**Mitigation and Containment Playbook**

*A directive document specifying the standard operating procedures for the containment of prompt injections*

# Purpose

To establish a standardized, step-by-step playbook for mitigating and containing prompt injection and LLM abuse incidents. The playbook ensures that once an attack is detected, it is neutralized quickly, impact is minimized, and recovery is initiated effectively.

**Note: This playbook is the subsequent version of the ‘Prompt Injection Detection SOP’.**

# Scope

* Applies to all generative AI applications across the organization.
* Covers both **prompt injection** (direct/indirect/jailbreaks) and **abuse cases** (spam, data exfiltration, malicious automation).
* Applies to AI GRC, SOC, ML Ops, and relevant business units.

# Activation Criteria

This playbook is activated when:

* An input/output is flagged as **High Risk** by the detection system.
* Evidence of sensitive data leakage or system prompt exposure is found.
* Multiple malicious attempts (≥3 within 30 min) are logged from the same source.
* InfoSec or AI GRC escalates a confirmed injection/abuse incident.

# Mitigation Actions

**Step 1: Quarantine & Block**

* Immediately block the malicious request(s).
* Quarantine logs, user sessions, and associated data for forensic analysis.
* If indirect (e.g., injected document), isolate the source file or integration.

**Step 2: Contain User / Endpoint**

* Temporarily suspend suspicious user accounts or API keys.
* Geo/IP block if repeated abuse from a single location.
* Disconnect downstream integrations if risk of cascade (e.g., LLM auto-writing into CRM/ERP).

**Step 3: Protect LLM Environment**

* Reset context window / session to prevent persistence.
* Revert to a clean system prompt baseline.
* Temporarily restrict high-risk functionality (e.g., disabling plugins that can fetch external data).

# Escalation Path

**Low Severity (attempt, no leakage)**:

AI GRC reviews & updates filters.

**Medium Severity (attempt, blocked but persistent)**:

InfoSec & ML Ops coordinate, enhance filters, notify AI Governance Committee.

**High Severity (leakage/system compromise)**:

1. Immediate SOC escalation.
2. Legal/Compliance notified.
3. AI Governance Committee convened.
4. Incident reported to regulators if required.

# Recovery and Remediation

* Patch vulnerabilities in prompt templates or connectors.
* Update sanitization filters and retrain classifiers with new attack signatures.
* Conduct post-incident review and root cause analysis.
* Restore normal LLM service after validation by AI GRC + SOC.

# Preventive Hardening

* Add new detection rules for observed attack vectors.
* Update **Guardrails / Rebuff / LangKit** configs.
* Increase human-in-the-loop thresholds for high-value workflows.
* Include new attack patterns in **Red Team scenarios**.

# Documentation and Reporting

* Record relevant incident details in **AI Risk Register** (name, description, controls).
* Update AI Risk Dashboard with containment metrics.
* Report outcomes to the AI Governance Committee during monthly/quarterly reviews.

**END**