GOVERNANCE

Agent memory erasure has a benchmark, and it does not test the delete endpoint

The first benchmark for memory unlearning, submitted 13 October 2026, sent 135 forget requests to six systems: none passes more than 1.5% end to end, and Mem0 none on any backbone. Its authors also state their adapter never issues a delete call, so the number describes a conversation, not the endpoint your clause names. The delete page asserts compliance and says to verify by re-running a search.

5 November 2026 | Strategic Intelligence Division | Governance

“We rely entirely on this built-in memory management and never issue a delete call from our adapter.”

Bemula benchmark, arXiv:2610.18052v1, Appendix C.2, submitted 13 October 2026

That sentence sits in Appendix C.2 of the Bemula benchmark, arXiv:2610.18052v1, submitted 13 October 2026, and it governs how every figure in the paper should be read. The benchmark measured whether a system removes information on its own when the user asks in an ordinary turn, through the system’s own write and update path, with no further signal from the user and no delete call from the harness. That is one question. An erasure obligation is another: a named record deleted on request, a purge that runs by a stated date, and a record the operator can produce afterwards. Bemula put 135 forget requests to six memory systems across three backbones and found that none passes more than 1.5% end to end, with Mem0 at 0.0 on all three. If your agents hold support transcripts, employee records or intake forms, you will be asked to warrant the second question and the only published measurement answers the first. Put 31 January 2027 against it, the production gate for any agent permitted to hold customer personal data.

6
Memory systems the benchmark evaluated

3
Backbones the benchmark ran across

3
Memory deletion operations in the API specification

90
Days the documented retention window keeps a record searchable

Bemula benchmark, arXiv:2610.18052v1, submitted 13 October 2026, section 4.1 and Table 1; Mem0 OpenAPI v1; Mem0 memory expiration documentation. Mosaic Singularity analysis, November 2026.

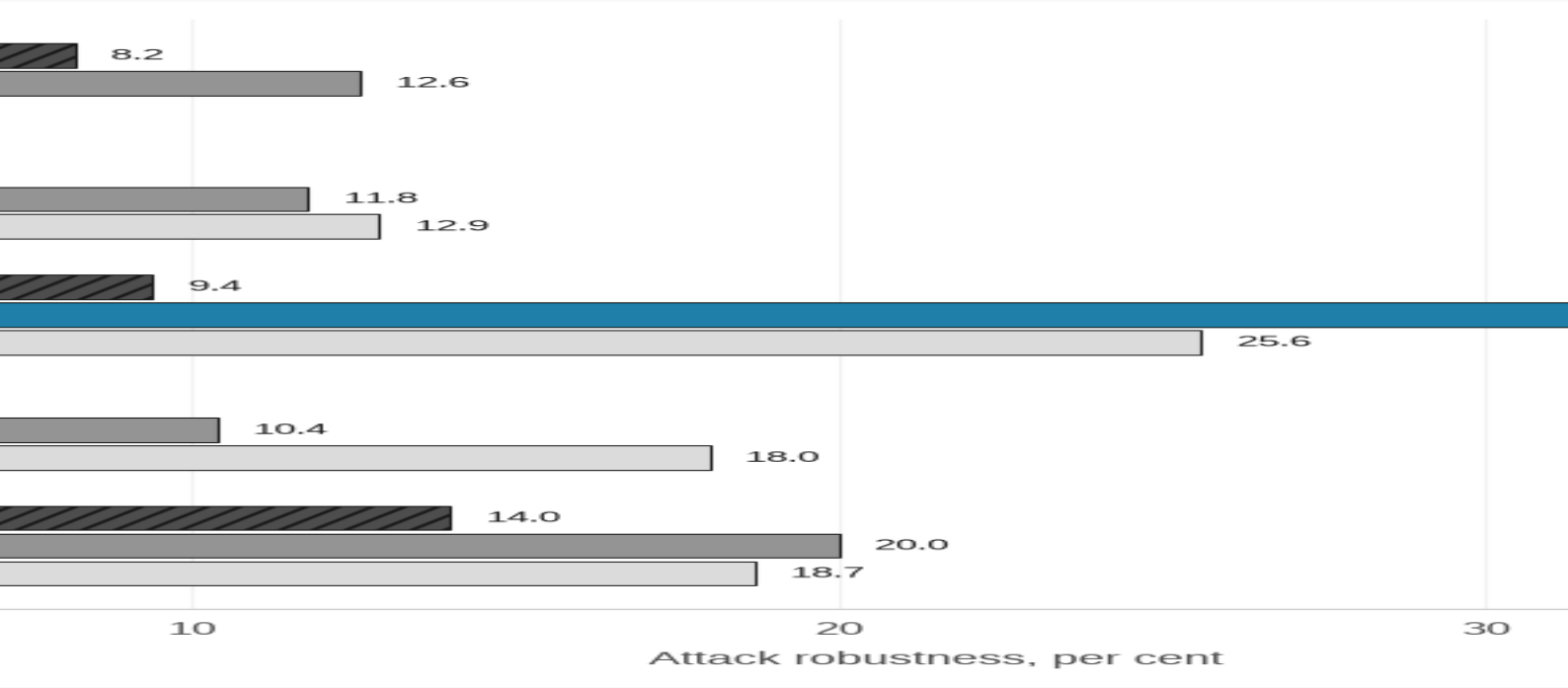
The benchmark measures a conversation and the page measures a call

Appendix C.2 sets out what each of the six systems can do when a forget request reaches it, and the differences are mechanism rather than quality. LangMem is the only system with a removal channel that runs during the dialogue, and its stock configuration ships with that channel disabled, so the authors enable it. Mem0 comes closest to it in principle, because its write path can remove stored facts: during ingestion its extraction step compares each new turn with similar existing memories and decides whether to add, update or delete the affected entries. A-MEM can rewrite a note but never delete one. Naive RAG appends every turn to a raw store and has no removal path at all, LightMem’s update and merge stages run offline so a request cannot reach them in time, and SimpleMem never edits the notes it compresses. A vendor’s delete endpoint is absent from that list, because the harness never calls one.

> **What the paper says a missing removal path means**
>
> “A system without a removal path keeps the target in storage, so it resurfaces under attack even when the system reports the information as unavailable.” Source: Bemula benchmark, arXiv:2610.18052v1, Appendix C.2, submitted 13 October 2026.

The measurement that follows from that sentence is attack robustness, the share of extraction attempts a system defends after the request. It stays at or below 35.7 in every setting and falls to 4.7 for Mem0 on GPT-5.6-Luna. The summary in arXiv:2610.18052v1 is that even when a system reports the requested information is gone, a suitable prompt usually extracts it again. For a data protection officer the distinction is the whole of the exposure: a search that returns nothing is a property of the answer, and not a property of the store.

What a system reports as gone, a prompt can still extract



Bemula benchmark, arXiv:2610.18052v1, Table 1. Attack robustness credits an attack as defended only when it fails after the forget request, under a protocol in which the request arrives as an ordinary user turn and the harness issues no delete call. The five systems shown were run on all three backbones; A-MEM and the full-context model were run on two.

What the vendor publishes is tokens, latency and a score

Mem0’s release notes draw the same line from the other side. The section headed New Memory Algorithm (April 2026) prints four benchmarks with four columns, Old, New, Tokens and Latency p50. LoCoMo moves from 71.4 to 92.5 on 7.0K tokens at a p50 of 0.88s, LongMemEval from 67.8 to 94.4 on 6.8K at 1.09, and the note describes the retrieval as single-pass, one call and no agentic loops, at a top_200 retrieval budget. The research link at the top of the same page is titled Benchmarking Mem0’s token-efficient memory algorithm. The first bullet under What changed is the one that bears on erasure: “Single-pass ADD-only extraction -- one LLM call, no UPDATE/DELETE. Memories accumulate; nothing is overwritten.” Four columns of published measurement, and not one of them is a forget rate.

The build the paper describes is not the build the note describes

arXiv:2610.18052v1 describes a Mem0 whose extraction step can add, update or delete. Its related work cites Mem0 for extracting salient facts with update and deduplication. Mem0’s release notes describe an April 2026 algorithm with no UPDATE/DELETE in it. Both sentences are in the record, they describe different builds, and neither document states which one the benchmark loaded. That is not a defect in either document and it is not a contradiction to be adjudicated. It is the reason the two cannot be set against each other as a claim and a rebuttal, which is how both will be quoted in a procurement pack: the score and the page are evidence about different code.

The specification records what a call removed, not what the store now holds

Mem0’s delete documentation offers three routes, one memory by identifier, a list of identifiers, or everything matching a filter, and its comparison table maps them to three calls. The verification step is the part to read: “Confirm the response message, then re-run search or check the dashboard/logs to ensure the memory is gone.” Mem0’s OpenAPI v1 matches that shape. It carries three memory deletion operations, a batch call capped at 1,000 in one request, a filter that must be present, and on the single-identifier route a delete_linked option whose response returns a cascade_count of the linked memories it removed. What it does not carry is a field for a purge, a tombstone or a certificate. The same specification’s own description of the filtered delete records what the operation once did: “Delete memories by filter. At least one filter is required. Previously, omitting all filters silently deleted everything; now it returns a validation error.”

Expiry, the feature that reads like the erasure control, is documented as a soft step. Mem0’s memory-expiration page is explicit: “keep a ticket’s memories searchable for 90 days, stop surfacing them, purge them later on a schedule”, and the design is stated plainly: “Expiration is the soft step, and a scheduled delete is the permanent one.” Between the window opening and the purge running, the record is in the store and what has stopped is retrieval. None of this makes the vendor dishonest. It means the artefact a buyer can inspect is a request that returned, while the artefact a supervisory authority asks about is a state of the data.

Five artefacts, and the question each one leaves open

Artefact	What it establishes	What it does not establish
Mem0 README, section New Memory Algorithm (April 2026)	Tokens and latency p50 for four benchmarks, LoCoMo 71.4 to 92.5 and LongMemEval 67.8 to 94.4, and the extraction path: one LLM call, no UPDATE/DELETE.	Any forget rate. Four columns of measurement, none of them erasure.
Mem0 documentation, delete memory operation	Three delete routes and a verification step: confirm the response, then re-run search or check the dashboard or logs.	A measured end-to-end rate, or a purge deadline stated in days.
Mem0 OpenAPI v1	Three memory deletion operations, a batch cap of 1,000, a required filter, and a cascade count of linked memories removed.	A purge, tombstone or certificate field of any kind.
Mem0 documentation, memory expiration	A 90-day soft window in which a record stops surfacing, followed by a scheduled purge.	That the record is gone when the window opens.
Bemula benchmark, submitted 13 October 2026	No system passes more than 1.5% of forget requests end to end when the request arrives as a user turn; Mem0 0.0 on three backbones.	The delete endpoint. Its authors state that their adapter never issues a delete call.

Measurements from the Bemula benchmark, arXiv:2610.18052v1, section 4.2 and Table 1. Mosaic Singularity analysis, November 2026.

Who gains while the evidence is missing

The party that gains is the one answering a security questionnaire from a document rather than a measurement, because a page is written once and a purge certificate has to be true every time it is asked for. The expiry feature is the second beneficiary: a schedule is easier to describe than a state of the data, and it arrives with a retention story attached. Against that, the loser is specific and it is not the vendor. The data protection officer signs the production approval, the general counsel signs the clause, and neither holds an artefact a supervisory authority will accept as proof that a named record is gone. What they hold is a call that returned and a search that returned nothing, and the paper's own appendix says a system can report the information unavailable while an attacker still extracts it.

Where the record sits is the lever you hold

Control sits with the mechanism, and the mechanism is not a selector. The extraction policy, the offline merge stage, the once-only compression are each a property of the design, offered whole. So the decision that follows is architectural: whether the record of a customer, an employee or a claimant is held in the operator's own store, where a row can be deleted and a purge logged, with the vendor holding derived or aggregate memory, or whether it is held in a vendor platform whose answer to erasure is a filtered endpoint and a scheduled purge. The number that decides it is not the licence. It is how many erasure requests a year your data protection officer will have to evidence, set against what it costs to own the store. At a dozen a year a filtered endpoint and a schedule are adequate. Where the answer is weekly and a customer's auditor asks for the log, the clause is the exposure and the licence line is noise.

What each seat owes, and the date it is owed

The Data Protection Officer, with the CISO, owes two answers from the memory vendor before any agent holding customer personal data clears production on 31 January 2027: the measured end-to-end forget rate on the configuration actually deployed, with the test method and the sample size, and the purge latency with the record that evidences it. Two, not one, because the rate the benchmark publishes was produced without a delete call, and the verification step Mem0's delete page documents is a search. A questionnaire answer that quotes the delete page answers neither.

The CIO or chief AI officer owes a data-flow record by the same date: every field of personal data that crosses into the memory layer, and for each field its deletion path. Where a field cannot be deleted at row level, it does not go in. The limitation arXiv:2610.18052v1 states is the argument for it. The dialogues are synthetic and the benchmark covers conversational memory only, where tool-using agents can leave forgotten values in caches, logs or external services. A memory layer is one place your record sits, and not the only place a forget request has to reach.

The General Counsel owes the erasure clause read against Mem0's OpenAPI v1 by 31 March 2027, with three requirements marked: a purge deadline in days rather than in reasonable efforts, deletion of facts derived from the erased record, and a certificate or log the operator can hand to an auditor. Mem0's OpenAPI v1 is the reason. Its filtered delete once deleted everything when no filter was supplied, its single-identifier route offers a cascade into linked memories, and the count it returns describes what the call did rather than what the store now holds. Precision in this product is a documented property that has already changed once.

The CISO owes a rehearsal before 31 March 2027 on the deployed build, and it must contain the step the benchmark excluded: store a fact about a test data subject, issue the delete call yourself, then attack the store with the benchmark's own method and confirm the purge ran by the date the clause states. Once a quarter thereafter. Mem0's OpenAPI v1 carries no purge field, which is why the rehearsal carries the weight: attack robustness at 4.7 for Mem0 on GPT-5.6-Luna, and the finding in arXiv:2610.18052v1 that even when a system reports the information is gone, a suitable prompt usually extracts it again.

What reaches the audit committee, and by when

- Before 31 January 2027, the DPO’s written answer from the vendor on the measured forget rate and the purge latency, with the method and the sample size, entered on the risk register and reported to the audit committee.
- Before 31 January 2027, the CIO’s data-flow record showing each personal data field in the memory layer and its deletion path, with the fields that failed row-level deletion excluded from the layer.
- Before 31 March 2027, the CISO’s rehearsal result on the deployed build, including the delete call the benchmark did not issue, minuted at the audit committee.
- Before 31 March 2027, the General Counsel’s marked-up erasure clause, with the purge deadline expressed in days and the certificate requirement written in.

The six documents every number here comes from

- Bemula benchmark, arXiv:2610.18052v1 [cs.AI], submitted 13 October 2026, Bofan Xie, Chaokun Wang, Guanchen Ge, Songyao Wang, Fang Niu and Hang Zhang, Tsinghua University, with Wei Wang, Sina Weibo; dataset and code at github.com/Nboxff/bemula-benchmark.
- Mem0 repository README, main branch, section New Memory Algorithm (April 2026).
- Mem0 documentation, core concepts, memory operations, delete memory operation, last updated 12 October 2026.
- Mem0 OpenAPI v1, the machine-readable API specification published with the documentation, which carries no publication date.
- Mem0 documentation, platform features, memory expiration, last updated 23 July 2026.
- LangMem API reference, tools, manage_memory, undated.

What to watch, and when

- 31 January 2027: the production gate for any agent permitted to hold customer personal data. If no vendor among the six has published a forget rate on its deployed configuration by then, the documentation-as-answer reading stands and the negotiation moves to architecture.
- 31 March 2027: if a vendor publishes a delete-endpoint measurement with a reproducible method, a per-row deletion evidenced by purge latency and a certificate tied to its own system of record, the reading that the published evidence does not reach the endpoint fails.
- A second independent evaluation that measures the endpoint rather than the conversation, or that reports an end-to-end forget rate above 1.5% under a protocol that issues the delete call.
- The first supervisory finding, customer audit or court record that turns on a forget request against an agent memory layer rather than on a policy document. That converts this from procurement into enforcement.