Closing the Agent Loop

Epistemic Context Packets as an Inference-Economics Primitive: Valid Handoff and the Agentic Memory Wall

Megan Anderson

AI ARMY, INC. · aiarmy.co

2026

Abstract

Agentic systems have inherited an unexamined assumption: that context moves between turns, sessions, and agents either as raw transcript or as lossy summary. Both shapes fail, and they fail from the same root — the unit of handoff carries no epistemics. Nothing in a transcript slice or a summary records what is settled versus hypothesized, what supersedes what, what the stakes are, or what must not be silently dropped. This paper argues that in non-trivial agentic workflows the relevant economic unit is not the token but the **valid handoff**, and that epistemic context packets — scoped, bounded, provenance-carrying export objects — reduce the cost of valid handoff by preserving status, provenance, supersession, and budget-aware priority. The claim is tier-conditional and falsifiable: packets do not always save tokens, and we say where they do not. Governance pays for itself when the cost of drift repair, fan-out duplication, and depth-compounded context failure exceeds the overhead of governed packet assembly. We decompose fidelity into four measurable dimensions (accuracy, consistency, precision, loyalty), name the failure taxonomy that token metrics cannot see (context rot, attribution drift, branch confusion, contradiction erasure, stale authority, depth-compounded telephone failure, ranker drift), present a tier ladder locating where each part of the thesis applies, state the Constellation design invariants that make the claim testable (reproducible packing, ranker versioning, governed edges, return-path packet identity, budget-aware degradation, legible staleness), and specify a minimal falsification benchmark. The position throughout is operational rather than theoretical: the value of the claim rests on measurement, and the conditions under which it would be falsified are stated in advance.

Keywords: valid handoff; context packets; agent memory; inference economics; context drift; governed agentic infrastructure; multi-agent systems; agentic memory wall

---

1. Introduction: The Agentic Memory Wall

The economics of frontier AI are shifting from a race for capacity to a race for efficiency. The industry phase now underway is defined less by who trains the largest model than by who

makes deployed models produce the most completed work per dollar — the unglamorous serving-stack engineering of key-value cache reuse, quantization, in-flight batching, and query routing [8].

All of those optimizations live below the API, and they belong to the providers. There is exactly one optimization surface the customer owns: **what context enters each call.** For agentic workloads — autonomous agents performing multi-step tasks, where measured token volumes run roughly four times a chat interaction for a single agent and fifteen times for multi-agent systems [7] — that surface is not marginal. It is the dominant controllable cost, and it is almost entirely unengineered.

Computing has hit this shape of wall before. When processor speed outran memory bandwidth, the bottleneck moved: the scarce resource was no longer computation but the ability to feed it. The agentic version of that wall is now visible. Models are no longer the binding constraint on long-horizon, multi-agent work; **context integrity is**. Sessions degrade not because the model weakens but because the context it receives has been lossily re-encoded one time too many. Parallel workflows blow budgets not because workers think expensively but because every worker receives the same undifferentiated transcript. Deep delegation chains fail not at the model boundary but at the handoff boundary, where each hop re-encodes and each re-encoding forgets. We name this bottleneck the **agentic memory wall**: memory and context are the bandwidth problem; model intelligence no longer is.

Runaway recursion is the wall's maximal case, and it is instructive because it exposes the two visible symptoms as one disease. Passing bloated context down an unbounded chain is a compute catastrophe; passing drifted context down the same chain is a correctness catastrophe. At depth, they are the same failure with the same root: **unstructured context passing**. A system that solves only the cost half (compress harder) buys its savings with drift; a system that solves only the drift half (transmit everything) buys its fidelity with cost. The wall is not climbed from either side alone.

This paper proposes the third class: context handoff as a governed, epistemically structured operation, using the context packet primitive already specified in this research program, and it makes the economic argument that the earlier papers in this series imply but do not prove — that governed context is not overhead but efficiency infrastructure. The vocabulary of the argument, in four lines:

> The agentic memory wall is the bottleneck. Valid handoffs are the economic unit. Epistemic context packets are the primitive. Constellation is the architecture.

2. The Economic Unit: Valid Handoff

Token counting is the natural first metric for context economics, and it is the wrong one to optimize directly. Tokens measure what a handoff *costs*; they do not measure whether the handoff *worked*. A handoff that arrives cheap and wrong is not cheap: it is a deferred expense with interest, paid later as corrective loops, re-derivation, arbitration between contradicting workers, or — most expensively — confident downstream action on degraded ground.

We therefore define the unit this paper prices: a **valid handoff** is a transfer of context from one execution locus to another (turn to turn, session to session, orchestrator to worker, worker to sub-worker) after which the receiving locus can act without violating what has already been settled, decided, prohibited, or superseded. A handoff is invalid to the degree that acting competently on what was received still produces contradiction of settled items, reliance on superseded items, or loss of binding constraints. Invalidity is a property of the *transfer*, not of the receiving agent: the most capable model in the world cannot honor a prohibition that was summarized away.

Three cost terms attach to handoff, and only the first is visible in a token ledger:

1. **Transmission cost** — tokens moved, multiplied across fan-out and depth.
2. **Repair cost** — the tokens, turns, and human attention spent detecting and correcting the consequences of invalid handoffs. One corrective loop routinely exceeds the cost of any provenance metadata that would have prevented it; in an N-worker fan-out the same omission is repaired N times, plus arbitration; and fan-out itself carries an order-of-magnitude token multiplier before any failure occurs [7].
3. **Risk cost** — the expected cost of invalid handoffs that are *not* detected: actions taken on stale authority, decisions rebuilt on a branch that was already rejected. This term dominates precisely in the high-stakes settings where agentic systems are supposed to earn their keep.

The dominant *hidden* cost of context handoff is drift incurred and later repaired. Not the only cost — transmission genuinely dominates in wide fan-out, and Section 6 treats it directly — but the hidden one, because it appears in no token dashboard and compounds silently. The economics claim of this paper, stated once and priced throughout:

> **Governance pays for itself when:** drift-repair cost + fan-out duplication cost + depth-failure cost **>** packet-assembly cost + governance overhead.

The break-even point is an empirical crossover in session length, handoff count, worker count, and chain depth. It is measured in Section 10, not asserted. Below the crossover — a single consumer chat, a trivial one-shot task — governed packets are honestly not the cheap option, and this paper does not claim they are.

3. The Two Bad Poles

Current practice distributes almost entirely across two failure poles.

Transcript bloat. Re-transmit the raw record — the whole transcript, or as much as fits. Fidelity is high by construction and the cost structure is ruinous: transmission scales with session length, multiplied by worker count in fan-out, and most of the spend is redundant, since the receiving agent needs a small, task-specific fraction of what it is sent. Bloat also carries its own quiet fidelity failure: burying the binding constraint on page 40 of a transcript is operationally similar to omitting it.

Lossy syntactic compression. Summarize, or retrieve by embedding similarity, or prune by salience. The cost structure improves and the selection function is the problem: it is *syntactic*. Summarization and token-pruning keep what is statistically salient and fluent [9]; retrieval keeps what is geometrically near the query [10]; recency-weighted schemes keep what is new. None of these criteria track epistemic load-bearing-ness, and the divergence between "compresses well" and "must survive" is exactly where drift enters. The newest fluent claim silently wins over the older settled one. The superseded-but-load-bearing decision drops because it reads as resolved. The "contested" marker is lost because markers are short and summaries abstract them away. Provenance evaporates because provenance is metadata and compression optimizes content.

A property of this pole deserves its own name, because it connects context engineering to a failure mode usually discussed as a model behavior. Syntactic compression **can become structurally sycophantic**: it tends to preserve what is fluent, recent, and agreeable-shaped while dropping contested status, disagreement markers, exclusions, and provenance. The result is a memory stream that systematically under-supplies the model with grounds for warranted corrective friction — the record of what was disputed, what was rejected, what constraint still binds. A model fed such a stream does not need a sycophantic disposition to behave sycophantically; the information space it acts in has been pre-agreed for it. This is a structural claim about the compression pipeline, not a universal claim about summarizers, and it is testable: Section 10's drift metrics operationalize precisely the survival of contested-status and constraint items under budget pressure.

The two poles look like opposites and share one root: **the unit of handoff carries no epistemics.** Neither a transcript slice nor a summary knows what it contains in the sense that matters — settled versus hypothesized, current versus superseded, binding versus background. Both poles were adopted by default, not chosen by test. The position of this paper is that neither may be assumed to be the shape of agentic computing's future, and that the field has simply not yet built and measured the structured third class. The remedy for drift and its relatives is not to wait for better models; it is to **engineer constraints on the information space itself**, so that the failure modes have less room to occur.

4. Fidelity per Token

If tokens are the denominator, the numerator must be defined or the ratio is rhetoric. We decompose handoff fidelity into four dimensions, each with a named measurement, and we treat the fourth as the differentiator.

Accuracy. Claims in the handoff verify against source state. *Measurement:* claim-verification rate — sampled packet claims checked against the canonical record.

Consistency. The handoff contains no internal contradictions, and contradictions present in the source arrive *marked as contradictions* rather than silently resolved in favor of one side. *Measurement:* contradiction-detection audit against the source's tagged conflicts.

Precision. The handoff supports the receiving agent's task as well as the full source would, within the handoff's declared scope. *Measurement:* question-answering parity at budget — task-relevant questions answered from the packet versus from the full record.

Loyalty. The handoff adheres to policy and to the original intent of the content being preserved. This dimension exists because the first three can all pass while the handoff betrays its source: every sentence accurate, no contradictions, questions answerable — and the emphasis inverted, a prohibition dropped, a stakes marker gone, the purpose subtly re-aimed. Generic summarization cannot measure loyalty because generic summaries carry no representation of intent. Packets can, because intent is a first-class object in the primitive: purpose binding, stakes, exclusions are fields, not vibes. *Measurement:* purpose-conditioned retention — the survival rate of stakes-bearing, exclusion-bearing, and constraint-bearing items required by the packet's declared purpose.

Fidelity per token is this four-vector evaluated against the handoff budget. It converts the paper's comparative claim into a scoreboard: at matched budget, which selection function preserves more of what the next action requires?

5. Epistemic Context Packets as the Third Class

The primitive is already specified. The ***Context Packet*** Specification defines a scoped, bounded, provenance-carrying, non-executing export object: a packet declares purpose and recipient; carries items tagged along orthogonal axes of kind, epistemic status, and promotion state; records confidence and stakes; preserves supersession lineage and bitemporal validity (what was true of the world, and when the system recorded it); and represents exclusions explicitly, with escalation paths, so that what was withheld is itself legible. The packet cannot command; it can only inform, propose, and mark. A packet is not a tool, skill, plugin, policy engine, authorization grant, or executable runtime instruction: it carries governed state for a declared recipient and task, and authority remains outside the packet. [Context Packet Specification, v0.4.0.]

The observation this paper adds is that fields designed for governance and audit double as a **selection function**. Where compression asks *"what can I summarize, shorter?"*, epistemic selection asks:

> **"What must survive budget pressure for the next action to remain valid?"**

That question has answers a packet can compute and a summary cannot: the governing decision survives because it is tagged as a decision in force; the superseded plan is excluded *as superseded* rather than forgotten; the contested claim travels with its contested marker; the prohibition survives because stakes and exclusions are structurally protected; the receiving worker gets the projection its declared purpose requires and nothing else. Selection by epistemic criteria — load-bearing, current, in scope — rather than by compressibility.

The related-work boundary is worth drawing precisely, because parts of this territory are crowded. Prompt-compression methods optimize token counts against perplexity or task loss and are orthogonal to epistemics; they can be applied *inside* a packet's items [9]. Retrieval-augmented memory selects by similarity and inherits the syntactic pole's failure modes [10]. Bitemporal knowledge stores model fact validity over time but does not close a governance loop from packing through action to promote write-back [13]. Agent-memory frameworks manage storage tiers and retrieval but do not model supersession, contested

status, or purpose-bound projection as first-class [11][12]. Provenance itself has a mature vocabulary [6]; what is missing is its integration into the handoff unit. The packet primitive's distinguishing property is not any single field but the closed set: status, provenance, supersession, stakes, exclusion, and reproducibility in one non-executing object that a governed loop can be built around.

6. Three Economics Channels

The efficiency claim decomposes into three channels with different mechanisms, different measurements, and different buyers.

Channel A — direct token economics (fan-out). In orchestrator-to-workers topologies, the incumbent shape is transcript broadcast: N workers, N copies of context that is mostly irrelevant to each — the architecture whose measured cost multiplier is roughly fifteen times a chat interaction [7]. Scoped packets replace broadcast with per-worker projections — the right context to the right agent, without bloat — and the savings scale with N. A secondary mechanism compounds it: the packet's reproducibility discipline (Section 9) implies deterministic serialization, and deterministic serialization yields stable prompt prefixes, which is precisely the shape provider-side prompt caching prices at a discount [15]. A design decision made for auditability converts into cache-hit economics. Channel A is boring, direct, and measurable in an afternoon.

Channel B — drift economics (continuity). Turn-by-turn and session-to-session, the incumbent shape is periodic lossy compaction, and every compaction seam is a drift-injection site. Channel B's prediction is not that packets transmit fewer tokens per handoff — with provenance metadata they may transmit more — but that at *matched budget* they produce fewer drift incidents and fewer recovery turns, and therefore lower **total cost to completion**. The repair term dominates the metadata term. This is the channel where the hidden cost of Section 2 becomes visible, and it is measured by the failure taxonomy of Section 8, not by a token dashboard.

Channel C — boundary economics (sovereignty). The third channel is the one the market prices as three separate products. Enterprises decline to load sensitive customer data and business intelligence into external systems; individuals increasingly want personal data local. The packet's boundedness is simultaneously the answer to both and the token mechanism of Channel A: **the same scoping that minimizes tokens is the data-minimization guarantee.** In the Constellation topology (Section 9), the canonical store lives with its owner; agents read and write against a cloud mirror; bounded packets are what travels — purpose-bound projections, never the corpus. Cost, privacy, and exposure reduction are delivered by one mechanism that the market currently buys as an efficiency tool, a compliance aid, and an exposure-reduction layer respectively. We state the security claim with discipline: boundary separation reduces blast radius and is a *layer*, not a guarantee; topology without policy at the gate is only a delay.

7. The Tier Ladder: Where the Thesis Applies

The thesis is not one claim at all scales. Stated as one sentence: **privacy at Tier 0, fidelity at Tiers 1–2, economics at Tier 3, safety–economics convergence at Tier 4, and trust**

operations at Tier 5. Making the claim tier-conditional is not hedging; it is the difference between a measurable thesis and a slogan.

Tier	Workload shape	What the primitive demonstrates	Dominant failure mode
0	Consumer chat, single session	Sovereignty and locality only. **Explicitly outside the economics claim** — packing overhead can exceed savings and we say so.	—
1	Single agent, long horizon	Compaction replacement: epistemic selection at the seams where summaries currently rot context.	Context rot; summary drift
2	Session-to-session continuity, same workstream	Bitemporality, supersession, and decision memory across gaps.	Attribution drift; branch confusion
3	Orchestrator → N parallel workers	Scoped projection versus broadcast; cache-stable prefixes. The API-customer segment where agentic token volumes run orders of magnitude beyond chat.	Fan-out token multiplication; omitted-critical-context
4	Deep chains and recursion, depth D	Provenance-carrying packets as the anti-telephone mechanism: what arrives at depth 3 is traceable to what left depth 0. Degradation otherwise compounds multiplicatively with depth while cost merely adds.	Depth-compounded drift
5	Organizational shared memory: many humans, many agents, cross-vendor, persistent	Audit, compliance, sovereignty, identity, legible contribution — trust operations, the layer that has not commoditized.	Ungoverned shared memory as operational liability

Tier 4 deserves one further sentence, because it is where this paper's economics meets this program's safety thesis: in the recursive limit, the cost claim and the safety claim are the same claim. A chain that cannot afford to transmit valid context is a chain that cannot afford to act safely; both are properties of the handoff structure, not of the model.

8. Failure Taxonomy

Token metrics cannot see the failures that matter. We name them, because named failures can be counted, and counted failures can price the repair term of Section 2.

Context rot. Progressive degradation of a long-horizon session through repeated lossy re-encoding; each compaction seam loses structure the next turn needed.

Attribution drift. A suggestion, hypothesis, or working assumption is later recalled as a decision; an agent's proposal is recorded as the human's commitment; "who did what when" collapses into "what was said." Provenance collapses in the actor dimension.

Branch confusion. A contested, superseded, or alternate branch is promoted into the active line as if settled; brainstorm content resurfaces as fact. Provenance collapses in the lineage dimension.

Contradiction erasure. Two conflicting claims enter a handoff; one silently wins by fluency or recency; the conflict — which was information — is destroyed.

Stale authority. An expired permission, revoked decision, or superseded constraint continues to govern action because its supersession did not survive the handoff.

Depth-compounded telephone failure. At chain depth D, per-hop re-encoding losses compound multiplicatively; the claim that survives to depth D bears decreasing resemblance to the claim that left depth 0, with no per-hop event large enough to flag.

Ranker drift. A failure class introduced by memory systems themselves: same store, same query, same budget — different packet, with no explanation. If the assembly path (ranking policy, packing policy) changes silently, memory behavior becomes untraceable precisely where it is trusted most. The mitigation is structural and stated as an invariant in Section 9: every packet records the policy and ranker versions that assembled it. For an enterprise operator the requirement is intuitive: "same question, different memory answer" must always be explainable.

Benchmark variable: sequence sensitivity. Beyond the taxonomy, the benchmark manipulates one condition: matched content delivered in different handoff orders. We adopt **sequence sensitivity** as the agent-operational form of **ordering sensitivity**, borrowing the sequence-comparison design from this program's adaptive-systems lane, whose stated scope includes artificial agents [4][5]. The inheritance is methodological only: agent-workflow results carry no evidential weight for that program's claims, nor the reverse.

9. Constellation Design Invariants

Section 5 defines the handoff object; this section defines the memory architecture required to produce, reproduce, and close the loop around it. The claims above are testable only if that architecture guarantees certain properties. This section states the invariants in their public form; they are design commitments of the Constellation memory model, and they are what make the benchmark's comparisons well-posed. (Implementation specifics — ranking formulas, promotion thresholds, decay functions, policy defaults — are deliberately out of scope; Section 11.)

Architecture in brief. Constellation is a sovereign memory system for agentic operations: the canonical store lives with its owner (locally, where required); a cloud mirror hosts agent read/write over open protocol; a **gated inbox** stands between the mirror and canonical state. Knowledge is held as a graph whose nodes carry epistemic tags (status, confidence, stakes, promotion state, bitemporal validity) and whose links are epistemic rather than associative. An index layer (Atlas) keeps the whole legible: clusters weighted, histories intact, packets emitted as bounded projections of full-fidelity stars — the storage-side resolution of the fidelity-per-token question, since keeping everything and

transmitting little stops being a tradeoff. Identity is part of the system for humans and agents alike, so contribution is attributable by construction. The shape is deliberately familiar: version control for knowledge — branches legible, blame answered by provenance, history immutable — with one upgrade at the core: **the diff unit is epistemic, not textual.** There is no rebase; history is bitemporal and does not rewrite. A merge is not textual conflict resolution but two tagged claims coexisting with lineage until promotion arbitrates.

Invariant 1 — Reproducible packing. For the same purpose, recipient, authority boundary, as-of time, source state, packing-policy version, ranker version, and budget, the assembler produces the same packet — or a traceable explanation for divergence. Every packet carries the metadata to check this: identifiers for source snapshots, policy and ranker versions, budget, omission policy, and an assembly trace hash. One object yields auditability (what exactly was this agent told, and why), cacheability (stable prefixes), and drift control (assembly-path changes cannot hide).

Invariant 2 — Versioned ranking. The ranking policy that selects and orders packet content is versioned, and its version is recorded in every packet it assembles. This is the structural mitigation for ranker drift.

Invariant 3 — Governed edges. The public edge vocabulary is small, closed, and behavior-bearing: *supersedes, supports, contradicts, derived_from, depends_on* (implementations may extend privately). Edges receive the same write governance as nodes — agents propose links; promotion confirms them. An epistemic graph is not governed if only its nodes are governed: ungoverned edges are the back door through which syntactic association re-enters the system.

Invariant 4 — Return-path packet identity. Worker output returns carrying the identity of the packet that scoped it: the producing packet's ID, the source claims used, new claims and proposed edges — all landing in the gated inbox as *proposals*. Loop closure becomes a data structure rather than a policy: every claim in the mirror traces to the exact projection that produced it, and output is evaluated against the packet that scoped it — integrity as a property of the coupling, made mechanical.

Invariant 5 — Budget-aware graceful degradation. Under budget pressure a packet sheds in declared order. MUST survive: governing decisions, current status, supersession markers, safety constraints, provenance anchors, unresolved conflicts, task-critical commitments. SHOULD survive: supporting rationale, recent relevant work, high-value examples, related decisions. MAY yield: background narrative, low-stakes history, stylistic preference, redundancy. This ladder is the operational difference between a system that compresses and a system that **prioritizes**.

Invariant 6 — Engineered forgetting without silent deletion. Memory objects carry freshness, access recency, stakes, and supersession state, so the system can reduce operational visibility of cold material without destroying provenance or audit history. (This is the concrete home of an open problem named in the companion memory paper.) The specific decay and tiering functions are implementation-protected.

Invariant 7 — Legible staleness. Constellation does not require perfect global immediacy; **it requires legible staleness.** Distributed operation with a local-canonical store and a cloud mirror makes eventual consistency unavoidable — and acceptable, on one condition: a stale packet must declare its as-of time, its source snapshot, and its known invalidation boundaries. Inconsistency that is visible and typed is a manageable property; inconsistency that is silent is drift. A contradiction between concurrent writers is, by construction, not a race to resolve but two tagged claims with provenance, held until promotion arbitrates. (Full synchronization semantics — partition behavior, promotion-queue policy under outage, backup responsibility — are specified separately.)

The topology consequence is worth stating once, plainly: with the gated inbox, proposal-first promotion stops being only a policy and becomes **structured**. Agents cannot write canonical state; they can only propose into a mirror whose gate is governed. Policy and topology then enforce the same invariant independently — defense in depth against the memory-poisoning class of attack — demonstrated against production-style agent memory and RAG stores at high success rates with poison rates below a tenth of a percent [14] — in which a hostile write arrives not as an overwrite but as a quarantined, tagged, provenance-carrying proposal.

10. Benchmark Design and the Minimal Falsification Set

The thesis is falsifiable, and this section states how. Three tiers form the minimal set; each targets one channel; success and failure conditions are declared in advance.

Tier 1 — long-horizon single agent (Channel B). Identical long-form tasks run under (a) syntactic compaction and (b) epistemic packets at matched context budget. Metrics: drift incidents by taxonomy class, recovery turns, decision preservation, fidelity-per-token (the four-vector of Section 4), total cost to completion including packing cost.

Tier 3 — orchestrator → N workers (Channel A). Identical parallel tasks under (a) transcript broadcast and (b) per-worker scoped packets. Metrics: total tokens to completion including packing, per-worker token distribution, task accuracy, omitted-critical-context incidents, cache-hit rate on packet prefixes.

Tier 4 — recursive depth D (the anti-telephone test). Identical deep-chain tasks under (a) summary chaining and (b) provenance-carrying packet chaining. Metrics: claim survival by depth, attribution accuracy by depth, contradiction preservation by depth, correction cost when depth failures surface.

Benchmark condition across arms: sequence sensitivity — matched content, varied handoff order — to measure order-dependence of outcomes under each handoff regime.

Accounting rules. Packing cost is charged to the packet arm in full, including any model calls used in assembly. Retrieval-time governance evaluation is charged in full. Raw packet size is reported but is not a success metric; the comparisons are fidelity-per-token and total cost to completion.

Success condition. At matched budget, the packet arm meets or exceeds the compression arm on task success with fewer drift incidents (Tiers 1, 4), and beats the transcript arm on

total cost to completion beyond a measured break-even point (Tier 3), with the break-even crossover reported as a function of session length, worker count, and depth.

Failure condition. If the packet arm does not reduce drift incidents at matched budget, or if total cost to completion never crosses below the transcript arm within realistic workload ranges, the thesis as stated is falsified — not rescued by re-definition.

Tier 2 is evidenced by longitudinal case study rather than controlled benchmark; Tier 5 by architecture demonstration and build log; Tier 0 is scoped out of the economics claim entirely. All results in the first release are reported as **measured on our runtime and benchmark harness**; third-party replication is explicitly future work, and the harness skeleton, synthetic task sets, and scoring schema are prepared for release to make replication possible.

11. Open Specification, Protected Implementation

The research program this paper closes follows one release doctrine: **open the accountability grammar; protect the authority loop.** Applied here:

Published: the valid-handoff construct; the tier ladder; the failure taxonomy; packet metadata categories including the reproducibility fields; the edge-governance principle; the return-path packet-identity principle; the degradation ladder; the benchmark design, harness skeleton, and scoring schema; the general Constellation architecture and its invariants; the legible-staleness doctrine.

Protected: ranking formulas; promotion thresholds; decay and cooling functions; enterprise policy defaults; trust scoring; authority-grant logic; runtime enforcement design; operational workspace flows.

What is published is everything required to *evaluate, audit, and replicate the claims*; what is protected is the policy content whose disclosure would weaken the very governance being evaluated or is more sensitive protected IP in our own products. The reproducibility invariant is the hinge, a packet declares *which* ranker version assembled it without disclosing the ranker, exactly as a signed artifact declares its signer without disclosing the key.

The primitive itself remains open. The reproducible-packing metadata and the return-path structure defined here are proposed as a draft extension to the open Context Packet Specification (v0.5.0-draft), leaving the stable v0.4.0 release untouched — the specification's own governance process, exercised in public, as promised.

12. Limits and Non-Claims

Stated plainly, because the claim's strength depends on its boundaries.

1. **Not claiming universal token savings.** Packets carry metadata; on short or trivial workloads a lazy summary is smaller and adequate. Tier 0 is outside the economics claim by construction.

2. **Packing is not free.** Assembly may spend inference to save inference; the break-even is empirical and reported, not assumed.
3. **Not semantic perfection.** Epistemic selection reduces named failure classes; it does not guarantee the receiving agent acts well, and it does not eliminate model error. It changes what the model has to work with, not what the model is.
4. **Not a replacement for human review.** Promotion through the gated inbox is where human judgment lives; the architecture routes judgment, it does not automate it away.
5. **Boundary separation is a layer, not a guarantee.** Topology without governance at the gate is a delay, not a defense.
6. **First results are single-runtime.** Measured on our runtime; replication is future work, and the published harness exists to invite it.
7. **Methodological lineage is not evidential transfer.** The sequence-sensitivity condition inherits an experimental design from this program's adaptive-systems work; agent-workflow results carry no evidential weight for that program's claims, nor the reverse.

13. Conclusion: Governance as Efficiency Infrastructure

The agent market prices governance as a tax: the compliance line item, the overhead accepted for audit's sake. This paper has argued the opposite ledger. In non-trivial agentic workflows — wherever work is continuous, delegated, or recursive — the relevant economic unit is the valid handoff, and the dominant hidden cost is drift incurred and later repaired. Epistemic context packets lower the cost of valid handoff by construction: they preserve status, provenance, supersession, and budget-aware priority, and they do it with an object small enough to travel and structured enough to audit. The same boundedness that saves tokens in fan-out minimizes data exposure across trust boundaries; the same reproducibility that makes packets auditable makes them cacheable; the same provenance that satisfies compliance is the anti-telephone mechanism that keeps depth-D chains coherent. These are not three products; they are one primitive, priced three ways. Governed memory is not overhead. It is how agentic systems preserve the conditions of valid action — and past the break-even that this paper's benchmark measures rather than assumes, it is also simply the cheaper way to run the agent loop.

References

[1] M. Anderson. "Context Packet Specification v0.4.0." AI ARMY, 2026. [2] M. Anderson. "Closing the Agent Loop. Governance Starts at Capture: An Epistemic Memory Architecture for Governed Agentic Operations." AI ARMY, 2026. [3] 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. [4] M. Anderson. "The Temporal Neuroscience Index (TNI): An Information Geometry of Subjective Time." AI ARMY, 2026. [5] M. Anderson. "Supplement II — Adaptive Systems Formal Supplement." AI ARMY, 2026. [6] P. Groth and L. Moreau, eds. "PROV-Overview: An Overview of the PROV Family of Documents." W3C

Recommendation, 2013. [7] Anthropic. “How we built our multi-agent research system.” Anthropic Engineering, June 2025. [8] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. “Efficient Memory Management for Large Language Model Serving with PagedAttention.” Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP), 2023. arXiv:2309.06180. [9] 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, pp. 13358–13376. arXiv:2310.05736. [10] 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. [11] 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. [12] 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. [13] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef. “Zep: A Temporal Knowledge Graph Architecture for Agent Memory.” arXiv:2501.13956, 2025. [14] 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. [15] Anthropic. “Prompt caching with Claude.” Product documentation, 2024.