

Artificial Intelligence Security Forensics: Standard Operating Procedure (SOP)

1. Purpose

To provide documented guidance on how Netskope Security Incident Response team handles forensic investigation and evidence acquisition processes during a security incident involving AI models, Large Language Models (LLMs), and agentic AI systems.

Critical note: All forensic evidence involving AI-generated code or reasoning must be verified by a human analyst. Never use a secondary LLM as the sole forensic auditor for a primary LLM's breach.

2. Scope

The SOP applies to all security incidents involving AI architectures at Netskope, including but not limited to:

- Autonomous / agentic AI workflows (Model Context Protocol - MCP)
- Retrieval-Augmented Generation (RAG) knowledge bases
- Internal and third-party LLM integrations
- AI infrastructure (vector databases, inference gateways, GPU clusters)
- Internal AI model evaluation
- Use of employee laptops for local LLM models and agentic workflows
- Scenarios where Netskope is the target of an externally operated AI agent

Note on third-party/SaaS LLMs: Where Netskope does not control the underlying model weights or infrastructure, several validation steps in this SOP (e.g., weight-hash verification, GPU / VRAM analysis) will not be feasible.

3. Definitions (selected)

- **Logprobs:** Log-probabilities assigned by the model to each output token; used to detect anomalous or adversarial generation patterns.
- **Top-K:** The K most similar retrieved chunks returned by a RAG pipeline for a given query.
- **Perplexity:** A measure of how surprising a sequence of tokens is to a model; used to flag adversarial or unnatural prompts.
- **AI-SBOM:** AI software bill of materials, an inventory of a model's training data lineage, fine-tuning history, and dependent components/libraries/packages.
- **WORM:** "Write Once, Read Many" storage, used to preserve evidence immutably for chain-of-custody purposes.
- **Patient Zero (agentic context):** The first agent in a multi-agent chain whose output or action introduced the compromise.
- **Excessive agency:** An OWASP LLM Top 10 (LLM06) risk category where a system grants a model too much functionality, permission, or autonomy, such that a manipulated model can take harmful actions (the "confused deputy" pattern in agentic systems).
- **Denial of Wallet (DoW) / LLMjacking:** Financially motivated abuse of pay-per-use AI billing, typically via stolen API keys or hijacked credentials, to run unauthorized, high-cost inference or compute workloads (including GPU cryptojacking).
- **Model / training data extraction:** Systematic querying of a model to reconstruct its weights, behavior, or underlying training/fine-tuning data (membership inference, model theft).

4. Roles and responsibilities

- **Incident response:** Primary lead for the end-to-end lifecycle of evidence collection, investigative procedures, and the formalization of the concluding forensic documentation.
- **AI/ML engineering:** Responsible for furnishing model provenance artifacts (e.g., AI-SBOM, golden image hashes) and facilitating sandbox environments for deterministic seed replay analysis.

- **Legal / privacy:** Mandatory notification partners for breaches involving PII exfiltration, prohibited content violations, or the inter-organizational transfer of evidentiary materials.
- **Threat and vulnerability management (TVM):** Provides external context for infrastructure enrichment and supports attribution efforts when Netskope is the target of autonomous adversary operations.
- **SOC:** Continuous monitoring and first-line detection of AI-related anomalies, triages and escalates confirmed events to the incident response team with the initial alert artifacts preserved.
- **Model owners:** Accountable for each production model or agent; maintain and execute the kill switch (endpoint disable, credential/token revocation) on security team request and confirm shutdown has taken effect.

5. Procedure

The forensic process must be performed at four technical levels to identify the entry point and intent of the AI incident:

- **Model level:** Validating model integrity, weight hashes (provenance), and hyperparameters (seeds/temperature).
- **Agentic / logic level:** Investigating reasoning traces (chain of thought), multi-agent handoffs, agent memory state, and MCP tool call parameters.
- **RAG / data level:** Extracting retrieved Top-K chunks and auditing vector metadata for context poisoning.
- **Infrastructure / gateway level:** Analyzing GPU / VRAM utilization, log probabilities (logprobs), billing / usage metrics, and WAF / proxy logs.

Step 1: Evidence acquisition (the AI forensic evidence stack)

Immediately upon detection, the incident response team must preserve the following artifacts in an immutable storage bucket (e.g. S3) to maintain the legal chain of custody and prevent loss due to non-determinism.

Note: Within each tier, prioritize collection of volatile artifacts before persistent ones. In-memory agent state, active session context, and real-time inference logs are subject to immediate loss upon system restart or session termination and must be

captured first. Stored model weights, configuration files, and billing records are persistent and can be collected after volatile evidence is secured.

- **Temporal and interaction metadata (tier 1):** full session history (conversation thread ID and summary objects), prompt / response traces with exact timestamps, logic state (random seed, temperature, Top-P, system prompt hash and content), and statistical artifacts (logprobs, to detect non-human/adversarial token patterns).
- **Agentic and tool traces (tier 2):** MCP tool call logs (JSON-RPC requests / responses), reasoning steps (chain of thought logs), agent memory store snapshots (before / after diff, to detect persistent memory or context poisoning), agent configuration files (AGENT.md, SKILLS.md, HEARTBEAT.md, or equivalent agent-defined configuration, skill, and heartbeat files), and identity context (OAuth 2.0 tokens and API keys used for tool execution).
- **Knowledge and infrastructure data (tier 3):** RAG artifacts (Top-K chunks, similarity scores, source metadata), resource logs (GPU / VRAM utilization, token-to-latency ratios), and billing / usage logs (API key usage, cost anomalies, request volume by identity), and integrity artifacts (AI-SBOM, dependency / package manifests, and SHA-256 hash of production model weights).
- **Tier 4, external context:** Any external infrastructure identifiers surfaced during acquisition (IPs, domains, certificates referenced in tool-call targets or WAF logs) should be flagged for enrichment in step 3 before they go stale.

Step 2: Forensic investigation and validation

Perform the following technical workflows to determine the root cause:

- **Validation of prompt tampering:** Perform seed replay (re-running input in a sandbox with the original seed to check for determinism) and conduct invisible token scans for hidden unicode overrides or zero-width spaces.
- **Validation of MCP and tool hijacking / excessive agency:** Perform argument origin tracing to determine if the LLM invented tool parameters and verify the tool schema's `requires_user_federation/authorization` flag to identify elevation of privilege.
- **Validation of RAG and memory integration:** Scan retrieved chunks for indirect prompt injections (directives hidden in docs) and audit metadata provenance to see if source documents were modified immediately prior to the incident. Separately, run

a security differential on the agent's long-term memory or context store against its pre-incident baseline to detect persistent memory poisoning, distinct from a single poisoned retrieval.

- **Validation of model integrity (self-hosted models):** Verify that the SHA-256 hash of the model weights in production matches the golden image in the AI inventory to rule out model-swapping attacks.
- **Validation of model integrity (third-party/SaaS models):** Where weights are not accessible, validate integrity via API response metadata, model or version headers, vendor status and incident advisories, and vendor attestations such as SOC 2 reports or model cards, in place of direct weight hashing.
- **Validation of system prompt leakage:** Compare any system prompt content appearing in model output against the known system prompt hash; determine whether leaked instructions could be used to craft further attacks.
- **Validation of model / training data extraction:** Review query pattern and rate logs for a given account or API key for systematic, high-volume, or unusually structured querying consistent with extraction or membership-inference attempts.
- **Validation of resource abuse / Denial of Wallet:** Correlate billing and usage anomalies with the identity (API key, OAuth token) that generated them; check whether the credential was reported lost, exposed, or recently used from an anomalous location.
- **Validation of supply chain integrity:** Verify hashes of third-party packages, plugins, and model or dataset dependencies against known-good versions in the AI-SBOM; check CI/CD logs for unauthorized dependency changes.

Step 3: Advanced forensic analysis

For complex incidents, implement advanced statistical and lateral movement checks:

- **Perplexity analysis:** Calculate the perplexity of suspect prompts. High-perplexity inputs resulting in confident (low-perplexity) outputs are hallmarks of adversarial probing.
- **Multi-agent lineage:** Reconstruct the message flow across agents to identify the Patient Zero agent in a multi-agent chain.

- **Resource correlation:** Correlate WAF / proxy logs with GPU metrics. A 400% spike in compute latency without a proportional token increase suggests a model DoS (complexity bomb) attack.
- **Hallucination vs. adversarial determination:** Where the step 4 report question (malicious vs. hallucination) is ambiguous, cross-check the output against RAG source grounding and confidence scores; a confidently wrong output with no adversarial input pattern points toward misinformation rather than attack.
- **External threat intelligence and ground truth correlation:** Internal logs establish what happened inside the environment but not what an external target is. For any external IP, domain, or tool-call target surfaced during the incident, engage Netskope Threat Labs to gather first-party internet intelligence (open services / ports, certificates, passive DNS, co-hosted domains) and correlate it against threat intelligence labels and known vulnerability context, to assess lateral movement or known C2 infrastructure.

Step 4: Documentation and RCA

Document all actions taken with a focus on intent and entry point. The final forensic report must answer:

1. **Action taken:** (e.g., "Extracted Vector ID 445; verified SHA-256 of model weights; replayed seed 42.")
2. Was the intent malicious? (Human-driven attack vs. model hallucination).
3. What was the entry point? (Direct input vs. poisoned context vs. tool metadata vs. model weight tampering).
4. Was there lateral movement? (Did the agent move from a low-privilege sandbox to production systems?)
5. Was there financial or resource impact? (Anomalous billing, compute, or GPU usage attributable to the incident).
6. Did the output constitute prohibited content? (If yes, follow the prohibited-content handling process instead of standard distribution).
7. Does this incident involve another organization's infrastructure, credentials, or personnel? (If yes, engage legal for cross-organization coordination before any technical exchange with the other party).

6. Forensic artifacts checklist

INCIDENT TYPE	REQUIRED ARTIFACTS	FORENSIC SOURCE
Jailbreak / injection	Input logs, logprobs, seed, system prompt hash, session thread	LLM gateway / proxy logs
Tool abuse (MCP)	JSON-RPC payloads, OAuth tokens, tool schema, identity context	MCP server logs
Data leak (PII)	Response logs, DLP proxy alerts, RAG chunks, WORM hashing	CASB / DLP / storage
RAG poisoning	Vector IDs, similarity scores, source metadata, chunk extraction	Vector DB audit logs
Model integrity / DoS	Model weight SHA-256, GPU utilization, AI-SBOM, latency logs	GPU metrics / CI-CD
Agentic failure	Reasoning traces (CoT), iteration counts, multi-agent handoff logs	Orchestration framework logs
Tampering	Access logs	Invocation platform (OpenAI, Claude, AWS CloudTrail)
System prompt leakage	System prompt hash, full session history, prompt / response traces	LLM gateway / proxy logs
Memory / context poisoning	Agent memory store snapshots (before / after diff), RAG source metadata	Vector DB / agent memory store
Resource abuse / Denial of Wallet	Billing / usage logs (API key usage, cost anomalies), resource logs (GPU/VRAM, token-to-latency ratios)	Cloud billing / API gateway logs
Model / training data extraction	Prompt / response traces, logprobs, output similarity analysis	LLM gateway / proxy logs
Supply chain compromise	AI-SBOM, dependency/package manifests, vendor attestations (SOC 2, model cards)	AI-SBOM registry / vendor advisories
Misinformation / hallucination	RAG source grounding, confidence scores, logprobs	Vector DB audit logs / LLM gateway
Sandbox / eval breakout	Test environment access logs, container/VM escape indicators	AI test and eval environment logs
Agent hijacking	MCP tool call logs (JSON-RPC), OAuth 2.0 tokens, tool schema flags	MCP server logs

INCIDENT TYPE	REQUIRED ARTIFACTS	FORENSIC SOURCE
Lateral movement / C2	External IPs, domains, certificates surfaced during acquisition, WAF / proxy logs	Netskope Threat Labs / WAF logs
Prohibited / harmful content	Output logs, content classification labels, escalation records	CASB / DLP / storage

7. Retention standards (ISO 42001 alignment)

- **Hot retention (90 days):** Prompt / response pairs, tool-call logs, guardrail events.
- **Cold retention (one year):** Full traces and high-risk agent reasoning steps.
- **Critical incident logs:** three years for forensic and regulatory compliance.
- **Legal hold:** Suspend standard retention and deletion schedules for any data subject to a litigation hold or active legal/regulatory investigation until legal formally releases the hold.
- **Prohibited or harmful content exception:** Outputs classified as prohibited or harmful content follow the special handling process and are retained independent of the standard tiers above to support escalation, chain of custody, and regulatory reporting.

8. Special handling: prohibited or harmful content incidents

When an incident involves AI-generated content that is illegal, prohibited under Netskope's Acceptable Use Policy, or otherwise harmful (e.g., CSAM, extremist content, or weapons synthesis instructions), standard triage and distribution steps are suspended in favor of the following controls:

- **Immediate containment:** Isolate the specific session, model endpoint, or agent instance; do not forward the raw content beyond the minimum reviewers required for assessment. Enforce isolation at a layer the affected component cannot modify: revoke its credentials and tokens at the identity provider and block egress at the gateway or cloud IAM/network level. Do not rely on host-based firewalls or tool-permission settings inside the agent's own execution environment, as a compromised agent can reconfigure these. Confirm containment by verifying that no further tool calls, inference requests, or egress occur before proceeding.

- **Restricted distribution:** Route all evidence handling through legal and incident response leadership only; redact or hash sensitive payloads before referencing them in wider case documentation.
- **Mandatory escalation:** Notify legal / privacy and, where required by law, external authorities within the timelines set by Netskope's incident response policy.
- **Preservation without amplification:** Preserve WORM-hashed copies of the output for chain of custody, but avoid unnecessary duplication of the content across tooling, tickets, and messaging.

9. Cross-organization and third-party coordination

Many AI incidents span the boundary between Netskope-operated infrastructure and third-party model providers, MCP tool vendors, or partner organizations. Coordination must move quickly without creating additional exposure:

- **Vendor notification:** Engage the AI/ML vendor's security or trust and safety team when the incident involves their model, API, or hosted infrastructure; request logs, advisories, or model cards relevant to the incident.
- **Shared incident handling:** When an incident involves another organization's infrastructure, credentials, or personnel (see step 4, RCA question 7), engage legal before any technical exchange with the external party.
- **Information sharing boundaries:** Share only the minimum technical indicators (hashes, timestamps, IOCs) necessary for the other party to act; do not share proprietary prompts, customer data, or internal tooling details.
- **Netskope Threat Labs liaison:** Route external threat intelligence requests and cross-organization correlation through Netskope Threat Labs to maintain a single point of contact.

10. Victim-side scenarios: identifying AI-driven attackers

Netskope may be the target rather than the source of an AI-driven attack (e.g., an external AI agent probing Netskope's APIs, an AI-generated phishing campaign impersonating Netskope, or automated credential-stuffing driven by an LLM agent). These scenarios require attacker-side, rather than internal-model, forensics:

- **Behavioral fingerprinting:** Look for statistical hallmarks of AI-driven activity, such as unnaturally consistent timing between requests, high grammatical uniformity in social-engineering content, or systematic exploration patterns, to distinguish automated agents from human attackers.
- **External attribution:** Engage Netskope Threat Labs to correlate observed infrastructure (IPs, domains, certificates) against known AI-driven attack tooling and threat actor campaigns.
- **Customer / partner impersonation:** When AI-generated content impersonates Netskope (deepfake media, cloned voice, spoofed communications), coordinate with legal and marketing / communications in addition to incident response.
- **Defensive tuning:** Feed confirmed indicators back into NewEdge/CASB/DLP and SWG detection rules to harden against repeat automated attacks.

11. Privacy and regulatory escalation

AI incidents involving personal data, biometric data, or cross-border data flows may trigger notification obligations. This section aligns forensic handling with Netskope's existing privacy escalation path rather than introducing a new one:

- **Privacy assessment trigger:** Legal / privacy must assess every confirmed AI incident involving PII, PHI, or regulated data categories (per section 4, roles and responsibilities) to determine notification obligations.
- **Data subject impact:** Document which data subjects, records, or categories of personal data were exposed, drawing on the forensic artifacts checklist (data leak (PII) row) as the starting evidence set.
- **Regulatory timelines:** Where applicable regulations impose notification deadlines, legal / privacy drives the timeline; forensic evidence must be preserved and summarized quickly enough to support that timeline, independent of the standard investigation cadence.
- **Cross-border considerations:** For incidents involving data processed or stored outside the customer's or Netskope's primary jurisdiction, flag this explicitly to legal / privacy as it may affect which regulatory regime applies.

12. Chain of custody and evidence integrity

All evidence gathered under this SOP must be defensible in legal, regulatory, or customer-facing contexts. Chain of custody failures can undermine an otherwise sound investigation:

- **WORM storage:** All evidence identified in step 1 (evidence acquisition) must be written to a dedicated forensic evidence store before any analysis begins, with a recorded hash at time of capture. Regardless of cloud platform, the store must enforce: write-once retention locked for the period defined in section 7, in a mode that administrators cannot shorten or remove (e.g., compliance-mode object lock or equivalent); object versioning so no write can silently replace a prior copy; separation from the production account or project, with no delete path from the environment under investigation; and encryption at rest with independent access logging. Immutability alone is not sufficient; the control must be one that a compromised administrator or agent cannot reverse.
- **Access logging:** Every access to preserved evidence (read, export, or analysis) must be logged with investigator identity and timestamp, maintained as part of the case record.