The Evidence-Anchored Organizational Graph

N71 Research

Technical Report TR-2026-01 June 2026

Abstract

Structured organizational memory is commoditizing. Write-time knowledge graphs, provenance tracking, and temporal awareness — capabilities that differentiated systems a year ago — are converging into the baseline, and we expect that convergence to complete. This report therefore makes a narrower and harder claim than "we built a graph." It specifies the Evidence-Anchored Organizational Graph as a system defined not by its feature list but by five **invariants** — properties that hold for every object and every answer, without exception — and by the **Grounded Answer Protocol** built on them: a fail-closed answering discipline under which no claim leaves the system without surviving validation against verbatim sources, and insufficient evidence produces refusal rather than generation. A feature list tells you what a system has; an invariant tells you what you can build on. The substrate (five layers, four retrieval channels) is specified as the minimum machinery the invariants require. Consistent with our disclosure policy, we publish contracts and invariants; extraction internals and ranking parameters are not disclosed.

1. The Argument

The failure modes of similarity-based organizational memory are settled science: interference between semantically adjacent facts, temporal flatness, and unaccountability are properties of the embedding geometry itself [1], and impossibility results establish that no purely semantic system escapes them without external symbolic structure [2]. The industry heard the diagnosis. Knowledge graphs, write-time extraction, and temporal metadata are now appearing across the category, and within a product cycle every serious memory system will claim them.

So the discriminating question is no longer *whether* a system has structure. It is what the structure **guarantees**. A graph whose answers cannot be audited is a similarity index with extra steps; a temporal model that decorates facts with timestamps but lets a confident stale answer through has solved the easy half of the problem. Enterprises do not adopt memory systems because recall improves — they adopt them when the output can be *trusted*: trusted to be derived from real sources, trusted to be current, and — hardest — trusted to say "I don't know" when that is the true answer.

This report specifies trust as a set of invariants, then specifies the machinery that maintains them.

2. The Five Invariants

These hold for every object and every answer in the system. They are enforced by construction — by storage-layer constraints, typed contracts, and protocol gates — not by convention or review.

I1 — Provenance totality. Every object above the raw-capture layer has a finite derivation path terminating in immutable captured source. There are no orphan facts. The question "why does the system believe this?" always has an inspectable answer ending in a verbatim passage.

I2 — Bi-temporal completeness. Every fact carries two timestamps: when it occurred in the world and when the system observed it. Queries may bind either. "What happened Monday?" and "what did we know on Tuesday?" are different questions with different correct answers, and a system that cannot distinguish them will eventually answer one with the other.

I3 — Non-destructive evolution. No operation deletes history. Facts are superseded with pointers to what replaced them; entity merges preserve recoverable pre-merge state; corrections append, with an audited record of who changed what and why. The graph at any past moment is reconstructible.

I4 — Validated emission. No claim leaves the system that has not survived citation validation against its anchored sources. Claims that fail validation are stripped before the answer is returned, and the validation record ships with the response. An answer is not text; it is text plus the audit of itself.

I5 — Closed-world honesty. When retrieval cannot assemble sufficient support, the system refuses — explicitly, with a machine-readable reason — rather than generating a plausible response. The refusal path is a first-class output, engineered and evaluated with the same rigor as the answer path.

I1–I3 are substrate invariants; versions of them are appearing elsewhere in the category, and we regard them as the emerging price of admission. I4 and I5 are emission invariants, and to our knowledge no other production system enforces both. They are where this report's claims concentrate, because they are the difference between a memory that informs decisions and a memory that merely furnishes text.

3. The Substrate

The invariants require machinery. We specify it compactly; the architecture here is deliberately conventional, because the substrate is not where novelty should live.

Layer 1 — Raw events. Immutable capture: source, actor, type, payload, and the bi-temporal pair (`occurred_at`, `observed_at`) that I2 requires. Append-only; never edited by any downstream process, including the system's own agents. The terminus of every I1 derivation path.

Layer 2 — Evidence. Events parsed into documents and paragraph-level chunks. The chunk is the system's unit of citation: stable URI, position, section path, source events. Every claim at every layer above resolves to chunk identifiers; every chunk resolves to a verbatim passage with capture context. Anchoring — retrieving the exact source view for any handle — is a single primitive operation.

Layer 3 — The graph. An extraction pipeline lifts typed entities and relationships from evidence: people, accounts, products, projects, decisions, commitments, and an extensible relation vocabulary. Extraction is graph-aware — candidates resolve against existing entities before writing, over the full identifier surface (names, addresses, handles, internal IDs), with confidence-scored, transactional, I3-compliant merges. Typing is governed by a per-organization ontology under an explicit adoption gate, specified fully in TR-2026-03. Every entity and edge carries the evidence chunks that justify it; an entity with no surviving evidence is a defect.

Layer 4 — Temporal state. Entities carry typed state timelines, not current-state snapshots. Each transition is timestamped and causally linked to its triggering event, making point-in-time queries (I2, I3) and "why is this entity in this state?" (I1) ordinary operations. Above the timelines, the system computes lifecycle dynamics — duration-in-state against the historical norm for the entity's type, momentum, behavioral baselines, and deviations from them — so that a deal stalled past its cohort's median with activity gone quiet is a structurally different object from an active one, before anyone asks.

Layer 5 — Synthesis. Continuous processes read across layers 1–4 and produce Thoughts: typed proactive-intelligence objects, each carrying a synthesis chain extending I1 to the system's own conclusions. Synthesis producers pass a typed promotion contract — candidates are validated and promoted, held as provisional, rejected, or deprecated, with typed reasons — so that derived knowledge enters the graph through a gate, not a side door.

Retrieval. Four channels run in parallel — graph traversal, semantic similarity, lexical match, and a freshness channel surfacing material too recent to be fully indexed — each covering the others' blind spots. Fusion policy and parameters are not disclosed.

4. The Grounded Answer Protocol

The protocol is the system's emission discipline — the machinery of I4 and I5 — and the central contribution of this report. Its defining commitment: **it fails closed.** The most damaging output of any memory system is a confident answer its corpus cannot support, and the protocol makes that output structurally difficult rather than merely discouraged.

Stage 1 — Retrieve. Candidate evidence is gathered through the four channels, bounded by the caller's governance scope (TR-2026-02): results are filtered by authority at the source, not redacted after.

Stage 2 — Synthesize. An answer is composed strictly from retrieved evidence, with inline citations binding each claim to specific chunks. Because the substrate has already resolved meaning, identity, and time, the synthesis model's task is reduced to reading comprehension over structured context — which is precisely what permits accurate operation with small, economical, sovereign-hostable models. The architecture, not the model, carries the intelligence.

Stage 3 — Validate. Every citation is checked against its anchored source. Claims that fail are stripped before the answer returns. The validation record — citations checked, failures, claims removed — ships as machine-readable metadata with every response (I4).

The refusal path (I5). Insufficient support yields an explicit refusal carrying a machine-readable reason and recommended next actions — broaden the query, ingest the missing source — never a degradation into plausible generation. We evaluate refusal quality as seriously as answer quality: a system that refuses too rarely is dangerous, and one that refuses too often is useless; the calibration between them is a measured property, not an aspiration (§5).

Two consequences of the protocol are worth stating because they are unusual. First, *the answer is evidence about itself*: a downstream agent consuming an N71 answer can programmatically inspect what was validated, what was stripped, and why — and make its own trust decision. Second, *the protocol composes*: because every claim resolves to chunk handles (I1), an agent can take any sentence of any answer and drill to the verbatim source in one call. There is no dead end between an assertion and its proof.

5. Evaluation

We evaluate where the invariants make falsifiable predictions. The MEME benchmark [3] isolates the two task classes on which practical-cost systems collectively fail: **Cascade** (when an upstream fact changes, do dependent facts update? — a direct test of I2/I3) and **Absence** (does the system know when it does not know? — a direct test of I5), with field averages of 3% and 1% respectively. A system claiming these invariants should expect to be measured exactly there; our full-suite run is in progress and will be published with a per-miss failure classification — *retrieval miss*, *ranking loss*, or *reasoning failure* — because the classification, not the score, determines whether a failure is fixed in the channels, the substrate, or the protocol. We will publish failures alongside successes. A vendor that publishes only wins should be presumed to have losses.

6. Limitations

Stated plainly. Extraction is performed by language models and inherits their errors; the mitigation is not perfect extraction but I1 plus I3 — every error is traceable and correctable without history loss. The freshness channel trades ranking quality for availability inside the indexing window. Refusal calibration is corpus-dependent and is weakest early in a deployment, for the same reason all of Layer 4 is: the system's distinctive properties compound with tenure, and a week-old workspace has invariants but not yet much for them to protect.

7. Disclosure

We publish invariants, contracts, and protocols — the claims we expect to be held to and measured on. Extraction prompt architecture, resolution scoring, confidence calibration, fusion weights, and synthesis triggering remain proprietary.

References

[1] Ray Barman, S., Starenky, A., Bodnar, S., Narasimhan, N., Gopinath, A. *The Geometry of Forgetting: How High-Dimensional Embeddings Reproduce Human Memory Phenomena.* arXiv:2604.06222, 2026.

[2] Ray Barman, S., Starenky, A., Bodnar, S., Narasimhan, N., Gopinath, A. *The Price of Meaning: Impossibility Theorems for Semantic Memory Systems.* arXiv:2603.27116, 2026.

[3] Jung, S., et al. *MEME: Multi-entity and Evolving Memory Evaluation.* KAIST AI · Tübingen · NAVER. arXiv:2605.12477, 2026.