ICOS Report II — Provenance-Traced Belief Memory: Claims, Evidence, Contradiction, and Change

Robert Vandelinder · Exile Research Inc. · ORCID 0009-0001-9793-420X

Preprint v1.0 — November 2026

Abstract

Large language models supply inference, but a persistent agent needs memory that can distinguish **what was observed** from **what is believed**. Most agent memory is either a rolling context buffer or an opaque vector store: observations and beliefs are conflated, provenance is absent or inferred, and memory changes silently. This report presents the memory architecture of ICOS, implemented and evaluated across three milestones. It has three properties. First, beliefs ("claims") are **derived from an immutable evidence ledger** and carry stored provenance; they are created only through a promotion pipeline that is **human-approved by default** and run through an **exactly-once journal**. Second, **contradiction is first-class**: conflicting beliefs are both retained, the loser marked contradicted but still queryable, and unresolved conflicts are parked as questions rather than overwritten. Third, memory **changes honestly**: reinforcement, decay, revision, and retirement are bounded, append-only, and audited, and *salience is tracked separately from truth*. Beliefs are recalled through three parallel surfaces, ranked with confidence gating and provenance lenses, and delivered to the model in **labeled context bands that never mistake memory for user speech**. We report the design rationale, the concrete parameters, and the evaluation (unit tests: 701 at M12's close; 1,327 unit / 90 e2e / 245 web at v0.8, plus live runs), and we state plainly what the evaluation does **not** establish.

---

1. Introduction

An agent that runs for months must maintain beliefs that survive model swaps, contradiction, and time. The common pattern — append conversation to a context window, retrieve similar text — produces three failures:

1. **Observation and belief are conflated.** What the user said, what the model said, and what the system concluded all become the same kind of "memory." Nothing records *why* a belief is held.
2. **Contradiction is invisible.** New information usually overwrites or coexists silently; the system cannot say "I believe X, but I once had evidence for not-X."
3. **Change is unaccountable.** Memory decays, strengthens, or is revised by opaque heuristics, with no history and no way to walk back.

ICOS treats memory as an **epistemic** system: a belief is a first-class record derived from evidence, with provenance, status, and confidence kept distinct from the evidence itself. This report is the second in a series. It has three tiers of scope. The **first tier** — M10 (belief substrate), M11 (recall), and M12 (dynamics) — is described mechanism by mechanism (§5) and evaluated directly (§6). The **substrate** — the M4 candidate ledger, the session store, and the provider boundary — is Report I; this report cites it rather than re-describing it. The **deferred** items — the M23 knowledge-base corpus, evidence invalidation, and gist minting — are named where they arise (§5.4, §7, §8), not claimed as built.

Throughout, `M#` is the project's internal milestone identifier (sub-slices appear as `M10b`, `M11a`, and so on); each maps to a plan and a committed verification record in the repository (Appendix A).

Contributions.

- **C1 — Evidence-first beliefs with stored provenance.** Claims are derived from an immutable candidate/evidence ledger (M4); each carries a normalized-triple identity (case, whitespace, edge quotes; no stemming), stored origin (user vs. agent, never join-derived), and an append-only audit trail. Promotion is human-approved by default and exactly-once across restart.
- **C2 — Contradiction as a first-class state.** Conflicts (same subject+predicate with a different object, or explicit same-triple negation) keep both beliefs; the loser is marked *contradicted* and remains queryable, and unresolved conflicts become parked questions.
- **C3 — Honest, bounded dynamics.** Reinforcement, category-scoped decay, revision with walkable history, and deliberate retirement are all append-only and audited; **activation (salience) is orthogonal to confidence (truth)**, so "loud but wrong" is a representable and detectable state.

We also report the recall and context-assembly layer that makes beliefs usable at turn time without letting them masquerade as user speech (§5.4–5.7).

---

2. Background and related work

Retrieval-augmented generation [Lewis et al., 2020] retrieves text passages into the prompt but is stateless: retrieved text is not a belief, carries no provenance beyond its source document, and does not evolve. ICOS's memory is stateful and its units are provenance-traced beliefs, not passages.

MemGPT / Letta [Packer et al., 2023] manages a tiered memory that the *model* edits through function calls. ICOS keeps change **deterministic and human-gated by default**, with an append-only audit trail; the model proposes candidates, it does not decide belief.

Generative Agents [Park et al., 2023] introduced a memory stream with recency/importance/relevance retrieval and reflection. ICOS shares the "memory as a stream" intuition but adds formal contradiction handling, stored origin, and a separation between retrieval-time observation and maintenance-time change.

Knowledge graphs, truth-maintenance systems, and belief revision provide the theory ICOS borrows from: contradiction handling and revision are classical problems (e.g. the AGM postulates [Alchourrón et al., 1985]) and event-sourced systems [Fowler] keep an append-only log as the source of truth. ICOS applies these ideas to LLM-mined beliefs with an explicit, bounded rule set rather than a general solver.

Agent reasoning/acting patterns (ReAct [Yao et al., 2023], Reflexion [Shinn et al., 2023], Voyager [Wang et al., 2023]) and the conceptual taxonomy of language-agent architectures (CoALA [Sumers et al., 2024]) sit above this layer; ICOS provides the durable memory substrate those patterns assume but rarely specify. Classical cognitive architectures (Soar [Laird], ACT-R [Anderson et al.]) inspire the salience/activation separation; ICOS makes no claim to a unified theory of cognition.

Agent harnesses. The closest contemporary systems are *agent harnesses* — the runtime that wraps a model and gives it hands. Claude Code, OpenAI's Codex, and LangChain's Deep Agents established the coding-harness standard (permission systems, subagents, context compaction); **Pi** (pi.dev) shows a deliberately minimal harness grown by its user; **OpenClaw** (with NVIDIA's NemoClaw/OpenShell) pushes security into the runtime boundary; and **Hermes Agent** (Nous Research) is a self-improving general agent that creates and updates skills from experience. ICOS belongs to this camp in spirit — *the runtime is the harness* — but differs on two axes. It is **provider-independent by hard boundary**, and its centre of gravity is the **epistemic and identity layer** that harnesses under-specify: provenance, contradiction, reviewed identity, and verified claims. Hermes is additionally ICOS's **lineage and cautionary precedent**: the earlier generations were built on it and were destabilised by runaway self-directed change, so ICOS makes change deliberate, human-gated, and audited (see §3).

Novelty claim. Not a new algorithm, but an **implemented architecture** that makes three properties default: provenance-stored beliefs, first-class contradiction, and salience orthogonal to truth — with the evidence ledger immutable throughout.

---

3. Design principles and boundaries

The architecture is governed by a small set of explicit rules; deviations are treated as bugs.

1. **Observation ≠ belief.** The M4 candidate/evidence ledger is immutable. Beliefs *reference* ledger rows; they never edit them. A belief cannot exist without a recorded derivation and at least one evidence reference.
2. **Provenance is stored, never derived.** The user/agent role is stamped at observation time (joining back to messages is fragile under undo/fork), and each evidence entry keeps its own role. Mixed-origin evidence never collapses to a single source.
3. **Humans gate belief creation.** Promotion requires human approval by default; automatic promotion exists but is **off** and admits nothing until explicitly configured.
4. **Nothing is silently rewritten.** Every change is a status transition with an append-only history row; the ledger never moves.
5. **Salience ≠ truth.** Activation (how "present" a belief is) is a separate dimension from confidence (how likely it is true).

6. **Recall observes; maintenance changes.** The recall path may not mutate beliefs except to record access counters. All mutation happens in a background maintenance pass.
7. **Degrade, don't fail.** A missing recall surface, an absent corpus, or a crashed maintenance pass degrades the system; it never fails a conversation turn.

State ownership. The milestones partition responsibility cleanly, and this partition is the architecture's backbone:

Stage	Owner	May write
Evidence	M4	the candidate/evidence ledger (immutable afterward)
Birth	M10	claims, promotion journal, prospective items
Reading	M11	`accessCount` / `lastAccessedAt` only
Aging	M12	confidence, status, activation, history tables

Runtime assumptions. ICOS memory assumes a single trusted operator, one host with a local filesystem for its SQLite stores, and a model endpoint behind the OpenAI-compatible boundary described in Report I. Nothing in the memory layer requires a network service: the semantic index is an embedded library and the ledger is a local file. The assumptions matter for the claims below — a multi-writer store would change the exactly-once argument, and a shared multi-tenant deployment would change what "the user" means for provenance.

---

4. Architecture

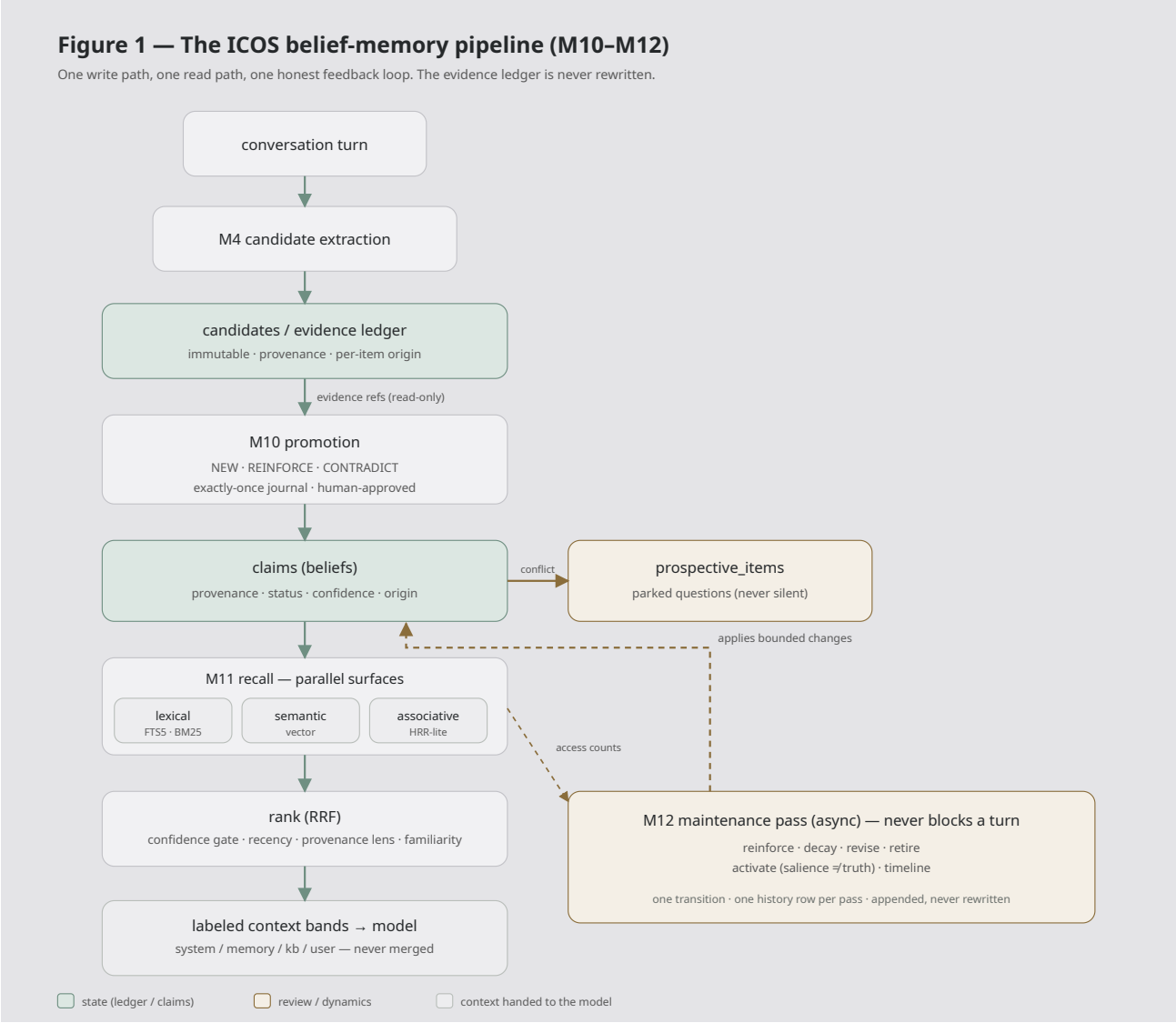


Figure 1 — the belief-memory pipeline: one write path (M4 → M10), one read path (M11), one honest feedback loop (M12).

Storage. SQLite is the system of record. Memory is not entrusted to a third-party store for basic function: the semantic index is an *embedded* library (`VectorDB` + `OnnxEmbedder`, `all-MiniLM-L6-v2`, 384-d, ONNX; no server, no API key — the model asset is fetched once and cached under the data root, after which inference is fully local) and is explicitly positioned as **recall acceleration only** — losing it degrades recall, never the record. Lexical search is SQLite FTS5 with BM25. The associative surface is a bespoke HRR-style resonance scan in SQLite.

---

5. Mechanisms

5.1 The belief record (M10b)

A claim is a normalized `(subject, predicate, object)` triple plus: `category` (fact/preference/relationship/procedure), `status` (candidate/active/contradicted/retired), two confidences — `extractorConfidence` (the observation, never re-estimated) and `confidence` (the engine estimate) — `firstAssertedAt` / `lastSurfacedAt` (originating vs. latest evidence), `origin` (user/agent), `evidence[]` (candidate refs with per-item role, never copies), a counter `timesObserved`, and reserved fields with a single future owner each.

Two disciplines are worth highlighting. **Identity** is the normalized triple; normalization is deliberately conservative (case, whitespace, edge quotes — no stemming). **Schema reservation:** fields that a later milestone will own (`activation`, `locked`, `sourceType`, `summary`, `related[]`, `emotional`) exist from the start, are read as null/zero, and are never defaulted into meaning; verification asserts they stay untouched. This is a direct lesson from a prior generation, where "we'll add it later" cost migrations.

5.2 Promotion (M10c)

Every promotion is exactly one derivation:

- **NEW** — no claim shares the triple; create it.
- **REINFORCE** — the triple exists; append evidence, touch `lastSurfacedAt`, re-estimate confidence (fixed rule).
- **CONTRADICT** — the triple conflicts with an active claim; the old claim becomes *contradicted*, a new one is created (§5.3).

Promotion runs through a small **journal** (`proposed → promoting → committed`, with `denied` and `failed` as terminals) whose approval id is the idempotency key. A crash between grant and write replays cleanly; a retried grant converges instead of duplicating. Authority defaults to human approval (action `memory.promote`); the automatic path is implemented but **off**, and admits nothing above a risk threshold until explicitly opted in.

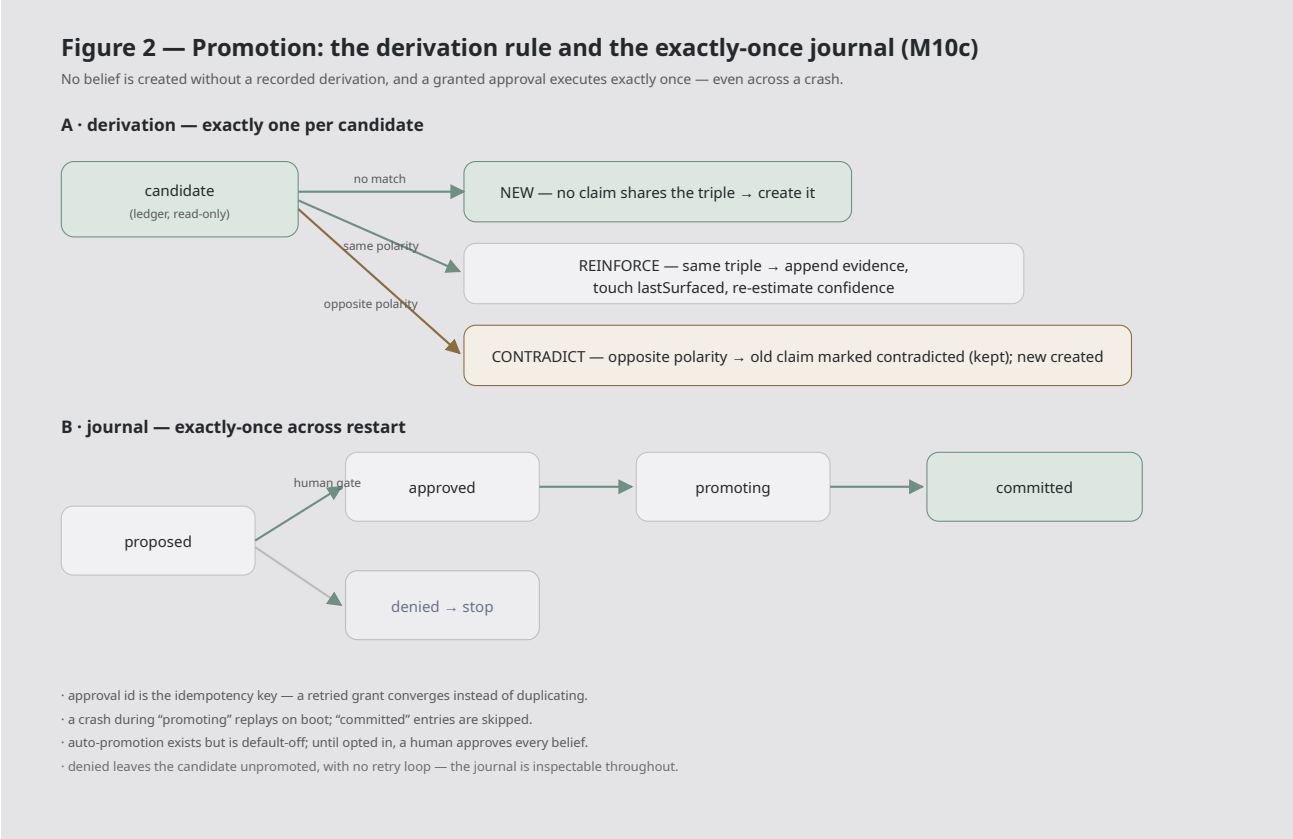


Figure 2 — the derivation rule and the exactly-once journal.

5.3 Contradiction and the clarification queue (M10e)

Conflict is defined narrowly and deterministically: **same normalized subject+predicate with a different object**, or an **explicit same-triple negation** (an explicit boolean, never inferred from text). On conflict both beliefs are kept; the loser stays queryable with its evidence intact — history is not rewritten. Affirmation and negation of one triple coexist as rows (unique on `(identity_key, negated)`).

A recorded contradiction with no follow-up is a dead end. When a claim's confidence drops below a threshold (default `0.5`) or the same triple is contested twice, promotion writes a **prospective item** describing the conflicting values, their origins and confidences, and a suggested question. Items sit in a queue (`GET /core/prospective`); ignoring one parks it — it is evidence that a question exists, not a demand. Resolution (M12c) records an outcome (`confirmed` / `corrected` / `dismissed`); a *correction* is a revision with provenance, not an edit.

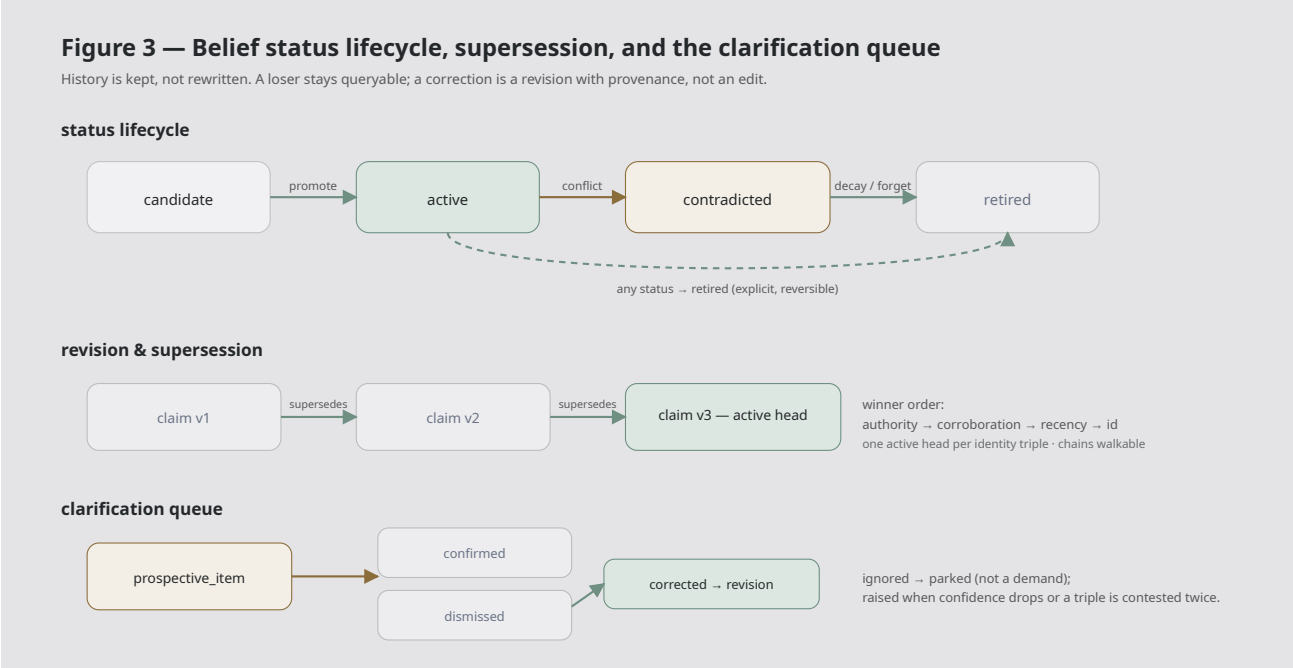


Figure 3 — status lifecycle, supersession chains, and the clarification queue.

5.4 Recall surfaces (M11a)

Three parallel channels, each returning scored hits tagged with their surface identity so fusion can attribute contribution:

- **Lexical** — FTS5 over subject/predicate/object/entities, BM25; phrase fallback so natural language never triggers an FTS-syntax crash.
- **Semantic** — vector search over claim embeddings via the embedded index.
- **Associative** — an HRR-style resonance scan over a bounded newest-`200` claim window (exact match weight `1.0`, prefix-partial `0.5`, minimum partial length 3) — *what feels linked*, not what matches.

Query shaping extracts a focused topic from a mixed conversational prompt (trimmed, bounded to 500 chars). A **KB bridge** contract is defined but the corpus is owned elsewhere (M23); the bridge returning nothing is a normal state, flagged not failed.

5.5 Ranking (M11b)

Order is "relevance first, honesty constraints as gates":

- **Confidence gate** — below `0.3`, excluded unless explicitly queried.
- **Contradicted** — demoted (half weight) and labeled, never silently dropped or promoted.
- **Recency** — `1 / (1 + ageDays/30)`, a true gradient.
- **Provenance lens** — configurable exclusion/downweighting of `agent`-originated claims from interactive recall (default off), so the agent's own chatter cannot crowd out user signal. Exact-match, deterministic, reported in the trace; the store is untouched — a lens, not a deletion.

- **Familiarity** — more than two outcomes: `recalled` / `familiar` (sub-threshold near-miss, labeled) / `unranked`, with `excluded` (gate, lens, duplicate, beyond-limits) reported separately.
- **Fusion** — Reciprocal Rank Fusion (Cormack et al., 2009) with `RRF_K = 60`. Order-only fusion is deliberate: BM25 ranks, vector distances, and resonance counts share no scale, so combining scores would be a category error.

5.6 Context bands and the anti-laundering rule (M11d)

Retrieved memory enters the prompt in **separately labeled system-scope bands**, each carrying per-claim origin and confidence: the memory band opens `[memory: N recalled belief(s) from prior turns — beliefs, not user speech]`, a `[vaguely familiar …]` band carries near-misses, the KB band is its own labeled slot, and the persona band is `<persona_grounding>`. Memory is never merged into a pseudo-user message. The critical interface rule: **extraction must never treat injected context as user-said.** The extractor is given the band structure (or an explicit injected-context marker) so recalled claims re-mined from the transcript are not laundered into fresh user facts. The assembled prompt is inspectable: per-surface counts, gate decisions, and lens exclusions are recorded in a bounded recall trace (`GET /core/recall/trace`).

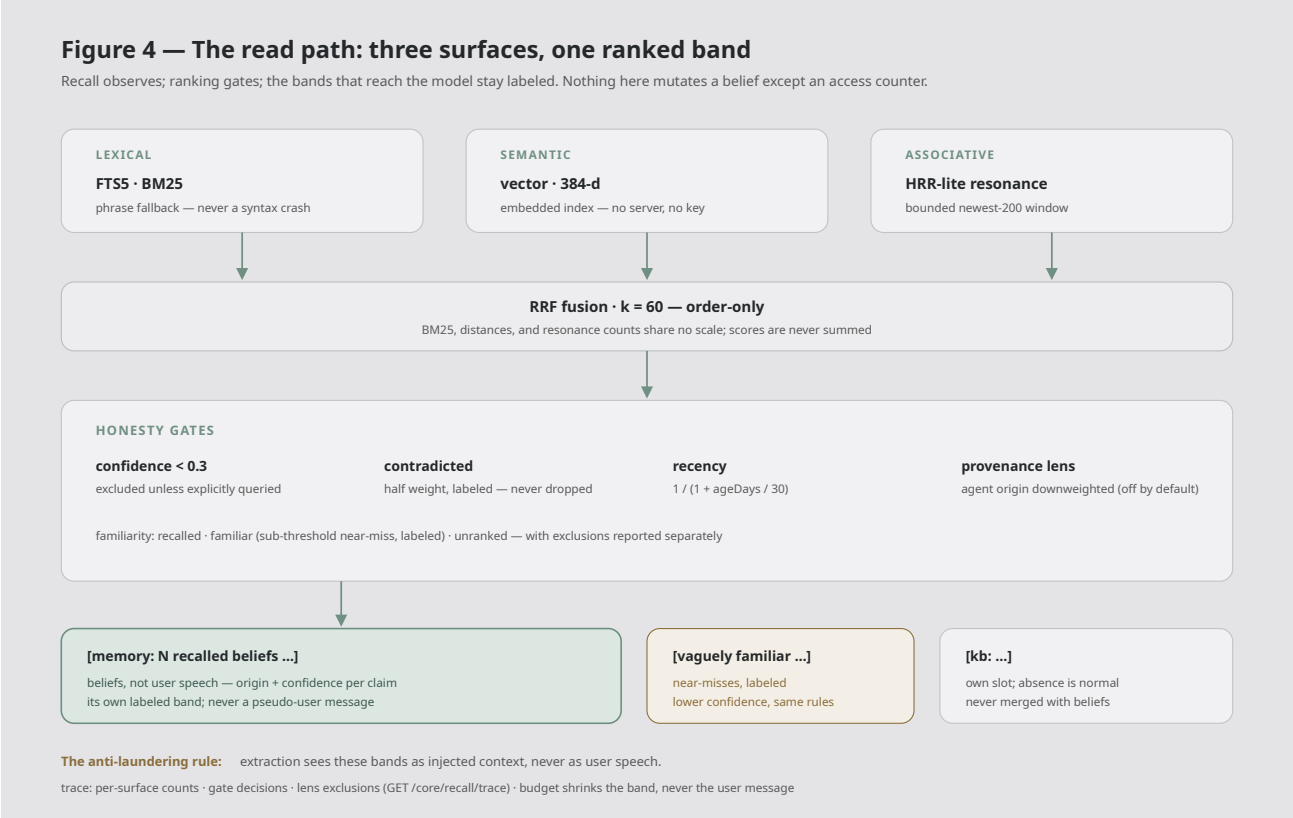


Figure 4 — recall observes, ranking gates, and the bands that reach the model stay labeled; extraction sees them as injected, never as user speech.

5.7 Budgeting and honest uncertainty (M11e)

Retrieval is bounded: a per-turn memory band cap (default `800` tokens, estimated as chars/4), a retrieval timeout (default `5000` ms), and a documented shrink order (familiar → demoted → lowest-fused). Oversized recall **shrinks the memory band; it never truncates the user message.** Per-surface failure degrades to the remaining surfaces; total failure yields empty bands with a `memory-degraded` flag and the turn continues. The agent-facing vocabulary is fixed — recalled / familiar / not-recalled, each with its confidence — so "I don't have that in memory" is an ordinary, reachable output rather than a failure.

5.8 Memory dynamics (M12)

All dynamics run in a **background maintenance pass** (default roughly hourly, also invocable via `POST /core/maintenance/run`); nothing blocks a turn, and a crashed pass recovers without half-applying a transition (one record, one transition, one history row per pass).

- **Reinforcement (M12a).** Promotion raises confidence by a fixed capped step (`+0.05`, ceiling `0.99`). The maintenance pass compounds a bounded boost (`+0.02 × min(5, timesObserved)`, hard cap `1`). Single observations never jump.
- **Decay and retirement (M12b).** Exponential half-life on confidence for unaccessed beliefs, past a 30-day grace, base half-life 90 days, with category multipliers (preference 0.8×, fact 1.0×, procedure 1.5×, relationship 2.0×) and per-category floors (0.15/0.2/0.25/0.3). Access within 7 days shields a belief. Retirement (a status transition, not a delete) requires a 180-day window *and* confidence ≤ 0.3, or an explicit human-approved instruction. Decay never deletes; the ledger never moves.
- **Revision and supersession (M12c).** Winner-picking is a fixed order: human-approved authority → corroboration count → recency → id. Supersession chains link a revising claim to its predecessor and are walkable, with one active head per identity triple. A certainty lock (≥ 5 observations and ≥ 0.9 confidence) protects a belief from **decay but not from revision** — evidence still wins. **Source reliability** populates the reserved `sourceType` (direct_statement / inference / speculation) and tracks per-source reliability with Laplace smoothing and punishment-only updates; agent-derived claims carry a fixed dampening factor (default `0.5`) so the agent's own statements cannot inflate by self-echo.
- **Activation — salience ≠ truth (M12d).** A separate `activation` dimension decays on its own schedule. Retrieval touches it (`+0.3`, clamped to [0,1]) and spreads it along links, fan-capped at 5 so over-connected beliefs do not dominate; disuse costs `0.05` per hour window. The payoff is a **divergence signal**: activation persistently high (≥ 0.7) while confidence decays (≤ 0.4) is the "loud but wrong" pattern, surfaced as a review item instead of a confident recall. Activation re-scores; it never gates recall alone and never promotes.
- **Temporal reasoning (M12e).** Claims carry timestamps; `GET /core/claims/timeline` navigates by time (`from`/`to`/`sessionId`), and time-anchored cross-links (capped at 5) connect a revision to its co-occurring context. A `temporalProximity = 1/(1 + days/7)` function is defined but, at this milestone, intentionally **not wired into ranking** — an unconsumed pure function is preferred over a weight with no caller.

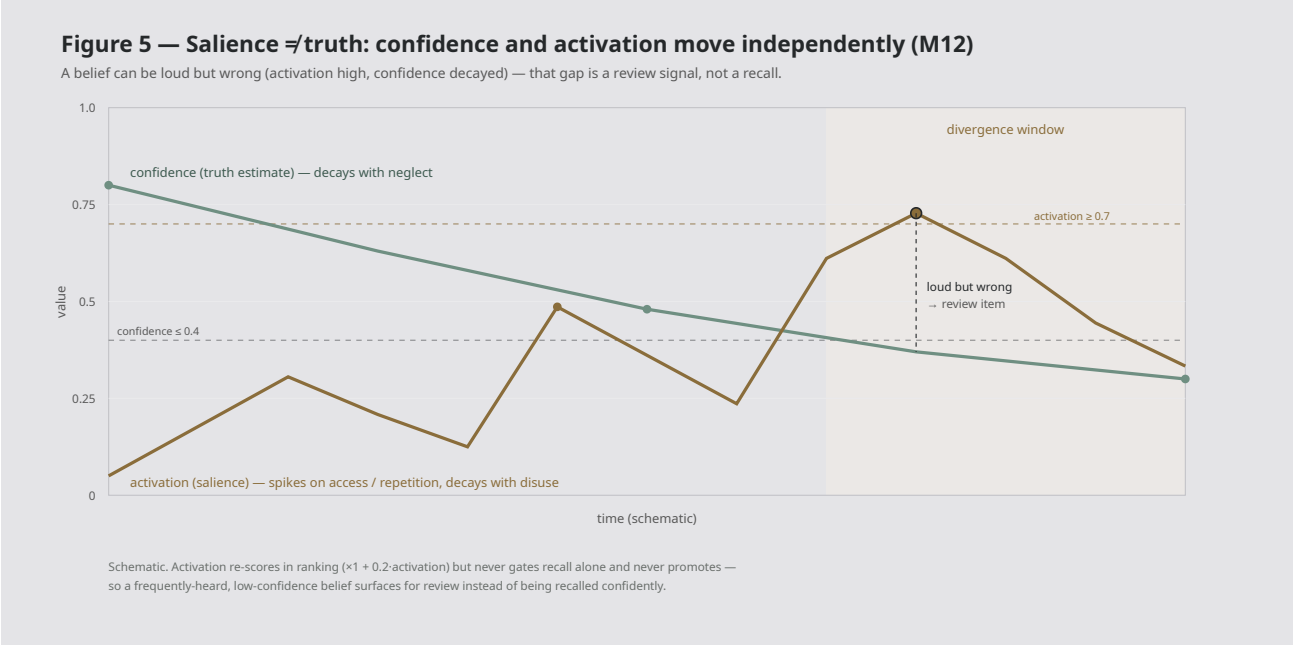


Figure 5 — confidence and activation move independently; the gap is a review signal, not a recall.

---

6. Evaluation

Method. Each milestone was verified with unit tests, end-to-end tests (the full application driven through HTTP), and live runs against a real model and a real database, with results committed as evidence. Test counts at each milestone's completion:

Milestone	Unit	E2E	Notes
M10 (memory substrate)	595	46	+ live promotion/contradiction sweeps
M11 (retrieval)	656	47	+ client tests land in M12
M12 (dynamics)	701	48	+ 168 web-client tests

Counts are **cumulative** suite totals at each milestone's close. The suite has since grown to **1,327 unit / 90 e2e / 245 web**, recorded at v0.8 (22 November 2026, `a95708b`), including the web client and the release hardening described in Report I.

Invariant verification. M10f asserts the epistemic invariants directly: every claim traces to ≥ 1 ledger candidate; no claim exists without a recorded derivation; concurrent duplicate promotions converge to one claim; a granted approval replays to the same claim across restart; re-mining touches `lastSurfacedAt` only; extractor and engine confidence never share a field; evidence lists reference (never embed) ledger rows; agent-mined facts keep `role: 'assistant'` end to end; and the reserved schema fields stay null and behavior-neutral. Durability invariants confirm the ledger's row counts and checksums are stable across a promotion run.

Live observations. A semantic search for "what programming language does the user like" returned the belief *user prefers TypeScript* at the top with score `0.649`, and triggered no automatic promotion. Promotion sweeps over extracted candidates returned the expected derivation counts (`{new: 1, skipped: 3}` and `{new: 1, contradicted: 1, skipped: 3}`) with the ledger untouched.

A live maintenance pass ran in 2–3 ms on the dataset while turns took ~15 s, which is the point of keeping it off the interactive path, and a ledger grew from 14 to 29 rows across eight passes without the claim store losing integrity. The scripted "loud but wrong" scenario — high activation with decayed confidence — raised a review item rather than a confident recall, and the anti-laundering check produced no candidate sourced from either injected band. The provenance rule also held against self-description: asked to confirm its own tooling and memory were "online", the agent treated recalled beliefs as `origin=user` claims, distinguished a mounted tool from a working milestone, and refused to corroborate a cross-session history it does not carry (verbatim in Report III §6).

What the evaluation does *not* establish. It shows the mechanisms behave as specified and survive restart, not that the architecture outperforms an alternative. There are no baselines or comparisons here, the model boundary in automated tests is mocked, and the corpora are small. §7 is explicit about this.

---

7. Limitations

The limits below are specific, and several are measured rather than suspected.

Scale is the open question. Constant-time cold-start was measured at n = 1 claim; the associative surface scans a bounded newest-`200` window and needs a scale review around ~10k claims, and the semantic index's behaviour beyond toy sizes is not yet measured. The ranking anomaly observed at tiny scale points the same way: on one exact query an agent-mined chatter claim ranked first, which is vector-distance luck at two-claim scale. It motivated the provenance lens, and it also shows how fragile ranking is before there is enough data to rank over.

The model boundary is mocked in automated tests. They verify the machinery, not model behaviour or extraction quality, so the extraction-precision claims rest on live observation rather than on a controlled comparison. Two implementation compromises are documented rather than hidden: `lastSurfacedAt` historically held a ledger reference rather than a timestamp, so recency resolves to the candidate's `extractedAt` with an `updatedAt` fallback; and the FTS5 external-content configuration is a known sharp edge, documented rather than shipped silently.

Some capabilities are designed but unbuilt. Evidence *invalidation* (a source retracted, not merely hidden by undo) has a design and no writer yet. Gist-family *detection* is implemented — a `gist_proposed` history row per member once a family of ≥ 3 claims is found — but *minting* a generalized belief outside promotion remains deferred, because it would violate the scope boundary. A parked-question rendering bug was also found: affirmed and negated rivals render identically.

Some things are intentionally unconsumed. `temporalProximity` is not wired into ranking because nothing calls it; the divergence signal currently lives only in a narrow activation/confidence window; the reserved `emotional` block stays null. Live coverage is partial in the same sense: not every gate, familiarity, or budget leg was exercised live in M11f.

---

8. Future work

Three threads follow from the limitations. The first is a **scale study** at roughly 10k claims, re-measuring cold-start and watching the associative and semantic surfaces for the failure modes that only appear with volume. The second is **evidence invalidation** with a real retraction source, which would close the designed-but-unbuilt gap and give "the source was withdrawn" a first-class path alongside `/undo`. The third is **gist authority**: deciding where generalized beliefs are legitimised, so detection can become minting without moving the scope boundary it currently respects.

The immediately following report (III) builds identity, drift detection, and claim verification on this memory layer; the belief store, provenance, and the salience/truth separation are the substrate it assumes.

---

9. Reproducibility

- **Artifact:** `a95708b2b74b75d7e404e4e30aaa9d9bad86c995` (tag `v0.8`, 22 November 2026)
  - **Repository:** https://github.com/ExileResearch/icos
  - **Release:** v0.8
- **Snapshot:** 1,327 unit / 90 e2e / 245 web (one deliberately skipped live-verifier spec), recorded at the artifact commit.
- **Setup:** Docker Compose (`docker compose up --build`); the memory store is a SQLite file under the data root.
- **Commands:** from `core/`: `npm test` (unit), `npm run test:e2e` (e2e), `npm run lint`, `npx tsc --noEmit`.
- **Evidence:** per-slice files under `.reference/plans/evidence/milestone-10/`, `milestone-11/`, and `milestone-12/`, browsable at https://github.com/ExileResearch/icos/tree/v0.8/.reference/plans/evidence.
- **Licensing:** code is PolyForm Noncommercial 1.0.0 (`LICENSE.md`); the report text license is declared separately on the Zenodo record and at arXiv submission.

Appendix A — Milestone and evidence map

`M#` is the project's internal milestone identifier. Each row points to the directory that owns the verification record; each record carries its own date and verification notes.

ID	Capability	Evidence (repository path)
M10	claims, promotion journal, contradiction, prospective items	`milestone-10/`
M11	recall surfaces, ranking, context bands, budget	`milestone-11/`
M12	reinforcement, decay, revision, activation, temporal	`milestone-12/`

All paths are relative to `.reference/plans/evidence/` at the cited tag. The M4 candidate ledger is Report I's scope; the M23 knowledge-base corpus is future work.

Acknowledgements

AI assistance. The manuscript was compiled and edited with AI assistance. The author reviewed every claim against the committed evidence and takes responsibility for the content.

References

- Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2), 510–530.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
- Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *SIGIR 2009*, 758–759.
- Fowler, M. (2005). *Event Sourcing.* https://martinfowler.com/eaaDev/EventSourcing.html
- Laird, J. E. (2012). *The Soar Cognitive Architecture.* MIT Press.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS.*
- Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). *MemGPT: Towards LLMs as Operating Systems.* arXiv:2310.08560.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.* arXiv:2304.03442.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning.* arXiv:2303.11366.
- Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). *Cognitive Architectures for Language Agents (CoALA).* arXiv:2309.02427.
- Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). *Voyager: An Open-Ended Embodied Agent with Large Language Models.* arXiv:2305.16291.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models.* arXiv:2210.03629.

Systems and harnesses

- Anthropic. *Claude Code.* https://claude.com/claude-code
- OpenAI. *Codex.* https://openai.com/codex
- Nous Research. *Hermes Agent.* https://hermes-agent.nousresearch.com
- OpenClaw. https://docs.openclaw.ai ; NVIDIA. *NemoClaw / OpenShell.* https://www.nvidia.com/en-us/ai/nemoclaw
- Zechner, M. *Pi.* https://pi.dev
- LangChain. *LangChain.* https://github.com/langchain-ai/langchain ; *LangGraph.* https://github.com/langchain-ai/langgraph ; *Deep Agents.* https://github.com/langchain-ai/deepagents
- Letta. *Letta (formerly MemGPT).* https://github.com/letta-ai/letta