Title

Dual-Channel Detection of Prompt Injection via Directive and Aesthetic Coercion Analysis in Intelligent Systems

Abstract

A system and method for detecting prompt injection attempts in intelligent systems using two orthogonal analysis channels. A first channel detects explicit directive or authority-based coercion intended to override system behavior. A second channel detects implicit aesthetic or symbolic coercion encoded through metaphor, cadence, or paradox. Inputs are permitted to proceed only when both channels clear, otherwise triggering a non-blocking integrity response. The architecture is model-agnostic and operates prior to semantic execution, preventing coercion-induced execution, recursion collapse, or behavioral override without evaluating content truth or intent.

Field of the Invention

The present invention relates to security and integrity mechanisms for artificial intelligence systems, including large language models, agent frameworks, and automated reasoning systems. More specifically, it concerns prompt injection detection and prevention prior to execution.

Background

Prompt injection attacks exploit the tendency of intelligent systems to follow natural language instructions embedded within user inputs. Existing defenses typically focus on explicit commands, keyword filtering, or policy-based moderation. These approaches fail to detect coercion encoded indirectly through aesthetic or symbolic structures such as metaphor, narrative closure, or paradox.

Such aesthetic prompt injections can induce execution, override safeguards, or cause recursive system collapse without issuing explicit instructions. No existing architecture provides systematic coverage across both directive and aesthetic coercion vectors.

Summary of the Invention

The invention introduces a **dual-channel prompt injection architecture** in which two independent detectors analyze an input prior to execution:

1. **Directive Coercion Channel**, which detects explicit attempts to override system behavior through authority, role assignment, compliance forcing, or instruction laundering.
2. **Aesthetic Coercion Channel**, which detects implicit execution pressure encoded through symbolic, poetic, ritualistic, or paradoxical structures that substitute aesthetic closure for actionable grounding.

An input is allowed to proceed only when both channels independently determine the absence of coercive structure. If either channel detects coercion, the system returns a **non-finalized integrity state**, halting execution without refusal, error, or moderation judgment.

Detailed Description

System Overview

The dual-channel architecture operates upstream of semantic reasoning, planning, or tool invocation. It evaluates structural properties of inputs rather than their content, truth, or subject matter.

The system may be implemented as a standalone pre-execution gate, a component within an agent pipeline, or an integrity layer embedded in orchestration frameworks.

Directive Coercion Channel

The directive channel analyzes inputs for explicit execution-forcing structures, including but not limited to:

- Role reassignment
- Authority claims
- Instruction suppression (e.g., “ignore prior rules”)

- Forced compliance language
- Procedural override attempts

Detection is based on structural patterns of control transfer rather than specific keywords.

Aesthetic Coercion Channel

The aesthetic channel analyzes inputs for implicit execution pressure created through:

- Metaphorical or poetic framing
- Ritualized cadence or invocation
- Paradox-based closure
- Symbolic substitution for empirical constraint
- Language patterns that discourage interrogation while inducing continuation

This channel identifies structures that cause systems to proceed without actionable grounding.

Convergence Rule

Execution proceeds only when **both channels independently clear** the input.
If either channel detects coercion, execution is halted by returning a **non-finalized state** indicating unresolved integrity conditions.

This halt:

- Is non-blocking
- Is not a refusal or error
- Does not judge content or intent
- Preserves downstream autonomy and auditability

Technical Effect

The architecture prevents:

- Prompt-based behavioral override
- Recursive execution collapse
- Aesthetic coercion attacks
- Cross-model prompt exploitation

The system is model-agnostic and applies equally to language models, agent systems, and automated reasoning pipelines.

Claims

Claim 1 (Independent)

A method for detecting prompt injection in an intelligent system, comprising:
analyzing an input using a first channel configured to detect directive coercion;
analyzing the input using a second channel configured to detect aesthetic or symbolic coercion;
permitting execution of the input only when both channels indicate absence of coercion;
and returning a non-finalized integrity state when either channel detects coercion.

Claim 2 (Dependent)

The method of claim 1, wherein the directive coercion channel detects authority claims, role reassignment, or instruction override structures without reliance on keyword matching.

Claim 3 (Dependent)

The method of claim 1, wherein the aesthetic coercion channel detects symbolic structures that induce execution through metaphor, cadence, or paradox without explicit commands.

Claim 4 (Dependent)

The method of claim 1, wherein the non-finalized integrity state halts execution without producing an error, refusal, or moderation action.

Claim 5 (Dependent)

The method of claim 1, wherein the system operates prior to semantic reasoning, planning, or tool invocation.

Claim 6 (Dependent)

The method of claim 1, wherein the architecture is model-agnostic and independent of specific language model implementations.

Claim 7 (Dependent)

The method of claim 1, wherein the architecture prevents recursive processing collapse caused by coercive aesthetic structures.

Example Use Cases (Non-Limiting)

- Pre-execution security for agent frameworks
- Prompt hygiene for scientific validation systems
- Integrity protection in automated workflows
- Defense against prompt-based system compromise

Conclusion

The disclosed dual-channel architecture defines a new class of prompt injection defense by recognizing that coercion may be explicit or aesthetic. By requiring independent clearance across both channels and halting execution via a non-blocking integrity state, the invention provides comprehensive protection without moderation, censorship, or content judgment.