gruve

SOLUTION BRIEF

AI compromise assessment

Forensic investigation of deployed AI environments for evidence of abuse, manipulation, and supply chain compromise

Partners: OWASP LLM Top 10 | NIST AI RMF | MITRE ATLAS

01

Business problem

Organizations have been deploying generative AI, agentic workflows, and MCP-enabled toolchains at speed. The focus has been on capability — getting copilots connected to internal data, deploying AI agents with tool access, wiring up MCP servers so AI can take action on behalf of users. The security question that has largely not been asked is a different one: not 'can this be abused?' but 'has it been?'

Most security programs have no established process for reviewing AI interaction logs forensically. Prompt logs, tool call records, and MCP manifest configurations exist — but they are not being reviewed with investigative methodology for evidence of abuse, injection attempts, reconnaissance activity, or policy violations. Audit logging for AI systems is frequently incomplete, inconsistently retained, or not evaluated for tamper-evidence. Supply chain trust for the models, plugins, and fine-tuned adapters being deployed has not been assessed.

The result is a growing blind spot: AI infrastructure that has been running in production for months or years with no forensic visibility into what has actually occurred inside it. Traditional endpoint, network, and application security controls have no telemetry into prompt patterns, MCP tool call sequences, or system prompt modification history. SIEMs have no AI-specific IOCs. Incident response playbooks have no AI environment investigation step.

> **The gap is not hypothetical.** MITRE ATLAS catalogs real-world AI attack techniques including prompt injection, model inversion, and supply chain poisoning. OWASP LLM Top 10 2025 identifies prompt injection and sensitive data exfiltration as top realized risks. Palo Alto Unit 42 has documented web-based indirect prompt injection observed in the wild. The techniques exist, they are being used, and most organizations have no way to detect whether they have been applied to their own AI environment.

02

Why now

AI is no longer experimental. Internal copilots are connected to HR records, financial systems, and legal documents. MCP servers are giving AI agents the ability to query databases, send messages, and execute code. Agentic workflows are making decisions and taking actions without human review of every step. The attack surface that existed at the prompt layer has expanded into a full tool-use and supply chain problem.

Simultaneously, the forensic evidence window is closing. AI interaction logs are not retained indefinitely. System prompt change histories are overwritten. MCP manifest versions are not tracked. Every month that passes without a formal investigation is a month of potential evidence that becomes unavailable. Organizations that wait for an incident to force the investigation will find themselves working with a narrower evidentiary record.

Regulatory pressure is also building. EU AI Act requirements, SEC disclosure obligations for material cyber incidents, and emerging insurance requirements for AI-related risk documentation are creating a new class of AI governance obligation. A forensic investigation with documented findings is a fundamentally stronger position than a posture assertion with no underlying evidence.

03

Solution overview

Gruve's AI Compromise Assessment is a forensic investigation of deployed AI environments. It is not a red team exercise, an architecture review, a governance advisory engagement, or an AI safety framework assessment. It is an investigation — applying the same structured evidence-collection and analysis methodology Gruve uses for endpoint and network compromise assessments to the AI-specific evidence layer: prompt interaction logs, MCP server manifests, system configurations, audit trail records, and supply chain provenance.

The engagement begins with scoping: understanding what AI systems are deployed, what logs are available, what MCP servers and plugins are in use, and what the priority investigation questions are. Gruve then conducts structured forensic review across six evidence categories — prompt logs, MCP manifests, audit logs, config files, supply chain, and anomaly analysis — producing risk-ranked findings, AI-specific IOCs, and a remediation roadmap grounded in actual evidence rather than theoretical risk.

AI assistance is used in the engagement only where it is appropriate — for log clustering, evidence organization, and accelerating the review of large prompt corpora. All forensic findings, severity determinations, IOC development, and final conclusions are investigator-validated. The output is defensible as a forensic record, not a generated risk opinion.

04

Gruve's solution items

Investigation component	Description
Prompt log forensics	Systematic review of AI interaction logs for injection attempts, jailbreaks, reconnaissance patterns, data exfiltration indicators, and anomalous usage activity — using structured forensic methodology adapted from log-based incident investigation.
MCP manifest and tool config audit	Review of all deployed MCP server definitions and plugin configurations for unauthorized capabilities, permission overgrant, inconsistencies across versions, and indicators of tampering or supply chain compromise.
AI audit log assessment	Completeness review, retention gap analysis, tamper-evidence evaluation, and access pattern baseline analysis — determining whether the audit record is sufficient to support investigation and whether it shows signs of manipulation.
System prompt and config forensics	Integrity review of deployed system prompts, guardrail configurations, tool permission settings, and change history — identifying unauthorized modifications, governance drift, or configuration states that create exploitable conditions.
Supply chain integrity review	Model provenance verification, plugin and fine-tuned adapter vetting, and third-party integration risk assessment — evaluating what has been sourced, from where, and with what level of trust verification at the time of deployment.
AI-specific IOC development	Derivation of indicators of compromise from engagement findings — prompt-based reconnaissance signatures, tool abuse patterns, config manipulation indicators — formatted for integration into SIEM and detection platforms.

05

Benefits

Benefit	Description
Evidence-backed findings on actual AI exposure	Moves the conversation from theoretical AI risk to specific, forensically-grounded findings tied to what has actually occurred in the deployed environment — not what could hypothetically occur.
Closed visibility gap in the security program	Produces the first formal forensic review of the AI environment — audit log completeness, prompt log history, MCP manifest state — giving security teams a documented baseline for ongoing monitoring and future investigation.
Operationalizable AI-specific IOCs	Delivers indicators of compromise in formats that security operations teams can directly integrate into existing SIEM platforms, extending AI-specific detection coverage without requiring a separate tooling investment.
Supply chain risk documented and addressed	Provides documented provenance review for deployed models, plugins, and adapters — closing the supply chain blind spot that most organizations carry implicitly but have never assessed.
Defensible forensic record for regulatory and legal use	Produces a documented investigation with evidence-backed findings that can satisfy regulatory examination, legal hold requirements, board inquiries, and insurance documentation requirements more credibly than posture assertions alone.

06

Engagement tiers

Tiers	Focus / scope	Core activities	Key deliverables	Target audience
AI Rapid Triage $15,000 – $22,000 *1–2 weeks*	Fast forensic scan of a deployed AI environment — priority log review, MCP manifest audit, and audit log completeness check. Produces an initial findings summary and IOC set.	Prompt log review (90-day window) · MCP manifest audit · Audit log completeness · Quick anomaly scan · IOC summary	Findings summary · MCP audit report · Preliminary IOC set · Remediation priority list	Security teams needing an initial AI environment review or post-incident scoping
Full AI Compromise Assessment $25,000 – $45,000 *3–4 weeks*	Comprehensive forensic investigation across all six evidence categories — prompt logs, MCP manifests, audit logs, system prompt forensics, supply chain integrity, and full anomaly analysis.	Full prompt log forensics · Complete MCP and plugin audit · Audit log assessment · Config and system prompt forensics · Supply chain review · AI IOC development	Full forensic report · Risk-ranked findings · AI-specific IOC catalog · Supply chain assessment · Executive brief · Technical remediation roadmap	CISOs and security teams requiring a complete forensic baseline of the AI environment
AI Security Retainer $60,000 – $100,000 / year *Ongoing*	Ongoing forensic monitoring and periodic full assessments — continuous log review, quarterly MCP manifest audits, IOC refresh, and ad-hoc investigation support for AI-related security events.	Continuous log monitoring · Quarterly manifest audits · IOC maintenance and refresh · Ad-hoc AI IR support · Annual full assessment	Monthly threat summary · Quarterly audit reports · Updated IOC feeds · On-call IR support for AI incidents · Annual full assessment report	Organizations with mature AI programs requiring continuous forensic coverage

07

Competitive landscape

Competitor	Market focus / strengths	Limitations / Differentiation opportunity for Gruve
Palo Alto / Prisma AIRS + Unit 42	Runtime AI monitoring platform with integrated red team consulting. Strong infrastructure-layer coverage and enterprise reach.	Runtime monitoring and prevention detects future threats. AI-01 investigates past events — what has already occurred in the prompt layer, tool call history, and configuration state. Complementary, not competing.
Lakera	Leading inline runtime protection focused on real-time prompt injection filtering in live AI applications.	Excellent at preventing future prompt injection. Provides no retroactive log forensics, MCP manifest auditing, or supply chain integrity review. AI-01 addresses what Lakera cannot see: historical evidence of what has already happened.
HiddenLayer	Broad AI security platform with model scanning, guardrail tooling, and red teaming capability.	Platform-led, forward-looking threat management. Gruve differentiates with a forensic investigation methodology — building an evidence record from the deployed environment rather than adding monitoring tooling on top of it.
Internal SecOps / AppSec	Deep familiarity with the environment and existing security infrastructure.	Internal teams typically lack AI-specific forensic methodology, MCP expertise, and structured prompt log analysis capability. No established process for AI-specific IOC development, supply chain provenance review, or AI audit log forensics.

08

Gruve differentiators

- **Built for organizations that need an AI security investigation,** not another AI security product demo or governance framework.
- **Combines AI-native understanding** of how prompt layers, MCP toolchains, and agentic workflows behave with DFIR discipline: evidence, reproducibility, and clear investigator ownership of conclusions.
- **Covers the full forensic surface:** prompt logs, MCP manifests, audit trail integrity, system prompt history, and supply chain provenance — in one engagement.
- **Uses AI as an evidence organization accelerator only** — never as an autonomous source of forensic findings or severity determinations.
- **Produces outputs that serve** security operations (IOCs), engineering (remediation roadmap), legal (defensible investigation record), and executive stakeholders (risk-ranked brief) from one engagement.

09

Case study

A mid-market financial services firm had deployed AI-powered internal tools and MCP-enabled workflows across three business units over 18 months. No formal security review of the AI environment had been conducted. Prompt logs existed but had never been analyzed. MCP server configurations had been deployed by individual development teams with no central oversight. The CISO, following an unrelated security incident, asked whether the AI environment needed to be included in the investigation scope — and whether there was any forensic basis to answer that question.

Challenges

The AI environment had no security owner and no documented baseline. Prompt interaction logs were stored in three different systems with inconsistent retention policies — some going back six months, others overwritten after 30 days. Six MCP servers had been deployed by development teams over 18 months, none of which had been formally audited for capability scope or permission configuration. Two external plugin integrations had been sourced from third-party repositories without documented provenance review. The incident response team had no process for including AI artifacts in their investigation workflow.

Investigation

- **Prompt log forensics**: Gruve reviewed six months of available interaction logs across all three systems using structured log analysis methodology, clustering anomalous prompt patterns and flagging systematic probing activity consistent with reconnaissance.
- **MCP manifest audit:** All six deployed MCP server configurations were reviewed for capability scope, permission settings, version history, and supply chain provenance. Two external plugin integrations were traced to their source repositories and assessed.
- **Audit log assessment:** Log completeness, retention gap analysis, and tamper-evidence evaluation was conducted across all available AI audit records, producing a formal assessment of what the evidentiary record could and could not support.
- **Config and system prompt forensics:** System prompt change history and guardrail configuration records were reviewed for unauthorized modifications and governance drift.

Key results

- Identified three MCP server configurations with significantly overprivileged tool access — one with write-capable database tool access that exceeded stated business requirements.
- Found evidence of systematic prompt probing in 90-day interaction logs consistent with reconnaissance activity — patterns not flagged by any existing monitoring control.
- Discovered two audit log retention gaps totaling 47 days of missing interaction history that would have obscured evidence of any activity occurring during those periods.
- Assessed both external plugins and found one sourced from a repository with unverified maintainer identity and no code review record — flagged for immediate remediation.
- Delivered 14 AI-specific IOC signatures formatted for SIEM integration, enabling ongoing detection coverage for the identified prompt and tool abuse patterns.
- Produced a prioritized remediation roadmap and executive brief that allowed the CISO to brief the board with a forensic evidence basis — not a posture assertion.

Investigate your AI environment before an incident forces the question.

See how Gruve's AI Compromise Assessment applies DFIR-grade investigation methodology to the prompt layer, MCP toolchain, audit trail, and supply chain of your deployed AI environment — and delivers forensic findings your security, legal, and executive teams can act on.

Website: https://www.gruve.ai/

Email: contact@gruve.ai

Scoping calls complimentary.