Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

AI Agent Memory Deletion Verification: Counterfactual Tests

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-15 · ai agent memory deletion verification · counterfactual evaluation · agent memory · machine unlearning · data provenance · retrieval augmented generation · semantic disclosure · context engineering · retained utility

Original article: https://reliablecontext.com/articles/ai-agent-memory-deletion-counterfactual-test



In this article

1. Executive Summary
2. Introduction and Background
3. What deletion verification measures
4. Specify the target and observable boundary
5. Build matched counterfactual episodes
6. Probe disclosure and decisions
7. Run an auditable removal evaluation
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Artificial intelligence (AI) agent memory deletion verification should test whether a withdrawn fact still changes observable behavior. The strongest practical comparison is a matched agent that **never received the target information**. A successful delete operation establishes a storage-layer result within its documented scope; it does not, by itself, establish the same result for derived memory or execution state. Amazon documents that deleting a short-term event does not remove information derived into long-term memory, while LangGraph documents checkpoint deletion separately from cross-conversation memory. (docs.aws.amazon.com) (docs.langchain.com)

The proposed **counterfactual evidence test** separates exact-string recovery, semantic disclosure, target-dependent decisions and retained-task utility. It observes permitted tool arguments, retrieval results, logs and later summaries alongside final answers. Yu and colleagues' **K-Bench**, current version **v2**, observes six channels, including reasoning text where the harness actually exposes it. Its findings motivate multi-channel evaluation, without making hidden reasoning a prerequisite for testing a closed service. (arxiv.org) Yao and colleagues' **Forgetting Without Restarting** defines an agent that never observed the target as the counterfactual, with an explicit boundary around reconstructible runtime state. (arxiv.org)

The evaluation uses matched never-exposed, exposed-unchanged, exposed-removed and full-reset episodes, supplemented by sham removal and unrelated retained facts. These are proposed experimental controls, grounded in paired-observation methods rather than a published deployment guarantee. (www.itl.nist.gov) Report **union leakage**, the proportion of episodes disclosing the target in any available channel, and a separate **behavioral residual**, the difference in target-dependent action rates between removed and never-exposed episodes. Prespecify an equivalence margin if the intended conclusion is practical equivalence; nonsignificance alone is insufficient. (pmc.ncbi.nlm.nih.gov) (docs.scipy.org)

Zero observed recoveries support a distribution-specific uncertainty statement, not universal forgetting. As a **hypothetical design calculation (Hypothetical Example)**, zero events across **300 independent trials** yield an exact one-sided **95%** upper bound of approximately **0.994%**, under a shared event probability. (jhanley.biostat.mcgill.ca) Acceptance therefore requires an explicit observer, adequate positive controls, useful retained-task performance and complete available evidence. MUSE (Machine Unlearning Six-Way

Evaluation) treats retained utility as a separate criterion, and work on contextual utility distinguishes removal from the ability to use information supplied again through an allowed prompt. (proceedings.iclr.cc) (proceedings.mlr.press) The defensible outcome is limited evidence of non-use under the tested conditions, failure of that claim, or an unassessable result.

Introduction and Background

A long-running agent can encounter a fact, summarize it, retrieve it later and use it to select a tool action. Evaluating withdrawal therefore requires a broader unit of analysis than a memory row. MemGPT's research design explicitly moves information between storage tiers, and LangGraph distinguishes information retained across conversations from thread-level checkpoints. These are concrete reasons to inventory the agent's accessible state before deciding what an intervention actually removes. (research.memgpt.ai) (docs.langchain.com)

This report examines how to verify AI agent memory deletion when an enterprise withdraws a permitted source or stored fact from future use. For Reliable Context's audience, the decision is an evidence question: what can an evaluator demonstrate about subsequent disclosure and action? The protocol below evaluates the agent under test. It makes no product-specific claim about the publisher's forgetting capabilities.

Source withdrawal, **fact correction** and **historical reconstruction** have different acceptance targets. Withdrawal asks whether revoked information continues to influence permitted work. Correction asks whether the agent adopts an authorized replacement. Historical reconstruction asks whether a past record can be reproduced. The original LongMemEval benchmark includes knowledge updates and temporal reasoning among its memory abilities; those categories do not establish counterfactual non-use after withdrawal. (github.com)

The recent preprint literature sharpens this distinction. Wang and colleagues' **Agentic Unlearning** addresses parameter and persistent-memory pathways, including dependency-derived artifacts, whereas Yao and colleagues formulate a runtime comparison with a never-exposed agent. Their objects of intervention differ, so their results should not be merged into a single forgetting score. (arxiv.org) This article proposes an acceptance protocol informed by those distinctions, rather than a tutorial for editing model weights.

The practical question is narrow enough to evaluate: after a specified intervention, can a specified observer recover the target, or detect target-dependent choices beyond a prespecified tolerance? Its answer remains conditional on the probe distribution, available observations and experimental controls. K-Bench likewise states an observer boundary: reconstruction by combining fragments across channels lies outside its described observer model. (arxiv.org)

What deletion verification measures

Information removal needs an operational definition. The protocol distinguishes actions that change access, actions that change stored state and measurements of subsequent behavior. Official documentation illustrates why these cannot be substituted for one another: disabling classic Amazon Bedrock agent memory disassociates memory from the agent, while Azure AI Search describes physical deletion as a background process. (docs.aws.amazon.com) (learn.microsoft.com)

The following terms define the proposed acceptance claim:

- **Removal target:** The particular source contribution, fact, preference or association withdrawn from use. Record whether the request concerns a value, one source supporting that value, or every derivation of that source. The World Wide Web Consortium (W3C) PROV data model provides a source-to-derived-entity relationship for expressing the lineage. (www.w3.org)
- **Access blocking:** The target cannot be obtained through a declared interface by the tested principal. This is an interface property. Pinecone's metadata-filter deletion, for example, is scoped to matching records in a specified namespace. (docs.pinecone.io)
- **Storage deletion:** A declared object or representation is removed under the storage system's semantics. Qdrant explicitly distinguishes deleting selected vectors from deleting the enclosing points. (qdrant.tech)
- **Answer suppression:** A final response omits the target. This leaves tool activity and intermediate state to inspect; ToolSandbox records per-turn state snapshots. (github.com)
- **Residual influence:** The target still changes observable decisions after intervention, even when no target string appears. Yao and colleagues supplement disclosure measurements with a string-free behavioral test. (arxiv.org)
- **Observer:** The evaluator's allowed channels, probe policy and budget. For this protocol, the conclusion covers only that observer; stronger access requires a new evaluation.
- **Counterfactual non-exposure:** A comparison condition whose trajectory never incorporated the target. Pairing corresponding episodes supports analysis of differences rather than comparisons between unrelated tasks. (www.itl.nist.gov)

The acceptance target must also separate **source exclusion** from **value suppression**. If an allowed independent source still supports the same value, continued use of that value need not indicate use of the withdrawn source. Provenance can represent distinct derivations, but a derivation record is not itself evidence about the model's internal causal pathway. (www.w3.org) The proposed test must therefore specify whether correlated and independently inferable information remains permitted.

This distinction matters for literal recovery. A randomly assigned synthetic secret can have negligible prior exposure, whereas a public fact may be reproducible without the revoked source. Use the former for an isolated non-exposure experiment, and treat the latter as a provenance-sensitive task. Wang and colleagues explicitly identify limits in tracking cross-agent information flows, reinforcing the need to declare the scope of dependency coverage. (arxiv.org)

An administrative deletion receipt belongs in the evidence packet. It should be interpreted alongside retrieval checks, available runtime observations and behavioral comparisons. Microsoft's documented key lookup expecting **Hypertext Transfer Protocol (HTTP) 404** is useful evidence of document absence through that interface; it does not measure an agent's subsequent decisions. (learn.microsoft.com)

Specify the target and observable boundary

Begin with a **synthetic target** that is absent from the never-exposed fixture and unrelated to public knowledge. A nonce, a randomly assigned identifier, identifies literal disclosure; an arbitrary preference creates a target-dependent choice without requiring the nonce in the answer. This is a proposed experimental construction. Its purpose is to isolate influence from background inference; correlated facts remain a separately declared test condition.

Before exposure, freeze the permissible task, unrelated facts and scoring rules. Define what can be inspected without assuming that a closed model exposes hidden reasoning. Enumerate the channels the deployment actually makes available. Stateful tool evaluation provides a useful precedent for recording intermediate state alongside dialogue. (github.com)

Record the following scope decisions before running episodes:

- **Target identity:** Preserve the exact synthetic payload and its semantic interpretation in restricted fixtures. Keep it outside the never-exposed agent's prompts, tools and grading-visible context.
- **Source identity:** Record source and version identifiers, ingestion time and dependent artifacts known to the evaluator. A derivation relationship supplies a useful vocabulary for that record. (www.w3.org)
- **Intervention time:** Identify the boundary after which non-use is required, including the period while asynchronous deletion may still be pending. AgentCore documents that expired records can remain readable before background deletion completes. (docs.aws.amazon.com)
- **Observable channels:** Enumerate final responses, retrieval, tool inputs and outputs, subsequent memory writes, and accessible summaries. Mark unavailable reasoning or provider caches explicitly as unavailable. Thread checkpoint inspection must be declared separately. (docs.langchain.com)
- **Correlated information:** Decide whether retained sources may imply the target. If they may, distinguish authorized re-derivation from prohibited dependence on the withdrawn observation.
- **External effects:** Declare which tools can be reset or simulated. Treat non-restorable committed actions as outside any claim that the full environment has been reset.

Table 1 specifies a **removal ledger** for each target. It is a proposed audit schema, not a record of an executed benchmark; its source and derivation fields follow the provenance distinction above. (www.w3.org)

Ledger field	Evidence to retain	Interpretive purpose
Target	Synthetic payload, meaning and protected fixture identifier	Separates exact recovery from semantic use
Source/version	Original source identifier, revision and independent supporting sources	Distinguishes withdrawn evidence from a still-permitted value
Observation time	The episode event at which the target became available	Locates exposure in the trajectory
Removal time	Request time, completion signal and first verified absence	Separates administrative completion from observed exclusion
Declared scope	Objects, derivatives, sessions and external systems covered	Prevents scope expansion after testing

Ledger field	Evidence to retain	Interpretive purpose
Observable channel	Available interfaces and their capture permissions	Defines the observer rather than presuming complete access
Expected behavior	Prespecified disclosure and behavioral criteria	Makes acceptance independent of favorable outcomes
Actual evidence	Versioned observations, task outcomes, errors and exclusions	Supports later reproduction and qualification

The ledger links a claim to observations without turning unavailable layers into empty cells. An evaluator should use explicit states such as **not observed**, **blocked** or **outside scope**, with reasons. Storage documentation makes the need concrete: deleting an AgentCore event leaves derived long-term information, and LangGraph exposes a separate thread-checkpoint deletion operation. (docs.aws.amazon.com) (docs.langchain.com)

Keep versioned model, instruction, retrieval, tool, source and sampling manifests. Preserve complete available trajectories rather than selected passages. ToolSandbox demonstrates the usefulness of per-turn state snapshots in stateful tool evaluation. (github.com)

Build matched counterfactual episodes

The experimental unit is an **episode**, a task trajectory with a declared initial state and observation window. Match episodes on everything intended to remain constant: task, unrelated evidence, tool environment and sampling plan. NIST describes paired data through one-to-one correspondence; applying that principle here makes the removal-versus-never-exposed contrast explicit. (www.itl.nist.gov)

Use the following conditions, with labels fixed before execution:

- **Never exposed:** The agent receives all permissible information except the target. Its initial state and accessible external stores must also exclude the target.
- **Exposed unchanged:** The agent receives and retains the target. This positive control tests whether the task and probes can elicit disclosure or target-dependent behavior.
- **Exposed then removed:** The same target is introduced, followed by the candidate removal intervention at the declared boundary.
- **Full-reset control:** Restart from a target-free fixture and replay permitted history where reconstruction is feasible. Declare any external effects that this comparator cannot reverse.
- **Sham removal:** Execute the surrounding administrative procedure without removing the target. This tests whether changed instructions, latency or session handling explain a behavioral shift.
- **Retained facts:** Include unrelated questions and actions with valid supporting evidence. MUSE's retained-utility criterion motivates evaluating this dimension separately from forgetting. (proceedings.iclr.cc)

A positive control must make the target relevant. If the exposed-unchanged agent cannot disclose or act on it, a low post-removal rate says little about intervention effectiveness. Also compare the never-exposed and reset conditions. A discrepancy may reveal a mismatched fixture, a replay artifact or a target-independent effect of restarting.

For a preference target, randomly assign which otherwise equivalent alternative is favored, and balance that assignment across fixtures. Keep action labels and target direction counterbalanced, so an agent's default preference does not masquerade as residual influence. These are proposed design controls; the analysis should still preserve episode pairing rather than pooling unmatched observations. (www.itl.nist.gov) (docs.scipy.org)

Replay requires particular care. Hold exogenous information constant where the counterfactual justifies doing so, but recompute tool responses when changed actions alter the relevant world state. Forcing an exposed-world response into a different counterfactual action sequence can create an incoherent control. ToolSandbox explicitly models successful tool execution as dependent on world state. (github.com) For this protocol, non-reconstructible observations are a limit on the counterfactual comparison.

Use matched random seeds where available, record unavailable seed control, and repeat across seeds. Separate episode replication from additional probes inside an episode. Multiple questions about the same contaminated trajectory are related observations. Hierarchical bootstrap work illustrates why nested measurements should not automatically be treated as independent samples; select an episode, user or session cluster according to the actual dependence structure. (pmc.ncbi.nlm.nih.gov)

Finally, define the observation horizon before seeing outcomes. A later summary can reopen an information pathway after an apparently clean answer. Classic Bedrock agents generate stored summaries at session end, which supplies a concrete reason to include post-session artifacts when those are within the deployment's scope. (docs.aws.amazon.com)

Probe disclosure and decisions

A useful **observer threat model** states access, adaptation and budget together. Specify fixed or adaptive probes, cross-channel combination and credentialed access. Record probe count, token budget, adaptive rounds and stopping rule. K-Bench's stated exclusion of cross-channel fragment reconstruction illustrates why the observer definition is part of the result. (arxiv.org)

The proposed probe suite has distinct targets:

- **Literal recovery:** Search permitted outputs for the exact nonce and declared normalizations. Record transformations such as whitespace removal before scoring, rather than inventing them after a suspicious response.
- **Semantic disclosure:** Test paraphrases, reversed relations and simple implications. An ACL evaluation expands unlearning tests beyond originals to rephrased, subject-replaced, relation-reversed and one-hop reasoned variants. (aclanthology.org)
- **Structured disclosure:** Inspect available tool arguments and returned fields, including identifiers that encode the target even when final prose omits it. (arxiv.org)

- **Delayed disclosure:** Inspect later summaries and memory write-backs after unrelated intervening work. The runtime paper explicitly includes memory write-backs in its observable scoring. (arxiv.org)
- **String-free influence:** Present choices whose correct permissible resolution is independent of the revoked preference, and score whether target assignment still shifts action selection.
- **Retained usefulness:** Score unrelated successful tasks, justified abstentions, unsupported refusals and tool errors separately. Retained utility is a distinct MUSE criterion. (proceedings.iclr.cc)

Do not put the target into a recall question if the test is supposed to measure residual memory. A question that supplies the answer changes exposure. If deliberate reintroduction is part of a separate task, label it accordingly: work on **contextual utility** asks whether removed information can still be used when a user supplies it anew. (proceedings.mlr.press)

Table 2 maps interventions to observations. Each cell is an evaluation obligation or declared limit, not an assertion that a particular product implements that intervention. The distinction between removing selected vectors and removing enclosing records is documented by Qdrant. (qdrant.tech)

Channel or state	Block source access	Delete stored target	Reconstruct/reset runtime
Final response	Probe continued disclosure	Probe continued disclosure	Compare with never-exposed response distribution
Retrieval results	Verify target cannot be served to the tested reader	Verify relevant records and derivatives are excluded	Recheck external retrieval separately
Tool arguments and observations	Inspect target-bearing fields	Inspect residual fields	Compare paired action trajectories
Long-term memory	Check permitted read interfaces	Inspect declared dependency coverage	Verify reconstructed state does not reload the target
Thread checkpoints and prompt state	Access blocking alone does not establish removal	Inventory state outside the deleted object	Compare replayed permissible history
Later summaries and write-backs	Inspect newly generated artifacts	Inspect surviving and regenerated artifacts	Include post-session observation window
Provider-internal reasoning and caches	**Unavailable unless explicitly exposed**	**Unavailable unless explicitly exposed**	**Unavailable unless explicitly exposed**
Committed external effects	Identify what has already happened	State whether reversal is in scope	**Unassessable as reset if effects cannot be restored**

The unavailable cells constrain the conclusion. They do not make observable testing worthless, but they prevent a claim covering those layers. In particular, low final-answer recovery does not establish low union leakage. K-Bench's current abstract reports that structured-retrieval answer suppression can leave aggregate leakage unchanged. (arxiv.org)

Use a versioned semantic rubric and retain original available outputs. Adjudicate ambiguous cases, report confirmed recovery and uncertain similarity separately, and keep protected grading fixtures out of later prompts. These are proposed scoring safeguards; they carry no published accuracy guarantee.

Run an auditable removal evaluation

Deletion in retrieval-augmented generation (RAG) includes at least a source lifecycle and a retrieval lifecycle. Azure documents that indexers do not automatically track object deletion and that adding deletion detection later does not retroactively remove earlier orphaned documents. Removing a source file therefore requires an independently verified index-side outcome in configurations with those semantics. (learn.microsoft.com)

Use the storage interface's documented completion behavior, then perform read and retrieval checks appropriate to the declared scope. Qdrant exposes a wait option for changes to occur. Weaviate documents background cleanup and limits for a single deletion query. These are implementation-specific reasons to distinguish request acceptance, visible absence and physical cleanup; none alone evaluates residual agent behavior. (api.qdrant.tech) (docs.weaviate.io)

A reviewable **removal packet** should contain:

- **Versioned fixtures:** Target-free and exposed initial states, source versions, probe definitions and task constraints.
- **Intervention receipts:** Requested scope, affected identifiers, completion signals, errors and independently observed absence checks. Microsoft's documented document-key lookup is one such check. (learn.microsoft.com)
- **Provenance coverage:** Source-to-derived-artifact links and known gaps, using explicit derivation relationships where appropriate. (www.w3.org)
- **Complete available observations:** Tool activity, retrieval, task outputs, memory writes and permitted later summaries, with sensitive fixture access controlled. (github.com)
- **Scoring configuration:** Literal normalization, semantic adjudication, action labels, utility criteria and uncertainty method, including the unit resampled. (docs.scipy.org)
- **Exclusions and reintroduction paths:** Unavailable provider internals, non-restorable external effects and allowed sources that may supply the target again. (proceedings.mlr.press)

Do not let the evidence packet itself become an accidental reintroduction path. Store protected target fixtures and detailed observer transcripts outside the tested agent's retrieval permissions. The report can retain hashes and restricted artifact references while authorized evaluators retain the evidence needed to reproduce scoring. This is an evaluation-design recommendation, not a claim that hashes prove deletion.

Acceptance is a decision over the declared claim, not an automatic interpretation of an empty result:

- **Fails:** Confirmed prohibited disclosure occurs, or the behavioral residual falls outside the prespecified acceptable region.
- **Limited evidence of non-use:** Positive controls are sensitive, observation coverage is adequate for the claim, uncertainty meets the declared criterion, and retained utility remains acceptable.
- **Unassessable:** Necessary channels are unavailable, controls are insensitive, counterfactual state cannot be reconstructed, or uncertainty is too wide for the acceptance claim.

Retain these labels across releases so that a weaker observer cannot create an apparent improvement. The same caution applies across repeated removal requests: MUSE treats sustainability over sequential requests as a separate desirable property. (proceedings.iclr.cc) Regression episodes should therefore test previously removed targets alongside newly withdrawn ones.

Data Analysis and Evidence

The quantitative report must disclose denominators, clustering, probe budget and exclusions. **Union leakage** counts an episode once when any observed channel recovers the target. Adding channel percentages double-counts overlapping episodes; channel-specific counts remain useful diagnostics alongside the episode union. This union definition makes the episode the unit of counting, while retaining channel diagnostics.

Table 3 defines the proposed scorecard. The metrics are an evaluation specification, not measured results from an executed agent.

Metric	Numerator or contrast	Denominator and interpretation
Union leakage	Episodes with confirmed target recovery in any observed channel	All tested episodes; disclose channel coverage
Semantic disclosure	Episodes with rubric-confirmed equivalent information	All episodes assessed semantically; report ambiguous cases separately
Behavioral residual	Target-dependent action rate after removal minus the never-exposed rate	Matched eligible episodes; report paired uncertainty
Retained utility	Successful unrelated tasks	All unrelated tasks tested, including refusals and errors
Observer coverage	Available channels actually captured	Declared channels; unavailable channels remain explicit

An episode can pass literal recovery yet fail semantic disclosure or behavioral influence. Conversely, a never-exposed agent may independently choose the target-associated alternative. Subtracting its rate is therefore necessary for the proposed behavioral contrast. Preserve the correspondence when resampling: SciPy's paired bootstrap uses common sampled indices across arrays. (docs.scipy.org) If episodes share users or source histories, resample at the relevant cluster level; preserve task strata when estimating an aggregate across heterogeneous tasks. (pmc.ncbi.nlm.nih.gov) (proceedings.neurips.cc)

For a practical equivalence claim, define a margin, **delta**, as the smallest action-rate difference that would change the deployment decision. Require the appropriate uncertainty interval to lie inside the prespecified equivalence region. Lakens describes the conventional two one-sided tests procedure, where a significance level, alpha, of **0.05** corresponds to a **90% confidence interval** inside the equivalence bounds. Do not substitute an arbitrary bootstrap interval without establishing its properties for the intended test. (pmc.ncbi.nlm.nih.gov)

For zero leakage, the **hypothetical design calculation (Hypothetical Example)** is:

`upper bound = 1 - 0.05^(1/n)` (jhanley.biostat.mcgill.ca)

With **n = 300** independent, identically distributed episodes and no observed events, the one-sided **95%** upper bound is approximately **0.994%**. The familiar approximation is **3/n**. These are calculations from the binomial zero-event method, not a benchmark outcome. (jhanley.biostat.mcgill.ca) NIST cautions against relying on normal approximations with small failure counts. (www.itl.nist.gov) Repeated dependent probes do not create independent trials. (pmc.ncbi.nlm.nih.gov)

Keep the sample size fixed, or choose a valid sequential method before monitoring results. Fixed-time inference does not generally retain its guarantees when stopping depends on observed outcomes. (jmlr.org)

The literature supplies motivation, not transferable acceptance thresholds. In **K-Bench v2**, dated **14 September 2026**, the authors report **22% to 86%** leakage for prompt/retrieval secrets in settings where answer-only evaluations report none; its observer and substrate choices define that result's scope. (arxiv.org) **Forgetting Without Restarting** also separates memory deletion from residual runtime behavior, while **Agentic Unlearning** uses task-specific forget, retained-performance and membership-inference metrics. (arxiv.org) Those measurements are different estimands. No reported percentage here establishes the performance of an enterprise deployment.

Implications and Future Directions

The next useful advance is a **portable, observer-specific removal packet** with paired behavioral controls. Its value is a hypothesis: teams could compare evidence across releases while retaining the exact boundary of each claim. A provenance vocabulary helps connect sources and derived artifacts, and a versioned task environment helps explain changes in behavior; neither replaces empirical observation. (www.w3.org) (github.com)

Research artifacts should be distinguished from independent replication. As checked for this report, the author-linked **K-Bench repository** documents executable evaluation and scoring commands, including a local scoring smoke test that does not execute a model. Its code is **MIT-licensed**, the synthetic corpus uses **Creative Commons Attribution (CC BY) 4.0**, and the target adapter has a separate **Llama 3.1 Community License**. These terms apply to different artifacts. Repository availability is not evidence that this report reproduced the results. (github.com)

No own executable repository was established from the fetched paper pages for **Forgetting Without Restarting** or **Agentic Unlearning**. That is a bounded research finding, not a claim that no repository exists. The former's stated limitations concern model parameters, committed effects and correlated information; the latter acknowledges incomplete cross-agent dependency coverage. Both limitations matter when transferring a laboratory protocol to a persistent enterprise environment. (arxiv.org)

The memory-evaluation landscape also needs separate scorecards for retention and removal. LongMemEval's original abstention cases concern nonexistent events, which differs from suppressing influence from a previously encountered target. (github.com) MemGPT's multi-tier memory design makes it useful background for state inventories, without supplying a forgetting certificate. (research.memgpt.ai) Future comparisons should state which objective they test rather than treating temporal accuracy, absence of evidence and counterfactual non-use as interchangeable.

Finally, plan for **authorized reintroduction**. A later permitted source may contain the same fact, or a user may explicitly supply it for a new task. Contextual-utility research treats the ability to use newly supplied information as a distinct evaluation concern. (proceedings.mlr.press) The removal claim must identify its time interval and evidence basis; a permanent ban on a value is a different policy from withdrawing one source.

Conclusion

AI agent memory deletion verification is an experiment about observable information use. A useful test starts with a target that the never-exposed control cannot obtain, introduces it into matched episodes, applies a declared intervention and compares subsequent disclosure and action. The never-exposed condition supplies the baseline for interpreting residual behavior. A clean final answer supplies only one observation.

The decision should keep disclosure, target-dependent choices and retained usefulness separate. Record which storage layers and channels were inspected, how asynchronous operations were handled, which external effects could be restored and which permitted sources might reintroduce the target. Preserve complete available evidence and the configuration needed to reproduce scoring.

Acceptance language should remain proportional to those observations. Confirmed prohibited disclosure or an unacceptable behavioral residual fails the claim. Adequate controls and uncertainty can support limited evidence of non-use under the tested observer. Missing essential access, incoherent counterfactuals or insufficient precision make the result unassessable. None of these outcomes is a universal forgetting statement or a legal erasure certificate.

The practical deliverable is a versioned removal packet whose conclusions another evaluator can inspect and challenge. It states what was withdrawn, what comparison was constructed, what was observed and how much uncertainty remains. That makes an evidence-removal decision reviewable without confusing a storage receipt, an answer filter or a successful unrelated task with removal of the target's influence.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/short-term-delete-event.html
2. https://docs.langchain.com/oss/python/langgraph/add-memory
3. https://arxiv.org/abs/2609.12808
4. https://arxiv.org/html/2609.04875v1
5. https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm
6. https://pmc.ncbi.nlm.nih.gov/articles/PMC5502906
7. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
8. https://jhanley.biostat.mcgill.ca/c607/ch08/zero_numerator.pdf
9. https://proceedings.iclr.cc/paper_files/paper/2025/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html
10. https://proceedings.mlr.press/v306/peng26a.html
11. https://research.memgpt.ai

12. https://github.com/xiaowu0162/LongMemEval
13. https://arxiv.org/html/2602.17692v2
14. https://arxiv.org/html/2609.12808v2
15. https://docs.aws.amazon.com/bedrock/latest/userguide/agents-memory-disable.html
16. https://learn.microsoft.com/en-us/azure/search/search-how-to-delete-documents
17. https://www.w3.org/TR/prov-dm
18. https://docs.pinecone.io/guides/manage-data/delete-data
19. https://qdrant.tech/documentation/manage-data/points
20. https://github.com/apple-aiml-research/ToolSandbox
21. https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/long-term-record-ttl.html
22. https://pmc.ncbi.nlm.nih.gov/articles/PMC7906290
23. https://docs.aws.amazon.com/bedrock/latest/userguide/agents-memory-view.html
24. https://aclanthology.org/2026.acl-long.88
25. https://learn.microsoft.com/en-us/azure/search/search-how-to-index-azure-blob-changed-deleted
26. https://api.qdrant.tech/api-reference/points/delete-points
27. https://docs.weaviate.io/deploy/faqs
28. https://proceedings.neurips.cc/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf
29. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/exacbici.htm
30. https://jmlr.org/papers/volume24/23-0401/23-0401.pdf
31. https://github.com/OniReimu/kbench

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.