Closing the Agent Loop

Open Problems at the Agentic Memory Wall

A Research Agenda for Governed Agentic Operations

Megan Anderson

AI ARMY, INC. · aiarmy.co

June 2026

Abstract

Agentic systems fail at scale in ways the field has named individually — context rot, memory poisoning, multi-agent token multiplication, sycophancy entrenchment. This agenda paper, the capstone of the Closing the Agent Loop series, argues that these are edge problems of a single undefined property: ***operational closure of the agent loop.***

We state the program's four positions:

- operational closure is not the same as task completion;
- context should be sovereign and portable at once;
- integrity is a property of the coupling between a model and its information and environment spaces;
- accountability must scale with capability for governed agentic operations,

and re-describe the field's named model degradation symptoms as failures of closure. We also pose seven open problems — the depth law of drift, measuring loyalty, ranker drift, the governance break-even, sequence sensitivity of handoffs, engineered forgetting with audit, and legible staleness as a consistency model — each stated falsifiably and mapped to the series paper and benchmark tier designed to test it. This is the research agenda and invitation paper of the series, not its proof paper: it claims that a class of failures currently studied separately can be studied as one, and that the proposed instruments make the claim testable.

Keywords: agent memory; context packets; governed agentic operations; operational closure; context drift; multi-agent systems; provenance; research agenda

---

The agent market spent the last two years naming its pain points. Context rot. Memory poisoning [10]. The multi-agent token multiplier [5]. Sycophancy that survives fine-tuning [11]. Every one of these has benchmarks, vendors, and a literature, and are being treated as a separate problem.

Our position is that they are edge problems of one thing the market has not defined properly: **t*he un-closed agent loop vs a closed agent loop*.** When the market says "agent loop," it often actually means the runtime task cycle — reason, plan, act, observe, perhaps even scheduled to recur — judged by whether a task got done. That is not the loop this series is named for. This series is about ***operational closure****:* whether agent work returns to a governed state where context, authority, evidence, outcome, and memory update are linked, reviewable, and accountable. ***Task loops complete. Governed loops close.*** Almost everything that goes wrong at scale in agentic operations goes wrong in the gap of understanding between those two sentences. Without understanding the constraints of the environment on model performance we cannot enable true accountability or governed agent operations.

Four core positions

1. Closure means something precise, and it hasn't been defined. Completion is a property of a task; closure is a property of an operation. An operation is closed when what was known, who authorized it, what was done, what resulted, and what memory now says are one traceable object. The series defines this and builds the grammar for it — because a market cannot converge on infrastructure for a property it cannot name.

2. Data should be sovereign and portable at the same time. We present the "context packet" as an object that refuses to trade one for the other. A context packet is a bounded, governed, portable assembly of context prepared for a specific recipient, task, and moment: *the right context, at the right time.* The corpus stays home; the projection travels. The same boundedness that makes a packet cheap makes it a privacy boundary; the same tracing that satisfies audit is the lineage that prevents drift from recurring. One primitive carries fidelity, security, and accountability at once — the market currently prices those as three products.

3. Integrity and governance live in the coupling — in the information space and the environment space, not in the model alone. Every model response is elicited from a unique state space, shaped by the prompt, the memory it was handed, the tools in reach, and the environment it runs in. This is why the same model answers a similar question differently for different users in different contexts: the answer was never a property of the model by itself. It is a property of the model *paired with* the information space it was given. The corollary is the program's design stance: drift, bloat, sycophancy, and stale authority are attacked by engineering constraints on what enters the loop — epistemic agentic memory, with status, provenance, supersession, stakes, and validity carried on the objects themselves — not by waiting for better models. A memory stream that silently drops contested markers manufactures agreement; a memory stream that preserves them gives any model, of any capability, the grounds for warranted correction.

***4. Humans stay in the loop for accountability; the loop stays traceable for auditability*.** Agents may read/and propose; but final promotion into a durable state is a governed, observable, human gated or policy gated act. Every claim traces to the packet that scoped the work that produced it. Judgment is continuously routed, not automated away, and the route is legible end to end.

Together these define the category the series builds toward: ***governed agentic operations,*** and the layer that enables them, ***governed agentic infrastructure.***

The symptoms, re-defined

Each pain point the field has named is what one part of an un-closed loop looks like from outside. ***Context rot*** is lossy re-encoding at handoff seams: the loop losing state each time it passes through a summary; long-horizon memory frameworks manage the storage problem [7][8][9] without governing what survives the seam.

Memory poisoning is an ungoverned write path: the loop accepting claims into authority without promotion; demonstrated attacks succeed at high rates with very small poison fractions [10].

The multi-agent token multiplier is broadcast in place of scoped projection: the loop paying transmission for context that carries no epistemics about what each worker needs, in the workload class where measured token volumes run an order of magnitude beyond chat [5].

Sycophancy entrenchment [11] has a structural accomplice in compression itself: selection by fluency, similarity, and recency [6][12] systematically drops the contested and the inconvenient, pre-agreeing the information space before the model ever speaks.

We see the different symptoms of one shared failure surface; and offer the grammar of repair: status, provenance [13], supersession, scope, and closure, carried on the objects that move through the agentic loop.

Seven open problems

These are the questions the program is instrumented to study. The series papers supply the definitions; the benchmark harness is being designed to supply the protocol, with task sets and scoring schema being prepared for release so that replication is possible.

1. The depth law of drift. Does context degradation compound multiplicatively with delegation depth while cost merely adds? and what is the exponent? If provenance-carrying packets flatten the curve, by how much, at what packing cost? *Instrument: the inference-economics paper's Tier 4 protocol (claim survival, attribution accuracy, contradiction preservation by depth) [4].*

2. Measuring loyalty. Accuracy, consistency, and precision have measurements; *loyalty* — purpose-conditioned retention of stakes, exclusions, and constraints — does not, because generic summaries carry no representation of intent. Can loyalty be standardized as a benchmark dimension, and does it predict downstream task validity better than the other three? *Instrument: the fidelity-per-token four-vector; Tier 1 [4].*

3. Ranker drift. Same store, same query, same budget, but a different packet passes with no explanation. Every memory vendor shipping learned retrieval is shipping this failure class, and it is essentially unstudied. How large is it in production systems, and does versioned,

recorded assembly (the Packing Reproducibility Invariant) eliminate it at acceptable cost? *Instrument: reproducible-packing metadata; a proposed draft extension to the Context Packet Specification [2].*

4. The governance break-even. Governance pays for itself when drift-repair plus fan-out duplication plus depth failure exceed packet-assembly plus overhead. Where does that crossover actually sit? as a function of session length, worker count, and chain depth? How does it vary across real workload shapes? *Instrument: Tier 3 protocol; the break-even inequality; pre-registered success and failure conditions [4].*

5. Sequence sensitivity of handoffs. Matched content, different delivery order: how order-dependent are agent outcomes, and does epistemic structuring reduce the dependence? The design is inherited methodologically from this program's adaptive-systems research; results in agent workflows carry no evidential weight for the cognitive claims, nor the reverse. *Instrument: the benchmark's cross-arm ordering condition [4].*

6. Engineered forgetting with audit. Systems must forget operationally without deleting historically; how to safely reduce the visibility of cold material while preserving provenance. What decay disciplines achieve this without silent loss, and how is "forgotten but auditable" verified? *Instrument: freshness, access-recency, stakes, and supersession state as first-class memory properties [3].*

7. Legible staleness as a consistency model. Sovereign local stores with cloud mirrors make eventual consistency unavoidable. The program's doctrine — perfect global immediacy is not required; *legible* staleness is — turns a distributed-systems weakness into a typed, visible property: a stale packet declares its as-of time, source snapshot, and invalidation boundaries. Is that sufficient for correct multi-agent operation in practice, and where does it break? *Instrument: the sync envelope and gated-inbox topology; the open protocol surface [2][3].*

Limits

This research agenda does not claim that governed loops or epistemic memory solves all agent failures, that context packets always reduce token cost, or that closure can eliminate all model errors. Tier 0 workloads — trivial single-session use — are explicitly outside the economics claim, and the break-even below which governance is not worth its overhead is treated as an empirical quantity, not assumed away. What the agenda claims is narrower and testable: that a class of agentic failures currently treated as separate symptoms can be studied as failures of operational closure, and that the proposed instruments make that claim testable.

The instruments

The series, in reading order:

1. *Closing the Agent Loop. Integrity is a Property of the Coupling & Trust Is Symbiotic with the Environment in Agentic Operations* [1] — defines the category and the closure claim.
2. *Context Packet Specification v0.4.0* [2] — defines the primitive that travels.
3. *Closing the Agent Loop. Governance Starts at Capture: An Epistemic Memory Architecture for Governed Agentic Operations* [3] — defines the epistemic agentic memory the loop writes into.
4. *Closing the Agent Loop. Epistemic Context Packets as an Inference-Economics Primitive: Valid Handoff and the Agentic Memory Wall* [4] — prices the loop, with valid handoff as the economic unit, and pre-registers the falsification conditions for the program.

First results will be reported as measured on our runtime; the harness, task sets, and scoring schema will be prepared for release. The field has named the pain points. This program names the shared failure surface, defines the instruments, and stakes the claim in a form built to be tested: ***governance is not overhead — it is how agentic systems preserve the conditions of valid action.***

References

[1] M. Anderson. "Closing the Agent Loop. Integrity is a Property of the Coupling & Trust Is Symbiotic with the Environment in Agentic Operations." AI ARMY, 2026.
[2] M. Anderson. "Context Packet Specification v0.4.0." AI ARMY, 2026.
[3] M. Anderson. "Closing the Agent Loop. Governance Starts at Capture: An Epistemic Memory Architecture for Governed Agentic Operations." AI ARMY, 2026.
[4] M. Anderson. "Closing the Agent Loop. Epistemic Context Packets as an Inference-Economics Primitive: Valid Handoff and the Agentic Memory Wall." AI ARMY, 2026.
[5] Anthropic. "How we built our multi-agent research system." Anthropic Engineering, June 2025.
[6] H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu. "LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models." Proceedings of EMNLP 2023. arXiv:2310.05736.
[7] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. "MemGPT: Towards LLMs as Operating Systems." arXiv:2310.08560, 2023.
[8] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav. "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory." arXiv:2504.19413, 2025.
[9] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef. "Zep: A Temporal Knowledge Graph Architecture for Agent Memory." arXiv:2501.13956, 2025.
[10] Z. Chen, Z. Xiang, C. Xiao, D. Song, and B. Li. "AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases." Advances in Neural Information Processing Systems 37 (NeurIPS 2024). arXiv:2407.12784.
[11] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, et al. "Towards Understanding Sycophancy in Language Models." The Twelfth International Conference on Learning Representations (ICLR 2024). arXiv:2310.13548.
[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang.

“Retrieval-Augmented Generation for Large Language Models: A Survey.” arXiv:2312.10997, 2023.
[13] P. Groth and L. Moreau, eds. “PROV-Overview: An Overview of the PROV Family of Documents.” W3C Recommendation, 2013.