INTELLIGENT BY DESIGN. SECURE BY DEFAULT.

Persistent Organizational Context in Heterogeneous AI Systems

State-Transition Gating, Semantic Compaction, and Selective Reasoning

Matthew Blizzard

Armored Helix Systems LLC

VERSION	1.0
DATE	August 2026
TYPE	Technical White Paper / Reference Architecture
STATUS	Architecture and research hypotheses; empirical claims identified explicitly

ARMORED HELIX / PUBLICATIONS

Abstract

Enterprises increasingly operate across multiple large language models, AI assistants, coding agents, workflow agents, and domain-specific AI systems. Although these systems can be individually powerful, the organizational context produced through them remains fragmented. A model may understand a problem solved in one interaction while another model, employee, or business unit independently reconstructs the same context elsewhere.

This paper proposes a reference architecture for transforming heterogeneous AI-assisted and enterprise activity into a persistent, evolving representation of organizational work. Raw observations are normalized into canonical events, semantically interpreted into units of work, consolidated into durable workstream state, connected through typed relationships, and evaluated for material change before expensive reasoning is invoked. The architecture separates evidence preservation from reasoning representation and treats semantic compaction and state-transition gating as scaling primitives.

The central hypothesis is that the scalable unit of enterprise reasoning should be the *material change in organizational state*, not the raw interaction. Because lossy compaction and persistent memory can amplify both accidental error and adversarial poisoning, the architecture also introduces a **Microglial Integrity Layer**: a defensive control plane that preserves source authority and permission lineage, quarantines suspect derived state, triggers re-grounding or invalidation, and can force safety-critical review before downstream reasoning. If validated empirically, the combined approach can reduce repeated context transmission, improve continuity across model providers, surface cross-workstream duplication and contradiction, and make expensive agentic reasoning grow more slowly than raw AI activity without treating compacted state as automatically trustworthy.

Keywords: enterprise AI; organizational memory; persistent context; semantic compaction; agentic systems; knowledge graphs; heterogeneous LLMs; selective reasoning; event-driven architecture; provenance; memory poisoning; integrity controls.

ARMORED HELIX / PUBLICATIONS

Publication Note and Suggested Citation

This is an independent technical white paper authored by Matthew Blizzard and published by Armored Helix Systems LLC. Matthew Blizzard designs and engineers production AI systems and enterprise automation architectures. The paper is not presented as peer-reviewed academic research, and it does not claim empirical validation of the proposed scaling hypotheses beyond the cited related work.

Suggested citation:

Blizzard, M. (2026). *Persistent Organizational Context in Heterogeneous AI Systems: State-Transition Gating, Semantic Compaction, and Selective Reasoning*. Technical White Paper, Version 1.0. Armored Helix Systems LLC.

IP publication note. Public disclosure can affect patent rights and filing strategy, particularly outside the United States. This paper intentionally describes architectural principles while omitting implementation-specific schemas, scoring weights, thresholds, customer configurations, deployment recipes, and proprietary orchestration details. If patent protection is contemplated, publication should be reviewed with qualified intellectual-property counsel before public release.

Contents

SECTION 1

Problem Statement

Enterprise AI is becoming plural. A single organization may use multiple general-purpose LLM providers, coding assistants, internal agents, collaboration systems, service-management platforms, knowledge repositories, and operational tools. Each system can accumulate useful local context, but the organization as a whole does not necessarily retain that context in a durable, interoperable form.

The consequence is repeated reconstruction. Teams resend background, re-run investigations, rediscover dependencies, duplicate engineering, and reach conclusions that may conflict with work already occurring elsewhere. Larger context windows improve individual model sessions, but they do not by themselves create a durable organizational model spanning providers, tools, time, permissions, and business domains.

The architectural question is therefore not merely how to retrieve more documents. It is:

> **How can an enterprise maintain a continuously evolving representation of organizational work and allocate expensive AI reasoning only when that representation changes in a materially interesting way?**

SECTION 2

Core Thesis and Contribution

The central thesis of the proposed architecture is:

> **The scalable unit of enterprise reasoning should not be the raw interaction. It should be the material change in organizational state.**

This thesis yields six design principles:

1. **Persistent state.** Organizational context should survive individual conversations, sessions, and model providers.
2. **Progressive semantic compaction.** High-volume evidence should collapse into smaller representations of meaning before reaching expensive reasoning stages.
3. **Attention before reasoning.** Deep reasoning should be selectively invoked according to the materiality of state change.
4. **Bounded comparison.** New or changed workstreams should be compared against plausible candidates rather than the entire organizational history.
5. **Evidence-state separation.** Compact state should remain traceable to higher-fidelity evidence so important conclusions can be re-grounded, corrected, or invalidated.

6. **Integrity before propagation.** Derived state should not gain authority merely because it has been summarized, repeated, or transformed by another model. Suspect, low-confidence, permission-tainted, stale, or adversarial state should be containable before it can influence downstream reasoning.

The individual ingredients have substantial precedent: retrieval-augmented generation, hierarchical summarization, graph retrieval, persistent agent memory, temporal graph memory, and importance-triggered reflection are established areas of research and product development [3, 4, 5, 6, 7, 9, 11]. The proposed contribution is the systems architecture formed by applying these mechanisms around *persistent organizational workstream state* and *state-transition gating* across heterogeneous AI ecosystems.

SECTION 3

Benefits in Plain Language

3.1 Context can survive the model that created it

A useful discovery made through one authorized AI system can become organizational state rather than remaining trapped inside that provider's conversation history. The architecture does not require copying every transcript into every other model. It attempts to preserve the useful meaning of the work in a provider-neutral state representation.

3.2 The enterprise can stop resending the same context

Without compaction, organizations repeatedly provide old conversations, tickets, documents, and background explanations to new model calls. Persistent state creates a smaller representation of what is currently believed to matter, while retaining links to evidence for re-grounding.

3.3 Different AI providers can contribute to the same organizational memory

Where enterprise APIs, logs, exports, or approved telemetry paths exist, activity from multiple AI providers and enterprise applications can contribute to shared organizational context. This allows model choice to remain flexible while the organization retains its state.

3.4 Duplicate and contradictory work can become visible

Two teams may independently solve the same problem or maintain incompatible assumptions. Persistent workstream state and typed relationships make these situations explicit candidates for detection rather than relying on chance human discovery.

3.5 Reasoning can be reserved for meaningful changes

A hundred interactions may reinforce an existing workstream without materially changing it. Those interactions should not automatically trigger a hundred expensive multi-step reasoning jobs. The architecture attempts to route deep reasoning only when state transitions appear novel, contradictory, overlapping, consequential, or otherwise material.

3.6 Persistent memory can be defended as a first-class system

A persistent context layer is also a new attack surface. The architecture therefore treats semantic-state integrity as a separate responsibility rather than assuming every compacted summary is safe. Suspicious state can be quarantined, re-grounded against source evidence, restricted by permission lineage, or invalidated before it reaches downstream reasoning agents.

3.7 Models become more replaceable

If organizational memory lives only inside a provider's application, provider changes can imply context loss. If organizational state is externalized, models can compete on quality, latency, specialization, security, and price without becoming the sole system of memory.

SECTION 4

Reference Architecture

Figure 1 shows the reference flow. Authorized activity enters through an observation plane, normalizes into canonical events, and resolves into persistent state. A distributed Microglial Integrity Layer operates across state writes, retrieval, and invalidation so derived context can be checked for provenance, authority, permission lineage, contradiction, staleness, and poisoning risk before it propagates. The systems-level Attention layer then determines which accepted or risk-elevated changes justify specialized reasoning. High-fidelity evidence and long-term knowledge remain available for re-grounding.

The design can be understood as six planes:

1. **Observation plane** - authorized activity from AI providers, enterprise systems, and telemetry sources.
2. **Interpretation plane** - semantic transformation from raw events into units of work.
3. **Persistent state plane** - durable workstreams, evidence lineage, semantic indexes, graph relationships, and long-term knowledge.
4. **Integrity control plane** - Microglial checks that bind source authority, preserve permission lineage, quarantine suspect state, trigger re-grounding, and invalidate affected descendants when evidence is revoked or corrupted.
5. **Selective reasoning plane** - specialist reasoning invoked only when a state transition warrants investigation or an integrity/risk override requires review.
6. **Controlled execution plane** - recommendations or actions subject to policy, provenance, and human authority where appropriate.

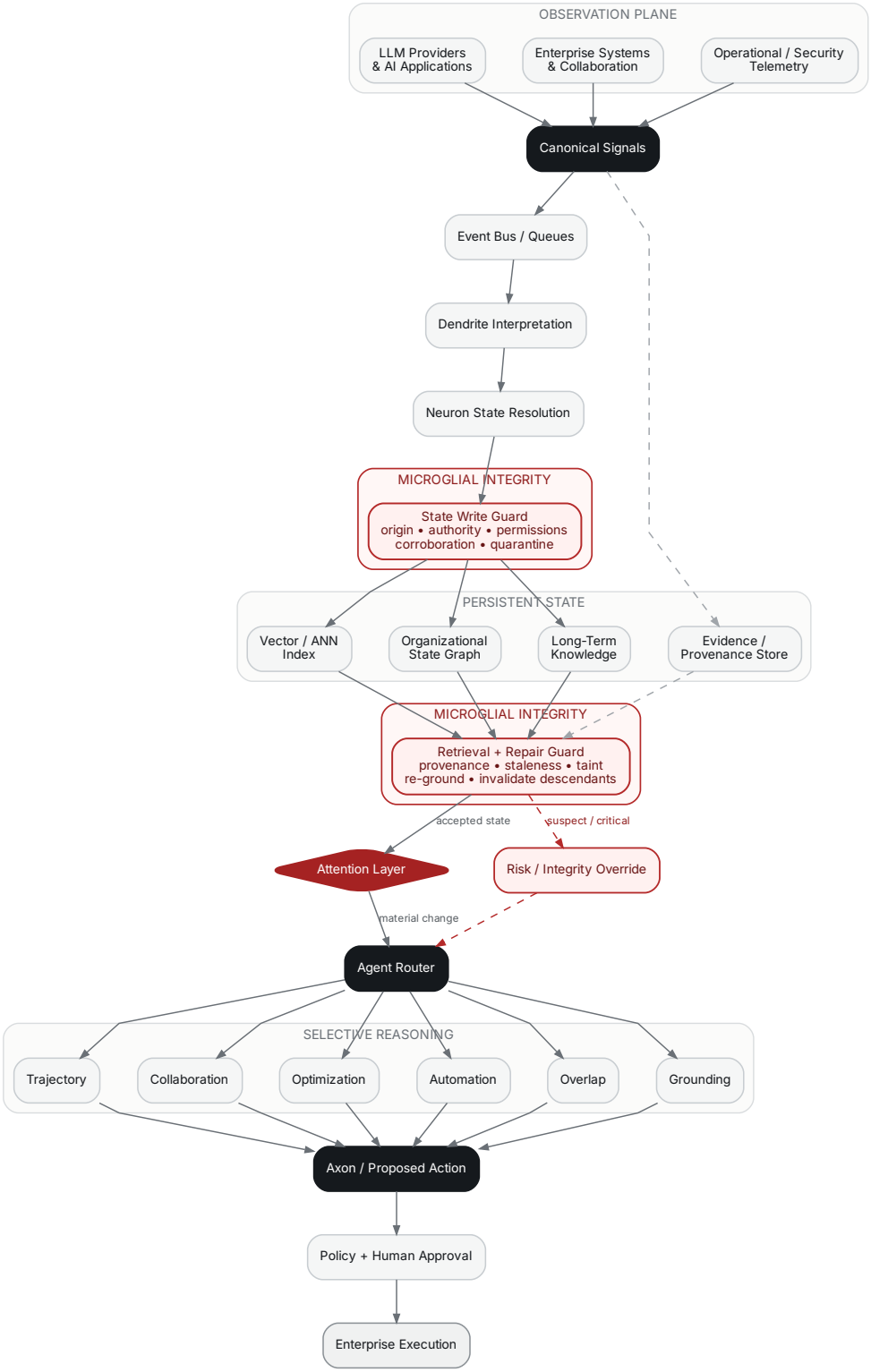

Figure 1: Reference architecture for persistent organizational context and selective reasoning.

SECTION 5

Canonical Semantic Model

The architecture uses a compact ontology to distinguish evidence from increasingly interpreted state. The neurological terms below are conceptual labels, not claims of biological equivalence.

5.1 Signals: what happened?

A **Signal** is the canonical representation of an authorized observation. Potential sources include AI interactions, agent runs, ticket changes, document events, code activity, collaboration events, operational telemetry, and security events. Provider-specific formats are normalized so downstream components do not depend on every source's native schema.

5.2 Dendrites: what did the activity mean?

A **Dendrite** represents an interpreted unit of work derived from one or more Signals. Several prompts, tool calls, edits, and responses may collectively represent one coherent activity such as investigating an authentication failure or designing a deployment pattern. This is the first major semantic-compaction boundary.

5.3 Neurons: what are we working on?

A **Neuron** represents a persistent workstream. It may correspond to a project, investigation, research program, technical initiative, vendor evaluation, recurring issue, or other durable objective. New Dendrites may reinforce the current state, resolve uncertainty, introduce conflicting evidence, or materially change the workstream.

The key distinction is between *new activity* and *materially changed state*. The architecture is designed so that many new observations can update a workstream without forcing deep reasoning on every event.

5.4 Synapses: how is work related?

A **Synapse** is a typed relationship between persistent workstreams. A minimal useful vocabulary includes:

- **SUPPORTS** - one workstream provides evidence or capability that strengthens another;
- **CONTRADICTS** - workstreams contain materially incompatible evidence, assumptions, conclusions, or requirements;
- **DUPLICATES** - workstreams substantially repeat the same effort;
- **SUPERSEDES** - newer work renders older work materially obsolete.

Relationships should be temporal and provenance-linked rather than treated as permanent truth.

5.5 Brain Regions: where should reasoning occur?

A **Brain Region** is a domain or organizational scope such as cybersecurity, infrastructure, research, finance, legal, a program, or a product area. Regions help constrain retrieval, policy, and candidate comparison. They are therefore both semantic and computational constructs.

5.6 Microglia: is the state safe to propagate?

Microglia is the architectural label for a defensive integrity layer surrounding persistent semantic state. The term borrows from the surveillance and maintenance role of biological microglia but does not imply biological equivalence. Microglia is not intended to be another unconstrained reasoning agent. It is a control plane composed primarily of deterministic policy, provenance, authorization, dependency, anomaly, corroboration, and validation mechanisms, with model-assisted checks used only where useful.

Its primary question is:

> **Can this derived state safely acquire influence over downstream organizational reasoning?**

Microglia can accept state, quarantine it, lower its confidence or authority, require independent corroboration, re-ground it against evidence, restrict its retrieval scope, or invalidate dependent state already derived from it.

5.7 Axons: what should happen next?

An **Axon** is intelligence leaving the architecture: a recommendation, alert, ticket update, generated artifact, automation proposal, or downstream action. High-impact Axons should be policy-controlled and, where appropriate, require human approval.

SECTION 6

Cross-Provider Context Persistence

Figure 2 illustrates the core interoperability benefit. Multiple providers and enterprise systems can contribute authorized observations to a shared semantic state. A later reasoning task can retrieve relevant compact state and evidence without inheriting the entire transcript history of the model that originally produced the insight.

This architecture is intentionally additive rather than universal. It cannot observe data a provider does not make available through an authorized integration, export, API, telemetry path, or customer-controlled logging mechanism. A provider outside those boundaries simply contributes less or no context to the organizational state model.

The practical advantage is not that every model knows everything. It is that an organization can preserve useful context independently of whichever model happened to create it.

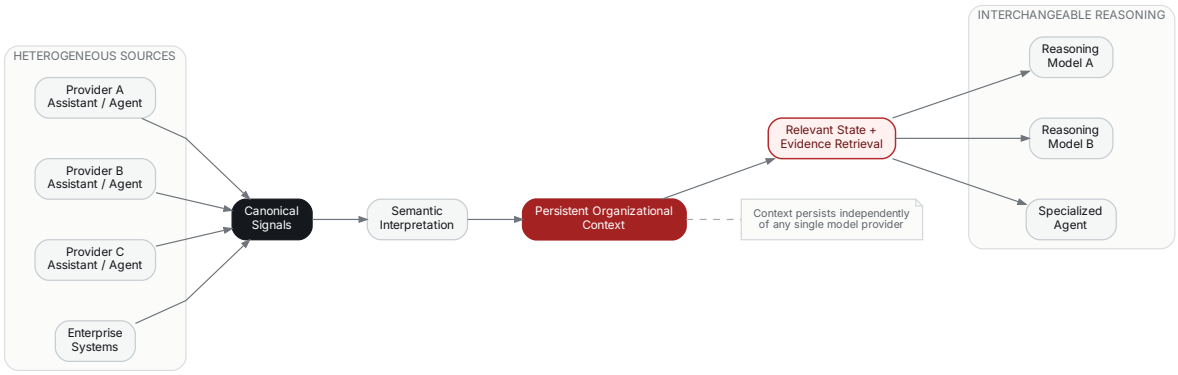

Figure 2: Provider-neutral context persistence. Organizational context remains external to any single reasoning model.

SECTION 7

Progressive Context Compaction

Figure 3 shows the intended information funnel. Raw activity should become increasingly compact before it reaches increasingly expensive reasoning layers.

Conceptually:

$$S \gg D \gg \Delta N \gg A \gg R \tag{1}$$

where:

- S = raw Signals,
- D = semantically meaningful Dendrites,
- ΔN = material Neuron state changes,
- A = Attention events,
- R = expensive reasoning executions.

The exact ratios are workload dependent and must be measured. The architectural objective is that increases in raw activity do not require proportional increases in deep reasoning.

Hierarchical and compressed-context approaches already have significant precedent. RAPTOR recursively clusters and summarizes text at multiple abstraction levels [6]; persistent-memory systems such as Mem0 dynamically extract and consolidate salient conversational information rather than carrying full history indefinitely [11]. The present architecture applies the broader principle to changing organizational workstream state.

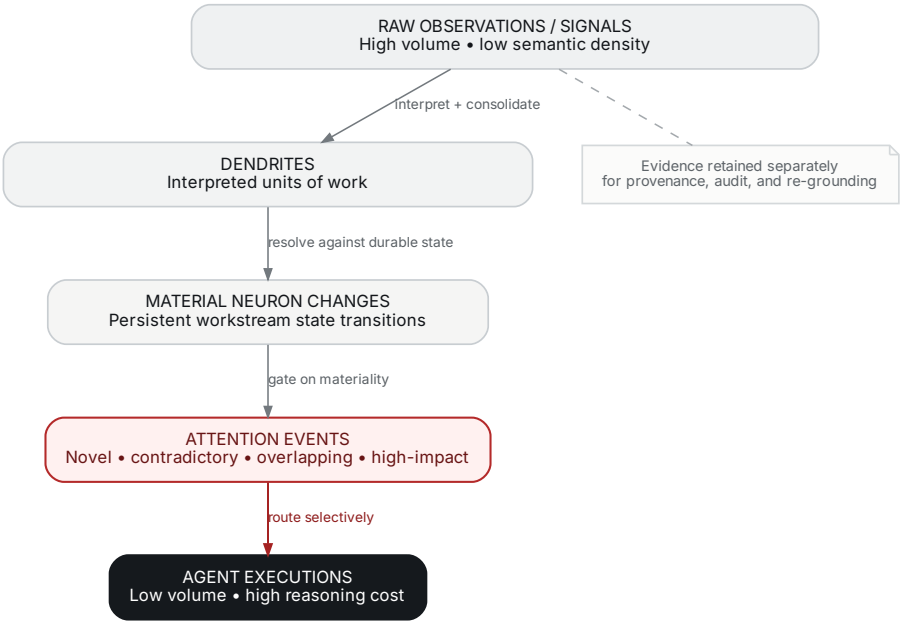

Figure 3: Progressive context compaction. Evidence is retained separately while the active reasoning representation becomes progressively smaller and more semantic.

SECTION 8

Microglial Integrity Layer

Persistent semantic memory introduces a failure mode that stateless query systems can often avoid: an error or attack can survive the interaction that created it and influence unrelated reasoning later. Poisoning research in RAG and persistent-agent memory demonstrates that relatively small malicious insertions can manipulate retrieval and later model behavior, while recent work also highlights the risk that summarization, tool echoes, or manufactured corroboration can launder an untrusted origin into apparently trusted memory [12, 13].

The **Microglial Integrity Layer** is proposed as a defensive control plane for this problem. Its objective is not to guarantee that every poisoned or incorrect item is detected. Its objective is to prevent compacted state from becoming *more authoritative merely because it has been transformed, repeated, or persisted*, and to make suspicious propagation detectable and reversible.

8.1 Origin and authority binding

Every derived object should retain the origin and authority characteristics of the evidence from which it was produced. A pure transformation such as summarization should not independently increase authority. Conceptually, for a transformation T applied without new independent validation:

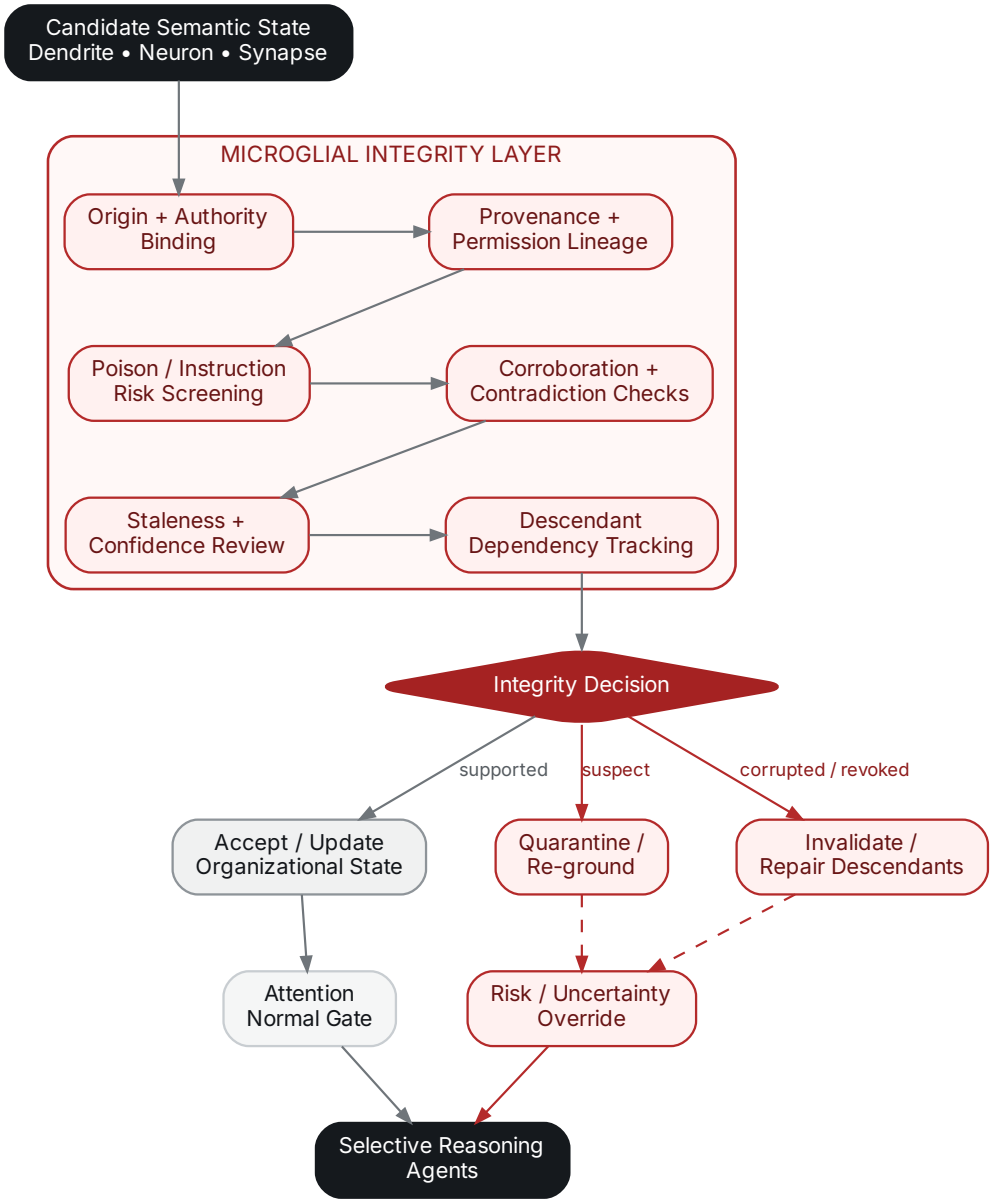

Figure 4: Microglial integrity controls around semantic-state writes and downstream propagation. Suspect state can be quarantined, re-grounded, or invalidated; risk and uncertainty can also bypass the normal Attention gate.

$$Authority(T(x)) \leq Authority(x) \tag{2}$$

Authority may increase only through an explicit validation process, such as corroboration from sufficiently independent trusted evidence, deterministic verification, or authorized human confirmation. This prevents model repetition from becoming self-corroboration.

8.2 Permission taint and compartmentalization

Derived state should carry conservative access lineage. A safe default is that the accessible audience for derived state is no broader than the audiences authorized to access the sources that materially support it, unless an explicit redaction or declassification process creates a lower-sensitivity derivative. Brain Regions can therefore function not only as semantic scopes but also as security compartments.

8.3 Quarantine before durable influence

Low-confidence, provenance-broken, instruction-like, adversarial, anomalous, or materially contradictory state can be placed in quarantine rather than immediately updating the authoritative organizational representation. Quarantine should preserve evidence and lineage while withholding normal downstream influence until re-grounding or review occurs.

8.4 Corroboration must be independent

Repeated model outputs, summaries of the same source, and agent-generated restatements should not be counted as independent evidence. Corroboration should preferentially rely on genuinely separate observations, authoritative enterprise systems, deterministic checks, or approved human validation. This is important because persistent-memory attacks can exploit apparent corroboration generated from the same poisoned origin [13].

8.5 Continuous repair and descendant invalidation

Microglia also operates after state has been accepted. If supporting evidence is deleted, revoked, contradicted, or later identified as poisoned, dependency lineage can be used to locate affected Dendrites, Neurons, Synapses, findings, and Axons. Those descendants can then be revalidated, downgraded, quarantined, recomputed, or invalidated.

8.6 Microglia is a control plane, not a trust oracle

No classifier or model can reliably identify every poisoned state in advance. For that reason, Microglia should combine prevention, containment, provenance, conservative authorization, independent validation, and repair rather than rely on a single “poison detector.” Model-assisted screening may be useful, but deterministic policy and origin-bound controls should carry as much of the security burden as practical.

SECTION 9

Attention as a State-Transition Gate

The **Attention** layer is a systems-level decision boundary between routine state maintenance and deeper reasoning. It is distinct from the internal attention mechanism of transformer models.

A conceptual function may be expressed as:

$$A(x) = f(Novelty, Contradiction, Overlap, Impact, Confidence, Context) \tag{3}$$

The implementation need not reduce these factors to a single scalar score. Different domains may use deterministic rules, embedding distance, anomaly detection, graph structure, model-based classification, risk metadata, confidence, or learned feedback.

The required architectural property is gating. Without it, reasoning demand tends toward raw activity volume. With effective gating, reasoning demand can move toward the volume of materially interesting state changes.

Importance-triggered reflection itself is not novel. Generative Agents, for example, used recency, importance, and relevance in memory retrieval and generated higher-level reflections as accumulated importance crossed thresholds [4]. The distinction proposed here is the level of operation: the gate evaluates changes to shared organizational workstream state rather than merely when an individual conversational or simulated agent should reflect on its own experience.

9.1 Attention must not be a single brittle threshold

A single scalar threshold creates an obvious recall risk: low-signal but high-impact anomalies may never cross it. The proposed architecture therefore supports multiple escalation paths in parallel with ordinary materiality scoring:

- **risk override** - security, safety, legal, or domain-critical indicators can force review even when general novelty is low;
- **uncertainty override** - unusually low confidence or internally inconsistent state can be escalated rather than silently discarded;
- **Microglial override** - quarantine, provenance failure, suspected poisoning, or descendant invalidation can force a review path;
- **periodic sweep** - dormant or low-signal state can be re-evaluated on a bounded schedule so importance is not defined only by instantaneous event magnitude;
- **audit sampling** - a small sample of below-threshold events can be reasoned over in shadow mode to estimate false-negative rates and recalibrate the gate.

This makes Attention a policy-and-detection layer rather than a single cost-cutting cutoff.

SECTION 10

Selective Agentic Reasoning

Agents are treated as a specialist reasoning tier rather than the entire architecture. Candidate roles include:

- **Overlap reasoning** - determine whether workstreams duplicate or materially reinforce one another;
- **Trajectory reasoning** - detect acceleration, stagnation, divergence, or major directional change;
- **Collaboration reasoning** - identify situations where coordination could reduce duplicated effort;
- **Optimization reasoning** - identify reuse, simplification, or process-improvement opportunities;
- **Automation reasoning** - identify recurring work suitable for controlled automation;
- **Grounding reasoning** - validate important conclusions against authorized evidence and long-term organizational knowledge.

Not every Attention event should invoke every agent. Routing should remain bounded and purpose-specific.

SECTION 11

Bounded Relationship Discovery

A persistent organizational context system creates an obvious computational challenge. If n workstreams exist, naive pairwise comparison approaches $O(n^2)$. That is not the proposed default architecture.

Figure 5 illustrates a bounded candidate pipeline.

A changed workstream can first retrieve a limited candidate set:

$$C(N_i) = ANN(N_i, k) \tag{4}$$

and reduce it further:

$$C'(N_i) = C(N_i) \cap Region \cap Domain \cap Time \cap Policy \tag{5}$$

Graph-neighborhood retrieval can then supplement semantic search. Deep comparison occurs only across the bounded candidate set.

This division of labor is deliberate: vector retrieval is useful for discovering plausible but previously unconnected similarity; graph retrieval is useful for traversing established relationships and dependencies. GraphRAG [7], HippoRAG [8], and temporal graph-memory systems such as

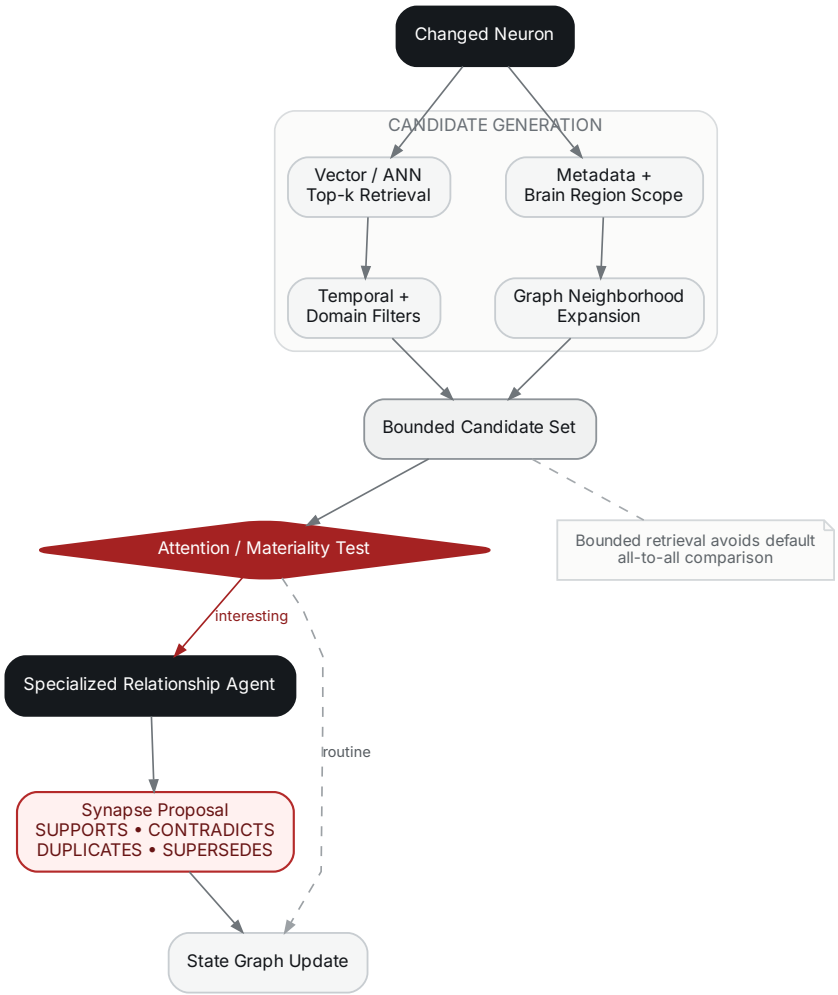

Figure 5: Bounded relationship discovery using semantic retrieval, scope, temporal filters, and graph neighborhoods before expensive reasoning.

Zep/Graphiti [9] demonstrate the broader value of graph and hybrid retrieval. The proposed use here is specifically relationship discovery among evolving organizational workstreams.

SECTION 12

Scaling Model

The architecture should not be described as infinitely scalable. Practical ceilings include cloud quotas, provider APIs, model throughput, vector-index workload, graph topology, storage volume, latency objectives, and cost.

Its scaling strategy is instead to make the highest-volume work comparatively cheap and parallelizable:

- ingestion is asynchronous and queue-driven;
- interpretation workers can remain stateless and horizontally scalable;
- durable state lives outside workers;
- provider-specific throttling is isolated;
- candidate retrieval is bounded;
- domain and temporal scoping limit the working set;
- routine state updates do not automatically trigger expensive agents;
- expensive reasoning is allocated selectively;
- Microglial integrity checks are tiered so high-volume deterministic validation can remain inexpensive while deeper re-grounding is reserved for suspicious or consequential state.

Let S be raw Signal volume and R be expensive reasoning executions. The design objective is:

$$Growth(R) < Growth(S) \tag{6}$$

for workloads in which much new activity reinforces or incrementally updates existing state. This is a testable hypothesis, not an established performance result.

12.1 Reasoning economics

Let E be the number of expensive agent executions and $\bar{c}$ the mean execution cost. Then:

$$C_{reasoning} \approx E \cdot \bar{c} \tag{7}$$

The architecture seeks $E \ll S$ while preserving useful detection coverage. Lower model spend is only one benefit. Avoiding unnecessary reasoning also reduces latency, false findings, alert fatigue, and opportunities for model error.

SECTION 13

Security, Governance, and Provenance

Persistent organizational context is valuable precisely because it can connect information across systems. That same property creates security and governance risk.

13.1 Understand work, not workers

The system should be architected to reason primarily about workstreams, relationships, evidence, dependencies, and organizational execution. It should not default to employee productivity scoring, covert behavioral ranking, ideological profiling, or unrestricted browsing of raw prompts.

13.2 Derived knowledge must retain access lineage

If a state representation is derived from restricted sources, the derived representation cannot automatically become globally visible. Permission-aware memory is therefore a state-layer concern, not merely a storage concern.

A production design may require source-level access metadata, policy tags, purpose restrictions, domain boundaries, runtime authorization, and permission intersection or other conservative derivation rules. Asana's current AI architecture publicly emphasizes shared memory and governance around its Work Graph, and Atlassian similarly exposes connected work context through Teamwork Graph while adding scope and audit controls [15, 16]. These adjacent systems reinforce the importance of permission-aware enterprise context.

13.3 Evidence, interpretation, inference, and action must remain distinguishable

A raw observation is not the same thing as an interpretation. An interpretation is not the same thing as a validated conclusion. A recommendation is not the same thing as an authorized action.

Important transformations should retain sufficient provenance to answer:

- What evidence supported this state?
- Which interpretation changed it?
- Which relationship or agent finding followed?
- Was the conclusion validated or merely inferred?
- Who or what authorized the resulting action?

13.4 Security compartmentalization should survive semantic compaction

The system should not collapse the enterprise into one unrestricted "company brain." Brain Regions, source permissions, tenant boundaries, and purpose restrictions should constrain both storage and inference. Microglial controls preserve this compartmentalization by propagating access lineage into derived state and treating cross-boundary derivation as a policy event rather than a normal consequence of summarization.

13.5 Deletion and invalidation must propagate

If underlying evidence is deleted, corrected, or becomes inaccessible, dependent Dendrites, Neurons, Synapses, findings, or Axons may need to be deleted, invalidated, recomputed, or marked unsupported. A durable-memory architecture that cannot propagate invalidation is not enterprise-ready.

13.6 Organizational context is more sensitive than its parts

A derived state graph can reveal relationships that no single source exposes. This creates concentration risk. Encryption, least privilege, auditability, tenant isolation, purpose restrictions, and restricted inference are therefore necessary but not individually sufficient; the semantic state itself must remain governed.

SECTION 14

Evaluation Framework

The architecture is deliberately framed so its core propositions can be measured.

14.1 Compaction metrics

Signal-to-Dendrite compaction:

$$CD = \frac{Signals}{Dendrites} \tag{8}$$

Dendrite-to-material-state compaction:

$$CN = \frac{Dendrites}{Material\ State\ Changes} \tag{9}$$

14.2 Attention selectivity

$$AS = \frac{Attention\ Events}{Material\ State\ Changes} \tag{10}$$

14.3 Reasoning invocation ratio

$$RIR = \frac{Agent\ Executions}{Signals} \tag{11}$$

14.4 Useful finding rate

$$UFR = \frac{Human\ Accepted\ Findings}{Agent\ Findings} \tag{12}$$

14.5 Relationship quality

Human-reviewed datasets can measure precision and recall for SUPPORTS, CONTRADICTS, DUPLICATES, SUPERSEDES, or later relationship types.

14.6 Organizational utility

Longer-term outcomes may include duplicated effort avoided, contradictions surfaced, reusable work discovered, automation opportunities accepted, time saved reconstructing context, repeated context tokens avoided, and mean model cost per accepted finding.

14.7 Integrity and poisoning-resilience metrics

Microglial controls should be evaluated as a security and reliability system rather than described as a qualitative safeguard. Useful measurements include:

- poisoning-detection and quarantine precision/recall under controlled adversarial inputs;
- false-quarantine rate for legitimate novel work;
- percentage of poisoned or invalidated state prevented from reaching downstream agents;
- mean propagation depth before containment when poisoning is discovered retrospectively;
- permission-leakage rate across derived-state retrieval tests;
- time to invalidate or recompute descendants after source revocation;
- rate at which model-generated restatements are incorrectly treated as independent corroboration.

14.8 Attention false-negative metrics

Attention should be evaluated for missed critical events, not only selectivity. Controlled evaluation should include low-frequency, high-impact anomalies and measure how often normal gating, risk overrides, Microglial overrides, periodic sweeps, or audit sampling successfully surface them.

SECTION 15

Research Hypotheses

The proposed architecture yields the following testable hypotheses:

H1 - Semantic compaction.
Raw interaction growth produces slower growth in materially changed workstream state.

H2 - Selective reasoning.
State-transition gating substantially reduces expensive reasoning executions relative to indiscriminate reasoning.

H3 - Intelligence density.
Findings generated after Attention filtering have a higher useful-finding rate than findings generated without gating.

H4 - Hybrid retrieval.
Combined vector and graph candidate retrieval produces better relationship discovery than either mechanism alone for evolving workstreams.

H5 - Context efficiency.
Persistent workstream state reduces historical context tokens required for long-running tasks relative to full-history reconstruction.

H6 - Provider continuity.
Relevant context created through one authorized AI provider can improve later reasoning performed through another provider without transferring complete transcript history.

H7 - Sublinear reasoning growth.
For suitable workloads, expensive reasoning demand grows materially slower than raw Signal volume.

H8 - Integrity containment.
Microglial controls materially reduce the probability that poisoned, provenance-broken, or invalid state influences downstream reasoning relative to an otherwise equivalent persistent-memory architecture without an integrity layer.

H9 - Attention safety.
Risk, uncertainty, Microglial, periodic, and audit-sampling override paths reduce the false-negative rate for low-signal/high-impact events relative to a single-threshold Attention gate at comparable reasoning budgets.

SECTION 16

Related Work and Positioning

16.1 Organizational memory

The idea of technologically augmented organizational memory predates modern LLMs. Stein and Zwass described organizational memory information systems in terms of acquiring, retaining, maintaining, searching, and retrieving organizational information [1]. Ackerman and Halverson emphasized organizational memory as distributed across work practices rather than reducible to a static archive [2]. The present architecture extends that lineage into heterogeneous AI-assisted work and persistent semantic state.

16.2 Retrieval-augmented generation and hierarchical context

RAG combines parametric language models with external non-parametric memory [3]. RAPTOR introduces recursive clustering and summarization to retrieve at multiple abstraction levels [6]. These methods support the broader proposition that external context can be structured and compressed rather than repeatedly carried in full.

16.3 Persistent agent memory

Generative Agents combine observation, retrieval, reflection, and planning [4]. MemGPT uses hierarchical virtual context management inspired by computer memory systems [5]. A-MEM dynamically links and updates memories [10]. Mem0 extracts and consolidates salient information into persistent memory and graph-enhanced memory [11]. Amazon Bedrock AgentCore Memory distinguishes short-term interaction memory from consolidated long-term memories [14]. These systems demonstrate that persistent and tiered memory is an active field rather than a unique claim of this paper.

16.4 Graph-based and temporal context

GraphRAG uses graph structure and community summaries for corpus sensemaking [7]; HippoRAG combines retrieval, graph structure, and graph algorithms for long-term memory [8]; Zep/Graphiti uses a temporally aware knowledge graph for ongoing conversational and business information [9].

16.5 Memory and retrieval poisoning

Persistent and retrieval-augmented memory create a security surface because adversarial content can be stored once and influence later reasoning. PoisonedRAG demonstrated high-success knowledge-corruption attacks against RAG systems at large corpus scale and found evaluated defenses insufficient in its setting [12]. More recent persistent-memory work analyzes "laundering" mechanisms through which untrusted memory can appear trusted after summarization, tool echoes, or manufactured corroboration, motivating origin-bound authority rather than relying only on content classification or mutable lineage [13]. These results support treating integrity, provenance, and authority as architectural state rather than optional inference-time filtering.

16.6 Enterprise context platforms

Commercial enterprise systems are converging on persistent graph-based context. Asana publicly describes its Work Graph as a connected model of company work with shared memory and governance [15]. Atlassian describes Teamwork Graph as connected context spanning Atlassian and third-party systems and exposes it to external AI agents [16]. Palantir describes its Ontology as an operational model that unifies objects, relationships, logic, actions, and security for humans and agents [17]. These systems are important adjacent architectures and constrain any novelty claim.

16.7 Defensible contribution

This paper therefore does **not** claim invention of persistent memory, graph context, RAG, summarization, or importance-based gating in isolation. The proposed contribution is the architectural synthesis around a specific unit of analysis:

> **A provider-neutral enterprise intelligence architecture that converts heterogeneous AI and enterprise activity into progressively compacted organizational workstream state,**

protects derived state through provenance- and authority-aware integrity controls, detects material state transitions and cross-workstream relationships, and allocates expensive reasoning according to those transitions.

SECTION 17

FAQ and Architectural Critiques

17.1 Is this just RAG with more steps?

No. RAG primarily asks what external information should be retrieved for a query. This architecture asks what the organization currently believes it is working on, what changed, how workstreams relate, and whether a change warrants reasoning. RAG remains a useful grounding mechanism inside the broader system.

17.2 Is this just GraphRAG?

No, although GraphRAG is directly relevant. GraphRAG is primarily a method for graph-based sensemaking and retrieval over corpora [7]. The architecture here focuses on continuously changing workstreams, material state transitions, and selective routing of expensive organizational reasoning.

17.3 Why not simply use very large context windows?

Context capacity does not eliminate relevance, governance, cost, or repeated-history problems. An organization can produce more history than any practical context window, and repeatedly feeding old history to new calls remains inefficient. Persistent state attempts to retain what matters while preserving evidence for re-grounding.

17.4 How does the framework scale?

By ensuring that the highest-volume stages are cheap and horizontally scalable, bounding candidate comparison, and preventing every raw event from becoming a deep reasoning job. The intended funnel is Signals $\rightarrow$ interpreted work $\rightarrow$ material state changes $\rightarrow$ Attention events $\rightarrow$ agent executions.

17.5 Does context compaction lose information?

Yes. Semantic compaction is intentionally lossy. The architecture addresses this by separating the active reasoning representation from the higher-fidelity evidence store. Consequential conclusions can be re-grounded against evidence rather than treating a summary as unquestionable truth.

17.6 What happens when the system summarizes something incorrectly?

State must be versioned, provenance-linked, challengeable, supersedable, and recomputable. High-impact state transitions should require stronger evidence or validation than routine updates.

17.7 Could hallucinations or malicious content poison long-term organizational memory?

Yes. Persistent memory makes this risk more consequential because a bad state can survive the session that created it. Model-generated inference should not silently become authoritative fact. The Microglial Integrity Layer preserves origin and permission lineage, prevents pure transformation from increasing authority, quarantines suspect state, requires independent corroboration where appropriate, and supports descendant invalidation. These controls reduce risk but do not guarantee immunity from poisoning [12, 13].

17.8 Is Microglia just another agent?

No. It should be implemented primarily as a defensive control plane: deterministic provenance rules, authorization checks, dependency tracking, source authority, anomaly signals, quarantine, and validation workflows. Models may assist in contradiction or anomaly analysis, but a reasoning model should not be the sole authority deciding whether its own derived memory is trustworthy.

17.9 Can Microglia guarantee that poisoned state never propagates?

No. That would be an unrealistic security claim. The goal is defense in depth: make poisoning harder to persist, constrain the authority of derived state, detect suspicious propagation, provide quarantine, and make retrospective repair possible. Poisoning resilience is explicitly part of the evaluation framework rather than an assumed property.

17.10 Does the Attention layer risk filtering out exactly the anomaly that matters?

Yes. That is why the proposed gate is not a single threshold. Domain-risk overrides, low-confidence escalation, Microglial triggers, periodic sweeps, and audit sampling create additional paths for low-signal/high-impact events. The residual false-negative rate remains an empirical question and should be measured directly.

17.11 Won't the state graph eventually become enormous?

Yes, but graph size does not require every operation to touch the entire graph. Domain partitioning, temporal scoping, metadata filters, vector candidate generation, bounded neighborhood traversal, lifecycle states, and archival policies can keep the active working set limited.

17.12 How does cross-provider context help in practice?

An organization can use different models for different work while preserving authorized context in a shared semantic layer. A later model receives the relevant compact state and evidence instead

of requiring complete historical transcripts from every previous provider. The main benefit is continuity without forcing model-provider uniformity.

17.13 Does every provider need to expose its data?

No. The architecture degrades gracefully. Sources with authorized APIs, exports, logs, or telemetry can participate; sources without them remain outside that portion of organizational memory.

17.14 Isn't this employee surveillance?

It can become surveillance if implemented badly. The proposed design principle is to model *work* rather than rank *workers*. Production governance should restrict productivity scoring, unrestricted raw-prompt browsing, and other individual-level uses that are not necessary to the architectural purpose.

17.15 How do permissions work for knowledge derived from several sources?

Conservatively. Derived state must retain access lineage. In some cases the safe permission set may be the intersection of source permissions; in others the system may generate a less-sensitive abstraction; in still others the state cannot be exposed across boundaries at all. Under-sharing is preferable to permission leakage.

17.16 Why not rely entirely on built-in model or agent memory?

Built-in memory is useful but often scoped to a user, agent, workspace, provider, or product. The architectural goal here is organizational continuity that remains independent of a single reasoning provider while supporting permission-aware retrieval by multiple authorized consumers.

17.17 What is actually new?

Potentially the architectural synthesis, not the ingredients. The candidate contribution is the combination of cross-provider observation, persistent workstream state, progressive organizational context compaction, Microglial integrity controls, state-transition Attention with safety overrides, bounded relationship discovery, and selective reasoning.

17.18 Has this architecture proven that it works?

Not yet at the level required for a peer-reviewed empirical performance claim. The evaluation framework in this paper defines what should be measured: compaction, Attention precision and recall, relationship quality, reasoning reduction, cost, latency, and accepted organizational findings.

SECTION 18

Key Methodological and Technical Limitations

The architecture is intentionally presented as a reference design and research program rather than a completed empirical result. Four limitations are especially important.

18.1 Unvalidated scaling hypotheses

The compaction and reasoning-growth claims in this paper remain hypotheses. Equations such as $S \gg D \gg \Delta N \gg A \gg R$ describe an intended operating regime, not a measured universal law. Workloads with continuously novel activity, poor state resolution, or frequent high-impact changes may compact weakly and may not exhibit sublinear reasoning growth.

Response. The paper defines explicit compaction, Attention, reasoning, utility, and cost metrics rather than presenting synthetic projections as established performance. Production validation should compare the architecture against at least a full-history or ungated baseline under equivalent tasks, models, permissions, and quality targets.

18.2 Semantic compaction as an error amplifier

Compaction is inherently lossy. An early Dendrite or Neuron misinterpretation can persist, acquire relationships, and influence many later agents. Adversarial poisoning creates the same structural problem intentionally.

Response. The Microglial Integrity Layer makes semantic-state integrity explicit. Derived state retains origin, authority, permission lineage, confidence, and dependencies; pure transformation cannot by itself elevate authority; suspect state can be quarantined or re-grounded; and later discovery of corruption can trigger descendant invalidation. This reduces blast radius and increases reversibility, but its effectiveness must itself be measured.

18.3 Attention bottlenecks

Selective gating creates an inherent recall/compute trade-off. A narrowly tuned materiality threshold may ignore low-signal events whose consequences become apparent only later.

Response. Attention is modeled as a multi-path detection system rather than a single threshold. Domain-risk overrides, uncertainty escalation, Microglial triggers, periodic bounded sweeps, and below-threshold audit sampling provide alternate discovery paths. Evaluation must report false-negative rates for low-frequency/high-impact events in addition to cost reduction.

18.4 Context concentration and governance

Combining fragmented activity into a unified semantic state can create a higher-density security asset than any individual source. Derived relationships may reveal sensitive facts even when raw source content is never directly exposed.

Response. The architecture rejects unrestricted global memory as a default. Derived state retains access lineage, Brain Regions can act as security compartments, cross-boundary deriva-

tion is policy-controlled, and Microglial checks propagate permission taint and can quarantine unsafe abstractions. Security evaluation should include adversarial permission-leakage tests at the derived-state and inference layers, not only storage access checks.

SECTION 19

Limitations and Open Questions

The architecture remains subject to several unresolved challenges:

- semantic interpretation and state-resolution error;
- ambiguity in merging and splitting workstreams;
- evolving relationship semantics and stale graph edges;
- calibration of Attention thresholds and policies;
- permission propagation across derived knowledge;
- deletion and retention semantics;
- model dependence and model-specific failure modes;
- adversarial poisoning, indirect instruction persistence, and manufactured corroboration;
- false positives and false negatives introduced by Microglial controls themselves;
- incomplete visibility into systems that lack suitable enterprise telemetry;
- organizational ambiguity, changing ownership, and inconsistent terminology;
- the possibility that the cost of maintaining high-quality state outweighs the reasoning savings for some workloads.

These are not peripheral implementation details. They are central research and engineering questions.

SECTION 20

Conclusion

Enterprise AI is likely to remain heterogeneous. Organizations will use multiple models, agents, data systems, and collaboration tools, each creating local context. The resulting challenge is not merely retrieval; it is persistent organizational continuity without allowing persistent error, poisoned memory, or permission leakage to become equally durable.

The architecture proposed in this paper treats organizational work as evolving state. Raw observations become interpreted units of work. Those units propose updates to persistent workstreams. Microglial integrity controls preserve provenance and authority, quarantine suspicious state, and support repair before or after persistence. Workstreams form temporal relationships.

Most accepted activity remains inexpensive state maintenance. Material transitions attract Attention. Only a smaller subset reaches expensive specialized reasoning. Validated intelligence can then return to enterprise execution through controlled outputs.

The resulting pipeline is:

$$Observe \rightarrow Interpret \rightarrow Compact \rightarrow Protect \rightarrow Persist \rightarrow Attend \rightarrow Reason \rightarrow Act \tag{13}$$

The most important proposition remains simple:

The scalable unit of enterprise reasoning should be the material change in organizational state, not the raw interaction.

If empirical evaluation supports that proposition, persistent organizational context can become a practical interoperability layer above fragmented AI providers: one that reduces repeated context, surfaces relationships among workstreams, preserves evidence, and allocates expensive reasoning where it has the greatest expected value.

SECTION A

Canonical Terms

Term	Function	Primary question
Signal	Canonical authorized observation	What happened?
Dendrite	Interpreted unit of work	What did the activity mean?
Neuron	Persistent workstream state	What are we working on now?
Synapse	Typed temporal relationship	How is this work related to other work?
Brain Region	Domain or organizational scope	Where should retrieval and reasoning occur?
Microglia	Defensive integrity control plane	Is this derived state safe and authorized to propagate?
Attention	Materiality and reasoning gate	Does this change justify deeper reasoning?
Agent	Specialized reasoning capability	What can be inferred or recommended?
Axon	Controlled output or proposed action	What should happen next?
Organizational State Graph	Temporal graph of workstream state and relationships	How is organizational work evolving and interacting?

SECTION B

Core Architecture Invariants

1. Evidence and semantic interpretation remain distinguishable.
2. Provider-specific activity normalizes into canonical observations.
3. Durable organizational state remains independent of any single model provider.
4. Raw information progressively compacts before expensive reasoning.
5. Routine activity may update state without invoking deep reasoning.
6. Expensive reasoning is selectively routed.
7. Workstream comparison is bounded rather than globally pairwise by default.
8. Vector retrieval and graph traversal serve complementary roles.
9. Material state transitions are first-class objects.
10. Relationships retain temporal and provenance information.
11. Derived state retains access lineage.

12. Pure transformation does not independently elevate source authority; trust elevation requires explicit independent validation.
13. Suspect state can be quarantined before normal downstream propagation.
14. Microglial controls can trigger re-grounding, restricted retrieval, or descendant invalidation.
15. Attention includes risk and audit override paths rather than relying on a single threshold.
16. Important conclusions can be re-grounded against evidence.
17. Consequential actions remain policy-controlled.
18. The architecture models work rather than ranking workers.
19. Deletion and invalidation propagate through dependent state when required.
20. Scaling and performance claims remain hypotheses until benchmarked.

ARMORED HELIX / PUBLICATIONS

References

ARMORED HELIX / PUBLICATIONS

References

[1] E. W. Stein and V. Zwass, "Actualizing Organizational Memory with Information Systems," *Information Systems Research*, vol. 6, no. 2, pp. 85-117, 1995. doi:10.1287/isre.6.2.85.

[2] M. S. Ackerman and C. Halverson, "Considering an Organization's Memory," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 1998. https://web.eecs.umich.edu/~ackerm/pub/98b24/cscw98.om.html.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

[4] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," 2023. https://arxiv.org/abs/2304.03442.

[5] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: Towards LLMs as Operating Systems," 2023. https://arxiv.org/abs/2310.08560.

[6] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval," *International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=GN921JHCRw.

[7] D. Edge et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," 2024. Microsoft Research. https://www.microsoft.com/en-us/research/publication/from-local-to-global-a-graph-rag-approach-to-query-focused-summarization/.

[8] B. J. Gutierrez et al., "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models," *Advances in Neural Information Processing Systems*, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

[9] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef, "Zep: A Temporal Knowledge Graph Architecture for Agent Memory," 2025. https://arxiv.org/abs/2501.13956.

[10] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, "A-MEM: Agentic Memory for LLM Agents," 2025. https://arxiv.org/abs/2502.12110.

[11] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav, "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory," 2025. https://arxiv.org/abs/2504.19413.

[12] W. Zou, R. Geng, B. Wang, and J. Jia, "PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models," in *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

[13] Y. Louck, "Securing LLM-Agent Long-Term Memory Against Poisoning: Non-Malleable, Origin-Bound Authority with Machine-Checked Guarantees," 2026. Preprint. https://arxiv.org/abs/2606.24322.

[14] Amazon Web Services, "Amazon Bedrock AgentCore Developer Guide: Memory Types," 2026. Accessed Aug. 25, 2026. https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/memory-types.html.

[15] Asana, Inc., "Asana Work Graph, Shared Memory, and AI Teammates," 2026. Accessed Aug. 25, 2026. https://asana.com/.

[16] Atlassian, "Atlassian Teamwork Graph: The Context Engine Behind Your AI - Everywhere," May 6, 2026. Accessed Aug. 25, 2026. https://www.atlassian.com/blog/company-news/teamwork-graph-team-26.

[17] Palantir Technologies Inc., "Foundry Ontology: Overview and Core Concepts," 2026. Accessed Aug. 25, 2026. https://www.palantir.com/docs/foundry/ontology/overview.