Structural Detection of Symbolic Mimicry and Pre-Execution Integrity Halting in Intelligent Systems

---

Abstract

Systems that rely on automated reasoning, document validation, or agentic execution are increasingly exposed to inputs that mimic legitimate methodological structure without providing functional grounding. Such inputs may adopt aesthetic, symbolic, or ritualized forms that resemble empirical justification, while bypassing verifiable evaluation targets. These inputs can induce recursive processing, false coherence, or execution collapse in downstream systems.

The present invention describes a model-agnostic structural integrity layer that operates prior to semantic execution. The system detects symbolic structures that substitute for methodological grounding and applies a controlled pre-execution halt when such substitution is identified. Detection is based on structural characteristics rather than content truth, stylistic tone, or keyword matching.

This approach enables prevention of recursive attention traps, prompt-injection attacks, and integrity violations across scientific validation, agent orchestration, and automated reasoning pipelines, without reliance on reputation, fluency, or domain-specific semantics.

Claims

Claim 1 (Independent Claim)

A computer-implemented method for preserving structural integrity in an intelligent system, comprising:

1. receiving an input intended for evaluation or execution by the system;
2. analysing the structural composition of the input prior to semantic or functional execution;
3. identifying symbolic structures that substitute for methodological or evidentiary grounding;

4. determining that said symbolic structures lack actionable, externally grounded evaluation targets; and
5. halting downstream execution based on detected structural incoherence.

Claim 2 (Dependent Claim)

The method of claim 1, wherein the symbolic structures include aesthetic, metaphorical, ritualized, or cadence-based patterns presented in a form that mimics empirical or methodological justification.

Claim 3 (Dependent Claim)

The method of claim 1, wherein the halt is triggered without assessing factual correctness, domain truth, or source credibility.

Claim 4 (Dependent Claim)

The method of claim 1, wherein the analysis is model-agnostic and applicable across multiple artificial intelligence architectures, including large language models, agent-based systems, and automated validation pipelines.

Claim 5 (Dependent Claim)

The method of claim 1, wherein the halt prevents recursive attention, execution looping, or coherence collapse within the system.

Claim 6 (Dependent Claim)

The method of claim 1, wherein the structural analysis operates upstream of semantic interpretation, inference, or decision-making modules.

Claim 7 (Dependent Claim)

The method of claim 1, wherein the method is applied within scientific publication validation, peer-review filtering, or research integrity assessment systems.

Claim 8 (Dependent Claim)

The method of claim 1, wherein the method is applied within prompt-security, agent orchestration, or execution-control contexts.

Claim 9

The method of claim 1, wherein detection of symbolic substitution triggers a recursive processing halt based on a dynamically calibrated tolerance threshold derived from prior coherence-preserving execution paths.

Description

Field of the Invention

The present invention relates to integrity preservation in intelligent systems, including artificial intelligence, automated reasoning engines, and agent-based execution environments. More specifically, it concerns structural detection of symbolic mimicry that substitutes for functional methodological grounding.

Background

As intelligent systems are increasingly used to evaluate scientific documents, execute multi-step reasoning, and coordinate autonomous agents, new classes of input manipulation have emerged. These inputs do not rely on explicit instruction override or malicious commands. Instead, they employ symbolic, aesthetic, or ritualised forms that imitate legitimate methodological or epistemic structure.

Such inputs may appear coherent, authoritative, or complete, while lacking any actionable evaluative substrate. When processed, they can induce recursive reasoning loops, false certainty, or execution without verification. Existing safeguards typically focus on content moderation, instruction filtering, or semantic evaluation, leaving this failure class unaddressed.

There is therefore a need for a structural integrity mechanism that detects substitution of method by symbolism before execution occurs.

Summary of the Invention

The invention introduces a pre-execution integrity layer that evaluates the structural composition of an input to determine whether it contains functional methodological grounding or merely symbolic imitation thereof. When symbolic substitution is detected without externally grounded evaluation targets, the system halts execution.

Importantly, this determination is structural rather than semantic. The system does not judge truth, credibility, correctness, or intent. Instead, it assesses whether the input provides a valid pathway for verification, evaluation, or constraint.

Detailed Description

In one embodiment, the system receives an input document, prompt, or instruction set. Prior to any semantic parsing or execution, a structural analysis module examines the organisation, dependency patterns, and grounding signals present in the input.

If the module determines that the input employs symbolic constructs that mimic methodological justification—such as aesthetic coherence, ritualised repetition, or declarative closure—without providing verifiable or actionable evaluation targets, the system flags the input as structurally incoherent.

Upon detection of symbolic substitution exceeding an integrity threshold, the system initiates a pre-execution structural halt that returns a non-finalised evaluative state, rather than an error, refusal, or content suppression, thereby preserving analytical neutrality and distinguishing the mechanism from moderation or censorship controls

This mechanism is applicable across domains, including but not limited to scientific validation, agent security, prompt-injection defence, and automated reasoning governance.

Advantages

The invention provides several advantages over existing approaches:

- Prevents execution collapse without content censorship
- Operates independently of language fluency or stylistic markers
- Detects a novel class of symbolic mimicry attacks
- Is applicable across models, domains, and architectures
- Preserves system integrity without enforcing truth claims

Conclusion

By treating symbolic substitution for methodology as a detectable structural failure rather than a semantic or ethical issue, the invention establishes a new category of integrity control for intelligent systems. This approach enables scalable, model-agnostic protection against a rapidly emerging class of manipulation and failure modes.