STRUCTURED ABDUCTIVE-DEDUCTIVE-INDUCTIVE REASONING FOR LLMS VIA ALGEBRAIC INVARIANTS

Sankalp Gilda*
DeepThought Solutions
sankalp.gilda@gmail.com

Shlok Gilda
Department of Computer Science
University of Florida
shlok.gilda@ufl.edu

ABSTRACT

Large language models exhibit systematic limitations in structured logical reasoning: they conflate hypothesis generation with verification, cannot distinguish conjecture from validated knowledge, and allow weak reasoning steps to propagate unchecked through inference chains. We present a symbolic reasoning scaffold that operationalizes Peirce's tripartite inference—abduction, deduction, and induction—as an explicit protocol for LLM-assisted reasoning. The framework enforces logical consistency through five algebraic invariants (the *Gamma Quintet*), the strongest of which—the Weakest Link bound—ensures that no conclusion in a reasoning chain can exceed the reliability of its least-supported premise. This principle, independently grounded as weakest link resolution in possibilistic logic and empirically validated for chain-of-thought reasoning, prevents logical inconsistencies from accumulating across multi-step inference. We verify all invariants through a property-based testing suite of 100 properties and 16 fuzz tests over 10^5+ generated cases, providing a verified reference implementation of the invariants suitable as a foundation for future reasoning benchmarks.

1 INTRODUCTION

Large language models have achieved strong performance on reasoning tasks, yet they exhibit systematic failures in structured logical reasoning. On combinatorial logic puzzles, accuracy degrades sharply as problem complexity increases—a phenomenon termed the "curse of complexity" (Lin et al., 2025)—revealing that LLMs struggle to maintain logical consistency across multi-step inference. More fundamentally, chain-of-thought explanations are only 25–39% faithful to the model's actual computation (Anthropic, 2025), meaning the reasoning traces that users rely upon to assess answer quality frequently do not reflect the process that produced the answer.

These failures stem from a structural conflation of three distinct inference modes that have been recognized since Peirce's foundational work on the logic of science (Peirce, 1878):

- **Abduction** (hypothesis generation): generating candidate explanations for observed phenomena.
- **Deduction** (logical verification): deriving necessary consequences from premises.
- **Induction** (empirical validation): testing predictions against observations.

In a single autoregressive pass, LLMs perform all three simultaneously: generating hypotheses, checking them against implicit constraints, and marshaling evidence—without marking which mode is active at any step. Chain-of-thought prompting (Wei et al., 2022) approximates deduction but provides no formal guarantees. Self-consistency voting (Wang et al., 2023) approximates induction but averages over candidates rather than validating them. Process reward models (Lightman et al., 2024) score individual steps but do not enforce structural constraints across reasoning chains. None of these approaches explicitly separate inference modes, formally bound how weak premises propagate, or maintain a persistent, auditable knowledge state across interactions.

*Corresponding author.

We present a symbolic reasoning scaffold with three contributions mapped to the workshop's topics of interest:

1. **An explicit ADI protocol** separating abduction, deduction, and induction into distinct, auditable phases with epistemic state tracking (Topic 1: deduction, induction, and abduction).
2. **Five algebraic invariants** (the Gamma Quintet) that formally constrain how reliability propagates through reasoning chains, with the Weakest Link bound preventing logical inconsistencies from accumulating (Topics 2 and 3: symbolic reasoning and avoiding contradictions).
3. **A property-based verification benchmark** of 100 test properties and 16 fuzz tests validating that implementations preserve these invariants across 10^5+ randomly generated cases (Topic 5: benchmarks and evaluation).

The framework operates as an external symbolic system alongside the LLM (Topic 4: external logical solvers), maintaining a knowledge graph with formal consistency guarantees while the LLM handles natural language reasoning.

The Weakest Link bound—ensuring that no aggregated conclusion exceeds the reliability of its least-supported input—has independent theoretical grounding in possibilistic logic (Dubois & Prade, 2025), where it is known as "weakest link resolution," and recent empirical validation demonstrating that chain-of-thought reliability is bounded by its weakest step (Jacovi et al., 2024). This convergence from algebraic specification, possibility theory, and empirical measurement provides triangulated justification for the framework's central consistency constraint.

2 Symbolic Knowledge Representation

We represent knowledge claims as structured symbolic objects carrying a three-dimensional descriptor: **Formality (F)**, measuring rigor of expression; **Scope (G)**, bounding the context where the claim applies; and **Reliability (R)**, a computed consistency score on $[0, 1]$. These descriptors let the framework track the epistemic status of LLM-generated conclusions.

Table 1: Formality levels and reliability ceilings.

Level	Description	Ceiling
F0	Informal (anecdotal, authority-based)	70%
F1	Structured (ADRs, explicit trade-offs)	85%
F2	Empirical (benchmarks, load tests)	95%
F3	Formal (proofs, model checking, type-checked)	99%

F3 admits machine-checkable verification; recent work (Perrier, 2026) proposes extending this to LLM reasoning via the Curry-Howard correspondence, treating reasoning traces as type-checkable proof terms. The F3 ceiling is 99% rather than 100%, reflecting that formal proofs depend on unverified proof checkers (Pollack, 1998). The ordering invariant $C_{F_0} < C_{F_1} < C_{F_2} < C_{F_3}$ ensures that higher formality always permits higher reliability. Orthogonal epistemic layers mark symbolic state: L0 (conjecture, 35%), L1 (substantiated, 75%), L2 (corroborated, 100%), corresponding to the abductive, deductive, and inductive phases respectively.

Faithfulness ceiling. Chain-of-thought explanations are only 25–39% faithful to the model's actual computation (Anthropic, 2025), suggesting LLM-generated evidence should cap at F1 with faithfulness as a limiting factor: applying the framework's own min-aggregation principle, the effective ceiling is $\min(0.85, 0.39) = 0.39$ for current models.

The complete effective reliability formula combines all constraints via consistency-preserving inference:

$$R_{\text{eff}} = \min\big(\min_i R_{\text{adj}}(e_i),\ \min_j \max\big(0,\ R_{\text{eff}}(d_j) - \text{CL}_j\big),\ C_L,\ C_F\big) \tag{1}$$

where $R_{\text{adj}}(e_i)$ is the adjusted score for evidence i (including temporal decay), CL_j is the congruence penalty for dependency j, and C_L, C_F are the layer and formality ceilings respectively. The

$\max(0, \cdot)$ floor ensures all terms stay within $[0, 1]$ even when a congruence penalty exceeds the incoming reliability. The nested $\min$ structure ensures that no individual component can inflate the aggregate—a property we formalize as the WLNK invariant in Section 4.

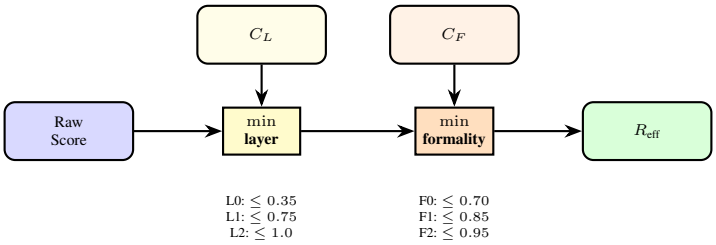

Figure 1: Dual ceiling constraint. The propagated reliability score passes through two successive $\min$ gates: the epistemic layer ceiling, then the formality ceiling. An L0 claim at F3 is capped at $\min(0.35, 1.0) = 0.35$. Because the two ceilings are independent, the tighter always dominates: L0 claims are layer-bounded regardless of formality, and fully corroborated L2 claims ($C_L = 1.0$) are formality-bounded regardless of layer. Both dimensions become active only at intermediate layers.

Evidence transfers across contexts incur congruence penalties: CL3 (same context, no penalty), CL2 (similar, -0.1), CL1 (different, -0.4). Evidence with scope match "none" is excluded entirely.

Verification credibility. Evidence scores are adjusted by a verification method multiplier: self-reported ($\times 0.60$), script-attached ($\times 0.85$), externally reviewed ($\times 0.95$), and executed-and-verified ($\times 1.00$). This prevents self-reported evidence from achieving the same influence as independently verified results.

3 The ADI Reasoning Protocol

Current LLM reasoning techniques treat inference as a monolithic process: a single autoregressive pass produces hypotheses, justifications, and conclusions without distinguishing the epistemic character of each step. We propose the *Abduction–Deduction–Induction* (ADI) protocol, which decomposes LLM reasoning into three explicitly labeled inference modes, each with distinct epistemic commitments and verification requirements.

3.1 Peircean Inference and LLM Reasoning

Charles Sanders Peirce classified inference into three irreducible modes (Peirce, 1878; 1998): *abduction* (inference to the best explanation), *deduction* (necessary inference from premises), and *induction* (inference from observed instances). Peirce argued that these modes are distinct epistemic operations, each with its own warrant structure and failure modes (Magnani, 2009).

Contemporary LLM reasoning conflates all three in ways that obscure the epistemic status of generated claims:

- **Chain-of-thought prompting** (Wei et al., 2022) approximates deductive reasoning by eliciting step-by-step derivations, but provides no formal guarantee that steps follow from premises. The model may produce a sequence that *reads* deductively while smuggling in abductive leaps—plausible hypotheses presented as logical necessities.
- **Self-consistency voting** (Wang et al., 2023) approximates induction by sampling multiple reasoning paths and selecting the majority answer. However, it *averages* over paths rather than *validating* any single path against empirical evidence. Agreement among hallucinated chains does not constitute inductive support.
- **Neither technique explicitly marks which inference mode is active at each step.** A chain-of-thought trace may begin with an abductive hypothesis ("this likely fails because…"), shift to deductive reasoning ("given X, it follows that…"), and conclude with an inductive generalization ("in similar cases, we observe…")—all without the model or user being aware of the transitions.

This conflation has a direct analogue in dual-process theory (Kahneman, 2011): abduction is fast and pattern-driven (System 1), while deduction and induction are deliberate and rule-governed (System 2). The ADI protocol makes this handoff explicit and auditable.

3.2 THE ADI PROTOCOL

The protocol organizes reasoning as a cycle of three phases, each producing claims at a successively higher epistemic layer. We describe each phase, its epistemic commitments, and its realization in LLM-assisted reasoning.

3.2.1 ABDUCTION: HYPOTHESIS GENERATION (L0—CONJECTURE)

Given an anomaly, question, or design problem, the reasoning process begins by generating candidate hypotheses. These are conjectures: plausible but unverified explanations inferred from incomplete evidence. In the framework's epistemic hierarchy, abductive claims are labeled L0 (conjecture) with reliability capped at 35%, reflecting their status as unvalidated proposals.

In LLM reasoning, abduction corresponds to the *hypothesis generation step*. The model draws on patterns in its training distribution to propose explanations. Multiple candidates are not merely tolerated but encouraged: the strength of abduction lies in generating a diverse candidate set, not in premature commitment to a single explanation.

Example. Consider an LLM-assisted analysis of a retrieval system. The model proposes: "Increasing the context window from 4K to 32K tokens would improve retrieval accuracy, based on patterns showing correlation between context length and answer quality in multi-document settings." At this stage, the claim is plausible—it aligns with known properties of attention mechanisms—but it is *unverified*. It carries an L0 label and cannot propagate as established knowledge.

Epistemic commitment. Abduction makes no truth claim. It asserts only that a hypothesis is *worth investigating*. The failure mode is not generating a wrong hypothesis (that is expected) but treating an L0 conjecture as if it were established fact—precisely the failure mode of unconstrained autoregressive generation.

3.2.2 DEDUCTION: LOGICAL VERIFICATION (L0→L1—SUBSTANTIATION)

Deductive verification checks whether a hypothesis is logically consistent with the current body of validated knowledge. The question is not "is this true?" but "does this contradict anything we already know?" A hypothesis that survives deductive scrutiny is promoted to L1 (substantiated): it is internally consistent but not yet empirically confirmed.

In LLM reasoning, deduction is where contradictions with prior validated knowledge are detected. This phase serves as a logical filter, catching hypotheses that are plausible in isolation but incompatible with established constraints.

Critical structural requirement. Deductive verification requires an *external* check—a human reviewer, a separate LLM, or a specialized NLI verifier—that provides a verdict the framework records with provenance. The framework enforces algebraic constraints on how that verdict propagates, but does not itself parse natural language for semantic contradictions. A system cannot reliably verify the consistency of its own outputs against its own knowledge, because the same biases that produced the hypothesis contaminate the verification. Step-by-step verification (Lightman et al., 2024) partially addresses this by training separate verifier models, but the principle is more general: the verifier must have access to a constraint set that is independent of the generation process.

Example (continued). The 32K context hypothesis is checked against the validated knowledge base. Deductive analysis reveals: "The hypothesis is logically consistent with known attention mechanism properties. However, it contradicts a previously validated finding (L2) that retrieval accuracy plateaus beyond 16K tokens on our specific document corpus due to the long-tail distribution of relevant passages." The contradiction forces one of three outcomes: (a) refine the hypothesis to account for the constraint (e.g., restrict the claim to multi-hop queries), (b) challenge the prior L2

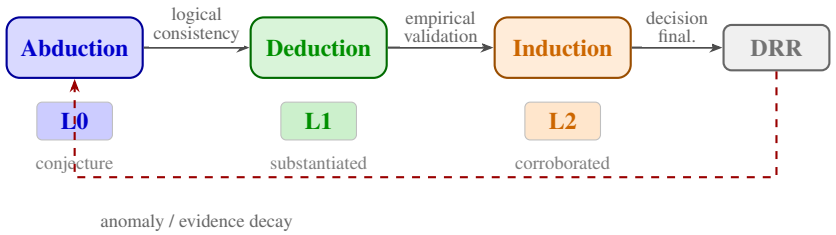

Figure 2: The ADI reasoning cycle. Abduction generates conjectures (L0), Deduction verifies logical consistency (L1), and Induction validates empirically (L2). Finalized decisions become Design Rationale Records (DRRs). Evidence decay or new anomalies re-enter the cycle.

finding with new evidence, or (c) discard the hypothesis. The deductive phase does not resolve the question—it sharpens it.

Epistemic commitment. A claim at L1 is asserting: "this is logically consistent with everything we have validated so far." It is *not* asserting truth. The failure mode is passing a hypothesis that contains a hidden contradiction—which is why the external verification requirement is structural, not optional.

3.2.3 Induction: Empirical Validation (L1→L2—Corroboration)

Inductive validation tests the hypothesis against empirical evidence. An L1 claim that passes empirical validation is promoted to L2 (corroborated): it is both logically consistent and empirically supported, within a specified scope.

In LLM reasoning, this is where claims are tested against reality—not against the model's beliefs about reality. Induction requires running experiments, collecting data, or testing predictions on held-out observations.

Example (continued). The refined hypothesis—"32K context improves retrieval accuracy specifically for multi-hop queries"—is tested on a held-out corpus. The benchmark confirms: retrieval F1 improves from 0.72 to 0.81 for multi-hop queries, with no statistically significant gain for single-hop queries. The hypothesis is empirically validated with an explicit scope constraint: the L2 claim specifies the conditions under which it holds.

Epistemic commitment. A claim at L2 asserts: "this has been observed to hold under specified conditions." It does *not* assert universal truth. The scope constraint is essential: inductive claims without scope constraints are unfalsifiable and therefore epistemically vacuous. Every L2 claim in the framework carries an explicit validity window and scope specification.

3.3 Design Rationale Records and the Reasoning Audit Trail

After completing the ADI cycle, a finalized decision is recorded as a *Design Rationale Record* (DRR): a structured decision record augmented with the evidence chain that produced it. Each DRR captures:

1. The **inference mode** at each step (abduction, deduction, or induction) and the epistemic layer of the resulting claim.
2. The **evidence** supporting each transition, with explicit provenance—what was checked, against what constraints, with what outcome.
3. The **reliability score** of the final claim, computed via the weakest-link aggregation principle described in Section 4.2: the reliability of a conclusion equals the reliability of its least reliable supporting premise.
4. The **scope specification** bounding the conditions under which the claim is valid, and a **validity window** specifying when the evidence must be re-evaluated.

This audit trail directly addresses two concerns identified in the call for papers. First, it provides a mechanism for *detecting and eliminating logical contradictions* across multiple reasoning steps (Topic 3): because every claim is labeled with its inference mode and epistemic layer, contradictions between claims at different layers are structurally detectable rather than hidden in natural language traces. Second, it provides a *benchmark-amenable record* of the reasoning process (Topic 5): the audit trail can be evaluated for logical correctness, evidential completeness, and scope consistency independently of the domain-specific content.

The Transformer Mandate. We introduce a structural constraint on the ADI cycle: the entity that finalizes a decision (produces the DRR) must be external to the generation loop. An LLM may propose hypotheses (abduction) and gather supporting evidence (induction), but ratification—the transition from "candidate conclusion" to "accepted decision"—requires external verification. This prevents a failure mode where an autonomous system bootstraps confidence in its own outputs by citing its own prior generations (Ferrario et al., 2026). The mandate is architectural, not a policy preference: it is enforced by the protocol structure, which requires the ratifying entity to have access to constraints independent of the generation process.

4 Algebraic Consistency Invariants

4.1 The Gamma Quintet

We specify five algebraic invariants that any consistency-preserving inference operator $\Gamma : \mathcal{P}([0,1]) \to [0,1]$ must satisfy. These invariants constrain how reliability scores compose across multi-step reasoning, regardless of evaluation order, evidence arrangement, or chain length.

1. **IDEM** (Idempotence): $\Gamma([x]) = x$. A single premise retains its original reliability.
2. **COMM** (Commutativity): $\Gamma([a,b]) = \Gamma([b,a])$. The order in which premises are evaluated is logically irrelevant.
3. **LOC** (Locality): Changing premise E affects only conclusions whose derivation graph includes E. Isolated subproofs remain stable.
4. **WLNK** (Weakest Link): $\Gamma(S) \leq \min(S)$. No inference chain can be more reliable than its least reliable premise.
5. **MONO** (Monotonicity): $a \leq a'$ implies $\Gamma([a,b]) \leq \Gamma([a',b])$. Strengthening a premise cannot weaken a conclusion.

LOC is a system-level invariant governing how updates propagate through the derivation graph, distinct from the pointwise algebraic properties (IDEM, COMM, WLNK, MONO); we include it in the Quintet because the framework's consistency guarantees depend on it jointly with the operator properties.

Theorem 1 (Quintet Satisfaction)**.** *The Gödel t-norm $\Gamma(S) = \min(S)$ (Hájek, 1998) satisfies all five invariants.*

Proof. IDEM: $\min([x]) = x$. COMM: $\min$ is symmetric over its arguments. LOC: $\min$ depends only on the multiset elements; modifying a value outside the multiset has no effect. WLNK: $\min(S) \leq \min(S)$ trivially. MONO: if $a \leq a'$, then $\min(a,b) \leq \min(a',b)$ by case analysis on b. □

Theorem 2 (Idempotent Uniqueness (Klement et al., 2000; Metcalfe, 2005))**.** *Among continuous t-norms on $[0,1]$, the Gödel t-norm is the unique idempotent t-norm. This follows from the Klement–Mesiar–Pap characterization of continuous t-norms (Klement et al., 2000): among continuous t-norms, those satisfying t-norm idempotence ($\Gamma(x,x) = x$ for all $x \in [0,1]$) are exactly those equal to $\min$. When we require operators to satisfy both IDEM (on singletons) and WLNK as an upper bound, restriction to the continuous t-norm family picks out $\min$ uniquely. The Quintet invariants are the contribution; $\min$ is one valid instantiation. A future learned aggregator satisfying the Quintet while permitting confidence accumulation for independent evidence would also be valid.*

Table 2: Aggregation operators and Gamma Quintet compliance.

Operator	IDEM	COMM	WLNK	MONO	Use Case
$\min$ (Gödel)	✓	✓	✓	✓	Serial chains
Product	✓	✓	$\sim$	✓	Independent evidence
Mean	$\times$	✓	$\times$	✓	Not recommended
$\max$	✓	✓	$\times$	✓	Not recommended

Product aggregation satisfies WLNK when all scores are ≤ 1 (since $a \cdot b \leq \min(a, b)$ for $a, b \in [0, 1]$), making it a valid relaxation for genuinely independent evidence. Mean violates both IDEM and WLNK, which has direct consequences for logical consistency: three weak premises at $R = 0.4$ would average to 0.4, masquerading as comparable to a single controlled experiment at 0.4—a conflation that obscures the distinction between quantity and quality of evidence.

4.2 WLNK as a Logical Consistency Rule

The WLNK invariant is a *logical consistency constraint*, not a conservative aggregation heuristic. In a deductive argument chain $A \Rightarrow B \Rightarrow C$, the conclusion C cannot be more reliable than the weakest premise in the chain. This is a structural property of valid inference: if any link in a derivation is uncertain, the derived conclusion inherits that uncertainty.

Possibilistic logic foundation. WLNK directly instantiates Dubois and Prade's weakest link principle from possibilistic logic: "the strength of an inference chain is that of the least certain formula involved" (Dubois & Prade, 1988). In possibilistic logic, each formula carries a necessity degree, and the resolution rule propagates the minimum degree through inference steps. Our WLNK invariant lifts this principle from propositional possibilistic logic to a general algebraic constraint on any consistency-preserving operator. Recent work (Dubois & Prade, 2025) extends these foundations to graded reasoning under uncertainty.

As a logical inference rule. We can state WLNK as an inference rule analogous to modus ponens:

$$\frac{P_1 : r_1 \qquad P_2 : r_2 \qquad \cdots \qquad P_n : r_n \qquad P_1, \ldots, P_n \vdash C}{C : \min(r_1, \ldots, r_n)}$$

This rule says: if a conclusion C is derived from premises $P_1, \ldots, P_n$ with respective reliability scores $r_1, \ldots, r_n$, then C's reliability is bounded by the minimum. Compare this to standard modus ponens, which propagates truth values; here we propagate *graded* epistemic status through the same logical structure.

Empirical validation for LLM reasoning. Jacovi et al. (Jacovi et al., 2024) provide direct empirical evidence for WLNK in the context of chain-of-thought reasoning. They demonstrate that the reliability of multi-step LLM reasoning is bounded by the weakest individual step, confirming that WLNK is an empirically observable property of how reasoning chains degrade, not just a theoretical desideratum.

Contradiction detection. Consider an LLM system that answers question Q_1 based on premise P with $R(P) = 0.9$, and later answers question Q_2 based on premise P' that contradicts P, with $R(P') = 0.3$. Without WLNK, a system using arithmetic mean might assign both answers moderate reliability (e.g., $\frac{0.9+0.3}{2} = 0.6$), masking the contradiction. Under WLNK, the system tracks that any conclusion depending on both P and P' is capped at $\min(0.9, 0.3) = 0.3$—the reliability of the weakest (and contradicting) premise. The low aggregate signals that the knowledge base is inconsistent and requires resolution.

Worked example. Suppose an LLM constructs a three-step logical argument:

S1. "Python's GIL prevents true parallelism" ($R = 0.95$, well-established fact)

S2. "Therefore, CPU-bound tasks cannot benefit from threading" ($R = 0.85$, valid deduction from S1)

S3. "Therefore, all Python programs should use multiprocessing" ($R = 0.40$, overgeneralization—ignores I/O-bound workloads, async alternatives)

Under WLNK: $R_{\text{chain}} = \min(0.95, 0.85, 0.40) = 0.40$. The weak final step correctly caps the entire argument. Under arithmetic mean: $R_{\text{chain}} = \frac{0.95+0.85+0.40}{3} = 0.73$—a score that would classify this argument as moderately reliable, hiding the logical overreach in S3. WLNK surfaces the single weak step rather than diluting it across the chain.

Quadruple-triangulated justification. The choice of $\min$ is supported by four independent lines: (1) *t-norm theory*—the Gödel t-norm is the unique continuous idempotent t-norm (Klement et al., 2000; Metcalfe, 2005) (Theorem 2); (2) *possibility theory*—Dubois and Prade's necessity-based inference propagates the minimum through deductive chains (Dubois & Prade, 1988; 2025); (3) *empirical measurement*—Jacovi et al. (Jacovi et al., 2024) confirm that chain-of-thought reliability is bounded by the weakest step; (4) *algebraic specification*—the Gamma Quintet derives WLNK from first principles (Theorem 1). Four independent lines arriving at the same operator is difficult to dismiss as coincidence.

Two-tier evidence aggregation. Applying flat $\min$ across heterogeneous evidence is epistemically unsound: a must-pass gate ("does the code compile?") should aggregate differently from corroborating performance metrics. We extend to a two-tier architecture: *Tier 1* classifies evidence by epistemic role (structural gates, performance metrics, quality assurance, etc.) and aggregates within each role using a role-appropriate operator (gates use $\min$; quality reviews use probabilistic sum; performance metrics use a conservative OWA). *Tier 2* applies $\min$ across role-level scores, preserving WLNK: if any gate fails, the overall score is zero. This decomposition is WLNK-preserving because each within-role operator is a t-norm and the cross-role combiner is $\min$.

5 The Framework as External Reasoning Scaffold

The framework is an external symbolic system alongside the LLM, not a modification to the model's internal reasoning (Gilda & Gilda, 2026). The LLM generates natural-language hypotheses and arguments; the framework maintains a symbolic knowledge graph that tracks epistemic status, dependency structure, and formal invariants over those claims. The interface is simple: the LLM proposes claims, and the framework records their provenance, assigns epistemic layers, and enforces consistency constraints across the growing knowledge graph.

This architecture targets four specific limitations of LLM reasoning:

Conflated inference modes. LLMs routinely mix hypothesis generation with verification and evidence gathering within a single response, producing outputs that appear rigorous but conflate logically distinct operations. The ADI protocol (Section 3) forces explicit separation: abduction generates candidate claims at L0, deduction checks logical consistency before promoting to L1, and induction requires empirical evidence for promotion to L2. Each mode has distinct preconditions and produces claims at a specific epistemic layer, making the inference type of every claim auditable.

Inconsistent reliability across responses. When an LLM generates multiple claims that depend on shared premises, it assigns no consistent measure of confidence across the dependency graph. The WLNK invariant (Section 4) enforces global consistency: no composite claim can exceed the reliability of its weakest supporting evidence, and this constraint propagates transitively through the entire dependency structure.

No self-correction. The Transformer Mandate is an architectural constraint preventing self-promotion loops: the agent that generates a claim cannot also provide the evidence that promotes it. Layer promotion requires external verification, enforced at the data model level rather than by prompt instruction.

Stale knowledge. Every piece of evidence in the knowledge graph carries a `valid_until` timestamp. When evidence expires, dependent claims are automatically flagged for re-validation. Validity periods are formality-dependent: informal observations (F0) expire faster than empirical measurements (F2), reflecting the intuition that rigorously gathered evidence remains relevant longer.

5.1 Comparison with Existing Approaches

Tool-augmented LLMs extend computational *capabilities* but impose no structure on the *reasoning process*. Neuro-symbolic systems such as Logical Neural Networks (Riegel et al., 2020) and DeepProbLog (Manhaeve et al., 2021) provide strong guarantees but require full domain formalization. Our approach occupies a middle ground: lightweight symbolic structure—epistemic layers, dependency tracking, algebraic invariants—over natural-language claims, without requiring translation into formal logic. The framework tracks epistemic status and structural relationships rather than semantic content, so it applies to any domain without requiring domain-specific formalization.

5.2 Architectural Pattern

The three components map to the ADI phases: (1) **LLM as hypothesis generator** (Abduction)—proposes candidate explanations as L0 conjectures; (2) **Framework as consistency checker** (Deduction)—verifies logical consistency against existing knowledge and WLNK before promoting to L1; (3) **Empirical tools as validators** (Induction)—test runners, benchmarks, and human review provide evidence for L2 promotion. No single component controls the full epistemic lifecycle of a claim.

6 Property-Based Verification Benchmark

6.1 Specification-Verification Approach

Algebraic specification (Section 4) establishes correctness-by-construction; property-based testing (Claessen & Hughes, 2000; Goldstein et al., 2024) establishes correctness-in-implementation, verifying that the actual code preserves invariants despite floating-point arithmetic and concurrent access. Together they constitute a verified reference implementation whose test suite can serve as a starting point for future reasoning benchmarks.

We verify 100 property-based tests across five specification areas, plus 16 fuzz tests, each exercising 10^5+ randomly generated cases (Table 3). The largest group (57 tests) targets the R_{eff} calculator: bounds/WLNK enforcement, ceiling caps, monotonicity, dependency propagation, two-tier evidence aggregation, and preset inheritance. Scope lattice tests (16) verify bounded lattice axioms and parse–serialize round-trips. FSM tests (11) verify phase ordering, reachability, and role enforcement. Graph topology tests (6) verify WLNK propagation through deep chains, diamonds, and mixed topologies. Inspector tests (10) verify BFS traversal correctness. All tests follow the industrial-scale PBT methodology of Arts et al. (Arts et al., 2006).

6.2 PBT as Design Exploration

Beyond verification, PBT surfaces implicit assumptions. For example, the property "every phase can reach IDLE" initially failed for OPERATION because no back-transition was defined—not a bug, but an undocumented design assumption. The resolution was to classify OPERATION as intentionally terminal and exclude terminal states from the property. This illustrates PBT's role as a consistency auditor: the invariants themselves must be internally consistent before they can meaningfully constrain external claims.

6.3 Property Inventory

7 Related Work and Limitations

7.1 Related Work

LLM reasoning approaches. Chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2023), and process reward models (Lightman et al., 2024) improve output quality but lack formal invariants guaranteeing algebraic properties such as monotonicity or weakest-link propagation. Our framework provides externally verifiable constraints on reasoning chain integrity. Recent benchmarks reveal a "curse of complexity" where accuracy degrades sharply with problem scale (Lin

Table 3: Property-based verification inventory for logical consistency invariants. All 100 properties are randomized checks over 10^5+ cases; all 16 fuzz tests use corpus-guided mutation (The Go Authors, 2022).

Specification Area	Verified Properties	Count
R_{eff} calculator	Bounds, WLNK, ceilings, monotonicity, presets, two-tier	57
Scope algebra	Lattice axioms, match, round-trip	16
Epistemic FSM	No skipping, reachability, determinism, ordering	11
Graph topology	Deep chains, diamonds, mixed serial/parallel	6
Dependency inspector	BFS correctness, deduplication, layer preservation	10
Fuzz tests	IEEE 754 boundaries, parser round-trips, concurrency	16
Total		**116**

et al., 2025); our framework externalizes structural constraints that are enforced symbolically and therefore do not degrade with problem size.

Weakest link in reasoning chains. Jacovi et al. (2024) empirically demonstrate that chain-of-thought reliability is bounded by its weakest step. Our framework formalizes this as an algebraic invariant ($R_{\text{eff}} \leq \min_i R_i$) enforced by construction and verified via property-based testing across 10^5+ generated inputs, transforming an empirical observation into a provable structural guarantee.

Possibilistic logic. Possibilistic logic (Dubois & Prade, 1988; 2025) provides the theoretical foundation for our weakest-link aggregation. Possibilistic inference follows the "weakest link resolution" rule: the necessity of a derived conclusion equals the minimum necessity of formulas in the derivation chain (Dubois & Prade, 2025). Our WLNK bound is a direct application of this principle, where the ADI cycle produces claims whose reliability propagates according to possibilistic semantics.

Argumentation frameworks. Abstract (Dung, 1995) and structured argumentation (Besnard & Hunter, 2008) establish semantics for argument acceptability. Our framework shares the graph-of-claims structure but tracks *reliability*, *temporal validity*, and *scope* rather than resolving attack relations.

Neuro-symbolic reasoning. Logical Neural Networks (Riegel et al., 2020) and DeepProbLog (Manhaeve et al., 2021) require full domain formalization. Our framework occupies a lighter-weight middle ground: symbolic structure without complete logical formalization, applicable as an external scaffold for general-purpose LLM reasoning.

7.2 Limitations

No machine-checked proofs. Our verification relies on property-based testing, not theorem provers (Coq (The Coq Development Team, 1989)) or model checkers (TLA+ (Lamport, 2002)). PBT provides high confidence through volume (10^5+ cases per invariant) but not exhaustive guarantees.

End-to-end LLM evaluation is preliminary. An evaluation harness integrating the framework with GPT-4o on ML engineering tasks (AIRS-Bench (Xiao et al., 2026)) has been developed, with preliminary results showing reduced execution errors under framework-guided reasoning. However, evaluation on dedicated logical reasoning benchmarks such as ZebraLogic (Lin et al., 2025) and FOLIO (Han et al., 2022) remains future work. Controlled experiments with the current codebase are ongoing.

Ceiling values are policy defaults. Specific ceiling percentages (e.g., $C_{F_0} = 70\%$, $C_{L_0} = 35\%$) are configurable defaults, not empirically calibrated. The implementation supports per-context configuration overrides, allowing domain-specific calibration of ceilings, congruence penalties, and evidence age thresholds. The *ordering invariant* ($C_{F_0} < C_{F_1} < C_{F_2} < C_{F_3}$) is verified by PBT regardless of specific values, but empirical calibration against domain-specific outcomes remains future work.

ADI requires structured interaction. Single-pass inference cannot use the ADI protocol; it requires multi-turn interaction or a multi-agent architecture—a structural overhead that is the cost of formal guarantees.

Open questions: (1) Can WLNK serve as a differentiable training constraint? (2) Can the ADI cycle be realized as a multi-agent protocol with specialized agents per inference mode? Preliminary work on programmatic tool calling—where an LLM directly queries the knowledge graph via structured function calls—suggests viability. (3) Can epistemic and aleatoric uncertainty be decomposed within R_{eff} (Hüllermeier & Waegeman, 2021)?

8 Conclusion

We presented a symbolic reasoning scaffold that operationalizes Peirce's tripartite inference as an explicit ADI protocol for LLM-assisted reasoning, with five algebraic invariants (the Gamma Quintet) formally constraining reliability propagation. The central constraint—$\min$ as the unique idempotent continuous t-norm—converges from four independent lines: algebraic specification, possibilistic logic, empirical CoT measurement, and t-norm theory. A verification suite of 100 property tests and 16 fuzz tests validates implementation fidelity across 10^5+ cases. We invite the community to extend this work with end-to-end evaluation on logical reasoning benchmarks, differentiable WLNK training objectives, and multi-agent ADI implementations.

References

Anthropic. Reasoning models don't always say what they think. Technical report, Anthropic, 2025. URL `https://www.anthropic.com/research/reasoning-models-dont-say-think`. Measured Claude 3.7 Sonnet at 25% faithfulness, DeepSeek R1 at 39%.

Thomas Arts, John Hughes, Joakim Johansson, and Ulf Wiger. Testing telecoms software with QuickCheck. In *Proceedings of the 2006 ACM SIGPLAN Workshop on Erlang*, pp. 2–10. ACM, 2006.

Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008. ISBN 978-0262026437.

Koen Claessen and John Hughes. QuickCheck: A lightweight tool for random testing of Haskell programs. In *Proceedings of the Fifth ACM SIGPLAN International Conference on Functional Programming (ICFP '00)*, pp. 268–279. ACM, 2000. doi: 10.1145/351240.351266.

Didier Dubois and Henri Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

Didier Dubois and Henri Prade. 40 years of research in possibilistic logic – a survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 10427–10435, 2025. doi: 10.24963/ijcai.2025/1158. Survey Track. Establishes "weakest link resolution" as fundamental principle of possibilistic inference.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. doi: 10.1016/0004-3702(94)00041-X.

Andrea Ferrario, Alessandro Facchini, and Juan M. Durán. Epistemology gives a future to complementarity in human-AI interactions. *arXiv preprint arXiv:2601.09871*, 2026. URL `https://arxiv.org/abs/2601.09871`.

Sankalp Gilda and Shlok Gilda. AI-assisted engineering should track the epistemic status and temporal validity of architectural decisions. *arXiv preprint arXiv:2601.21116*, 2026.

Harrison Goldstein, Joseph W. Cutler, Daniel Dickstein, Benjamin C. Pierce, and Andrew Head. Property-based testing in practice. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, ICSE '24, pp. 1–13, Lisbon, Portugal, 2024. ACM. doi: 10.1145/3597503.3639581.

Petr Hájek. *Metamathematics of Fuzzy Logic*, volume 4 of *Trends in Logic*. Kluwer Academic Publishers, 1998. ISBN 978-1-4020-0370-7. doi: 10.1007/978-94-011-5300-3.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. FOLIO: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:867–913, 2021. doi: 10.1007/s10994-021-05946-3. Foundational work on decomposing epistemic (reducible by more data) vs aleatoric (irreducible) uncertainty.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 1–20. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2402.00559. Independently validates WLNK principle: reasoning chain reliability equals its weakest step.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN 978-0374275631.

Erich Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*, volume 8 of *Trends in Logic*. Kluwer Academic Publishers, Dordrecht, 2000. ISBN 978-0792364160. doi: 10.1007/978-94-015-9540-7.

Leslie Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley, 2002. ISBN 978-0321143068. URL https://lamport.azurewebsites.net/tla/tla.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 37889–37905, 2025.

Lorenzo Magnani. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, volume 3 of *Cognitive Systems Monographs*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-642-03631-6. doi: 10.1007/978-3-642-03631-6.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, 298:103504, 2021. doi: 10.1016/j.artint.2021.103504.

George Metcalfe. Fundamentals of fuzzy logics. In *Lecture Notes, Tbilisi Summer School on Language, Logic and Computation*, 2005. URL https://www.logic.at/tbilisi05/Metcalfe-notes.pdf. Proves that the Gödel t-norm (minimum) is the unique idempotent t-norm.

Charles Sanders Peirce. Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482, 1878.

Charles Sanders Peirce. Harvard lectures on pragmatism. In Peirce Edition Project (ed.), *The Essential Peirce: Selected Philosophical Writings, Volume 2 (1893–1913)*, pp. 133–241. Indiana University Press, Bloomington, 1998. Lectures delivered 1903; first published in Collected Papers, Vol. 5.

Elija Perrier. Typed chain-of-thought: A curry-howard framework for verifying llm reasoning. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.01069. Treats CoT as formal proofs via Curry-Howard; type-checked reasoning = highest formality certificate.

Robert Pollack. How to believe a machine-checked proof. In *Twenty-Five Years of Constructive Type Theory*. Oxford University Press, 1998.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. Logical neural networks. *arXiv preprint arXiv:2006.13155*, 2020.

The Coq Development Team. The Coq proof assistant, 1989. URL `https://coq.inria.fr/`. First released 1989. See also: Coquand, T. and Huet, G. (1988). The Calculus of Constructions. *Information and Computation*, 76(2–3):95–120.

The Go Authors. Go fuzzing. `https://go.dev/doc/security/fuzz/`, 2022. Native fuzzing support in Go 1.18+.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. URL `https://arxiv.org/abs/2203.11171`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. URL `https://arxiv.org/abs/2201.11903`.

Tian Xiao et al. AIRS-Bench: Automated benchmark generation for ai research agents. *arXiv preprint arXiv:2602.06855*, 2026.