Anamnesis Benchmarks

Evidence, not claims

Persistent memory for AI, measured with a pre-registered, reproducible benchmark suite. Three arms that isolate memory from everything else, blind grading, and honest negatives reported next to the wins.

8 TASKS · 5 RUNS EACH · OPUS-4-8 SUBJECT · BLIND HAIKU JUDGE · 2026-08

Method

Each task asks for a specific, non-derivable fact about the system, chosen so a model cannot guess it from reasoning alone (validated: a blinded model scores near zero). We compare three arms:

- **A. Source present** the relevant source document is in context (the theoretical best).
- **B. Blinded** no memory and no source.
- **C. Anamnesis** the relevant memory is injected.

Answers are graded blind against a fixed rubric, shuffled so grading order carries no signal, with human spot-checks. We report method and results, not implementation internals. Comparisons are against a naive, flat vector-retrieval baseline; head-to-head results against named products are not yet available and are not implied here.

T1. Fidelity: delete the source, keep the knowledge

With the source removed, a blinded model scores near zero. Injected memory restores near-oracle answer quality while carrying a fraction of the context.

Arm	Fidelity	Input tokens
A. Source present	0.983	7,820
B. Blinded	0.075	143
C. Anamnesis	0.942	431

Memory closes almost the entire quality gap of the full-document baseline at roughly an 18x token discount. Mean of 8 tasks, 5 runs each.

T5. Cost and the search tax

Without memory, an assistant rediscovers a fact the expensive way: it searches and reads across many turns before it can answer. With the fact already in context, it answers on the first turn. Measured on real headless sessions run against a live repository.

Metric	With memory	No memory	Ratio (off ÷ on)
Turns	1.0	5.0	5.0x
Input tokens	30,395	138,891	4.6x
Time-to-value	5.4 s	27.6 s	5.1x
Cost per query	see note	see note	2.2x to 7.6x

Cost per query is cheaper with memory by a factor that ranges from about 2x on a cold session to roughly 8x on a warm one, where prompt caching is in effect. The size of the tax scales with how deeply the fact is buried; we report the per-task spread rather than a single averaged figure.

T2. Scale and noise

Vector search looks strong in a demo and degrades in production as the pool fills with near-duplicate and contradictory records. We tested this deliberately and report where it breaks.

Recall holds, then crowds. We buried 8 known facts under up to 1,000 realistic distractors, 25 percent of them adversarial near-misses (same wording, wrong value). Retrieval recall stays perfect through 500 records. At 1,000, under that hostile mix, the most-contested fact is pushed out of the top five. Latency stays flat throughout (13 to 42 ms).

Pool size	Recall @ 5
20	1.00
100	1.00
500	1.00
1,000	0.88

Recall understates the cost. What matters is whether the answer is correct. Even when the true fact was still retrieved, the context also carried contradictory wrong values, and the model could not tell which was authoritative: it either picked a wrong value or refused to answer. Answer fidelity for flat retrieval fell from 0.94 to 0.42. This is the classic vector-store failure at scale.

Recovery: trust in ranking

The fix is provenance. When a contradiction arrives from a lower-trust source while the real decision is self-authored, trust-aware ranking surfaces the authoritative memory first.

Ranking at 1,000 memories	Arm C fidelity
Clean pool (baseline, ~20)	0.942
Naive / flat retrieval	0.421
Trust down-weighting	0.648
Trust exclusion	0.942

Down-weighting helps but is partial: a down-ranked contradiction is still injected and still confuses the model. Excluding uncorroborated records from the decision set restores full fidelity.

Honest scope. This recovers the common case, where a contradiction is lower-trust than the authoritative fact. It does not resolve contradictions that are equally trusted and equally recent; those require temporal belief revision (T3) or human review. Trust-in-ranking is a validated capability currently behind a feature flag and not yet enabled in production.

T3. Temporal belief revision

A plain vector store has no notion of a decision changing: the old record stays in the index and can be served as current. Anamnesis treats truth as temporal. When a decision is superseded, a contradiction sweep detects it and retires the old record. From then on, the injected context carries only the current value; the superseded record is inhibited but preserved for audit and point-in-time history.

In direct testing, the sweep detected the supersession automatically, the current value was served, and the stale value was excluded from the current-truth context while remaining available for historical reads. All checks passed.

Security and IP containment

Isolation is enforced at the architectural level, not in prompt instructions. This section describes the model and its outcomes and omits implementation internals.

- **Read-only guest access.** A read-only credential is carried by a scoped service token and enforced at the API layer on every route, so it cannot write, even through REST.
- **One-way mirror.** A local model reads the shared pool for context; its own output stays local and never flows back, containing IP.
- **Local-first execution.** Inference defaults to a local, quantized model, so sensitive queries can be answered without leaving the machine.
- **Provenance and trust tiers.** Every memory carries a provenance-derived trust tier (established, provisional, unverified) that is never self-claimed. A provisional memory may not override an

established one.

Method and results are reproducible from a fixed task set and harness. Figures reflect the measured runs described above. Comparisons against named third-party products are planned and are not represented in this report.