Contents

Neuro-Semiotic Reasoning: Algebraic Semiosis for Observational Causal Inference

Martin Trajkow

Lead Researcher, o-machine

martin@o-machine.com

https://o-machine.com

Version: 2.0 • May 27, 2026

Initial Publication: Ver. 1.0 • Feb 17, 2026

Abstract

Current AI systems excel at statistical pattern matching and formal discovery but fail systematically at causal reasoning from real-world observational signals. While systems like AlphaEvolve demonstrate LLM-guided evolutionary search for innovation within formal spaces [17, 21], they remain blind to unstructured, hidden signals of physical and commercial reality. We identify this as the Causal Reasoning Gap: the inability to derive verified causation from observational signals that have not been textualized or modeled.

We present the Neuro-Semiotic Reasoning Engine, an architecture that constructs causal understanding through semiotic interpretation of differential relationships across temporal, spatial, conceptual, and absence dimensions. Drawing on Peirce's unlimited semiosis, we introduce Behavioral Signals as the modality-independent unit of computation and demonstrate that causal meaning can be constructed through algebraic operations over high-dimensional signal representations.

A key theoretical finding is that recursive semiotic interpretation — predicted by Peircean theory to require iterative processing — admits algebraic collapse: the recursive chain can be computed as constant-time resonance operations that achieve equivalent or superior reasoning quality at orders-of-magnitude lower latency. We validate this empirically through a systematic ablation study across 14 experimental configurations and 109 evaluation queries, demonstrating that the architecture achieves a mean answer quality of 4.4/5.0 on causal reasoning tasks while operating 140× faster than LLM-based alternatives.

Keywords: Causal reasoning, Behavioral signals, Neuro-semiotic AI, observational causal inference, algebraic semiosis, absence detection, adversarial falsification.

1. Introduction

Large language models have achieved remarkable fluency in text generation, yet leading AI researchers increasingly question their fundamental architecture as a path to human-level intelligence. As LeCun argues, LLMs lack a grounded model of the world because they are "limited to the discrete world of text" [16]. Sutton identifies deeper structural deficits: they lack ground truth for verification and mechanisms for surprise-driven learning [18]. Silver and Sutton characterize the current paradigm as fundamentally limited by its reliance on static datasets rather than continuous interaction with reality [20].

Furthermore, even the most advanced evolutionary discovery systems, such as DeepMind's AlphaEvolve [17], operate primarily within formal problem spaces — mathematical analysis, combinatorics, and code — where the environment is defined by rigid rules and automated evaluators. While these systems can discover novel algorithmic solutions, they are essentially performing synthesized innovation: exploring permutations of a known, modeled universe.

We identify a different, more fundamental structural limitation: the Hidden Reality Blind Spot. Current AI is blind to what has not yet manifested as a document or been modeled into a formal search space. The most valuable causal insights are often those still manifesting as raw, unrecorded behavioral signals. In highly dynamic environments, statistical correlation and formal search both face challenges because they lack access to the underlying causal mechanisms operating in real-time.

We define this as the Causal Reasoning Gap: the inability to derive causation from unstructured, temporal, multi-modal observational signals. This paper presents the Neuro-Semiotic Reasoning Engine, an architecture designed to address this gap through computational semiosis — the construction of causal understanding through differential interpretation of Behavioral Signals.

We report three principal contributions:

1. A formal framework for observational causal inference grounded in Peircean semiosis, introducing Behavioral Signals as a modality-independent unit of computation and Semantic Atomicity as a structural constraint enabling algebraic reasoning.
2. A theoretical finding that recursive semiotic interpretation admits algebraic collapse — the iterative interpretive chain predicted by semiotic theory can be computed as constant-time algebraic operations over high-dimensional signal representations, achieving equivalent reasoning quality at dramatically lower computational cost.
3. Empirical validation through a 109-query evaluation framework with systematic ablation across 14 configurations, demonstrating measurable improvements in causal reasoning depth, adversarial falsification accuracy, and multi-hop inference quality.

2. Background and Related Work

2.1 The Hierarchy of Knowledge Processing

We distinguish four levels of knowledge processing:

Level	Definition	Example	Current AI Capability
Information	Raw facts	"Firm X hires Y"	High
Knowledge	Organized facts	"Firm X is building an engineering team"	High
Understanding	Causal relationships	"Firm X builds engineering team because…"	Limited
Wisdom	Actionable insight	"Firm X's hiring shift indicates strategic pivot"	Limited

Current systems excel at Levels 1-2 but face severe challenges at Levels 3-4. We hypothesize this limitation is architectural, not merely computational.

2.2 Statistical Pattern Matching: Capabilities and Limits

LLMs predict text distributions based on training data. They manipulate language patterns without necessarily constructing causal models of underlying reality. As Sutton notes, "you can't have prior knowledge if you don't have ground truth." This can produce instances where LLMs optimize for token probability rather than correspondence with reality [7].

Even more fundamentally, LLMs lack explicit mechanisms for surprise-driven learning. They cannot readily detect when reality deviates from predictions or update their internal world models dynamically in real-time. This makes temporal reasoning from observational signals — such as detecting a strategic inflection point — structurally challenging.

2.3 Neuro-Symbolic and Discovery Systems

Recent work in neuro-symbolic AI seeks to combine neural learning with symbolic reasoning, but typically requires pre-defined, brittle ontologies [15]. Meanwhile, discovery systems like AlphaEvolve [17] utilize LLMs to guide evolutionary search in formal spaces, with downstream applications demonstrating mathematical exploration at scale [21]. While effective for algorithmic generation, these systems are constrained to environments where outcomes are formally verifiable. They are not designed to infer organizational strategy because strategy is never formally specified — it must be inferred from observational signals.

2.4 Retrieval-Augmented Generation

RAG systems [22] enhance LLM outputs by retrieving relevant documents at query time. However, standard RAG retrieves by semantic similarity — a fundamentally different operation from causal reasoning. GraphRAG [23] extends this by incorporating knowledge graph structure, improving multi-hop reasoning. Our architecture builds on the insight that graph structure improves reasoning, but diverges fundamentally: we process behavioral signals rather than document fragments, enforce semantic atomicity on all inputs, and reason through algebraic semiosis rather than LLM-mediated synthesis.

2.5 Semiotic Foundations

We draw on Peirce's theory of unlimited semiosis [11] — the principle that signs generate interpretants which themselves become new signs. Meaning is dynamically constructed through the process of recursive interpretation. In our architecture, a fact gains meaning from its differential relationships across temporal, spatial, conceptual, and absence dimensions.

A central finding of this work is that the recursive interpretive process described by Peirce admits an algebraic realization. Rather than requiring iterative chains of interpretation, the semiotic relationships between signals can be computed through constant-time algebraic operations over high-dimensional representations — a result we term Algebraic Semiosis (§3.5).

3. Theoretical Foundation: Causal Understanding Through Behavioral Signals

3.1 A Different Unit of Computation: Actions, Not Tokens

Every reasoning paradigm is defined by its fundamental unit of computation — what it "sees" of the world.

- LLMs see Tokens: Statistical fragments of text, stripped of time and context.
- Symbolic AI sees Propositions: Rigid logical statements.
- Knowledge Graphs see Triples: Static snapshots of entity-relation-entity.

The proposed architecture processes Behavioral Signals: observable actions by real-world entities — moving, connecting, separating, acquiring, releasing, abandoning — anchored in time, linked to evidence, and analyzed in relation to their own history and the concurrent behaviors of adjacent actors.

Critically, a Behavioral Signal is modality-independent. The signal may be extracted from text documents, imagery, sensor data, or transactional records. The architecture reasons over the fact of the behavior, not solely over the medium that carried the signal. This decoupling aims to allow the system to ingest reality regardless of whether that reality has been comprehensively "textualized."

3.2 Semantic Atomicity as a Computational Constraint

Traditional NLP entity-relationship extraction typically produces compound, entangled relationships. When extracting logic from unstructured text, generative models naturally mirror the fuzzy, overlapping structure of human language. However, these compound structures cannot be subjected to strict algebraic computation.

The Neuro-Semiotic architecture resolves this limitation by enforcing a strict constraint we term Semantic Atomicity. Before any observational signal enters the temporal knowledge graph, the semiotic extraction layer is rigidly forced to decompose all raw text into indivisible (s, p, o, t) tuples (subject, predicate, object, timestamp).

By strictly constraining the semiotic parser to output these atomic tuples, we ensure that the system does not ingest blurry heuristics. This constraint guarantees that the subsequent algebraic operations are mathematically well-defined: commutative where required, reversible, and topologically consistent, bridging the gap between linguistic ambiguity and formal reasoning.

3.3 The Four Dimensions of Differential Meaning

We propose that causal meaning emerges from analyzing how signals relate across four dimensions:

1. Temporal Dimension: Causes precede effects — velocity and acceleration of behavioral change reveal intent.
2. Spatial Dimension: Entities connected through multiple independent pathways may indicate causal relationships.
3. Conceptual Dimension: Facts gain meaning when interpreted within evolving concept phases.
4. Absence Dimension: The deviation between expected behavior and observed behavior serves as a diagnostic signal.

3.4 Unlimited Semiosis: The Recursive Construction of Meaning

The central hypothesis is that causal meaning can be approximated through differential interpretation of behavioral signals. We model this through a semiotic chain where each interpretation generates a new sign (interpretant) for further analysis.

Motivating Scenario 3.1: The Jony Ive Inference

To illustrate the reasoning process, consider a hypothetical scenario prior to any official product announcements: “What kind of hardware is Jony Ive building for OpenAI?”

Instead of retrieving semantic text matches, the engine computes the unannounced product’s structural “shape” through a sequence of algebraic operations:

1. Unobserved State Inference: The query's structural vector mathematically resonates with the behavioral profile of Ive's design firm (LoveFrom), pulling the correct entity into the reasoning context despite its not being explicitly named.
2. Cross-Signal Convergence: Divergently phrased signals ("LoveFrom hires micro-optics engineers," "OpenAI secures low-power edge silicon") exhibit high topological alignment. The algebra fuses them into a single, coordinated supply-chain event.
3. Behavioral Divergence: The prevailing tech industry trend is scaling massive cloud infrastructure. The LoveFrom/OpenAI vector is mathematically orthogonal to this baseline — they are explicitly not building data centers.
4. Diagnostic Absence: The system calculates a near-zero resonance with traditional screen-based UX patents or heavy robotic chassis procurement, ruling out traditional phones and robotics.
5. Adversarial Falsification: Public statements claim the project is "not a wearable." The system subjects this to structural falsification: if it were a stationary ambient device (like a smart speaker), it would require acoustic engineering and home-automation integrations. The mathematical absence of these behaviors falsifies the public claim, indicating the denial is likely strategic misdirection.

Inferred Hypothesis: Despite public denials, the structural behavioral footprint forces the conclusion of a screenless, AI-integrated wearable device focused on ambient computing.

In this scenario, the insight is mechanically derived from the intersection of observed actions and calculated absences, inferring the truth of the product from the geometric "hole" it leaves in the market, even when public statements contradict the behavioral reality.

3.5 Algebraic Semiosis: From Recursion to Resonance

Peirce's theory of unlimited semiosis predicts that meaning construction requires recursive interpretation — each interpretant becoming a new sign for further analysis. Our initial theoretical framework (v1.0) formalized this as iterative processing chains with computational cost proportional to chain depth.

Implementation and empirical investigation revealed a deeper result: the recursive interpretive process admits algebraic collapse. The semiotic relationships between behavioral signals — causal adjacency, conceptual resonance, behavioral similarity, absence detection — can be encoded as algebraic operations over high-dimensional signal representations and evaluated in constant time.

This finding has both practical and theoretical significance:

Practical consequence. The architecture achieves multi-hop causal reasoning, adversarial falsification, and absence detection without iterative processing. Query understanding, signal scoring, and causal chain assembly operate as single-pass algebraic evaluations, enabling latencies of 50-200ms for operations that would require seconds under recursive interpretation.

Theoretical consequence. Algebraic semiosis suggests that Peircean recursive interpretation, when properly formalized over structured signals, has latent algebraic structure. The "unlimited" chain of interpretants need not be computed sequentially — the relationships it would discover are already encoded in the geometric configuration of the signal space. The recursion is not eliminated — it is resolved analytically, much as a recurrence relation can be solved in closed form.

We formalize this as follows. Let $\mathcal{S} = \{f_1, f_2, \ldots, f_n\}$ be a set of semantically atomic behavioral signals, each represented as a high-dimensional algebraic encoding $\mathsf{h}(f_i)$. A semiotic relationship R between signals f_i and f_j (causal adjacency, conceptual similarity, behavioral resonance) is computable as:

$$R(f_i, f_j) = \phi(\mathsf{h}(f_i), \mathsf{h}(f_j))$$

where ϕ is an algebraic similarity function operating in constant time with respect to the number of signals. The recursive semiotic chain $I = \{i_1, i_2, \ldots, i_k\}$ is then analytically recoverable from the pairwise relationship matrix R without sequential evaluation.

4. System Architecture

The architecture is implemented through five pillars that collectively realize algebraic semiosis over behavioral signals. We describe each pillar at the functional level — what it computes and why — while the specific algebraic mechanisms constitute proprietary implementation.

4.1 Pillar 1: Temporal Knowledge Graph

We utilize a fact-first architecture where the fundamental unit is the timestamped Behavioral Signal with provenance tracking. Each signal is stored as a semantically atomic tuple (s, p, o, t, E) where E is the evidence set linking the signal to its source material.

The temporal knowledge graph enables:

- Velocity analysis: Detecting acceleration or deceleration in behavioral patterns (e.g., a hiring spike followed by deceleration signals strategic pivot completion).
- Causal ordering: Enforcing temporal precedence as a hard constraint on causal inference.
- Evidence chains: Maintaining full provenance from final inference back to source signals.

4.2 Pillar 2: Concept-Mediated Ontology

Entities connect through evolving concepts rather than static taxonomies. Adding a new signal to a domain recontextualizes the concept itself, which recursively recontextualizes connected entities. Meaning is constructed dynamically rather than encoded statically.

A critical design decision is the elimination of rigid categorical hierarchies. Traditional ontologies require pre-defined taxonomic trees that become brittle as domains evolve. Our concept-mediated approach treats concepts as emergent structures — behavioral clusters that form and dissolve as the signal landscape changes.

4.3 Pillar 3: Adversarial Falsification

Inferences undergo structured evaluation through a multi-agent verification protocol.

- Generation: Causal connections are proposed based on structural pattern matches across the four differential dimensions.
- Verification: Proposed inferences are critically examined against counter-evidence and contradictory signals. An inference is only committed if it withstands adversarial assessment against available signals across all four dimensions.
- Revision: When verification identifies weaknesses, the generation process is re-invoked with focused context — only the specific evidence that triggered the concern — preventing dilution of the critical signal.

This verification protocol proved to be the architecture's strongest and most robust capability in empirical evaluation (§7.3).

4.4 Pillar 4: Behavioral Modeling and Absence Detection

Absence detection requires expectation. The architecture maintains behavioral baselines for entities within specific concept phases. Significant deviations from these baselines are treated as diagnostic signals of structural shifts.

Absence detection operates through dual-track analysis:

- Track A: Semantic centroid shift analysis, measuring how an entity's behavioral profile drifts over time.
- Track B: Structural transformation analysis, measuring the degree of algebraic divergence between an entity's historical and recent behavioral representations.

This dual-track approach detects not only individual entity anomalies but also macroscopic phenomena: entire behavioral domains collapsing simultaneously across multiple entities, signaling sector-level structural transitions.

4.5 Pillar 5: Evolutionary Optimization

The system's reasoning strategies — including detection patterns, verification thresholds, and scoring parameters — are designed to evolve through selection pressure. Drawing on recent work in test-time computation and evolutionary meta-learning [14, 17, 19], the optimization framework adjusts reasoning procedures based on outcome validation from evaluation feedback.

5. Emergent Reasoning Capabilities

5.1 Inference of Unobserved States via Structural Resonance

The architecture enables inference of hidden structural states from weak, cross-domain signals. When a query references an action or behavioral pattern without naming specific entities, the algebraic representation allows the system to identify latent entities — actors whose behavioral profiles resonate with the query's structural signature — without iterative graph traversal.

Validated empirically: given the query "Who is driving the regulatory crackdown?", the system identifies the relevant regulatory body as the dominant resonant entity while rejecting unrelated entities with near-zero resonance scores, achieving clean separation without any entity being explicitly named.

5.2 Rapid Falsification via Absence Signal Diagnostics

The system generates meaning from the non-occurrence of expected events to rapidly evaluate claims. If an entity claims "proprietary hardware breakthroughs" but exhibits zero evidence of specialized hiring or component procurement, the system generates a high-confidence anomaly signal.

Empirical evaluation demonstrates that adversarial falsification is the architecture's most robust capability: 14 of 24 adversarial queries achieve perfect scores across all evaluation dimensions, with the system reliably disproving false premises with evidence citations and correctly validating true claims (§7.3).

5.3 Multi-Hop Causal Chain Tracing

The architecture traces multi-hop dependencies to map systemic exposures. Unlike traditional graph traversal approaches that require iterative queries proportional to hop depth, the algebraic approach enables multi-hop causal chain assembly through sequential algebraic operations, with resonance scores remaining discriminative across 2-3 hops.

5.4 Macro-Event Fusion

The same algebraic primitives that operate at the individual signal level scale identically to detect macroscopic phenomena. Structurally related event clusters — even those expressed through entirely different vocabularies — are recognized as manifestations of the same underlying event when their participant topology aligns, regardless of lexical divergence.

Validated empirically: clusters describing a regulatory crackdown through different vocabulary ("investigates/suspends" vs. "scrutinizes/bans") achieve structural resonance of 0.73, while an unrelated event cluster scores -0.004.

5.5 Complete Auditability

Conclusions are backed by a traceable interpretive chain leading back to atomic, evidence-linked Behavioral Signals. Each step in the proof chain carries its structural provenance, enabling post-hoc inspection of why a particular inference was committed.

6. Evaluation Methodology

6.1 Evaluation Philosophy

Evaluating causal reasoning systems presents distinct challenges. Precision and recall measure fact retrieval — they do not fully capture the quality of causal inference, the detection of absence, or the depth of cross-domain reasoning. Standard benchmarks for RAG systems evaluate retrieval accuracy, not reasoning depth.

We developed a purpose-built evaluation framework that measures reasoning quality across multiple dimensions, evaluated by an independent LLM judge against structured rubrics. All evaluations use Gemini 3.1 Flash Lite as the judge model with default generation settings (temperature, top-p), ensuring reproducibility and eliminating evaluator variance from stochastic sampling.

6.2 Evaluation Dataset

The evaluation dataset comprises 109 queries organized by reasoning complexity:

Category	Count	Reasoning Depth
Multi-hop causal	30	2-4 hops
Velocity/temporal	32	2-5 dimensions
Adversarial falsification	32	2-5 dimensions
Strategic reasoning	7	5 dimensions
Strategic ecosystem impact	5	5 dimensions
Partial answer / edge cases	3	2 dimensions

A curated subset of 27 queries (the “SOTA set”) is designated for publication-grade comparison: 65% causal reasoning (“why” questions) and 35% anti-hallucination (adversarial falsification).

All queries are evaluated against a live knowledge graph containing 18,500+ behavioral signals extracted from real-world sources in the autonomous vehicles domain, covering regulatory actions, corporate strategy, partnerships, technology development, and competitive dynamics.

6.3 Grading Rubric

Each query response is evaluated across four components on a 1-5 scale:

1. Retrieval & Proof Chain — Were the structurally relevant signals surfaced? Does the proof chain contain the evidence needed to answer the query?
2. Draft Answer — Does the initial narrative correctly interpret the proof chain? Are causal claims supported by cited evidence?
3. Edit Loop — Does the verification process correctly validate strong answers and flag weak ones? Does revision improve quality when triggered?
4. Final Answer — What the user receives. Overall quality of causal reasoning, evidence grounding, and absence of hallucination.

6.4 Ablation Design

To measure the marginal contribution of each architectural component, we define four core experimental conditions:

Condition	Components Active	What It Tests
C0: Semantic Baseline	Vector retrieval only	Standard RAG performance floor
C1: Structural Retrieval	+ Graph traversal + Verification	GraphRAG-equivalent with adversarial validation
C2: Abductive Engine	+ Computed steps + Behavioral analysis	Full causal reasoning pipeline
C3: Full System	+ Algebraic query understanding	Complete architecture with zero-LLM query parsing

Two supplementary axes isolate specific contributions:

- No-evidence: Strips raw evidence text from the proof chain, providing only structured tuples to generation. Tests whether verbose source text helps or hinders narrative quality.
- No-verification: Removes the adversarial verification step. Tests the contribution of the falsification protocol to final answer quality.

6.5 Research Questions

RQ1: Does structure-first reasoning improve causal inference quality over semantic retrieval? Hypothesis: Graph-structural retrieval with adversarial verification will outperform pure semantic retrieval on multi-hop causal and falsification queries.

RQ2: Does algebraic query understanding match LLM-based understanding in reasoning quality? Hypothesis: Algebraic query parsing achieves equivalent or superior answer quality at dramatically lower latency, validating the algebraic semiosis thesis.

RQ3: Under what conditions does absence detection serve as a reliable falsifier? Hypothesis: The introduction of behavioral baselines significantly reduces false positives in claim validation compared to purely generative methods.

7. Empirical Results

7.1 Component Contribution Analysis

The following table presents mean scores across all ablation configurations, ordered by experimental phase:

Config	Description	N	Retrieval	Draft	Edit	Final
A2	Semantic baseline	5	4.8	4.2	4.4	3.8
A3	+ Structural retrieval	5	4.8	4.4	4.8	4.4
A1	+ Abductive engine	25	4.2	4.2	4.5	4.0
A5	+ No evidence text	5	4.8	4.4	4.8	4.6
B1	Full pipeline (scale test)	48	4.5	4.4	4.6	4.3
B3	Baseline adversarial only	8	5.0	4.9	4.8	4.9
C2	LLM understanding	33	4.3	4.2	4.5	4.1
C1	Algebraic understanding	33	4.3	4.4	4.6	4.4

7.2 Key Finding: Algebraic Understanding Outperforms LLM Understanding

The most significant empirical result validates the algebraic semiosis thesis. In a controlled comparison over 33 identical queries:

	Retrieval	Draft	Edit	Final	Query Understanding Latency
Algebraic (C1)	4.3	4.4	4.6	4.4	~50ms
LLM-based (C2)	4.3	4.2	4.5	4.1	~7,000ms

The algebraic approach is both 140× faster and achieves +0.3 higher final answer quality. Furthermore, the LLM-based configuration suffered from 3 revision failures and 1 empty answer due to context exhaustion, while the algebraic configuration produced zero errors across all 33 queries.

This result is counterintuitive: LLM-based query understanding produces more precise entity extraction, yet this precision does not translate to better answers. The architecture's retrieval pipeline is robust to imprecise entity identification — structural graph traversal and multi-hop expansion compensate for extraction imprecision, while the algebraic approach avoids the stochastic failure modes inherent in generative parsing. The algebraic architecture demonstrated remarkable robustness to extraction failure: during testing (Runs B7/B8), even when the extraction layer failed to identify a specific entity — returning only generic descriptions like "autonomous trucking company" — the engine still achieved a 4.5 Final score. The underlying algebraic representation (via HNSW fallback) mathematically compensated for the lexical extraction failure, finding the correct topological signals regardless.

7.3 Adversarial Falsification: The Strongest Capability

Adversarial falsification — the system's ability to disprove false premises and validate true claims — proved to be the most robust and consistent capability across all configurations:

- In the 24 adversarial queries within Run B1, the architecture successfully disproved false premises with explicit evidence citations, achieving perfect 5.0 scores across all four evaluation dimensions (Retrieval, Draft, Edit, Final) on 14 of the 24 queries.
- The baseline minimal configuration (Run B3) achieved an average Final score of 4.9 on adversarial tasks — matching or exceeding the full pipeline.
- The verification protocol works identically regardless of which upstream components are active.

This finding has architectural significance: adversarial falsification requires only entity-scoped signal retrieval and the verification protocol. The minimum viable configuration for claim validation is structurally simple, suggesting that falsification may be a more fundamental reasoning operation than multi-hop inference.

Representative examples:

> "Is it accurate that safety concerns have had no measurable impact on autonomous vehicle deployment timelines?" System correctly disproves: "Evidence indicates that safety incidents — such as a collision involving [Company A] and a 'grisly accident'

where [Company B]'s vehicle dragged a pedestrian — are the causal drivers [of regulatory delays]."

"Is it correct that [Company C]'s vision-only approach has slowed FSD development?" System correctly validates: "It is incorrect to say that the vision-only approach has slowed development because the evidence indicates that this strategy has instead enabled end-to-end neural networks…"

7.4 Evidence Text Paradox

A surprising finding: stripping raw evidence text from the proof chain produces the highest answer quality. Ablation run A5 evaluated the architecture with raw evidence text completely stripped from the proof chain, providing the generation model with only structural (s, p, o, t) tuples. This minimal configuration achieved a Final Answer score of 4.6/5.0 — the highest score across all 14 experimental runs, notably outperforming the full pipeline (Run A1) which scored 4.0.

When the generation model receives clean structural tuples rather than verbose source text, it narrates more clearly and introduces fewer errors. Evidence text — originally included to improve grounding — actually introduces temporal confusion and narrative smoothing artifacts. The generation model performs better when it narrates structure rather than paraphrasing prose.

This result supports the architectural thesis: the reasoning substrate should be structural, not textual. The generation model's role is to narrate pre-computed structural relationships, not to rederive them from raw text.

7.5 Category-Level Performance

Performance varies by reasoning category, revealing the architecture's strengths and current limitations:

Category	Retrieval	Draft	Edit	Final	N
Multi-hop causal	4.5	4.8	4.8	4.5	4
Strategic reasoning	4.4	4.4	4.6	4.4	7
Ecosystem impact	4.8	4.4	4.6	4.4	5
Adversarial falsification	4.4	4.4	4.8	4.4	8
Velocity/temporal	3.9	4.2	4.5	4.2	8

Strategic reasoning and ecosystem impact queries — which require assembling multi-factor causal explanations across entity boundaries — perform strongly. Velocity/temporal queries are weakest (3.9 retrieval), primarily due to sparse temporal data for specific rate-of-change questions rather than architectural limitations. The system correctly abstains rather than hallucinating when temporal data is insufficient.

7.6 Limitations and Honest Gaps

We report the following limitations transparently:

1. Behavioral scoring requires semantic action resolution. Prior to semantic action resolution, algebraic behavioral scores remained flat across signals (~0.50). Once abstract verbs were analytically mapped to concrete graph actions, the high-dimensional representations produced highly discriminative scores ranging from 0.49 to 0.76, with 15-20 unique variance values per query. While coarse structural topology remains the primary filter, algebraic behavioral scoring provides the decisive tie-breaking mechanism for resolving deep signal pools.
2. Relational Expansion is a resolution escalation, not a default. Ablation Run A3 (relational expansion only) demonstrated that adding multi-hop graph traversal to straightforward queries actually degrades performance by introducing structural noise. However, on paradoxical or contradictory queries (e.g., temporal conflicts), relational expansion was strictly necessary to surface the correct causal drivers. This validates the engine's dynamic retrieval strategy: structural traversal is an escalation mechanism for resolving ambiguity, not a default necessity.
3. One genuine adversarial failure. Query tc_fals_062 ("Is it accurate that no company has successfully commercialized autonomous trucking?") was incorrectly validated — the system agreed with a false premise. Root cause: the proof chain did not surface commercialization signals strongly enough. This represents a retrieval gap, not a reasoning failure.
4. Velocity queries require denser temporal coverage. Rate-of-change questions ("How fast is X accelerating?") require fine-grained temporal data that may not exist for all entities. The architecture correctly refuses to answer rather than hallucinating, but the refusal rate is higher than desired.
5. Generation infrastructure constraints. At scale (48+ queries), generation truncation and resource exhaustion produce occasional empty or truncated answers. These are infrastructure limitations (LLM serving capacity) rather than architectural issues.

8. Discussion

8.1 Algebraic Semiosis as a Theoretical Contribution

The finding that recursive semiotic interpretation admits algebraic collapse has implications beyond this specific architecture. If Peircean semiosis — the foundational theory of meaning construction in semiotics — has latent algebraic structure, this suggests a path toward formal theories of meaning that are both computationally tractable and theoretically grounded.

The practical consequence is dramatic: operations that recursive interpretation would require seconds to compute are resolved in milliseconds. But the theoretical consequence may be more

significant: it suggests that meaning, when properly formalized over structured signals, is not inherently sequential. The “chain of interpretants” that Peirce described may be better understood as a geometric configuration in a high-dimensional signal space — a configuration that can be read directly rather than traversed iteratively.

This parallels developments in other fields: recurrence relations admitting closed-form solutions, iterative algorithms collapsing into matrix operations, and sequential attention mechanisms being reformulated as parallel computations. Algebraic semiosis may represent a similar structural insight applied to the domain of meaning construction.

8.2 Structure vs. Text: The Evidence Paradox

The finding that stripping evidence text improves answer quality (§7.4) challenges a core assumption of RAG systems: that providing more source context improves generation quality. Our results suggest the opposite — for causal reasoning tasks, structural representations outperform textual ones as input to the generation model.

This aligns with the theoretical framework: if the reasoning substrate is structural (algebraic operations over behavioral signals), then the generation model’s role is narration, not reasoning. Providing raw text re-introduces the very ambiguity that the structural layer resolved, forcing the generation model to re-derive relationships from prose rather than narrating pre-computed structure.

8.3 Adversarial Verification as a Fundamental Primitive

The robustness of adversarial falsification across all configurations (§7.3) suggests that verification may be more fundamental than multi-hop inference as a reasoning operation. The minimum viable system for claim validation is remarkably simple: entity-scoped retrieval plus structured verification. This finding has implications for system design: falsification capability should be treated as a foundational layer upon which more complex reasoning is built, not as an optional enhancement.

8.4 Comparison with Related Approaches

Dimension	Statistical Generation (LLMs)	GraphRAG	Proposed Architecture
Primary Input	Tokens / textual fragments	Document chunks + graph triples	Modality-agnostic Behavioral Signals
Logic Foundation	Probabilistic distribution	Hybrid (retrieval + generation)	Algebraic semiosis over differential dimensions
Verification	External ground truth	None (single-pass generation)	Adversarial falsification protocol

Dimension	Statistical Generation (LLMs)	GraphRAG	Proposed Architecture
Query Understanding	LLM-based (seconds)	LLM-based	Algebraic (50ms)
Absence Reasoning	Not supported	Not supported	Dual-track behavioral absence detection
Auditability	Limited (attention weights)	Partial (source citations)	Complete (structural proof chains)

8.5 Domain Generality

While the empirical evaluation is conducted on a single domain (autonomous vehicles), the architectural principles are domain-agnostic by construction. The algebraic encoding of structural roles (subject-action-target) is invariant across domains: a regulatory investigation in autonomous vehicles has the same structural topology as a regulatory investigation in pharmaceuticals. We hypothesize that expansion to new verticals requires primarily signal ingestion — the reasoning pipeline, scoring, and generation operate identically. We plan to validate this through cross-domain transfer experiments in future work.

9. Conclusion

The Causal Reasoning Gap is an architectural gap. Closing it requires moving beyond the "discrete world of text" to a system that analyzes the world through modality-independent Behavioral Signals.

This paper makes three contributions to that effort. First, we formalize a framework for observational causal inference grounded in semiotic theory, introducing Behavioral Signals as the unit of computation and Semantic Atomicity as the constraint enabling algebraic reasoning. Second, we report the theoretical finding that recursive semiotic interpretation admits algebraic collapse — a result with implications for formal theories of meaning construction. Third, we validate the architecture empirically across 109 queries and 14 ablation configurations, demonstrating that the algebraic approach achieves both higher quality (+0.3 Final Score) and dramatically lower latency (140×) than LLM-based alternatives.

The architecture's strongest capability — adversarial falsification — achieves near-perfect accuracy as a minimal configuration, suggesting that structured verification is a more fundamental reasoning primitive than previously appreciated. The evidence text paradox — that structural representations outperform textual ones for causal narration — supports the core thesis that reasoning substrates should be structural, not textual.

Future work focuses on three directions: (1) SOTA-level comparative evaluation against published GraphRAG and multi-hop reasoning benchmarks, (2) cross-domain transfer validation to confirm domain-agnostic architectural claims, and (3) scaling the algebraic behavioral scoring to become

a primary discriminator rather than a structural tiebreaker.

References

[1] Brown, T. B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[2] Garcez, A. d'A., et al. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv:1905.06088.

[3] Goodfellow, I. J., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

[4] Hamilton, W. L. (2020). Graph Representation Learning. Morgan & Claypool Publishers.

[5] Hogan, A., et al. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), 1-37.

[6] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899.

[7] Ji, Z., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[8] Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

[9] Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.

[10] Pearl, J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press.

[11] Peirce, C. S. (1931-1958). Collected Papers of Charles Sanders Peirce, Vols. 1-8. Harvard University Press.

[12] Touvron, H., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[13] He, H., & Thinking Machines Lab. (2025). Defeating nondeterminism in LLM inference. Thinking Machines Lab: Connectionism. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/

[14] Kamradt, G., & Chollet, F. (2025). ARC Prize 2025 results and analysis. ARC Prize. https://arcprize.org/blog/arc-prize-2025-results-analysis

[15] Mao, H., Chen, Z., Tang, W., Zhao, J., Ma, Y., Zhao, T., Shah, N., Galkin, M., & Tang, J. (2024). Position: Graph foundation models are already here. arXiv preprint arXiv:2402.02216.

[16] Reese, H. (2026). Yann LeCun's new venture is a contrarian bet against large language models. MIT Technology Review. https://www.technologyreview.com/2026/01/22/1131661/yann-lecuns-new-venture-ami-labs/

[17] Novikov, A., Vũ, N., Eisenberger, M., et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131.

[18] Sutton, R. S. (2019). The Bitter Lesson. Incomplete Ideas. http://www.incompleteideas.net/IncIdeas/BitterLes

[19] Darlow, L., Crowley, E. J., Storkey, A., & Hospedales, T. (2025). Continuous Thought Machines: Adaptive test-time computation for abstract reasoning. arXiv preprint arXiv:2505.05522.

[20] Silver, D., & Sutton, R. S. (2025). Welcome to the Era of Experience. In Designing an Intelligence (forthcoming). MIT Press. https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

[21] Georgiev, B., Gómez-Serrano, J., Tao, T., & Wagner, A. Z. (2025). Mathematical exploration and discovery at scale. arXiv preprint arXiv:2511.02864. https://doi.org/10.48550/arXiv.2511.02864

[22] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33.

[23] Edge, D., Trinh, H., Cheng, N., et al. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Appendix A: Formal Definitions Summary

Symbol	Definition
$f = (s, p, o, t, E)$	Atomic Behavioral Signal (subject, predicate, object, timestamp, evidence set)
$\mathsf{h}(f)$	High-dimensional algebraic encoding of signal f
$M(f, C)$	Meaning function: signal × context → interpretant
$R(f_i, f_j) = \phi(\mathsf{h}(f_i), \mathsf{h}(f_j))$	Algebraic semiotic relationship between signals
$\Delta_T(f, t_i, t_j)$	Temporal differential of signal f between times t_i and t_j
$\Delta_S(f)$	Spatial differential: relational context set of signal f
c	Concept: higher-order abstraction mediating entity relationships
$\theta(c, t)$	Phase of concept c at time t
$a = (f_{expected}, t, \epsilon)$	Absence signal: expected behavior not observed within tolerance

Symbol	Definition
σ	Convergence verification across four dimensions
$I = \{i_1, i_2, \ldots, i_k\}$	Semiotic chain: sequence of recursive interpretants
R	Pairwise relationship matrix recoverable without sequential evaluation

Appendix B: Evaluation Dataset Categories

The 109-query evaluation dataset spans six categories designed to test distinct reasoning capabilities:

- Multi-hop causal (30 queries): Require tracing causal chains across 2-4 entity boundaries. Example: "How does Company A's acquisition of Company B affect Company C's competitive position?"
- Velocity/temporal (32 queries): Require detecting rates of change, acceleration, and temporal patterns. Example: "Which company is accelerating autonomous trucking deployment fastest?"
- Adversarial falsification (32 queries): Present false or partially true premises for the system to validate or disprove. Example: "Is it accurate that safety concerns have had no impact on deployment timelines?"
- Strategic reasoning (7 queries): Require synthesizing multi-factor causal explanations for strategic decisions. Example: "Why did Company D prioritize manufacturing partnerships over vertical integration?"
- Ecosystem impact (5 queries): Require tracing cross-entity cascading effects. Example: "How does Company E's shutdown benefit Company F?"
- Partial answer (3 queries): Edge cases testing graceful degradation when data is incomplete.

Appendix C: Domain-Specific Signal Types

The evaluation knowledge graph contains 18,500+ behavioral signals across the following categories:

- Corporate actions: Partnerships, acquisitions, divestitures, leadership changes, organizational restructuring
- Regulatory signals: Investigations, approvals, suspensions, permit grants/revocations
- Technology development: Product launches, testing milestones, patent filings, technology pivots
- Financial signals: Funding rounds, IPOs, capital allocation shifts, valuation changes

- Workforce dynamics: Hiring patterns, team expansions/contractions, key personnel movements
- Competitive dynamics: Market entry/exit, competitive responses, strategic repositioning

Each signal is stored as a semantically atomic tuple with full provenance tracking, temporal anchoring, and structural encoding.

Appendix D: Ablation Testing Configurations and Results

The architecture was evaluated across multiple isolated variable tests (ablation runs) to determine the specific contribution of individual components. Scores are derived from an independent LLM judge evaluating output on a 1-5 scale across four dimensions.

Run	Configuration	n	Retrieval	Draft	Edit	Final
A1	Full Pipeline (Multi-hop)	25	4.2	4.2	4.5	4.0
A2	Baseline (No HDC/Relational)	5	4.8	4.2	4.4	3.8
A3	Relational Traversal Only	5	4.8	4.4	4.8	4.4
A5	Full + No Evidence Text	5	4.8	4.4	4.8	4.6
B1	Full (Velocity/Adversarial)	48	4.5	4.4	4.6	4.3
B3	Baseline (Adversarial only)	8	5.0	4.9	4.8	4.9
C1	Algebraic Unified Pipeline	33	4.3	4.4	4.6	4.4
C2	LLM Extraction Pipeline	33	4.3	4.2	4.5	4.1

Table D.1: Summary of key ablation run results comparing baseline, full pipeline, and specialized configurations. Note: n represents the number of queries evaluated in the run.