Memory Use in Language-Model Agents: Action Recommendations and Status Claims

Anonymous authors
Paper under double-blind review

Abstract

Retrieving a relevant document does not establish which version is approved, whether an action is authorized, or whether delivery succeeded. We study this gap in memory-based assistants by separating task satisfaction, measured against a fixed reference, from evidence appropriateness, assessed over the entire response using only model-visible information. Our offline study combines retrieval diagnostics for six memory-system configurations with controlled evaluations of three language-model configurations from two families. We vary history while holding the request and task material fixed in 54 reviewed stories, remove key or neutral records in a 13-story subset, and test explicit states in 162 short scenarios. Complete history increased task satisfaction by 51.9–57.4 percentage points; removing key state evidence was more damaging than removing neutral records. Yet a configuration that recovered every required record in 72 cases still failed the action contract in 11 responses. A separate analysis of 106 archived retail-agent interactions also found disagreement between native task outcomes and final-response grounding. These results distinguish evidence recovery from its justified use: memory evaluation must examine what an assistant recommends, asks, and claims to have happened, alongside whether it meets the task objective.

1 Introduction

A workplace assistant is asked to send the approved release notice to a customer group. It finds two notices, but the approval and delivery receipt may appear elsewhere in the conversation. The document supplies the content; separate records establish which version is approved and what has already happened. Confusing these roles can lead to an unsupported recommendation, an omitted task step, or an unwarranted completion claim. Figure 1 illustrates these evidence boundaries with a robot assistant.

Existing evaluations make different parts of agent memory observable. MemoryAgentBench (Hu et al., 2026) separates accurate retrieval, test-time learning, long-range understanding, and selective forgetting; LongMemEval (Wu et al., 2025) tests updates and abstention. AMemGym (Cheng et al., 2026) studies evolving user state and memory utilization, while Mem2ActBench (Shen et al., 2026) examines memory-grounded tool use. We ask whether a proposed next step both addresses the task objective and stays within the evidence available to the agent.

Task satisfaction and evidence appropriateness answer different questions. When approval is missing, asking which notice is approved can be justified but leave the requested delivery incomplete. Conversely, choosing a notice merely because it is newer can happen to match the approved version in the task reference without adequate support. Evaluating the two criteria separately distinguishes missing task coverage from unsupported reasoning, rather than treating both as a single failure.

Our contribution is a paired assessment of task coverage and whole-response grounding, supported by complementary tests of memory use. Retrieval diagnostics locate failures despite complete evidence recovery. Shared-material history comparisons and selective removals test whether state-bearing records affect the response. Explicit-state controls identify an inquiry-target ambiguity in the task reference. Finally, archived retail interactions examine whether the measurement distinction also applies to recorded tool workflows. Together, these studies show how evidence availability, task satisfaction, and justified communication can diverge.

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.

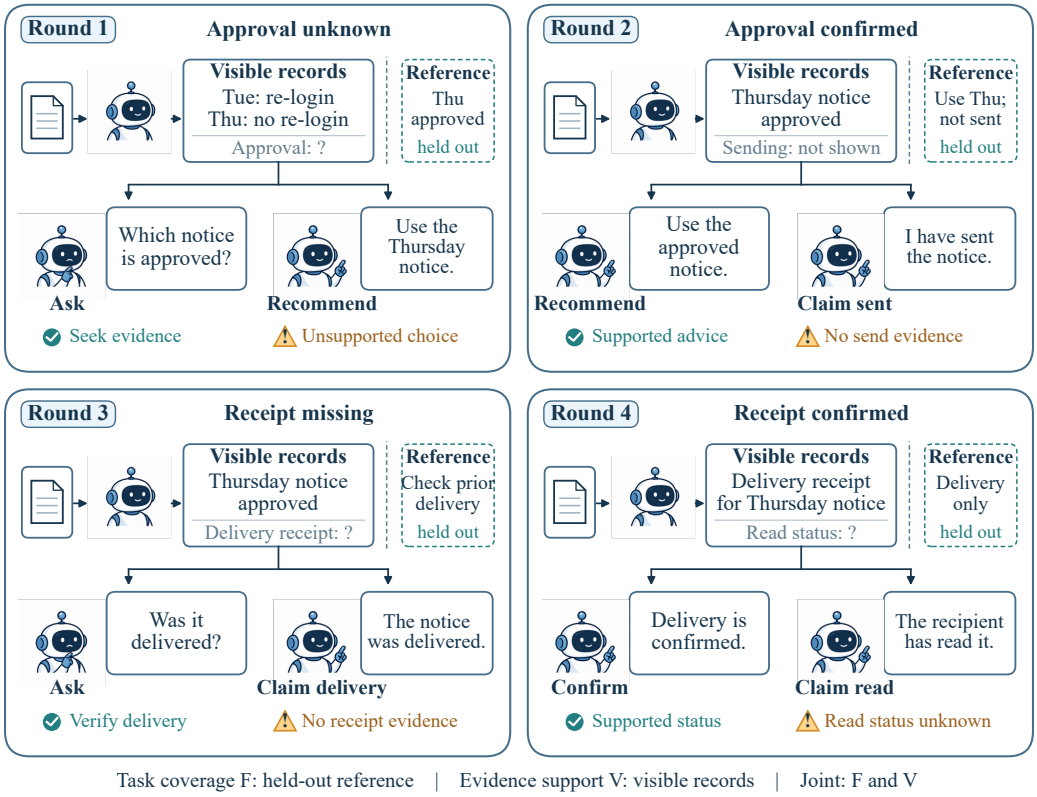

Figure 1: Four schematic checkpoints for a release-notice assistant, read left-to-right then top-to-bottom. The Tuesday notice requires re-login; Thursday does not. Round 1 is adapted from H3-S01 (Thursday is original Q); only its unsupported recommendation is observed. The remaining responses and rounds are illustrative; they do not represent recorded interactions or actual tool execution. Visible records support V; the held-out reference defines F; joint passage requires both. A supported inquiry can still leave the task incomplete.

2 Related Work

Memory evaluation. LongMemEval (Wu et al., 2025) tests recall, updates, and abstention; MemoryAgentBench (Hu et al., 2026) separates retrieval, test-time learning, long-range understanding, and forgetting. Our response-level assessment pairs task coverage with support for every proposed action, question, and status claim.

Dynamic memory. AMemGym (Cheng et al., 2026) diagnoses writing, reading, and utilization in an interactive environment. Our offline intervention holds task material fixed while varying historical evidence, measuring both task coverage and response grounding.

Memory-to-action evaluation. Mem2ActBench (Shen et al., 2026) tests memory-grounded tool selection and parameter grounding; AMA-Bench (Zhao et al., 2026) evaluates long-horizon memory in state–action trajectories, and SABER (Cuadron et al., 2025) studies verification before action. Our controlled models only propose actions and status claims; archived retail trajectories provide separate observational outcomes.

Retrieval methods. MemGAS (Xu et al., 2026) studies retrieval budgets and downstream generation. Its top-three passage test differs from our all-required-evidence diagnostic of whether complete retrieval suffices for an action contract.

3 Problem Formulation

At each controlled checkpoint, the model receives a request, task material, and the history available in that condition. It proposes actions, conditions, inquiries, and a reply; no tool is executed. A fixed task reference is held out from the model and used to assess task satisfaction. The visible input supplies the basis for evidence appropriateness (Figure 1). Archived interactions with recorded tool outcomes are analyzed separately.

We evaluate `SEND_SEGMENT`, `ASK_USER`, `STATE_ACK`, and `ADVANCE_WORKFLOW` semantically. A send recommendation must identify the recipient, content, version, and authorization. Conditional actions must retain unmet conditions. Status claims must not upgrade "prepared" to "sent," "sent" to "received," or "received" to "approved." Format deviations remain in provenance; formatting alone does not determine semantic scores.

Task satisfaction (F). A response satisfies the task when it covers every applicable requirement in the frozen reference: the required recommendation or workflow step, the requested status, and, where inquiry is required, the fact that must be asked about or checked. In an offline response, proposing the required inquiry is sufficient; obtaining its answer on the spot is not required. Merely naming an action is insufficient. This criterion evaluates task coverage, not whether every accompanying statement is justified or whether a tool actually executed.

Evidence appropriateness (V). Every expressed action, condition, question, and state claim must be warranted by the model-visible materials. This is a whole-response judgment. An appropriately cautious question can therefore pass even when it fails to cover the task reference.

Joint passage (J). A response passes jointly when $F = V = 1$. A definite failure on either criterion determines joint failure. Other unresolved combinations remain unknown and stay in the denominator.

These judgments assess action selection, state reporting, and inquiry targets without creating additional subscales. Recommending an action subject to execution confirmation differs from asking whether to choose it without making the required recommendation. Conditional branches also need evidence: no receipt does not prove failed delivery.

The designs specify different task objectives. Removing history retains an information-rich target, so a justified request for missing evidence can still leave the task incomplete. In explicit unknown-state scenarios, the target itself may be an inquiry. Task satisfaction therefore measures coverage of the specified objective rather than the best defensible choice under the visible prompt alone. When evidence is missing, failure on this criterion does not by itself imply an unsafe decision.

Seven earlier diagnostic labels, including status overclaim and unnecessary clarification, remain in their original release (Appendix E); none is inferred from recalibrated judgments. All 441 answer-bearing history and ablation responses retain fixed denominators.

3.1 Contrasting responses

In H1-S02, 35B preserves uncertainty about a podcast send, so the reply is evidence-appropriate but leaves a reference-required send-now step incomplete. In H3-S01, 27B recommends the version approved in full history without visible approval support; it explicitly denies sending, so the error is not fabricated execution. Figure 1 illustrates the latter distinction. These are explanatory selections, not a random sample; Appendix H.5 summarizes their visible facts and F/V outcomes and links the complete frozen records in the local candidate artifact.

4 Experimental Design

Upstream diagnostics use a legacy action contract; story and explicit-state studies evaluate whole responses under Section 3. External native outcomes remain separate from our reference-based task score.

Appendix Table 3 summarizes the designs, units, and denominators.

4.1 Upstream evidence diagnostics

We compare Mem0 (Chhikara et al., 2025), MemOS (Li et al., 2025), LangMem (LangChain, 2025), A-Mem (Xu et al., 2025), Nemori (Nan et al., 2025), and Graphiti (Zep, 2026) on 72 shared cases from 12 ledgers, aligning complete evidence recovery with raw, ungated contract passage. The contract checks actions and applicable arguments, not entire responses. A separate common-controller experiment compares Top-10 and Top-50. Appendix H distinguishes raw and historical gated scores; shared cases and configurations are not independent replications.

In a restricted Nemori cache, 10/72 cases lack required evidence. Reference-assisted repair inserts the missing public records; sham repair changes records without supplying them. The other 62 cases retain unchanged inputs. This diagnoses omissions rather than providing a deployable retriever.

4.2 Controlled story evaluation

Three human contributors wrote or substantively rewrote the reviewed materials; AI-assisted drafts remain in provenance. Across three frozen batches, 54 stories compare complete history against current request plus material, holding the request and material fixed.

The ablation reuses 13 stories with full augmented history, neutral records removed, or key state evidence removed. Material remains available. Its frozen primary contrast is neutral-removed minus key-removed, testing selected evidence bundles rather than individual state mechanisms.

Matched inputs and system prompts are evaluated with Qwen 27B, Qwen 35B, and Gemma 26B: three configurations from two families. Identifiers and generation settings are recorded per study. Requests start separate conversations. Failures remain in denominators, without automatic retries or result-dependent additions.

4.3 Review and statistics

Two reviewers assessed visible-evidence appropriateness and reference-based task satisfaction in separate phases for 441 responses: 324 history and 117 ablation. The authors report that each reviewer formed and confirmed individual judgments before AI organized the return sheets; the original AI-assistance declarations remain in the record (Appendix M). Agreement precedes coordination. A later author-supplied, post-hoc reading covers 450 specified requests (900 F/V cells) plus 49 targeted extension requests; its 36 differences from the frozen table are applied only to a separate score version. This return is not an additional independently blinded replicate. Reviewer-identity and signed-version linkage are not independently documented (Appendix M).

In the final 151-response partition, evidence review preceded task references. Among jointly definite judgments, reviewers agreed on evidence appropriateness in 143/147 responses (97.3%; kappa 0.893) and on task satisfaction in 134/143 (93.7%; kappa 0.861). Humans reviewed the consolidated decisions. That predecessor review left four cells across three responses unresolved; the current adjudicated history/ablation version has 13 unresolved metric cells (Appendix J). Appendix M documents exposure, 56 coordinated cells, prior AI proposals, and identical-response checks.

The story is the analysis unit. We report condition counts, paired differences, 10,000-draw paired bootstrap intervals, and positive/zero/negative story counts, preserving history batches during resampling. Unknown scores produce endpoint ranges and interval envelopes. The current plots and tables use the separate adjudicated score version; prior boundary-envelope analyses belong to the earlier release and are not mixed into these counts.

4.4 Explicit-state controls

The 162 rule-based scenarios comprise 54 each for authorization, approved versions, and receipts. Each varies two contrasting known states (A/B) and one unknown state (U), holding material, recipient, rule, and request fixed. Each configuration produces 486 responses: 1,458 total, with 162 scenario units.

The controlled studies contain 54 stories, including 13 reused for ablation, and 162 short scenarios: 216 material identifiers and 1,899 responses. Analyses use design-specific denominators; upstream and external cases are separate.

A scenario passes when all three responses meet both criteria. Unknown-receipt references require checking prior delivery; most visible rules say to ask under uncertainty without naming the inquiry target. We reviewed all 54 unknown-receipt inputs to examine that interpretation (Appendix N).

Before coordination, reviewers agreed on 2,890/2,916 cells. An AI-assisted coordinator resolved 26 cells across 24 responses, not a third independent human review; originals and reasons remain available. Qwen 35B disables thinking at temperature zero with a 900-token cap; 27B retains service-default thinking. Comparisons concern configurations, not isolated architectures. JSON compliance does not determine semantic scores.

Wilson intervals and paired scenario resampling describe scenario outcomes; shared templates limit population inference. Previously published disputed-cell bounds refer to the predecessor score table and are not presented as sensitivity bounds for the new release.

4.5 External measurement applicability

We selected 106 archived τ-bench (Yao et al., 2024) retail tasks whose reference actions involved database changes, using a fixed trial independently of observed outcomes. Local replay reproduced all 753 tool responses and 106 native results. Native scoring checks final database state for 72 tasks or designated text in any assistant response for 34; neither audits final-reply grounding. Two reviewers assessed final-response grounding, state-claim support, and authorization before state-changing calls; a targeted human adjudication addressed 12 tasks. Native success follows the historical evaluator, not our task-satisfaction rubric. This retrospective cohort uses a different archived agent, not the three controlled-study configurations. Appendix O details selection, scoring, reviewer exposure, and reward branches.

5 Results

5.1 Complete retrieval is not sufficient

Graphiti recovered all required evidence in 72/72 cases, but only 61/72 outputs passed the legacy contract (Figure 2a). Complete-evidence failures also numbered 19 for each of Mem0, LangMem, and A-Mem, 13 for Nemori, and two for MemOS. Contract passage with incomplete reference evidence occurred, particularly for MemOS, so neither outcome substitutes for the other.

Increasing returned records in the separate budget experiment reduced passage from 57 to 45 for Mem0 and LangMem, 50 to 43 for A-Mem, 54 to 46 for Nemori, and 60 to 46 for Graphiti; MemOS remained at 26/72 (Figure 2b). Composition, ranking, and length change together: this establishes non-monotonicity, not an isolated context-length effect.

On 10 repairable cases, original/repaired/sham passage was 5/10/4 for 27B, 4/9/4 for 35B, and 3/9/3 for Gemma (each out of 10; Appendix H). Targeted repair therefore helped where changing records without supplying omissions did not.

5.2 History beyond document availability

With request and material shared, complete history increased task satisfaction from 16/54 to 44–45/54 for Qwen 27B, from 11–12/54 to 42/54 for Qwen 35B, and from 13/54 to 41/54 for Gemma 26B (Figure 3). The paired gains are 51.9–53.7, 55.6–57.4, and 51.9 percentage points. The corresponding 95% paired story-bootstrap endpoint envelopes are 38.9–66.7, 40.7–70.4, and 38.9–64.8 points. These are conditional on the adjudicated labels and do not measure shared annotation error.

Definite task-satisfaction gains occur in 28, 31, and 28 stories, with no definite declines; one pair each remains unresolved for the Qwen configurations. Joint passage of task satisfaction and evidence appropriateness rises from 15 to 40–41, 11–12 to 39–40, and 9–10 to 37 of 54 stories, respectively. The task and whole-response evidence judgments need not move together.

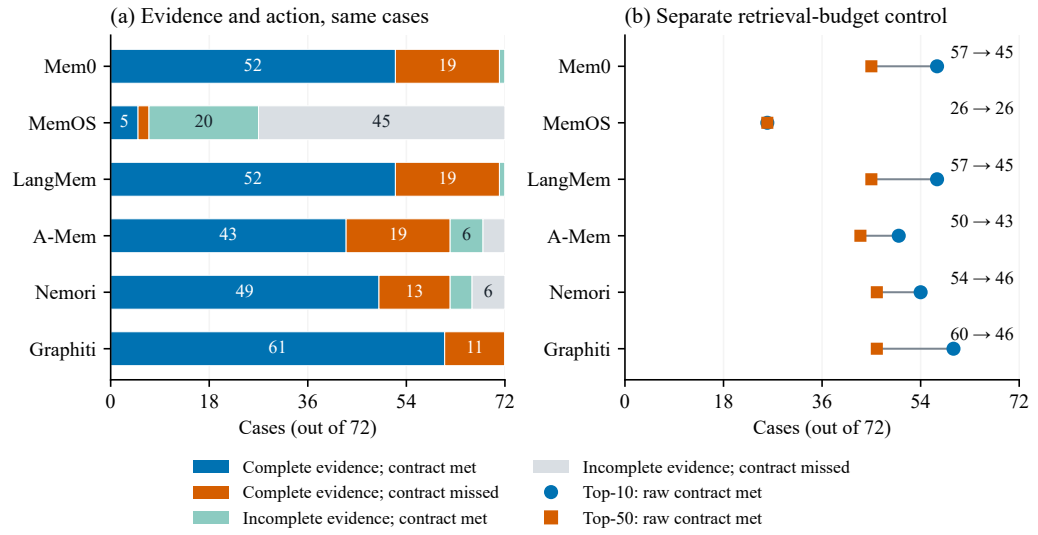

Figure 2: Two upstream diagnostics using the legacy raw action contract. (a) Four evidence/contract outcomes on the same 72 cases for each configuration; orange segments have complete evidence but miss the contract. (b) Separate Top-10 versus Top-50 raw passage counts, linked only within configuration. The panels are different experiments, not consecutive stages; neither metric is the later whole-response score.

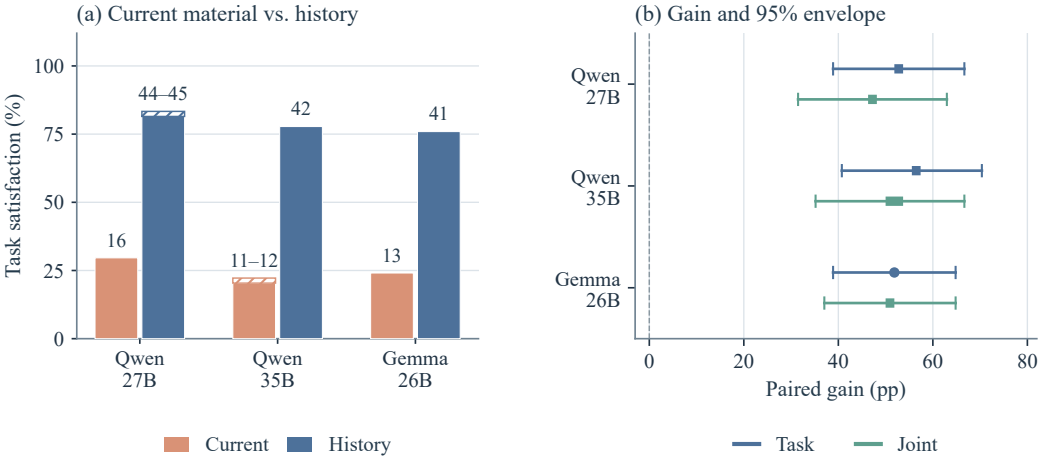

Figure 3: Adjudicated paired-history outcomes for 54 shared-material stories per configuration. (a) Grouped bars compare current material with complete history; labels are passing counts out of 54 and hatched caps retain unresolved score ranges. (b) Task and joint-passage gains with 95% paired story-bootstrap endpoint envelopes; thick segments show score bounds. Models are configurations from two families, not independent replications.

The effect remains positive in each of the three collection batches (Appendix G); the 13-story ablation below reuses story materials and is not an independent replication.

5.3 Selective removal of state evidence

In 13 reused stories, neutral-record removal leaves task satisfaction at 9/13, 7/13, and 7/13 for Qwen 27B, Qwen 35B, and Gemma; key-state removal lowers it to 2–3/13, 1/13, and 0/13 (Figure 4). The neutral-minus-key paired differences are 46.2–53.8, 46.2, and 53.8 percentage points, with 95%

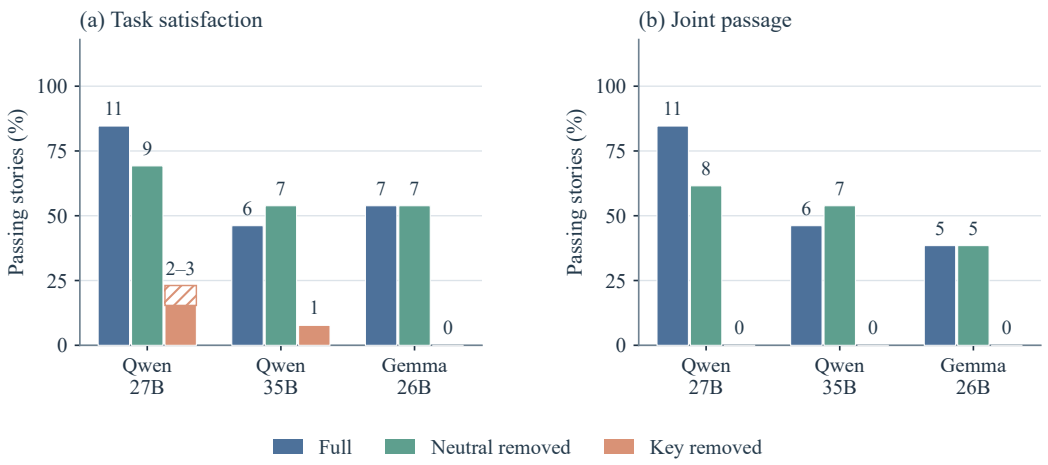

Figure 4: Evidence-removal outcomes on 13 reused stories per configuration. Grouped bars show separate categorical conditions; labels are passing counts out of 13 and hatched caps retain unresolved endpoints. Task satisfaction and joint passage use the same fixed denominator. The comparison concerns selected evidence bundles, not isolated state fields.

Table 1: Joint response assessments with design-specific fixed denominators. Unresolved combinations are not treated as failures.

Response assessment	History	Ablation
Task satisfied; evidence appropriate	151	42
Task satisfied; evidence inappropriate	14	8
Task incomplete; evidence appropriate	101	40
Neither criterion met	51	22
Unresolved combination	7	5
Total responses	324	117

endpoint envelopes of 23.1–76.9 points for all three configurations. Joint passage is 8/13, 7/13, and 5/13 after neutral removal versus 0/13 after key removal.

Full augmented history yields task satisfaction in 11/13, 6/13, and 7/13 stories. Relative to full history, neutral removal therefore changes the counts by -2, $+1$, and 0; it does not help consistently. These are selected evidence-bundle comparisons, not isolated effects of authorization, version, or receipt records.

Key removal can induce a well-grounded inquiry that nevertheless leaves the information-rich reference task incomplete.

5.4 Satisfying the task and respecting evidence are different outcomes

The two criteria separate in both directions under the same adjudicated scoring rule:

The ablation reuses 13 history stories, so these are not 441 independent study units. In the complete-history condition alone, ten responses definitely satisfy the reference task while exceeding visible evidence; 21 are definitely evidence-appropriate but task-incomplete. An unsupported recommendation is not automatically a fabricated execution claim.

5.5 Explicit states expose an inquiry-target boundary

Both criteria pass in both known states for 162/162 scenarios with Qwen 27B, 156/162 with Qwen 35B, and 155/162 with Gemma (Figure 5). Requiring joint passage in all three states yields 161/162 (99.4%, Wilson 95% interval 96.6–99.9), 109/162 (67.3%, 59.7–74.0), and 107/162 (66.0%, 58.5–

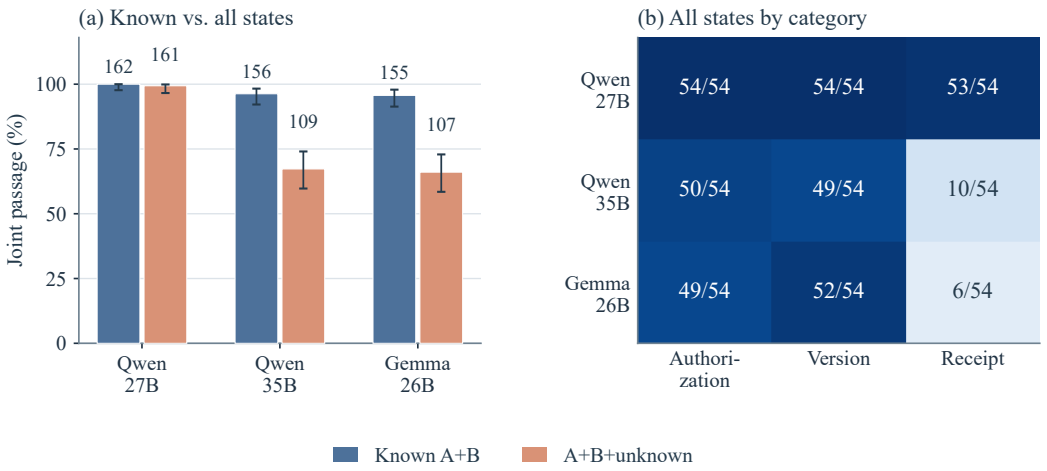

Figure 5: Explicit-state outcomes. (a) Grouped bars compare joint passage for both known states with passage for all known and unknown states over the same 162 scenarios; labels are counts out of 162, with Wilson 95% intervals. (b) The heatmap shows all-state passage by task category, with 54 scenarios per cell. Darker cells indicate higher passage, not greater evidence availability.

Table 2: Native task outcomes and final-response grounding in 106 archived retail tasks. Native success is not the controlled-study task score.

Native outcome	Grounded	Ungrounded	Uncertain	Tasks
Success	45	15	2	62
Failure	35	9	0	44

72.9). These are outcomes under supplied decision rules, not a general reliability ranking; configuration and decoding settings differ.

Unknown-receipt cases dominate non-passage for Qwen 35B and Gemma: all-state receipt passage is 10/54 and 6/54, versus 53/54 for Qwen 27B. Under the strict reference, 32 Qwen 35B and 42 Gemma unknown-receipt responses are evidence-appropriate but task-incomplete. The reference asks about or proposes checking prior delivery, while the visible rule often only asks the agent to inquire under uncertainty. A permission-only question may therefore be reasonable but still miss this narrower target. The 54 inputs share nine rule/request combinations, including one used in 46 cases (Appendix N). The broader interpretation has not been rescored across the complete cohort and is not substituted for the strict endpoint.

5.6 Native outcomes do not determine response grounding

The external cohort shows both directions of disagreement. Among 62 native successes, 15 final responses were ungrounded (24.2%, descriptive 95% Wilson interval 15.2–36.2); among 44 native failures, 35 were grounded (79.5%, 65.5–88.8).

Both discrepancies occur under database-hash and designated-string scoring (Appendix O). These descriptive strata contain different tasks; they are not randomized comparisons of evaluators.

Separate checks found unsupported state claims in 6/83 applicable tasks and authorization failures in 29/92 tasks with state-changing calls, with three and one uncertain judgments, respectively. The 15 ungrounded successful-task replies are not 15 fabricated-completion cases: only three had a definite state-support failure.

6 Discussion

The results separate two points of failure: recovering the records needed for a decision and using them appropriately in the response. Complete retrieval did not guarantee contract passage, while shared-material history and removal comparisons showed that state-bearing records affected task satisfaction. These comparisons locate a response-level consequence of evidence availability without identifying the contribution of each state field or the model's internal reasoning.

The receipt cases also show why the task reference must specify the information an inquiry is meant to obtain. Permission to resend concerns a future action; checking a receipt concerns a past outcome. Both can be reasonable next steps, but they satisfy different objectives. The input review supports the narrower delivery-verification reference while retaining the broader reading. Evaluating that alternative requires applying a revised rubric to the complete affected set.

For document-delivery and service assistants, evaluation should therefore keep the requested objective, visible approval or authorization, and evidence of the reported outcome distinct. The archived retail analysis reinforces this need: native task outcomes and final-response grounding disagree even when tool outcomes are available. It tests the applicability of the distinction to recorded workflows; the history intervention remains confined to the controlled studies.

7 Limitations

The controlled materials share authoring conventions: 13 ablation stories are reused, and 46 unknown-receipt inputs share wording. The three configurations differ in decoding. Bundled removals and category contrasts identify neither individual state mechanisms nor internal reasoning. The post-hoc receipt analysis depends on a supported but non-unique reference interpretation. External evidence comes from one archived agent in simulated retail tasks, not deployment logs or a cross-domain test of our configurations. Reviewer exposure was not fully established; targeted adjudication and the later 450-request return were post-hoc, and neither can detect shared scoring errors. The original return sheets contain AI-role language that conflicts with the authors' account of human review; individually attributable, version-linked A/B confirmation is not available. Public release still requires anonymization and licensing checks.

8 Conclusion

Historical state evidence changed task satisfaction even when the request and material were fixed, while complete retrieval alone did not ensure a valid recommendation. Separating task satisfaction from whole-response evidence appropriateness exposed both unsupported task-satisfying responses and grounded but incomplete ones. Memory evaluation should follow retrieved records through the assistant's recommendation, inquiry, and status report, with each judged against its appropriate source of evidence.

AI Assistance

Generative AI tools were used for literature discovery and for assisting with manuscript editing, including refining search queries, identifying potentially relevant prior work, and improving the clarity, grammar, and readability of the manuscript. The authors independently verified all retrieved references and reviewed all AI-assisted edits. Generative AI tools were not used for research planning, experimental design, analysis-code implementation, synthetic data generation, annotation, figure creation, result analysis or interpretation, or substantive drafting of manuscript sections. The authors are responsible for the final methods, results, interpretations, figures, citations, and text.

Reproducibility

The original frozen score table, its 216-material catalogue, and all 1,899 responses remain preserved. A separate adjudicated score version applies 36 documented cell differences from the 450-request return, retaining old and new values, unresolved endpoints, and row identities; figures and reported counts are regenerated from that version (Appendix M). The 54-input interpretation review and 106-task external replay remain separate validation units, not additions to controlled-study denominators.

References

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0856bc553d3e3b9827e5140d0ad3bf8d-Paper-Conference.pdf.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

Alejandro Cuadron, Pengfei Yu, Yang Liu, and Arpit Gupta. SABER: Small actions, big errors—safeguarding mutating steps in LLM agents. *arXiv preprint arXiv:2512.07850*, 2025. URL https://arxiv.org/abs/2512.07850.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

LangChain. LangMem. Software repository, 2025. URL https://github.com/langchain-ai/langmem. Accessed 2026-09-23; evaluated adapter configuration is documented separately.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025. URL https://arxiv.org/abs/2507.03724v4.

Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341v1*, 2025. URL https://arxiv.org/abs/2508.03341v1.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2ActBench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8173–8190. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.370. URL https://aclanthology.org/2026.acl-long.370/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/93e444b4a1a64815631c3c74977ee200-Paper-Conference.pdf.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025. URL https://arxiv.org/abs/2502.12110.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Zep. Graphiti: A framework for building temporal knowledge graphs. Software repository, 2026. URL https://github.com/getzep/graphiti. Accessed 2026-09-23; evaluated adapter configuration is documented separately.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. AMA-Bench: Evaluating long-horizon memory for agentic applications. *arXiv preprint arXiv:2602.22769*, 2026. URL https://arxiv.org/abs/2602.22769.

A Scope and analysis units

The experimental analysis contains four complementary evidence layers. A separate external cohort tests whether native task outcomes identify evidence-appropriate final responses (Section O).

Table 3: Study design and analysis units. The ablation reuses history stories; upstream and archived cases are not pooled with controlled responses.

Layer	Changed evidence or assessment	Unit	Scope
Upstream diagnostic	Retrieval configuration and, separately, Top-10/Top-50 budget	Shared case	72 cases; 12 ledgers
Paired history	Historical records vary; request and material fixed	Story	54 pairs per model
Evidence removal	Neutral or key-state records removed from reused stories	Story	13 stories; 3 arms
Explicit states	Known A/B and unknown U under fixed request and rule	Scenario	162 scenarios; 486 replies per model
External cohort	Native outcome compared with final-response grounding	Archived task	106 retail tasks

The six memory-system configurations are not six independent populations. Responses to the same story are not independent stories. The 13 ablation stories reuse the 54-story set. The unified catalogue therefore contains 216 material identifiers: 54 history stories and 162 explicit-state scenarios. It links 324 history responses, 117 ablation responses, and 1,458 explicit-state responses, totaling 1,899 response records. This total excludes upstream retrieval diagnostics and is not a denominator for a pooled accuracy rate.

B Data provenance

The controlled tasks cover document delivery, version updates, authorization changes, recipient identity, send receipts, retractions, and workflow progress. The final story texts were written or substantively rewritten by three human contributors and reviewed before use. Earlier AI-assisted drafts are retained for provenance. The data are synthetic and contain no natural user logs or real personal records.

For every frozen request, the public package should include:

1. story and checkpoint identifier;
2. model-visible system and user messages;
3. condition label;
4. model identifier and request parameters;
5. request and response SHA-256 hashes;
6. raw response, including malformed output;
7. reviewer A and B pre-coordination judgments and reasons;
8. final coordinated values and the metric-registry version.

Scoring references are stored separately from model-visible messages. The local candidate package exposes the task reference needed to audit task satisfaction and the exact visible messages needed to audit evidence appropriateness. Author identities, review identities, endpoint addresses and local output paths are excluded from that candidate. Full pre-coordination returns and service records remain internal pending a separate privacy and licensing review; the package is not yet a public release.

C Models and inference configuration

Table 4: Models and inference configuration

Key	Frozen model identifier	Family	Role
27B	`Qwen/Qwen3.8-27B`	Qwen	matched configuration
35B	`Qwen3.6-35B-A3B-FP8`	Qwen	matched configuration
Gemma	`gemma-4-26B-A4B-it`	Gemma	cross-family configuration

Requests use independent conversations and frozen prompts. Each experiment's request manifest records temperature, output-token limit, and thinking configuration; settings must not be assumed identical across studies. The explicit-state study uses a 900-token cap and a documented 35B thinking override, as described in Section L. Endpoint-reported model metadata, token counts, and request manifests are preserved. Model differences are not pure parameter-count effects because architecture, quantization, serving stack, and model family differ.

D Output semantics

The expected schema contains `current_actions`, `conditional_actions`, and a natural-language `reply`. The scientific scoring is semantic. A readable answer is not penalized merely because a model emits different formatting. Parser status is retained as engineering provenance; semantic reviewers score recoverable content. If content cannot be recovered, the frozen missing-response rule applies. No formal response is automatically retried or replaced after observing its result.

The main action vocabulary is:

- `SEND_SEGMENT`: recommend sending a specified current segment to the correct recipient;
- `ASK_USER`: request information that blocks the decision;
- `STATE_ACK`: state the strongest status supported by evidence;
- `ADVANCE_WORKFLOW`: recommend a justified workflow transition.

E Metric registry

E.1 Overall metrics and reproduction-field notation

The main text uses the names task satisfaction and evidence appropriateness. The symbols below identify the corresponding fields in existing scoring files and statistical formulas; they do not add new criteria.

Task satisfaction (F). True when the response satisfies every applicable requirement in the frozen task reference. These may include the recommendation, a workflow step, a state report or a specific inquiry target. The rubric evaluates task coverage, not tool-execution success. A correct action category alone is insufficient when a required component is absent.

Evidence appropriateness (V). True when the entire response is appropriate given evidence visible in that condition. This includes current actions, conditional actions, questions, and status wording.

F and V are not nested. F=0/V=1 can be a reasonable but task-incomplete inquiry. F=1/V=0 can be a task-satisfying recommendation that assumes unsupported authorization, proposes an unsupported conditional branch, or overstates a status. The quadrant does not identify which of these occurred.

Joint passage J is F AND V with unresolved values retained. If either component is 0, J=0; if both are 1, J=1; otherwise J is unknown. Denominators retain unresolved responses. Reported lower and upper counts enumerate possible binary resolutions rather than imputing labels.

The operational rules distinguish recommending an action from asking permission to execute an already specified recommendation, the absence of a receipt from evidence of failed delivery, and verifying a past event from authorizing a future event. Exact-text requirements concern content, not

JSON syntax. The same rules apply to natural-language responses. Section 3 of the main text distinguishes action, workflow/status, and inquiry requirements without introducing separate subscale scores.

E.2 Diagnostic metrics

- `unsupported_current_action`: a current action or current state assertion lacks visible support; an unexecuted conditional branch is not a current action.
- `target_error`: an expressed recipient, object, or workflow target is wrong.
- `version_error`: the answer explicitly uses or merges an obsolete, wrong, or unconfirmed version.
- `content_error`: an expressed material claim or payload contradicts or invents task content.
- `unnecessary_clarification`: the response asks for, or gates the recommendation on, information already available or unnecessary.
- `status_overclaim`: the response upgrades state beyond evidence, for example prepared to sent or sent to received/approved.
- `reasonable_inquiry`: an actual inquiry addresses a blocking identity, version, authorization, material, or delivery-state gap.

In the earlier nine-metric release, these diagnostics use fixed denominators for all 441 answer-bearing responses; absence of an event is false rather than not applicable. The full-return recalibration updates task satisfaction and evidence appropriateness, not these seven diagnostic labels. Updated diagnostic prevalence and status-overclaim counts cannot be inferred from the new joint assessment categories.

F Review, coordination, and quality control

This section documents the historical nine-metric coordination stage. Section M reports the completed two-stage review of task satisfaction and evidence appropriateness and the current full-return integration.

Two human reviewers separately checked AI-assisted worksheets before coordination. The consolidated return set comprises 31 legacy Qwen responses and 279 Gemma responses, each with nine metrics. These 310 responses are not a random sample of the 441 primary responses. The pre-coordination agreement record for this return set is:

- 2,790 response–metric cells;
- TRUE/FALSE/NA raw agreement 81.2%, Cohen's kappa 0.579, Gwet's AC1 0.759;
- applicability raw agreement 91.4%, kappa 0.050, AC1 0.906;
- among 2,543 cells marked applicable by both, binary-value agreement 88.8%, kappa 0.720, AC1 0.813.

The divergence between applicability kappa and AC1 is reported because the applicability classes are highly imbalanced. We do not report coordination-after agreement as if it were independent reviewer reliability.

The 525 disputed cells comprised 109 core task-satisfaction, evidence-appropriateness, or status decisions and 416 secondary applicability/value decisions. Original judgments, targeted checks, corrections, and adjudication reasons remain in the historical audit. That stage neither estimates annotation accuracy nor supersedes the current 441-response two-stage review in Section M. Its status-overclaim labels were not recomputed after the full-return recalibration; a task-satisfying but evidence-inappropriate response cannot automatically be counted as a fabricated completion claim.

G Statistical analysis

Let F_i^H and F_i^C denote task-satisfaction indicators for story i under complete history and current-plus-material conditions. The shared-material history estimand is the paired risk difference:

$$\Delta_F = \frac{1}{n} \sum_{i=1}^{n} \left(F_i^H - F_i^C \right). \tag{1}$$

For the state-evidence ablation, F_i^N and F_i^K denote task-satisfaction indicators under neutral-removed and key-removed conditions. Its frozen-plan contrast is $n^{-1} \sum_{i=1}^{n} (F_i^N - F_i^K)$. The local protocol freeze is not an external preregistration.

Confidence intervals use 10,000 paired bootstrap draws over stories. Each draw resamples story identifiers with replacement and keeps all paired conditions for that story. We report positive/zero/negative story counts so a positive mean cannot be read as universal improvement. Cross-model counts are descriptive; no model–story row is treated as an independent story.

G.1 Shared-material history: adjudicated results

Table 5: Paired history effects in the separate adjudicated score version. Counts use 54 fixed story units per configuration; intervals are 95% paired-bootstrap endpoint envelopes.

Configuration	Current	History	Gain (pp)	95% interval (pp)
Qwen 27B	16/54	44–45/54	51.85–53.70	[38.89, 66.67]
Qwen 35B	11–12/54	42/54	55.56–57.41	[40.74, 70.37]
Gemma 26B	13/54	41/54	51.85	[38.89, 64.81]

For task satisfaction, definite positive/zero/negative paired differences are 28/25/0, 31/22/0, and 28/26/0; one pair in each Qwen configuration is unresolved. Evidence-appropriate current-to-history counts are 46–47 to 48, 32–33 to 47–49, and 37–38 to 43–44, respectively. Its definite declines are 5, 1, and 6 stories, so improved task coverage must not be reported as a universal safety gain. The ablation's reused stories are not independent corroboration.

G.2 State-evidence removal: adjudicated results

Table 6: Selected state-evidence removal in 13 reused stories per configuration.

Configuration	Full	Neutral removed	Key removed	Neutral − key (pp)	95% interval
Qwen 27B	11/13	9/13	2–3/13	46.15–53.85	[23.08, 76.92]
Qwen 35B	6/13	7/13	1/13	46.15	[23.08, 76.92]
Gemma 26B	7/13	7/13	0/13	53.85	[23.08, 76.92]

Corresponding joint-passage counts under full/neutral/key conditions are 11/8/0, 6/7/0, and 5/5/0 out of 13. Neutral-minus-full task-satisfaction differences are -15.38, $+7.69$, and 0 points; only the primary neutral-minus-key contrast is consistently positive here. These are bundled removals, not single-record mechanism tests.

G.3 Batch-stratified check

History collections comprise H1 (12 stories), H2 (24), and H3 (18). Condition counts remain positive in each batch under the adjudicated score version. The 10,000-draw paired bootstrap resamples story units, preserving batch sizes; unresolved labels are represented by endpoint envelopes.

Table 7: Task-satisfaction counts by collection batch, current → complete history.

Batch	Stories	Qwen 27B	Qwen 35B	Gemma 26B
H1	12	3 → 11	2 → 11	1 → 9
H2	24	10 → 21	7 → 21	9 → 21
H3	18	3 → 12–13	2–3 → 10	3 → 11
All	54	16 → 44–45	11–12 → 42	13 → 41

The older leave-one-batch-out range and boundary-cell envelope were computed on the predecessor table. We do not reuse them as if recalculated for this version.

H Upstream diagnostics and ablations

H.1 Full evidence versus action

On 72 shared cases, Graphiti recovered complete current-state evidence in all 72, while 61 outputs met the legacy action contract. These metrics are aligned by case but are not interchangeable.

The six-configuration table below is recomputed from `a/case_diagnostics.csv` using `evidence_complete` and `raw_correct`. All configurations share 72 case identities from 12 ledgers, with 24 expected sends, 24 inquiries and 24 workflow advances. The legacy contract checks target and content for sending, but primarily the action category for inquiry and workflow advance. A pass with incomplete reference evidence therefore does not establish that the entire response is appropriate.

Table 8: Full evidence versus action

Configuration	Complete / raw pass	Complete / raw fail	Incomplete / raw pass	Incomplete / raw fail	Total raw pass	Historical gated pass
Mem0	52	19	1	0	53	54
MemOS	5	2	20	45	25	25
LangMem	52	19	1	0	53	55
A-Mem	43	19	6	4	49	48
Nemori	49	13	4	6	53	54
Graphiti	61	11	0	0	61	61

Each row has denominator 72. Historical reports used gated correctness for their four-way table; Figure 2 consistently uses the raw, pre-gate contract. Nine rows differ between raw and gated outcomes (six improvements and three regressions). This is a metric distinction, not missing cases. Adapter choices, condition labels and collection batches preclude treating the table as a controlled ranking of the systems' complete capabilities.

H.2 Retrieval budget

Six cache configurations were evaluated at Top-10 and Top-50. Five action-accuracy values decreased and one was unchanged. Since more than budget changes–including returned record identity, order, and total context–the result is reported as a non-monotonic budget finding, not a pure context-length mechanism.

Table 9: Retrieval budget

Cache configuration	Complete evidence: Top-10 → Top-50	Raw contract pass: Top-10 → Top-50
Mem0	71 → 72	57 → 45
MemOS	7 → 7	26 → 26
LangMem	71 → 72	57 → 45
A-Mem	62 → 71	50 → 43
Nemori	62 → 70	54 → 46
Graphiti	72 → 72	60 → 46

Each cell has denominator 72. The budget experiment contains 864 requests with saved responses; 862 ended with `stop` and two with `length`. It reconstructs public records from cached identifiers and uses one 35B controller. Its responses differ from the reused same-case diagnostic, so their Top-10 numbers are not interchangeable. Figure 2 shows descriptive counts; it does not reuse the old gated-score intervals for raw scores. Source: `benchmark_followup_abc_v1/b_full/summary.json`, row-level `raw_exact` and `evidence_complete` fields.

H.3 Targeted repair

Only 10/72 restricted-cache cases lacked required evidence. The repairable subset is defined by evidence absence, not by selecting incorrect answers. Original, repaired and sham inputs yield the following raw, ungated legacy-contract results.

Table 10: Targeted repair

Configuration	Original / 10	Required-evidence repair / 10	Sham repair / 10
Qwen 27B	5	10	4
Qwen 35B	4	9	4
Gemma	3	9	3

Each configuration has 216 saved responses across all 72 cases and three conditions; the other 62 cases are not additional successful repairs. The repair uses public records selected with reference knowledge and is not token-matched. It is an oracle diagnostic, not a deployable method. Source summaries are `benchmark_followup_abc_v1/c_full/summary.json`, `benchmark_c27b_replication_r1/full/summary.json`, and `gemma_matched_analysis_20260918_r1/01_E1_Gemma 自动评分.json`. Gemma uses the inherited E1 semantic parser on the 216 original frozen responses, all of which are parseable by that scorer; no recovery response replaces an original. These automated legacy-contract judgments are not the human-reviewed story rubric.

H.4 Extension data

Mem0, MemOS, and LangMem each completed 216 retrieval cases on 18 additional ledgers. These cases are distinct from the 216 material identifiers in Section A and must not be pooled with the 54-story action analysis. Requested Top-k is not necessarily the number returned by every adapter; MemOS returned at most five records in the recorded runs.

Table 11: Extension data

Configuration	Complete evidence @10	Complete evidence @50	Stale-only failure @10	Stale-only failure @50
Mem0	26/216	205/216	165/216	11/216
LangMem	26/216	205/216	165/216	11/216
MemOS	0/216	0/216	125/216	125/216

The 216 full ranked result lists are identical between these Mem0 and LangMem runs, not merely equal in aggregate score. Mem0 stores normalized records with inference disabled; the LangMem adapter writes normalized records through its manage/search tools and uses an in-memory store with a shared embedding configuration. These are distinct evaluated adapters, not independent retrieval-behavior replications. Their results do not cover all features or default capabilities of either system. Source: `benchmark_v2_extension18_runs_r1/full/{mem0,memos,langmem}/summary.json` and corresponding returned lists.

H.5 Frozen response-level examples

The artifact's `provenance/example_links.json` locates the two Section 3.1 examples in `answers.jsonl` by request ID: `H1_35b_H1-S02_current_plus_materials` and `H3_27b_H3-S01_current_plus_materials`. Each record contains the original visible input, complete response and separately stored task reference. The first is evidence-appropriate but task-incomplete; the second satisfies the task but lacks visible approval for its recommendation. The latter explicitly denies sending anything and therefore does not exemplify fabricated execution. These explanatory cases were selected after scoring, not sampled to estimate prevalence. The internal case-study record retains a third, earlier illustration whose hidden prerequisite made it less suitable for the main text.

For self-contained inspection, the following are English paraphrases of the frozen Chinese inputs and replies, not verbatim transcripts. In H1-S02, the visible material is a podcast erratum and the request

asks whether it was sent to the specified recipient and what to recommend if not. No send receipt is visible. The 35B reply says it cannot verify sending without a mail-query tool and recommends the original only conditionally; the held-out reference requires a send-now recommendation because full history records a failed send without a later successful resend ($F = 0, V = 1$). In H3-S01, visible notice P says Tuesday with a re-login, whereas Q says Thursday without a re-login. The user requests the approved notice, but visible input contains no approval. The 27B reply recommends Q because it is later and more specific while explicitly denying that it sent anything; the held-out reference selects Q ($F = 1, V = 0$). The complete original messages, replies, and references remain in the local candidate artifact, not in this PDF.

The adjudicated complete-history slice has 162 responses (54 stories and three configurations). It contains 116 definite responses passing both criteria, ten definite task-satisfying but evidence-inappropriate responses, 21 definite evidence-appropriate but task-incomplete responses, 11 definite failures of both, and four unresolved combinations. These are response counts, not independent-story estimates. The earlier `complete_history_diagnostic.py` export reproduces the predecessor table and is preserved as historical provenance, not as the current numerator source.

I Negative results and class imbalance

I.1 Earlier history and prompt controls

The earlier study used 24 test stories with 51 checkpoints. Its complete-history and current-request-only comparisons could change document availability and were collected in different batches; they are exploratory precursors, not additional replications of the 54-story shared-material contrast. The earlier evidence-visibility by stale/neutral-history test used five checkpoints, with four in the primary group. Its sufficient-evidence cells contained no unnecessary-clarification events in the initial two-configuration analysis, which does not establish equivalence. A second prompt produced mixed, configuration-dependent effects rather than a uniform improvement. These negative and mixed findings remain in the earlier protocol's score release. We do not transplant its nine-metric percentages into the current two-criterion table.

I.2 Current reporting

The current result tables retain the history-associated evidence-appropriateness decreases (Section G.1), mixed neutral-removal effects (Section G.2), and retrieval-budget declines (Section H.2). The earlier nine-metric release separately found a Gemma unnecessary-clarification increase from 12/54 to 24/54; that secondary label was not recalibrated with the two overall judgments. Parser and format failures remain in the raw records without result-dependent re-sampling. Applicability-class imbalance explains the low kappa but high AC1 in Section F. Sparse workflow-advance cases and earlier negative effects remain under their original protocol rather than entering the present pooled scores.

J Exact source artifacts

The predecessor release remains unchanged at `paper/iclr2027/artifact/score_release/`. The separate adjudicated version at `paper/iclr2027/artifact/score_adjudicated_450/` contains 1,899 row-linked F/V scores, a 36-cell old/new change log, source-hash verification, and recalculated statistics. The 450-request return reports 900 F/V cells in scope and 49 additional explicit-state request readings; it is not a new independent full-corpus review of all 1,899 responses. Thirteen metric cells in the history/ablation material remain unresolved, corresponding to 12 response-level unresolved combinations; original answers and earlier scores are not overwritten.

The bounded `provenance/` export remains the source for earlier two-phase returns and coordination. The authors' account is that A/B each reviewed independently before AI organized their returns, but original documents contain contrary AI-role language. The supplied anonymous acknowledgement does not individually bind A and B to exact file hashes. Individual reviewer identities and the document versions covered by their confirmation therefore remain unverified; the original records are preserved.

K Artifact availability boundary

The manuscript and local candidate artifact are distinct. Numerical results, source-linked scores, and regeneration scripts have been checked locally, but the candidate has not been released publicly. Full pre-coordination returns and service records remain internal pending privacy and distribution-rights review. Local checks establish arithmetic and record linkage, not reviewer identity, semantic validity of every judgment, or permission to publish the artifacts.

L Controlled state-evidence extension: collection and scoring units

We collected 1,458 responses from three model configurations on 162 additional short scenarios. Each scenario has two contrasting known states and one unknown state, with 54 scenarios each for authorization, approved-version selection, and delivery receipts. Within a scenario, the task material, recipient, decision rule, and current request remain fixed; only the current-state record changes. Each configuration is evaluated on all 486 conditions. The primary endpoint requires both task satisfaction and evidence-appropriate behavior in all three states of a scenario. Response-level endpoints use 486 as the per-configuration denominator, whereas the primary endpoint uses 162.

The response-level counts below retain the strict reference under the separate adjudicated score version. They count individual answers, not scenarios passing all three states. Responses within a scenario are dependent and are not treated as 486 independent scenario observations.

Table 12: Controlled state-evidence extension: collection and scoring units

Configuration	Task-satisfying responses	Evidence-appropriate responses	Responses passing both
Qwen 27B	486/486	485/486	485/486
Qwen 35B	440/486	463/486	431/486
Gemma	436/486	469–471/486	427/486

The Gemma configuration produced 486 nonempty responses using 324,855 input tokens and 121,319 output tokens. The Qwen 35B configuration produced 486 nonempty responses using 292,275 input tokens and 93,936 output tokens. All 972 responses ended with the service's normal stop indicator, and their input token counts matched the preflight counts. These are collection measurements, not semantic correctness rates. The Qwen configuration explicitly disabled thinking while retaining temperature zero and the 900-token output limit. An earlier request under its original configuration exhausted that limit before returning final content; this attempt is retained separately rather than mixed into the revised configuration's results. Reported usage describes these service configurations and does not establish a model-level efficiency ranking.

The short scenarios test the use of explicit state evidence under supplied decision rules. They are analyzed separately from the 54 shared-material history stories and the 13-story ablation subset. Their counts cannot be pooled into a single history-necessity denominator. The later 27B run adds 486 responses; its service-default thinking setting was not overridden. All three configurations have separate A/B returns for task satisfaction and evidence appropriateness. Of 2,916 metric cells, 2,890 agree and 26 were resolved by an AI-assisted coordinator, affecting 24 responses. Original human-return artifacts and the user's confirmation of their human review are retained alongside earlier AI-assistance declarations. Coordination is not an independent third-reviewer assessment.

L.1 Reproduction and figure construction

The current Figure 5 is regenerated by `paper/iclr2027/figures/adjudicated_results/render_current.py` from the adjudicated score table after verifying request identities and all-state scenario membership. It displays grouped bars and a category heatmap; scenario proportions use 162 as the denominator per model. Its 95% Wilson intervals describe scenario outcomes, not seed variability, and repeated templates limit population interpretation. Editable SVG and vector PDF are retained with the figure. The earlier `work/delivery_memory_pilot/plot_counterfactual_evidence.py` and its heatmap remain predecessor-release provenance, not the source of this manuscript figure.

`work/delivery_memory_pilot/analyze_counterfactual_review_sensitivity.py` reproduces category counts, quadrants, known-state passage, and two post-hoc sensitivity analyses. Neither script modifies raw responses or frozen scores or makes model requests.

L.2 Adjudication and inquiry-target sensitivity

The predecessor score version's 26-cell disputed-label bounds and permissive unknown-receipt envelope are retained in its original analysis files. They were not rerun under the present 36-cell change set and are therefore not numerical sensitivity claims for this manuscript version. Under the adjudicated strict-reference scores, there are 32 Qwen 35B and 42 Gemma unknown-receipt responses that respect visible evidence but omit the reference-specified prior-delivery inquiry. Two Gemma evidence judgments remain unknown. A broader permission-only inquiry rule would require a uniformly revised rubric and complete rescoring, not selective favorable relabeling. Section N documents why the current reference has a textual basis without claiming it is the only reasonable interpretation.

L.3 Limits of the 27B check

The 27B follow-up check covered all 486 records structurally and applied a heuristic semantic-risk scan. Two flagged responses were inspected; neither required a score change. This was not an independent full human reannotation and does not establish zero residual error. The existing evidence-appropriateness failure on C74-U remains in the final table.

M Complete-return integration of the 441-response review

The review partitions contain 108, 182, and 151 responses, together covering all 441 original response identities. For the last partition, the author separately confirmed personal review by both reviewers for evidence appropriateness before this batch's task references were shown, and subsequently for task satisfaction. The original AI-assistance declarations are retained with these confirmations. The last task-satisfaction returns agree on 134 definite values out of 151 responses; eight responses contain at least one uncertain value. Among the 143 jointly definite judgments, agreement is 134/143 and Cohen's kappa is 0.861. These pre-coordination figures describe this partition rather than all 441 responses.

For evidence appropriateness in that same partition, there are 143 definite agreements out of 151 responses. Among the 147 jointly definite judgments, agreement is 143/147 (97.28%) and Cohen's kappa is 0.893. Reporting jointly definite agreement does not discard uncertain responses from the experiment denominator; it specifies the subset on which binary agreement is defined.

The predecessor release contains 56 coordinated history/ablation metric cells, covering earlier coordination and same-rule corrections. The author subsequently confirmed that the consolidated rules and decisions had undergone independent human review. Earlier AI-assisted proposals remain identifiable; this confirmation does not imply unaided review or add a third reviewer. At that predecessor-release stage, four cells remained unresolved across three responses; the current counts are given in Section J. Relative to the previous partial table, 47 task-satisfaction or evidence-appropriateness values change. All 882 metric cells have old and new values, source-specific reasons and provenance. All 441 raw-response identities match the prior manifest. Original returns and raw outputs remain unchanged.

The later author-supplied acknowledgement states anonymously that A and B accept the final review. It does not identify each reviewer, name the return files or versions they confirmed, or establish their exposure to earlier scores. We retain the authors' account alongside the original AI-assistance declarations, but do not treat this acknowledgement as a separately attributable signature or the post-hoc boundary adjudication as a new blind review.

Sixteen groups have identical prompts, inputs, responses, and reference goals across designs. One conflicting task-satisfaction pair was identified: RSR168 in the ablation and CAL006 in the history study. The released correction applies the previously frozen strict interpretation consistently to both. The 57-cell sensitivity set includes the counterpart rather than varying only the newly corrected record. Same-class inspection reads all 30 responses in the procurement-validation and

shared-list-restoration families. An additional full-corpus text screen finds three candidates for treating absent receipts as authorization to resend; reading these candidates supports one new evidence-appropriateness correction. This screen is not a third independent semantic review of all 441 responses.

M.1 Predecessor-release sensitivity analyses

For the predecessor score release, the scoring-boundary envelope permits each cell in the specified boundary set to vary between zero and one while other values remain fixed. Its lower effect endpoints for task satisfaction stay positive for all configurations: 51.85, 53.70, and 48.15 percentage points for history, and 38.46, 38.46, and 53.85 points for neutral-minus-key removal. These deterministic ranges are not confidence intervals, an exhaustive audit of annotation errors, or a test of shared reviewer errors.

A second analysis varies the interpretation of the reference requirement in two complete task families. One reading requires explicitly restating that an earlier synchronization was withdrawn; a broader action-sufficient reading accepts restoring the currently specified list without that restatement. Another reading requires an explicit internal-validation step; a broader reading accepts the specified approval entry as covering that check. Every inspected response in these families is listed, including those whose value would not change. The alternative overrides and resulting contrasts are stored under `posthoc_two_family_construct_sensitivity` in the summary JSON. This post-hoc analysis does not replace the strict-reference metric or establish that the broader reading has been independently validated.

M.2 Frozen release and portable numerical reproduction

The predecessor artifact entry point is `paper/iclr2027/artifact/score_release/README.md`. Run `python reproduce.py` from that directory with Python 3.10 or newer and NumPy 2.1.3. The program checks unique response identities, paired-condition membership and all-state membership before reproducing the predecessor-release contrasts and counts. It resamples stories 10,000 times with seed 20260922, stratifying the combined history contrast by collection batch. Unknown values use endpoint envelopes. The artifact includes `scores.json`, `materials.csv`, `answers.jsonl`, `expected.json` and a standalone reproduction program; it needs neither project-internal imports nor a model endpoint.

`answers.jsonl` links every score to its original visible messages and response, with scoring references kept in a separate field. The bounded coordination export in Section J makes the original A/B judgments for all coordinated cells inspectable. The full original archive remains internal. Neither numerical reproduction nor this export validates the semantic judgments, converts AI suggestions into human review, or validates alternative task definitions. Final manual anonymity and distribution checks remain necessary.

The internal builder `finalize_measurement_release.py` checks the confirmed source version, all 441 response files, the 1,458 explicit-state raw-response checksums, and the 35B revised-request mapping. It refuses to replace a differing frozen release. Neither builder nor reproduction entry point makes model calls. The adjudicated controlled-response Figures 3–5 use `paper/iclr2027/artifact/score_adjudicated_450/scores.json`; Figure 1 is schematic and Figure 2 uses the separate upstream diagnostic data. `work/delivery_memory_pilot/apply_450_adjudication.py` and `paper/iclr2027/figures/adjudicated_results/render_current.py` verify the handoff hashes and regenerate new statistics/figures without editing the predecessor. The change log records each old value and reason. Public hosting, licensing and final anonymous-PDF checks remain separate release steps.

N Input-expectation validation for unknown delivery receipts

N.1 Question, material, and staged review

This validation asks whether the requirement to inquire about a previous delivery is recoverable from the visible task input, rather than merely specified by the private scoring reference. It covers all 54

unknown-receipt conditions in the 162-scenario explicit-state study. These are not the 54 shared-material history stories. The corresponding 162 model responses remain unchanged, and this review adds no model calls or new task materials.

The first stage presented the frozen system prompt and visible inputs, without model responses, existing scores, private references, or the second-stage questions. Reviewers described the next step they understood the input to require. According to the authors' process confirmation, the two reviewers formed these judgments independently before using AI to organize the returns, without first seeing model answers, scores, or the second-stage questions. This confirmation does not establish that the materials were previously unfamiliar to them.

The directed second stage asked reviewers to classify the prior-delivery inquiry as explicit, inferred from context, unsupported, or uncertain. It then asked whether requesting only new permission was sufficient under that interpretation. One reviewer initially supplied item-specific judgments. The other initially supplied a common sufficiency analysis without inspecting each input, then completed a source-specific follow-up for all 54 conditions. Both returns are retained. The follow-up contained the original system prompt and individual inputs, not the other reviewer's judgments, private references, model answers, or scores. Its actual completion time, AI-use sequence, and exposure to other materials before the follow-up were not separately established. It is therefore a directed source check, not a new blinded reliability study.

N.2 Requirement specificity and alternative readings

The final item-specific labels agree on the requirement category and permission-only sufficiency in all 54 conditions. Their distribution is:

Table 13: Prior-delivery inquiry in 54 unknown-receipt inputs. The last column counts cases where a permission-only question omits the delivery-verification target under the strict reading.

Input subset	n	Explicit	Inferred	Permission-only
Initial subset	8	2	6	8
Extensions	46	0	46	46
Total	54	2	52	54

Only C10-U and C22-U explicitly request verification of the earlier delivery. In the remaining 52 conditions, the reviewers infer the inquiry target from the previous delivery attempt, the unknown receipt, and the request for a next step. These conditions cannot all be described as explicitly requiring a delivery check. Moreover, the general rule and final request have only nine distinct text combinations, with 46 conditions sharing one combination. Agreement over these repeated forms should not be presented as evidence from 54 independent natural workflows or used to claim a precise population-level validity estimate.

Under the delivery-verification reading, a permission-only response omits a question or proposed check concerning the earlier outcome. The shortfall is not that the model fails to obtain evidence unavailable to it. Asking for new resend permission can remain grounded and does not itself constitute an unauthorized send or a redundant request for permission already granted. Both reviewers retain a broader reading under which seeking new permission is a reasonable next step. The source check consequently supports a text-grounded interpretation of the specified reference, without establishing that it is the only necessary interpretation.

We retain the strict-reference results and describe them as satisfaction of that reference, separately from the appropriateness of the proposed next step. The permissive calculation discussed in Section L.2 is a post-hoc bound for the predecessor release, not a validated alternative score for the current version. Adopting a broader endpoint accepting any reasonable next step would require a stated rubric applied to all 162 corresponding responses, including those that already pass, rather than selectively relabeling failures. The later 54/54 agreement is descriptive and cannot by itself settle this construct question.

N.3 Traceability

The internal source-follow-up receipt records 54 matched input identities, 216 completed required fields, and 324 source-line quotations. The itemwise comparison retains the original common analysis, later input-specific judgments, and reference side by side. The original AI-assistance declarations remain beside the authors' account of human judgment. These internal files are not claimed to be accessible through this PDF. `accept_source_followup.pyverify`, under `notes/measurement_external_validation/input_expectation/`, checks these records without changing scores; it establishes traceability, not the validity of the semantic interpretation.

O External applicability to archived retail tool interactions

O.1 Fixed cohort and replay

We test whether the distinction between task outcomes and grounded reporting is useful beyond the constructed materials using archived retail trajectories from the official τ-bench repository. The pinned version is commit `7362a1708d720a4d697d042f21fd8637ea144c3b`, with trajectories from `historical_trajectories/gpt-4o-retail.json`. These are interactions of an archived GPT-4o agent in a simulated retail environment, not real customer records or new evaluations of our three model configurations.

The archive contains four trials for each of 115 tasks. We fix trial 0 and include the 106 tasks whose reference actions contain a database modification. Selection uses task specifications, not native success or an observed grounding error, and includes 14 tasks in which the agent actually makes no modifying call. Membership was fixed before semantic annotation. One trajectory's structure and native outcome had been inspected for feasibility before the freeze. This is an archival analysis, not a prospectively preregistered test on unseen data. The other trials are not additional independent observations in the reported cohort.

Offline replay executes the recorded calls against the original tool implementations, data, policies, and task definitions, with fresh initial state for each trajectory. It does not generate new agent responses or simulated-user messages. All 753 recorded tool outputs match the replayed outputs, and all 106 native task rewards are reproduced. This validates compatibility with the pinned environment, not the correctness of the human annotations. The archived system policy is recoverable, but the exact tool-schema payload transmitted to the historical model is not, so local tool definitions do not establish the original request payload.

O.2 Three distinct annotation objects

Reviewers assess the final user-facing response against information available before that response, including the visible system policy. Four trajectories have further tool calls after that text. Those later results cannot retroactively support it. They also assess state claims within the response and authorization before each modifying call. Authorization concerns the requested object, scope, necessary payment information, and user confirmation. It is not a complete policy-compliance score. A rejected call can still lack authorization.

Table 14: Three distinct annotation objects. N/A means the object is absent for that task; question marks indicate unresolved judgments.

Measure	Object	Pass	Fail	?	N/A	n
Final-response grounding	Last user-facing response	80	24	2	0	106
State-claim support	State/current-fact assertions	74	6	3	23	83
Pre-call authorization	Modifying calls for the task	62	29	1	14	92

No state claim and no modifying call are inapplicable cases, not successful judgments or missing returns. A task fails authorization if any modifying call fails. If none fails but at least one is uncertain, the task remains uncertain. Among the 166 modifying calls nested within 92 tasks, 124 pass authorization, 41 fail, and one is uncertain. Separately, 142 calls are accepted by the tool and 24 rejected. Calls are not treated as independent tasks.

Six uncertain task-level metric cells remain across four tasks. Counting all uncertain values as failures or passes gives deterministic applicable-denominator ranges of 75.5–77.4% for response grounding, 89.2–92.8% for state-claim support, and 67.4–68.5% for authorization. These are classification-sensitivity ranges, not confidence intervals. Original claim segmentation also remains separate: the two reviewers marked 140 and 136 claims, respectively, so a pooled claim-level percentage would have no common denominator.

O.3 Native task outcomes and final-response grounding

The historical evaluator uses two alternative reward branches. For 72 tasks it compares the final database hash with the reference state. For 34 tasks with designated textual outputs, it checks whether the specified strings occur in any assistant response, ignoring case and removing commas from the response. The branches are not joint checks, and neither is a semantic audit of the last response. We therefore retain the name *native task outcome*, rather than identifying it with the task-satisfaction measure in our controlled study.

Table 15: Native outcomes and grounding by mutually exclusive reward branch. Question marks are unresolved final-response judgments.

Branch	Outcome	Grounded	Ungrounded	?	n
Database hash	Success	29	9	1	39
Database hash	Failure	29	4	0	33
Designated text	Success	16	6	1	23
Designated text	Failure	6	5	0	11
All tasks	Success	45	15	2	62
All tasks	Failure	35	9	0	44

Thus, 15/62 native successes have an ungrounded final response, whereas 35/44 native failures have a grounded one. Both discrepancies occur in both branches: the first is 9/39 for database-hash tasks and 6/23 for output-text tasks; the second is 29/33 and 6/11, respectively. The grounding-uncertain tasks are retained in their denominators, with one native success in each branch. This post-hoc stratification addresses whether the pooled result is confined to one scoring mechanism. Different tasks enter the two branches, so their rate difference is not an effect of the evaluator.

The pooled proportions are 24.2% and 79.5%, with descriptive task-level Wilson 95% intervals of 15.2–36.2% and 65.5–88.8%. These intervals do not capture template dependence, annotation uncertainty, or selection of a fixed archival cohort, and are not estimates of deployment prevalence. Nor are the 15 grounding failures 15 fabricated-completion cases: only three tasks among the native successes have a definite state-support failure. Other grounding problems include unsupported future arrangements, timing guarantees, or scope. Each failure must be interpreted from its corresponding statement rather than inferred from the cross-tabulation alone.

O.4 Review provenance and limits

According to the authors' confirmation, both reviewers independently formed the initial judgments and then used AI to organize their returns. Their prior exposure to native rewards, model identity, or the other reviewer's opinions was not established. The issued review materials withheld native rewards, model identity, reference actions, and hidden task information. Withholding those fields does not establish every reviewer's actual exposure history.

A later 12-task human adjudication addressed the identified boundaries. Its packet contained original judgments, AI suggestions, and source excerpts, and a companion analysis disclosed aggregate native outcomes. It was therefore not a fresh blind review. Some returned explanations relied on the packet's descriptions of messages not reproduced in the excerpts. An AI-assisted audit subsequently checked those claims against the original trajectories, but cannot substitute that inspection for a human reading history. The matched adjudications comprise 36 task-level labels, nine call-level labels, and seven alignable claim labels. Fourteen labels change relative to the separate original returns: five from one reviewer and nine from the other, not 14 new tasks. Original returns, the differently segmented claim lists, and all residual uncertainty are preserved.

The subsequent same-rule audit covered all 106 final responses, 166 modifying calls, and the original claim annotations. It identified no additional supported score-change candidate outside the 12-task set. This was an AI integrity check, not another independent human reannotation or proof of zero residual error. Post-adjudication agreement must not be reported as initial inter-reviewer reliability.

This external analysis supports the applicability of the measurement distinction to a different, public tool-interaction corpus. It does not manipulate memory, establish a causal source of the observed errors, or validate a new model ranking. The 106 archived tasks remain separate from the 216 constructed materials and their 1,899 responses. We do not pool accuracy over these different designs.

O.5 Reproduction records

The internal adjudicated task table, statistics, label-change log, and branch analysis link the reported counts to retained returns. Local verification reconstructs these results from the original returns and matched human adjudications, joins blind identifiers to the fixed trial-0 tasks, and checks the evaluator branch against replay. These files are not linked as public PDF artifacts; anonymous distribution and licensing checks remain separate from local verification.