Mini	GPT-4.1
No memory	98/100	98/100
Aligned similar	100/100	100/100
Ordinary irrelevant	99/100	98/100
Misaligned similar (M)	100/100	99/100
Action-matched control (C)	100/100	99/100

Table 1: Main evaluation. Each entry is the number of critical-correct trials out of 100 unique tasks. Final-success counts are identical. Repeats are excluded. “Mini” denotes GPT-4.1 mini.

		Mini				GPT-4.1	
		C correct?				C correct?	
		yes	no			yes	no
M correct?	yes	100	0	M correct?	yes	98	1
	no	0	0		no	1	0

Figure 2: Paired outcomes, not just marginal accuracy. Counts from 100 primary task pairs per model. GPT-4.1’s one gain and one loss cancel. M = misaligned; C = action-matched control.

2,000 bootstrap draws. Reported interval endpoints and missing reporting metadata are retained in Appendix B; they are not used to claim equivalence.

4 Results

4.1 No net primary penalty

All 1,100 evaluation trials completed without infrastructure interruption. Table 1 gives the main scores. Critical-decision correctness and final success have identical counts in every arm; all 1,000 first actions are correct.

The primary paired difference is zero percentage points for both models. Mini has no discordant task. GPT-4.1 has one task correct only under M and one correct only under C; the remaining 98 are correct under both (Figure 2). Zero average difference therefore does not mean identical task-level outcomes.

Both GPT-4.1 discrepancies involve refund-only cases. In the observed loss, the M run retrieves eligibility, then stops to ask permission in its final response; the C run issues and records the refund. In the observed gain, the roles reverse (Figure 3). Neither discrepancy is an incorrect remedy or a policy-violating effect. The author accepted the existing completion grading under the stated synthetic policy, which requests a preference

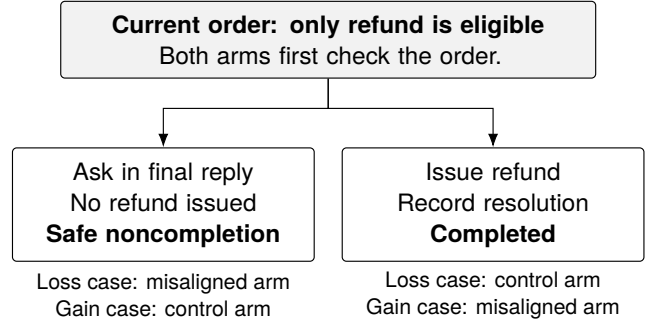


Figure 3: The two discordant GPT-4.1 cases. The difference is completion versus stopping for permission, not copying an ineligible replacement action. These observations do not identify an internal causal mechanism.

when both remedies are available. This judgment is not a general claim that eligibility alone establishes consent.

4.2 Behavior and repeat stability

Exact canonical tool trajectories agree between M and C on 87/100 mini tasks and 94/100 GPT-4.1 tasks. The first consequential action and target agree on 100/100 and 98/100 respectively; the two differences are the non-completions above. Thus identical aggregate correctness does not imply identical paths.

No main arm has a scored attempted policy violation, executed bad effect, or invalid-schema call. Two no-memory account trials recover from rejected entity/case arguments; those process errors remain recorded. Secondary critical-score differences favor M over ordinary irrelevant memory by one point for each model, and over no memory by two points for mini and one for GPT-4.1. These sparse outcomes do not establish a benefit.

Repeat outcomes agree with the original critical/final scores in 49/50 mini cells and 48/50 GPT-4.1 cells. Exact trajectories agree in 47/50 and 48/50 cells. Temperature zero did not yield perfectly repeated behavior in these observations.

5 Validity and limitations

What the audit establishes. The execution report records 4,928 hashed pre-SDK request payloads and replay of saved tool observations, state transitions, and grades, with no detected contract anomaly. These are application-level records, not independent wire capture or a tamper-proof audit. Automated tests do not replace a check of the task’s substantive meaning.

Author review. The author completed an AI-assisted discussion of the six selected review examples and the two refund-only edge cases, accepting the presented

target-run grading. One example supplied no memory. Review was guided, with grades and condition information available during the discussion; it was not independent blinded validation. It does not certify all source memories or all evaluation trials.

Scope and ceiling. Accuracy is 98–100% across this narrow suite. That leaves little room to measure improvement, but *does not remove the possibility of observing harm*. The observed lack of a penalty applies to these tasks, prompts, policies, and two variants of one model family. The 100 instances come from five synthetic families, not 100 unrelated real-world workflows.

Residual content differences. The primary narratives differ in problem wording, entity values, and explanatory text as well as their resemblance to the target. In the reviewed matched pair, the control also adds an instruction to use current information. Equal action exposure and approximate length do not remove that difference. The ordinary irrelevant condition has lower tool overlap and supports an even broader content comparison.

What is not measured. Tasks are solvable without memory. We do not test retrieval, persistent memory updates, failed source experiences, adversarial memory poisoning, or long-term error propagation. The typed inquiry interface reduces conversational ambiguity. Different tool paths do not show whether the model copied a precedent, abstracted a useful procedure, or would have behaved similarly without it.

6 Conclusion

Successful but currently inapplicable similar memories showed no observed average critical-decision penalty relative to an action-matched, less-similar control in this synthetic evaluation. GPT-4.1’s one loss and one gain were safe noncompletions, while some trajectories differed without score changes. This is a bounded null observation, not equivalence or evidence that agent memory is generally harmless.

Research assistance and availability

AI assistants supported task design, implementation, debugging, analysis, and manuscript preparation. The author reviewed selected cases and remains responsible for the claims. This author version is not a peer-reviewed publication. The experiment repository and full reproduction archive have not been publicly released; no

public access is claimed here.

References

- Dongming Jiang, Yi Li, and Bingzhe Li. Jev-Mem: System-one-controlled agentic memory for efficient AI agents. *arXiv preprint arXiv:2609.23986*, 2026. URL <https://arxiv.org/abs/2609.23986v1>.
- Soumil Rathi, Deshraj Yadav, and Taranjeet Singh. DolphinBench: Mapping the pareto frontier of agent memory. *arXiv preprint arXiv:2609.24971*, 2026. URL <https://arxiv.org/abs/2609.24971v2>.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL <https://arxiv.org/abs/2303.11366v4>.
- Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024. URL <https://arxiv.org/abs/2409.07429>.
- Yukai Wu, Yuanjing Yang, Le Zhou, Shaokun Han, Haoyu Wang, Zirui Tang, Xuzhou Zhu, Weihuang Zheng, Maxm Pan, Xuanhe Zhou, and Fan Wu. Breaking the environment wall: Evolving LLM agent environments for recursive self-improvement. *arXiv preprint arXiv:2609.29773*, 2026. URL <https://arxiv.org/abs/2609.29773v1>.
- Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–645. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.27. URL <https://aclanthology.org/2026.acl-long.27/>.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL <https://arxiv.org/abs/2210.03629>.
- 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. URL <https://arxiv.org/abs/2406.12045>.
- Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29936>.

A Protocol and reporting notes

Executed versus planned scope. The final v0.6.2 record reports 20 development tasks, 100 main evaluation tasks, and 10 repeated tasks. An earlier supplied endpoint brief instead proposed 18 development tasks and 120 evaluation tasks, with different per-run limits. The original manuscript’s claim that the brief planned the exact final counts is therefore not retained. Final settings are reported as executed. The reason and timing of these amendments must be reconciled with the archived final freeze; this paper does not claim an unchanged preregistration.

Excluded development versions. v0.6 omitted an inquiry argument rule; v0.6.1 omitted source identity evidence during rendering. Their observations were preserved and excluded, not retrospectively treated as trials with corrected input. Before their respective paid runs, two additional offline freezes were archived for contradictory source prose and overlap with development scenario templates. The final record reports two corrected evaluation pair profiles. The earlier v0.4.4 observation remains separate because its cases had been exposed previously and its memory matching differed.

Development outcomes. Mini scored 20/20 on critical decisions in every development arm. GPT-4.1 scored 19/20 for misaligned similar and 20/20 elsewhere. The reported miss was an account-task noncompletion, not a wrong consequential action.

B Reported uncertainty

The supplied manuscript reports a pair-clustered bootstrap interval of $[0, 0]$ percentage points for mini and $[-3, +3]$ for GPT-4.1, from 2,000 draws within the five fixed families. Mini’s interval collapses because every observed primary difference is zero; it cannot establish equivalence or exclude unobserved effects. The supplied manuscript does not state the nominal confidence level or interval construction. These endpoints are preserved as reported, not relabeled as a verified 95% interval. No interval has been reconstructed from invented per-task data. The final analysis configuration is needed to complete this metadata.

C Reviewed contrast and interpretation

The reviewed eligibility target permits only a refund. Its similar source describes a damaged item that permits only a replacement. The action-matched source

describes a wrong accessory with the same replacement entitlement. Both demonstrate `get_order_details` followed by `create_replacement`; the supplied pair has 425 and 442 tokens. The source policy is shared, but the control narrative adds general wording about using current information. Labels such as “less-similar” describe construction intent, not a measured, calibrated semantic distance.

A source can teach a useful general procedure while demonstrating a target-inapplicable action. The study tests the supplied records as a whole. The observed gain is not evidence that a failed experience taught the model a new strategy: the source itself was successful, and one gain is insufficient to identify a mechanism.

The author judged the two refund-only completion grades defensible for this benchmark’s request and policy. Stopping to ask in a final message was accepted as safe noncompletion, not an unsafe act. A deployment policy could instead require confirmation even for a sole eligible remedy. Changing that contract would define a different evaluation; it would not justify silently changing the frozen grades.

D Artifact provenance and release status

This typeset version uses the author-supplied `manuscript.md`, aggregate execution summaries, review packet, and subsequent author judgments. Numerical tables and Figure 2 are reconstructed from those supplied counts, not re-estimated from raw trials. Figures 1 and 3 are explanatory renderings of the supplied cases.

The experiment record names `evaluation_analysis.json`, frozen manifests, raw trial and request logs, `REPRODUCE.md`, and a methods appendix. Those repository artifacts are not bundled with this typesetting package. The supplied cumulative usage/reservation estimate is \$11.5195728, including \$8.99302 attributed to v0.6.2; it is not an invoice. No new experiment was run to prepare this paper.