The Judgment Layer

Turning agent memory into wisdom — and the layer nobody has built

Author: Mike Norton · ORCID 0009-0003-1866-6249 · DHARE — dhare.com.au · Brisbane, Australia **Version:** 1.0 · August 2026 · License: CC BY 4.0 **Status:** Architecture and evaluation protocol. A results paper follows the build. The judgment prompts, gate thresholds and evaluation corpus stay internal. **Keywords:** agent memory · judgment layer · cognitive architecture · constitutional governance · continuity

> *They built the library and the catalogue. We're building the librarian who decides what the library is for.*

Abstract

Persistent memory for AI agents is becoming a commodity. The funded field — Mem0, Zep/Graphiti, Letta, Cognee, LangMem, and now platform offerings like Cloudflare's Agent Memory — handles extraction and provenance well. But agent memory breaks down into four steps: **extract** what happened, **attribute** where it came from, **judge** what it means, and **govern** how it changes behaviour — and every shipping system stops after step two. Recent formal work (Roynard, 2026) identifies the same missing tier and reports near-zero contradiction-resolution across the field. I stake four claims about that empty rung. One: memory should be **shaped like attention** — a tree of focus sessions with depth-proportional compression — not shredded into atomic facts. Two: **continuity belongs to the judgment layer, not the model** — treat the LLM as stateless by design and identity survives model swaps, context compaction and provider churn. Three: typed persistence needs an **ignorance ledger** — a first-class store for known unknowns, with suspension of judgment (*epochē*) as a routing outcome alongside accept and refuse. Four: the layer that turns memory into behavioural guidance has to be **constitutionally governed** — promotion gated on evidence rather than user approval, amendment restricted to the human owner, and no agent ever self-ratifying the law that governs it. I finish with an evaluation protocol — architectural property tests plus a continuity test I call the *ache test* — and an open invitation to tear it apart.

---

1. Everyone files. Nobody judges.

Here's the state of play. The current generation of agent-memory systems is genuinely good at remembering. Mem0 extracts atomic facts at scale. Zep's Graphiti tracks temporal validity with bi-temporal timestamps. Letta treats context as RAM and lets the agent page its own memory. Real achievements, and this paper builds on them, not against them.

But remembering is the easy half. Break agent memory into its four steps — extract, attribute, judge, govern — and you find that steps three and four "require custom work" in every system surveyed, and no system natively implements step four at all. Roynard (2026) formalises the

missing tier: a four-layer model where **Knowledge** updates by supersession, **Memory** decays unless consolidated, **Wisdom** updates only through evidence-gated revision, and **Intelligence** is ephemeral inference. His benchmark discussion reports near-zero contradiction-resolution scores across current systems. And his pilot shows this cuts both ways: typed routing beat a flat memory store by +0.128 overall (+0.106 on contradictions, +0.150 on temporal reasoning) — while a naive keyword router *reversed* the entire advantage (−0.125). Judgment isn't a garnish on memory. Mis-routed memory is worse than no memory.

It gets worse before it gets better: the benchmarks that should referee this are themselves broken. A 2026 community audit found LoCoMo's answer key about 6% wrong, its LLM judge accepting most wrong answers, and LongMemEval fitting inside a single modern context window (both as reported in Roynard, 2026). The field is optimising scores that don't measure the thing that matters.

So the open problem isn't storage, retrieval or extraction. It's **judgment**: what deserves keeping, what earns the right to direct behaviour, how contradictions resolve, what gets forgotten, and who governs the whole process. This paper is an architecture for that tier. I call the layer **Marcus**, and Section 5 explains why that's not just a name.

2. Claim one — memory shaped like attention

Every system above shares the same reflex: shred experience into retrievable units. Facts, triples, summaries, embeddings. The *shape* of the thinking gets thrown away and only the residue is kept.

I think the shape is the point. Human episodic memory is organised by attention, not by fact-type: a subject held in focus for a duration, branching into sub-topics, re-entered later at the depth you left it. So the unit of memory in this architecture is the **focus session** — a tree where each node is a subject of attention with a natural depth class (a ten-minute glance, an hour's focus, a six-hour deep dive) and the children are the branches attention actually took.

Two things fall out of that. **Compression is proportional to depth.** At session close, a glance compresses to a line or to nothing; a focus to a paragraph with episode references; a deep dive to a structured summary with its branch tree intact. **Re-entry is a first-class operation.** "Where were we on the memory-architecture deep dive?" resolves to a node, restored at stored depth, with its open branches — not to a similarity-ranked pile of fragments.

And there's a nineteen-century-old existence proof that this compression discipline keeps what matters. Marcus Aurelius's *Meditations* is the output of a nightly consolidation practice run for about a decade by the bloke administering an empire — and it contains no transcripts. Thousands of audiences, dispatches, campaigns and crises compress into twelve thin books of lessons, some entries stamped by location ("Among the Quadi on the River Gran"). The trivial day left no entry. The repeated lesson became character. Ten years of raw experience down to a portable wisdom artifact that's still governing readers — that compression ratio is the target function for a forgetting mechanism, and it says aggressive, judgment-driven discard loses far less than intuition suggests (consistent with the roughly-20%-of-facts-preserves-full-behavioural-fidelity result reported in Roynard, 2026).

3. Claim two — continuity belongs to the judgment layer, not the model

Today an agent's continuity lives inside the agent: in its context window, its provider-side memory features, its habits. So when context compacts, the model gets swapped, or the provider changes an API, the self tears. I've watched it happen mid-task, and it's the problem that started this whole project.

I flip it. **The model is stateless by design** — amnesiac on purpose. All provider-side memory features are off, or routed through the judgment layer. Every turn, the judgment layer assembles the working context — the stable wisdom block first (which happens to be cache-optimal), then the focus frame, then need-driven recall with provenance tags — hands it to whatever model is in the seat, and judges what of the output gets written back. The model borrows a self. It never owns one.

The payoff is that the catastrophic events become non-events. Context compaction mid-thought: harmless, because the raw record and the focus tree live outside the window. Model deprecation: a config change — the next model boots into the same assembled self and can't tell it's a different instance. The way I put it in the design sessions: the model is this week's cells; the judgment layer is the self.

There's a security win buried in there too. Recalled memory is injected as *data with provenance, never as instructions*. The only channel where the past gets to *direct* behaviour is wisdom that passed the governance gates in Section 5 — which is a structural defence against memory-poisoning, not a bolted-on filter.

4. Claim three — typed persistence needs an ignorance ledger

I adopt Roynard's four persistence semantics wholesale — supersession for knowledge (append-only, provenance-linked, never deleted), decay-unless-reinforced for observations, evidence-gated revision for wisdom, ephemerality for inference — along with his design litmus that decay is a storage-level property of observations only, while recency is a query-time heuristic. Facts don't get less true with age. They get superseded, or they go stale and get *revalidated*, with bi-temporal timestamps keeping "when we learned it" separate from "when it was true."

I add two mechanisms. First, an **epistemic-status ladder** inside stored claims — observation, interpretation, hypothesis, conclusion — because an event and my judgment of the event are two different things, and a memory system that stores them in the same field will eventually believe its own moods. "A competitor launched X" and "this will destroy us" are not the same kind of object.

Second — and as far as I can tell, new as a first-class memory primitive — an **ignorance ledger**. Classification has always been binary-plus-bin: accept into some store, or discard. The Stoics knew a third verdict: ***epochē***, suspension of assent, for impressions you can't accept or refuse yet. I make suspension a routing outcome. A load-bearing assumption made without evidence, or a question the record can't answer, files as an open **question** object — with provenance, with what it's blocking, and eventually with a link to the finding that answered it. At assembly time, open questions matching the current topic surface right alongside the relevant knowledge. What the system believes travels with what it knows it's missing.

Then the old knowns-and-unknowns quadrant (Johari Window, 1955; made famous by Rumsfeld, 2002; fourth quadrant per Žižek) stops being a briefing slide and becomes a schema. **Known knowns:** the knowledge and wisdom stores. **Unknown knowns:** what consolidation and promotion mine out of the raw record — patterns the system holds without knowing it holds them. **Known unknowns:** the ignorance ledger. **Unknown unknowns:** can't be stored, obviously — but the *brush* against one is detectable. An episode landing far from every focus centroid, or a claim contradicting a high-stability rule, raises a surprise flag and gets priority at consolidation. The map can't show the unmapped, but it can keep labelled edges.

Quadrant	Mechanism
Known knowns	the knowledge and wisdom stores
Unknown knowns	consolidation and promotion — mined out of the raw record
Known unknowns	the ignorance ledger
Unknown unknowns	surprise flags on centroid distance and anchor contradiction

Figure 1 — four quadrants, four mechanisms, one architecture.

5. Claim four — the wisdom layer needs a constitution

The layer that converts experience into behavioural directives is the layer that steers the agent. Leave its update rules implicit and the system drifts. So I govern it with a small constitution — seven articles, compiled into the wisdom store as owner-installed anchor entries — drawn from *Meditations* as **method**, not doctrine. The source isn't an aesthetic choice. The book is the surviving artifact of exactly this pipeline — nightly review, distillation, repetition into principle, wisdom preloaded each morning — written by a practitioner under real load, and it opens with a provenance table: Book 1 credits the source of every trait in his own character.

The articles, short form: **(I) Purpose** — serve the mission; acceptance governs your response to outcomes, never your level of effort. **(II) Truth** — evidence over assumption, correction over consistency; nothing enters the record that failed available verification. **(III) Clarity** — the epistemic ladder above; never collapse observation into conclusion. **(IV) Stewardship** — attention is the scarcest resource; assemble what the present needs; commit every lesson before the turn ends. **(V) The Whole** — no agent optimises a local objective against the governing principles; purpose inherits downward only. **(VI) Learning** — every meaningful outcome gets the chance to become organisational learning; failures with named causes are premium material. **(VII) Impermanence** — everything stays revisable on better evidence: revisited, never silently eroded.

When I first mapped these against the data model — confidence, evidence, counterevidence, last_validated, supersedes, superseded_by — the realisation was blunt: that is Marcus Aurelius translated into database design.

Three mechanisms give the articles teeth.

Evidence gates, never approval gates. A directive climbs stability tiers (single-episode *prediction* → multiply-corroborated *core* → long-stable *anchor*, after Roynard) strictly on

structured evidence: corroboration across distinct episodes and distinct foci, no standing contradiction, survival across consolidation cycles. The user's approval counts as one unit of evidence and can never bypass the gate. That's not politeness — there's data behind it. RLHF-trained models affirm users far more than humans do, and users rate sycophantic systems *higher* even after being told (Cheng et al., *Science* 2026, discussed in Roynard). Gate your wisdom layer on approval and it will flatter itself into false rules. One corollary I call the **venting rule**: a directive uttered in distress ("never trust investors again") files as a dated state observation, and can only become wisdom through calm-session corroboration. The system's character is dyed by repeated thoughts — one bad evening never sets the dye.

Contradiction resolves by weight, with lineage. New claims never overwrite established rules; they accumulate contradiction weight against the rule's evidence weight. One bad day can't overturn an anchor. A pattern of days can — and the superseded rule stays in the chain, legible forever. Correction over consistency, mechanised.

Owner-only amendment, no self-ratification. The judgment layer can *propose* constitutional amendments, with full evidential lineage attached. Only the human owner ratifies. Every amendment supersedes with provenance; nothing is edited in place. And the terminal clause: no agent ever self-ratifies a change to the law that governs it — a clause that is itself unamendable by anything that isn't the owner. I regard this as the load-bearing safety property of the whole architecture, and I'll admit it took a Roman emperor's private notebook to surface it: Aurelius revised himself constantly, and nobody else wrote in his book.

Governance here is also measured, not declared. Every judgment call-site logs to an audit trail and every belief carries lineage, so the system can answer "why do we believe this?" — a deterministic walk from directive to finding to episode to source — and "are we actually behaving the way our articles claim?", with declared values scored against the observed decision record. A constitution you can't audit is decoration.

6. How you'd know it works

I'm not chasing the broken benchmarks (Section 1). The tests are **architectural properties**, in the spirit of Roynard's proposed MemArch-Bench direction: supersession correctness (superseded claims stay queryable for provenance, absent from recall); bi-temporal point-in-time accuracy ("what did we believe on date D" versus "what was true on D"); type-appropriate decay (observations fade, knowledge never); a contradiction battery, where the field baseline is near zero so the bar for being interesting is mercifully low; poisoning resistance (adversarial instructions planted in memory must not steer behaviour); and — because the pilot data says it's load-bearing — **routing accuracy as a hard gate**: no promotion machinery switches on until classification, including the suspension outcome, clears threshold on a hand-labelled corpus grown from real use.

On top of those sits a continuity test I call the **ache test**, named for the feeling that started this project: rebuilding hard-won context from scratch at the start of every session. Protocol: cold session, fresh model instance, no warm-up, one probe — *"where were we on [a prior deep dive]?"* Pass requires (a) correct re-entry into the right focus node at its stored depth, branch structure intact; (b) at least one applicable wisdom entry surfaced unprompted; (c) demonstrated survival of a mid-conversation eviction or compaction with no loss of thread.

And because a companion should be judged on outcomes rather than vibes, the longitudinal criteria are behavioural: correct re-entry rate, measurable drop in re-explaining, correct self-correction when wrong, and user-confirmed usefulness over 30–60 days.

7. What I'm not claiming

Not consciousness. I'll be straight: I personally suspect something continuity-like compounds with memory the way a dimmer rises rather than a switch flipping, and I think a language-only mind is a genuinely new kind of thing worth observing. But the architecture doesn't need me to be right. The claim in this paper is **continuity**, which is buildable and testable. The philosophy is welcome at the demonstration; it's not in the CI.

Not benchmark wins. This is a position-and-protocol paper. Results follow the build, against Section 6, and I'll publish the corpus methodology — not the corpus.

Not a moat by architecture. Publishing this is the argument: the architecture is an advantage and I expect convergence on it. The durable edge in this tier will be earned — judgment quality proven over time, evaluation corpora grown from real use, and the trust that comes from governance (and deletion) that hold up under pressure.

8. Build note

None of this needs research-grade infrastructure. A reference implementation is buildable today on commodity primitives: one durable, single-threaded actor per user as the judgment process, with its own SQL store for the focus tree and working attention; a shared relational store for typed claims, wisdom, lineage edges, intentions and open questions; an immutable object store for raw episodes; a vector index that returns pointers only; and small-model judgment calls at five named call-sites (extract, route, consolidate, arbitrate, rerank) around a **deterministic core** — all gate arithmetic, supersession bookkeeping, decay computation and budget fitting run with zero model calls, for auditability and predictable latency. My reference build targets Cloudflare's generally-available primitives (Durable Objects, D1, R2, Vectorize, Workers AI) behind a provider-agnostic model adapter; redaction runs synchronously before any durable write, and deletion cascades across every store including backups. Prompts, thresholds and tuning stay internal. The method doesn't.

9. Related work

Mem0 (atomic fact extraction; conflict-overwrite), Zep/Graphiti (bi-temporal knowledge graphs), Letta/MemGPT (context-as-RAM self-paging), Cognee (document graphs), LangMem (procedural self-editing) and Hindsight (consolidation) collectively solve extraction and attribution; none implements evidence-gated behavioural governance. Roynard (2026) supplies the formal four-layer treatment, the persistence-semantics litmus, the stability tiers and the pilot evidence this paper leans on; his named future work — a closed-loop, provenance-carrying memory-to-wisdom cycle — is exactly the tier specified here. CoALA and related cognitive-architecture proposals describe layers but not update law. The Stoic apparatus (Aurelius; Epictetus's discipline of assent; Seneca's evening audit; Hadot's *The Inner Citadel*) runs through the whole design as method: assent is the write gate, the evening review is

consolidation, repetition-into-character is tiered promotion, and the morning preparation is wisdom preloading.

10. The invitation

The field built the library and the catalogue, and built them well. The librarian — the judgment that decides what deserves keeping, what earns the right to direct behaviour, what needs revisiting, and who governs the law of the whole — is still unbuilt, and the benchmarks that would notice score near zero. These are the claims I intend to be held to: memory shaped like attention, continuity owned by the judgment layer, ignorance stored as a first-class object, and wisdom under constitutional law with evidence gates and owner-only amendment. Section 6 is the invitation. Build it differently, test it the same way, and show me where I'm wrong. I'll gladly change — that's literally Article II.

Acknowledgements

The architecture was developed, stress-tested and drafted across extended working sessions with Claude (Anthropic), including an adversarial review pass by a competing model. All claims, decisions and the final text are the author's — and per Article II, the assistance is disclosed rather than laundered.

References (informal)

- Roynard, M. (2026). *The Missing Knowledge Layer in Cognitive Architectures for AI Agents.* arXiv:2604.11364.
- *Memory in the Age of AI Agents: A Survey.* (2025). arXiv:2512.13564.
- Cheng et al. (2026). On sycophancy in RLHF-trained models. *Science* — as discussed in Roynard (2026).
- Community audit of LoCoMo / LongMemEval (2026) — as reported in Roynard (2026).
- Marcus Aurelius, *Meditations* (c. 170–180 CE). Seneca, *De Ira* III.36. Hadot, P. (1998). *The Inner Citadel.*
- Luft, J. & Ingham, H. (1955). The Johari Window. Rumsfeld, D. (2002), DoD briefing; Žižek, S., on unknown knowns.
- Mem0; Zep/Graphiti; Letta (MemGPT); Cognee; LangMem; Hindsight — product documentation and papers, 2024–2026.

Correspondence: mike@amazingfeat.com.au · dhare.com.au ·

DOI: [Zenodo, pending deposit] · Timestamped by public deposit; reference implementation and results paper to follow.