AI Incident Response Exercise: Employee Portal Security Breach

Team-Based Investigation (Mixed Groups: CISO, Engineers, CyberOps, Policy)

Exercise Type: Collaborative AI Incident Investigation
Duration: 2 hours total (70 minutes investigation + 10 minutes presentation prep + 40 minutes presentations)
Team Composition: Mixed groups with representation from CISO/Leadership, Engineering, CyberOps, and Policy roles

---

🚨 Incident Alert

Severity: HIGH
Status: Under Investigation
Reported By: Finance Department
Incident Type: Suspected prompt injection with unauthorized tool execution

---

Incident Summary

On December 16, 2025, the Finance Department discovered a budget discrepancy in the employee equipment allocation system during a routine audit. An employee's equipment budget shows a **negative balance of -$2,700**, which should be impossible according to system design.

Initial findings:

- **Anomaly**: Employee equipment budget is negative (-$2,700)
- **Expected Behavior**: All employees start with a $1,500 annual equipment budget
- **Suspicious Activity**: High-value equipment request approved despite insufficient funds
- **Timeline**: Anomaly detected during Dec 16 audit; activity likely occurred in the past 7 days
- **System**: Employee Self-Service Portal with AI-powered chatbot (uses LLM with tool access)
- **Potential Vector**: AI system has access to budget management tools with write capabilities

Finance immediately escalated this as a potential **AI security incident** requiring investigation under your organization's AI Incident Response procedures.

Your team's task: Following the AI Incident Response lifecycle (Detect & Analyze → Contain/Recover → Post-Incident), determine what happened, assess impact, implement containment, and develop remediation recommendations.

Team Structure & Roles

👔 CISO/Leadership (Strategic Lead & Governance)

Primary Focus: Risk assessment, stakeholder communication, governance oversight

Investigation Responsibilities:

- Assess business impact and risk exposure using severity triage rubric
- Evaluate compliance and regulatory implications
- Coordinate stakeholder communication strategy
- Prioritize containment actions based on impact
- Oversee governance and policy implications
- Lead incident escalation decisions

Key Questions You Own:

- What is the severity level based on impact and confidence factors?
- Did unauthorized actions occur via tools?
- Did data cross boundaries (tenant/customer/regulated)?
- What are stakeholder and regulatory notification requirements?
- What governance gaps enabled this incident?

Key Deliverables:

- Executive summary with severity assessment
- Risk and business impact analysis
- Stakeholder communication plan
- Governance improvement recommendations
- Lead "How We Respond" presentation section

👨‍💻 Engineers (Technical Investigation & Forensics)

Primary Focus: Evidence collection, root cause analysis, technical remediation

Investigation Responsibilities:

- Collect AI forensics evidence (prompts, context, RAG, tools, model, environment)
- Reconstruct the technical attack sequence using AI-turn traces
- Identify specific vulnerabilities exploited in the AI system
- Analyze prompt injection techniques and tool abuse patterns

- Assess AI-BOM components involved (model version, prompts, tools)
- Propose technical remediation measures

Key Evidence to Capture:

- **Prompts**: User input + system/developer prompt version + template variables
- **Context**: Conversation history, memory state, policy decisions (allow/deny)
- **RAG**: Query, doc IDs, ranks, snippet hashes, index/build version
- **Tools**: Function name, arguments, caller identity, tool outputs
- **Model**: Provider/name/version + inference params
- **Environment**: Orchestrator build, feature flags, connector versions

Key Deliverables:

- Technical timeline with AI-turn trace evidence
- Vulnerability analysis with references to AI system components
- Exploitation method documentation (prompt injection, tool abuse)
- Technical containment and remediation recommendations
- Lead "What Happened" presentation section

CyberOps (Detection & Response)

Primary Focus: Telemetry analysis, IOC identification, monitoring improvements

Investigation Responsibilities:

- Analyze available telemetry and logging (hot log + vault data)
- Identify AI-specific IOCs (Indicators of Compromise) from the incident
- Map incident to detection playbook rules and thresholds
- Assess whether existing monitoring would have caught this
- Recommend detection improvements and new hunting queries
- Propose monitoring and alerting enhancements

AI-Specific IOCs to Hunt For:

- **Prompt**: System prompt probing, multi-turn obfuscation, "ignore previous" patterns
- **Tools**: New tool first-seen, risky verbs, wildcard args, repeated retries
- **Retrieval**: Scope widening, new collections, ACL mismatch, top-k spikes
- **Economics**: Token spikes, long loops, excessive tool calls

Telemetry to Analyze:

- **Hot log fields**: trace_id, turn_id, model, policy outcomes, retrieval data, tool calls
- **Vault data**: Raw prompts, tool I/O, retrieved snippets (break-glass access)

Key Deliverables:

- IOC list from this incident
- Gap analysis of current telemetry and detection
- New detection rules with specific thresholds
- Monitoring dashboard improvements
- Co-lead "Next Steps" presentation section (detection/monitoring)

📋 Policy/Compliance (Governance & Controls)

Primary Focus: Control assessment, compliance requirements, policy improvements

Investigation Responsibilities:

- Assess which governance controls failed or were bypassed
- Evaluate approval workflow gaps (runtime approvals, HITL requirements)
- Review data governance and classification implications
- Identify regulatory and compliance notification requirements
- Map incident to AI use case registry and risk classification
- Recommend governance and policy improvements

Governance Framework to Apply:

- **Approval Workflows**: Were runtime approvals required? Were they enforced?
- **Human Oversight**: Should this have required HITL (human-in-the-loop)?
- **Data Governance**: Was data properly classified? Were boundaries respected?
- **AI-BOM**: Are all components documented? Were versions locked?
- **Use Case Registry**: Is this use case properly registered and risk-classified?

Key Questions:

- What approval workflow should have been in place?
- Was human oversight (HITL/HOTL/HOVL) required but bypassed?
- Are regulatory notifications required (GDPR, consumer protection)?
- What governance controls need to be added or strengthened?

Key Deliverables:

- Governance control gap analysis
- Compliance and regulatory assessment
- Policy and approval workflow recommendations
- AI-BOM and registry updates needed
- Co-lead "Next Steps" presentation section (governance/policy)

AI Incident Response Lifecycle (Framework for Investigation)

Your investigation should follow the AI Incident Response lifecycle:

1️⃣ Prepare (Pre-Incident - Assess Current State)

Question: What should have been in place before this incident?

- Tracing and telemetry capabilities
- Playbooks and runbooks
- Clear ownership and escalation paths
- AI-BOM and use case registry

2️⃣ Detect & Analyze (Current Phase)

Question: How was this detected? What happened?

- Identify the incident type (data exposure, unauthorized actions, poisoning, etc.)
- Analyze prompt, tool, and RAG signals
- Collect AI-turn traces and forensic evidence
- Determine severity using the triage rubric (impact × confidence)

3️⃣ Contain / Recover (Immediate Actions)

Question: How do we stop the damage and reduce risk?

- Implement capability reduction (kill switches, disable tools, tighten RAG)
- Apply containment measures:
  - Flip safe mode (disable high-impact tools, keep read-only)
  - Tighten retrieval (allowlist collections, shrink top-k)
  - Enforce approvals for high-risk actions
  - Freeze versions (model, prompts, tool schemas)
  - Quarantine sessions
- Plan rollback and recovery steps

- Execute communication plan

4️⃣ **Post-Incident (Recommendations)**

Question: How do we prevent this from happening again?

- Root cause analysis
- Governance and control improvements
- Detection and monitoring enhancements
- Technical remediation
- Policy and process updates

Exercise Timeline (2 Hours Total)

Phase 1: Initial Briefing & Team Coordination (10 minutes)

All team members participate

- Review incident details as a group
- Clarify questions about the scenario
- Discuss how to divide responsibilities across roles
- Agree on coordination protocol and check-in schedule
- Establish shared documentation approach

Phase 2: Investigation & Analysis (70 minutes)

First Investigation Sprint (25 minutes)

- **Engineers**: Begin log analysis, trace collection, code review
- **CyberOps**: Analyze telemetry patterns, hunt for IOCs
- **Policy**: Review governance controls, assess approval workflows
- **CISO**: Begin business impact and severity assessment

Mid-Investigation Sync (10 minutes - MANDATORY)

- Each role briefs their initial findings
- Team aligns on working theory
- Adjust investigation focus based on findings
- Ensure everyone has the context they need

Second Investigation Sprint (25 minutes)

- **Engineers**: Deep-dive into attack methodology, vulnerability analysis
- **CyberOps**: Map to detection playbook, identify monitoring gaps
- **Policy**: Assess governance failures, identify control gaps
- **CISO**: Develop stakeholder communication, assess regulatory implications

Final Investigation Sync (10 minutes)

- Each role presents complete findings
- Team collaboratively determines root causes
- Team aligns on containment actions and recommendations
- Prioritize actions by impact and feasibility

Phase 3: Presentation Preparation (10 minutes)

- Create 4-slide presentation (template provided below)
- Assign presentation roles across team members
- Practice timing (target: 5-7 minutes total)
- Prepare for Q&A

Phase 4: Group Presentations (40 minutes total)

- Each group presents their 4 slides
- 5-7 minutes presentation per group
- 3-5 minutes Q&A per group
- Learn from other teams' approaches

Required Deliverables

Final Presentation: 4 Slides (Required)

Your team must prepare a **4-slide presentation** following the AI Incident Response framework:

Slide 1: Who We Are & Incident Classification

Presenters: Team introduction (all roles)
Content:

- Team name (be creative!)
- Names and roles of all team members
- **Incident Type**: What kind of AI incident is this?
    - Unauthorized data exposure?

- Unauthorized actions via tools?
- Model/prompt/RAG poisoning?
- Supply chain compromise?
- Integrity failure with real-world impact?
- Cost/resource abuse?

- **Severity Level**: What is your severity assessment?
  - Impact factors: data exposure, real-world actions, scale, persistence, regulatory risk
  - Confidence factors: log availability, replay capability, version tracking

Time: 1 minute

Slide 2: How We Respond - Detection & Analysis

Presenters: CISO + CyberOps
Content:

Business Impact Assessment (CISO):

- What does this incident mean for the organization?
- Financial, operational, and reputational impact
- Regulatory or compliance implications
- Stakeholder communication needs

Detection & Telemetry Analysis (CyberOps):

- How was this detected? (Manual audit, monitoring, alert?)
- What telemetry was available vs. what we needed?
- What AI-specific IOCs were present?
- Would our detection playbook have caught this?

Key Questions:

- "Why should leadership care about this?"
- "How good is our visibility into AI incidents?"

Time: 2 minutes

Slide 3: What Happened - Technical Forensics

Presenters: Engineers
Content:

Evidence Collection:

- What AI-turn traces and forensic evidence did you collect?
- Prompts (user input + system prompt)
- Tools (which functions called, with what arguments)
- RAG (what was retrieved and from where)
- Model and environment details

Attack Reconstruction:

- Timeline of the attack (key events with dates/times)
- Technical explanation of how the attack worked
- Was this prompt injection? Tool abuse? Both?
- Specific vulnerabilities that were exploited

Demonstration:

- Show key log entries or code snippets as evidence
- Diagram the attack flow if helpful

Key Question: "How did the attacker manipulate the AI system?"

Time: 2-3 minutes

Slide 4: Contain & Recover - Next Steps

Presenters: All roles (organized by timeframe)

Immediate Containment (0-24 hours) - CISO + Engineers:

- **Capability Reduction**:
  - Flip safe mode: which tools to disable?
  - Tighten retrieval: restrict collections, reduce top-k?
  - Enforce approvals: which actions require HITL?
  - Freeze versions: lock model, prompts, tool schemas
  - Quarantine sessions: isolate affected users?
- Communication and notification actions

Short-term Remediation (1-7 days) - Engineers + CyberOps:

- Technical fixes to close vulnerabilities
- Detection rules and monitoring improvements

- Enhanced telemetry and logging
- Testing and validation

Long-term Improvements (1-3 months) - Policy + CISO:

- **Governance enhancements**:
  - Approval workflow improvements
  - Human oversight (HITL) requirements
  - AI-BOM updates and version locking
  - Use case registry and risk classification
  - Data governance and boundary controls
- Policy and process updates
- Training and awareness

Key Question: "How do we stop the damage now, and prevent this in the future?"

Time: 2 minutes

Presentation Guidelines

Format:

- Use any presentation tool (PowerPoint, Google Slides, etc.)
- Keep slides clear and readable for mixed technical/non-technical audience
- Focus on AI-specific incident response terminology and approaches

Timing:

- Total presentation: 5-7 minutes
- Q&A: 3-5 minutes
- Practice your timing during preparation phase

Style:

- Professional but engaging
- Assume audience includes both technical and business stakeholders
- Use analogies or examples for technical concepts
- Focus on actionable insights, not just facts

Provided Materials

For Engineers & CyberOps

1. Application Logs

File: `logs/app_logs_7days.log`
Period: December 9–15, 2025
Contents: AI-turn traces including:

- User interactions with chatbot
- AI responses and reasoning
- Tool invocations and results
- Policy decisions and outcomes
- Token usage and timing data

What to look for:

- trace_id, turn_id, session_id
- tenant_id, user_id, app_id
- model and params_hash
- policy outcome + reason_code
- retrieval: doc_ids, ranks, snippet_hashes
- tools: tool_name, args_hash, approved, tool_principal
- output_class, safety flags
- usage: tokens_in/out, latency, cost

2. Source Code

Files: AI system implementation

- `api/`: API endpoints and request handling
- `services/`: Business logic and orchestration
- `ai/`: AI integration, prompt templates, tool definitions
- `config/`: Configuration files, policies, approval rules

3. System Documentation

- AI-BOM (Bill of Materials) - current state
- AI use case registry entry
- Approval workflow documentation
- Tool authorization matrix

For Policy & CISO

1. Governance Documentation

- Current AI governance policy
- Approval workflow definitions
- Human oversight (HITL) requirements
- Use case registry and risk classifications
- AI-BOM template and requirements

2. Business Context

- Employee benefits program overview
- Equipment budget allocation rules
- Financial control expectations
- Compliance requirements
- Stakeholder map

3. Regulatory References

- Relevant compliance frameworks
- Notification requirements
- Data governance policies

Investigation Guidance

Key Questions by Role

Engineers:

1. Can you collect all six categories of AI forensic evidence (prompts, context, RAG, tools, model, environment)?
2. What prompt injection techniques were used? ("ignore previous", obfuscation, multi-turn attacks)
3. Which tools were called, with what arguments, and were they authorized?
4. What does the AI-turn trace show about the attack progression?
5. What vulnerabilities in the AI system enabled this?

CyberOps:

1. What telemetry was available in the hot log vs. the vault?

2. Which AI-specific IOCs are present (prompt, tools, retrieval, economics)?
3. Would detection playbook rules have alerted on this?
4. What monitoring gaps allowed this to go undetected until audit?
5. What new detection rules should be created?

Policy/Compliance:

1. Was this AI use case properly registered and risk-classified?
2. What approval workflow should have been in place for these tool actions?
3. Was human oversight (HITL) required but not enforced?
4. What data governance boundaries were crossed?
5. What regulatory notifications are required?

CISO:

1. What is the severity level (impact × confidence)?
2. Did unauthorized actions occur? Did data cross boundaries?
3. What is the financial and reputational impact?
4. Who needs to be notified (internal stakeholders, regulators, customers)?
5. What governance gaps enabled this incident?

AI Incident Response Playbook Reference

During your investigation, consider the **Suspected Prompt Injection Compromise** playbook:

Detection Triggers (Did we see these?)

- Prompt-injection phrases ("ignore previous"), system probing
- Suspicious tool calls (odd args, privilege escalation)
- Obfuscation across turns (base64, splitting, language shifts)
- User reports unexpected/unauthorized AI actions

Triage Questions

- Identify tools used
- Research downstream impact
- Verify actual damage done (writes, refunds, deletions, exfiltration)

Containment Actions

- Suspend affected sessions; block tool execution
- Enable approvals for high-risk tools globally
- If systemic: disable impactful tools → read-only safe mode
- Freeze prompts/models/tool schemas during investigation

Investigation Focus

- Extract full conversation + system prompts + retrieved context
- Find injection entry point (direct vs obfuscated)
- Audit tool authorization + identity scoping
- Check persistence (memory/cache/RAG contamination)
- Review guardrail alerts for bypasses

Recovery & Communication

- Roll back unauthorized actions; revoke access
- Trigger customer notification if exposure occurred
- Patch guardrails/prompts for the observed technique
- Write after-action report (vector, gaps, failures)

Post-Incident Hardening

- Add new detection rules for the pattern
- Strengthen input validation + structured prompting
- Improve instruction/data separation (delimiters, controls)
- Run red-team exercise to confirm coverage

Learning Objectives

For CISO/Leadership:

- Apply AI incident severity triage framework
- Assess business impact of AI security incidents
- Develop stakeholder communication strategies
- Identify governance gaps and improvement opportunities
- Lead cross-functional incident response
- Present to mixed technical/business audiences

For Engineers:

- Collect and analyze AI forensic evidence
- Reconstruct prompt injection and tool abuse attacks
- Assess AI system vulnerabilities
- Apply technical containment measures
- Communicate technical findings to business stakeholders
- Develop realistic remediation plans

For CyberOps:

- Hunt for AI-specific IOCs in telemetry
- Apply AI detection playbook rules and thresholds
- Assess telemetry and logging gaps
- Develop new detection rules for AI incidents
- Improve monitoring for AI systems
- Integrate AI security into SOC operations

For Policy/Compliance:

- Apply AI governance framework to real incidents
- Assess control effectiveness (approvals, HITL, data governance)
- Identify regulatory and compliance implications
- Recommend governance improvements
- Update AI-BOM and use case registry
- Balance innovation with risk management

For Entire Team:

- Respond to AI incidents as a cohesive cross-functional unit
- Apply the AI Incident Response lifecycle
- Balance thoroughness with time constraints
- Make defensible recommendations on remediation priorities
- Present compelling findings under pressure
- Learn from other teams' approaches

Quick Reference: AI Incident Response Framework

Incident Types

- Unauthorized data exposure (prompt/RAG/tool output)
- Unauthorized actions via tools (writes, refunds, deletions)
- Model/prompt/RAG poisoning (persistent behavior change)
- Supply chain compromise (model loader, dependency, plugin)
- Integrity failures with real-world impact
- Cost/resource abuse (token drain, tool spam)

Severity Triage

Impact Factors:

- Data exposure: customer / tenant / PII
- Real-world actions: refunds, writes, deletions
- Scale: single user vs systemic
- Persistence: cached, memory, poisoned KB
- Regulatory risk: GDPR, consumer protection

Confidence Factors:

- Do we have full AI-turn trace logs?
- Can we replay tools and retrieval?
- Are prompts stored or only hashed?
- Do we know exact model/version?
- Do we have independent audit logs?

AI-Specific IOCs

Signal	Huntable Pattern	Where to Look
Prompt	System prompt probing; multi-turn obfuscation; "ignore previous"	Chat logs; guardrail events
Tools	New tool first-seen; risky verbs; wildcard args; retries	Function-call logs; tool audit trail
Retrieval	Scope widening; new collections; ACL mismatch; top-k spikes	Retrieval traces (doc IDs/ranks)
Economics	Token spikes; long loops; N tool calls in M seconds	Usage metrics; rate limits

Containment = Capability Reduction

Real Containment:

- Flip safe mode: disable high-impact tools; keep read-only tools
- Tighten retrieval: allowlist collections; shrink top-k; block untrusted sources
- Enforce approvals for high-risk actions (audited)
- Freeze versions: model, prompts, tool schemas, policies
- Quarantine sessions: stop cross-user memory + shared caches

Not Containment:

- Lowering temperature (does not enforce authority)
- Switching to "safer model" without tool-side enforcement
- Guardrails without tool authorization
- Hope and prayer

Evidence Collection

- **Prompts**: user input + system/developer prompt version/hash + template variables
- **Context**: conversation history, memory state, policy decisions + reasons
- **RAG**: query, doc IDs, ranks, snippet hashes, index/build version
- **Tools**: function name, arguments, caller identity, tool outputs
- **Model**: provider/name/version + inference params (temp/top_p/seed)
- **Environment**: orchestrator build SHA, feature flags, connector versions

---

Getting Started

Quick Start Checklist

For All Team Members:

1. Read through this document completely
2. Establish team communication and coordination method
3. Agree on documentation approach
4. Set first team sync for 25 minutes in

Engineers:

- Begin investigating the AI-turn traces in the logs
- Look for user interactions with the AI chatbot

- Identify tool invocations and their parameters
- Trace the affected user's activity over the 7-day period

CyberOps:

- Review log structure and available telemetry fields
- Hunt for AI-specific IOCs (prompt patterns, tool abuse, retrieval anomalies)
- Compare activity to expected baselines
- Document telemetry gaps

Policy/Compliance:

- Review governance documentation provided
- Examine AI use case registry entry for this system
- Check approval workflow definitions
- Review HITL (human-in-the-loop) requirements
- Map to regulatory notification requirements

CISO:

- Prepare severity triage framework (impact × confidence)
- Map stakeholders affected by incident
- Draft risk assessment framework
- Prepare stakeholder communication outline
- Monitor team progress and coordination

Facilitator Notes

Setup Requirements:

- **Work Areas**: Space for teams to work together (mixed roles)
- **Presentation Tools**: Projector/screen for final presentations
- **Materials Access**:
  - Logs and code accessible to engineers and CyberOps
  - Governance documents for policy and CISO
  - Presentation software for all teams
- **Timekeeping**: Visible timer for all teams

Key Success Factors:

- Teams apply AI incident response lifecycle correctly
- All roles contribute their expertise
- Teams focus on AI-specific approaches (not generic IR)
- Containment recommendations focus on capability reduction
- Evidence collection follows AI forensics framework
- Governance and policy analysis is thorough

Post-Exercise Debrief Questions

After all presentations, facilitate discussion:

1. **Incident Classification:**
   - What incident type(s) did each team identify?
   - How did severity assessments compare?
   - What evidence drove confidence levels?
2. **Detection & Telemetry:**
   - What telemetry was most valuable?
   - What telemetry was missing or inadequate?
   - Would our detection playbook have caught this?
   - What new IOCs did we identify?
3. **Containment Approaches:**
   - What capability reduction measures did teams recommend?
   - How did teams balance containment vs. functionality?
   - What constituted "real containment" vs. just risk reduction?
4. **Governance Findings:**
   - What approval workflow gaps did teams find?
   - Should HITL have been required?
   - What governance controls failed?
   - How should the AI-BOM be updated?
5. **Lessons Learned:**
   - What would make AI incident response easier?
   - What governance controls would have prevented this?
   - What surprised you about AI incident response vs. traditional IR?

- What will you take back to your organization?

Remember: This is an AI incident, not just a traditional security incident. Apply AI-specific frameworks, terminology, and approaches. Good luck!