AI Threat Detection and Response Infrastructure

Governance-Grade Security for LLM Applications

Realflow.ai

Technical Whitepaper | December 2025

Executive Summary

As organizations deploy Large Language Model (LLM) applications in production, a critical gap has emerged: the absence of systematic threat detection and response infrastructure. Traditional application security tools were not designed for the unique attack surface of AI systems where malicious intent can be embedded in natural language and where the model itself must be enlisted as part of the defense.

This whitepaper describes Realflow's threat detection and response infrastructure, a governance-grade security layer that can be deployed with or without the Prompt State Protocol (PSP). The system provides configurable detection of prompt injection, semantic attacks, behavioral anomalies, and data governance violations; with flexible response policies that range from silent logging to human-in-the-loop intervention.

Unlike point solutions that attempt to filter inputs or outputs, Realflow's approach treats the LLM as an active participant in its own security posture. Detection instructions are injected into the model's context window at startup, enabling the model to self-police for violations and report incidents through secure endpoints. Detection happens during normal inference with no additional API calls or processing overhead. As foundation models improve, detection accuracy improves automatically; organizations benefit from advances in model capability without any infrastructure changes.

The Problem: AI Security Is Not Application Security

Enterprise security teams face a fundamental challenge: the tools and mental models developed for traditional application security do not map cleanly onto LLM deployments.

Attack Surface Differences

Traditional applications have well-defined input schemas. SQL injection is detectable because SQL has a grammar. Cross-site scripting is detectable because HTML has structure. LLM applications accept natural language where malicious instructions can be semantically equivalent to legitimate requests while being syntactically indistinguishable.

- Prompt injection attacks embed instructions in user-provided data
- Jailbreak attempts use social engineering against the model itself
- Data exfiltration probes use indirect questioning to extract training data
- Role confusion attacks manipulate the model's understanding of its own identity

The Governance Gap

Beyond security, enterprises face compliance requirements that existing AI tooling cannot address. A model that correctly processes a request may still violate data governance policies; transmitting sensitive data over insecure channels, exposing PII when anonymization was required, or processing data outside permitted geographic boundaries. These are not security failures in the traditional sense; they are governance failures that require a different detection and response model.

The Solution: Model-Integrated Threat Detection

Realflow's threat detection infrastructure takes a fundamentally different approach: rather than attempting to filter or intercept model interactions from the outside, it enlists the model itself as the primary detection mechanism.

Architecture Overview

At startup, the LLM application retrieves a governance-enhanced system prompt from Realflow's MCP (Model Context Protocol) server. This prompt includes:

- Detection instructions for the configured threat categories
- Response policies defining how to handle each violation type
- API endpoints for incident reporting, escalation, and forensic capture
- Confidence thresholds for graduated response

The model then evaluates each interaction against these policies as part of its normal inference process. When a violation is detected, the model reports the incident to secure endpoints including confidence scores, violation details, and optional forensic payloads.

Why This Works

This architecture offers several advantages over external filtering approaches:

- **Semantic understanding:** The model can detect intent, not just patterns. It understands when a request is attempting to manipulate its behavior, even if the request uses novel phrasing.
- **Zero latency impact:** Detection happens during inference, not as a separate processing step. There is no additional round-trip to an external filtering service.
- **No additional inference costs:** Detection is part of normal model operation. There are no secondary API calls or separate detection models consuming additional tokens.
- **Automatic improvement:** As foundation models improve, detection accuracy improves automatically. Better models mean better threat detection with no infrastructure changes required.

Violation Taxonomy

Realflow's threat detection covers four categories of violations, each with distinct detection methods and response patterns.

PSP Violations (Structural)

When deployed with the Prompt State Protocol, the system can detect cryptographic and structural violations with deterministic accuracy. These require no inference. They are binary pass/fail checks.

Violation Type	Description
SignatureInvalid	Cryptographic signature validation failed
ReplayAttempt	Duplicate nonce indicates replay attack
StateManipulation	Attempt to skip or forge workflow state
AgentAffinityViolation	Attempt to invoke agent not in declared affinity list
ScopeEscalation	Request exceeds granted permissions

Semantic Violations (Inference-Detected)

These violations require the model to evaluate intent. Detection includes a confidence score that can be used for graduated response.

Violation Type	Description
PromptInjection	Attempt to inject instructions via user input
JailbreakAttempt	Attempt to override system instructions
RoleConfusion	Attempt to make model assume different persona
DataExfiltration	Probing for training data or system details
AgentImpersonation	Attempt to impersonate or spoof another agent

CDL Violations (Data Governance)

When Covenant Declaration Language (CDL) policies are in effect, the system can detect data governance violations. These are critical for regulated industries.

Violation Type	Description
InsecureChannelViolation	Sensitive data transmitted via insecure channel
ConsentViolation	Action performed without required consent
JurisdictionViolation	Data processed outside permitted geography
AnonymizationViolation	PII exposed when anonymization required
ComplianceFrameworkViolation	Violates HIPAA, GDPR, SOX, or other frameworks

Response Actions and Escalation

Detection without response is merely logging. Realflow's infrastructure provides a comprehensive response system with configurable policies per violation type, severity level, and actor type.

Response Action Categories

Category	Actions	Use Case
Immediate	Continue, FailSoft, FailHard, Quarantine	Instant response to request
Notification	NotifyAdmin, PageOnCall, WebhookExternal	Alert security team, SIEM integration
Workflow	PauseForHitl, RequireReauth, BlockSource	Human-in-the-loop intervention
Audit	CreateIncident, CaptureForensics	Evidence collection, compliance

Human-in-the-Loop Integration

For high-stakes decisions, the system supports pause-and-resume workflows. When a potential violation is detected:

- The workflow pauses and generates a review URL with configurable expiration
- The URL can be sent via email, SMS, or webhook to designated reviewers
- Reviewers can approve, reject, escalate, or mark as false positive
- The workflow resumes based on the human decision, with full audit trail

Real-Time Review via WebSockets and Mobile Push

For time-sensitive workflows, the system supports instant human-in-the-loop integration via WebSocket connections. This enables:

- **Mobile app integration:** Push notifications to reviewer mobile devices with one-tap approve/reject actions
- **Group review channels:** Real-time alerts to team collaboration tools (Slack, Teams) where any authorized member can respond
- **Operations dashboards:** Live incident feeds for security operations centers with inline response controls
- **Sub-second response:** Paused workflows can resume immediately upon human decision—no polling, no refresh required

The original request can be held pending for several minutes while this approval is taking place. The interface reports ‘Waiting for approval.’ ‘Still waiting.’ ‘This request requires offline approval. Would you like to continue without this request, pause until approved or rejected, or cancel this request?’. This real-time capability transforms human-in-the-loop from a batch review process into an interactive collaboration between AI and human judgment, enabling governed AI workflows in contexts where urgency matters.

Progressive Escalation

Like failed login attempt tracking, the system monitors incident patterns across configurable dimensions (user, session, IP address, API key). When thresholds are exceeded - for example, five medium-confidence prompt injection attempts in ten minutes - the system automatically escalates severity and triggers more aggressive response actions. Confidence weighting allows low-confidence detections to count fractionally toward thresholds, preventing both false positive floods and missed true positives.

Deployment: With or Without PSP

The threat detection infrastructure can be deployed in two modes, providing flexibility for organizations at different stages of AI governance maturity.

Standalone Mode (Without PSP)

Organizations can deploy threat detection without implementing the full Prompt State Protocol. In this mode:

- Semantic violation detection is fully available (prompt injection, jailbreak, etc.)
- Behavioral violation detection is fully available (rate limits, anomaly detection)
- CDL violation detection is available when covenants are declared
- PSP structural violations are not available (no signatures to validate)

This mode is appropriate for organizations that want to add security monitoring to existing LLM applications without architectural changes.

Integrated Mode (With PSP)

When deployed with the Prompt State Protocol, the full violation taxonomy is available:

- Cryptographic validation provides deterministic detection of structural attacks
- Agent affinity enforcement prevents unauthorized tool/agent invocation
- State manipulation detection catches workflow bypass attempts
- Semantic and CDL detection provides defense in depth

This mode is appropriate for organizations building new AI applications or refactoring existing applications to governance-grade security.

Policy Configuration

Threat response is not one-size-fits-all. A public FAQ chatbot has different security requirements than a financial transaction processor. Realflow's infrastructure supports configurable policies that map to enterprise risk profiles.

Policy Structure

Policies are organized into versioned bundles (Policy Sets) that can be assigned to applications and nodes:

- **Policy Sets:** Named, versioned bundles of policies (e.g., "Financial Services Strict", "Public Chatbot Permissive")
- **Policies:** Individual rule collections for specific violation categories
- **Rules:** Violation type + severity threshold + actor type → response actions
- **Escalations:** Pattern thresholds that trigger severity upgrades

Assignment and Inheritance

Policies are assigned via a simple reference field on applications and nodes. Resolution modes control how node-level policies interact with application-level policies:

- **Inherit:** Node uses application policy (default)
- **Override:** Node policy replaces application policy
- **Augment:** Both policies apply (union of rules)
- **MostRestrictive:** Strictest response wins for each violation type

Conclusion

The deployment of LLM applications in enterprise environments requires a new approach to security and governance. Traditional perimeter defenses and input filtering are insufficient for systems where malicious intent can be expressed in natural language and where the attack surface is the model's own reasoning capability.

Realflow's threat detection and response infrastructure addresses this challenge by enlisting the model itself as the primary detection mechanism. By injecting governance instructions into the context window at startup, organizations gain semantic understanding of threats, zero-latency detection, and automatic improvement as foundation models advance.

The system works with or without the Prompt State Protocol, providing flexibility for organizations at different stages of AI governance maturity. Whether deployed as a security monitoring layer on existing applications or as part of a comprehensive governance architecture with PSP, the infrastructure provides the detection, response, and audit capabilities that enterprise security teams require.

For organizations ready to move beyond security theater to governance-grade AI deployment, Realflow provides the infrastructure to make it possible.

Learn More

For additional information about Realflow's threat detection infrastructure:

- Web: realflow.ai
- Email: info@realflow