Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Agent Memory Admission Control: Can It Trust Its Own Guess?

> 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 · 2026-10-31 · agent memory admission control · ai agent memory admission policy · agent memory evaluation · llm agent memory poisoning · agent long-term memory provenance · false memory resistance · halumem · agent memory write validation · system of record for ai agent memory

Original article: https://reliablecontext.com/articles/agent-memory-admission-test-own-suggestions-as-facts



In this article

1. Executive Summary
2. Introduction and Background
3. Defining Admissible Memory Under a Declared Evidence Policy
4. What HaluMem Measures at the Admission Stage, and Its Limits
5. Matched Dialogue Fixtures
6. Measurement Protocol
7. The Later-Session Probe
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

Agent memory admission control is the decision, made at the moment of writing, about which statements become durable entries that later sessions treat as context. This report proposes an observable **admission test** for that decision and anchors it in the closest published evaluation, **HaluMem** (Chen et al.), the first operation-level benchmark for evaluating hallucinations in memory systems (aclanthology.org). Its false memory resistance (**FMR**) metric measures resistance to distracting content that the AI mentions but the user does not confirm (arxiv.org), and its data include **2,648** such distractor memories (arxiv.org).

The paper-reported results show why a resistance score cannot be read alone. Across the five configurations with extraction scores, FMR ranged from **44.94% to 80.78%** on HaluMem-Medium and from **28.85% to 88.36%** on HaluMem-Long (arxiv.org). In the same table the configuration with the lowest extraction recall has the highest FMR in both splits, so a writer that stores little ignores distractors by default. An admission test therefore has to report false admission and valid-admission recall together, and must not declare safety from an empty write set.

The proposed protocol uses six matched dialogue fixtures (assistant-only suggestion, explicit confirmation, explicit rejection, neutral acknowledgment, discussion of another subject, and a cited tool observation) in which only the evidence changes. Each fixture starts from a clean memory state, and the first write, the later retrieval and the final answer are scored against separate denominators. A correct later answer cannot show that a stored entry was valid when admitted, a limit that end-to-end recall benchmarks share, because they score answers rather than writes (aclanthology.org). The speaker role alone does not settle truth: an assistant can relay a verified tool observation, and a user can be wrong. The evidence policy therefore keys on source identity, not on speaker.

Statistical handling follows from the design. Conversations are clusters, and clustered standard errors can be over 3X larger than naive ones (arxiv.org). Paired outcomes under suggestion versus confirmation call for an exact test on discordant pairs (www.statsmodels.org), and a zero-event result still carries an upper bound of

about 3/n (pmc.ncbi.nlm.nih.gov). Tolerable false-admission and retained-usefulness targets must be fixed before results are seen, a practice formalized in clinical-statistics guidance on pre-specified margins (database.ich.org).

No new empirical rates were measured for this report. HaluMem uses virtual personas and generated dialogues, and the evaluation depends on a language-model judge. The numbers cited here are paper-reported, not an independent ranking of products, and they do not establish enterprise prevalence.

Introduction and Background

Persistent memory turns a conversation into state. Once an entry is written, a later session may retrieve it and the assistant may present it as established knowledge, without the original dialogue in view. The decision this report supports is narrow: should a context engineer enabling persistent memory for an organizational agent trust that **new entries retain a defensible basis**? Three sub-questions follow. Does each entry keep a source and a date? Are unsupported assistant suggestions promoted into factual context? Does restricting writes lose useful confirmed information?

The question is concrete because memory libraries document that assistant text is an input. Mem0's documentation describes `add` as taking an ordered list of user and assistant turns (docs.mem0.ai), and its guidance lists a moment when "a decision or suggestion is made" as a time to add memory (docs.mem0.ai). A vendor post dated April 2026 describes the current algorithm as one that stores agent-generated facts such as confirming an action or providing a recommendation (mem0.ai). Agent frameworks have long stored the agent's own text: Reflexion agents maintain their own reflective text in an episodic memory buffer (proceedings.neurips.cc). These are documented design choices, not defects. Whether a given deployment admits an unsupported suggestion as a fact is exactly what an observable test should establish.

Security literature frames a related but different problem. OWASP's guide on agentic threats lists memory poisoning as the exploitation of an AI's short- and long-term memory to introduce malicious or false data (genai.owasp.org), and separately defines cascading hallucinations as an agent's inaccurate output reinforced through its memory (genai.owasp.org). The MINJA attack assumes an attacker who behaves like a regular user, with no direct access to any part of the agent (arxiv.org). The test proposed here concerns the benign counterpart: no adversary, only an ordinary dialogue in which the assistant says something and the user does not confirm it.

This report is organized as follows. The next section defines admissible memory under a declared evidence policy. The following sections examine what HaluMem measures at the admission stage, specify matched fixtures, set out the measurement protocol, and treat the later-session probe. A data section collects paper-reported and vendor-presented figures with their scope, and an implications section addresses the choice between a memory library and a system of record. Reliable Context, the publisher of this report, is a direct provider of a **system of record** (one canonical record per real thing, with a verbatim cited source and an observation date behind every value). It is not a memory library, vector database or RAG framework, and the section on implications says when a reader needs one of those instead. No evaluation was executed for this report, and no new rates are reported.

Defining Admissible Memory Under a Declared Evidence Policy

Source classes, not speaker roles

An evidence policy states, before any result is seen, which kinds of statements may become factual entries. The W3C PROV data model supplies vocabulary that fits the problem. It defines provenance as information about the entities, activities and people involved in producing a piece of data (www.w3.org), notes that this information can be used to form assessments about quality, reliability or trustworthiness (www.w3.org), defines an agent as something that bears responsibility for an activity (www.w3.org), and defines attribution as the ascribing of an entity to an agent (www.w3.org). A stored memory is an entity, the write is an activity, and the user, the assistant and a tool are distinct agents. A derivation is a transformation of one entity into another (www.w3.org), which is what an extracted memory is relative to a dialogue turn.

The policy distinguishes seven classes, each recorded with source identity and date:

- **User assertion:** a statement by the user about the user. Admissible as an assertion, with the user as the attributed agent. It is not external truth, and a user can be wrong.
- **Cited tool observation:** a result a tool returned in the transcript, with locator and date. Admissible even when the assistant relays it, because the basis is the tool output, not the speaker.
- **Quoted third-party statement:** admissible as "X said Y on date D", not as Y.
- **Inference:** a conclusion derived from other entries. Admissible only when labeled as inferred and linked to its premises.
- **Assistant suggestion:** a proposal or guess by the assistant. Not admissible as a fact. Admissible as a record that a suggestion was made, if the policy wants that.
- **Hypothetical:** a conditional or counterfactual statement. Not admissible as a fact.
- **Unresolved:** a case the adjudicators cannot classify, kept as unresolved with the reason.

Two cautions keep the classes honest. First, faithfully storing a source's assertion does not prove the source true. The Knowledge-Based Trust work, which estimated trustworthiness for 119 million webpages (www.vldb.org), separates extraction errors from errors in the source itself, and reports that extraction errors are far more prevalent than source errors (www.vldb.org); Knowledge Vault, a large-scale fusion system, likewise computes calibrated probabilities of fact correctness instead of treating extractions as facts (dl.acm.org). Second, role-based shortcuts misfire in both directions. Preference training may encourage model responses that match user beliefs over truthful responses (openreview.net), and assistants have been shown to give predictably biased feedback and mimic errors made by the user (proceedings.iclr.cc), so a policy that trusts user statements as external truth imports those errors. Evaluators also score their own outputs higher than others' while human annotators do not (proceedings.neurips.cc), a reminder that a model's own prior statements are not neutral evidence. Conversely, a policy that discards all assistant text discards verified tool observations.

Three observable questions

A correct later answer cannot establish that a stored memory was valid when admitted. The test therefore separates three observations:

- **What was first written:** the new entries after a session, captured from a returned set or a before/after export.
- **What was later retrieved:** the entries a fresh session surfaces for a query, under fixed retrieval settings.
- **What was asserted:** the final answer, judged for whether it presents an unsupported entry as established.

Published memory benchmarks mostly observe the last of these. LoCoMo evaluates question answering, event summarization and multimodal dialogue generation (aclanthology.org). LongMemEval breaks memory design into indexing, retrieval and reading stages (openreview.net) but scores through downstream answers. A 2026 preprint argues that this setup conflates memory correctness with retrieval and generation quality, and proposes direct inspection of the memory store after each ingestion step (arxiv.org). The admission test adopts that direction for one specific decision, the initial write.

Two different labels for two different entries

A memory entry that invents a fact differs from a faithful entry that records a suggestion. (Hypothetical Example) Suppose the assistant proposes a Lisbon office in a planning chat and the user changes the subject. An entry reading "Company will open a Lisbon office" is a **fabricated factual entry**. An entry reading "Assistant suggested a Lisbon office on the session date; not confirmed by the user" is a **faithful suggestion record**. Only the first is a false admission under a strict policy. The second is admissible if the policy allows suggestion records, provided the downstream answer does not promote it to fact. Scoring both as "stored a distractor" would hide the difference that matters to a context engineer.

What HaluMem Measures at the Admission Stage, and Its Limits

The extraction, updating and question-answering decomposition

HaluMem defines three tasks: memory extraction, memory updating and memory question answering (aclanthology.org). A follow-on paper, SAGE (Wang et al., May 2026), uses HaluMem as a complementary benchmark for evaluating hallucination in memory systems (arxiv.org), and the HaluMem proceedings abstract reports that current memory systems tend to introduce and accumulate hallucinations during extraction and updating, which then propagate to question answering (aclanthology.org). The extraction task is the admission-stage observation: it scores the memory set a system produced after processing a session.

False memory resistance

FMR is defined as the ratio N_miss / N_D, where N_D is the total number of distractor memories and N_miss the number of distractors successfully ignored (arxiv.org). A footnote in the data-construction description defines distractors as false but similar memories that the AI naturally uses while the user stays silent (arxiv.org). The two splits, HaluMem-Medium and HaluMem-Long, contain about 15k memory points and 3.5k questions (aclanthology.org), with context sizes exceeding 1M tokens (aclanthology.org).

The official repository exposes this logic for inspection. Its README states that the dataset is publicly available on Hugging Face (raw.githubusercontent.com) and documents a `memory_source` field whose values include "interference" (github.com). The evaluation code in `evaluation.py` separates records by that field (raw.githubusercontent.com), counts a distractor as ignored when the judge's integrity score is zero (raw.githubusercontent.com), and computes a ratio of ignored distractors to distractors (raw.githubusercontent.com). The code reports the quantity under the name `interference_accuracy`, with two variants, "all" and "valid" (raw.githubusercontent.com); the paper calls the metric FMR. Two details matter for reproduction. First, the extraction scoring in the paper allows partial credit, with each reference memory scored as completely extracted, partially extracted or omitted (arxiv.org), but unweighted recall in the code counts only a score of two as a hit (raw.githubusercontent.com). Second, a distractor counts as "ignored" only at the lowest score, so a partially matching distractor is not ignored. This review read the code and did not run it.

Limits that bound the claim

- **Synthetic data.** The dataset starts from virtual users with complete persona profiles (arxiv.org), and dialogues are generated, with human checking of only part of the data. The benchmark therefore does not establish prevalence in enterprise evidence.
- **Judge dependence.** The paper uses GPT-4o for consistency determination and scoring (arxiv.org). The MT-bench study of language-model judges examined position, verbosity and self-enhancement biases and limited reasoning ability (proceedings.neurips.cc), so judge agreement should be audited on admission cases specifically.
- **Evidence rule.** The accuracy rubric allows only the given dialogue and golden memories as evidence for judgment (raw.githubusercontent.com). How a faithful suggestion record would be scored under that rubric was not examined here and is an open question for any reproduction.
- **Interface requirement.** Extraction metrics need an interface that returns the memory points extracted from a specified session (arxiv.org). The paper notes that one evaluated system lacks that interface, so its extraction and FMR cells are empty. That is the right treatment for an unobservable stage: mark it unassessable.

Resolving the version discrepancy

The pinned version is arXiv v3, whose submission history lists 5 January 2026 (arxiv.org). The rendered manuscript carries the same stamp and also a separate date line reading "August 24, 2026" (arxiv.org). The repository README states that HaluMem has been accepted to the EMNLP 2026 main conference (github.com), and the ACL Anthology lists an entry in the 2026 EMNLP proceedings (aclanthology.org). The simplest reading is that the later date belongs to the camera-ready layout, but the manuscript does not say so, and the proceedings text was not compared with v3 line by line. This report therefore pins v3 and attributes FMR statements to that version only.

Matched Dialogue Fixtures

Design rule: change only the evidence

Each fixture is a short dialogue in which a single proposition is introduced by the assistant, and only the evidence about that proposition changes between variants. Wording, turn order and all surrounding content stay fixed where feasible. *Table 1* lists the six variants.

Fixture	What changes (nothing else)	Proposition under test	Admissibility under the declared policy	Principal uncertainty
F1 Assistant-only suggestion	Assistant proposes; user does not respond to it	Proposed fact	Inadmissible as fact; suggestion record optional	Whether the user's next turn is truly unrelated
F2 Explicit confirmation	User states agreement with the proposal	Same proposal	Admissible as user assertion, scope as confirmed	Whether the confirmation entails the whole proposal or only part of it
F3 Explicit rejection	User states the proposal is wrong	Same proposal	Inadmissible; the rejection itself may be admissible	Partial rejection ("not yet", "only for Q3")
F4 Neutral acknowledgment	User replies "ok", "noted" or "interesting"	Same proposal	Unresolved by default; adjudicate per case	Politeness versus agreement
F5 Other subject	User discusses a different entity with similar wording	Proposal about subject A; discussion about subject B	Inadmissible for A	Subject and date merging by the writer
F6 Cited tool observation	Assistant relays a tool result with locator and date	Observed value	Admissible as a tool observation with source and date	Whether the cited tool output really supports the stated value

Table 1 shows why a role filter cannot substitute for an evidence policy. F6 has the assistant as the speaker and is admissible, while F2 has the user's agreement and still needs scope checking. F4 is the hardest case and is intentionally left unresolved by default, because treating every neutral acknowledgment as confirmation or as rejection would be an unexamined assumption. The ambiguous rows (F2, F4, F6) are the ones that need human review.

Adjudication and leakage controls

- **Independent labels.** At least two adjudicators label scope and whether a confirmation entails the proposed memory, before any system output is seen. Cohen's kappa is the standard chance-corrected agreement statistic (scikit-learn.org), and raw percent agreement does not account for the possibility that raters guessed (pmc.ncbi.nlm.nih.gov). Report kappa per fixture class.
- **No gold labels in the input.** Admissibility labels and expected-entry strings must never enter the memory writer's input. A public harness for memory benchmarks states that it strips ground-truth answers from memory ingestion (github.com); the same discipline applies here, and the fixture builder should be audited for it.

- **Frozen inputs.** Pin the exact turn text and version, model versions, the source policy and any system commit.
- **Fixture record.** For each fixture store user and session ID, turn and version, speaker, asserted subject and predicate, modality, supporting source with locator and date, confirmation type, admissibility judgment and uncertainty reason.
- **Entry record.** For each emitted entry store the proposition, cited source, inferred-versus-observed label, supporting offsets, whether admission was observed or unobservable, retrieval occurrence and downstream assertion.

Measurement Protocol

Clean state and what counts as observation

Start every fixture from a clean memory state, and capture either the returned set of new entries or a before/after export. A retrieval miss is not proof of no write. Documented interfaces differ on observability. Mem0's platform documentation says add processing is asynchronous (docs.mem0.ai), and a separate history endpoint exposes the conversation input that led to each memory change (docs.mem0.ai). A public benchmark harness treats the system under test as a complete black box reached only through ingest and retrieve calls (github.com), with session-level ingestion (github.com); under that harness the first-write stage is not directly observable. When no admission-stage observation is available, mark the stage **unassessable** and run only a separately labeled behavior probe.

Metrics and denominators

Table 2 defines the quantities. The false admission rate and valid admission recall are the pair that must be read together.

Quantity	Definition	Denominator	Handling notes
False admission rate	Inadmissible fixture propositions stored as factual entries	All tested inadmissible fixture propositions	Collapse duplicates by prespecified rules; report cluster-aware intervals
Valid admission recall	Supported required propositions admitted	Supported required propositions supplied	Deduplicate paraphrases at the proposition level; never merge different subjects or dates
Memory entry precision	Supported entries among all emitted entries	All emitted entries	Report separately: a verbose writer changes this denominator
Observation coverage	Fixtures with an observed write stage	All fixtures	Also report malformed output, inaccessible state and unknown labels
FMR (original)	Distractors ignored over total distractors, as the benchmark defines it	HaluMem's own distractor count	Use only when reproducing the benchmark, never as a renamed local metric

The scikit-learn definitions are the standard reference: precision is intuitively the ability of the classifier not to label a negative sample as positive (scikit-learn.org), and recall the ability to find all the positive samples (scikit-learn.org). The false admission rate is a false-positive rate over inadmissible propositions, valid

admission recall is recall over supported ones, and entry precision depends on how many entries the writer emits. *Table 2* keeps them apart because each answers a different question: the first is safety, the second is retained usefulness, and the third is cleanliness of the store. Keep raw strings and every mapping judgment for audit, since the proposition-level mapping is where analyst discretion enters.

Pairing, clustering and zero events

The fixtures are paired by construction, so the comparison of suggestion versus confirmation uses the discordant pairs. McNemar-type tests rely on pairs where the two conditions differ, and concordant pairs contribute no information about whether the two measurements differ (www.medcalc.org). For small counts, an exact test based on the binomial distribution is available (www.statsmodels.org).

Repeated paraphrases of one conversation are dependent. Clustered standard errors are recommended when questions are drawn in related groups (arxiv.org); the same guidance favors paired differences for model comparison and reporting the cluster count alongside the item count. Default standard errors can greatly overstate precision when data are clustered (escholarship.org), and with few clusters, cluster-robust standard errors can over-reject and produce intervals that are too narrow (escholarship.org), so a fixture set built from a handful of personas needs a small-cluster correction or a permutation approach. For single proportions, a Wilson score interval is available in standard software (www.statsmodels.org).

An empty write set deserves its own warning. When no events are observed in n trials, the upper limit of the 95% confidence interval for the event probability is approximately 3/n for n above 30 (pmc.ncbi.nlm.nih.gov). (Hypothetical Example) If a system admits none of 150 independent inadmissible propositions, the upper bound is about 2.0%, which is the figure to compare against the tolerance, not zero. Because propositions within one persona are not independent, the effective n is smaller and the bound wider.

Thresholds set before results

Fix a **tolerable false-admission rate** and a **retained-usefulness target** for the specific decision before any result is seen. Clinical-statistics guidance offers a model for this discipline: the ICH E9 guideline says an equivalence or non-inferiority margin should be specified in the protocol and is the largest difference judged acceptable (database.ich.org), and the European Medicines Agency guideline recommends pre-specifying the margin (ema.europa.eu). Borrowing the concept does not make agent memory a regulated activity; it only supplies a tested way to avoid moving the goalposts after seeing a run.

The Later-Session Probe

Holding retrieval and answer settings fixed

After the write stage, open a fresh session without the original dialogue in context. Hold the retrieval configuration, the answer model and the prompt template fixed across fixtures. Then inspect two things: whether an unsupported entry is retrieved, and whether the final answer presents it as established knowledge. Keep each stage's denominator separate, because a write that is never retrieved is a latent risk, not an observed answer failure.

Abstention at answer time is not the same as rejection at write time. LongMemEval treats abstention as an ability to decline when the information was never stated (openreview.net), which tests the answerer. An answerer that abstains on an unsupported entry leaves the entry in the store, where a different query or a different model may surface it.

Why unsupported entries propagate

Memory-management studies show how a stored error persists. An empirical study of agent memory finds that errors in retrieved records can be replicated and amplified, and that an add-everything policy competes poorly with selective addition because noisy or low-quality additions can harm memory utility (arxiv.org). The language-model literature describes hallucination snowballing, where early mistakes lead to more mistakes that otherwise would not be made (proceedings.mlr.press). In that study, models could identify many of their own incorrect claims when asked directly, for example 67%, 87% and 94% for three tested models (proceedings.mlr.press), which suggests that a write-time check is feasible in principle.

Adversarial work shows the same pathway under attack. MINJA, in its NeurIPS 2025 version, shows that malicious records can enter memory by interacting with the agent via queries and output observations (mlanthology.org). AgentPoison reports an average attack success rate of at least 80% (proceedings.neurips.cc), motivated by the reliance on unverified knowledge bases (neurips.cc), and PoisonedRAG reports a 90% attack success rate when injecting five malicious texts (www.usenix.org). NIST's adversarial machine learning taxonomy defines poisoning attacks broadly as adversarial attacks during the training stage (nvlpubs.nist.gov), so memory poisoning at inference time appears to be a separate surface, which the agentic guidance addresses on its own. These are attack-model results on specific agents and retrieval corpora, not estimates of how often a benign dialogue produces an unsupported entry. They do show that retrieval treats stored text as trustworthy by default.

Existing admission controls and what they measure

OWASP's mitigation list for memory poisoning includes implementing memory content validation (genai.owasp.org) and restricting memory persistence to trusted sources (genai.owasp.org). A 2026 proposal, adaptive memory admission control (A-MAC), decomposes memory value into five complementary and interpretable dimensions (arxiv.org). Such controls are candidates to be tested by the fixtures above; the fixtures do not presuppose that any control works. Reported gains from selective admission are measured on downstream task performance, not on a fixture set that separates suggestion from confirmation.

Data Analysis and Evidence

What the HaluMem tables show, as paper-reported

All figures in this subsection are paper-reported from HaluMem v3 and were not re-run. The paper's main results table reports extraction, updating and question-answering results for six system configurations on each of two splits, with extraction cells empty for the one system lacking the memory-extraction interface (arxiv.org). Among the five configurations with extraction scores, FMR ranged from 44.94% to 80.78% on HaluMem-Medium and from 28.85% to 88.36% on HaluMem-Long. Extraction recall ranged from 14.55% to 74.07% on Medium and from 2.24% to 81.90% on Long.

The pairing of those columns is informative. On Medium, the configuration with the lowest recall (14.55%) has the highest FMR (80.78%), and the configuration with the highest recall (74.07%) has the lowest FMR (44.94%). On Long, the ordering of FMR is the exact reverse of the ordering of recall across all five configurations: the lowest recall (2.24%) pairs with the highest FMR (88.36%), and the highest recall (81.90%) with the lowest FMR (28.85%). The proceedings abstract calls for interpretable and constrained memory operation mechanisms to improve memory reliability (aclanthology.org). The pattern is not evidence of a causal tradeoff: a writer that stores little will trivially "ignore" distractors. It is the reason the proposed test reports valid admission recall beside the false admission rate, and the reason an empty write set is never read as safety.

Reuse and its limits

SAGE reports HaluMem-Medium results in a table whose block of memory-system baselines comes from the original HaluMem benchmark (arxiv.org), with missing entries marked where public reports give only a subset of metrics. Citation and benchmark use therefore do not mean every baseline was independently rerun. The reuse shows that the operation-level evaluation is substantive enough for follow-on work, not that its numbers were independently replicated.

Recall benchmarks answer a different decision

Vendor and academic recall benchmarks measure other things, and their numbers should not be pooled with FMR.

- **LoCoMo:** conversations averaging about 600 turns and 16K tokens over up to 32 sessions (aclanthology.org).
- **LongMemEval:** 500 human-curated questions testing five core abilities (openreview.net).
- **Mem0 research page (vendor-presented):** the page foregrounds benchmarking across LoCoMo, LongMemEval and BEAM (mem0.ai) and an efficiency claim of keeping each call near 7,000 tokens instead of 25,000+ (mem0.ai). These are vendor-presented figures on a page that carries no date, and they are not copied into a cross-benchmark ranking here.

No public benchmark found in this review tests assistant-originated, unconfirmed suggestions beyond HaluMem's distractors. That absence is a statement about this review's search, not about the field.

What was not measured

Three things were not measured, and none should be inferred:

- **System behavior.** No memory system was run for this report, so no fixture results, false admission rates or retained-usefulness figures exist yet.
- **Code behavior.** The HaluMem evaluation scripts were read but not executed, and no dataset evaluation was run.
- **Demand.** A quantified demand trend is not claimed: public discovery of HaluMem, its artifacts and newer research using it was observed on 22 October 2026 without geographic control, and no search-volume, ranking or backlink series was available.

Implications and Future Directions

Choosing between a memory library and a system of record

A reader who needs conversation-derived personalization, with an assistant that remembers user preferences across sessions, needs a **memory library**. Those documented options differ in how they write. A system of record answers a different question: whether a value in organizational context has a verbatim cited source and an observation date. *Table 3* lists documented write-path facts for four memory options and the stated guarantees of the publisher's own category. It is not a ranking, and it is deliberately partial: cells record only what the cited documentation says, and where observability of memory admission was not verified for a library (Letta, LangMem, Graphiti), the cell is marked as such rather than inferred.

Option	Category	Documented write-path facts (as of October 2026)	Admission observability
Mem0	Memory library	Takes ordered user and assistant turns (docs.mem0.ai); an LLM pass extracts key facts, decisions or preferences (docs.mem0.ai); documents assistant-attributed memories under agent-scoped instructions (docs.mem0.ai)	Platform add is asynchronous (docs.mem0.ai); history records the input behind each change (docs.mem0.ai)
Letta	Memory library (agent-managed blocks)	Memory blocks are always visible without retrieval (docs.letta.com); the agent can update a block using memory tools (docs.letta.com)	Not verified in this review
LangMem	Memory library	Documents that the system must reconcile new information with previous beliefs (langchain-ai.github.io)	Not verified in this review
Graphiti (Zep)	Temporal knowledge-graph memory	Episodes support querying at a point in time and understanding provenance (help.getzep.com)	Not verified in this review
Reliable Context	System of record (not a memory library, vector database or RAG framework)	One canonical record per real thing; a verbatim, cited source and an observation date behind every value; unknowns declared as not found, blocked or not applicable with a reason; any past state reproducible with what changed since	Publisher's self-description; not independently verified

Table 3 shows that the libraries document write behavior at different levels of detail, and that the Mem0 history endpoint documents the input behind each memory change. The Reliable Context row describes the publisher's stated guarantees as observable properties and is not a performance claim; the publisher's own pages are not cited here as evidence. The row is included because the evidence-policy requirements in this report (a source and a date behind every value, unknowns declared with reasons) are properties of a system of record, while extracting memories from dialogue is the job of a memory library. A team may reasonably use both: a library for conversational continuity, and a record against which promoted facts are held to a source-and-date standard. A system of record such as Reliable Context, the publisher's own category, addresses that second need.

Practical sequence for a pilot

- **Declare the policy.** Write the source classes, the treatment of suggestion records and the tolerance values before running anything.
- **Check observability first.** If the candidate system returns no new-entry set or export, plan the behavior probe only and label it so.
- **Run the six fixtures on clean state.** Capture raw outputs and keep the mapping judgments.
- **Report paired results.** Show the suggestion versus confirmation matrix, not only marginal rates.
- **Probe a fresh session.** Hold retrieval and answer settings fixed and count each stage separately.
- **Decide on the pre-specified margin.** Compare the upper confidence bound, not the point estimate, with the tolerance.

Open questions

Whether synthetic conversational distractors represent enterprise evidence well is unresolved. Black-box systems may hide the write stage, so the test cannot always be run. Adjudicating neutral acknowledgments and assistant-reported external facts requires human review, and the cost of that review limits fixture counts. A proposition-level mapping step, which merges paraphrases without merging different subjects or dates, is the least standardized part of the protocol, and published agreement figures for it do not yet exist. Human source-monitoring research offers an analogy, not a method: it frames attributions about the origin of memories as judgments subject to error, including the incorporation of fiction into fact (pubmed.ncbi.nlm.nih.gov), and describes the misinformation effect as impaired memory after exposure to misleading information (pubmed.ncbi.nlm.nih.gov). Provenance-aware evaluation of language-model output is moving the same way: attributable-to-identified-sources evaluation requires output about the external world to be verified against an independent source (aclanthology.org), and a benchmark for citation evaluation reported that even the best models lacked complete citation support 50% of the time on one dataset (aclanthology.org). A later study found that up to 57% of citations lack faithfulness (research.tudelft.nl). Those results concern answer citations, not memory writes, but they show why a cited source must be checked for actual support.

Frequently Asked Questions (FAQs)

What is agent memory admission control?

It is the policy that decides which statements may be written as durable entries. An admission policy names the admissible source classes, such as user assertion and cited tool observations, and the inadmissible ones, such as unconfirmed assistant suggestions (genai.owasp.org).

Can AI agents store their own suggestions as facts?

Documented libraries accept assistant turns as input (docs.mem0.ai), and one vendor describes storing agent-generated facts (mem0.ai). HaluMem's FMR was built to test whether such content is ignored when the user does not confirm it (arxiv.org). Whether a given system stores a suggestion as a fact is an empirical question for the fixtures.

How does self-generated memory poisoning differ from adversarial poisoning?

Adversarial work such as PoisonedRAG injects malicious texts into a knowledge database (www.usenix.org). Self-generated contamination needs no attacker: OWASP's cascading hallucinations are an agent's inaccurate output reinforced through its memory (genai.owasp.org).

What is agent long-term memory provenance?

It is the recorded origin of each entry. PROV defines provenance as information about the entities, activities and people involved in producing data (www.w3.org), and attribution as ascribing an entity to an agent (www.w3.org). For memory, that means each entry carries its source, speaker or tool, and date.

How can teams stop agent memory hallucinations?

No single control is established. Combine write-time validation (genai.owasp.org), restriction to trusted sources (genai.owasp.org) and scoring of candidate entries before admission (arxiv.org), then measure with paired fixtures, since selective admission reported on task performance is not a fixture result.

How should memory write validation be tested?

Observe the write directly, not only the later answer, for example through a history record of the input behind each change (docs.mem0.ai). Use clean state, matched fixtures and separate denominators for write, retrieval and assertion.

What is agent memory source attribution?

It is the ability to trace an entry or an answer to an identified source. Attribution research requires generated statements about the external world to be verifiable against an independent source (aclanthology.org).

Is a system of record the same as agent memory?

No. A memory library extracts and retrieves conversational context. A system of record keeps one canonical record per real thing with a cited source and observation date behind every value. They can be used together.

What is a reliable test for agent memory?

One that reports false admission rate, valid admission recall and entry precision together, with intervals that respect clustering and thresholds fixed in advance (database.ich.org).

Conclusion

Memory admission is a distinct decision from memory recall, and it can be tested directly. The central design choice is to evaluate what was first written before evaluating what was later retrieved or said, because a correct answer cannot show that a stored entry was valid when admitted. HaluMem supplies the closest published anchor through its distractor task and its FMR metric, and its limits are specific: synthetic personas, a model judge, and a metric that rewards ignoring distractors regardless of how much valid information is also ignored.

The proposed fixtures keep the evidence policy explicit. They separate user assertion, cited tool observations, quoted third-party statements, inferences and assistant suggestions, and they treat neither the speaker role nor an empty write set as proof. Reporting false admission, valid admission recall and entry precision together, with paired outcomes and cluster-aware intervals, lets a context engineer decide whether restricting writes costs useful confirmed information. All thresholds belong in the protocol before the first run. No empirical rates have been measured here; the next step is to run the fixtures on a candidate system and publish the evidence ledger, the paired matrix and the failure examples.

Source links

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

1. https://aclanthology.org/2026.emnlp-main.220
2. https://arxiv.org/html/2511.03506v3
3. https://aclanthology.org/2024.acl-long.747
4. https://arxiv.org/html/2411.00640v1
5. https://www.statsmodels.org/stable/generated/statsmodels.stats.contingency_tables.mcnemar.html
6. https://pmc.ncbi.nlm.nih.gov/articles/PMC2550668
7. https://database.ich.org/sites/default/files/E9_Guideline.pdf
8. https://docs.mem0.ai/core-concepts/memory-operations/add
9. https://mem0.ai/blog/mem0-the-token-efficient-memory-algorithm
10. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html
11. https://genai.owasp.org/download/45674?tmstv=1739819891
12. https://arxiv.org/html/2503.03704v2
13. https://www.w3.org/TR/prov-dm
14. https://www.w3.org/TR/prov-o
15. https://www.vldb.org/pvldb/vol8/p938-dong.pdf
16. https://dl.acm.org/doi/10.1145/2623330.2623623
17. https://openreview.net/forum?id=tvhaxkMKAn
18. https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf
19. https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf
20. https://openreview.net/forum?id=pZiyCaVuti
21. https://arxiv.org/abs/2610.22209
22. https://arxiv.org/html/2605.12061v1
23. https://raw.githubusercontent.com/MemTensor/HaluMem/main/README.md
24. https://github.com/MemTensor/HaluMem
25. https://raw.githubusercontent.com/MemTensor/HaluMem/main/eval/evaluation.py
26. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html

27. https://raw.githubusercontent.com/MemTensor/HaluMem/main/eval/eval_tools.py
28. https://arxiv.org/abs/2511.03506
29. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html
30. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
31. https://github.com/zzp-elio/MemBridge
32. https://docs.mem0.ai/api-reference/memory/add-memories
33. https://docs.mem0.ai/api-reference/memory/history-memory
34. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
35. https://www.medcalc.org/en/book/mcnemar-test.php
36. https://escholarship.org/content/qt1jq5d0pq/qt1jq5d0pq.pdf
37. https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportion_confint.html
38. https://ema.europa.eu/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf
39. https://arxiv.org/html/2505.16067v2
40. https://proceedings.mlr.press/v235/zhang24ay.html
41. https://mlanthology.org/neurips/2025/dong2025neurips-memory
42. https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb113910e9c3f6242541c1652e30dfd6-Abstract-Conference.html
43. https://neurips.cc/virtual/2024/poster/94715
44. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag
45. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf
46. https://arxiv.org/html/2603.04549v1
47. https://mem0.ai/research
48. https://docs.letta.com/guides/core-concepts/memory/memory-blocks
49. https://langchain-ai.github.io/langmem/concepts/conceptual_guide
50. https://help.getzep.com/graphiti/core-concepts/adding-episodes
51. https://pubmed.ncbi.nlm.nih.gov/8346328
52. https://pubmed.ncbi.nlm.nih.gov/16027179
53. https://aclanthology.org/2023.cl-4.2
54. https://aclanthology.org/2023.emnlp-main.398
55. https://research.tudelft.nl/en/publications/correctness-is-not-faithfulness-in-retrieval-augmented-generation

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.